UNIVERSIDAD AUTÓNOMA METROPOLITANA
Unidad Iztapalapa
Casa abierta al tiempo

Evaluación de la capacidad discriminativa de las máquinas restringidas de Boltzmann

Tesis realizada como requisito para obtener el grado de
Doctor en Ciencias (Ciencias y Tecnologías de la Información)

Máximo Eduardo Sánchez Gutiérrez

Posgrado en Ciencias y Tecnologías de la Información
Ciencias Básicas e Ingeniería

Octubre 2018

Dirigida por .....

Dr. John Goddard
Director
Universidad Autónoma Metropolitana–Iztapalapa
Ciudad de México, México

Dr. Enrique Marcelo Albornoz
Codirector
Instituto sinc(i), Universidad Nacional del Litoral/CONICET
Santa Fe, Argentina

Esta Tesis fue defendida en una disertación pública en la
UAM-Iztapalapa. Los miembros del jurado calificador fueron:

Presidente .....

Dra. María del Carmen Gómez Fuentes,
Universidad Autónoma Metropolitana–Cuajimalpa
Ciudad de México, México

Secretario .....

Dr. Pedro Pablo González Pérez,
Universidad Autónoma Metropolitana–Cuajimalpa
Ciudad de México, México

Vocal .....

Dr. Jorge Cervantes Ojeda,
Universidad Autónoma Metropolitana–Cuajimalpa
Ciudad de México, México

Vocal .....

Dr. Enrique Marcelo Albornoz,
Instituto sinc(i), Universidad Nacional del Litoral/CONICET
Santa Fe, Argentina

Vocal .....

Dr. César Martínez,
Instituto sinc(i), Universidad Nacional del Litoral/CONICET
Santa Fe, Argentina

Evaluación de la capacidad discriminativa de las máquinas restringidas de Boltzmann

por

Máximo Eduardo Sánchez Gutiérrez

Una tesis doctoral presentada para obtener el grado de
Doctor en Ciencias (Ciencias y Tecnologías de la Información)

Resumen

Durante la primera década del 2000, el reconocimiento de dígitos escritos a mano fue una de las primeras aplicaciones más notables utilizando redes neuronales profundas. Posteriormente, a inicios de la década de 2010, estas mismas redes mostraron resultados prometedores en múltiples áreas de aplicación, incluyendo el reconocimiento de emociones en el habla. En este trabajo se hace énfasis en la aplicación de las redes neuronales a problemas de clasificación que incorporan la señal de habla debido a su importancia en el proceso de comunicación entre personas. Este tipo de señal se caracteriza por su alta variabilidad temporal, pues su producción queda condicionada por la morfología y movimiento de los elementos en la cavidad oral y el rostro, y puede verse influenciada por el acento regional, la condición social o el estilo personal, entre otros. La expresión de emociones es otro de los elementos que enriquecen la comunicación humana, incluso se ha comprobado que las palabras por sí mismas no aportan el significado completo del mensaje para un escucha. En este sentido, el análisis de los componentes paralingüísticos como la prosodia, la calidad de la voz, el ritmo e incluso el estado emocional con el cual se expresa un mensaje son muy importantes y deben tenerse en cuenta en sistemas que deben interactuar con las personas. Consecuentemente, en esta interacción, el sistema debe ser capaz de interpretar esta información complementaria a las palabras, por ejemplo realizando la detección/clasificación de las emociones del hablante. Gracias a su versatilidad y a los resultados mostrados, las máquinas restringidas de Boltzmann han sido empleadas como principal bloque de construcción en el aprendizaje profundo, provocando que el interés de diversos grupos de investigación se viera dirigido hacia su perfeccionamiento.

Con esto en mente, en esta Tesis se abordan las redes neuronales, en particular, de las máquinas restringidas de Boltzmann en un esquema multi–capa con el objetivo de mejorar su arquitectura mediante la evaluación y poda de aquellas neuronas menos útiles para el proceso de clasificación. Este tema ha cobrado mucha importancia en épocas recientes debido al surgimiento de un gran número de algoritmos de aprendizaje profundo para diversas aplicaciones, con especial atención en aquellas que funcionan en dispositivos móviles pues existe la necesidad de mejorar estas arquitecturas profundas mediante la reducción de variables, lo que permitiría aminorar

los costos de implementación y de procesamiento.

En el presente trabajo se presenta una metodología que permite mejorar los resultados obtenidos mediante el aprendizaje profundo y las máquinas restringidas de Boltzmann en un esquema de clasificación. En una primera etapa, se evaluaron múltiples arquitecturas en tareas de clasificación y luego se desarrolló un método para evaluar la importancia relativa de cada neurona de estas redes. Los resultados muestran que utilizar las máquinas restringidas de Boltzmann en una arquitectura profunda mejora las tasas de error, y además, se ha descubierto que podar aquellas neuronas que contribuyen menos a la solución de la tarea de clasificación, produce redes con arquitecturas menos densas sin sacrificar la capacidad de generalización y, en la mayoría de los casos, mejora las tasas de error obtenidas con las técnicas tradicionales. Cada una de estas propuestas de evaluación y poda precisaron la codificación de simulaciones que permitieran determinar su eficacia. Esta experimentación fue dividida en cuatro partes, la primera de ellas investigó la pertinencia de utilizar el aprendizaje profundo para el reconocimiento de emociones en la voz, esta tarea se acotó a dos emociones. La segunda también abordó el reconocimiento de emociones en la voz, aunque en esta ocasión investigando siete emociones. En la tercera parte se investigó la poda de las neuronas evaluadas como menos discriminantes empleando dos idiomas y dos emociones, la evaluación se llevó a cabo utilizando cinco medidas de disimilitud. En última instancia se investigó el uso de las técnicas propuestas de poda para las máquinas restringidas de Boltzmann, en otras bases de datos no relacionadas con la señal del habla, para ésto se utilizaron cinco bases de datos y hasta siete clases. En todos los casos se obtuvieron resultados favorables con propuestas innovadoras.

Director: Dr. John Goddard

Codirector: Dr. Enrique Marcelo Albornoz

There is no such thing as a "self-made man". We are made up of thousands of others. Everyone who has ever done a kind deed for us, or spoken one word of encouragement to us, has entered into the makeup of our character and of our thoughts, as well as our success.

—*George Matthew Adams*

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1

PRESENTACIÓN

Cada vez con más frecuencia las personas se comunican con las máquinas. Si el interlocutor es otro humano o una máquina, el problema de transmitir el mensaje correcto y tener la interpretación correcta por parte del interlocutor es de suma importancia. La computación afectiva tiene como objetivo mejorar la interacción humano–computadora permitiendo a las computadoras adaptar sus respuestas de acuerdo a las necesidades humanas. Así pues, uno de los objetivos es investigar cómo reconocer las emociones mediante la captura de señales biométricas obtenidas de las personas.

Por naturaleza, los seres humanos utilizan todos sus sentidos para la percepción e interpretación de los mensajes involucrados en la comunicación. Escuchan el sonido, leen labios, interpretan gestos y expresiones faciales e interpretan la semántica de las expresiones. Usando todos estos sentidos y habilidades, una persona puede percibir una frase como divertida, puede entender la ironía y el sarcasmo, y puede, en consecuencia, reaccionar de manera apropiada. Consecuentemente, las personas perciben el estado emocional del interlocutor y, por lo tanto, son capaces de adaptarse a él. A medida que la interacción humano–computadora (HCI) evoluciona, algunos esfuerzos se orientan a permitir a las computadoras reconocer las emociones humanas y reaccionar en consecuencia. Aún cuando la detección de emociones es un proceso implícito para los seres humanos, es una tarea muy difícil para las computadoras. Se han logrado

identificar más de 300 emociones humanas [1, 2], algunas de ellas se exhiben en la Tabla 1.1:

Aburrido	Ambivalente	Atónito	Desobediente	Impresionado
Aceptar	Animado	Aventurero	Disgustado	Insatisfecho
Afectuoso	Antagónico	Avergonzado	Disgustado	Molesto
Agradable	Apático	Calmado	Divertido	Perplejo
Agresivo	Aprehensivo	Cauteloso	Enojado	Presumido
Alegre	Asombrado	Desconfiado	Estupefacto	Previsor
Amargado	Atento	Desinteresado	Eufórico	Vacío

Tabla 1.1: *Ejemplos de emociones identificadas por J.D. O'Connor y Gordon Frederick Arnold en [2], 1973*

Incluso después de varias décadas de investigación, diversos aspectos necesitan ser cuidadosamente formulados ya que representan retos importantes para la correcta clasificación de las emociones. Tal es el caso de las peculiaridades del habla, donde es importante analizar la duración de las palabras, la cadencia, el volumen, la velocidad, etc. La variabilidad entre oraciones, estilos de habla, e incluso la variabilidad entre individuos, es otro reto que debe ser tomado en cuenta. Como sabemos, existen diferentes formas en las que se expresan las personas, esto se debe entre otras cosas a las distintas regiones donde radican los hablantes y sus estratos sociales, lo que indica que éstos expresan emociones de acuerdo con su cultura y medio ambiente [3]. Además, nuestro estado emocional no es constante ni periódico y puede variar en cualquier sentido, lo que impacta en la comunicación. Por estas razones se ha sugerido que las palabras por sí mismas no aportan el significado completo del mensaje para un escucha, consecuentemente el análisis de los componentes paralingüísticos como la prosodia, la calidad de la voz, el ritmo e incluso las emociones con las que son dichas las palabras se ha vuelto importante [4].

Merece la pena dejar claro que las emociones no tienen una definición teórica aceptada por todos y por ende, existen diferentes modelos que intentan conceptualizar y explicar las emociones. Lo que sí sabemos sobre las emociones es que se pueden medir, por ejemplo, en el tono de la voz, en su energía, en la duración de las distintas estructuras gramaticales, etc. La señal acústica generada para el mismo enunciado o

frase cambia principalmente debido a cambios físicos provocados por las emociones. El propósito principal de un algoritmo de reconocimiento de emociones en el habla es detectar el estado emocional de un hablante a partir de señales de voz. Haciendo uso de estos conceptos, podemos explicar la idea de 'emoción en el habla' como un estado, en donde los parámetros medidos pueden ser evaluados por un sistema automático que permita su clasificación.

1.1. Motivación

El reconocimiento de emociones en el habla es útil para aplicaciones que se benefician de la interacción hombre–máquina, ejemplos de esto son: la evaluación de enfermedades como el Parkinson donde el estado neurológico de los pacientes puede estimarse por medio del deterioro en la producción del habla, en la evaluación sobre si un hablante está comiendo con el objetivo de adaptar el reconocimiento de voz a texto [5], en la detección de mentiras y evaluación de la sinceridad que se han asociado con manifestaciones de miedo y euforia y, en general, con el comportamiento verbal y no verbal, en las tareas de identificación de idiomas e identificación de dialectos o acentos [6], en determinar el destinatario del mensaje como en la interacción entre un adulto y un menor, en la evaluación de la calidad del habla bajo condiciones externas como el frío o incluso, internas como la salud, por ejemplo el análisis del ronquido para sugerir un tratamiento médico dirigido [7] o incluso, en la generación de voces que tienen un énfasis emocional adecuado en la industria de los juguetes y videojuegos [8]. Con todas estas áreas vinculadas al reconocimiento de emociones, resulta de gran importancia abordar el tema haciendo uso de técnicas novedosas que han mostrado ser efectivas, como las redes neuronales que, en conjunto con el aprendizaje profundo, han logrado producir mejores resultados que otras no profundas como clasificadores lineales o máquinas de soporte vectorial [9, 10].

En un estudio anterior [11] se consideró la aplicación de las máquinas restringidas de Boltzmann (RBM) y redes de creencia profunda (DBN) en la tarea de reconoci-

miento automático de las emociones en el habla en español. Esto permitió obtener resultados comparables, y en los casos explorados, mejores que los resultados de otros clasificadores cuando los parámetros son optimizados correctamente. A pesar de ello, debido a que el uso de las DBNs y RBMs es relativamente nuevo y aún más en el área del reconocimiento de emociones, existen pocas aproximaciones a la interpretación del funcionamiento de las capas de la red [12]. Es por esto que, como resultado de la experimentación realizada en ese trabajo, surgieron algunas ideas que son desarrolladas detalladamente en esta Tesis.

1.2. Contribución

En este trabajo nos centramos en el entrenamiento del modelo que, en la Figura 1-1, está dentro del subsistema de reconocimiento maquinal. En este punto analizamos la utilidad de podar una red neuronal, en particular una máquina restringida de Boltzmann (RBM) [13, 14, 15, 16, 17]. Esta poda puede verse como una reducción de dimensión o incluso, una selección de características [18, 19, 20]. Estas formas alternativas y la forma de evaluarlas son posteriormente analizadas. Uno de los principales desafíos en el área de redes neuronales artificiales, y que se aborda en este trabajo, es la identificación de una arquitectura adecuada para un problema específico, pues la elección de una topología inadecuada puede aumentar exponencialmente el coste de entrenamiento e incluso obstaculizar la convergencia de red [21, 22, 23, 24]. Por otro lado, investigaciones recientes indican que redes más grandes o más profundas pueden mapear las características del problema en un espacio más apropiado, y por lo tanto mejorar el proceso de clasificación [25, 26, 27, 28, 29]. En este sentido, es interesante investigar si existen medidas de similitud que proporcionen una pista para encontrar las neuronas más discriminativas en una red. En el presente trabajo, exploramos esta cuestión empleando diferentes medidas para realizar la poda posterior al entrenamiento. De las neuronas determinadas como las más discriminativas, se obtienen las activaciones que sirven de entrada a un clasificador con el objetivo de determinar la factibilidad de poda. Encontramos que dos medidas en particular parecen ser buenos

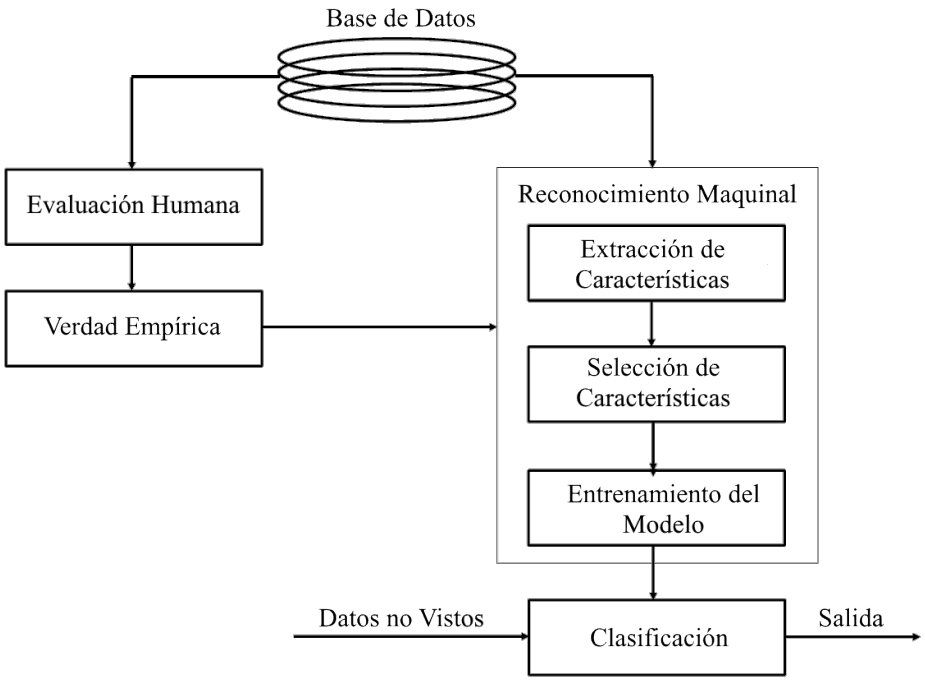


Figura 1-1: *Sistema de Reconocimiento de Emociones en el Habla, adaptada de Ali Hassan [30], 2012*

indicadores de las neuronas más discriminativas, produciendo ahorros de más del 50 % de las neuronas, manteniendo al mismo tiempo una tasa de error aceptable. Además, demostramos que comenzar con una arquitectura de red más grande y luego podar es más ventajoso que usar desde el inicio una red más pequeña. Finalmente, se introduce un índice cuantitativo que puede proporcionar información sobre la elección del tamaño de neuronas a podar.

El esquema conceptual del proceso de experimentación y poda se puede ver en la Figura 1-2, el primer paso consiste en extraer características de la señal de voz. En el segundo paso, un RBM se entrena de manera no supervisada con un conjunto de entrenamiento dado. Luego, ese conjunto de entrenamiento se propaga a través de la RBM para obtener las activaciones de las unidades ocultas y, posteriormente, cada unidad oculta es evaluada para podar la RBM según las medidas discriminativas.

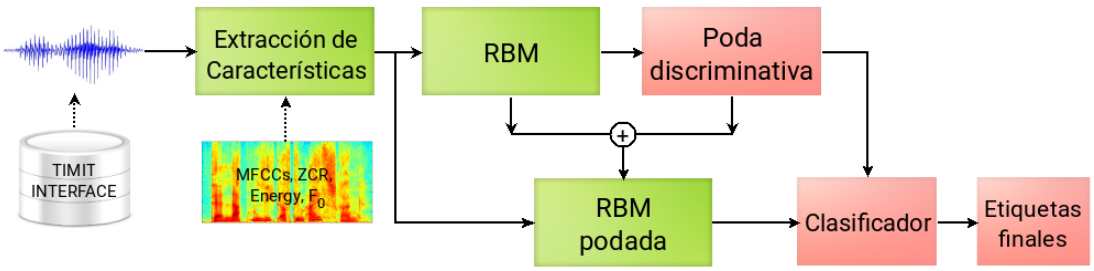


Figura 1-2: *Esquema conceptual del proceso general de los experimentos de poda.*

1.3. Estructura del documento

Los capítulos de este trabajo están organizados en tres grandes partes. La primera comprende el estudio de las emociones en el habla, en el Capítulo 2 se describe el modelo de producción del habla, los aspectos biológicos y fisiológicos, las "seis grandes emociones" o emociones prototípicas y el efecto que tienen en la producción del habla así como las características de su señal.

La segunda parte incluye las propuestas realizadas para mejorar las arquitecturas de las redes neuronales, en particular de las máquinas restringidas de Boltzmann (RBM). La descripción general de las RBMs y el aprendizaje profundo se proporciona en el Capítulo 3. En el Capítulo 4 se presenta la poda neuronal discriminativa a través de distintas medidas de disimilitud binarias y multiclase.

En la tercera parte se presenta la experimentación de los modelos propuestos previamente para optimizar las arquitecturas tradicionales en un esquema de clasificación. En el Capítulo 5 se muestra la aplicación de las máquinas restringidas de Boltzmann en la tarea de clasificación binaria y multiclase, así como los conjuntos de datos utilizados y las características extraídas de ellos para su análisis. En el Capítulo 6 se extiende la experimentación para evaluar los efectos de la poda discriminativa en tareas de clasificación binaria y multiclase en actividades que involucran tanto datos de habla emocional, como datos pertenecientes a otras áreas de conocimiento. Finalmente, en el Capítulo 7 se presentan las conclusiones y perspectivas de futuras investigaciones.

CAPÍTULO 2

LA SEÑAL DEL HABLA Y LAS EMOCIONES

Cada vez con más frecuencia las personas se comunican con las máquinas y, ya sea que el interlocutor sea otro humano o una máquina, el problema de transmitir el mensaje correcto y obtener la interpretación correcta es de suma importancia. Las personas perciben el estado emocional del interlocutor a través de todos los sentidos, debido a esto son capaces de adaptarse a él. A medida que la interacción humano–computadora (HCI) evoluciona, la importancia de permitir a las computadoras reconocer las emociones humanas y reaccionar en consecuencia ha ido en aumento. A pesar de que la detección de emociones es algo natural para los seres humanos, resulta una tarea muy difícil para las computadoras. El propósito ulterior de los sistemas de reconocimiento de emociones (ERS) es la aplicación del conocimiento relacionado con las emociones de tal forma que la comunicación entre las computadoras y los humanos sea más satisfactoria. Sin duda, la mayoría de la gente ha experimentado muchas veces la frustración causada por la mala interacción con los sistemas automatizados. Al permitir que las computadoras reaccionen al estado emocional del usuario, esta interacción puede transformarse en una experiencia más productiva. Sin embargo, mejorar la interacción con las máquinas no es la única aplicación del reconocimiento emocional, los sistemas especializados pueden ser utilizados para problemas aún más serios como detectar la agresividad, el estrés o la frustración.

2.1. Las características de la señal del habla

Para producir el sonido del habla es necesario generar presión de aire. Estos sonidos se pueden clasificar según la fuente de presión de aire utilizada para producirlos como [31]: (1) Pulmonares, donde la presión del aire es generada por los pulmones, (2) Velares, donde la presión de aire se genera cerrando el tracto bucal, levantando la parte posterior de la lengua contra el velo y (3) Glotales, donde la presión del aire se genera al cerrar el tracto bucal en la glotis. Otra forma de clasificar el sonido del habla, y la que podría ser las más utilizada, es la lograda a través de la vibración de los pliegues vocales, de esta forma se puede clasificar el sonido como sordo y sonoro, donde en este último intervienen las cuerdas vocales al tensarse generando el sonido que será filtrado por el resto del aparato fonador. Por lo tanto, los sonidos sordos se generan al obstruir el flujo de aire y relajar las cuerdas vocales [32].

Consecuentemente, el sonido del habla es producido por la acción coordinada de tres sistemas fisiológicos [32]: el sistema respiratorio, el sistema vocal o fonatorio y el sistema de resonancia. El sistema respiratorio está compuesto por los pulmones, la tráquea, la caja torácica y el diafragma. En conjunto, las fuerzas ejercidas por la presión del aire dentro de los pulmones y las ejercidas por los músculos inspiratorios y espiratorios son reguladas por el sistema respiratorio que, a su vez, impulsa al sistema de fonación. El sistema de fonación está compuesto principalmente por: la laringe, los pliegues vocales y la glotis. Durante la respiración normal, los pliegues vocales se encuentran separados permitiendo que el aire fluya libremente a través de la glotis mientras que, durante la producción del habla, los pliegues vocales se acercan y tensan obstruyendo el aire haciendo que la presión de aire se acumule debajo de los pliegues vocales, eventualmente forzándolos a separarse. A medida que el aire comienza a fluir a través de la glotis, la presión del aire entre los pliegues vocales cae, haciendo que éstos se cierren. El resultado es una fluctuación periódica de la presión de aire, esta fluctuación produce un sonido que, entre otras características, tiene frecuencia base y armónicos. Los armónicos son multiplos de la frecuencia base que es frecuentemente denominada frecuencia fundamental (F_0). Cualquier cambio en la presión del aire

debido, por ejemplo, a un cambio en la función respiratoria o la tensión y posición de los pliegues vocales, afecta la forma en que los pliegues vocales se abren y cierran. Estos cambios producen variaciones en la intensidad de F_0 y por ende, los armónicos del sonido. Por ejemplo, cuando el esfuerzo espiratorio es intenso, los pliegues vocales se cierran rápidamente, lo que produce un aumento no sólo en la intensidad global, sino también de F_0 y la energía de los armónicos [33, 34, 35].

Por otro lado, el sistema de resonancia que comprende el resto del tracto vocal que se extiende desde la glotis, a través de la faringe y hasta las cavidades oral y nasal, puede o no, obstruir el flujo de aire y participar en el filtrado del sonido. La forma y longitud de los articuladores como la lengua, el velo, los dientes y los labios, determinan cómo se amplifican o atenúan ciertos armónicos. Existen patrones específicos de armónicos atenuados y amplificados llamados formantes, ellos corresponden a las distintas vocales y consonantes vocalizadas en el lenguaje hablado (para un tratamiento más completo de la producción del habla, véase [36, 37]). De esta última etapa de la producción del habla se tiene un control voluntario mayor que de las etapas anteriores, aunque es todavía susceptible a la perturbación involuntaria. Por ejemplo, debido a la ansiedad a muchas personas se les seca la boca cuando hablan en público, lo que se podría esperar afecte el sonido de la voz, aunque estos efectos aún no son bien conocidos.

En principio, la codificación de la emoción en la voz puede medirse en cualquiera de una serie de etapas, desde los cambios fisiológicos en varias partes del sistema de producción vocal, hasta la percepción del interlocutor sobre la calidad del habla. Decidir cuál de estas etapas medir depende del área de aplicación y objetivos específicos, para el caso de las emociones en el habla, evaluar las propiedades físicas del sonido es lo más apropiado. Las técnicas de medición y análisis acústico están bien desarrolladas, pueden aplicarse utilizando equipos relativamente baratos y fácilmente disponibles. Aunque las relaciones entre la fisiología de la producción del habla y la señal acústica resultante son poco claras, existe un incremento constante en la literatura sobre el tema (ver [38] para una visión general de la percepción del habla).

El análisis acústico permite así, relacionar la codificación emocional del habla con la decodificación de la misma.

Dado que el habla es una señal dinámica y rápidamente cambiante, la mayoría de las variables acústicas se obtienen de segmentos de habla sobre los cuales se puede asumir que la señal es relativamente estable, estos segmentos podrían corresponderse con fonemas, sílabas o incluso segmentos más largos. La evaluación de los segmentos cortos del habla está más enfocada al contenido fonético, por otro lado, la modulación del habla emocional se espera que sea en gran medida suprasegmental pues los cambios fisiológicos que se creen sustentan el habla emocional son relativamente lentos. Por este motivo, la mayor parte de la investigación sobre el habla emocional ha incluido las variables acústicas a corto plazo en marcos de tiempo más largos para obtener medidas suprasegmentales.

Las características del habla a continuación descritas se han subdividido en tres categorías: prosódicas, espectrales y de calidad. La siguiente sección presenta una breve introducción a las tres categorías acústicas y sus asociaciones con la producción y percepción del habla; muchos textos ofrecen un tratamiento más detallado, por ejemplo, [38, 39].

Características prosódicas

La señal de voz consiste en una secuencia temporal de diferentes tipos de sonidos correspondientes a vocales, consonantes e interjecciones además de silencios, todos los cuales pueden ser portadores de información afectiva, sin embargo, no existen reglas establecidas sobre qué tamaño de segmentos de la señal del habla son los más apropiados para el análisis emocional. Las características que se espera varíen según las diferentes emociones expresadas y, que al mismo tiempo son utilizadas con mayor frecuencia están relacionadas con el tiempo: la velocidad, la duración de los sonidos y las pausas.

La medida acústica que se encuentra más relaciona con la percepción de la sonoridad del habla es la intensidad. La intensidad es una medida de la potencia de una señal de voz, y refleja tanto el sonido producido en la glotis, como la amplificación y atenuación de los armónicos en el tracto vocal. Por tanto, varía en función del esfuerzo vocal y de la resonancia del tracto, los cuales pueden ser alterados fisiológicamente por los efectos de las emociones. Por ejemplo, una mayor fuerza espiratoria, tal vez correspondiente a la preparación para una respuesta de lucha, tenderá a aumentar la intensidad del habla aumentando la intensidad del sonido producido en la glotis. La falta de saliva correspondiente al miedo conducirá a una disminución de las amplitudes de los formantes, disminuyendo así la intensidad del habla. La dependencia de la intensidad del habla en estos dos mecanismos independientes implica que la interpretación de los cambios de intensidad del habla con la emoción, en términos de los mecanismos subyacentes, sea difícil. Al examinar la intensidad del habla en combinación con otras características, debería ser posible comprender mejor la fuente de los cambios de intensidad.

Una de las características más utilizadas de la señal de la voz es la frecuencia fundamental (F_0), ésta se mide en ciclos por segundo o Hertz (Hz) y mide la velocidad a la cual los pliegues vocales se abren y cierran a través de la glotis, también ayuda a determinar el tono percibido de la voz. Esta frecuencia varía continuamente en función de los aspectos lingüísticos y paralingüísticos del habla, así como de cambios fisiológicos. A lo largo de una frase o enunciado, la F_0 se puede cuantificar en términos de su nivel promedio, mínimo, máximo y su variación en el tiempo. El piso o nivel de reposo de F_0 está determinado principalmente por el tono muscular de la laringe, en particular de los pliegues vocales. Por lo tanto, se puede esperar que cualquier cambio en el tono muscular laríngeo, producido como parte de una respuesta emocional, tendrá un impacto significativo en el valor de F_0. Para estimar esta característica, se toma como base la ventana:

$$f_s(n;m) = s(n)w(m-n) \tag{2.1}$$

donde $s(n)$ es la señal de voz y $w(m-n)$ es la ventana de longitud N_w terminando en la muestra m.

Frecuencia fundamental

Aunque existen múltiples métodos para calcular la frecuencia fundamental, probablemente el más utilizado es el método del espectro donde se toma una porción periódica de la señal de la que se mide el periodo T_0, teniendo en cuenta que la frecuencia fundamental se encuentra en $F_0 = \frac{1}{T_0}$. Es por esto que para encontrar F_0 con este método, es necesario que la señal sea periódica. Así pues, un periodo es la unidad de repetición mínima en la señal. Esto quiere decir que para los múltiplos de un periodo T el valor de la señal es el mismo por lo que $F_0 = \frac{1}{T_0}$ se cumple cuando $x(t) = x(t+T) = x(t+2T) = x(t+3T) = \ldots$ donde $x(t)$ es la onda.

Cruces por Cero

La tasa de cruces por cero (ZCR), es la tasa de cambios de signo en una señal, es decir, la tasa con la que la señal cambia de positivo a negativo o viceversa y está definida como:

$$zcr = \frac{1}{T-1}\sum_{t=0}^{T-1} A(s_t \cdot s_{t-1}) \; donde$$

$$A(s_t \cdot s_{t-1}) = \begin{cases} 1 & si \; (s_t \cdot s_{t-1}) < 0 \\ 0 & en \; otro \; caso \end{cases} \tag{2.2}$$

donde s es una señal de longitud T y, s_t es la señal en el tiempo t.

Energía

Por naturaleza, la energía asociada con el habla es variable en el tiempo. De ahí el interés por saber cómo esta característica está cambiando, más específicamente, cómo lo está haciendo en una región del segmento hablado, por ejemplo, en el caso

de una palabra. Como sabemos, la señal de voz consiste en segmentos hablados y no hablados, es en estas regiones que la energía asociada a la presencia de la voz es grande en comparación con la región sin voz. Por lo tanto, la energía en segmentos cortos se puede utilizar para determinar aquellas regiones que nos sean de interés. La relación para encontrar la energía a corto plazo se puede derivar de la relación de energía total. Esta última está dada por [40]:

$$E_T = \sum_{t=0}^{T-1} s^2(n) \tag{2.3}$$

Donde s es una señal de longitud T.

Para el caso del cálculo de energía en una ventana utilizamos la Ecuación 2.1 como lo hemos venido haciendo en los párrafos anteriores. En consecuencia se puede escribir la relación para encontrar la energía a corto plazo como:

$$e(n) = \sum_{t=0}^{T-1} [s(n)w(m-n)]^2 \tag{2.4}$$

donde w es la ventana.

Características espectrales

Como se ha mencionado antes, los cambios en el tracto vocal conducirán a cambios en la resonancia y el rango de frecuencias. Otros cambios, como los de la fuerza espiratoria y la tensión de los músculos laríngeos, afectan la manera en que los pliegues vocales se abren y cierran, y por lo tanto, afectan el sonido producido por la glotis. Tales cambios se manifiestan en el espectro de la frecuencia de la señal del habla. La principal dificultad que debe resolverse en el análisis espectral del habla es, separar las características de la forma de onda glotal de las características del tracto vocal y los articuladores. Si se realizan ciertas aproximaciones sobre la linealidad y las interacciones entre los sistemas glotales y de resonancia, es posible estimar las características del filtro del tracto vocal y, mediante un proceso de filtrado inverso, la excitación glotal de la señal acústica del habla. Este enfoque ha sido ampliamente

utilizado en la investigación del habla para estimar formantes y la forma de onda glotal (por ejemplo, [41, 42, 43, 44]).

El objetivo del análisis de la forma de onda glotal es cuantificar aspectos que pueden estar directamente relacionados con la fisiología y la mecánica de la apertura y cierre de los pliegues vocales, esta relación con la fisiología glotal hace que sea un buen candidato para el estudio del habla afectiva. Un reto que se presenta es que la forma de la onda, a partir de la cual se desea estimar parámetros tales como el tiempo de apertura y de cierre glotal, es altamente sensible tanto a la parametrización del análisis como a las condiciones de grabación debido a que, la estimación precisa de los formantes requiere que se realicen grabaciones con ciertas características, por ejemplo, en entornos controlados que posibiliten la eliminación o reducción del ruido. Además, la parametrización vuelve la estimación de formantes en el habla espontánea menos confiable debido a que generalmente, dichos parámetros dependen de la experiencia previa de quien los elige.

Coeficientes cepstrales en las frecuencias de Mel

Ya que el oído humano percibe el sonido en una escala logarítmica [45], es necesario utilizar una transformación de las frecuencias a la escala perceptual. Un método para realizar esta transformación es hacer uso de la escala de Mel que es una aproximación a la escala perceptual humana; el punto de referencia entre ésta y la medición normal de la frecuencia se define mediante la asignación de un campo perceptivo de 1000 Mels a un tono de 1000 Hz, 40 dB por encima del umbral del oyente y, a partir de los 500 Hz se definen intervalos cada vez más grandes para producir incrementos iguales de tono, como resultado, cuatro octavas en la escala de hercios son alrededor de dos octavas en la escala Mel.

Los Mel frequency cepstral coefficients (MFCC) son coeficientes para la representación del habla basados en la percepción auditiva humana. Se derivan de la Transformada de Fourier y de la Transformada Discreta del Coseno, la diferencia radica

en que en MFCC las bandas de frecuencia están situadas logarítmicamente según la escala Mel. Para convertir de Hertz a Mels se utiliza la siguiente fórmula:

$$m = 2595 \log_{10} 1 + \frac{hz}{700} \tag{2.5}$$

Y, al contrario, para convertir de Mels a Hertz:

$$hz = 700(10^{\frac{m}{2595}} - 1) \tag{2.6}$$

Características de calidad de la voz

Como hemos visto, existe una fuerte relación entre las características de la calidad de voz y las emociones. Ejemplos de cualidades vocales son voz neutra, susurrante, agitada, chirriante y áspera o incluso, de falsete. El desafío aquí es estimar la fuente glotal, para ello, las medidas más comunes son Jitter y Shimmer. El Jitter mide la variación de ciclo a ciclo de la longitud del período mientras que el Shimmer mide las variaciones de ciclo a ciclo de la amplitud pico. Otra característica de calidad de voz es la relación de armónicos a ruido (HNR), ésta mide el grado de periodicidad en un sonido. Las implicaciones de utilizar estas características se pueden encontrar en [46, 47, 48, 49].

2.2. Los efectos de las emociones en el habla

Uno de los problemas que ha obstaculizado la investigación del habla emocional es la falta de un marco teórico firme que sirva para estructurar la investigación empírica, sobre todo porque la investigación del habla emocional suele abordar el tema desde diferentes ángulos. Así, mientras que las investigaciones aportan modelos sofisticados de producción del habla, a menudo éstas requieren una mayor comprensión de la psicología emocional. La investigación, desde el punto de vista psicológico, requiere un grado superior de sofisticación en la producción y el análisis del habla ya que, como resultado del estado actual, los resultados de las investigaciones han sido difíciles de

comparar y comprender. Como se señaló en [50], la investigación debe guiarse por un tratamiento teórico más completo de la psicología y la fisiología del habla emocional, comenzando por comprender la producción normal del habla. Existen varias revisiones detalladas de la historia y el estado actual de la psicología emocional en la literatura científica ([51, 52]). Esta sección presenta un breve resumen de ésta, con énfasis en las diversas teorías sobre las respuestas fisiológicas y sus efectos sobre la producción vocal.

Charles Darwin fue el primero en reconocer explícitamente la importancia de las expresiones vocales y faciales específicas de las emociones en animales, tanto en términos de su importancia comunicativa como de las manifestaciones de los cambios corporales que ocurrían durante los episodios emocionales [53]. La teoría de James–Lange de la emoción propone que las situaciones emocionales propician actividad en el sistema nervioso que es percibida por el cerebro a través de un proceso de retroalimentación periférica [54]. Con esto se entiende que la percepción de patrones específicos de actividad fisiológica periférica da lugar a la percepción subjetiva de diferentes emociones. El autor de [54] se opuso a algunas ideas de la teoría de James–Lange, incluida la idea de que las emociones tienen patrones correspondientes y diferenciados de la actividad fisiológica periférica. Desde entonces, las teorías sobre las emociones han diferido con respecto a la especificidad de la excitación fisiológica emocional.

Teorías de la emoción

El estudio de las emociones fue relegado durante la primera mitad del siglo XX, período donde predominó el conductivismo. No fue sino hasta la segunda mitad del siglo, con el surgimiento del cognitivismo, que se retomó la investigación de las emociones. Las ideas de la psicología cognitiva llevaron a [55] a proponer que la excitación fisiológica provocada durante los episodios emocionales es confusa y poco diferenciada pues las emociones surgen debido a un aumento de la excitación general, que después es interpreta en dependencia del contexto o situación. Por lo tanto, cuando ocurre una excitación elevada en una situación amenazante, la excitación es interpretada

como miedo, mientras que cuando la misma excitación se produce en respuesta a un estímulo positivo, por ejemplo ganar un premio, se interpreta como alegría. Aunque la evidencia empírica de la teoría de Schachter y Singer ha sido criticada tanto desde el punto de vista metodológico como conceptual, los argumentos en contra siguen siendo en gran medida teóricos y no empíricos.

A pesar de las múltiples investigaciones, la evidencia de la existencia de respuestas fisiológicas específicas a una emoción determinada sigue siendo poco concluyente ([56, 57, 58]). Alrededor de estos estudios, existe un número de teorías en las que las emociones se caracterizan dimensionalmente como existentes en un espacio de dos o más dimensiones, sosteniendo que la excitación fisiológica emocional es esencialmente no específica. Por ejemplo, [59, 60], proponen que cuando la expectativa a la ocurrencia de un evento concluye, aumenta la actividad simpática que es percibida e interpretada de manera similar a las propuestas de [55], dando inicio al análisis de emociones cualitativamente diferentes.

En contraste con los modelos de excitación emocional no específica, las teorías de [61, 62, 57] postulan la existencia de respuestas neuromotoras preestablecidas que se corresponden con un número reducido de emociones básicas o arquetípicas que, combinadas en "mezclas emocionales", dan lugar a un mayor número de emociones secundarias. Según estas teorías, la activación de dichas respuestas preestablecidas produce una respuesta expresiva y fisiológica coordinada, emocionalmente específica. La idea de que existe un número limitado de emociones básicas biológicamente motivadas, cada una con una respuesta emocional bien definida, es apoyada por una serie de estudios que han demostrado cómo, a pesar de las diferencias culturales de diversos grupos de individuos, un pequeño número de emociones son compartidas y reconocidas universalmente [63, 58].

Aunque existe evidencia de respuestas fisiológicas específicas para algunas emociones, es escasa, por ejemplo [57, 64]. Curiosamente, algunas de las pruebas más sólidas de la existencia de respuestas fisiológicas provienen de estudios donde se han utilizado variaciones de la técnica de retroalimentación facial [65]. En esta técnica se instruye

a los sujetos, de manera implícita, a producir configuraciones faciales correspondientes a emociones básicas, sin embargo, los sujetos no sólo informan sentir una mayor intensidad de la emoción correspondiente a la expresión facial producida, sino que, también, muestran actividad identificable en el sistema nervioso, presumiblemente debido a la retroalimentación entre los sistemas de respuesta facial y nervioso. Por otro lado, estos resultados han sido debatidos en [66] al sugerir que las variaciones apreciadas en el sistema nervioso no son más que un reflejo del esfuerzo requerido para producir los cambios faciales.

Las teorías de "evaluación" o appraisal, en inglés, son otro conjunto de teorías que postulan la existencia de emociones asociadas a respuestas fisiológicas diferenciadas, las primeras teorías cognitivas de "evaluación" de la emoción fueron presentadas en [67, 68]. Estas teorías iniciales fueron continuadas entre los años 70 y 80 ([52, 50, 69, 70]). Todas ellas coinciden en que las emociones se producen en respuesta a las evaluaciones de una situación y sus consecuencias para el bienestar y las necesidades del individuo. Como se señala en [71], existen correspondencias entre las diferentes versiones de la teoría de la evaluación, en particular con respecto a los criterios de evaluación o dimensiones de evaluación, que se cree subyacen a la mayoría de las emociones.

Sin embargo, la mayoría de los resultados obtenidos con la técnica de evaluación se han analizado de manera individual, es decir, se evalúan los efectos específicos por separado para cada dimensión de evaluación, por ejemplo [50, 72]. La teoría indica que cada dimensión de evaluación proporciona información al organismo sobre la mejor manera de adaptarse a una situación o evento y así prepararse para tomar la acción apropiada. Por lo tanto, una persona, al evaluar las acciones de otro individuo como positivas o negativas, se preparará para responder de diferentes maneras, por ejemplo de acercamiento, alejamiento, lucha o rendición. Es esta preparación la que constituye la respuesta emocional medible en el habla pues provoca cambios en la tonificación muscular del tracto vocal, dando lugar a diferentes patrones acústicos. Prever maneras de comprobar empíricamente si una respuesta emocional se produce

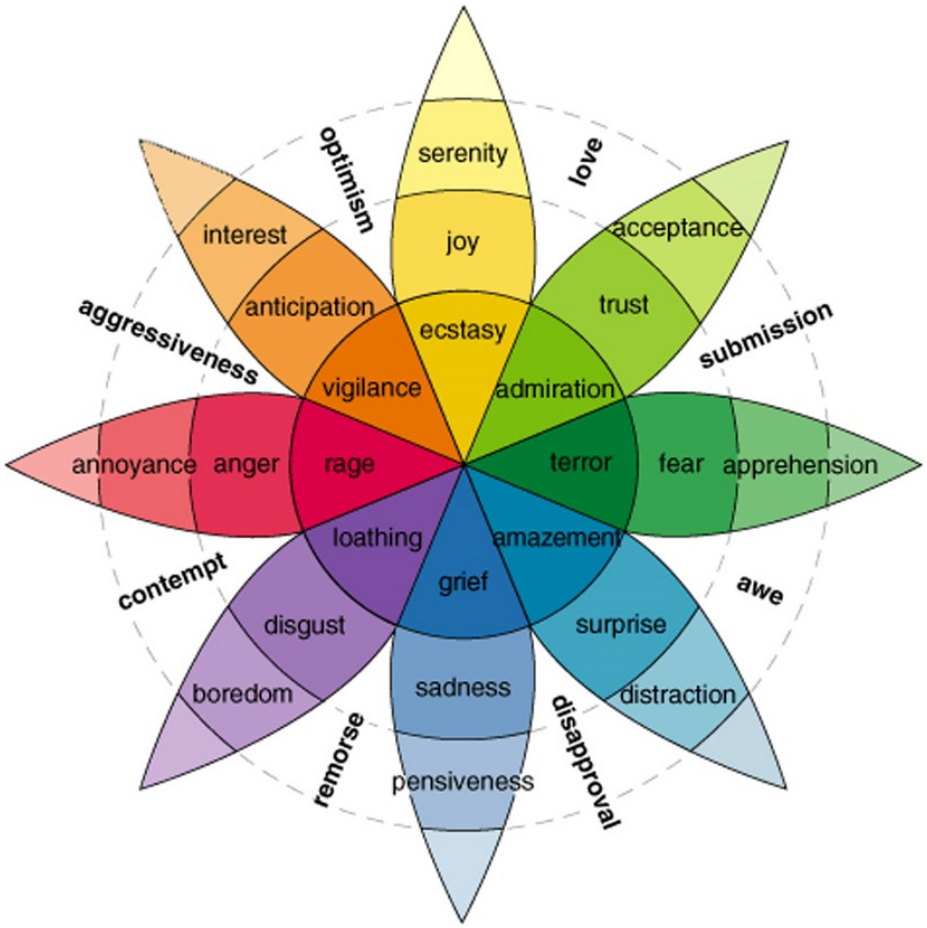


Figura 2-1: *Rueda de emociones por Robert Plutchik, tomada de [73], 2001*

como una única respuesta o como una combinación de respuestas a un evento resulta complicado. Sin embargo, encontrar parámetros vocales acústicos que sirven como marcadores individuales es posible, de la misma manera que el ceño fruncido y los cambios de la frecuencia cardíaca se han vinculado a la frustración y el enojo.

Considerar la existencia de un conjunto finito de emociones es uno de los conceptos más populares utilizados en el reconocimiento de la emociones. Tomkins identificó la existencia de nueve emociones básicas reconocibles por expresiones faciales específicas [74], por otro lado Ekman identificó sólo seis emociones básicas: la ira, el miedo, el disgusto, la felicidad, la tristeza y la sorpresa [58]. Ambos realizaron amplios estudios para probar la universalidad de estas emociones y las correspondientes expresiones faciales. Al no existir consenso sobre si las emociones humanas pueden ser agrupadas en un conjunto discreto, la investigación se orientó hacia la mezcla de emociones

básicas que den paso a otras emociones secundarias. Para tal efecto Plutchik consideró ocho emociones básicas que representó en una rueda de emociones esquematizada en la Figura 2-1, aquí las emociones opuestas se neutralizan entre sí y cuando se mezclan las adyacentes, resultan emociones secundarias por ejemplo mezclando sorpresa y tristeza se produce decepción.

Las emociones básicas fueron ampliamente criticadas al negar la universalidad y los fundamentos de los estudios de Ekman proponiendo el uso del modelo de excitación y valencia. Este enfoque se basa en el principio de que las emociones pueden ser consideradas como puntos en un espacio n–dimensional. Hubo varios intentos de clasificar las emociones en una o en un pequeño número de dimensiones, por ejemplo [75]. El enfoque más popular es el de dos dimensiones, ya sea agradabilidad (el grado de placer percibido) y activación (el grado de excitación), o valencia y excitación [76].

La eficacia de utilizar el enfoque dimensional ha sido señalada por distintos autores, entre ellos Schlosberg, Russell, Plutchick, Cowie y otros [77]. A pesar de que existen variaciones en los nombres de las dimensiones, se puede notar que en general una dimensión describe si la emoción percibida es positiva o negativa, mientras que la otra dimensión describe la intensidad de la emoción [78]. En la Figura 2-2, se muestra un mapa esquemático de las emociones en un modelo de valencia y excitación.

Los efectos de las emociones en la señal del habla

Como hemos visto, el estado emocional de un hablante afectará la calidad de su discurso de múltiples maneras. Estas variaciones pueden ser provocadas por cambios involuntarios en el sistema de producción del habla o incluso, por la adopción más controlada de estilos específicos y culturalmente aceptados. En paralelo con el desarrollo del lenguaje hablado, se han establecido formas prototípicas de expresión emocional o actitudes emocionales. Estas representaciones afectivas están determinadas, al menos en parte, por normas culturales tales como reglas de cortesía y etiqueta y, por lo tanto, varían según la cultura y el contexto social. Esto indica que en ciertas

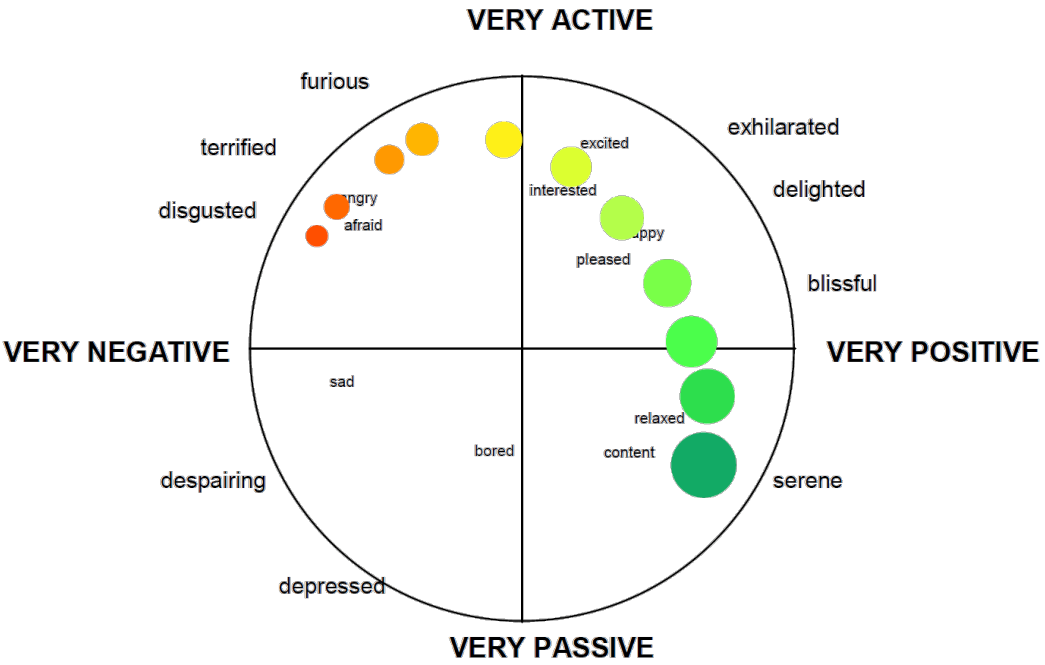


Figura 2-2: *Modelo bidimensional tomado de Schröder Marc et al., [79] 2000*

situaciones sociales, los hablantes pueden controlar o enmascarar la expresión "natural" de su estado emocional [80], por ejemplo, mientras que las expresiones de alegría entre amigos o familiares pueden ser muy animadas, la misma emoción experimentada entre extraños conducirá a una expresión emocional más tenue. Una teoría intrínsecamente relacionada es la teoría de adaptación comunicativa de Giles [81], ésta sugiere que el estilo comunicativo de un hablante converge o diverge del de su interlocutor dependiendo de la relación entre ambos [82].

La producción del habla comienza con los procesos cognitivos de planificación y estructuración del contenido de lo que se va a decir y la activación de los sistemas de producción del habla para producir los sonidos apropiados. Aunque la planificación del contenido y la estructura del enunciado es un proceso en gran parte automático, también se basa en la capacidad de atención y memoria que son recursos limitados, haciéndola susceptible a interferencias que dan como resultado un cambio en la fluidez del habla [83, 84, 85, 86]. Estos efectos han sido estudiados ampliamente mostrando cambios en F_0, la velocidad de habla y en el espectro como respuesta a las fluctuaciones de la carga cognitiva en los hablantes [87].

Existen razones para creer que las situaciones emocionales impactan la atención y la memoria, se ha demostrado que la presencia de información nueva e inesperada, conduce a una respuesta que redirige la atención hacia el nuevo estímulo [88]. Emociones como la ansiedad afectan negativamente la planificación, estructuración y ejecución del habla [80, 50], las personas susceptibles a esta emoción muestran un sesgo de procesamiento cognitivo hacia estímulos amenazantes que impacta el desempeño en tareas no relacionadas [89, 90]. La prueba clásica para evaluar este sesgo consiste en pedir a los sujetos que nombren los colores con los que están escritas las palabras que se les presentan secuencialmente. Los resultados de esta prueba se presentan mediante la medición del tiempo requerido para decir el color entre una palabra y otra, esto ha permitido develar que el tiempo requerido para decir el color de palabras que representan amenazas es mayor que para palabras no relacionadas con amenazas [89]. La interpretación es que las palabras relacionadas con amenazas son procesadas con mayor prioridad por sujetos propensos a sufrir ansiedad, interfiriendo con la tarea de identificación del color.

El impacto de otras emociones en la planificación y estructuración del habla es menos claro. Por ejemplo, el mismo experimento aunque llevado a cabo con sujetos deprimidos no ha mostrado una tendencia similar [89]. En otro estudio, [91] encontraron un aumento en la velocidad del habla y una disminución en las pausas entre palabras cuando pacientes deprimidos salieron de una sesión de terapia, sin embargo, hay que señalar que cualquier posible afectación en la planificación, estructuración y producción del habla como consecuencia de las emociones, dependerá del contenido del discurso pues es probable que las frases bien practicadas se memoricen y ejecuten de forma automática y por lo tanto, los efectos podrían resultar enmascarados.

2.3. Características del habla emocional

Como hemos explorado antes, los tres sistemas de producción del habla actúan bajo el control del sistema nervioso. La musculatura que es el sistema que controla

la respiración, la posición de la laringe y la de los articuladores, también se ve influenciada por las emociones. Los músculos de la boca y los labios, que cambian la longitud y la forma del tracto vocal, también son utilizados para la expresión facial de las emociones mientras que la producción de saliva y mucosa, que afecta la resonancia del tracto vocal, depende de la actividad parasimpática.

En situaciones en las que el cuerpo se encuentra en una condición relajada, no se imponen limitaciones al funcionamiento de los sistemas de producción del habla. Así, cuando hablamos en la vida cotidiana, la respiración puede producir una vocalización con intensidad bien controlada durante la duración de una frase o oración, sin comprometer las necesidades respiratorias [92, 93]. Asimismo, los músculos pueden adaptarse a los requerimientos de la producción del habla, produciendo así una articulación precisa y clara. Sin embargo, cuando una persona es introducida en una situación no relajada, los efectos se reflejan con más intensidad sobre los tres subsistemas del habla. Como hemos visto, las emociones están acompañadas por respuestas adaptativas del sistema nervioso cuya finalidad es preparar al cuerpo para tomar una acción [56, 94]. Esto implica que las situaciones emocionalmente cargadas podrían provocar un patrón de cambios fisiológicos que, de alguna manera no arbitraria, perturba los sistemas de producción del habla. Por ejemplo, la ira, que desencadena la preparación para el conflicto, produce una mayor tensión en la musculatura junto con la elevación en la frecuencia cardíaca y respiratoria, lo que a su vez provocaría un cambio en la producción de sonido y en la calidad de la voz [50]. La posición teórica alternativa es que las emociones producen un cambio en la fisiología general, con la diferenciación entre emociones dependientes de factores cognitivos, en lugar de fisiológicos (por ejemplo, [60, 55]).

A continuación se proporciona un breve resumen de los hallazgos principales para las emociones comúnmente estudiadas. Para una revisión detallada sobre los estudios empíricos que han medido las propiedades acústicas del habla emocional, se puede consultar [95, 50, 96], al tiempo que los detalles de las medidas mencionadas, como F_0, se detallan en la Sección 2.1.

Aburrimiento

El habla aburrida es generalmente lenta y monótona, con un nivel y rango bajo de F_0 así como de variabilidad y velocidad de articulación. De manera opuesta, el habla interesada o entretenida cuenta con un rango amplio de F_0 y una velocidad incrementada del habla.

Disgusto

Los resultados para el disgusto son poco consistentes, esto puede deberse a las técnicas de inducción de disgusto en los sujetos de prueba pues podrían considerarse cuestionables. En algunos estudios se utilizan actores que interpretaron la emoción mientras que en otros, se evaluaron las reacciones a películas poco placenteras. Los estudios que utilizaron el primer procedimiento encontraron una disminución en el promedio de F_0 mientras que los que utilizaron el segundo, encontraron un aumento del promedio. Esto implica que el disgusto no es universalmente reconocido en el habla [97].

Felicidad

La felicidad es una de las pocas emociones positivas frecuentemente estudiadas. Ésta se corresponde con niveles altos de excitación medidos en F_0, su rango y variabilidad. Además, hay evidencia, aunque no concluyente, de un aumento en la energía y en la velocidad del habla. Sin embargo, algunas formas de la emoción parecen caracterizarse por niveles bajos de intensidad y F_0, menor intensidad y articulación más lenta.

Enojo

En general, se ha observado un aumento de F_0 y de la intensidad en el habla donde se presenta esta emoción. Aunque teniendo en cuenta que la media de F_0 no

es una medida acústica singular, es decir que se calcula con base en un conjunto de valores, no está claro si el habla enojada tiene un valor mayor de F_0 o un rango más amplio de F_0 o ambos. Es posible que, como en las emociones anteriores, los estudios que han encontrado un aumento de los valores de F_0 han medido el enojo "caliente" o rabia, mientras que los estudios en los que este aumento no se vió reflejado pueden haber medido el enojo "frío" o molestia, según lo encontrado en [96]. El enojo también parece estar caracterizado por altas frecuencias que, junto con el aumento de intensidad, refleja un mayor esfuerzo vocal.

Miedo

Los altos niveles de excitación esperados para el miedo son consistentes con la evidencia que muestra incrementos en la intensidad y promedio de F_0 así como en la tasa de velocidad del habla. Los resultados para el rango de F_0 son, sin embargo, menos consistentes. Al igual que con el enojo, el aumento de la intensidad del habla es acompañado por un incremento en las frecuencias altas. La ansiedad o la preocupación, que son emociones relacionadas, también muestran una articulación más rápida aunque los datos sobre las otras variables son menos consistentes. En [96] encontraron una disminución en la media, el valor mínimo y el rango de F_0 para la ansiedad, lo que permitió hacer una distinción clara entre ambas emociones.

Tristeza

Al igual que con el miedo, los resultados coinciden reportando disminuciones en los valores de F_0, la intensidad y la tasa de articulación. La mayoría de los estudios publicados en la literatura parecen haber estudiado las formas más silenciosas y resignadas de esta emoción, en lugar de las formas más excitadas como la desesperación, donde se encuentran correlatos que reflejan la excitación, como el aumento de la intensidad, de la energía y de los valores de F_0.

Problemas metodológicos

Es evidente, a partir de este breve resumen, que cuando existe una consistencia considerable en los hallazgos, suele estar vinculada a la excitación, independientemente de la de la emoción específica que se investigue. Así, la alegría, el enojo y el miedo, son todas emociones de excitación alta que se expresan con perfiles vocales similares. Pocos o ninguno de los estudios encontraron patrones acústicos que pudieran diferenciar las principales dimensiones no emocionales como la valencia y el control ejercido. Sin embargo, esto no significa que las emociones discretas no se puedan diferenciar, de hecho, dado el alto reconocimiento de las emociones actuadas y evaluadas en los estudios de percepción, hay razones para creer que existen patrones acústicos específicos a las emociones.

A pesar de los muchos estudios en el tema, hay varios problemas con el diseño y la ejecución del análisis de la prosodia afectiva que podrían explicar por qué tales patrones acústicos todavía no se han encontrado consistentemente. Ejemplo de esto son las diferencias entre el habla emocional espontánea e inducida.

Habla emocional inducida y espontánea

Es posible que las altas tasas de reconocimiento en los estudios de percepción reflejen el hecho de que en estos estudios se ha utilizado el habla emocional actuada. Tales expresiones emocionales actuadas reflejan, al menos parcialmente, la adopción de estereotipos vocales sociales o culturales, denominados “efectos de atracción” por [98, 99]. En los estudios de percepción esto podría significar tasas de reconocimiento artificialmente altas, ya que el propósito es comunicar eficazmente el estado emocional del orador a una audiencia. Las emociones reales, por el contrario, están acompañadas por cambios en la señal del habla que no poseen un propósito de comunicación intencional, sino que reflejan cambios no controlados en la producción del habla denominados “efectos de empuje” por [98, 99].

Cuando la tarea es clasificar el habla emocional “real”, que refleja los efectos de

empuje de la emoción, la puntuación de reconocimiento de emociones son mucho más bajas, consistente con las características del habla no actuada. Desafortunadamente, los pocos estudios que han intentado algún tipo de inducción emocional real, por ejemplo [100, 101], o que han trabajado con grabaciones del habla cotidiana, se han limitado a dos emociones, como el estrés alto y bajo o feliz y triste.

2.4. Extracción de características

La extracción de características de la voz para el reconocimiento de emociones resulta de gran importancia ya que son estas características las que serán examinadas para determinar la clase a la que pertenecerá el sonido. Para extraer estas características, primero se deberá pensar en las que resulten más adecuadas para resolver el problema, en particular el de emociones, ya que podría no ser suficiente extraer el tono y la frecuencia. Incluso se debe considerar la duración del intervalo que se analizará para determinar si dichas características serán de utilidad en la clasificación; tomando en cuenta la duración de este intervalo, pueden ser divididas en dos: las locales, que son extraídas de secciones o ventanas de la señal y las globales que son obtenidas mediante cálculos estadísticos de los elementos locales.

Diversos estudios han mostrado que el uso de características globales tienden a producir mejores resultados en lo que concierne a la exactitud y tiempo de clasificación [102], una de las razones es que el número de características globales es mucho menor en comparación con las características locales y por ende el tiempo requerido por las técnicas de validación es menor. Así mismo muestran que si bien las características globales ayudan a producir mejores clasificaciones, éstas sólo son eficientes para distinguir entre emociones de alta y baja energía como en el caso de la ira y la tristeza. De igual manera se ha visto que la dimensionalidad de los vectores de entrenamiento que serán obtenidos mediante el análisis de características globales será menor, lo que dificultará el uso de clasificadores como las SVM y los Modelos Ocultos de Markov (HMM). Para estos tipos de clasificadores que funcionan mejor con

vectores de alta dimensionalidad, será preferible utilizar las características extraídas mediante el análisis local [102]. Esto lleva a cuestionarnos el tamaño de la ventana. Una opción a seguir es segmentar la señal de la voz en los fonemas que la conforman y otra es la de segmentarla por cada frase, con esto, el tamaño de la ventana y el tipo de análisis, local o global, influirán en el número de elementos de los vectores de entrenamiento. Tener muchos o pocos de éstos no implica necesariamente que se clasificarán correctamente las señales de voz ya que también son importantes las características representadas en ellos, es por esto que debemos cerciorarnos que éstas describan correctamente el contenido emocional de la voz.

Las características de la voz pueden ser agrupadas en tres categorías: continuas, espectrales y cualitativas [102]. Se ha estipulado que elementos tales como el tono y la energía transmiten la mayor parte del contenido emocional de un enunciado, esta propiedad permite diferenciar aquellas emociones que tienen una alta activación de las que poseen una baja [103], las características más usadas en esta categoría son: la frecuencia fundamental, la energía, la duración y los formantes. Las características espectrales podrán extraerse mediante distintas técnicas entre las que destacan: codificación lineal predictiva (LPC), coeficientes cepstrales en las frecuencias de Mel (MFCC) y coeficientes de potencia de la frecuencia logarítmica (LFPC). Estos elementos resultan importantes debido a que 'el contenido emocional de una frase tiene impacto en la distribución de la energía espectral a lo largo del rango de frecuencias de la voz' [102]. Algo más que sabemos sobre la energía en la voz es que la frecuencia fundamental y sus armónicos cambian bajo distintos estados emocionales, bajo esta premisa Teager expone que 'escuchar es el proceso de detectar energía' [104].

Como se ha venido exponiendo en esta sección, seleccionar los atributos que describan correctamente una emoción es una tarea de importancia. Es para esta selección que se han explorado distintas opciones como es el caso del volumen en la voz pues suele ser un indicador intuitivo de la emoción, algunos autores [104] la miden como una función directa del voltaje del micrófono aunque normalmente depende de la distancia entre éste y el hablante, de la dirección y del ambiente. Para abordar este

	Enojo	Felicidad	Tristeza	Miedo	Disgusto
Velocidad	Más rápida	Rápida o lenta	Más lenta	Muy rápida	Mucho más rápida
Tono promedio	Muy alto	Más alto	Más bajo	Muy alto	Muy bajo
Intensidad	Alta	Alta	Baja	Normal	Baja
Calidad	Agitada	A todo volumen, jadeante	Resonante, resoplante	Irregular	Quejumbrosa
Cambios en el tono	Abruptos	Suaves, hacia arriba	Suaves, hacia abajo	Normales	Amplios

Tabla 2.1: *Emociones y parámetros de la voz tomado de Cowie et al. [105], 2001*

problema, se ha propuesto prestar atención al estrés del hablante ya que esto ayuda a distinguir el volumen [106]. Otra opción es utilizar una distribución de energía ya que los brincos hacia las vocales y en contra de las consonantes son indicador de un alto volumen.

La duración de las pausas entre las palabras, los fonemas y los atributos parece ser otro buen indicador de las emociones ya que la velocidad del habla se puede medir en palabras por minuto, esto nos lleva a la necesidad de detectar los límites entre las palabras lo que es una tarea complicada. Para aminorar este problema se proponen otras formas de medir la velocidad del habla que funcionan mejor para la extracción automática, tal es el caso de la detección de fonemas mediante el análisis del núcleo de la sílaba o encontrar los límites entre las características del habla como el de la 'explosión fricativa'.

Habiendo descrito de manera general algunas características que pueden ser extraídas de la señal del habla se manifiesta su importancia pues, con base en la experimentación realizada, creemos que ayudan a caracterizar de manera eficaz las emociones que se presentan en la voz.

CAPÍTULO 3

APRENDIZAJE PROFUNDO

Algunos autores sugieren que las arquitecturas profundas son mucho más eficientes, en términos de los elementos computacionales requeridos, que las arquitecturas de poca profundidad como las redes neuronales tradicionales con una sola capa oculta u otras técnicas como los modelos ocultos de Markov y las máquinas de soporte vectorial, entre otras [107]. No obstante, este tipo de redes profundas son muy difíciles de entrenar, para superar este problema en el 2006 se introdujo un algoritmo de aprendizaje no supervisado que hizo posible dicho entrenamiento [9]. En las siguientes secciones se dará una breve introducción a estas arquitecturas profundas y a las máquinas restringidas de Boltzmann (RBM), al mismo tiempo que se revisarán sus algoritmos de entrenamiento.

Una motivación clave para el aprendizaje profundo o deep learning (DL) se basa en la biología, ésta nos dice que el cerebro funciona de forma 'profunda'. Esta naturaleza jerárquica, proviene de la observación de que las capas superiores representan conceptos cada vez más abstractos, lo que lleva a pensar que la estructura jerárquica empleada por el Neocórtex (capa racional del cerebro, controla las emociones y las capacidades cognitivas), es la respuesta a gran parte de su poder [108, 109, 110].

El aprendizaje profundo es un paradigma del aprendizaje maquinal que se centra en el uso de los modelos jerárquicos de datos, éste plantea la hipótesis de que con el fin

de aprender las representaciones de alto nivel de los datos, se necesita una jerarquía de representaciones intermedias. Por ejemplo, en el caso de la visión computacional, el primer nivel de representación podría ser la obtención de los pixeles, el segundo nivel podría reconocer líneas y contornos, mientras que las representaciones de nivel superior podrían reconocer partes y objetos como se puede ver en la Figura 3-1. La hipótesis es que, si se permite que la red encuentre representaciones en varios niveles de abstracción, se obtendrán mejores resultados, pues cada capa irá encontrando patrones en las capas más bajas y representando conceptos más abstractos en las superiores. Aunque parece ser una buena idea, en la práctica no resulta tan simple como apilar muchas capas, no obstante, los recientes avances en los algoritmos de aprendizaje para arquitecturas profundas, han hecho posible que estos sistemas sean factibles [9].

En lugar de utilizar las redes neuronales tradicionales, en varios trabajos se ha propuesto hacer uso de una arquitectura profunda que utilice como bloque de construcción las máquinas restringidas de Boltzmann (RBMs) que pueden ser apiladas para obtener distintos niveles de especialización [111]. Para ejemplificar esta arquitectura profunda, se pueden utilizar las redes neuronales artificiales (ANN) tradicionales de la siguiente forma: teniendo como base una ANN en la primer capa y deseando extenderla a una nueva, se plantea utilizar la salida de esta primera como entrada para la siguiente. Este concepto se irá explicando a lo largo de las siguientes secciones aunque, si se desea, se puede ver representado en la Figura 3-5.

3.1. Máquinas Restringidas de Boltzmann

La máquina de Boltzmann (BM), representada en la Figura 3-2, fue desarrollada por Geoffrey Hinton y Terry Sejnowski en 1983 [112]. Ésta es un tipo de red neuronal donde todas las neuronas están conectadas entre si y tiene la particularidad de que toma decisiones estocásticas sobre si una neurona estará activada o no, es decir, son construidas introduciendo variaciones probabilistas a los pesos de la red. A esta má-

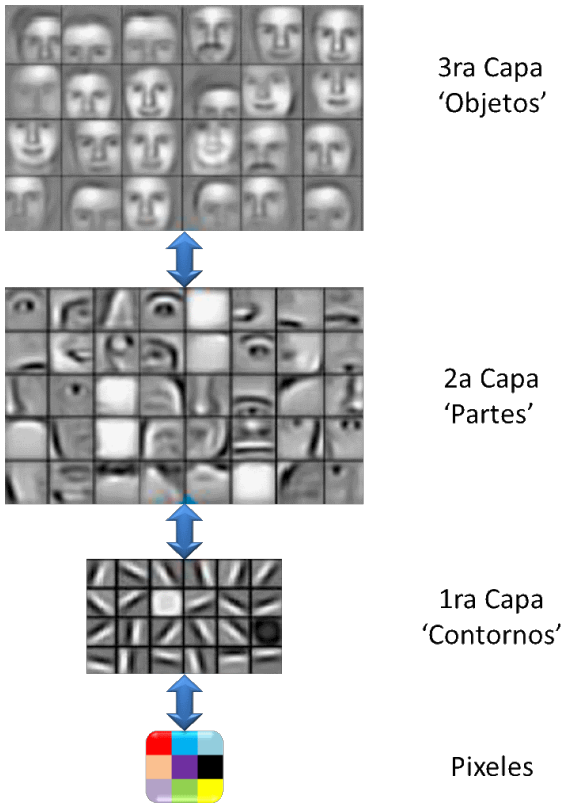


Figura 3-1: *Visualización de las representaciones en las RBMs*

quina se le presenta un conjunto de vectores de entrenamiento que deberá aprender a clasificar con alta probabilidad, para lograrlo la BM debe encontrar los pesos de las conexiones que logren que los vectores de datos con los que fue entrenada presenten un 'costo' o valor bajo en relación con otros ejemplos. No obstante, las BM sin restricciones de conectividad no han demostrado ser útiles para resolver los problemas que se dan en la práctica ya que, como es de esperarse, el proceso de aprendizaje es lento en redes de gran tamaño debido a la forma en la que están construidas como se puede ver en la Figura 3-2. Es por esto que se propusieron las máquinas *restringidas* de Boltzmann (RBM), esta 'restricción' se basa en reducir el número de conexiones impidiendo que las neuronas o unidades de la misma capa se 'vean', como en la Figura 3-3. En ambas figuras, v son las unidades visibles o capa de entrada, h las unidades ocultas o capa de salida, W las conexiones o pesos entre v y h. La diferencia radica en que las conexiones o pesos J y L entre las unidades ocultas y visibles respectivamente, presentes en la Figura 3-2, son eliminadas de la topología presentada en la Figura 3-3.

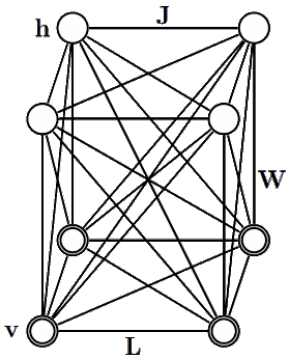


Figura 3-2: *Máquina de Boltzmann*

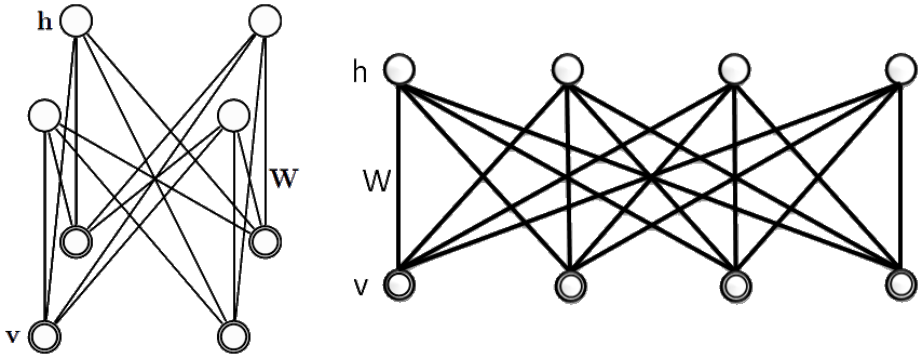


Figura 3-3: *Máquina Restringida de Boltzmann*

Sean v_i y h_j los estados de la unidad visible i y la unidad oculta j y a_i y b_j sus respectivos sesgos con los pesos $w_{i,j}$ entre v_i y h_j, la función de energía de las RBMs está dada por [113]:

$$E(v,h) = -\sum_{i \in visible} a_i b_i - \sum_{j \in oculta} b_j h_j - \sum_{i,j} v_i h_i w_{i,j} \tag{3.1}$$

O de manera simplificada:

$$E(v,h) = -a'v - b'h - h'Wv \tag{3.2}$$

La red asigna una probabilidad a cada par, unidad visible y un vector de unidades ocultas, a través de esta función:

$$p(v,h) = \frac{\exp^{-E(v,h)}}{Z} \tag{3.3}$$

donde la función de partición Z, está dada por la suma de todos los pares de vectores visibles y ocultos:

$$Z = \sum_{v,h} \exp^{-E(v,h)} \tag{3.4}$$

La energía libre de la entrada, es decir, de las unidades visibles, es la que se debe modificar para aumentar o reducir las probabilidades como se explica en las siguientes tres ecuaciones [10]:

$$FE(v) = -\sum_{i} a_i v_i - \sum_{i} log - \sum_{h_i} \exp^{h_i W_i x} \tag{3.5}$$

La probabilidad de que la red clasifique a un vector de unidades visibles v, está dada por sumatoria de todos los vectores ocultos:

$$p(v) = \frac{1}{Z} \sum_{h} \exp^{-E(v,h)} \tag{3.6}$$

Esta probabilidad se puede elevar mediante el ajuste de los pesos para reducir la energía de ese vector y así aumentar la de los otros, en la siguiente ecuación con ϵ siendo la tasa de aprendizaje, se muestra esa modificación de los pesos.

$$\Delta w_{i,j} = \epsilon(\langle v_i h_j \rangle_{datos} - \langle v_i h_j \rangle_{modelo}) \tag{3.7}$$

El gradiente de la probabilidad logarítmica de un vector de entrenamiento con respecto a un peso donde los paréntesis angulares, $\langle\rangle$, denotan la distribución de los *datos* (datos de entrada) y del *modelo* (datos propagados) es:

$$\frac{\partial \log p(v)}{\partial w_{i,j}} = \langle v_i h_j \rangle_{datos} - \langle v_i h_j \rangle_{modelo} \tag{3.8}$$

Debido a la estructura específica de estas redes, las unidades visibles y ocultas son condicionalmente independientes [114]:

$$p(v|h) = \prod_i p(v_i|h)$$
$$p(h|v) = \prod_j p(h_j|v) \tag{3.9}$$

Usando esta propiedad, podemos escribir:

$$p(v_j = 1|h) = \sigma(a_i + \sum_j h_j w_{i,j})$$
$$p(h_j = 1|v) = \sigma(b_j + \sum_i v_i w_{i,j}) \tag{3.10}$$

donde σ es la función sigmoidea:

$$\sigma(x) = \frac{1}{1 + \exp^{-x}} \tag{3.11}$$

Este modelo cuenta con dos fases; la fase positiva disminuye la energía de los datos de entrenamiento y la fase negativa aumenta la energía de todos los demás estados visibles que el modelo puede generar. La fase positiva es fácil de calcular debido a la Ecuación (3.9), por el contrario, la fase negativa no es fácil de calcular ya que implica sumar todos los estados posibles del modelo, por este motivo en lugar de calcular la fase negativa exacta se realizan muestras del modelo.

En resumen, la idea para entrenar el modelo es hacer que éste genere datos parecidos a aquellos que le fueron presentados como de entrenamiento, o dicho de otra manera, queremos maximizar la probabilidad logarítmica de los datos de entrenamiento o reducir al mínimo la probabilidad logarítmica negativa de estos.

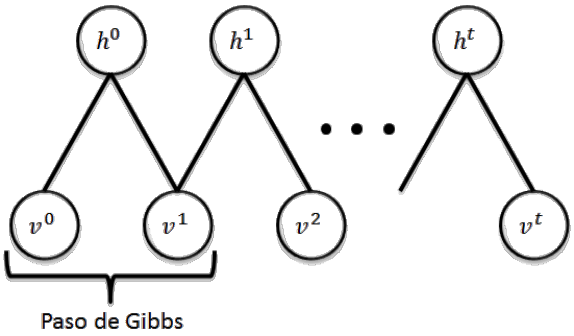


Figura 3-4: *Paso de Gibbs*

Muestreo y divergencia contrastiva

Como dijimos, las RBMs son un tipo de red neuronal, por lo que también funcionan actualizando los estados de algunas unidades en dependencia de otras, para actualizar el estado de una unidad i se necesita hacer uso de las ecuaciones (3.7) y (3.8). Nótese que no podemos garantizar que una unidad será activada, en todo caso podemos decir que será activada con alta probabilidad. Las muestras de $p(x)$ se pueden conseguir mediante la ejecución de una cadena de Markov hasta la convergencia utilizando el muestreo de Gibbs [115].

El muestreo de Gibbs para N variables aleatorias $S = (S_1, \ldots, S_N)$ se realiza a través de una secuencia de muestreo con N sub-pasos de la forma $S_i \sim p(S_i|S_{-i})$ donde S_{-i} contiene las $N-1$ variables aleatorias de S excluyendo S_i. Para las RBMs, como se puede ver esquematizado en la Figura 3-4, S consiste en el conjunto de unidades visibles y ocultas y ya que son condicionalmente independientes, se puede realizar el muestreo por bloques:

$$\begin{aligned} h^0 &= p(h|v^0) \\ v^1 &= p(v|v^0) \\ h^1 &= p(h|v^1) \\ &\ldots \\ v^n &= p(v|h^{n-1}) \end{aligned} \tag{3.12}$$

Con esta configuración, las unidades visibles se muestrean simultáneamente con los valores fijos en las unidades ocultas, de forma similar, las unidades ocultas se muestrean simultáneamente dados los valores de las unidades visibles. Un paso en la cadena de Markov se toma como sigue:

$$h^{n+1} \sim \sigma(W'v^n + c)$$
$$v^{n+1} \sim \sigma(Wh^{n+1} + b) \tag{3.13}$$

donde h^n se refiere al conjunto de todas las unidades ocultas en el paso n-ésimo de la cadena de Markov. Esto quiere decir que para el caso particular de h_i^{n+1}, se verá activada con probabilidad $\sigma(W_i'v^n + c_i)$ y, de manera similar v_j^{n+1} se verá activada con probabilidad $\sigma(W_jh^{n+1} + b_j)$. Cuando $t \to \infty$, las muestras v^t, h^t modelan correctamente a $p(v, h)$ aunque por supuesto, hacer los cálculos para $t \to \infty$ resulta computacionalmente prohibitivo. Es por esto que se propuso la Divergencia Contrastiva (CD) [116]. La CD hace uso de dos técnicas para acelerar el proceso de muestreo:

- Ya que se desea que la $p(v) \approx p_{entrenamiento}(v)$, la cadena de Markov se inicializa con un vector de entrenamiento con lo que se logra que la distribución sea cercana a p y por ende, la cadena esté próxima a converger.
- Además, la CD no espera a que la cadena converja, en principio se requiere una cantidad k de pasos de Gibbs, aunque en la práctica se demostró que con $k = 1$ es suficiente [10].

Para utilizar la DBN para clasificación, un vector de entrada se le presenta a la capa visible del primer nivel, pasando hacia arriba las salidas a través de la DBN hasta que se llega a la última capa oculta. En la RBM superior se elige la unidad que tiene la menor energía libre [9], una forma más simple de usar la DBN para clasificación es simplemente añadir una última capa consistente en un clasificador estándar y entrenar todo el modelo como si se tratase de una red neuronal feedforward con backpropagation.

3.2. Redes de creencia profunda

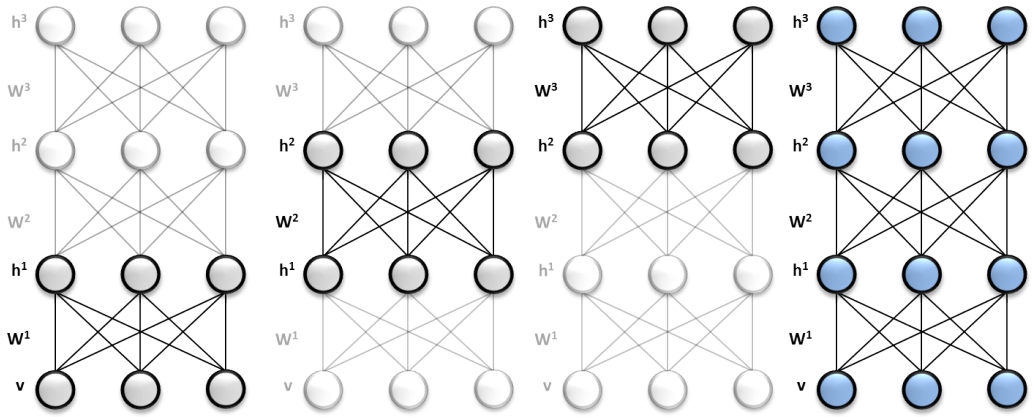


Figura 3-5: *Pasos en el entrenamiento de una DBN*

Una Red de Creencia Profunda (DBN) es un tipo red neuronal con una arquitectura profunda, es decir, con muchas capas ocultas. Se compone de una capa de entrada que contiene las unidades visibles, un número ℓ de capas ocultas y finalmente una capa de salida que tiene una unidad para cada clase a clasificar. En la Figura 3-5 se muestra una DBN con $\ell = 3$ a la que se le puede agregar o enlazar un clasificador estándar a la salida de ésta. Las DBNs modelan la distribución entre el vector de unidades visibles v y las ℓ capas ocultas h^k de la siguiente forma:

$$P(v, h^1, \ldots, h^\ell) = \left(\prod_{k=0}^{\ell-2} P(h^k | h^{k+1}) \right) P(h^{\ell-1}, h^\ell) \tag{3.14}$$

Donde $v = h^0$ y $P(h^{k-1}|h^k)$ es una distribución para las unidades visibles condicionada a las unidades ocultas de la máquina restringida de Boltzmann, en el nivel k, y $P(h^{l-1}|h^l)$ es la distribución para la RBM del nivel superior. Los parámetros de una DBN son los pesos w^j entre las unidades de las capas $j-1$ y j y el sesgo b^j de la capa j. Para entrenar una DBN se sigue un algoritmo voraz no supervisado que se aplica a las RBMs que la constituyen [117], el proceso es como sigue:

- **Paso 1.** Entrenar la primer capa como una RBM.
- **Paso 2.** Obtener la salida de (1) y utilizarla como fuente de datos para la siguiente capa.
- **Paso 3.** Entrenar la siguiente RBM, utilizando los datos obtenidos en (2) como vectores de entrenamiento para la capa visible.
- **Paso 4.** Repetir los pasos (2 y 3) para cada capa.
- **Paso 5.** Aplicar retropropagación a los pesos de la DBN como se haría con un perceptrón multi–capa (MLP).

CAPÍTULO 4

INFORMACIÓN Y PODA DISCRIMINATIVA

En los últimos años, muchos desafíos de reconocimiento a gran escala han sido resueltos con éxito utilizando los recientes avances en el reconocimiento de patrones, el aprendizaje maquinal y las redes neuronales artificiales. En el caso de redes neuronales artificiales, existen varios criterios utilizados para evaluar la calidad de una red, por ejemplo: el tiempo de entrenamiento, la escalabilidad y la habilidad de generalización, entre otros. Analizar el tamaño de la red es especialmente relevante ya que los trabajos recientes muestran que las redes más grandes o más profundas pueden asignar las características a un espacio más apropiado. En consecuencia, han surgido nuevas complicaciones asociadas con los algoritmos de entrenamiento complejos y computacionalmente exigentes. En este contexto, las máquinas restringidas de Boltzmann (RBM) han recibido mucha atención.

La idea detrás del paradigma de aprendizaje profundo sugiere que para aprender las representaciones de alto nivel de datos, se requiere una jerarquía de representaciones intermedias [10]. Estas representaciones intermedias en una arquitectura profunda se traducen en una red neuronal artificial con varias capas de unidades ocultas entre las capas de entrada y de salida. Sin embargo, estas capas ocultas son difíciles de optimizar. Los mejores resultados obtenidos en las tareas de aprendizaje supervisado implican un componente de aprendizaje no supervisado, generalmente en una fase de

pre–entrenamiento voraz no supervisado. Esto significa que si se permite a la red descubrir representaciones en varios niveles de abstracción, obtendrá mejores resultados ya que en las capas inferiores la red encontrará características básicas, mientras que en las capas superiores se representarán conceptos más complejos [9, 10].

Cuando se diseñan las redes neuronales es necesario decidir la arquitectura óptima y número de parámetros necesarios para resolver una tarea específica. Como se ha introducido antes, las grandes redes tienen la capacidad de aprender funciones más difíciles, sin embargo, esto se produce a costa de una mayor complejidad computacional. Sin embargo, la complejidad computacional de una red neuronal depende no sólo del número de parámetros u operaciones aritméticas sino de la arquitectura, tipos de capas y patrones de conectividad. Para ello, se han propuesto varias ideas para diseñar redes "ligeras" que proporcionan mejor rendimiento. Un enfoque común para determinar el tamaño de la red es mediante heurísticas, por lo general buscando buen rendimiento y capacidad de generalización en un conjunto de validación, especialmente si el tamaño del problema es grande. Otro enfoque considera formas de "crecimiento" de una red neuronal artificial hasta que se logra un rendimiento satisfactorio [13]. Una técnica diferente utiliza métodos de "poda" [15, 16, 17] que, en general, comienzan por el entrenamiento de una red neuronal artificial, que es lo suficientemente grande para asegurar un rendimiento satisfactorio para, posteriormente, eliminar de la red entrenada algunas neuronas (por ejemplo, las que tienen los pesos más pequeños) para después reentrenar la red. Este procedimiento también podría repetirse hasta que se alcance algún criterio de convergencia, de lo contrario se asume que la red más pequeña que se ha producido tiene la topología más adecuada para el conjunto de datos dado. Este tipo de poda se denominó poda de post–entrenamiento (PTP) [15, 118]. Sin embargo, el sobre–entrenamiento puede ocurrir si el número de parámetros de la red es mayor que el número óptimo (lo que sucede con frecuencia pues el número óptimo es desconocido), por otro lado, tener un número pequeño de parámetros limitará la capacidad de aprendizaje de la red. En [119], se utilizó un *análisis discriminante generalizado* en conjunto con una DBN mostrando una mejora

significativa con respecto a las máquinas de soporte vectorial. Sin embargo, Brueckner [120] encontró que la RBM fue la que ayudó en la tarea de clasificación y no la DBN.

A pesar de que estos resultados son alentadores, la selección de una topología adecuada para la red sigue siendo un área abierta, además, las técnicas utilizadas frecuentemente para evaluar dichos mecanismos son computacionalmente costosas, y en general no responden a una valoración objetiva del aprendizaje particular de las neuronas. Un enfoque es mediante la selección, de acuerdo con algún criterio, de las neuronas que más contribuyen al objetivo de la red y luego podar las que contribuyen con menos. Para ello, en este trabajo investigamos el uso de cinco medidas discriminativas binarias y ocho medidas multi–clase, como una forma de evaluar la utilidad de cada unidad oculta de una RBM en un problema de clasificación. Aquí utilizamos la información sobre las clases ya que el objetivo es medir cuánto del poder discriminativo de la red es aportado por cada neurona oculta con el fin de encontrar un tamaño de red adecuado, manteniendo al mismo tiempo un rendimiento de clasificación apropiado. Una de las medidas que pueden ser estudiadas es la varianza en la activación de las neuronas, la lógica subyacente es que las neuronas con una activación constante no transmiten información discriminativa sobre las muestras. Sin embargo, como las máquinas de Boltzmann son estocásticas, medir la varianza de activación no resulta apropiado pues una neurona con probabilidad constante de activación, podría tener una varianza de activación alta. Existen varias medidas que pueden ser utilizadas con este objetivo, el de medir la actividad relevante de las neuronas dado un vector de entrenamiento. En las siguientes secciones se exploran varias medidas en los contextos binarios y multi–clase.

4.1. Medidas discriminativas binarias

En un sentido general estas desemejanzas miden la *distancia* entre dos distribuciones discretas p y q. Los histogramas de las activaciones de salida de cada neurona se utilizan para aproximar las distribuciones de probabilidad de las dos clases: p y q y, al mismo tiempo, calcular las probabilidades que se utilizan en las siguientes medidas discriminativas. La primera medida que introducimos es la *información mutua* (MI) calculada para las activaciones de cada unidad oculta para ambas clases.

Información mutua

La información mutua de dos variables aleatorias es una medida de la dependencia entre ellas. Más específicamente, cuantifica la cantidad de información obtenida sobre una variable aleatoria a través de otra. Esta medida está estrechamente relacionada con la entropía H de una variable aleatoria X, con función de masa de probabilidad p, que mide la aleatoriedad de la variable dada, es decir, la media de la información proporcionada por un evento:

$$H(X) = -\sum_{x} p(x) \log_2 p(x) \tag{4.1}$$

La idea detrás de esta definición es que, si uno de los eventos es más probable que otros, la observación de ese evento es menos informativa. Por el contrario, los eventos más raros proporcionan más información cuando ocurren. En este sentido, es posible definir la información para un evento particular como $I(x) = -\log_2 p(x)$, por lo que su valor esperado sobre todos los valores posibles de x conduce a la entropía de Shannon (4.1). A partir de la entropía de Shannon podemos definir la entropía condicional de una variable aleatoria X dada la variable aleatoria Y por:

$$H(X|Y) = \sum_{x,y} p(x,y) \log_2 p(x|y) \tag{4.2}$$

donde $p(x, y)$ es la probabilidad conjunta de que $X = x$ y $Y = y$.

Otra definición que necesitamos para introducir el concepto de información mutua es la entropía conjunta, que mide cuánta incertidumbre hay en las dos variables aleatorias X y Y tomadas en conjunto y se define por:

$$H(X,Y) = -\sum_{x,y} p(x,y) \log_2 p(x,y) \tag{4.3}$$

De 4.2 y 4.3 se puede obtener:

$$H(X|Y) = H(X,Y) - H(Y) \tag{4.4}$$

Con lo anterior, la información mutua se define entonces como:

$$\begin{aligned} MI(X,Y) &= \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \\ &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{4.5}$$

Con esta ecuación en mente, en nuestra implementación, la información mutua está definida de la siguiente manera:

$$MI\left(Clase, Atributo\right) = H\left(Clase\right) + H\left(Atributo\right) - H\left(Clase, Atributo\right) \tag{4.6}$$

Divergencia de Kullback–Leibler

La divergencia de Kullback–Leibler (KL) es una medida discriminativa entre dos distribuciones de probabilidad. Dado que existen dos variables aleatorias discretas X y Y, descritas por las distribuciones de probabilidad $p(x)$ y $q(y)$, el modelo definido por p se evalúa en términos de *proximidad* a la distribución q. En otras palabras, esta divergencia mide la relación entre la probabilidad o incertidumbre de que una muestra de p se comporte como una muestra de q. Antes de definir la divergencia KL,

observemos que la *entropía cruzada* se define como:

$$\begin{aligned} H(X;Y) &= E_X[\log_2 \frac{1}{q(y)}] \\ &= -\sum_x p(x) \log_2 q(y) \end{aligned} \tag{4.7}$$

donde E_X representa la expectativa con respecto a la distribución de probabilidad p. La información KL, o entropía relativa de p con respecto a q, se puede definir como:

$$\begin{aligned} D_{KL}(p \parallel q) &= \sum_x p(x) \log_2 \frac{p(x)}{q(y)} \\ &= -\sum_x p(x) \log_2 q(y) \\ &\quad + \sum_x p(x) \log_2 p(x) \\ &= H(X;Y) - H(X) \end{aligned} \tag{4.8}$$

con $H(X;Y)$ siendo la entropía cruzada de X y Y mientras que $H(X)$ es la entropía (4.1) de X. Sin embargo, la divergencia de KL no es simétrica, aunque en nuestra implementación consideramos una medida simétrica, a la que nos referimos como KLS y a la que también se le conoce como divergencia de Jeffreys, definida como:

$$D_{KLS}(p \parallel q) = \frac{(D_{KL}(p \parallel q) + D_{KL}(q \parallel p))}{2} \tag{4.9}$$

Distancia de Wasserstein

La distancia de Wasserstein, o *Earth Mover's Distance* (EMD) [121], se basa en el costo mínimo a pagar, o trabajo por hacer, para transformar una distribución en otra. Es más robusto que otras técnicas que utilizan histogramas ya que opera con representaciones de distribuciones de tamaño variable, evitando así problemas con intervalos o bins que son típicos cuando se trabaja con histogramas. Intuitivamente hablando, dadas dos distribuciones, una puede ser considerada como una masa de tierra, mientras que la otra se puede ver como agujeros en el suelo que deben ser lle-

nados. Esto significa que la EMD mide el trabajo necesario para mover o transformar una distribución en otra, donde una unidad de trabajo corresponde a transportar una unidad de suelo en una unidad de distancia. La medida de distancia entre "lugares" se conoce como la distancia de suelo, que se introduce en (4.10). La EMD se define para histogramas de la forma $(\mu, p(x))$, donde μ es la media del histograma mientras que $p(x)$ es el número de ocurrencias de x. Los histogramas pueden o no normalizarse, de modo que la masa total de dos histogramas puede no ser igual. De esta forma, dados dos histogramas X y Y, la EMD se define en términos de *flujo óptimo* como $F = (f_{ij})$ que minimiza el trabajo W:

$$W(X, Y, F) = \sum_{i,j} f_{ij}\delta_{ij} \tag{4.10}$$

donde $\delta_{ij} = dist(\mu_i, \mu_j)$ es una cierta distancia entre μ_i y μ_j, por ejemplo la distancia euclidiana, mientras que $W(X, Y, F)$ es el trabajo necesario para mover la tierra de un histograma a otro. El flujo (f_{ij}) debe cumplir con las siguientes restricciones:

$$\begin{aligned} f_{ij} &\geq 0 \\ \sum_j f_{ij} &\leq \mu_i \\ \sum_i f_{ij} &\leq \mu_j \\ \sum_{i,j} f_{ij} &= min(\sum_i \mu_i, \sum_j \mu_j) \end{aligned} \tag{4.11}$$

La primera restricción nos permite mover la tierra de X a Y y no al contrario, la segunda limita la cantidad de tierra que se puede enviar desde X, la tercera limita la cantidad máxima de tierra que Y puede recibir y, finalmente, la cuarta nos obliga a mover la mayor cantidad posible de tierra *(flujo total)*. Una vez que el problema de transporte se resuelve y se encuentra el flujo óptimo F, la EMD se define como el trabajo W normalizado por el flujo total definido en la ecuación 4.11:

$$EMD(X,Y) = \frac{\sum_{i,j} f_{ij}\delta_{ij}}{\sum_{i,j} f_{ij}} \tag{4.12}$$

Diferencia absoluta de la frecuencia de activaciones

En [122], los autores describen un método para seleccionar los átomos más discriminativos de un diccionario fijo con el fin de mejorar el rendimiento de la clasificación de una red neuronal en una representación rala. Un diccionario se define como una matriz, $\Phi \in \mathbb{R}^{M\times N}$, cuyas columnas $\vec{\phi_j}$ son los átomos por lo que, una señal particular puede ser descrita por una combinación lineal de estos átomos, también conocidos como características de la señal.

La idea detrás de este método es seleccionar los átomos más discriminativos del diccionario usando la probabilidad de "activación" del átomo dada una clase. Se supone que un átomo está activo para una señal particular (de una clase dada) si el coeficiente correspondiente en su representación es diferente de cero. Los candidatos considerados son aquellos átomos con mayor diferencia absoluta entre las probabilidades de activación para cada clase. Es decir, un átomo es más discriminativo si está activo más veces para las señales que pertenecen a una clase, que para las señales pertenecientes a la otra clase. En nuestro trabajo podemos aplicar una idea similar a las neuronas en lugar de los átomos, nos referimos a ella como la *diferencia de la frecuencia de activación condicional* o *difference of conditional activation frequency* (DCAF):

Sea $p_i \triangleq p(x_i \neq 0|\vec{x} \in C_1)$ y $q_i \triangleq q(y_i \neq 0|\vec{y} \in C_2)$ las probabilidades de activación de la neurona i para la Clase 1 y la Clase 2, respectivamente. Nuestra implementación de este criterio es la siguiente:

$$D_{DCAF}(X,Y) = |p_i - q_i| \tag{4.13}$$

donde p_i y q_i se calculan utilizando la frecuencia de activación relativa para cada

neurona:

$$p_i \approx \frac{\#\text{ activación de la neurona } i \text{ para datos} \in C_1}{\#\text{ todos los datos} \in C_1}$$
$$q_i \approx \frac{\#\text{ activación de la neurona } i \text{ para datos} \in C_2}{\#\text{ todos los datos} \in C_2} \tag{4.14}$$

En el caso particular de las clases igualmente representadas (es decir, en un conjunto de datos equilibrado), ambos denominadores son iguales, por lo que los criterios pueden simplificarse sólo computando la diferencia absoluta entre el número de activaciones por cada clase.

Prueba t de varianzas desiguales

La prueba t de Welch es una adaptación de la prueba de t de Student [123, 124] que compara las medias de dos grupos. Es un buen enfoque cuando no se cumple la suposición de homogeneidad de varianzas, especialmente con tamaños de muestras desiguales. La idea general es que las medias de las activaciones de las neuronas pueden utilizarse para estimar que tan lejos están las distribuciones entre sí o, de alguna manera, probar si las unidades estadísticas que subyacen a las dos muestras se superponen.

Sea μ_i la media de las muestras de las activaciones para una neurona, *i=1,2* la clase, σ_i la varianza y n_i el tamaño del grupo. Entonces nuestra implementación de la prueba t de Welch es definida como:

$$WT = \frac{(\mu_1 - \mu_2)^2}{\frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}} \tag{4.15}$$

En el contexto de la clasificación de patrones en un problema de dos clases con distribuciones normales, esta prueba está relacionada con la denominada relación de Fisher [125], ya que en la ecuación 4.15 el numerador refleja la varianza inter–clase mientras que el denominador considera la varianza intra–clase. De esta forma, μ_1 es la media de las activaciones de salida de una neurona para patrones de la Clase 1 y μ_2

es la media de las activaciones de salida de una neurona para la Clase 2. Asimismo, σ_1 es la varianza de las activaciones de salida de una neurona para la Clase 1 y σ_2 es la varianza de las activaciones de salida de una neurona para la Clase 2.

4.2. Medidas discriminativas multiclase

En esta sección se revisan las distancias utilizadas para medir la discriminación de las neuronas en un contexto multi–clase. Las distancias presentadas a continuación se pueden dividir en dos grupos, modelos de filtro y modelos de envoltura. Basándose en las características de los datos, los modelos de filtro evalúan las características sin utilizar ningún algoritmo de clasificación. Un algoritmo de filtro típico consta de dos pasos: en el primer paso, ordena las características basadas en ciertos criterios mientras que en el segundo paso, las características en los lugares más altos se eligen para inducir modelos de clasificación. Estos modelos seleccionan características independientemente de un clasificador específico. Por lo tanto, el enfoque de filtro es independiente del algoritmo de aprendizaje, lo que lo vuelve computacionalmente simple, rápido y escalable. Sin embargo, la principal desventaja del enfoque es que ignora los efectos del subconjunto de característica seleccionado en el rendimiento global del algoritmo de clasificación. El subconjunto de características óptimas dependerá entonces de los sesgos específicos y la heurística del algoritmo de clasificación. Tomando en cuenta esta suposición, los modelos de envoltura utilizan un clasificador específico para evaluar la calidad de las características seleccionadas y ofrecen una forma sencilla y poderosa de abordar el problema de selección de características, independientemente del algoritmo de aprendizaje elegido. Sin embargo, los modelos de envoltura son muy costosos desde el punto de vista computacional por lo que el subconjunto de características seleccionadas está inevitablemente predispuesto al clasificador específico en comparación con los modelos de filtro. Este componente de búsqueda pretende encontrar un subconjunto de características con la evaluación más alta, usando una función heurística para guiarlo. Las evaluaciones de rendimiento se realizan típicamente utilizando validación cruzada.

El procedimiento de selección de características se divide en dos partes: evaluación de atributos y método de búsqueda. La evaluación de atributos es la técnica mediante la cual cada atributo o característica del conjunto de datos se evalúa en el contexto de la clase. El método de búsqueda es la técnica mediante la cual, se prueban las diferentes combinaciones de atributos en el conjunto de datos para obtener una lista de características ordenadas. En esta Tesis, el proceso de evaluación de atributos se lleva a cabo mediante el cálculo de las medidas dicriminativas para las neuronas, mientras que el de búsqueda se realiza estableciendo un orden sobre esos valores. Algunas de las medidas que se introducen en esta sección, en particular las que guardan una relación con información mutua y la correlación de Pearson, hacen uso intensivo del concepto de búsqueda a través de la evaluación iterada entre los valores discriminativos de cada atributo (neurona) y la variable de salida (clase). Esto quiere decir que se calculan las medidas de forma similar a como se haría con una medida binaria pero el proceso de búsqueda iterada posibilita la evaluación multiclase.

Anova

El propósito del ANOVA unidireccional es determinar si los datos de varios grupos tienen una media común. Es decir, el ANOVA unidireccional permite averiguar si existen diferencias estadísticamente significativas entre las medias de dos o más grupos independientes. El ANOVA de un solo sentido prueba si la varianza explicada en un conjunto de datos es significativamente mayor que la varianza inexplicada, esta técnica, sin embargo, no puede decir qué grupos específicos fueron estadísticamente significativos entre sí, sólo indica que hay diferencias estadísticamente significativa entre los grupos en su conjunto. Específicamente, prueba la hipótesis nula:

$$H_0 = \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_j \tag{4.16}$$

donde μ = media del grupo y J = número de grupos. El ANOVA unidireccional prueba la diferencia en la media del grupo dividiendo la variación total en los datos en dos componentes:

- Variación de la media del grupo de la media general, es decir, $\bar{y_j} - \bar{y}$ (variación entre grupos), donde $\bar{y_j}$ es la media muestral del grupo j, y $\bar{y}$ es la media global de la muestra.
- Variación de las observaciones en cada grupo de sus estimaciones de la media del grupo, $y_{ij} - \bar{y_i}$ (variación dentro del grupo).

Entonces, ANOVA compara la variación entre grupos y la variación dentro de los grupos. Si la relación entre la variación dentro del grupo y la variación entre grupos es significativamente alta, podemos concluir que las medias del grupo son significativamente diferentes entre sí. Esto se puede medir usando una estadística de prueba que tiene una distribución F con $(J-1, N-J)$ grados de libertad donde N es el número total de observaciones:

$$F = \frac{SSG/(J-1)}{SSE/(N-J)}, \tag{4.17}$$

donde la suma de cuadrados entre-grupos (SSG) y la suma de errores cuadráticos (SSE) son:

$$SSG = \sum_j n_j \left(\bar{y_j} - \bar{y}\right)^2 \tag{4.18}$$

$$SSE = \sum_i \sum_j \left(y_{ij} - \bar{y_j}\right)^2, \tag{4.19}$$

donde n_j es el tamaño de la muestra para el grupo j-th.

Correlación de Pearson

Aunque existen al menos 30 formas de interpretar y calcular el coeficiente de correlación de Pearson (PCC) [126], en este trabajo entendemos esta correlación como la covarianza normalizada, de forma que su rango esté entre 1 (correlación directa) y -1 (correlación inversa) con 0 denotando la ausencia de alguna relación. La idea detrás del uso de este coeficiente es que la correlación de las muestras dentro de una clase se espera mayor que la correlación entre clases [127], de esta forma, se impone un

orden a las neuronas mediante los valores de correlación. Este coeficiente, en términos de la covarianza de dos variables aleatorias (x, y), se puede expresar como:

$$\begin{aligned} r = r_{xy} &= \frac{cov\,(x,y)}{\sigma_x \sigma_y} \\ &= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)}\,\sqrt{(\sum y_i^2 - n\bar{y}^2)}} \end{aligned} \tag{4.20}$$

donde σ es la desviación estándar, n es el tamaño de la muestra y $\bar{x}, \bar{y}$ son las medias. Cuando x y y provienen de la misma clase, a este coeficiente se le interpreta como la correlación intra–clases mientras que, cuando provienen de clases diferentes, como la correlación entre–clases.

Puntaje de Fisher

La idea central del Fisher Score es encontrar distancias entre los valores de activación de las neuronas (o puntos de datos) de forma tal que, para la misma clase sean lo más pequeñas posible mientras que, para diferentes clases, sean lo más grandes posible. Esto se logra al evaluar los datos individualmente donde el 'score' o puntuación del j-ésimo punto será calculado como[128]:

$$S_j = \frac{\sum_{k=1}^{K} n_k \left(\mu_k^j - \mu^j\right)^2}{\sum_{k=1}^{K} n_k \left(\sigma_k^j\right)^2} \tag{4.21}$$

donde μ_k^j y σ_k^j son la media y la desviación estándar de la clase k-ésima para el dato j-ésimo respectivamente y, μ^j la media del conjunto de datos en la posición j-ésima de todas las clases. Después de calcular el puntaje de Fisher para cada punto de datos, se seleccionan aquellos con el puntaje más alto, permitiendo así ordenar los datos como 'más significativos'.

Ganancia de información[1]

La ganancia de información es un criterio para medir la "bondad" de las salidas de las neuronas en la tarea de clasificación, ésta mide la 'información ganada' al evaluar la presencia o ausencia de una neurona en la disminución de la entropía global. La entropía se considera una medida de la impredecibilidad del sistema por lo que evaluarla permite conocer la aleatoriedad de la variable dada, es decir, la cantidad de información proporcionada por un evento. La información mutua, definida en 4.5 utilizando la entropía H de la variable aleatoria X proporcionada por otra variable aleatoria Y, representa la cantidad en la que la entropía de X disminuye y está dada por:

$$MI\left(X,Y\right) = H\left(X\right) - H\left(X|Y\right) \tag{4.22}$$

Dada esta ecuación, la ganancia de información implementada en este trabajo está definida por:

$$IG\left(Clase, Atributo\right) = H\left(Clase\right) - H\left(Clase|Atributo\right) \tag{4.23}$$

Razón de ganancia[1]

Esta medida ofrece una puntuación normalizada de la contribución de las salidas de las neuronas en una tarea de clasificación, esto quiere decir que una *razón de ganancia* alta para una neurona, implica que será útil para la clasificación. Esta medida introduce un término de normalización a la ganancia de información definida en 4.23 dividiéndola por la entropía del *Atributo* que esta siendo evaluado. Debido a esta normalización, los valores de la razón de ganancia caen en el rango $[0, 1]$. Un valor de $GainR = 1$ indica que el *Atributo* predice completamente la *Clase*, y cuando $GainR = 0$, significa que no hay relación entre la *Clase* y el *Atributo*, esta razón se

[1]Esta medida está basada en Información Mutua, descrita en la subsección 4.1 de las medidas discriminativas binarias.

presenta como:

$$GainR\left(Clase, Atributo\right) = \frac{\left(H\left(Clase\right) - H\left(Clase|Atributo\right)\right)}{H\left(Atributo\right)} \tag{4.24}$$

Incertidumbre simétrica[2]

De forma similar a la *razón de ganancia* (4.24), el criterio de la *incertidumbre simétrica* normaliza la *ganancia de información* (4.23) dividiéndola por la suma de las entropías de la *Clase* y del *Atributo*. La ganancia de información es una medida simétrica, es decir, la cantidad de información obtenida sobre Y después de observar X es igual a la cantidad de información obtenida sobre X después de observar Y. Aunque la simetría es una propiedad deseable para una medida de correlación, la ganancia de información no normaliza los valores para garantizar que sean comparables y tengan el mismo efecto. Por este motivo la incertidumbre simétrica compensa el sesgo al normalizar su valor al rango $[0, 1]$ donde un valor de $SU = 1$, indica que un *Atributo* predice completamente la *Clase* (o al contrario) y $SU = 0$, significa que no están correlacionados. Este coeficiente de incertidumbre simétrica viene dado por:

$$symmUncert\left(Clase, Atributo\right) = 2 \times \frac{\left(GainR\right)}{H\left(Clase\right) + H\left(Atributo\right)} \tag{4.25}$$

Relief F

La medida Relief–F selecciona n instancias de activaciones neuronales al azar, calcula sus vecinos más cercanos y optimiza un vector de pesos. Esta evaluación de atributos asigna un peso a cada característica dependiendo de su capacidad para distinguir entre las clases. Específicamente, el cálculo del peso se basa en la probabilidad de que los vecinos más próximos de diferentes clases tengan valores diferentes para una instancia y la probabilidad de que los vecinos más cercanos de la misma clase

[2]Esta medida está basada en Información Mutua, descrita en la subsección 4.1 de las medidas discriminativas binarias.

tengan el mismo peso.

$$W_f = p(\text{diferentes valores de } f|\text{instancias cercanas de diferentes clases}) - p(\text{diferentes valores de } f|\text{instancias cercanas de la misma clase}) \quad (4.26)$$

La idea es que el algoritmo debe estimar la capacidad de los atributos para separar cada par de clases independientemente de cuáles sean las dos clases más cercanas entre sí.

One R

El algoritmo One–R construye una regla para cada atributo en los datos de entrenamiento y luego selecciona la regla con el menor error. Para crear una regla para un atributo es necesario construir una tabla de frecuencias para cada atributo en la clase, dichas reglas están basadas en una sola característica. Aunque es una forma mínima de clasificador, puede ser útil para determinar un punto de referencia para otros esquemas de aprendizaje. El algoritmo general es el siguiente:

Algoritmo 1 One–R

```
1: for each neurona
2:     for each valor de activación
3:         contar la frecuencia del valor de activación para cada clase
4:         encontrar la clase más frecuente
5:         crear una regla que asigne esa clase a dicho valor
6:     end for
7:     calcular el error total de las reglas de cada neurona
8: end for
9: escoger la neurona con el menor error total
```

4.3. Optimización de la estructura de una red neuronal

Uno de los problemas más frecuentemente encontrados en el ámbito de las redes neuronales y el aprendizaje profundo es encontrar una topología adecuada que mejore la representación de la información. Existen dos enfoques que pueden emplearse para encontrar este tipo de topologías: el enfoque constructivo o de crecimiento y, el enfoque destructivo o de poda. Por lo general, el método de poda comienza el entrenamiento con una red neuronal suficientemente grande como para asegurar un entrenamiento exitoso, posteriormente, algunas conexiones entre neuronas o incluso neuronas completas son removidas para continuar con el entrenamiento. Si el entrenamiento converge, se reanuda el ciclo de entrenamiento y poda. Si el entrenamiento falla, se asume que la red más pequeña que cumplió el criterio de convergencia tiene la topología más adecuada para el conjunto de datos dado. De manera similar, el enfoque constructivo o de crecimiento consiste en comenzar con sólo las neuronas de entrada y de salida y añadir nuevas neuronas en una capa oculta, hasta que se obtenga un resultado adecuado.

Independientemente del método de poda, evaluar las redes resultantes es difícil de lograr, principalmente por falta de criterios de optimización de las redes neuronales [118]. Un criterio que puede ser definido es la topología de red mínima, desafortunadamente, determinar cuándo una topología es realmente mínima es, por lo general, inviable. Además, el criterio de minimalidad varía para diferentes implementaciones y aplicaciones, por ejemplo, puede basarse en el número de capas, neuronas y conexiones, o una mezcla de éstas. Además del criterio de topología mínima, otro criterio que puede ser empleado es el de generalización que, como el anterior, es dependiente de otras características[129]. Para evaluar este criterio se debe tener en cuenta el tiempo de entrenamiento y el tamaño de la red. Por lo general, la eficiencia de los algoritmos de poda y crecimiento no se compara entre sí debido principalmente a que hacerlo requeriría que se establecieran criterios de optimalidad. Por lo que comparar métodos

de poda y optimización, de una manera teórica, es poco factible.

Antecedentes

Aún cuando la comparación teórica de los métodos de poda es difícil debido principalmente, a la necesidad de criterios de optimalidad y de un marco experimental unificado, se han propuesto algunas técnicas que buscan determinar una topología óptima a través de la evaluación de las salidas de la red. Este tipo de técnicas son conocidas como de análisis de sensibilidad. Estas técnicas permiten podar una red basándose en la relevancia o análisis de sensibilidad de la función de error E con respecto a un peso w. Esta medida se utiliza para cuantificar la contribución que los pesos o nodos individuales aportan a la solución para así poder eliminar los menos relevantes. La sensibilidad normalizada se define como:

$$S_w^E = \lim_{\Delta w \to 0} \frac{\frac{\Delta E}{E}}{\frac{\Delta w}{w}} = \frac{\partial \ln E}{\partial \ln w} = \frac{w}{E}\frac{\partial E}{\partial w} \tag{4.27}$$

Otras medidas de sensibilidad han sido propuestas, entre ellas se encuentran la técnica de "skeletonización" [130] que define la importancia de E con respecto de w como $S_w^E = -w\frac{\partial E}{\partial w}$. Esta medida ha sido utilizada en el método propuesto en [131] donde, las conexiones con menor importancia son podadas sin requerir un reentrenamiento posterior.

Un método basado en el análisis de sensibilidad se describe en [132]. La salida de cada neurona de la capa oculta es analizada para cada vector de entrenamiento después que la red ha convergido. Si la salida de una neurona oculta es constante para todos los vectores, entonces ello indicaría que dicha neurona funciona como sesgo y, por lo tanto, puede ser podada. De manera similar, si dos neuronas ocultas producen entre si los mismos valores de activación para los vectores, una de ellas puede ser podada. De igual manera, los pesos que son suficientemente pequeños son candidatos de poda. Posteriormente a la poda de neuronas y pesos, la red es reentrenada.

Alternativamente, un método basado en el análisis de sensibilidad que utiliza mo-

delos lineales para evaluar las unidades ocultas se describe en [133]. En este método, cuando una neurona oculta puede ser aproximada como un modelo lineal de la capa de entrada, puede ser podada. Una ventaja de este método es que después de podar, no es necesario reentrenar la red. Una extensión de este método lineal se puede encontrar en [134], ahí se presenta un algoritmo denominado poda de dependencia lineal que utiliza conjuntos de ecuaciones lineales, aunque esta propuesta mejora el modelo de [133], requiere el reentrenamiento de la red después de la poda. El procedimiento de poda descrito en [15] elimina iterativamente las unidades ocultas y luego ajusta los pesos restantes de tal manera que se preserve la eficiencia general de la red. Entonces, con estos métodos, la poda se formula como la resolución de un problema de ecuaciones lineales.

En [135] se propone un enfoque en dos fases para podar tanto las unidades de entrada de un MLP como las ocultas basándose en Información Mutua (MI). Todas las características de los vectores de entrada se clasifican de acuerdo con su relevancia. Así, las unidades de entrada de un MLP son evaluadas de acuerdo con su relevancia y contribución al desempeño de la red para, posteriormente, eliminar las características menos relevantes de los vectores de entrada. De la misma forma, las unidades ocultas que resulten ser redundantes se eliminarán iterativamente del MLP de acuerdo con la información mutua. Estos métodos evalúan la presencia o ausencia de un determinado peso o nodo. Dentro de este grupo, se encuentran la *Lesión Cerebral Óptima (Optimal Brain Damage)* [20] y la *Cirugía Cerebral Optima (Optimal Brain Surgeon)* [18, 19], que son dos métodos de poda neuronal basados en el análisis de perturbaciones de la expansión de Taylor de segundo grado de la función de error. En la siguiente ecuación, $\overrightarrow{w}$ representa el vector concatenado de los pesos W. Cuando el proceso de entrenamiento converge, el gradiente será cercano a cero y, por lo tanto, el aumento en el error E debido a un cambio en $\overrightarrow{w}$ está dado por:

$$\Delta E \simeq \frac{1}{2}\Delta\overrightarrow{w}^T\mathbf{H}\Delta\overrightarrow{w} \tag{4.28}$$

donde $\mathbf{H}$ es la matriz Hessiana, $\mathbf{H} = \frac{\partial^2 E}{\partial \overrightarrow{w}^2}$. Esta ecuación también se interpreta como

importancia (*saliency*), esto no es otra cosa que el estado o la calidad por la que se destaca un elemento en relación con sus vecinos. En este caso particular indica el cambio del error en relación con $\overrightarrow{w}$. Eliminar un peso w_i equivale a igualar ese peso a cero. Por lo tanto, la eliminación de un subconjunto de pesos, S_{prune}, da lugar a un cambio en E al asignar $\Delta w_i = w_i$ si $i \in S_{prune}$, si no, $w_i = 0$.

Por otro lado, la *Lesión Cerebral Óptima (OBD)* es un caso especial de la *Cirugía Cerebral Optima (OBS)* que está basado en 4.28, donde la matriz Hessiana $\mathbf{H}$ se asume diagonal. En este procedimiento se eliminan los pesos con la menor importancia (saliency) calculada como:

$$(\Delta E)_i \simeq \frac{1}{2} w_i^2 \mathbf{H}_{ii} \tag{4.29}$$

La *Poda de Componentes Pricipales (PCP)* [136] es un método intermedio entre OBD y OBS que utiliza como base el *Análisis de Componentes Principales (PCA)*, éste aplica PCA a las activaciones de los nodos de las capas de una red y, aunque se requiere la matriz de correlación de dichas activaciones en cada capa, se evita el cálculo de la matriz Hessiana de la función de error. Este método elimina los "eigen–nodos" menos importantes y, además, no requiere el reentrenamiento la red. Por otro lado, el *Daño Celular Óptimo (OCD)* [137] extiende la OBD permitiendo eliminar entradas irrelevantes y unidades ocultas. Para OBS también existe una extensión que permite eliminar unidades ocultas y entradas redundantes llamada *nodo-OBS* [138].

La principal desventaja de los métodos OBD y OBS, y algunas de sus extensiones, es que en caso de no encontrarse un mínimo local, por ejemplo cuando el entrenamiento se detiene prematuramente, el término de primer orden en la expansión de la serie Taylor no será cero, lo que convierte a estos métodos en poco adecuados, además, el reentrenamiento de la red suele ser necesario.

Poda discriminativa

El objetivo general de la poda discriminativa es proporcionar un mecanismo mediante el cual se identifiquen y poden los pesos y neuronas menos importantes de

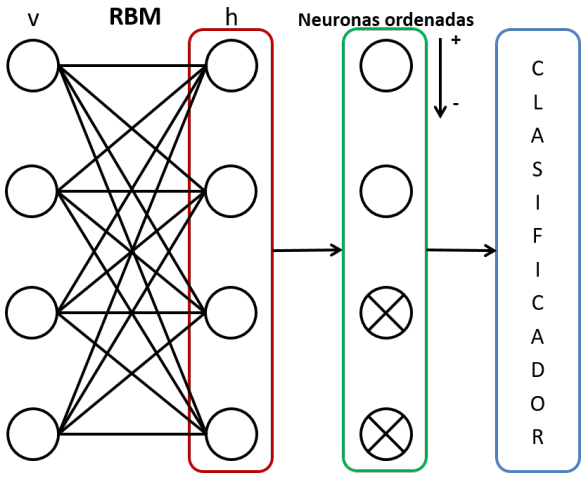
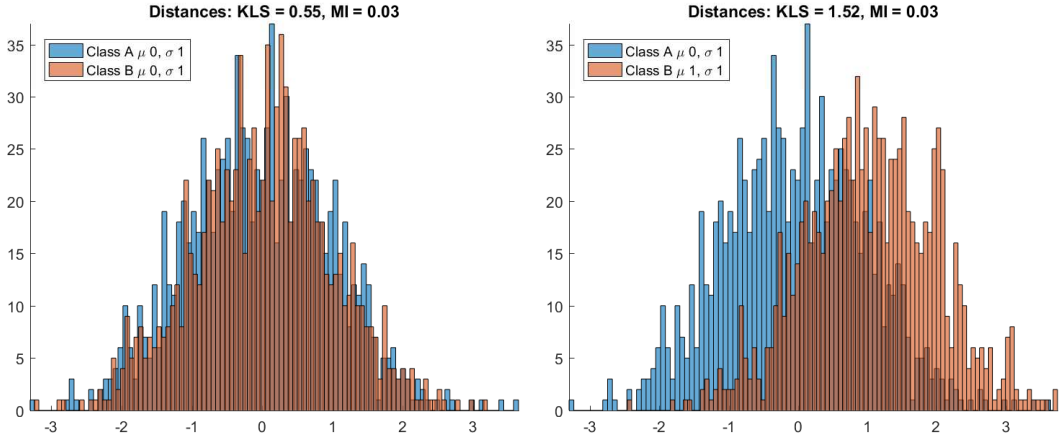


Figura 4-1: *Esquema general de poda discriminativa. Incluye un RBM entrenado de manera no supervisada, un método para clasificar las neuronas más discriminativas y un clasificador final.*

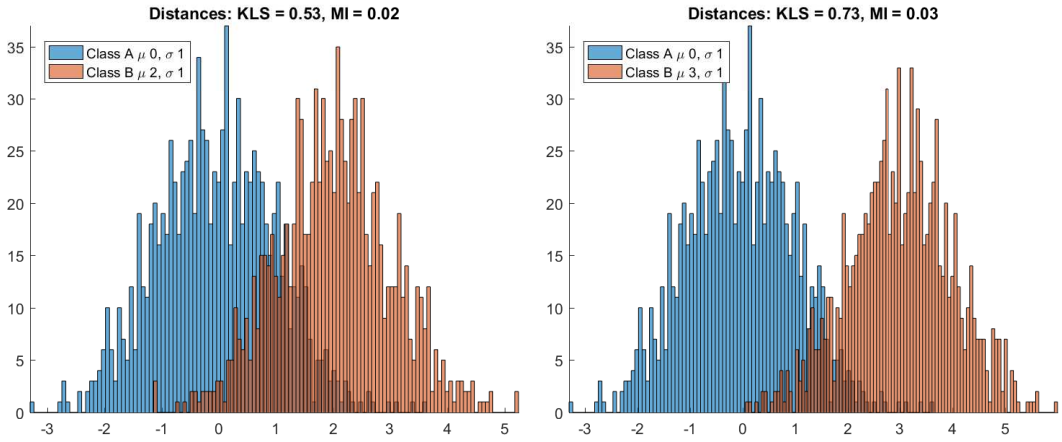
una red neuronal, este proceso se puede ver esquematizado en la Figura 4-1. En ella se puede ver que las activaciones de las neuronas seleccionadas como más discriminativas (cuyo número puede ser seleccionado trivialmente) sirven de entrada a un clasificador final que podría ser, incluso, otra red neuronal. En esta aproximación, la importancia de las unidades o neuronas se describe por su capacidad discriminativa, dicha capacidad es evaluada mediante las medidas descritas en las Seciones 4.1 y 4.2.

Determinar el éxito del entrenamiento de una RBM o DBM puede lograrse al observar el desempeño de las características aprendidas por la máquina de Boltzmann (BM), esto es, al finalizar el entrenamiento. Una problemática que resulta interesante analizar es el comportamiento de estas redes durante la fase de entrenamiento, ya que será beneficioso tener una comprensión más profunda de la dinámica de aprendizaje del modelo, más allá de analizar solamente el desempeño final, específicamente, es interesante analizar cómo las características estadísticas de cada neurona evolucionan y cómo se puede analizar durante el entrenamiento. Para realizar este análisis, en un sentido orientado a la poda neuronal, se realizaron varios experimentos que consistieron en examinar las distribuciones de las salidas de cada neurona en la capa de salida. Para ello se generaron artificialmente dos distribuciones Gaussianas pensando que cada una representa las activaciones de una neurona para una clase particular.

Al cambiar la media y desviación estándar de los datos fue posible analizar el comportamiento y el efecto que ejercen dichos cambios a algunas distancias o medidas de discriminación, además, es interesante notar que la información mutua no fue susceptible a variaciones de la media (Fig. 4-2), aunque sí a variaciones de la desviación estándar (Fig. 4-3). El resultado de este análisis dio pié a los experimentos posteriores.

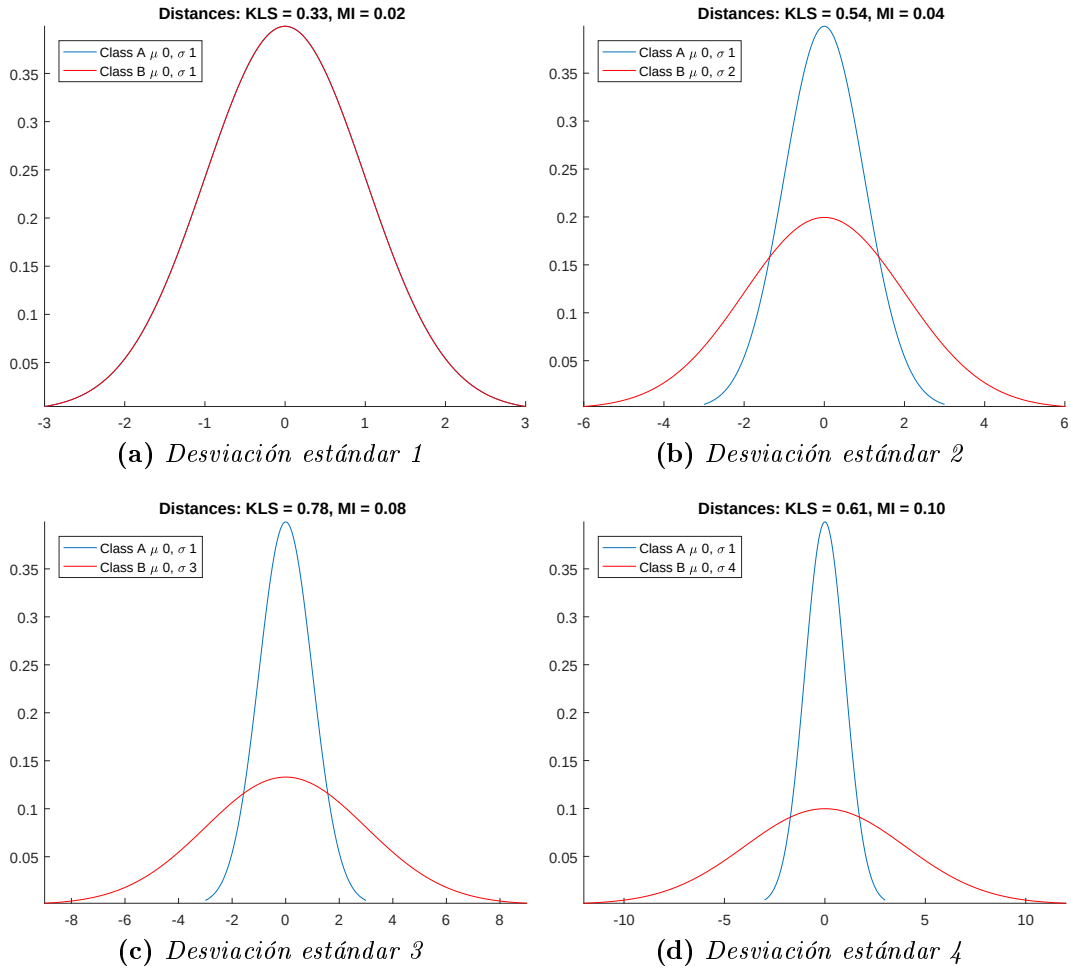


(a) *Diferencia entre medias de 0* **(b)** *Diferencia entre medias de +1*



(c) *Diferencia entre medias de +2* **(d)** *Diferencia entre medias de +3*

Figura 4-2: *Pruebas con datos artificiales (distribución normal con desviación estándar 1) modificando la media*



(a) *Desviación estándar 1* **(b)** *Desviación estándar 2*

(c) *Desviación estándar 3* **(d)** *Desviación estándar 4*

Figura 4-3: *Pruebas con datos artificiales (distribución normal con media 0) modificando la desviación estándar*

En el método propuesto, posterior al entrenamiento, se evalúan todas las neuronas para, en un paso posterior, seleccionar las mejores i neuronas. El algoritmo propuesto en esta Tesis para la evaluación, identificación y poda de las neuronas menos discriminativas se muestra a continuación:

Algoritmo 2 Evaluación discriminativa

Input: una RBM entrenada de manera no supervisada
1: **for each** clases
2: calcular los valores propagados en la capa oculta para cada vector de entrenamiento
3: **end for**
Output: activaciones de la capa oculta

Input: las salidas de la RBM anterior (los vectores propagados)
4: **for each** neurona i
5: estimar el histograma de los datos de salida para cada clase
6: calcular el poder discriminativo D_i de la neurona i para la medida discriminativa elegida.
7: **end for**
8: ordenar la neuronas de a cuerdo a su poder discriminativo en orden descendente.
Output: neuronas con un valor discriminativo asociado a cada una de ellas

Input: neuronas ordenadas
9: **for** $i \leftarrow 1$ **hasta** *número total de neuronas*
10: utilizar las primeras i neuronas para clasificar los datos mediante K vecinos más cercanos.
11: **end for**
Output: RBM podada

El siguiente capítulo presenta la experimentación que permitió elegir los parámetros necesarios para entrenar la RBM que sirve como primera entrada al algoritmo descrito antes. Esta experimentación toma como entrada los archivos de audio de dos bases de datos de habla emocional, una en idioma español y otra en alemán para resolver un problema de clasificación.

CAPÍTULO 5

EXPERIMENTACIÓN SOBRE CLASIFICACIÓN

En este capítulo se presenta la metodología de selección de parámetros utilizados en los experimentos realizados que tuvieron como objetivo principal, investigar la tarea de clasificación. A continuación se discuten los resultados obtenidos, se revisa la forma en que los datos fueron separados en los conjuntos de entrenamiento, prueba y validación y, se examinan las características extraídas de los datos. Las bases de datos emocionales que se utilizaron como parte de nuestros experimentos se describen en detalle.

5.1. Corpus de datos de habla española

Para llevar a cabo estos experimentos se utilizó la base de datos *INTERFACE*, esta base de datos fue creada en el Center for Language and Speech Technologies and Applications (TALP), de la Universidad Politécnica de Catalunya (UPC), con el propósito de estudiar el habla con emociones y la síntesis de voz. Las frases aquí contenidas expresan seis emociones más cinco variaciones del estado neutral, en cuatro idiomas de los cuales se utiliza el español junto con las transcripciones de las frases habladas. Cabe mencionar que los hablantes son actores profesionales, un hombre y una mujer [139].

Las emociones que se muestran a continuación son las más utilizadas en el análisis y síntesis del habla emocional; además, los estilos neutros también fueron definidos como una referencia a la expresión emocional, estas variaciones en el estilo son: lento, suave, fuerte y rápido como se muestra en la Tabla 5.1.

El corpus consiste de 184 oraciones incluyendo palabras aisladas, oraciones y el extracto de un texto en un contexto emocional neutral, también se han incluido las formas afirmativas e interrogativas de esas mismas oraciones. La distribución se puede ver en la Tabla 5.2.

Español	
6 emociones	A = enojo D = disgusto F = miedo J = alegría S = sorpresa T = tristeza
Variaciones de 'Neutral'	H = neutral/fuerte L = neutral/suave N = neutral/normal W = neutral/lento Z = neutral/rápido

Tabla 5.1: *Tipos de emociones en la base de datos INTERFACE*

Identificador (yyy)	Tipo de oración
001 - 100	Oraciones afirmativas
101 - 134	Oraciones interrogativas
135 - 150	Párrafos
151 - 160	Dígitos y números
161 - 184	Palabras aisladas

Tabla 5.2: *Tipos de oraciones en la base de datos INTERFACE*

Los creadores de esta base ponen a nuestra disposición el estudio que realizaron para evaluarla y que se muestra a continuación [139]. Dicha evaluación consistió en realizar pruebas subjetivas donde participaron 16 estudiantes de ingeniería de la UPC como oyentes no profesionales, en estas pruebas fueron reproducidas 56 oraciones, ocho

de ellas por cada una de las siete emociones. Estos datos son de mucha utilidad pues contra ellos podremos comparar los resultados obtenidos en esta propuesta. Cada oyente decidió qué emoción correspondía a cada expresión y la intensidad percibida en una escala de uno a cinco. Una segunda opción podía ser seleccionada en el caso de que la primera no fuera clara, además, para evitar que los oyentes tuvieran una referencia inmediata, las grabaciones fueron alternadas entre el hablante femenino y el masculino. Los resultados de esta prueba subjetiva muestran que más del 80 % de las frases fueron clasificadas correctamente con la primera elección y, de considerarse la segunda elección, más del 90 %, esto se constata en la Tabla 5.3. Cabe mencionar que cada expresión fue correctamente clasificada por al menos la mitad de los oyentes y que los errores fueron cometidos en las palabras o frases cortas, mientras que todas las oraciones y textos largos fueron clasificadas acertadamente en el primer intento por todos los oyentes.

	S	J	A	F	D	T	N	
S	89	20	7	0	6	2	4	128
J	0	115	7	0	2	2	2	128
A	2	14	85	2	5	5	15	128
F	4	1	1	103	5	13	1	128
T	2	1	2	5	106	3	9	128
D	1	3	1	16	3	101	3	128
N	0	2	2	1	4	1	118	128
	98	156	105	127	131	127	152	896

Tabla 5.3: *Resultados de la prueba subjetiva tomados de [139]. Los valores en las columnas representan el número de oraciones reconocidas contra las emociones reales en cada fila. A=enojo, D=disgusto, F=miedo, J=alegría, S=sorpresa, T=tristeza y N=neutral*

Características y parámetros

A pesar de que las características más utilizadas para el reconocimiento de emociones en la voz son los MFCCs [140], se ha discutido ampliamente el uso de otras como las prosódicas, que en conjunto, han reportado una importante mejora en la discriminación de emociones [141]. Teniendo esto en mente, propusimos el siguiente

conjunto de características que fueron extraídas de las grabaciones de audio utilizando la herramienta OpenSMILE [142]. Con esto, obtuvimos un vector de 30 dimensiones:

- 12 MFCCs y su primera derivada
- Promedio de F_0 y su primera derivada
- Promedio de los cruces por cero y su primera derivada
- Energía y su primera derivada

Cada uno de estos valores son una entrada del vector. Para comprender mejor los datos que componen los vectores que alimentan la red, se presenta el siguiente vector de ejemplo normalizado entre 0 y 1. Los valores en las posiciones 1-12 se corresponden con los 12 MFCCs, en las posiciones 13-24 está la primera derivada de los 12 MFCCs, la posición 25 es el promedio de F_0, la siguiente posición, 26, es el valor de la primera derivada del promedio de F_0, en la posición 27 está el promedio de los cruces por cero mientras que en la 28 se encuentra la primera derivada de los cruces por cero, en la penúltima posición está la energía y al final, en la posición 30, la primera derivada de la energía.

$$[0.7009, 0.5599, 0.5846, 0.5603, 0.3382, 0.3451, 0.3350, 0.2584, 0.4344, 0.4550,$$
$$0.7354, 0.1608, 0.4279, 0.5484, 0.3486, 0.4781, 0.4690, 0.4935, 0.4970, 0.5480,$$
$$0.3791, 0.4210, 0.4202, 0.5695, 0.2847, 0.2143, 0.1723, 0.6341, 0.4810, 0.6412\ ]$$

Para los experimentos de RBM y DBN modificamos la herramienta desarrollada por Drausin Wulsin [143], lo que permitió llevar a cabo un gran número de pruebas con el fin de determinar las mejores configuraciones y parámetros. Estos experimentos consistieron en diferentes combinaciones de tamaño de lote, tasa de aprendizaje, número de unidades ocultas y número de RBMs. En la Tabla 5.4 se muestran los valores de los parámetros con los que se realizaron los experimentos de manera exhaustiva, con todas las posibles elecciones de los parámetros. Las particiones de entrenamiento (70 %), prueba (25 %) y validación (5 %) se eligieron de manera aleatoria cuidando que se tuviera un número balanceado de emociones.

Parámetros	Valores
Tamaño del lote	[6, 12, 18, 24, 30, 36, 42, 48, 54, 60]
Tasa de aprendizaje	[0.01, 0.001, 0.0001, 0.00001]
Unidades ocultas	[28, 56, 84, 112, 140, 168]
Número de capas	[1, 2, 3, 4, . . . , 13, 14, 15]

Tabla 5.4: *Parámetros de configuración para el entrenamiento de las RBM y DBN*

Los experimentos con DBN se realizaron mediante la adición de una red RBM a un DBN previamente entrenado, y utilizando los parámetros mostrados en la Tabla 5.4. La capa de clasificación consta de siete unidades de salida, una para cada clase. La clase más probable fue considerada como la unidad con el nivel de activación más alto.

Para los experimentos realizados con máquinas de soporte vectorial (SVM) se utilizó la herramienta LIBSVM [144] y, para el resto de los clasificadores: K vecinos más cercanos (KNN), árboles de decisión (DT) y perceptrón multicapa (MLP) se utilizaron las herramientas disponibles a través de Matlab [145]. El SVM se utilizó con un núcleo radial. Para KNN se utilizaron tres vecinos y la medida de distancia de coseno. El DT fue construido usando el índice de diversidad de Gini, y luego podado con el fin de obtener mejores capacidades de generalización. El MLP fue entrenado de una manera tradicional con una capa oculta de diez unidades, todas las funciones de activación fueron sigmoides excepto en la última capa donde se usaron las funciones "tansig". También realizamos algunos experimentos "mixtos" donde los clasificadores: SVM, KNN, DT y MLP fueron alimentados con las salidas de un RBM.

Resultados y Discusión

Los resultados para las diferentes configuraciones presentadas en la Tabla 5.4 se muestran en la Figura 5-1. En ella se observa la tasa de error para cada arquitectura de red con una sola capa; con una neurona oculta, dos y hasta 168, siempre con 30 neuronas en la capa visible que se corresponden con el vector 30–dimensional. La combinación de parámetros que produjo el mejor resultado fue: 112 unidades ocultas, un tamaño de lote de 42 y una tasa de aprendizaje de 0.00001. Con esta configuración,

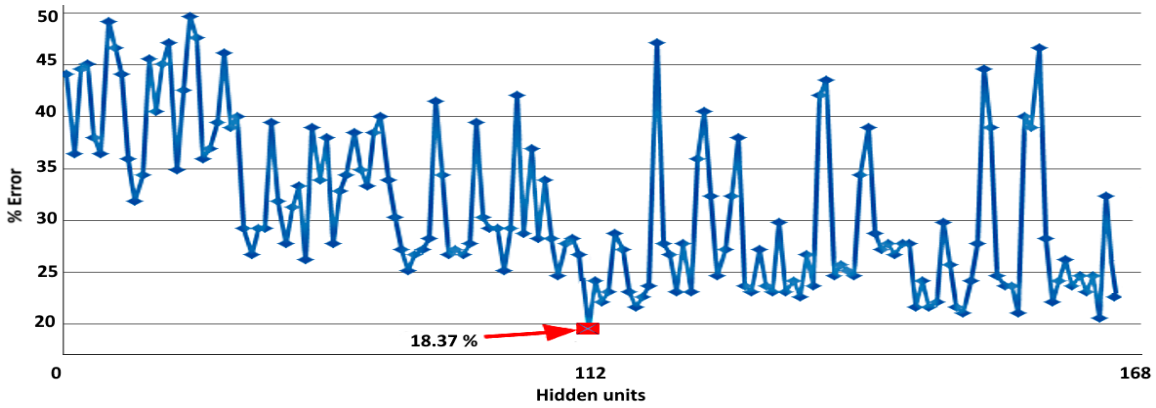


Figura 5-1: *Tasas de error de los experimentos con una RBM para las diferentes combinaciones de configuraciones vistas en la Tabla 5.4.*

el DBN alcanzó una tasa de error del 18.37 %.

Con el fin de realizar un segundo conjunto de experimentos, con varios RBM apilados, hemos utilizado el RBM con la mejor configuración encontrada en el experimento anterior, estos resultados se pueden ver en la Figura 5-2. En ella se observa la tasa de error para las diferentes arquitecturas de una DBN; con una, dos y hasta quince capas que preprocesan los vectores 30–dimensionales. En el caso de los clasificadores mixtos, se puede observar que la salida de la DBN ayuda a los clasificadores estándar a lograr un mejor rendimiento, como se puede ver en la tabla 5.5. Estos resultados se obtuvieron alimentando al clasificador con la salida de la DBN con tres capas entrenadas como se ha descrito anteriormente.

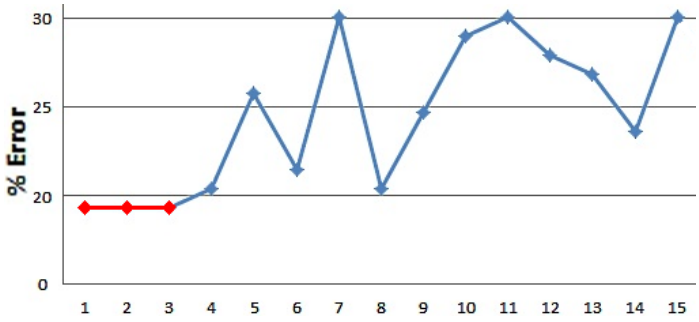


Figura 5-2: *Tasas de error de los experimentos con DBNs y varias RBMs apiladas. Los mejores resultados se obtuvieron con 1, 2 y 3 RBMs.*

Se puede ver que el mejor resultado se logró con sólo una, dos y tres capas de RBMs para después empeorar. Una posible explicación para este resultado es que se utilizó

Tabla 5.5: *Peor desempeño obtenido para cada clasificador.*

Clasificador	Tasa de error (%)
DBN - RBM	**18.37**
K-nn	31.63
DBN - K-nn	24.49
DT	34.69
DBN - DT	23.47
MLP	40.82
DBN - MLP	20.41
SVM	25.43
DBN - SVM	18.97

un pequeño conjunto de datos; cuantas más RBMs apiladas, más parámetros libres para entrenar, requiriendo así datos de entrenamiento adicionales. Es necesario realizar más investigación para poder demostrar esta afirmación. Los resultados obtenidos son comparables, y de hecho mejores que los resultados de otros clasificadores seleccionados, cuando los parámetros son elegidos correctamente. Estos resultados fueron presentados en el congreso *Mexican Conference on Pattern Recognition (MCPR)* y posteriormente publicados en [146].

5.2. Corpus de datos de habla alemana

En esta segunda aproximación empleamos una conocida base de datos emocional actuada en idioma alemán. La base de datos "Berlin Database of Emotional Speech" fue desarrollada en el Instituto de Ciencias de la Comunicación de la Universidad Técnica de Berlín [147] y es probablemente la base de datos más utilizada en este contexto. Contiene grabaciones en idioma alemán habladas por diez personas, cinco hombres y cinco mujeres a los que se les pidió que fingieran seis emociones diferentes (enojo, alegría, tristeza, miedo, disgusto y aburrimiento), así como un estado neutro en diez enunciados cada uno. Cinco de los diez enunciados consistían en una frase y las otras cinco en oraciones más largas. Debido a que el objetivo de de la base de

datos es el análisis fonético de las emociones y la síntesis del habla emocional, las grabaciones se realizaron bajo condiciones controladas y, por lo tanto, se caracterizan por una buena calidad de audio.

Después de las grabaciones se realizó una prueba de audición con 20 sujetos humanos que debían reconocer la emoción de cada enunciado y valorarlo por su naturalidad. Se descartaron las expresiones que fueron clasificadas erróneamente por más del 20 % de los oyentes o clasificadas como poco naturales por más del 40 %. El número de expresiones final fue de 535 donde la distribución de las emociones es, a excepción del enojo, relativamente equilibrada, como puede verse en la Tabla 5.6.

Tabla 5.6: *Distribución de emociones de la EmoDB*

Emoción	Número de frases	Tasa de reconocimiento subjetivo
Enojo	127	96.9 %
Aburrimiento	81	86.2 %
Disgusto	46	79.6 %
Miedo	69	87.3 %
Alegría	71	83.7 %
Tristeza	62	80.7 %
Neutral	79	88.2 %

Características y parámetros

Tal y como en el caso anteriormente estudiado, nuevamente se propuso el siguiente vector de características (vector de 30 dimensiones). Las arquitecturas, parámetros y procedimientos para encontrarlos fueron exactamente iguales a los discutidos en la sección anterior.

- 12 MFCCs y su primera derivada
- Promedio de F_0 y su primera derivada
- Promedio de los cruces por cero y su primera derivada
- Energía y su primera derivada

La diferencia radica en la propuesta de dos metodologías de validación para los clasificadores, una para asegurar la independencia del hablante y otra para tratar la independencia del texto o mensaje. Consecuentemente, considerando las características del corpus, se obtuvieron diez particiones para realizar experimentos independientes del hablante y ocho particiones para experimentos independientes del texto (frase). Ambos casos se realizaron bajo el esquema de "dejar uno fuera" [148], LOTO (dejar un texto fuera) y LOSO (dejar un hablante fuera). El MLP, que se utilizó como línea de base, tiene una capa oculta con ((# características + # classes) / 2) neuronas. Un 10 % del conjunto de entrenamiento se dejó para la prueba de generalización. El entrenamiento del MLP fue detenido en 500 épocas o cuando la red alcanzó el pico de generalización con los datos de prueba, este entrenamiento fue realizado con la ayuda de la herramienta *Weka Toolkit* [149].

Los experimentos con DBNs se realizaron mediante la adición de una capa RBM a un DBN previamente entrenado y utilizando los parámetros óptimos encontrados en el experimento anteriormente definido con un corpus en español (Tabla 5.4). Los parámetros para el entrenamiento de las RBMs y DBNs fueron *batch size* = 42, *tasa de aprendizaje* = 0.00001, *unidades ocultas* = 112 y *número de capas* = (1 + RBM). La capa de clasificación cuenta con siete unidades de salida, una para cada clase. Los clasificadores profundos fueron entrenados, con datos de prueba balanceados, hasta que el pico de generalización se alcanzó.

Resultados y Discusión

En esta sección, se presentan y discuten los resultados de los clasificadores propuestos en ambos esquemas. La Tabla 5.7 muestra el desempeño de los clasificadores para los experimentos de LOSO y LOTO. En la primera columna, se muestra el clasificador mientras que, en la segunda y tercera columnas se presentan los aciertos medios de cada clasificador para las tareas LOSO y LOTO. Los resultados indican que los clasificadores profundos tienen mejores resultados que los MLP en ambos esquemas. Además, en el esquema de LOSO la mejora es realmente significativa (8.67 % sobre la

línea base). Como puede verse, las emociones son bastante dependientes del hablante, además, los resultados son mejores (LOTO) cuando la independencia del hablante no se toma en cuenta. Estos resultados sugieren que el DBN podría ser utilizado en los esquemas más difíciles y que hay una correlación importante entre la emoción elicitada y los hablantes específicos.

Tabla 5.7: *Resultados de clasificación para los esquemas LOTO y LOSO.*

Clasificador	aciertos LOTO (avg)	aciertos LOSO (avg)
Perceptrón multicapa	**68.10** [%]	**51.65** [%]
DBN-RBM	**69.14** [%]	**60.32** [%]

También hemos evaluado si los resultados son estadísticamente significativos mediante el cálculo de la probabilidad de que un experimento dado sea mejor que nuestro clasificador [150]. Para realizar esta prueba asumimos la independencia estadística de los errores de clasificación para cada enunciado, de esta manera, para el esquema de LOSO tenemos que la confianza de la relación obtenida entre las tasas de error de DBN y MLP es $\mathsf{Pr}(err < err\_ref) > 99.85\,\%$. Por otro lado, la mejora utilizando el esquema LOTO no es significativa.

En este experimento se evaluaron las máquinas restringidas de Boltzmann y las redes de creencia profundas en el reconocimiento de emoción en el habla mediante dos metodologías de validación para asegurar la independencia del hablante y la independencia del mensaje. Los resultados muestran que los clasificadores profundos son mejores que un MLP tradicional, tanto en los esquemas de LOSO como de LOTO. Estos resultados fueron publicados en [151] y presentados en el congreso *Iberoamerican Congress on Pattern Recognition (CIARP).*

CAPÍTULO 6

EXPERIMENTACIÓN CON PODA DISCRIMINATIVA

En esta sección se presentan y discuten los resultados de la aplicación del método de poda neuronal en esquemas de clasificación binaria y multiclase. Los pasos generales utilizados en un enfoque multiclase se describen en el Algoritmo 2, mientras que su representación esquemática se puede ver en la Figura 6-1. Para ver una descripción más detallada sobre las medidas discriminativas se puede consultar el Capítulo 4.

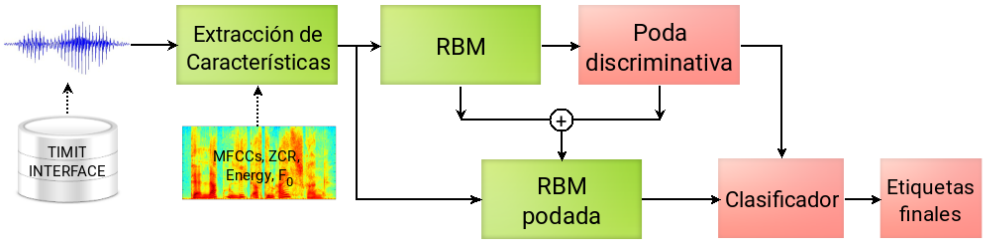


Figura 6-1: *Esquema conceptual del proceso general de los experimentos.*

6.1. Experimentos de poda binaria

En esta sección se describen los experimentos llevados a cabo mediante el uso de medidas discriminativas aplicadas a la poda neuronal en un contexto de clasificación binaria. El procedimiento general se presenta en la figura 6-1, donde el primer paso consiste en extraer las características de la señal de voz. En el segundo paso, una

RBM con N unidades ocultas se entrena de manera no supervisada. Los valores de N utilizados en este trabajo se especifican en la sección 6.1. Luego, el conjunto de entrenamiento se propaga a través de la RBM para obtener las activaciones de las unidades ocultas y, de acuerdo con el Algoritmo 2, cada unidad oculta es priorizada según las medidas discriminativas. Una vez ordenadas por capacidad discriminativa, se componen vectores que son clasificados en la etapa de clasificación final.

Es importante señalar que se implementó un método de validación cruzada aleatoria [152] para las bases de datos en los experimentos binarios. Esto nos permite obtener resultados más estables y evitar las estimaciones sesgadas del error de reconocimiento que suelen estar presentes en experimentos con sólo una partición de entrenamiento y prueba. Experimentar con un conjunto reducido de clases tiene como objetivo ejercer un mayor control sobre los experimentos mediante la reducción de variables, esta reducción permitió tener una mejor idea sobre el comportamiento del sistema.

A continuación se describen brevemente las bases de datos utilizadas para evaluar el enfoque binario.

Bases de datos

Base de datos TIMIT

TIMIT es un corpus de habla creado para el desarrollo y evaluación de sistemas automáticos de reconocimiento de voz por el Instituto de Tecnología de Massachusetts, SRI International y Texas Instruments, Inc. [153]. El corpus tiene enunciados de 630 hablantes expresados en los ocho principales dialectos del inglés americano, que incluyen transcripciones temporalmente alineadas ortográficas, fonéticas y de palabras. En los experimentos expuestos en este trabajo, la alineación fonética se utiliza para obtener archivos individuales de cada fonema. Además, se consideran todos los dialectos regionales, incluidos los hablantes de ambos sexos.

Como es de esperar, ciertos grupos de fonemas son más difíciles de clasificar que otros. Por ejemplo, el conjunto de fonemas en inglés: /b/, /d/, /eh/, /ih/ y /jh/ son difíciles de identificar [154, 155, 156]. De estos fonemas, aquí se consideran las vocales /eh/ y /ih/ debido a su cercanía en el espacio de formantes.

Los subconjuntos de prueba y entrenamiento definidos en esta base de datos se encuentran balanceados de manera fonética y dialectica, el conjunto de entrenamiento contiene muestras de 8904 audios, mientras que el conjunto de prueba tiene 3149 muestras.

INTERFACE

Este corpus, creado para estudiar el habla emocional, fue descrito en la sección 5.1. Para los experimentos se usaron dos clases (enojo y neutral) particionados en conjuntos de entrenamiento y prueba con el 70 % y 30 % de los datos respectivamente, utilizando como esquema de validación la validación cruzada aleatoria. Las particiones están equilibradas con respecto a los hablantes y las clases.

Base de datos de cáncer de seno

El cáncer de seno es uno de los más comunes, aunque ocurre en hombres y mujeres, el cáncer de mama masculino es raro. Algunos de los factores de riesgo son conocidos, por ejemplo: el envejecimiento, la genética, antecedentes familiares, períodos menstruales, no tener hijos y la obesidad. Todos ellos aumentan las probabilidades de que se desarrolle cáncer de mama, sin embargo aún no se sabe qué causa la mayoría de los cánceres de mama o exactamente cómo algunos de estos factores de riesgo hacen que las células se vuelvan cancerosas [157]. En este trabajo se utilizó la base de datos conocida como "Wisconsin Breast Cancer" (WBCD). Los datos fueron recogidos por el Dr. William H. Wolberg en los Hospitales de la Universidad de Wisconsin–Madison. En esta base de datos hay 569 registros, cada registro en la base de datos cuenta con 32 atributos (10 atributos por cada núcleo celular, un identificador y un diagnóstico)

que se detallan en la Tabla 6.1. En esta base de datos, 212 registros son diagnósticos malignos y 357 son benignos [158, 159].

Tabla 6.1: *Atributos de la base de datos "Wisconsin breast cancer"*

#	Atributo
1	ID
2	Diagnóstico
3	Radio
4	Textura
5	Perímetro
6	Área
7	Suavidad
8	Compacidad
9	Concavidad
10	Simetría
11	Fractalidad

Características y configuración

Para la base de datos de cáncer de seno, se utilizaron los 30 atributos que originalmente se encuentran en el corpus de los que se puede obtener información detallada en [158]. El proceso fue esquematizado en la sección anterior (Figura 6-1), donde las características sirven como entradas al RBM, y luego son enviadas desde la RBM al clasificador. Las salidas exactas recibidas por el clasificador dependen de la poda aplicada a las unidades ocultas, para poder evaluar la eficacia de este procedimiento es necesario aplicar un clasificador a la red podada. Varios clasificadores estándares podrían ser aplicados en este bloque: K-vecinos más cercanos (KNN), árboles de decisión, perceptrones multicapa (MLP), y máquinas de soporte vectorial (SVM), entre otros [148]. En este trabajo aplicamos un clasificador 1-NN, los resultados presentados a continuación son calculados sobre la media de clasificación de las salidas de las redes podadas. Con el fin de determinar si el proceso de poda es beneficioso para lograr resultados de clasificación adecuados, se definió una línea de error base *(línea*

base) utilizando un RBM no podado como entrada al clasificador. Adicionalmente, el proceso propuesto también puede ser visto como selección de características en RBMs pre-entrenadas.

Resultados y discusiones

En esta sección se presentan y discuten esencialmente tres tipos de resultados para cada base de datos, los cuales consideran diferentes aspectos del proceso de poda de acuerdo con el Algoritmo 2. El primer resultado, considera el error de clasificación en relación con el número de unidades ocultas que se utilizan en la etapa de clasificación luego de la poda, respecto de las distintas medidas discriminativas (Figuras 6-2 – 6-4). En el segundo, se presenta un análisis que permite estimar qué porcentaje de neuronas se requiere para obtener un error de clasificación razonable (Figuras 6-5 – 6-7 para dos de las medidas). Por último se presenta un índice, inspirado en el análisis de componentes principales, para proporcionar más información sobre el proceso de poda (Figura 6-8). La linea base propuesta consiste en clasificadores RBM+Knn, utilizados en su forma estandar (sin poda), y los resultados están representados en las figuras por una línea continua, mientras que los intervalos de confianza (CI) están representados por una línea punteada.

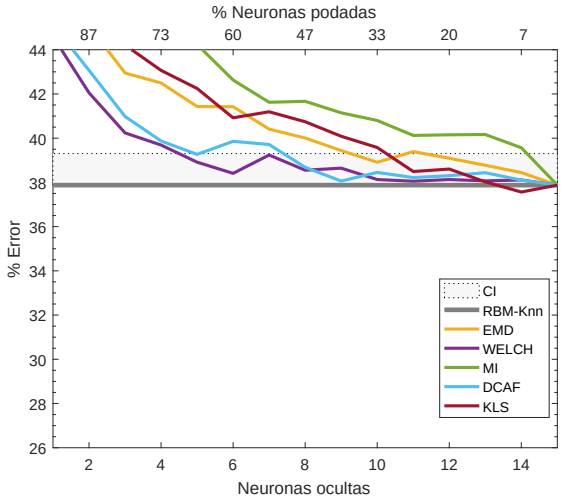
Los resultados que se exponen en esta sección se computaron promediando 10 experimentos, que hacen uso de la metodología de validación cruzada con el objetivo de obtener comportamientos y tendencias más estables y fiables en las tasas de error. Para las bases de datos de habla se implementaron inicialmente cinco configuraciones de RBM: la entrada de la red siempre consta de 30 unidades visibles, mientras que el número de unidades ocultas fue de 15, 30, 60, 120 y 240 (es decir, 0.5, 1, 2, 4 y 8 veces el número de unidades visibles) por otro lado, para la base de datos de cáncer de seno se utilizaron hasta 480 neuronas ocultas. Dado que los resultados presentados aquí son resultados promediados, utilizamos un *intervalo de confianza*, definido por la ecuación 6.1, y calculado con un nivel de confianza del 95 %. Los intervalos de

confianza se calculan utilizando:

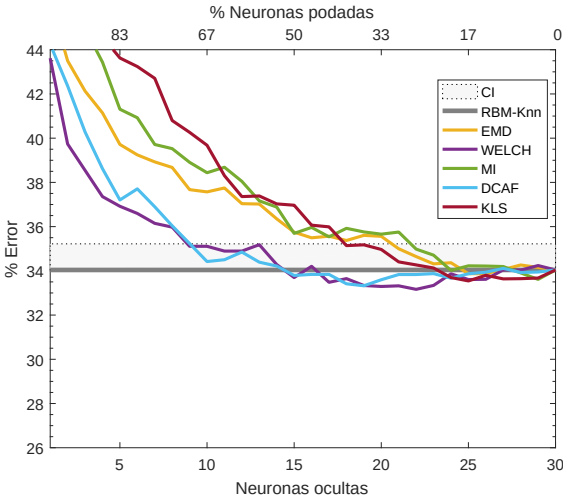
$$CI = \left(\bar{x} - t\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + t\frac{\sigma}{\sqrt{n}} \right) \tag{6.1}$$

donde $\bar{x}$ es la media de las muestras, σ es la desviación estándar de las muestras, n es el número de muestras y t es la *t-estadística* al 95 % con 9 grados de libertad.

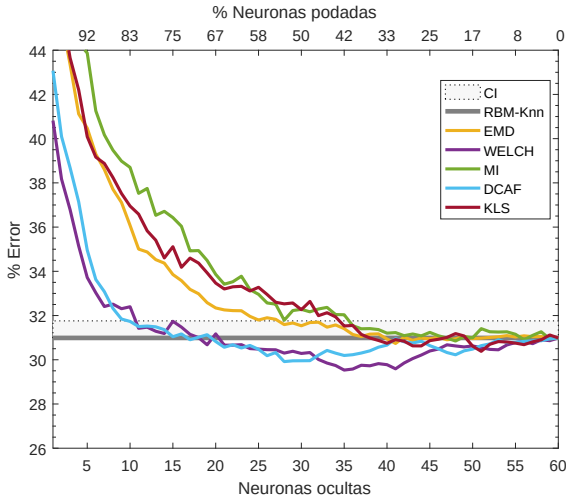
Los experimentos con el corpus TIMIT se muestran en la Figura 6-2. El primero (6-2a), muestra que cuando la arquitectura de la RBM inicial tiene menos neuronas en la capa oculta que la dimensión de los vectores de entrada, no se mejora el resultado de la línea de error base. El segundo (6-2b), muestra que utilizar 30 unidades ocultas es mejor que utilizar 15 pues el error, con respecto a la línea de error base, se mejora. Al mismo tiempo, la primera vez que esto ocurre es alrededor de 15 unidades, el mismo número de unidades ocultas que en el experimento (6-2a). En las Figuras 6-2c–e se puede observar que aún se puede disminuir el error, esto indicaría que es correcta la hipótesis de que una red más grande puede elegir mejores fronteras de decisión, mientras que la poda ayuda a elegir la información más relevante.



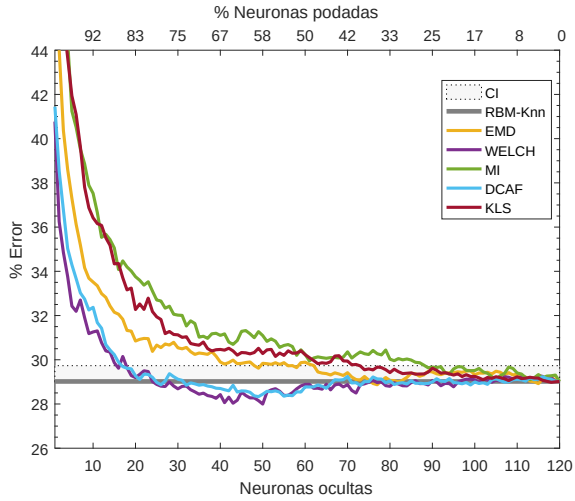
(a) *Poda de la RBM con 15 unidades ocultas.*



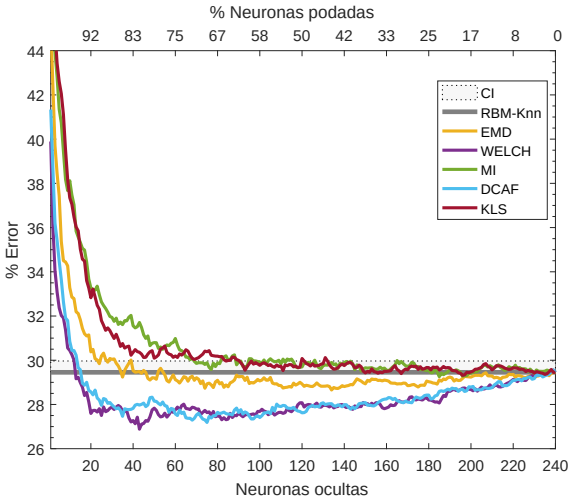
(b) *Poda de la RBM con 30 unidades ocultas.*



(c) *Poda de la RBM con 60 unidades ocultas.*



(d) *Poda de la RBM con 120 unidades ocultas.*

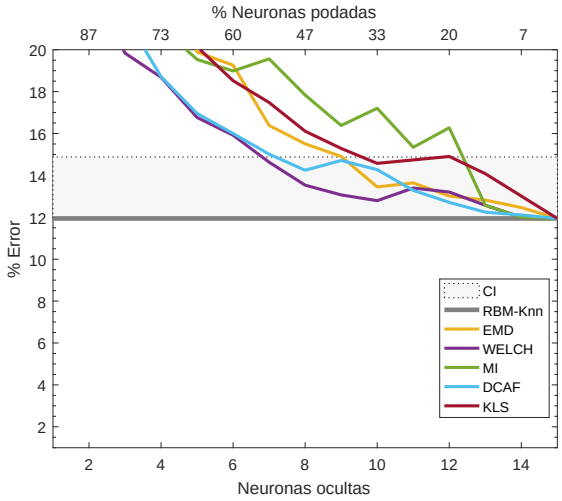


(e) *Poda de la RBM con 240 unidades ocultas.*

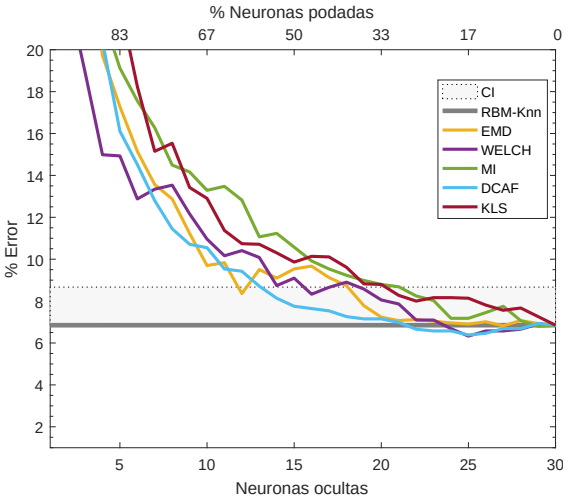
Figura 6-2: *Resultados de la clasificación sobre la base de datos TIMIT. Las pruebas se realizaron utilizando 15, 30, 60, 120 y 240 neuronas iniciales.*

Los resultados de aplicar la misma metodología de los experimentos anteriores, aunque ahora utilizando la base de datos INTERFACE, se presentan en la Figura 6-3. El comportamiento general de estos experimentos es similar a los observados con el conjunto de datos TIMIT: el error base es mayor cuando se usan inicialmente 15 unidades ocultas y, la poda de esta red no mejora el error. Los experimentos que utilizan más unidades ocultas en la configuración inicial (Figuras 6-3b, c, d, e) permiten lograr menores errores de clasificación. Sin embargo, a diferencia del segundo experimento de TIMIT (Figura 6-2b), 30 unidades ocultas no son suficientes para mejorar el rendimiento. Esto puede ser causado por varios factores, por ejemplo, que se utilizan frases completas y no fonemas para calcular los vectores de características por lo que la complejidad del corpus de INTERFACE es mayor.

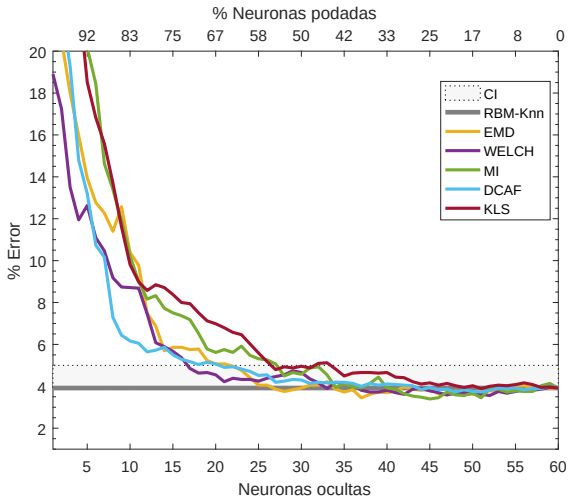
En los experimentos con la base de datos de cáncer de seno se puede observar una tendencia similar a los resultados obtenidos con los otros corpus, éstos son presentados en la Figura 6-4. En dicha Fig. 6-4a se puede apreciar que al iniciar con 15 neuronas no se alcanza una mejora apreciable en la tasa de error. Por otro lado, cuando se comienza con 30 neuronas (Fig. 6-4b) se reduce de manera importante el número de neuronas requeridas para obtener menores tasas de error, aunque no es hasta que la arquitectura de la red comienza con 120 neuronas (Fig. 6-4d) que el error disminuye en un 2 % y se puede podar significativamente la RBM. Es importante destacar que comenzar con 240 y 480 neuronas ocultas (Fig. 6-4e y Fig. 6-4f), no se mejoran los resultados de la configuración de 120, por lo que se podría decir que se ha alcanzado el punto máximo entre generalización y especialización con estas técnicas para este caso.



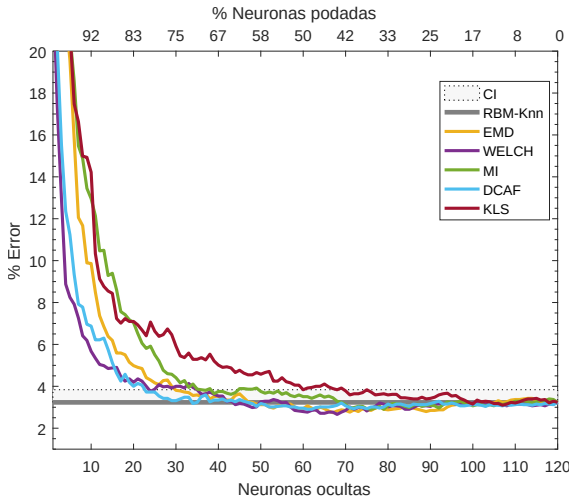
(a) *Poda de la RBM con 15 unidades ocultas.*



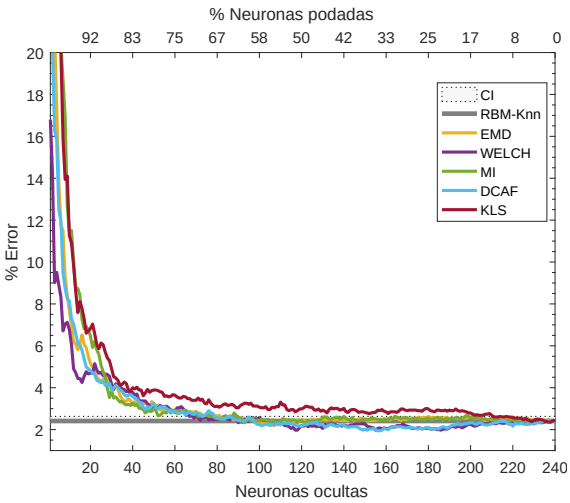
(b) *Poda de la RBM con 30 unidades ocultas.*



(c) *Poda de la RBM con 60 unidades ocultas.*

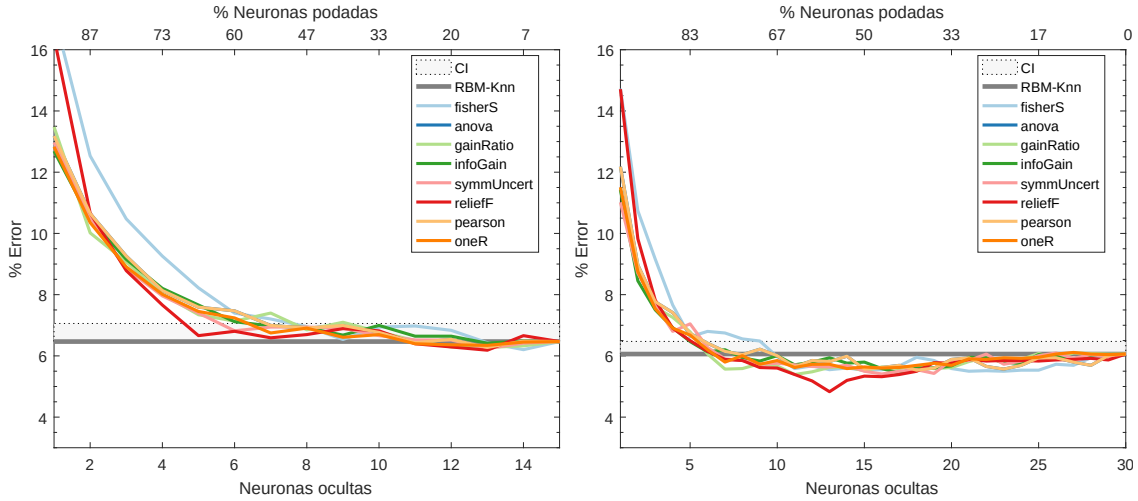


(d) *Poda de la RBM con 120 unidades ocultas.*

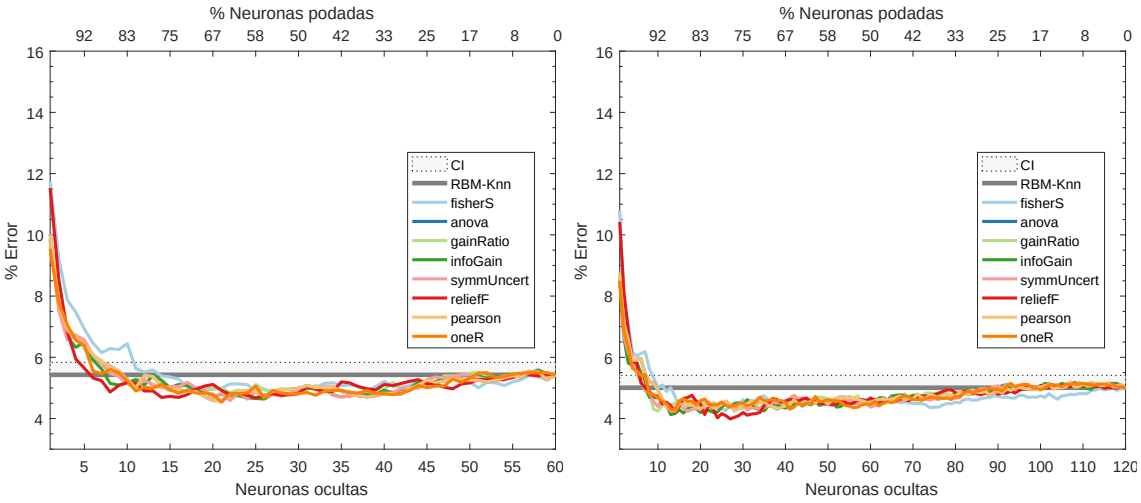


(e) *Poda de la RBM con 240 unidades ocultas.*

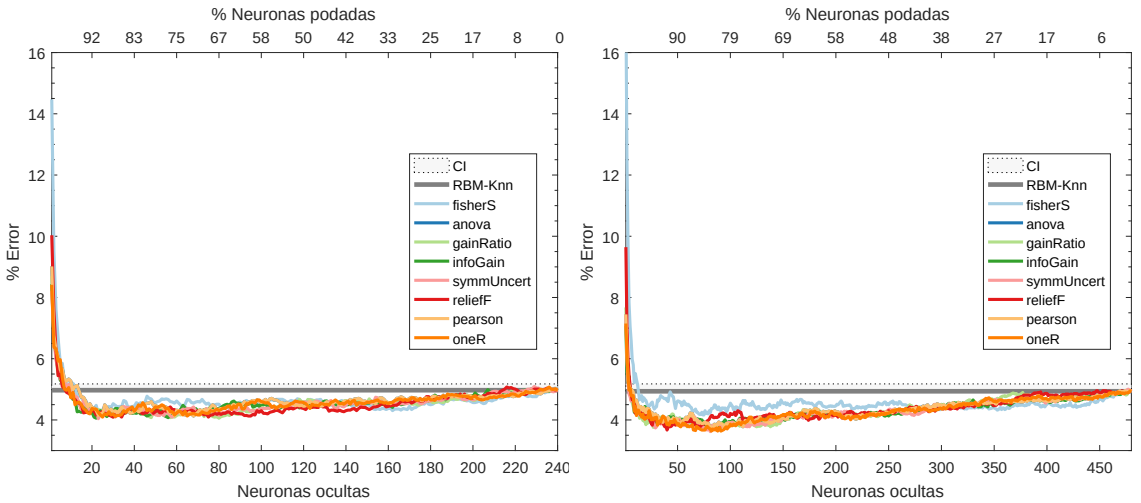
Figura 6-3: *Resultados de la clasificación sobre la base de datos INTERFACE. Las pruebas se realizaron utilizando 15, 30, 60, 120 y 240 neuronas iniciales.*



(a) *Poda de la RBM con 15 unidades ocultas.* **(b)** *Poda de la RBM con 30 unidades ocultas.*



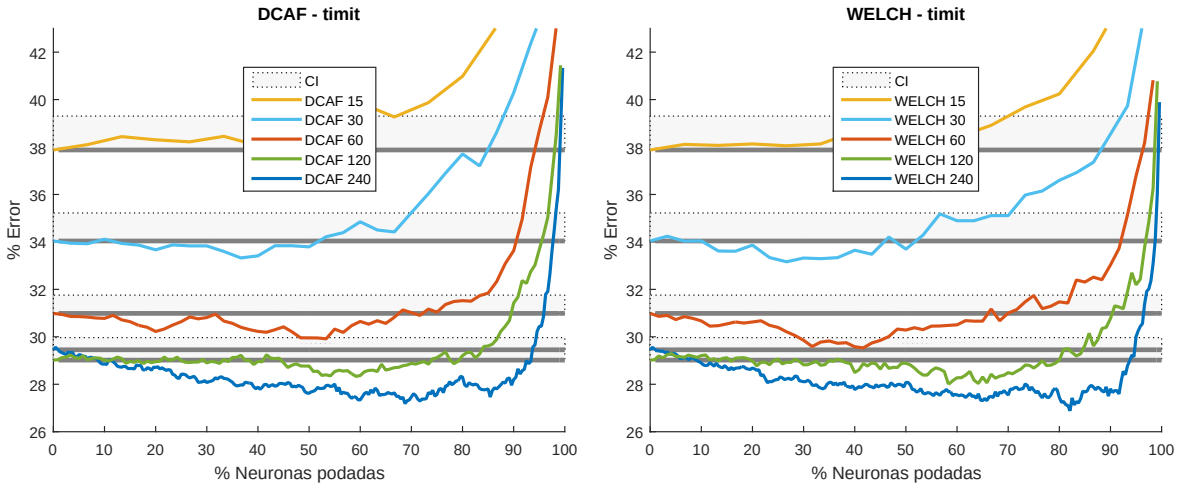
(c) *Poda de la RBM con 60 unidades ocultas.* **(d)** *Poda de la RBM con 120 unidades ocultas.*



(e) *Poda de la RBM con 240 unidades ocultas.* **(f)** *Poda de la RBM con 480 unidades ocultas.*

Figura 6-4: *Resultados de la clasificación sobre el corpus Breast Cancer. Las pruebas se realizaron utilizando 15, 30, 60, 120, 240 y 480 neuronas iniciales.*

Se puede observar una tendencia en los experimentos presentados en las Figuras 6-2, 6-3 y 6-4 para las distancias Welch y DCAF y, hasta cierto punto para EMD, donde alrededor del 20 % - 30 % de las mejores unidades son suficientes para alcanzar e incluso mejorar la línea de error base, dando así una posible reducción de al menos el 70 % de las unidades ocultas. Por lo tanto, las medidas de Welch y DCAF producen redes con menos neuronas que alcanzan tasas de error aceptables más rápidamente. El comportamiento es similar para la base de datos de cáncer de seno, aunque para las medidas Fisher Score y relief-F. Para analizar estas tendencias, se muestran las Figuras 6-5, 6-6 y 6-7, donde es posible ver en detalles las dos mejores medidas para cada base de datos utilizada. En todas las figuras se pueden observar los resultados de utilizar un RBM no podado como entrada al clasificador (línea de error base) representados por una línea continua, mientras que los intervalos de confianza (CI), definidos en la ecuación 6.1, están representados por una línea punteada.

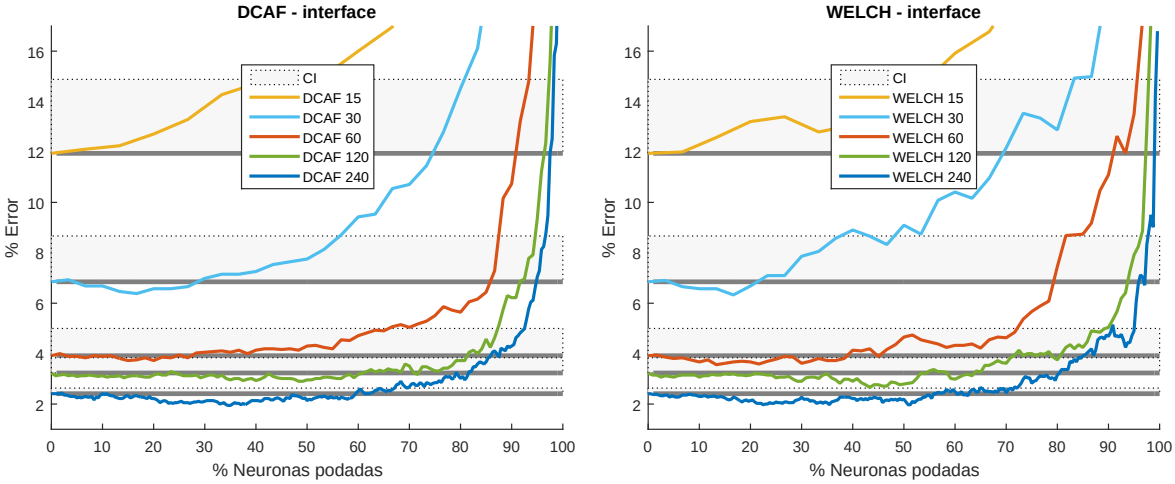


(a) *DCAF con 15, 30, 60, 120 y 240 unidades.* **(b)** *Welch con 15, 30, 60, 120 y 240 unidades.*

Figura 6-5: *Curvas de error para las medidas DCAF y Welch para la base de datos TIMIT. Las pruebas se realizaron utilizando 15, 30, 60, 120 y 240 neuronas iniciales.*

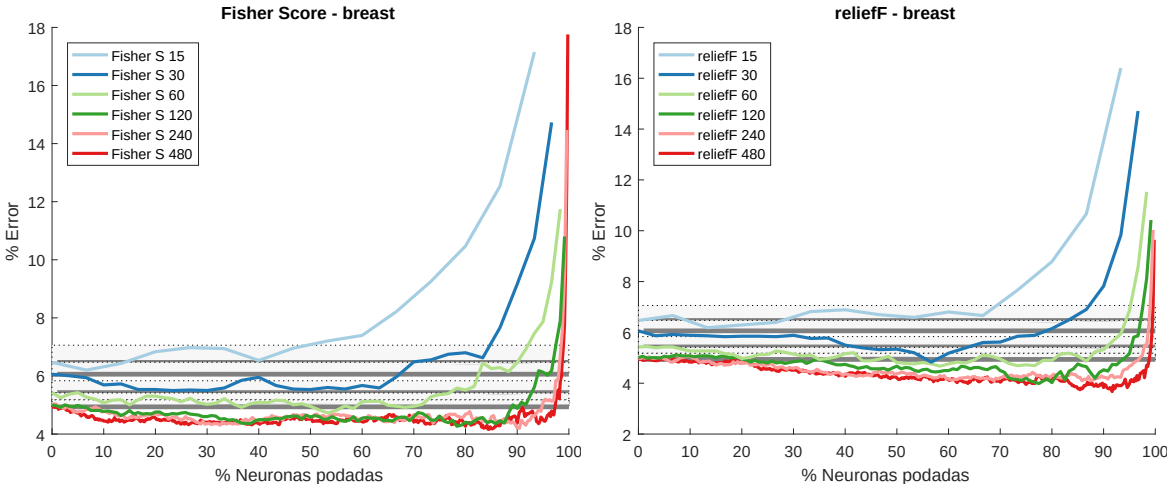
Las Tablas 6.2, 6.3 y 6.4 proporcionan más información sobre estas medidas, en ellas se muestra la arquitectura inicial de la red, el error obtenido con esta red inicial, los detalles de la red más pequeña dentro del intervalo de confianza y la mejor tasa de error para esa arquitectura en particular. Al inspeccionar las Tablas podemos

observar que utilizar más unidades iniciales en las redes resulta, generalmente, en menores tasas error. Los resultados presentados hasta este punto, para ambos corpus de habla, fueron publicados en [160].



(a) *DCAF con 15, 30, 60, 120 y 240 unidades.* **(b)** *Welch con 15, 30, 60, 120 y 240 unidades.*

Figura 6-6: *Curvas de error para las medidas DCAF y Welch para la base de datos INTERFACE. Las pruebas se realizaron utilizando 15, 30, 60, 120 y 240 neuronas en la RBM.*



(a) *Fisher Score con 15, 30, ..., y 480 unidades.* **(b)** *relief-F con 15, 30, ..., y 480 unidades.*

Figura 6-7: *Curvas de error para las medidas Fisher Score y relief-F para el corpus Breast Cancer. Las pruebas se realizaron utilizando 15, 30, 60, 120, 240 y 480 neuronas en la RBM.*

El número de unidades podadas después de lo cual sólo se obtienen errores de clasificación aceptables, proporciona un punto de poda adecuado, aunque de ninguna manera es el *mejor*. Es posible decir que para aprovechar de forma correcta la metodología propuesta en este trabajo, es necesario utilizar al menos la misma cantidad de neuronas ocultas que de neuronas visibles, ya que utilizar menos no beneficia los resultados de clasificación, como puede verse en las Figuras 6-2 y 6-3 para todas las medidas. Con esto en mente, tratamos de estimar un mejor punto de poda usando la ecuación 6.2 al proporcionar un equilibrio entre la tasa de error y el número de unidades podadas, en lugar de centrarse únicamente en el número de neuronas podadas. Esta ecuación calcula la Ganancia Acumulada Discriminativa Relativa (RDCG) y es análoga a una ecuación similar utilizada en el Análisis de Componentes Principales [161] empleada para reducción de dimensionalidad. La RDCG se calcula utilizando:

$$RDCG_j = \frac{\sum_{i=1}^{j} d_i}{\sum_{i=1}^{I} d_i} \tag{6.2}$$

donde j es el número elegido de unidades retenidas (no podadas) y d_i es el valor de una de las medidas para la neurona i.

Tabla 6.2: *Resultados de clasificación sobre la base de datos TIMIT utilizando las mejores dos distancias.* ⋆ *Configuración inicial,* † *Neuronas no podadas.* [1]*Primer configuración que entra al intervalo de confianza,* [2]*Mejor rendimiento para poda.*

Welch							
⋆	% Error base	†	% Error[1]	% Ahorro	†	% Error[2]	% Ahorro
15	**37.87**	7	39.23	53.33	15	**37.87**	0.00
30	34.03	13	35.18	56.66	22	**33.15**	26.60
60	30.98	15	31.74	75.00	35	**29.53**	41.60
120	29.01	18	29.64	85.00	50	**28.00**	58.30
240	29.45	12	30.00	95.00	43	**26.88**	82.08
DCAF							
15	**37.87**	5	39.26	66.66	15	**37.87**	0.00
30	34.03	9	35.23	70.00	19	**33.32**	36.60
60	30.98	10	31.73	83.33	28	**29.92**	53.30
120	29.01	17	29.68	85.83	49	**28.33**	59.16
240	29.45	13	30.22	94.58	75	**27.19**	68.75

Tabla 6.3: *Resultados de clasificación sobre la base de datos INTERFACE utilizando las mejores dos distancias. ⋆ Configuración inicial, † Neuronas no podadas. [1]Primer configuración que entra al intervalo de confianza, [2]Mejor rendimiento para poda.*

Welch							
⋆	% Error base	†	% Error[1]	% Ahorro	†	% Error[2]	% Ahorro
15	**11.94**	7	14.63	53.33	15	**11.94**	0.00
30	6.84	17	8.65	43.33	25	**6.32**	16.66
60	3.91	17	4.84	71.66	52	**3.56**	13.33
120	3.23	35	3.83	70.83	68	**2.65**	43.33
240	2.41	67	2.63	72.08	117	**1.97**	51.25
DCAF							
15	**11.94**	7	15.01	53.33	15	**11.94**	0.00
30	6.84	13	8.71	56.66	25	**6.38**	16.66
60	3.91	18	5.04	70.00	48	**3.72**	20.00
120	3.23	23	3.72	80.83	61	**2.90**	49.16
240	2.41	73	2.63	69.58	156	**1.94**	35.00

Tabla 6.4: *Resultados de clasificación sobre la base de datos Breast Cancer utilizando las mejores dos distancias. ⋆ Configuración inicial, † Neuronas no podadas. [1]Primer configuración que entra al intervalo de confianza, [2]Mejor rendimiento para poda.*

Fisher Score							
⋆	% Error base	†	% Error[1]	% Ahorro	†	% Error[2]	% Ahorro
15	6.47	8	6.96	46.67	14	**6.20**	6.66
30	6.06	10	5.97	66.67	21	**5.49**	30.00
60	5.43	11	5.64	81.67	28	**4.72**	53.33
120	5.01	9	5.32	92.50	26	**4.26**	78.33
240	4.97	7	5.16	97.08	23	**4.19**	90.41
480	4.94	12	5.17	97.50	75	**4.16**	84.37
relief-F							
15	6.47	5	6.66	66.67	13	**6.18**	13.33
30	6.06	6	6.15	80.00	13	**4.83**	56.66
60	5.43	5	5.64	91.67	25	**4.67**	58.33
120	5.01	6	5.16	95.00	27	**3.98**	77.50
240	4.97	6	5.09	97.50	30	**4.09**	87.50
480	4.94	5	5.13	98.96	40	**3.67**	91.66

En la Figura 6-8 se muestra la RDCG para las medidas discriminativas utilizadas en la poda de cada base de datos. En el eje de las abscisas se observa el número de neuronas mientras que en el de las ordenadas, lo que podría interpretarse como el

porcentaje de explicación de los datos. Así pues, un valor de RDCG de 0.7 y un número de neuronas de 200 significaría que este número de neuronas es el requerido para explicar el 70 % de los datos, un valor de RDCG de 0.8 y 150 neuronas querría decir que el 80 % de la información es representado por 150 neuronas y así sucesivamente. Aunque la evolución de la RDCG es monótona creciente, las curvas de información mutua y Anova se comportan de manera opuesta debido a que, si nos remontamos a sus definiciones en las secciones 4.1 y 4.2 respectivamente, vemos que los valores de ambas medidas son negativos.

La idea de que este índice proporciona información útil para determinar un posible punto de poda se puede explorar con más detenimiento al analizar las Tablas 6.5, 6.6 y 6.7, que son similares a Tablas 6.2, 6.3 y 6.4, utilizando tres valores de RDCG, 0.7, 0.8 y 0.9 para las medidas de Welch y DCAF para los corpus de habla y, Fisher Score y relief-F para el de cáncer de seno. Aquí, 0.8 parece ser un compromiso razonable entre el número de neuronas y el porcentaje de explicación.

Tabla 6.5: *Resultados de clasificación sobre la base de datos TIMIT utilizando las mejores dos distancias y RDCG como estimador del punto de poda. ⋆ Configuración inicial, † Neuronas no podadas.*

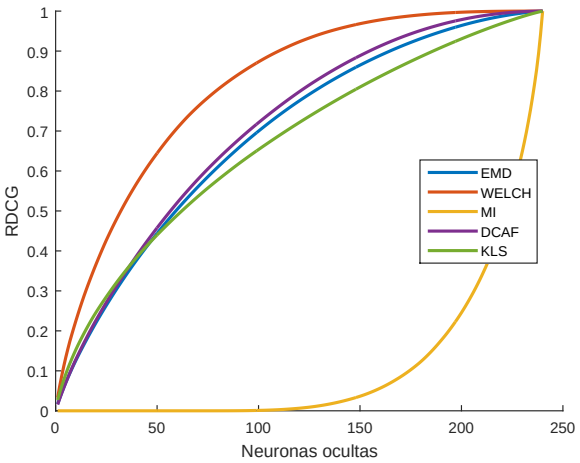
Welch										
		RDCG 0.7			RDCG 0.8			RDCG 0.9		
⋆	% Error base	†	% Error	% Ahorro	†	% Error	% Ahorro	†	% Error	% Ahorro
15	37.87	4	39.69	73.33	5	**38.91**	66.67	7	39.24	53.33
30	34.03	8	35.98	73.33	10	**35.11**	66.67	14	34.28	53.33
60	30.98	15	31.74	75.00	20	**31.17**	66.67	28	30.03	53.33
120	29.01	30	28.69	75.00	41	**28.06**	65.83	56	28.39	53.33
240	29.45	59	27.64	75.42	79	**27.44**	67.08	109	27.64	54.58
DCAF										
		RDCG 0.7			RDCG 0.8			RDCG 0.9		
⋆	% Error base	†	% Error	% Ahorro	†	% Error	% Ahorro	†	% Error	% Ahorro
15	37.87	6	39.86	60.00	7	**39.71**	53.33	9	38.06	40.00
30	34.03	11	34.50	63.33	14	**34.22**	53.33	18	33.41	40.00
60	30.98	24	30.65	60.00	30	**29.96**	50.00	39	30.57	35.00
120	29.01	47	28.53	60.83	60	**28.77**	50.00	77	29.13	35.83
240	29.45	96	27.34	60.00	121	**27.63**	49.58	155	27.99	35.42

Tabla 6.6: *Resultados de clasificación sobre la base de datos INTERFACE utilizando las mejores dos distancias y RDCG como estimador del punto de poda. ⋆ Configuración inicial, † Neuronas no podadas.*

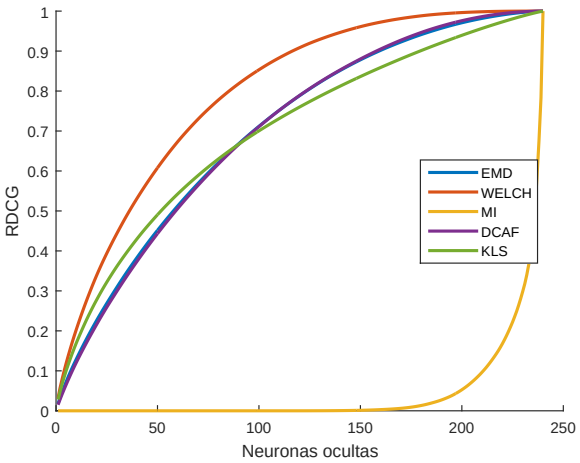
Welch										
		RDCG 0.7			RDCG 0.8			RDCG 0.9		
⋆	% Error base	†	% Error	% Ahorro	†	% Error	% Ahorro	†	% Error	% Ahorro
15	11.94	5	16.77	66.67	7	**14.63**	53.33	9	13.07	40.00
30	6.84	9	12.16	70.00	11	**10.16**	63.33	15	9.09	50.00
60	3.91	17	4.84	71.67	22	**4.38**	63.33	30	4.65	50.00
120	3.23	32	3.91	73.33	42	**3.26**	65.00	58	2.84	51.67
240	2.41	65	2.82	72.91	86	**2.46**	64.16	117	1.97	51.25
DCAF										
		RDCG 0.7			RDCG 0.8			RDCG 0.9		
⋆	% Error base	†	% Error	% Ahorro	†	% Error	% Ahorro	†	% Error	% Ahorro
15	11.94	6	16.00	60.00	7	**14.01**	53.33	10	14.27	33.33
30	6.84	11	9.53	63.33	14	**8.13**	53.33	18	7.26	40.00
60	3.91	23	4.82	61.67	29	**4.32**	51.67	38	4.16	36.67
120	3.23	47	3.15	60.83	60	**2.95**	50.00	76	3.01	36.67
240	2.41	98	2.32	59.17	124	**2.30**	48.33	158	1.97	34.17

Tabla 6.7: *Resultados de clasificación sobre la base de datos Breast Cancer utilizando las mejores dos distancias y RDCG como estimador del punto de poda. ⋆ Configuración inicial, † Neuronas no podadas.*

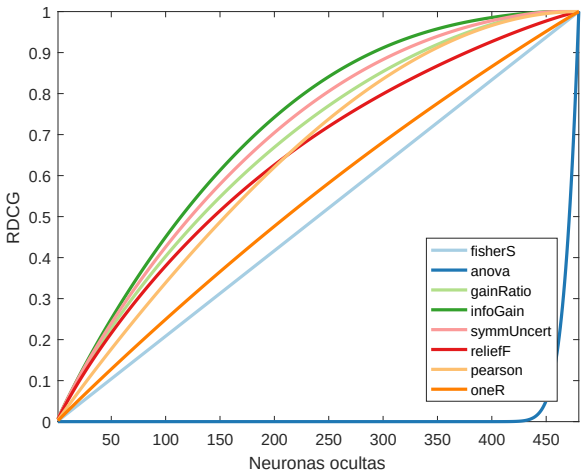
Fisher Score										
		RDCG 0.7			RDCG 0.8			RDCG 0.9		
⋆	% Error base	†	% Error	% Ahorro	†	% Error	% Ahorro	†	% Error	% Ahorro
15	6.47	11	6.98	26.67	12	**6.83**	20.00	14	6.20	6.67
30	6.06	21	5.50	30.00	24	**5.53**	20.00	27	5.69	10.00
60	5.43	42	5.02	30.00	48	**5.30**	20.00	54	5.08	10.00
120	5.01	84	4.60	30.00	96	**4.76**	20.00	108	4.78	10.00
240	4.97	168	4.36	30.00	192	**4.71**	20.00	216	4.78	10.00
480	4.94	336	4.41	30.00	384	**4.55**	20.00	432	4.44	10.00
relief-F										
		RDCG 0.7			RDCG 0.8			RDCG 0.9		
⋆	% Error base	†	% Error	% Ahorro	†	% Error	% Ahorro	†	% Error	% Ahorro
15	6.47	8	6.70	46.67	10	**6.82**	33.33	12	6.29	20.00
30	6.06	16	5.32	46.67	20	**5.76**	33.33	24	5.85	20.00
60	5.43	30	4.81	50.00	38	**4.97**	36.67	48	4.97	20.00
120	5.01	60	4.67	50.00	75	**4.78**	37.50	94	4.83	21.67
240	4.97	120	4.37	50.00	151	**4.34**	37.08	189	4.80	21.25
480	4.94	239	4.16	50.21	301	**4.39**	37.29	377	4.78	21.46



(a) *Comportamiento del índice RDCG para TIMIT.*



(b) *Comportamiento del índice RDCG para INTERFACE.*



(c) *Comportamiento del índice RDCG para Breast Cancer.*

Figura 6-8: *Ganancia Acumulada Discriminativa Relativa (RDCG) de una RBM con 240 unidades iniciales para TIMIT e INTERFACE, y 480 para Breast Cancer.*

6.2. Experimentos de poda multiclase

En la sección previa se investigó el uso de cinco medidas discriminativas diferentes como una forma de evaluar la utilidad de cada unidad oculta en una RBM para resolver un problema de clasificación binario. En dichos experimentos se utilizó información sobre las clases ya que el objetivo fue medir cuánto del poder discriminativo de la red es aportado por cada neurona de la capa oculta manteniendo, al mismo tiempo, un desempeño de clasificación adecuado. Los resultados mostraron ahorros (poda) de más del 50 % de las neuronas en la red neuronal de tamaño completo, reduciendo el vector que será clasificado en una etapa posterior, al mismo tiempo que se mantuvo una tasa de error aceptable y en algunos casos mejorándola.

En contraste con el problema de clasificación binario, los resultados de esta sección se presentan para problemas de varias clases aplicando un esquema específico de validación. La validación cruzada dejando uno fuera o Leave-one-out cross-validation separa el conjunto de datos disponibles en dos subconjuntos, uno utilizado para entrenar el modelo y otro para realizar el test de validación. En el caso de las bases de datos de habla, este esquema separa para validación del proceso de entrenamiento, los audios de un solo hablante *leave-one-speaker-out* (LOSO) o, una sola frase *leave-one-text-out* (LOTO). Para el caso de otro tipo de bases de datos, se elige una clase que no formará parte del conjunto de entrenamiento. Este proceso se repite tantas veces como clases distintas se tengan, eligiendo una de ellas para dejar fuera (leave-one-out) y ser utilizada para validación.

Bases de datos

Tres conjuntos de datos se utilizaron en estos experimentos, todos ellos con diferentes propiedades en términos del número de clases, de características, de instancias y tipos de características. Algunos de estos datos han sido utilizados en investigaciones similares [162, 157], de forma análoga, las distancias relacionadas con *información mutua* han sido abordadas en contextos discriminativos [101, 162, 163].

Tabla 6.8: *Características de las Bases de datos utilizadas.*

#	Nombre	Instancias	Atributos	Clases
1	Interface	5,113	30	7
2	Berlín	535	30	7
3	Sensores de gas	13,910	128	6

Las primeras dos bases de datos se han descrito detalladamente en el Capítulo 5 aunque, a diferencia de la experimentación anterior, en esta ocasión se utilizan las siete clases de ambos corpus. La tercera base de datos, conocida como "Gas Sensor Array Drift Dataset at Different Concentrations" se describe a continuación.

Base de datos de sensores de gas

La base de datos "Gas Sensor Array Drift Dataset at Different Concentrations" contiene 13,910 mediciones de 16 sensores químicos utilizados en una tarea de discriminación de 6 gases a varios niveles de concentración; Amoniaco, Acetaldehído, Acetona, Etileno, Etanol y Tolueno. El conjunto de datos fue recolectado durante 36 meses en una plataforma de suministro de gas situada en el laboratorio ChemoSignals del Instituto BioCircuits de la Universidad de California en San Diego [164]. En la siguiente Tabla 6.9 se concentran los detalles de esta base de datos.

Tabla 6.9: *Distribución de gases*

#	Gas	Número de mediciones
1	Amoniaco	1,641
2	Acetaldehído	1,936
3	Acetona	3,009
4	Etileno	2,926
5	Etanol	2,565
6	Tolueno	1,833

Características y configuración

Para las dos bases de datos de habla, Interface y Berlín, de la misma forma que en el experimento binario presentado en este mismo capítulo, se eligió un conjunto estándar de características representado por un vector 30-dimensional. La otra base de datos tiene su propio conjunto de características establecido y, ya que el objetivo del estudio aquí presentado no se centra en el censado de químicos, una descripción detallada de las características de la base de datos se puede encontrar en [164].

En estos experimentos se utilizó el esquema de validación leave-one-out, donde el total de los datos se dividen en k subconjuntos, de manera que aplicamos el método *leave-one-out* k veces, utilizando cada vez un subconjunto distinto para validar el modelo entrenado con los otros $k-1$ subconjuntos. En particular, para estos experimentos se eligió $k = 10$.

El esquema presentado en la Figura 6-1 esboza el enfoque de la poda de la red. Para poder evaluar la eficacia de este procedimiento es necesario aplicar un clasificador a la red podada. Varios clasificadores estándares podrían ser aplicados en este bloque: K-vecinos más cercanos (KNN), árboles de decisión, perceptrones multicapa (MLP), y máquinas de soporte vectorial (SVM), entre otros [148].

Con el objetivo de comparar los resultados y tendencias de los experimentos multiclase con los binarios, se eligió el mismo clasificador K-vecinos más cercanos con $K = 1$ para la tarea de clasificación multiclase. Los resultados presentados a continuación son calculados como la media de clasificación de las salidas de las redes podadas. Con el fin de determinar si el proceso de poda es beneficioso para lograr resultados de clasificación adecuados, se definió una línea base utilizando un RBM no podado como entrada al clasificador 1-NN.

Para las bases de datos se implementaron distintas configuraciones de RBM que se detallan en la Tabla 6.10, la elección de la arquitectura de la red se basa en la experiencia obtenida con la experimentación de clasificación, en los resultados de poda binaria, el número de instancias, clases y atributos, entre otras.

Tabla 6.10: *Distribución de neuronas para cada base de datos.*

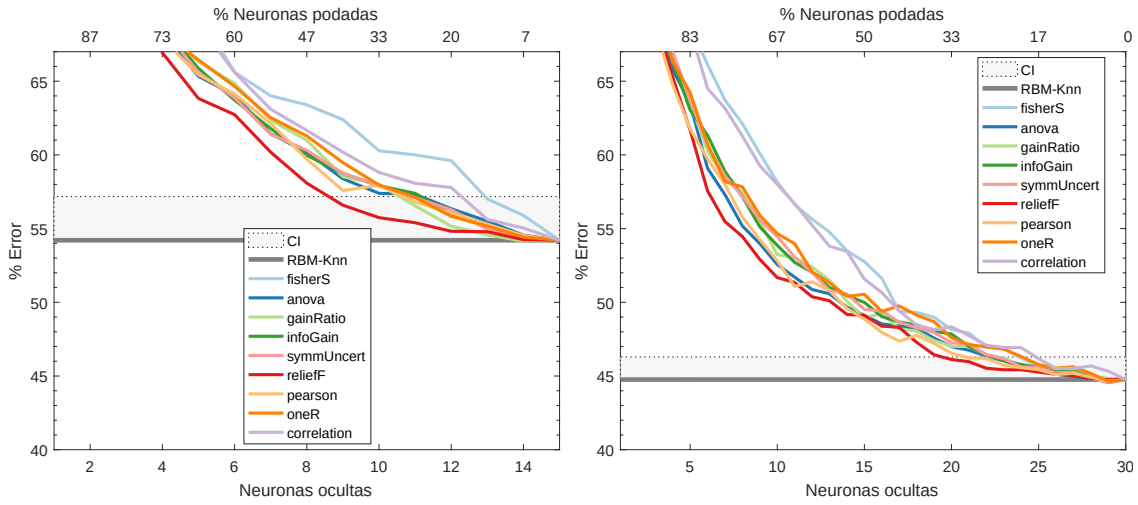
#	Nombre	Neuronas de entrada	Neuronas para poda
1	Interface	30	15, 30, 60, ..., 1920
2	Berlín	30	15, 30, 60, ..., 960
3	Sensores de gas	128	15, 32, 64, ..., 512

Resultados y discusiones

De forma similar a como se hizo en la subsección de resultados de poda binaria (6.1), aquí se presentan y discuten los resultados del proceso de poda multiclase de acuerdo con el Algoritmo 2. El primer resultado considera el error de clasificación en relación con el número de unidades ocultas que se utilizan en la etapa de clasificación luego de la poda, respecto de las distintas medidas discriminativas (Figuras 6-9, 6-10 y 6-11). El segundo, presenta un análisis que permite estimar qué porcentaje de neuronas se requiere para obtener un error de clasificación razonable (Figuras 6-12, 6-13 y 6-14 para dos de las medidas). Por último, se utiliza RDCG como índice cuantitativo para proporcionar más información sobre el proceso de elección del punto de poda (Figura 6-15). También, como en la sección anterior, se utiliza el *intervalo de confianza (CI)*, definido por la ecuación 6.1, y calculado con un nivel de confianza del 95 %. La linea base propuesta consiste en clasificadores RBM+Knn, utilizados en su forma estandar (sin poda), y los resultados están representados en las figuras por una línea continua, mientras que los intervalos de confianza están representados por una línea punteada.

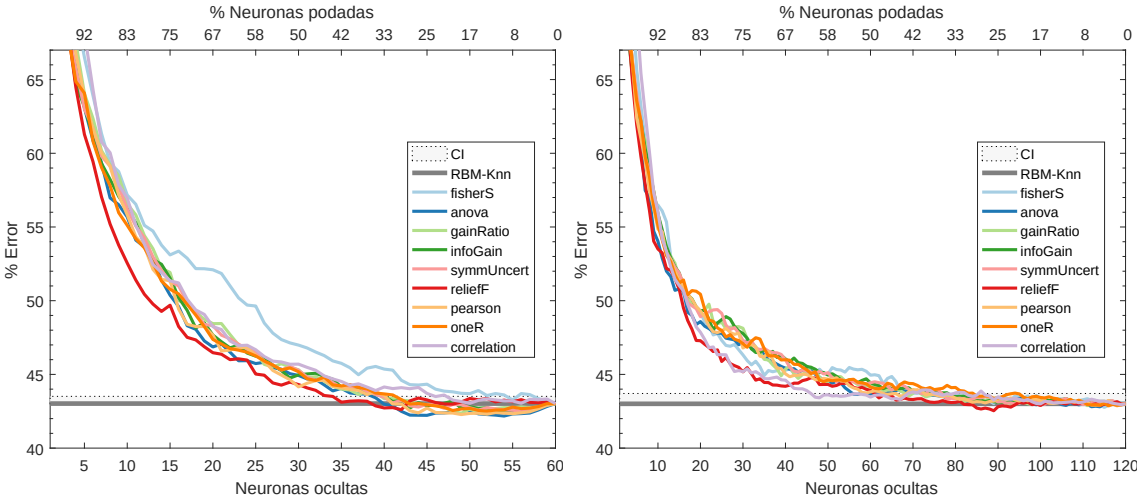
Los resultados para la base de datos *INTERFACE*, utilizando las siete clases, se presentan en la Figura 6-9. Dada la experiencia obtenida con los resultados binarios donde se hipotizó que el corpus tenía una alta complejidad, se agregaron los experimentos con 480, 960 y 1920 neuronas en la capa oculta (Figuras 6-9f, g, h). El comportamiento general de estos experimentos es similar a los observados con los otros conjuntos de datos: el error base es mayor cuando se usan inicialmente 15 uni-

dades ocultas y, al mismo tiempo, se requieren todas las neuronas para lograrlo, los resultados muestran una tendencia favorable donde se obtienen ahorros del 50 % de neuronas ocultas al mismo tiempo que se mejora la tasa de error base. De manera similar a la anterior, los experimentos que utilizan más unidades (Figuras 6-9b, c, d, e) se benefician más. Podemos ver en la Fig. 6-9c, que utiliza 60 neuronas iniciales, se obtienen por primera vez resultados que mejoran la linea base con ahorros de hasta el 33 %. En estas figuras existe un comportamiento que, aunque general, resulta interesante pues se puede apreciar que las curvas de algunas distancias se aproximan con diferente velocidad a la linea base aunque, algunas de ellas sin lograr mejorarla. Por este motivo se eligieron para su análisis posterior *Fisher Score* y *relief-F*, aunque su comportamiento no fuera el más rápido.

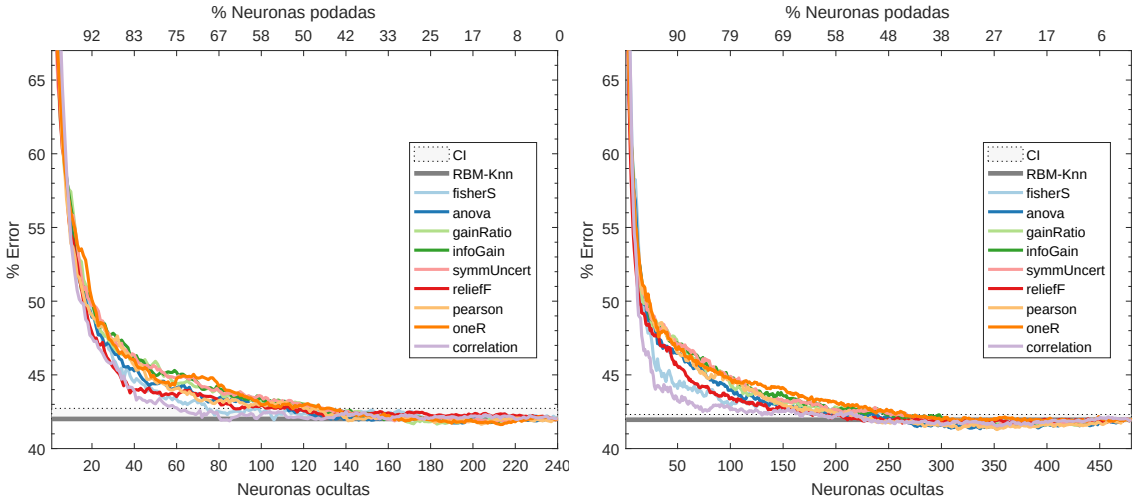


(a) *Poda de la RBM con 15 unidades ocultas.* **(b)** *Poda de la RBM con 30 unidades ocultas.*

Figura 6-9: *Resultados de la clasificación sobre el corpus INTERFACE. Las pruebas se realizaron utilizando 15, 30, 60, 120, 240, 480, 960 y 1920 neuronas iniciales.*

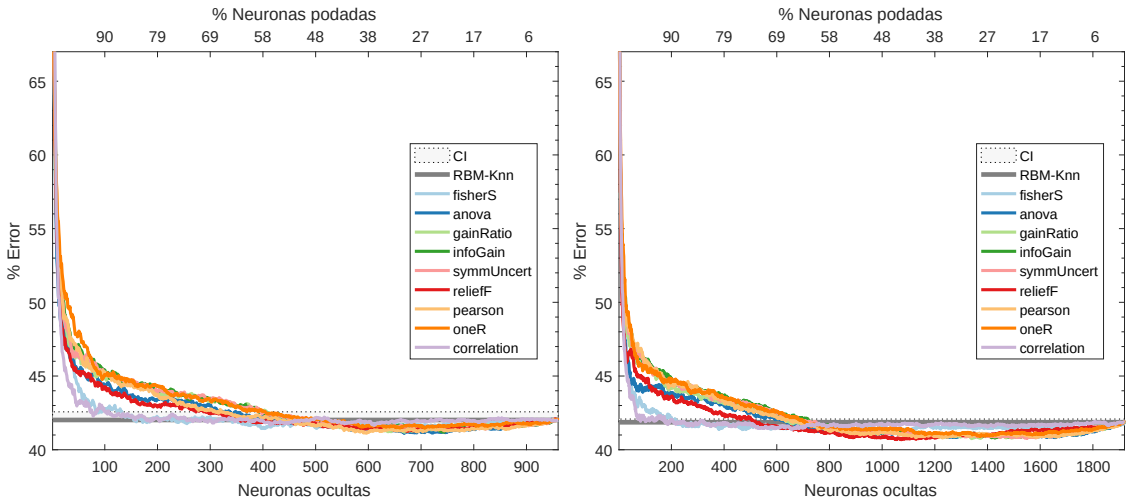


(c) *Poda de la RBM con 60 unidades ocultas.* **(d)** *Poda de la RBM con 120 unidades ocultas.*



(e) *Poda de la RBM con 240 unidades ocultas.* **(f)** *Poda de la RBM con 480 unidades ocultas.*

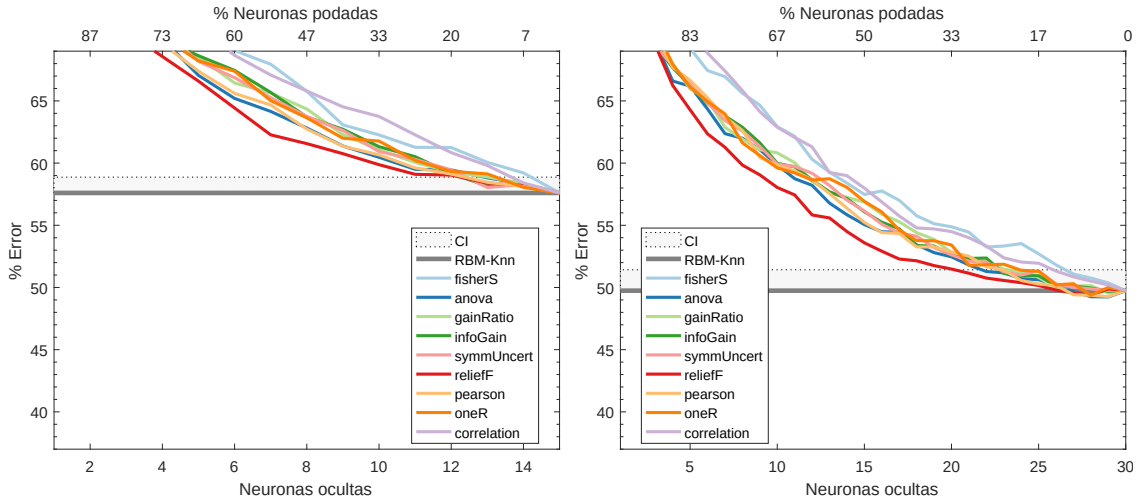
Figura 6-9: *(cont.) Resultados de la clasificación sobre el corpus INTERFACE. Las pruebas se realizaron utilizando 15, 30, 60, 120, 240, 480, 960 y 1920 neuronas iniciales.*



(g) *Poda de la RBM con 960 unidades ocultas.* **(h)** *Poda de la RBM con 1920 unidades ocultas.*

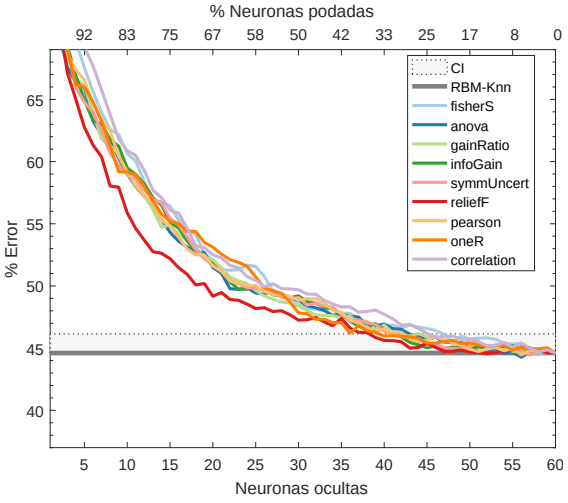
Figura 6-9: *(cont.) Resultados de la clasificación sobre el corpus INTERFACE. Las pruebas se realizaron utilizando 15, 30, 60, 120, 240, 480, 960 y 1920 neuronas iniciales.*

Para el corpus de habla *Berlín* se utilizaron redes con diferentes topologías que van desde 15 hasta 960 neuronas iniciales, los resultados de los experimentos se muestran en la Figura 6-10. Estos resultados se convalidan con los encontrados con las otras bases de datos. El primero de ellos, presentado en la Figura 6-10a, muestra que cuando la dimensión de los vectores de entrada es mayor que el número de unidades ocutltas de la RBM, ninguna de las redes podadas mejora el resultado de la línea base. El segundo experimento, que se presenta en la Figura 6-10b, muestra nuevamente que emplear 30 unidades ocultas es más beneficioso que 15, pues el error, con respecto a la línea base, se mejora aunque marginalmente. Esta tendencia continúa hasta la red con 120 unidades ocultas(Fig. 6-10d) donde los resultados a partir de esta red muestran ahorros considerables, sobre el 60 % de unidades ocultas podadas y hasta alcanzar el 95 % de poda para la red de 960 neuronas ocultas (Fig. 6-10g).

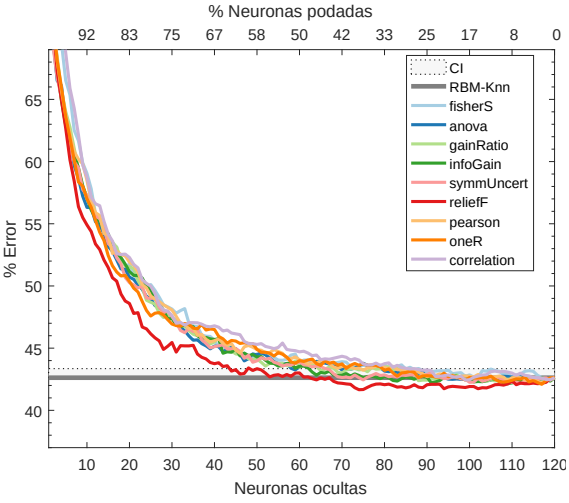


(a) *Poda de la RBM con 15 unidades ocultas.*

(b) *Poda de la RBM con 30 unidades ocultas.*

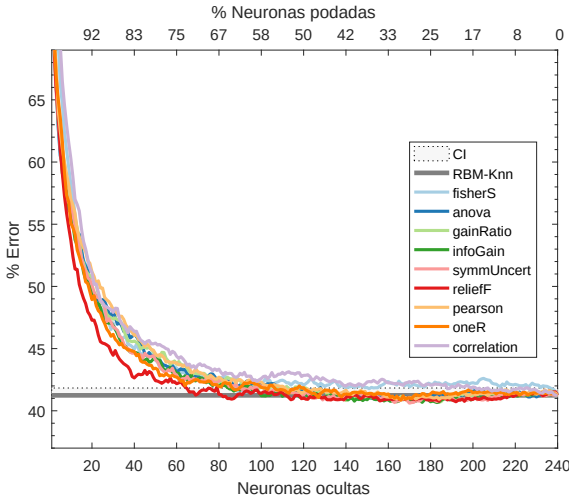


(c) *Poda de la RBM con 60 unidades ocultas.*

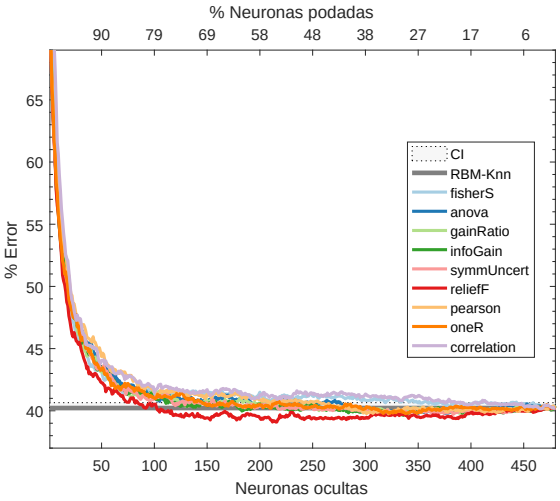


(d) *Poda de la RBM con 120 unidades ocultas.*

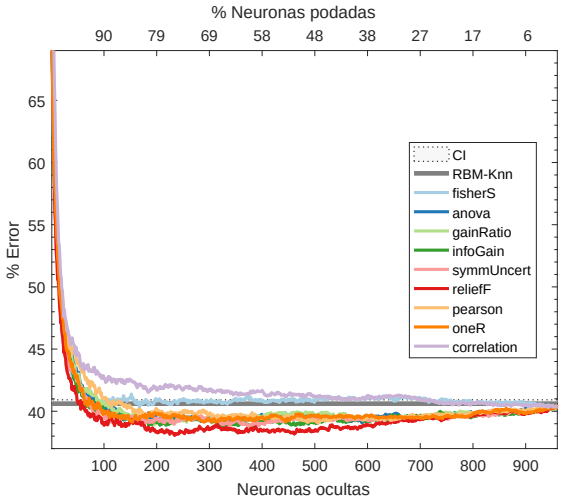
Figura 6-10: *Resultados de la clasificación sobre el corpus EMODB. Las pruebas se realizaron utilizando 15, 30, 60, 120, 240, 480 y 960 neuronas iniciales.*



(e) *Poda de la RBM con 240 unidades ocultas.*



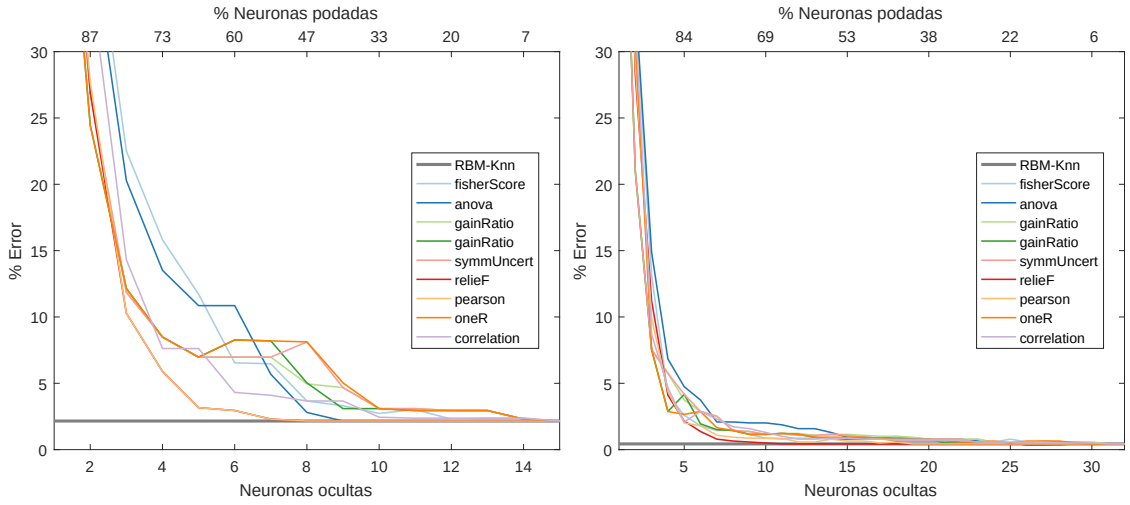
(f) *Poda de la RBM con 480 unidades ocultas.*



(g) *Poda de la RBM con 960 unidades ocultas.*

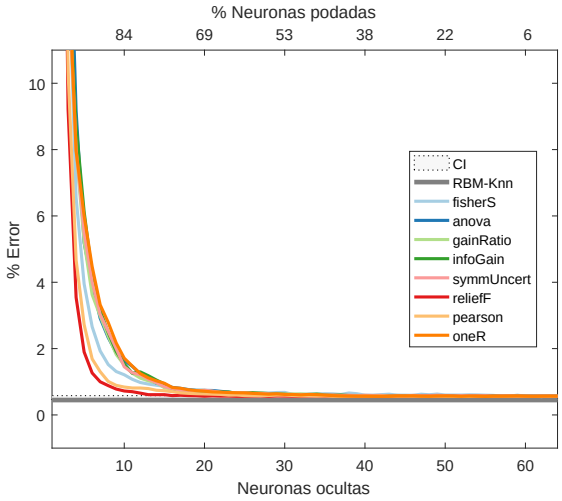
Figura 6-10: *(cont.) Resultados de la clasificación sobre el corpus EMODB. Las pruebas se realizaron utilizando 15, 30, 60, 120, 240, 480 y 960 neuronas iniciales.*

Para la base de datos *sensores de gas* se utilizaron configuraciones iniciales de 64, 128, 256 y 512 neuronas logrando una rápida convergencia y alcanzando niveles muy bajos de error. Como se puede apreciar en la Figura 6-11, las neuronas requeridas para alcanzar un porcentaje de error cercano a cero son muy pocas, en las Figuras 6-11a y b se puede apreciar que incluso con 5 neuronas se alcanza un porcentaje de aciertos del 98 %. Como se ha hecho notar en ocasiones anteriores, comparar los resultados de distintos métodos de clasificación es, desde muchas perspectivas, inviable. Sin embargo, en aras de presentar un panorama completo, podemos decir que 11 neuronas (Fig. 6-11b) son suficientes para superar la tasa de error reportada por los autores de la base de datos [164]. En este trabajo, la poda de 9 neuronas en la red de 32 unidades iniciales (11 neuronas útiles) son suficientes para alcanzar la linea de error base y producir una tasa de error del 0.43 % y con 26 neuronas, llegar al 0.36 % de error mejorando la linea base sin que las redes con más neuronas iniciales obtengan mejores resultados. Los resultados reportados por los autores del corpus, provienen del clasificador conocido como *máquina de soporte vectorial* [165] obteniendo, en el mejor de los casos, 5.08 % y en el peor, 82.98 % de error.

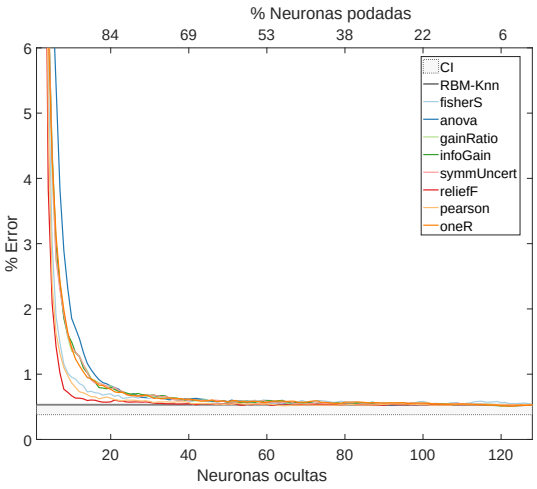


(a) *Poda de la RBM con 15 unidades ocultas.* **(b)** *Poda de la RBM con 32 unidades ocultas.*

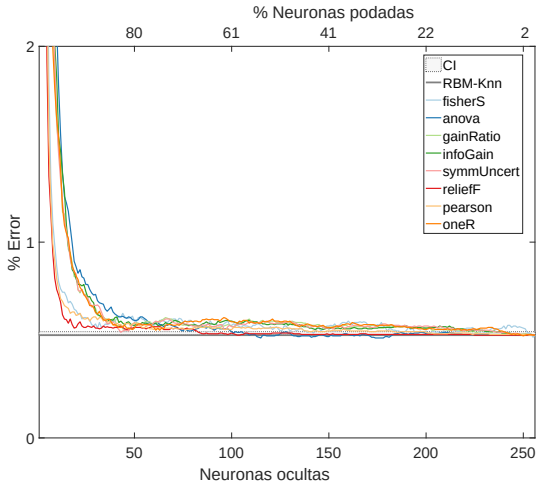
Figura 6-11: *Resultados de la clasificación sobre el corpus Gas Sensor. Las pruebas se realizaron utilizando 15, 32, 64, 128, 256 y 512 neuronas iniciales.*



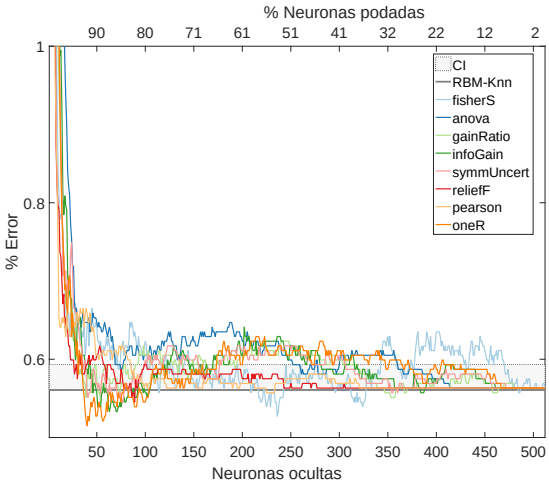
(c) *Poda de la RBM con 64 unidades ocultas.*



(d) *Poda de la RBM con 128 unidades ocultas.*



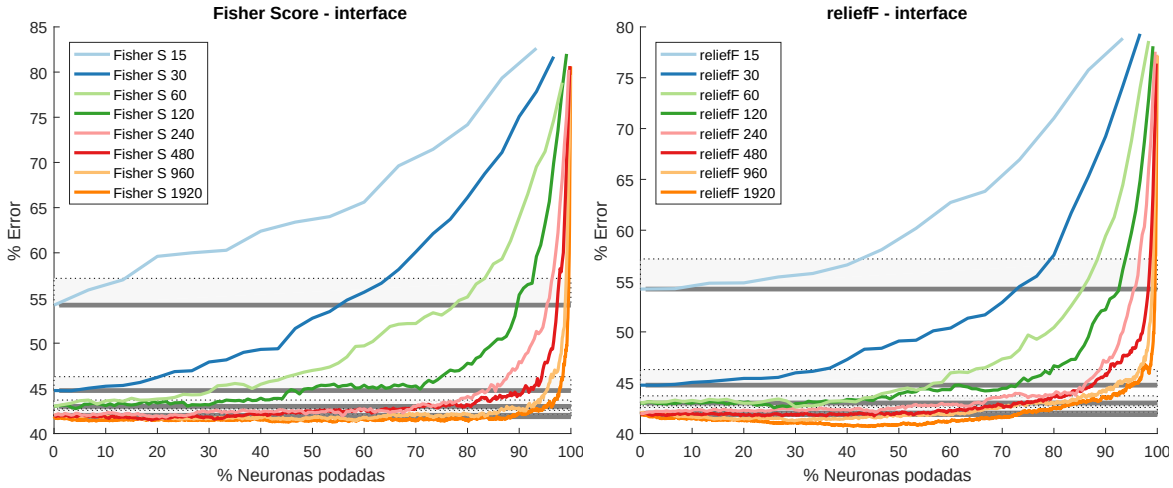
(e) *Poda de la RBM con 256 unidades ocultas.*



(f) *Poda de la RBM con 512 unidades ocultas.*

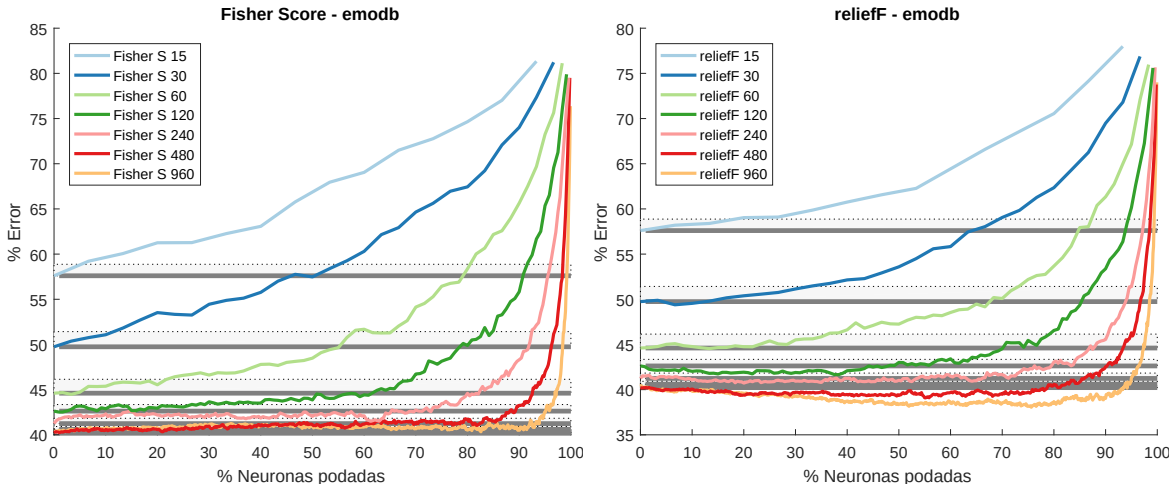
Figura 6-11: *(cont.) Resultados de la clasificación sobre el corpus Gas Sensor. Las pruebas se realizaron utilizando 15, 32, 64, 128, 256 y 512 neuronas iniciales.*

En los experimentos, y los resultados obtenidos de ellos, para cada uno de los tres corpus antes presentados, se puede observar una tendencia general para las distancias *Fisher Score* y *relief-F*, donde alrededor del 25 % de las unidades seleccionadas como más discriminativas son suficientes para alcanzar e incluso mejorar la línea de error base, dando así una posible reducción de al menos el 75 % de las unidades ocultas en la arquitectura inicial al emplear las distancias *Fisher Score* y *relief-F*. Debido a esto, en la Figura 6-12 se analizan particularmente estas distancias para el corpus INTERFACE, en ella se puede observar que el porcentaje de ahorro o poda descrito por el eje de las abscisas, se incrementa cuando la arquitectura inicial es mayor, de la misma forma que la tasa de error se mejora. Este análisis se repite en la Figura 6-13 para la base de datos de Berlín, en la subfigura 6-13b podemos apreciar que, aunque por poco, la distancia relief-F mejora la tasa de error y la poda (eje de ordenadas y abscisas respectivamente) de Fisher Score para la RBM con una arquitectura inicial de 960 neuronas. El caso de la base de datos de sensores de gas se muestra en la Figura 6-14 y es parecido a los otros experimentos aunque, como vimos antes, se obtuvo una convergencia rápida con pocas neuronas, particularmente la red con 32 unidades iniciales alcanza la misma tasa de error que las redes más grandes y se confirma, nuevamente, que utilizar menos neuronas que la dimensión del vector de entrada es peor.



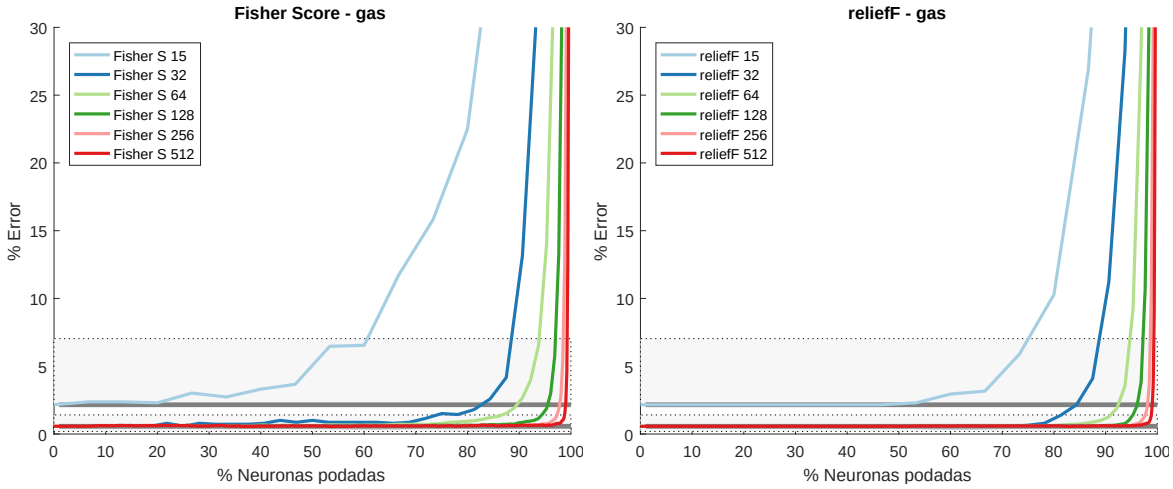
(a) *Fisher Score con 15, 30, ..., y 1920 unidades.* **(b)** *relief-F con 15, 30, ..., y 1920 unidades.*

Figura 6-12: *Curvas de error para las medidas Fisher Score y relief-F para el corpus INTERFACE. Las pruebas se realizaron utilizando 15, 30, 60, 120, 240, 480, 960 y 1920 neuronas iniciales.*



(a) *Fisher Score con 15, 30, ..., y 960 unidades.* **(b)** *relief-F con 15, 30, ..., y 960 unidades.*

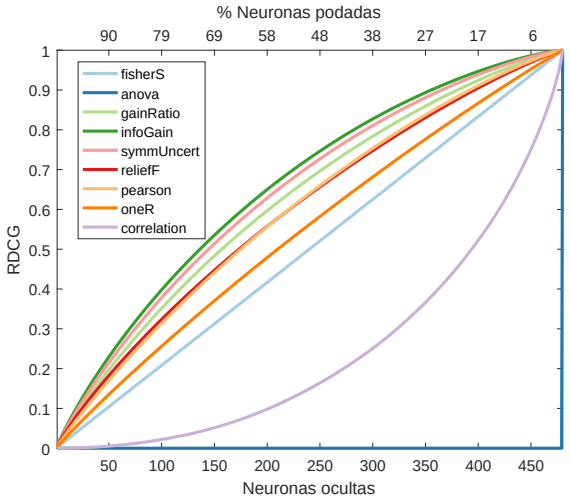
Figura 6-13: *Curvas de error para las medidas Fisher Score y relief-F para el corpus EMODB. Las pruebas se realizaron utilizando 15, 30, 60, 120, 240, 480 y 960 neuronas iniciales.*



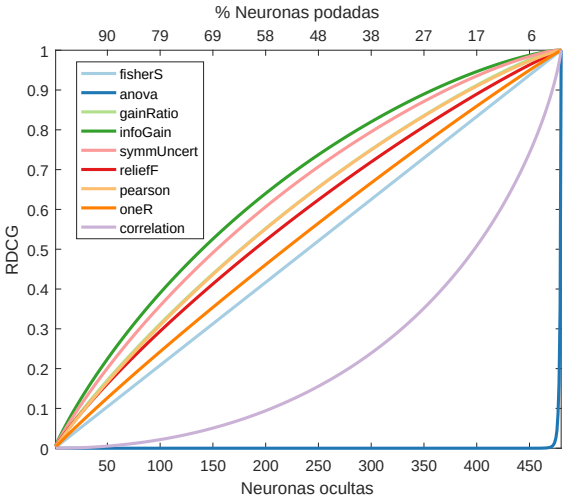
(a) *Fisher Score con 15, 32, ..., y 512 unidades.* **(b)** *relief-F con 15, 32, ..., y 512 unidades.*

Figura 6-14: *Curvas de error para las medidas Fisher Score y relief-F para el corpus Gas Sensor. Las pruebas se realizaron utilizando 15, 32, 64, 128, 256 y 512 neuronas iniciales.*

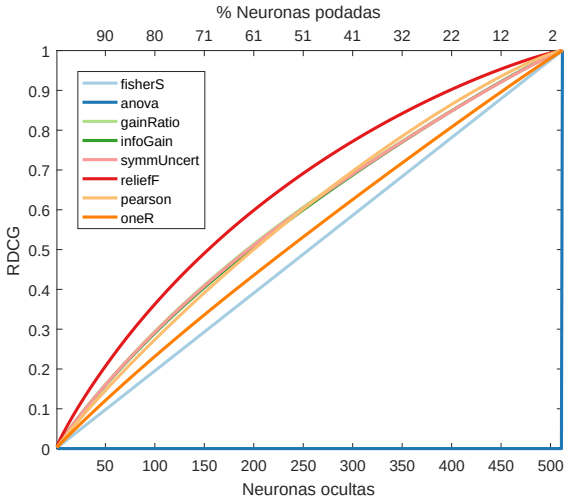
Como en los experimentos de poda binaria, el índice RDCG mostrado en la Figura 6-15, expone la RDCG de las medidas discriminativas, para facilitar el análisis de los resultados se presenta el índice para una RBM con 480 neuronas iniciales para las bases de datos INTERFACE y Berlín mientras que para el corpus de sensores de gas se utilizan 512. En el eje de las abscisas se observa el número de neuronas mientras que en el de las ordenadas, lo que podría interpretarse como el porcentaje de explicación de los datos. Así pues, un valor de RDCG de 0.8 y 150 neuronas querría decir que el 80 % de la información es contribuida por 150 neuronas. Esto proporciona información útil para determinar un posible punto de poda sin que sea, de ningún modo, el mejor. Esta idea es posteriormente analizada en las Tablas 6.14–6.16 donde tres valores de RDCG (0.7, 0.8 y 0.9) se presentan para las medidas *Fisher Score* y *relief-F*.



(a) *Comportamiento del índice RDCG para INTERFACE.*



(b) *Comportamiento del índice RDCG para EMODB.*



(c) *Comportamiento del índice RDCG para Gas Sensor.*

Figura 6-15: *Ganancia Acumulada Discriminativa Relativa (RDCG) de una RBM con 480 unidades iniciales para INTERFACE y EMODB, mientras que para Gas Sensor se utilizaron 512.*

Las Tablas 6.11 para INTERFACE, 6.12 para Berlín y 6.13 para la base de datos de sensores de gas, proporcionan más información sobre las dos medidas *Fisher Score* y *relief-F*. La información proporcionada en ellas es sobre el número más pequeño de unidades después de lo cual sólo se obtienen errores de clasificación aceptables sin que esto signifique que son las menores tasas de error obtenidas, en cualquier caso, se interpretan como las redes de menor tamaño que obtienen resultados comparables a los de la red sin podar. Con propósitos comparativos, también se muestra la mejor tasa de error obtenida con esa arquitectura particular. En los detalles de estas tablas se muestra el número de unidades iniciales en la red, el error base obtenido al utilizar todas las neuronas en la clasificación, el número de neuronas no podadas, el porcentaje de error después de la poda y el porcentaje de ahorro para las distintas bases de datos. Después de inspeccionar las Tablas, podemos observar que al usar más unidades iniciales en la red, las tasas de error base son, generalmente, mejores y utilizan menos neuronas.

Tabla 6.11: *Resultados de clasificación sobre la base de datos INTERFACE utilizando las mejores dos distancias.* ⋆ *Configuración inicial,* † *Neuronas no podadas.* [1]*Primer configuración que entra al intervalo de confianza,* [2]*Mejor rendimiento para poda.*

Fisher Score							
⋆	% Error base	†	% Error[1]	% Ahorro	†	% Error[2]	% Ahorro
15	**54.20**	13	57.01	13.33	15	**54.20**	0.00
30	44.76	24	46.22	20.00	29	**44.73**	3.33
60	**43.02**	55	43.26	8.33	60	**43.02**	0.00
120	43.00	73	43.67	39.17	115	**42.76**	4.16
240	42.01	75	42.59	68.75	188	**41.82**	21.66
480	41.96	183	42.31	61.88	390	**41.49**	18.75
960	42.02	84	42.48	91.25	397	**41.43**	58.64
1920	41.86	177	42.08	90.78	1053	**41.26**	45.15
relief-F							
15	**54.20**	9	56.59	40.00	15	**54.20**	0.00
30	44.76	20	46.12	33.33	29	**44.74**	3.33
60	43.02	34	43.51	43.33	42	**42.60**	30.00
120	43.00	61	43.68	49.17	89	**42.53**	25.83
240	42.01	87	42.72	63.75	239	**41.94**	0.41
480	41.96	178	42.25	62.92	263	**41.72**	45.20
960	42.02	273	42.56	71.56	673	**41.24**	29.89
1920	41.86	463	42.04	75.89	1073	**40.68**	44.11

Tabla 6.12: *Resultados de clasificación sobre la base de datos EMODB utilizando las mejores dos distancias. ⋆ Configuración inicial, † Neuronas no podadas. [1]Primer configuración que entra al intervalo de confianza, [2]Mejor rendimiento para poda.*

Fisher Score							
⋆	% Error base	†	% Error[1]	% Ahorro	†	% Error[2]	% Ahorro
15	**57.60**	15	**57.60**	0.00	15	**57.60**	0.00
30	**49.74**	27	51.08	10.00	30	**49.74**	0.00
60	44.61	47	46.12	21.67	58	**44.48**	3.33
120	42.62	85	43.33	29.17	118	**42.53**	1.67
240	**41.24**	88	41.76	63.33	240	**41.24**	0.00
480	**40.22**	319	40.64	33.54	480	**40.22**	0.00
960	40.61	68	40.81	92.92	167	**40.34**	82.60
relief-F							
15	**57.60**	13	58.39	13.33	15	**57.60**	0.00
30	49.74	21	51.14	30.00	28	**49.39**	6.67
60	44.61	39	45.85	35.00	52	**44.54**	13.33
120	42.62	44	43.15	63.33	74	**41.66**	38.33
240	41.24	65	41.78	72.92	192	**40.73**	20.00
480	40.22	82	40.61	82.92	212	**39.09**	55.83
960	40.61	49	40.85	94.90	234	**38.05**	75.63

Tabla 6.13: *Resultados de clasificación sobre la base de datos Gas Sensor utilizando las mejores dos distancias. ⋆ Configuración inicial, † Neuronas no podadas. [1]Primer configuración que entra al intervalo de confianza, [2]Mejor rendimiento para poda.*

Fisher Score							
⋆	% Error base	†	% Error[1]	% Ahorro	†	% Error[2]	% Ahorro
15	**2.16**	15	**2.16**	0.00	15	**2.16**	0.00
32	**0.43**	31	**0.43**	3.13	31	**0.43**	3.13
64	0.45	57	0.57	10.94	62	**0.45**	3.13
128	0.50	49	0.53	61.72	114	**0.49**	10.93
256	0.52	139	0.54	45.70	255	**0.51**	0.39
512	0.56	73	0.58	98.57	236	**0.53**	53.91
relief-F							
15	**2.16**	8	**2.16**	46.67	8	**2.16**	46.67
32	**0.43**	11	**0.43**	65.63	11	**0.43**	65.63
64	**0.45**	19	0.58	70.31	52	**0.45**	18.75
128	0.50	11	0.53	91.40	61	**0.48**	52.34
256	**0.52**	83	0.54	67.58	103	**0.52**	59.77
512	0.56	26	0.59	94.92	89	**0.54**	82.62

Es posible decir que para aprovechar de forma correcta la metodología propuesta en este trabajo, es necesario utilizar al menos la misma cantidad de neuronas ocultas que de neuronas visibles, ya que utilizar menos no beneficia los resultados de clasificación. El número de unidades podadas después de lo cual sólo se obtienen errores de clasificación aceptables, proporciona un punto de poda que considera sólo el número de neuronas no podadas, otra forma de estimar el punto de poda pero que tome en cuenta la tasa de error y el número de neuronas podadas, aunque de ninguna manera el *mejor*, es utilizar la ecuación de Ganancia Acumulada Discriminativa Relativa (RDCG) definida en 6.2 pues con ella se intenta proporcionar un equilibrio entre la tasa de error y el número de unidades podadas, en lugar de centrarse únicamente en uno de ellos. Este índice es análogo a una ecuación similar utilizada en el Análisis de Componentes Principales [161] empleada para reducción de dimensionalidad. Los detalles de este índice para las distancias Fisher score y relief-F con la base de datos INTERFACE se puede encontrar en la Tabla 6.14, en la Tabla 6.15 para Berlín y en la Tabla 6.16 para la base de datos de sensores de gas. Aquí, 0.8 parece ser un compromiso razonable entre el número de neuronas y el porcentaje de explicación pues se obtienen valores cercanos a la linea de error base con menos neuronas.

Tabla 6.14: *Resultados de clasificación sobre la base de datos INTERFACE utilizando las mejores dos distancias y RDCG como estimador del punto de poda. ⋆ Configuración inicial, † Neuronas no podadas.*

Fisher Score										
		RDCG 0.7			RDCG 0.8			RDCG 0.9		
⋆	% Error base	†	% Error	% Ahorro	†	% Error	% Ahorro	†	% Error	% Ahorro
15	54.21	11	60.00	26.67	12	**59.61**	20.00	14	55.89	6.67
30	44.76	21	47.92	30.00	24	**46.22**	20.00	27	45.23	10.00
60	43.02	42	44.60	30.00	48	**43.82**	20.00	54	43.61	10.00
120	43.00	84	43.18	30.00	96	**43.38**	20.00	108	43.09	10.00
240	42.01	168	42.39	30.00	192	**41.96**	20.00	216	42.21	10.00
480	41.96	336	41.82	30.00	384	**41.67**	20.00	432	41.78	10.00
960	42.02	672	41.69	30.00	768	**41.70**	20.00	864	41.90	10.00
1920	41.86	1344	41.47	30.00	1536	**41.52**	20.00	1728	41.47	10.00
relief-F										
		RDCG 0.7			RDCG 0.8			RDCG 0.9		
⋆	% Error base	†	% Error	% Ahorro	†	% Error	% Ahorro	†	% Error	% Ahorro
15	54.21	9	56.59	40.00	10	**55.74**	33.33	13	54.79	13.33
30	44.76	17	48.29	43.33	21	**45.97**	30.00	25	45.27	16.67
60	43.02	34	43.51	43.33	41	**42.75**	31.67	49	43.13	18.33
120	43.00	69	43.31	42.50	83	**42.95**	30.83	100	42.89	16.67
240	42.01	136	42.25	43.33	165	**42.33**	31.25	198	42.35	17.50
480	41.96	274	41.95	42.92	331	**41.94**	31.04	397	41.97	17.29
960	42.02	544	41.61	43.33	660	**41.39**	31.25	792	41.50	17.50
1920	41.86	1092	40.76	43.13	1323	**41.09**	31.09	1586	41.38	17.40

Tabla 6.15: *Resultados de clasificación sobre la base de datos EMODB utilizando las mejores dos distancias y RDCG como estimador del punto de poda. ⋆ Configuración inicial, † Neuronas no podadas.*

Fisher Score										
		RDCG 0.7			RDCG 0.8			RDCG 0.9		
⋆	% Error base	†	% Error	% Ahorro	†	% Error	% Ahorro	†	% Error	% Ahorro
15	57.60	11	61.27	26.67	12	**61.25**	20.00	14	59.21	6.67
30	49.74	21	54.45	30.00	24	**53.53**	20.00	27	51.08	10.00
60	44.61	42	46.61	30.00	48	**45.54**	20.00	54	45.38	10.00
120	42.62	84	43.66	30.00	96	**42.95**	20.00	108	42.88	10.00
240	41.24	168	42.01	30.00	192	**42.19**	20.00	216	42.32	10.00
480	40.22	336	40.96	30.00	384	**40.56**	20.00	432	40.59	10.00
960	40.38	672	41.13	30.00	768	**40.77**	20.00	864	40.74	10.00
relief-F										
		RDCG 0.7			RDCG 0.8			RDCG 0.9		
⋆	% Error base	†	% Error	% Ahorro	†	% Error	% Ahorro	†	% Error	% Ahorro
15	57.60	9	60.76	40.00	11	**59.10**	26.67	13	58.39	13.33
30	49.74	18	52.14	40.00	21	**51.14**	30.00	25	50.16	16.67
60	44.61	36	46.62	40.00	43	**44.99**	28.33	51	44.66	15.00
120	42.62	73	41.82	39.17	87	**42.06**	27.50	102	41.74	15.00
240	41.24	145	41.29	39.58	173	**40.89**	27.92	204	40.98	15.00
480	40.22	290	39.44	39.58	346	**39.68**	27.92	407	39.79	15.21
960	40.38	579	38.64	39.69	690	**39.22**	28.13	813	39.71	15.31

Tabla 6.16: *Resultados de clasificación sobre la base de datos Gas Sensor utilizando las mejores dos distancias y RDCG como estimador del punto de poda. ⋆ Configuración inicial, † Neuronas no podadas.*

Fisher Score										
		RDCG 0.7			RDCG 0.8			RDCG 0.9		
⋆	% Error base	†	% Error	% Ahorro	†	% Error	% Ahorro	†	% Error	% Ahorro
15	2.16	11	3.02	26.67	12	**2.30**	20	14	2.37	6.67
32	0.43	23	0.79	28.13	26	**0.58**	18.75	29	0.58	9.38
64	0.56	45	0.6	29.69	52	**0.60**	18.75	58	0.56	9.38
128	0.53	90	0.57	29.69	103	**0.55**	19.53	116	0.57	9.38
256	0.52	180	0.55	29.69	205	**0.54**	19.92	231	0.54	9.77
512	0.57	359	0.59	29.88	410	**0.62**	19.92	461	0.61	9.96
relief-F										
		RDCG 0.7			RDCG 0.8			RDCG 0.9		
⋆	% Error base	†	% Error	% Ahorro	†	% Error	% Ahorro	†	% Error	% Ahorro
15	2.16	7	2.3	53.33	8	**2.16**	46.67	11	2.16	26.67
32	0.43	16	0.43	50	19	**0.43**	40.63	24	0.43	25
64	0.56	29	0.57	54.69	36	**0.56**	43.75	46	0.56	28.13
128	0.53	57	0.54	55.47	72	**0.54**	43.75	92	0.53	28.13
256	0.52	119	0.53	53.52	151	**0.53**	41.02	192	0.53	25
512	0.57	256	0.57	50	320	**0.57**	37.5	399	0.57	22.07

CAPÍTULO 7

SUMARIO Y FUTURAS LÍNEAS DE INVESTIGACIÓN

Gracias a años de investigación, las tecnologías del habla se han convertido en parte de nuestra vida diaria. Por un lado, el rápido aumento en la cantidad de contenido multimedia ha creado la necesidad de concebir nuevas tecnologías que procesen este tipo de información de forma automática, por otro lado, estos avances tecnológicos han permitido la inclusión de diversas herramientas en el día a día. Así, las tecnologías del habla son de gran importancia en el estudio de problemas relacionados con la salud mental, con problemas en la producción de habla y, en general, en las áreas que se relacionan con el comportamiento humano. Por tal motivo, en esta Tesis se realizó un análisis extenso de las redes profundas (DBN) que se utilizaron con diferentes configuraciones para la tarea de reconocer emociones, para luego proponer un método novedoso que permite mejorar su rendimiento mediante la poda neuronal con base en medidas discriminativas aplicadas sobre la red. Para evaluar el modelo propuesto se utilizaron distintas configuraciones experimentales que permitieron analizar la evolución de las redes neuronales probándolas en el área del reconocimiento de emociones en la voz con resultados muy satisfactorios.

Con el objetivo de explorar esta problemática se investigaron tres cuestiones principales en el contexto de la poda neuronal y clasificación: (1) las emociones y el efecto que tienen en la producción del habla así como las características de su señal (2)

el aprendizaje profundo, en particular de las máquinas restringidas de Boltzmann (RBM) y (3) la poda discriminativa en problemas binarios y multi-clase.

En esta contribución se presentó la identificación de una arquitectura de red neuronal adecuada a través de la evaluación de la información discriminativa de cada neurona. Se investigó si las medidas discriminativas podrían proporcionar una pista para encontrar las neuronas más discriminativas en una máquina restringida de Boltzmann, con el fin de podar aquellas que contribuyan menos. En el documento se compararon nueve medidas discriminativas en el enfoque multi-clase y cinco en el binario. Los resultados indican que, en general, más del 50 % de las neuronas se pueden podar manteniendo una tasa de error aceptable. Además, se confirma que comenzar con una arquitectura de red más grande y luego podar es más ventajoso que usar una red más pequeña para empezar. Finalmente, se introdujo un índice cuantitativo que puede proporcionar información sobre cómo elegir un punto de poda adecuado.

Los resultados obtenidos en este trabajo confirman los encontrados previamente en [11], ahí se hipotizó que hacer uso de arquitecturas profundas sería beneficioso para la clasificación de emociones en la voz. Sin embargo, dichos resultados, mostraron preliminarmente que utilizar más de tres capas resulta poco provechoso y, en la mayoría de los casos, contraproducente. En este trabajo se ha encontrado el mismo comportamiento, utilizar más de tres capas no supone una mejora, incluso después de la poda neuronal, incrementar la profundidad no resulta en la disminución de la tasa de error. Por otro lado, se confirma que comenzar con una arquitectura de red más grande y luego podar es más ventajoso que usar una red pequeña para empezar.

Esta simplificación de la arquitectura de la máquina restringida de Boltzmann y redes de creencia profunda, a través del análisis discriminativo de las neuronas, se logró mediante un enfoque que descarta las neuronas menos discriminativas para la conformación del vector que alimenta el clasificador final. Existen otras técnicas ampliamente utilizadas para podar las redes neuronales artificiales, generalmente MLP, pero precisan el reentrenamiento de la red después de cada poda de pesos. Por otro lado, la propuesta descrita en este trabajo es más cercano a las técnicas de selección de

características pues descarta neuronas en lugar de pesos individuales, para lograrlo, las unidades se clasifican según una medida determinada (en el documento se comparan nueve medidas diferentes en el enfoque multi-clase y 5 en el binario) y no son consideradas las unidades menos discriminativas en la etapa posterior de clasificación final. Una ventaja importante es que esta propuesta no requiere el reentrenamiento de la red después de la poda.

La adopción del enfoque de clasificación en la metodología de poda presentada en este trabajo es muy prometedora pues los resultados indican que una vez que se ha elegido un número adecuado de neuronas iniciales, las redes podadas producen resultados mejores que los valores iniciales utilizando menos del 50 % de las neuronas y, en algunos casos, produciendo ahorros de hasta el 70 %. Los resultados también muestran que las mejores medidas para problemas binarios, en términos de tasa de error contra el número de neuronas, son la prueba t de Welch y DCAF, mientras que las mejores dos distancias para el problema multi-clase son Fisher score y relief-F. Esto es interesante dado que en investigaciones previas sobre la poda de redes neuronales artificiales, la medida o distancia preferida ha sido información mutua. Esta redefinición del esquema de clasificación tradicional expone evidencia que sugiere que la metodología propuesta podría ser empleada en diferentes sistemas de reconocimiento de patrones, y no sólo aquellos referentes al reconocimiento de emociones y máquinas restringidas de Boltzmann.

Como producto de esta investigación se encontró que podar el mayor número de neuronas no proporciona necesariamente la menor tasa de error, esto implicaría que tratar de encontrar una arquitectura "perfecta" es menos efectivo que entrenar una red grande para después podarla. Por tal motivo se introdujo en el documento, el índice RDCG como una forma alternativa de encontrar un punto de poda adecuado, los resultados sugieren que utilizar un valor de RDCG de 0.8 permite obtener una tasa de error aceptable al mismo tiempo que proporciona un ahorro significativo para las arquitecturas de las RBMs.

7.1. Trabajos futuros

Este trabajo abordó una serie de problemas existentes en el campo del reconocimiento de emociones en el habla y la poda neuronal, en el proceso se identificaron ciertos aspectos que podrían beneficiarse de los resultados hallados:

- Utilizar este método de poda en contextos multi-lenguajes, es decir, evaluar su utilidad para modelar emociones que han sido elicitadas en diferentes idiomas y elegir el mejor clasificador común que clasifique las emociones en todos los lenguajes utilizados.

- Aplicar la metodología de poda neuronal como mecanismo de selección de características al introducirse datos crudos a la red para posteriormente, reducir la dimensión y extraer las características más significativas de ellos. Estos datos crudos podrían provenir de la señal del habla, de señales biomédicas o de otras fuentes.

- Utilizar el enfoque de poda en otras redes neuronales de gran tamaño, como las Extreme Learning Machines (ELM) pues su arquitectura y buen desempeño al hallar fronteras de decisión complejas, justifican su uso.

- Explorar el aspecto generativo de estas redes para obtener datos pertenecientes a las distribuciones de las diferentes clases para, posteriormente, introducir esa información a alguna herramienta de síntesis del habla que permitiera generar entonaciones de habla emocional [166].

7.2. Trabajos publicados

Este trabajo ha sido un esfuerzo continuado durante varios años, en ese tiempo se han publicado los siguientes resultados:

- Reyes-Vargas M., Sánchez-Gutiérrez M., Rufiner L., Albornoz M., Vignolo L., Martínez-Licona F., Goddard-Close J., "Hierarchical Clustering and Classifi-

cation of Emotions in Human Speech Using Confusion Matrices", Speech and Computer, 162-169, 2013. [167]

- Albornoz E.M., Sánchez-Gutiérrez M., Martinez-Licona F., Rufiner H.L., Goddard J., "Spoken Emotion Recognition Using Deep Learning", Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, 104-111, 2014. [151]
- Sánchez-Gutiérrez M., Albornoz E.M., Martinez-Licona F., Rufiner H.L., Goddard J., "Deep Learning for Emotional Speech Recognition", Pattern Recognition, 311-320, 2014. [146]
- Sánchez-Gutiérrez M., Goddard J., Albornoz E.M., Rufiner H.L., Martinez-Licona F., "Redes de Creencia Profunda y Emociones", Komputer Sapiens, 12-18, 2017. [168]
- Sánchez-Gutiérrez M., Albornoz E.M., Rufiner H.L., Goddard J., "Post-training discriminative pruning for RBMs", Soft Computing, 1-15, 2017. [160]

BIBLIOGRAFÍA

[1] M. Schubiger, *English Intonation. Its Form and Function.* Niemeyer Verlag, 1958.

[2] J. O'Connor and G. F. Arnold, *Intonation of Colloquial English.* Prentice Hall Press, 1973.

[3] M. Benzeghiba, R. D. Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10, pp. 763 – 786, 2007. Intrinsic Speech Variations.

[4] A. Mehrabian, "Communication without words," *Psychology Today*, vol. 2, pp. 53–56, 1968.

[5] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The interspeech 2015 computational paralinguistics challenge: nativeness, parkinson's & eating condition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Interspeech 2016*, pp. 2001–2005, 2016.

[7] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, *et al.*, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, pp. 3442–3446, 2017.

[8] C. Magerkurth, A. D. Cheok, R. L. Mandryk, and T. Nilsen, "Pervasive games: bringing computer entertainment back to the real world," *Computers in Entertainment (CIE)*, vol. 3, p. 4A, 2005.

[9] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[10] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[11] M. E. S. Gutiérrez, *Reconocimiento de Emociones Utilizando Técnicas de Aprendizaje Maquinal*. Universidad Autónoma Metropolitana (Thesis), 2013.

[12] E. Schmidt and Y. Kim, "Learning emotion-based acoustic features with deep belief networks," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 65–68, 2011.

[13] X. L. Guo, H. Y. Wang, and D. H. Glass, "A growing bayesian self-organizing map for data clustering," in *2012 International Conference on Machine Learning and Cybernetics*, vol. 2, pp. 708–713, July 2012.

[14] K. O. Stanley and R. Miikkulainen, "Evolving neural networks through augmenting topologies," *Evolutionary computation*, vol. 10, no. 2, pp. 99–127, 2002.

[15] G. Castellano, A. M. Fanelli, and M. Pelillo, "An iterative pruning algorithm for feedforward neural networks," *IEEE Transactions on Neural Networks*, vol. 8, no. 3, pp. 519–531, 1997.

[16] K. Suzuki, I. Horiba, and N. Sugie, "A simple neural network pruning algorithm with application to filter synthesis," *Neural Processing Letters*, vol. 13, no. 1, pp. 43–53, 2001.

[17] S. Hussain and A. A. Alili, "A pruning approach to optimize synaptic connections and select relevant input parameters for neural network modelling of solar radiation," *Applied Soft Computing*, 2016.

[18] B. Hassibi, D. G. Stork, and G. J. Wolff, "Optimal brain surgeon and general network pruning," in *Neural Networks, 1993., IEEE International Conference on*, pp. 293–299, IEEE, 1993.

[19] B. Hassibi, D. G. Stork, G. Wolff, and T. Watanabe, "Optimal brain surgeon: Extensions and performance comparisons," *Advances in neural information processing systems*, pp. 263–263, 1994.

[20] Y. LeCun, J. S. Denker, S. A. Solla, R. E. Howard, and L. D. Jackel, "Optimal brain damage.," in *NIPs*, vol. 2, pp. 598–605, 1990.

[21] I. Sutskever and G. E. Hinton, "Learning multilevel distributed representations for high-dimensional sequences.," in *AISTATS*, vol. 2, pp. 548–555, 2007.

[22] F. J. Huang, Y.-L. Boureau, Y. LeCun, *et al.*, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *2007 IEEE conference on computer vision and pattern recognition*, pp. 1–8, IEEE, 2007.

[23] F. Cao, B. Liu, and D. S. Park, "Image classification based on effective extreme learning machine," *Neurocomputing*, vol. 102, pp. 90 – 97, 2013. Advances in Extreme Learning Machines (ELM 2011).

[24] B. Lu, G. Wang, Y. Yuan, and D. Han, "Semantic concept detection for video based on extreme learning machine," *Neurocomputing*, vol. 102, pp. 176 – 183, 2013. Advances in Extreme Learning Machines (ELM 2011).

[25] K.-L. Du and M. Swamy, *Neural Networks and Statistical Learning*. Springer London, 2014.

[26] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, (New York, NY, USA), pp. 609–616, ACM, 2009.

[27] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[28] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[29] G.-B. Huang, D. H. Wang, and Y. Lan, "Extreme learning machines: a survey," *International Journal of Machine Learning and Cybernetics*, vol. 2, no. 2, pp. 107–122, 2011.

[30] A. Hassan, *On Automatic Emotion Classification Using Acoustic Features*. University of Southampton: Electronics and Computer Science, 2012.

[31] K. N. Stevens, *Acoustic phonetics*, vol. 30. MIT press, 2000.

[32] K. Johnson, *Acoustic and Auditory Phonetics*. Wiley-Blackwell, 2011.

[33] J. Iwarsson, M. Thomasson, and J. Sundberg, "Effects of lung volume on the glottal voice source," *Journal of voice*, vol. 12, no. 4, pp. 424–433, 1998.

[34] J. D. Hoit and T. J. Hixon, "Age and speech breathing," *Journal of Speech, Language, and Hearing Research*, vol. 30, no. 3, pp. 351–366, 1987.

[35] P. Ladefoged, "Linguistic aspects of respiratory phenomena," *Annals of the New York Academy of Sciences*, vol. 155, no. 1, pp. 141–151, 1968.

[36] R. Kent, J. Kent, J. Rosenbek, H. Vorperian, and G. Weismer, "A speaking task analysis of the dysarthria in cerebellar disease," *Folia Phoniatrica et Logopaedica*, vol. 49, no. 2, pp. 63–82, 1997.

[37] P. Lieberman and S. E. Blumstein, *Speech physiology, speech perception, and acoustic phonetics*. Cambridge University Press, 1988.

[38] L. J. Raphael, G. J. Borden, and K. S. Harris, *Speech science primer: Physiology, acoustics, and perception of speech*. Lippincott Williams & Wilkins, 2007.

[39] J. Deller, P. J.R., H. J.G., and J.H., "Discrete time processing of speech signals," *Prentice Hall PTR*, 1993.

[40] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*. Pearson Education, 2011.

[41] G. Fant, "Glottal source and excitation analysis," *STL-QPSR*, vol. 1, no. 1979, pp. 85–107, 1979.

[42] G. Fant, "Vocal source analysis-a progress report," *STL-QPSR*, vol. 20, no. 3-4, pp. 31–53, 1979.

[43] G. Fant, "Some problems in voice source analysis," *Speech Communication*, vol. 13, no. 1-2, pp. 7–22, 1993.

[44] M. Rothenberg, "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing," *The Journal of the Acoustical Society of America*, vol. 53, no. 6, pp. 1632–1645, 1973.

[45] F. J. S. Campos, *Modelos Ocultos de Markov: Del Reconocimiento de Voz a la Musica, 978-1847536778*. LuluPress, 2009.

[46] J. Krajewski and B. Kröger, "Using prosodic and spectral characteristics for sleepiness detection," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[47] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Combining frame and turn-level information for robust recognition of emotions within speech," *Interspeech*, vol. 8, pp. 2249–2252, 2007.

[48] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," in *Ninth International Conference on Spoken Language Processing*, 2006.

[49] A. A. Razak, R. Komiya, M. Izani, and Z. Abidin, "Comparison between fuzzy and nn method for speech emotion recognition," in *Information Technology and Applications, 2005. ICITA 2005. Third International Conference on*, vol. 1, pp. 297–302, IEEE, 2005.

[50] K. R. Scherer, "Vocal affect expression: A review and a model for future research.," *Psychological bulletin*, vol. 99, no. 2, p. 143, 1986.

[51] R. R. Cornelius, *The science of emotion: Research and tradition in the psychology of emotions*. Prentice-Hall, Inc, 1996.

[52] N. H. Frijda, *The emotions*. Cambridge University Press, 1986.

[53] C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[54] W. B. Cannon, "The james-lange theory of emotions: A critical examination and an alternative theory," *The American journal of psychology*, vol. 39, no. 1/4, pp. 106–124, 1927.

[55] S. Schachter and J. Singer, "Cognitive, social, and physiological determinants of emotional state.," *Psychological review*, vol. 69, no. 5, p. 379, 1962.

[56] J. T. Cacioppo, G. G. Berntson, J. T. Larsen, K. M. Poehlmann, T. A. Ito, *et al.*, "The psychophysiology of emotion," *Handbook of emotions*, vol. 2, pp. 173–191, 2000.

[57] P. Ekman, R. W. Levenson, and W. V. Friesen, "Autonomic nervous system activity distinguishes among emotions," *Science*, vol. 221, no. 4616, pp. 1208–1210, 1983.

[58] P. E. Ekman and R. J. Davidson, *The nature of emotion: Fundamental questions*. Oxford University Press, 1994.

[59] A. Graesser and G. Mandler, "Recognition memory for the meaning and surface structure of sentences.," *Journal of Experimental Psychology: Human Learning and Memory*, vol. 1, no. 3, p. 238, 1975.

[60] K. A. MacDowell and G. Mandler, "Constructions of emotion: Discrepancy, arousal, and mood," *Motivation and Emotion*, vol. 13, no. 2, pp. 105–124, 1989.

[61] S. Tomkins, *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company, 1962.

[62] C. E. Izard, *Human emotions*. Springer Science & Business Media, 2013.

[63] P. Ekman, "Universals and cultural differences in facial expressions of emotion," *Nebraska Symposium on Motivation*, vol. 19, pp. 207–283, 1971.

[64] R. W. Levenson, "Autonomic nervous system differences among emotions," 1992.

[65] E. Paul, V. Wallace, and C. Joseph, "Facial action coding system: The manual on cd rom. a human face," 2002.

[66] F. Boiten, "Autonomic response patterns during voluntary facial action," *Psychophysiology*, vol. 33, no. 2, pp. 123–131, 1996.

[67] M. B. Arnold, "Emotion and personality.," 1960.

[68] J. R. Jennings, J. R. Averill, E. M. Opton, and R. S. Lazarus, "Some parameters of heart rate change: Perceptual versus motor task requirements, noxiousness, and uncertainty," *Psychophysiology*, vol. 7, no. 2, pp. 194–212, 1970.

[69] I. J. Roseman, "Cognitive determinants of emotion: A structural theory.," *Review of personality & social psychology*, 1984.

[70] K. R. Scherer, A. Schorr, and T. Johnstone, *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press, 2001.

[71] T. L. Gehm and K. R. Scherer, "Factors determining the dimensions of subjective emotional space.," 1988.

[72] T. Wehrle, S. Kaiser, S. Schmidt, and K. R. Scherer, "Studying the dynamics of emotional expression using synthesized facial muscle movements.," *Journal of personality and social psychology*, vol. 78, no. 1, p. 105, 2000.

[73] R. Plutchik, "The nature of emotions," *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[74] S. S. Tomkins, "Affect theory," *Approaches to emotion*, vol. 163, no. 163–195, 1984.

[75] H. Schlosberg, "The description of facial expressions in terms of two dimensions.," *Journal of experimental psychology*, vol. 44, no. 4, p. 229, 1952.

[76] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology*, vol. 17, no. 3, pp. 715–734, 2005.

[77] A. Batliner, S. Steidl, C. Hacker, and E. Nöth, "Private emotions versus social interaction: a data-driven approach towards analysing emotion in speech," *User Modeling and User-Adapted Interaction*, vol. 18, no. 1-2, pp. 175–206, 2008.

[78] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant.," *Journal of personality and social psychology*, vol. 76, no. 5, p. 805, 1999.

[79] S. Marc and C. R. et.al, "Feeltrace: An instrument for recording perceived emotion in real time," *ISCA Workshop on Speech and Emotion*, pp. 19–24, 2000.

[80] F. J. Tolkmitt and K. R. Scherer, "Effect of experimentally induced stress on vocal parameters.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 12, no. 3, p. 302, 1986.

[81] H. Giles, N. Coupland, and I. Coupland, "Accommodation theory: Communication, context, and," *Contexts of accommodation: Developments in applied sociolinguistics*, vol. 1, 1991.

[82] M. Willemyns, C. Gallois, V. J. Callan, and J. Pittam, "Accent accommodation in the job interview: Impact of interviewer accent and gender," *Journal of Language and Social Psychology*, vol. 16, no. 1, pp. 3–22, 1997.

[83] W. Schneider and R. M. Shiffrin, "Controlled and automatic human information processing: I. detection, search, and attention.," *Psychological review*, vol. 84, no. 1, p. 1, 1977.

[84] R. M. Shiffrin and W. Schneider, "Controlled and automatic human information processing: Ii. perceptual learning, automatic attending and a general theory.," *Psychological review*, vol. 84, no. 2, p. 127, 1977.

[85] W. J. Levelt, A. Roelofs, and A. S. Meyer, "A theory of lexical access in speech production," *Behavioral and brain sciences*, vol. 22, no. 1, pp. 1–38, 1999.

[86] B. Roberts and K. Kirsner, "Temporal cycles in speech production," *Language and Cognitive Processes*, vol. 15, no. 2, pp. 129–157, 2000.

[87] S. E. Lively, D. B. Pisoni, W. Van Summers, and R. H. Bernacki, "Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences," *The Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2962–2973, 1993.

[88] E. N. Sokolov, "Higher nervous functions: The orienting reflex," *Annual review of physiology*, vol. 25, no. 1, pp. 545–580, 1963.

[89] C. MacLeod and E. M. Rutherford, "Automatic and strategic cognitive biases in anxiety and depression.," 1998.

[90] A. Mathews and C. MacLeod, "Cognitive approaches to emotion and emotional disorders," *Annual review of psychology*, vol. 45, no. 1, pp. 25–50, 1994.

[91] H. Ellgring and K. R. Scherer, "Vocal indicators of mood change in depression," *Journal of Nonverbal Behavior*, vol. 20, no. 2, pp. 83–110, 1996.

[92] J. C. Bunn and J. Mead, "Control of ventilation during speech," *Journal of Applied Physiology*, vol. 31, no. 6, pp. 870–872, 1971.

[93] S. A. Shea, J. D. Hoit, and R. B. Banzett, "Competition between gas exchange and speech production in ventilated subjects," *Biological psychology*, vol. 49, no. 1-2, pp. 9–27, 1998.

[94] R. W. Levenson, P. Ekman, K. Heider, and W. V. Friesen, "Emotion and autonomic nervous system activity in the minangkabau of west sumatra.," *Journal of personality and social psychology*, vol. 62, no. 6, p. 972, 1992.

[95] R. W. Frick, "Communicating emotion: The role of prosodic features.," *Psychological Bulletin*, vol. 97, no. 3, p. 412, 1985.

[96] R.Banse and K. Scherer, "Acoustic profiles in vocal emotion expression," *J. Personality Social Psych*, vol. 70, no. 3, pp. 614–636, 1996.

[97] T. Johnstone and K. R. Scherer, "Vocal communication of emotion," *Handbook of emotions*, vol. 2, pp. 220–235, 2000.

[98] K. R. Scherer and A. Kappas, "Primate vocal expression of affective state," in *Primate vocal communication*, pp. 171–194, Springer, 1988.

[99] K. R. Scherer, "What does facial expression express?," in *Parts of the argument in this chapter have been presented at the 36th congress of the German Society of Psychology in Berlin, 1988.*, John Wiley & Sons, 1992.

[100] M. Alpert, R. L. Kurtzberg, and A. J. Friedhoff, "Transient voice changes associated with emotional stimuli," *Archives of General Psychiatry*, vol. 8, no. 4, pp. 362–365, 1963.

[101] T. Johnstone, C. M. van Reekum, K. Hird, K. Kirsner, and K. R. Scherer, "Affective speech elicited with a computer game.," *Emotion*, vol. 5, no. 4, p. 513, 2005.

[102] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572 – 587, 2011.

[103] C. Williams and K. Stevens, "Vocal correlates of emotional states, speech evaluation in psychiatry," *Grune and Stratton*, pp. 189–220, 1981.

[104] H. Teager, "Some observations on oral air flow during phonation," *IEEE Trans. Acoust. Speech Signal Process*, vol. 5, no. 28, pp. 599–601, 1990.

[105] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, 2001.

[106] G. Izzo, "Multiresolution techniques and emotional speech," *PHYSTA Project Report*, 1998.

[107] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, p. 153, 2007.

[108] E. Kandel, J. Schwartz, and T. Jessell, "Principles of neural science," *McGraw-Hill Medical*, 2000.

[109] H. Ghashghaei, C. Hilgetag, and H. Barbas, "Sequence of information processing for emotions based on the anatomic dialogue between prefrontal cortex and amygdala," *NeuroImage*, vol. 34, no. 3, pp. 905–923, 2007.

[110] M. Bar, R. Tootell, D. Schacter, D. Greve, B. Fischl, J. Mendola, B. Rosen, and A. Dale, "Cortical mechanisms speci," *Neuron*, vol. 29, no. 2, pp. 529–535, 2001.

[111] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for boltzmann machines*," *Cognitive science*, vol. 9, no. 1, pp. 147–169, 1985.

[112] Hinton and Sejnowski, "Optimal perceptual inference," *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 448–453, 1983.

[113] G. Hinton, "A practical guide to training restricted boltzmann machines," *Department of Computer Science, University of Toronto*, 2010.

[114] L. Arnold, S. Rebecchi, S. Chevallier, and H. Paugam-Moisy, "An introduction to deep learning," *inproceedings*, 2011.

[115] C. Andrieu, N. de Freitas, A. Doucet, and M. Jordan, "An introduction to mcmc for machine learning," *Machine Learning*, no. 50, pp. 5–43, 2003.

[116] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[117] G. Hinton, "Learning multiple layers of representation," *TRENDS in Cognitive Sciences*, vol. 11, no. 10, pp. 428–434, 2007.

[118] R. Reed, "Pruning algorithms-a survey," *IEEE transactions on Neural Networks*, vol. 4, no. 5, pp. 740–747, 1993.

[119] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," *ICASSP, IEEE*, pp. 5688–5691, 2011.

[120] R. Bruckner and B. Schuller, "Likability classification - a not so deep neural network approach," *Proceedings INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association*, 2012.

[121] O. Pele and M. Werman, "Fast and robust earth mover's distances," in *Computer vision, 2009 IEEE 12th international conference on*, pp. 460–467, IEEE, 2009.

[122] R. Rolon, L. Di Persia, H. L. Rufiner, and R. Spies, "Most discriminative atom selection for apnea-hypopnea events detection," in *Anales del VI Congreso Latinoamericano de Ingeniería Biomédica (CLAIB 2014)*, pp. 709–712, oct 2014.

[123] R. R. Wilcox, "Anova a paradigm for low power and misleading measures of effect size," *Review of Educational Research*, vol. 65, no. 1, pp. 51–77, 1995.

[124] H. Keselman, A. R. Othman, R. R. Wilcox, and K. Fradette, "The new and improved two-sample t test," *Psychological Science*, vol. 15, no. 1, pp. 47–51, 2004.

[125] S. Hegde, K. Achary, and S. Shetty, "Feature selection using fisher's ratio technique for automatic speech recognition," *arXiv preprint arXiv:1505.03239*, 2015.

[126] J. Lee Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *The American Statistician*, vol. 42, no. 1, pp. 59–66, 1988.

[127] X. Wei and K.-C. Li, "Exploring the within-and between-class correlation distributions for tumor classification," *Proceedings of the National Academy of Sciences*, vol. 107, no. 15, pp. 6737–6742, 2010.

[128] Q. Gu, Z. Li, and J. Han, "Generalized fisher score for feature selection," *arXiv preprint arXiv:1202.3725*, 2012.

[129] L. P. Prechelt *et al.*, "A set of neural network benchmark problems and benchmarking rules," 1994.

[130] M. C. Mozer and P. Smolensky, "Using relevance to reduce network size automatically," *Connection Science*, vol. 1, no. 1, pp. 3–16, 1989.

[131] E. D. Karnin, "A simple procedure for pruning back-propagation trained neural networks," *IEEE transactions on neural networks*, vol. 1, no. 2, pp. 239–242, 1990.

[132] J. Sietsma and R. J. Dow, "Creating artificial neural networks that generalize," *Neural networks*, vol. 4, no. 1, pp. 67–79, 1991.

[133] X. Jiang, M.-S. Chen, M. T. Manry, M. S. Dawson, and A. K. Fung, "Analysis and optimization of neural networks for remote sensing," *Remote Sensing Reviews*, vol. 9, no. 1-2, pp. 97–114, 1994.

[134] H. Chandrasekaran, H.-H. Chen, and M. T. Manry, "Pruning of basis functions in nonlinear approximators," *Neurocomputing*, vol. 34, no. 1-4, pp. 29–53, 2000.

[135] H.-J. Xing and B.-G. Hu, "Two-phase construction of multilayer perceptrons using information theory," *IEEE Transactions on Neural Networks*, vol. 20, no. 4, pp. 715–721, 2009.

[136] A. U. Levin, T. K. Leen, and J. E. Moody, "Fast pruning using principal components," in *Advances in neural information processing systems*, pp. 35–42, 1994.

[137] T. Cibas, F. F. Soulié, P. Gallinari, and S. Raudys, "Variable selection with neural networks," *Neurocomputing*, vol. 12, no. 2-3, pp. 223–248, 1996.

[138] A. Stahlberger and M. Riedmiller, "Fast network pruning and feature extraction by using the unit-obs algorithm," in *Advances in neural information processing systems*, pp. 655–661, 1997.

[139] V. Hozjan, Z. Kacic, A. Moreno, A. Bonafonte, and A. Nogueiras, "Interface databases: design and collection of a multilingual emotional speech database," *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC*, pp. 2024–2028, 2002.

[140] L. Rabiner and B.-H. Juang, "Fundamentals of speech recognition," *Prentice Hall PTR*, 1993.

[141] E. Albornoz, D. Milone, and H. Rufiner, "Spoken emotion recognition using hierarchical classifiers," *Computer Speech and Language*, vol. 25, pp. 556–570, 2011.

[142] F. Eyben, M. Wollmer, and B. Schuller, "opensmile - the munich versatile and fast open-source audio feature extractor," *ACM Multimedia (MM)*, pp. 1459–1462, 2010.

[143] D. Wulsin, "Dbn toolbox v1.0, department of bioengineering, university of pennsylvania," 2010.

[144] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 1–27, April 2011.

[145] Mathworks, "Matlab users guide (r2017b)." `http://www.mathworks.com`, 2017.

[146] M. E. Sánchez-Gutiérrez, E. M. Albornoz, F. Martinez-Licona, H. L. Rufiner, and J. Goddard, *Pattern Recognition*, ch. Deep Learning for Emotional Speech Recognition, pp. 311–320. Cham: Springer International Publishing, 2014.

[147] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in *Proc. of 9th European Conference on Speech Communication and Technology (Interspeech)*, pp. 1517–1520, Sep. 2005.

[148] S. S. Haykin, S. S. Haykin, S. S. Haykin, and S. S. Haykin, *Neural networks and learning machines*, vol. 3. Pearson Upper Saddle River, NJ, USA:, 2009.

[149] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[150] H. L. Rufiner, M. E. Torres, L. G. Gamero, and D. H. Milone, "Introducing complexity measures in nonlinear physiological signals: application to robust speech recognition," *Physica A: Statistical Mechanics and its Applications*, vol. 332, no. 1, pp. 496–508, 2004.

[151] E. M. Albornoz, M. Sánchez-Gutiérrez, F. Martinez-Licona, H. L. Rufiner, and J. Goddard, *Spoken Emotion Recognition Using Deep Learning*, pp. 104–111. Cham: Springer International Publishing, 2014.

[152] D. Michie, D. Spiegelhalter, and C. Taylor, *Machine Learning, Neural and Statistical Classification*. London: Ellis Horwood, University College, 1994.

[153] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[154] K. N. Stevens, *Acoustic phonetics*, vol. 30. MIT press, 2000.

[155] C. Martínez, J. Goddard, D. Milone, and H. Rufiner, "Bioinspired sparse spectro-temporal representation of speech for robust classification," *Computer Speech & Language*, vol. 26, no. 5, pp. 336–348, 2012.

[156] L. D. Vignolo, H. L. Rufiner, and D. H. Milone, "Multi-objective optimisation of wavelet features for phoneme recognition," *IET Signal Processing*, vol. 10, no. 6, pp. 685–691, 2016.

[157] E. D. Übeyli, "Implementing automated diagnostic systems for breast cancer detection," *Expert systems with Applications*, vol. 33, no. 4, pp. 1054–1062, 2007.

[158] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in *Biomedical Image Processing and Biomedical Visualization*, vol. 1905, pp. 861–871, International Society for Optics and Photonics, 1993.

[159] A. M. Abdel-Zaher and A. M. Eldeib, "Breast cancer classification using deep belief networks," *Expert Systems with Applications*, vol. 46, pp. 139–144, 2016.

[160] M. Sánchez-Gutiérrez, E. M. Albornoz, H. L. Rufiner, and J. G. Close, "Post-training discriminative pruning for rbms," *Soft Computing*, pp. 1–15, 2017.

[161] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[162] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *Journal of machine learning research*, vol. 13, no. Jan, pp. 27–66, 2012.

[163] A. El Akadi, A. El Ouardighi, and D. Aboutajdine, "A powerful feature selection approach based on mutual information," *International Journal of Computer Science and Network Security*, vol. 8, no. 4, p. 116, 2008.

[164] A. Vergara, S. Vembu, T. Ayhan, M. A. Ryan, M. L. Homer, and R. Huerta, "Chemical gas sensor drift compensation using classifier ensembles," *Sensors and Actuators B: Chemical*, vol. 166, pp. 320–329, 2012.

[165] M. M. Adankon and M. Cheriet, "Support vector machine," in *Encyclopedia of biometrics*, pp. 1303–1308, Springer, 2009.

[166] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 3, pp. 1315–1318, IEEE, 2000.

[167] M. Reyes-Vargas, M. Sánchez-Gutiérrez, L. Rufiner, M. Albornoz, L. Vignolo, F. Martínez-Licona, and J. Goddard-Close, "Hierarchical clustering and classification of emotions in human speech using confusion matrices," vol. 8113, pp. 162–169, 2013.

[168] M. Sánchez-Gutiérrez, J. Goddard, E. Albornoz, H. Rufiner, and F. Martinez-Licona, "Redes de creencia profunda y emociones," *Komputer Sapiens*, vol. 1, no. 9, pp. 12–18, 2017.

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA

ACTA DE DISERTACIÓN PÚBLICA

No. 00009
Matrícula: 2141801981

EVALUACIÓN DE LA CAPACIDAD DISCRIMINATIVA DE LAS MÁQUINAS RESTRINGIDAS DE BOLTZMANN

MAXIMO EDUARDO SANCHEZ GUTIERREZ

ALUMNO

REVISÓ

LIC. JULIO CESAR DE LARA ISASSI
DIRECTOR DE SISTEMAS ESCOLARES

En la Ciudad de México, se presentaron a las 10:00 horas del día 11 del mes de octubre del año 2018 en la Unidad Iztapalapa de la Universidad Autónoma Metropolitana, los suscritos miembros del jurado:

DRA. MARIA DEL CARMEN GOMEZ FUENTES
DR. JORGE CERVANTES OJEDA
DR. ENRIQUE MARCELO ALBORNOZ
DR. CESAR MARTINEZ
DR. PEDRO PABLO GONZALEZ PEREZ

Bajo la Presidencia de la primera y con carácter de Secretario el último, se reunieron a la presentación de la Disertación Pública cuya denominación aparece al margen, para la obtención del grado de:

DOCTOR EN CIENCIAS (CIENCIAS Y TECNOLOGIAS DE LA INFORMACION)

DE: MAXIMO EDUARDO SANCHEZ GUTIERREZ

y de acuerdo con el artículo 78 fracción IV del Reglamento de Estudios Superiores de la Universidad Autónoma Metropolitana, los miembros del jurado resolvieron:

Aprobar

Acto continuo, la presidenta del jurado comunicó al interesado el resultado de la evaluación y, en caso aprobatorio, le fue tomada la protesta.

DIRECTOR DE LA DIVISIÓN DE CBI

DR. JESUS ALBERTO OCHOA TAPIA

PRESIDENTA

DRA. MARIA DEL CARMEN GOMEZ FUENTES

VOCAL

DR. JORGE CERVANTES OJEDA

VOCAL

DR. ENRIQUE MARCELO ALBORNOZ

VOCAL

CANCELADO

DR. CESAR MARTINEZ

SECRETARIO

DR. PEDRO PABLO GONZALEZ PEREZ