ARTICULO 7

SISTEMA DE RECONOCIMIENTO AUTOMÁTICO DEL HABLA

SUCCI AGUIRRE CLOVIS GUSTAVO
Universidad Autónoma Juan Misael Saracho, Tarija -Bolivia
gustsucc@gmail.comw

1. RESUMEN

Los primeros intentos de desarrollo de sistemas de ASR (Answer Speech Recognition), datan de los años 50. Estos primeros trabajos abordaban el reconocimiento de un vocabulario reducido, del orden de 10 palabras, emitidas por un único locutor. La década de los 60 marca el inicio de tres proyectos que han tenido gran repercusión en el área. Estos proyectos fueron desarrollados por Martin (RCA Labs.), en el campo de la normalización de la voz; Vintsyuck (URSS), en métodos de programación dinámica, y Reddy (CMU), quien introdujo la Inteligencia Artificial en la Interacción Ordenador/Persona.

El reconocimiento de la voz mediante diversas técnicas tales como cadenas ocultas de Markov y Redes Neuronales es tema de investigación constante, obteniendo resultados de distinta performance según el método elegido. En el presente artículo se comentan los resultados de una experiencia en reconocimiento de voz de un individuo, tomando como patrones a ser reconocidos las cifras decimales (0- 9), y utilizando como método una red neuronal de Kohonen. Luego de una fase de entrenamiento y sintonización, produce con solo cien neuronas un aceptable resultado de reconocimiento (65%).

2. PALABRAS CLAVE

Reconocimiento automático del habla - interfaces hombre-máquina - inteligencia artificial - señal de voz (procesamiento y análisis) Rede Neuronal. ASR.

3. INTRODUCCIÓN

A pesar de la sencillez que parece presentar el problema del habla para los humanos, el estudio de la misma muestra, de forma inmediata, una enorme complejidad. en ella aparecen mezclados varios niveles de descripción, que interactúan entre sí. de esta forma, el problema del reconocimiento del habla presenta una naturaleza interdisciplinaria, y para solucionarlo es necesario aplicar técnicas y conocimientos procedentes de las siguientes áreas (rabiner&juang): procesamiento de señales, física (acústica), reconocimiento de patrones, teoría de la información y comunicaciones, lingüística, fisiología, informática y psicología. además de la interdisciplinariedad expuesta, existen algunos aspectos prácticos relacionados con el habla que hacen del asr una tarea difícil. estos se pueden agrupar en seis categorías (VARILE&ZAMPOLLI):

*1. **Continuidad:*** en el lenguaje natural no existen separadores entre las unidades, ya que no existen silencios, en algunos casos, ni entre las palabras.

*2. **Dependencia del contexto:*** cada sonido elemental en los que se puede dividir el habla (fonema) es modificado por el contexto en el que se encuentra. de esta forma, se produce el efecto denominado coarticulación, según el cual los fonemas anterior y posterior a uno dado modifican el aspecto del mismo. aparecen también efectos de orden superior, dependiendo la pronunciación de un fonema, de su situación en una palabra o incluso en una frase.

*3. **Variabilidad:*** se pueden distinguir dos tipos de variabi-

lidad. la variabilidad intra-hablante está relacionada con las modificaciones introducidas por un mismo hablante sobre diferentes pronunciaciones de los mismos fonemas o palabras. incluso en idénticas condiciones, cada pronunciación presentará diferencias con las restantes debido a la diferente duración temporal. la variabilidad inter-hablante se debe a aspectos relacionados con el locutor y el entorno, ya que la señal obtenida dependerá de los dispositivos utilizados en su captación, del entorno donde se obtiene y, principalmente, de aspectos anatómicos particulares del aparato fonador de cada hablante.

4. Necesidades de almacenamiento: debido a las causas anteriores, se hace necesario procesar y almacenar grandes cantidades de datos.

5. Estructuración: la misma señal contiene información sobre varios niveles de descripción. de esta forma, una frase puede ser descrita en el nivel semántico, sintáctico o fonético. por otra parte, una señal de voz contiene información sobre el locutor que la emite. así, es posible distinguir el sexo y la identidad de la persona a partir de la propia señal. un sistema de ASR tendría que determinar qué información de las citadas es de interés para lograr su objetivo.

6. Inexistencia de reglas de descripción y redundancia: no existen reglas precisas capaces de describir los diferentes niveles en los que se presenta la información. es más, cada uno de los niveles citados anteriormente aparece fuertemente relacionado con los demás, dificultando el análisis de la voz.

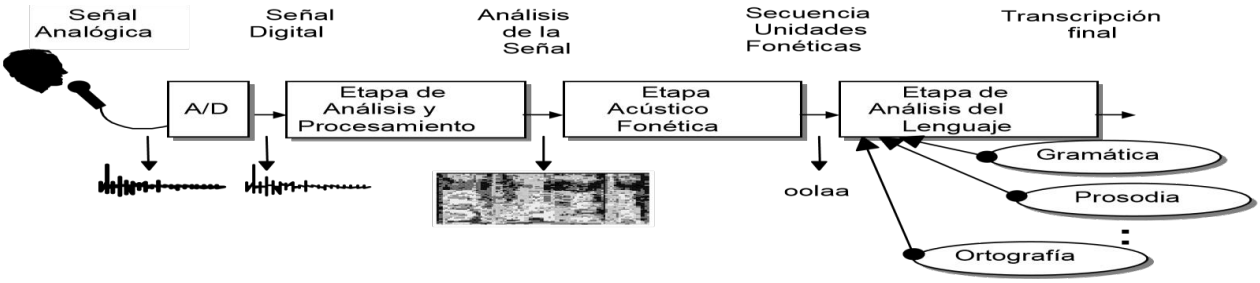

Figura 1: Esquema Reconocimiento del Habla

4. RECONOCIMIENTO AUTOMÁTICO DEL HABLA

El reconocimiento de la voz tiene como dificultad principal reconocer el habla sorteando la gran disparidad entre los registros vocales de distintas personas, según su sexo, edad y pronunciaciones típicas correspondientes a distintas zonas geográficas. Todas estas razones hacen que dos personas que pronuncian la misma palabra, posean patrones frecuenciales y temporales en dicha pronunciación radicalmente distintos, y que en apariencia, no guardan similitud alguna.

De los múltiples métodos que pudieran ser usados para el reconocimiento de la voz [3][4][6][11], como análisis espectrales [5][8], estadísticos [12], etc., en el presente trabajo se optó por comentar una solución mediante una red neuronal de Kohonen [1][10]. La elección de esta red se debe a que su algoritmo de entrenamiento es bastante simple, lo que determina tiempos de aprendizaje aceptables, en comparación con otro tipo de redes, para una misma cantidad de neuronas. Por otra parte, su característica de mapa autoorganizable aporta una vía natural para asociar neuronas cercanas a pronunciaciones similares dentro del espacio de muestras. El espacio de muestras que debe ser aprendido por la red es el constituido por las diez cifras decimales, y se realizó la experiencia con registros sonoros de tres individuos de sexo masculino adultos. Para no tener tiempos de entrenamiento ni de reconocimiento muy extensos, se ha optado por implementar la red con solo 100 neuronas. El sistema ha sido implementado y probado en una plataforma PC, y con una placa de sonido Sound Blaster

4.1. Técnicas más utilizadas aplicadas al Reconocimiento del Habla

Las técnicas que más se utilizan en el reconocimiento automático del habla son:

Técnicas de Programación Dinámica (DTW)

Esta técnica consiste en realizar una comparación entre los patrones o plantillas de las que dispone el sistema con la señal acústica recibida como entrada, de esta forma se obtienen posibles candidatos a los que puede pertenecer la señal recibida. Para realizar esta tarea tan compleja se parametriza la señal recibida y se transforma la señal de entrada en coeficientes espectrales para analizarla de forma correcta. Una vez se obtiene los espectros de la señal comienza el proceso de reconocimiento comparándolo con los patrones almacenados. Esta técnica, es utilizada tanto para resolver problemas de reconocimiento de habla continua como aislada. Sin embargo esta técnica suele tener algunos problemas debido a: la duración de la palabra no tiene que ser de una duración determinada, por lo que puede que no coincida con la de la plantilla; y el ritmo con el que se realiza la pronunciación no tiene que mantenerse constante por lo que no se ajustará a la plantilla en ese sentido, ya que este depende de la persona.

Modelos Ocultos de Markov (HMM)

Un modelo oculto de Markov se puede considerar como una especie de autómata finito, ya que está formado por una serie de estados que tienen una conexión directa mediante transiciones. El proceso va a dar comienzo en un estado inicial, específicamente diseñado para ello, y cada uno de los estados va a tener asociado un conjunto de probabilidades sobre un grupo de símbolos salientes. Por cada una de las ejecuciones, se va a elegir una transición hacia un estado nuevo, y se va a generar un símbolo de salida relacionado con dicho estado. La elección en cada ejecución de cada transición y símbolo se va realizar en función de probabilidades, y por tanto va a ser una elección completamente aleatoria. La característica principal de los modelos de Markov es no se va a conocer nunca el conjunto de estados por los que el proceso ha realizado el recorrido hasta llegar al conjunto de símbolos obtenidos en la salida, y este es el motivo fundamental por el que se le conoce como Modelo oculto de Markov. Al aplicar los modelos ocultos de Markov al reconocimiento del habla, cada estado va a indicar cuáles son aquellos sonidos que son más probables para cada segmento del habla, mientras que las transiciones van a ser restricciones temporales para cada uno de esos sonidos, indicando cuáles son sus secuencias de apariciones.

Redes Neuronales

El estudio de las redes neuronales fue abandonado prácticamente desde que aparecieran, debido a que no se podía llevar a cabo su entrenamiento con algoritmos que fuesen eficientes. Sin embargo, en la actualidad ha quedado perfectamente demostrado que los modelos basados en las redes neuronales cuentan con una gran potencia desde el punto de vista computacional. Las redes neuronales son una estructura de procesamiento y aprendizaje de información, que está formada por un conjunto de nodos que se denominan neuronas, las cuales están conectadas mediante una serie de pesos. Cada neurona va a recibir una entrada a partir de las conexiones que tiene con el resto de neuronas, y va a producir una salida. Gracias a las ventajas que tienen (capacidad de aprendizaje, tolerancia ante fallos, capacidad de producir respuestas en tiempo real...),las redes neuronales han pasado a ser una de las mejores soluciones para abordar el problema del reconocimiento automático del habla. Sin embargo, los sistemas basados en redes neuronales también tienen algunos inconvenientes como puede ser el elevado tiempo de entrenamiento necesario o el desconocimiento previo del número de nodos que se necesitan para abordar un problema. Esto implica que se haga necesario combinar dichos sistemas con técnicas basadas en programación dinámica y en modelos ocultos de Markov.

5. CASO PRÁCTICO

La red utilizada es una red de Kohonen, como ya se menciono. No obstante para el aprendizaje de la misma se divide el proceso en dos fases [1], a saber: una primera fase de entrenamiento clásico donde se presentan en forma continua los distintos patrones a la red, y se modifican las posiciones de las neuronas en función de estos; una segunda y ultima fase de ajuste fino, también conocida como algoritmo de Cuantizacion del Vector de Aprendizaje [7][9][13] (Learning Vector Quantization - LVQ) en el cual

primero se caratula cada neurona con una etiqueta con el mismo numero del patrón más próximo, y que en teoría debería representar, y a continuación se van eligiendo al azar pares patron-neurona, corrigiéndose la posición de la neurona mediante un sistema de "premio-castigo", que se explicara mas adelante. Ambos pasos se realizan en el modulo 'SOM' (por "Self-Organizing Map").

NUMERO PRONUNCIADO	TASA DE RECONOCIMIENTO
0	9/10
1	5/10
2	6/10
3	7/10
4	6/10
5	6/10

NUMERO PRONUNCIADO	TASA DE RECONOCIMIENTO
6	8/10
7	6/10
8	4/10
9	7/10

Figura 2: Tasas de Reconocimuento Inicial

5.1. Datos y validación de sistemas de ASR

En esta sección discutimos brevemente los datos utilizados para la experimentación.

Bases de datos

1. Peterson: esta es una pequeña base pública con datos sobre las formantes del inglés que se utilizó en algunas pruebas iniciales de los clasificadores.

2. TIMIT: esta base de datos es la más utilizada en el campo del ASR, está formada por unas 6000 frases en idioma inglés. Fue la base utilizada en las primeras pruebas.

*3. TIMEX:*se realizaron tareas conjuntas con la UAM (México) a fin de lograr el corpus de fonemas en castellano que culminó en el diseño, la grabación y etiquetado de un corpus formado por 10 hablantes mexicanos. Las tareas de etiquetado se llevaron a cabo en el Laboratorio de Audiología de la UAM. Se trabajó sobre estas grabaciones etiquetadas de manera de adaptarlas y organizarlas en forma similar al corpus TIMIT (inglés), ya que todas las rutinas de análisis y clasificación se desarrollaron basándose en esta estructura. Para esto, primero se submuestrearon las grabaciones, que originalmente se habían adquirido a 22050 Hz, obteniéndose archivos de señales muestreadas a 16000 Hz. Una vez organizada la base de oraciones, se procedió a la obtención de patrones de entrenamiento y prueba para cinco fonemas, a partir de las rutinas para procesamiento con MFCC, de manera de poder realizar una comparación con los resultados obtenidos para los cinco fonemas más cercanos de TIMIT. Estos patrones se utilizaron para entrenar una red neuronal tipo TDNN, obteniéndose resultados preliminares muy buenos (mejores que para TIMIT) aunque poco significativos debido al pequeño número de hablantes de este corpus.

4. Latino 40: esta base está formada por 40 hablantes latinos y desarrollada por el Stanford Research Institute (SRI). Fue la primera base en castellano de tamaño mediano utilizada en el PID. Los experimentos con POF se realizaron con esta base. Un problema que presenta es la gran variabilidad dialéctica de los datos.

5. Albayzin: esta es una base en idioma español desarrollada por un grupo de universidades españolas. Se ha podido tener acceso a la misma por colaboración con el Grupo de Investigación en Procesamiento de Señales y Comunicaciones de la Universidad de Granada y se describe adecuadamente en la separata sobre el sistema de referencia, ya que fue utilizada para su desarrollo.

6. CSLU: recientemente se adquirieron una serie de datos pertenecientes a conversaciones telefónicas en múltiples idiomas que permiten atacar los problemas de degradación que presenta este canal.

Obtención de resultados

Para la obtención de las medidas finales de desempeño de los sistemas se debe estimar el error cometido en alguna tarea particular. Para ello, el método clásico en bases de datos grandes consiste en dividir la base en dos partes y utilizar una para entrenamiento y la otra para prueba. Si bien los primeros resultados se obtuvieron de esta forma, el método clásico presenta limitaciones dado que introduce sesgos debidos a la evaluación de error en la partición particular elegida. Para evitar estos sesgos se puede recurrir al método denominado dejar-k-afuera (LKO, leavek-out) que permite realizar particiones múltiples de los datos y luego estimar el error basándose en el promedio sobre todas estas particiones. Mediante LKO se realizaron los experimentos más recientes.

6. PROCESOS DE APRENDIZAJE

El reconocimiento automático del habla consta de dos procesos de aprendizaje diferenciados:

Por un lado un **aprendizaje deductivo** que consiste en la transferencia de conocimientos del hombre a un sistema informático;

Y por otro lado un **aprendizaje inductivo,** aquí se trata de que el sistema sea capaz de obtener esos conocimientos a través de ejemplos. La evolución de los sistemas de reconocimiento de voz hacen que la máquina pueda interpretar como un "sí" no solo si oye esa palabra, sino también si escucha expresiones equivalentes. Lo que facilita el reconocimiento del lenguaje natural tal y como lo haría un ser humano, con una exactitud por encima del 90%. En este segundo tipo, los ejemplos los constituyen aquellas partes de los sistemas basados en los modelos ocultos de Márkov o en redes neuronales artificiales que son configuradas automáticamente a partir de muestras de aprendizaje.

7. ¿POR QUÉ SE DEBE ESTUDIAR EL RECONOCIMIENTO DEL HABLA?

Hablar es el medio más natural para las personas a la hora de comunicarnos. Para interactuar del mismo modo con el mundo digital que nos rodea, lo primero que necesitamos es que las máquinas (el ordenador, el teléfono, el coche, etc.) nos entiendan. El reconocimiento del habla proporciona las herramientas necesarias para transformar la voz en conceptos que después puedan utilizar las máquinas para emprender acciones.

Los sistemas comerciales han estado disponibles desde 1990. A pesar del aparente éxito de estas tecnologías, muy pocas personas utilizan el sistema del reconocimiento del habla en sus computadoras. Parece ser que muchos de los usuarios utilizan el ratón y el teclado para guardar o redactar documentos, porque les resulta más cómodo y rápido a pesar del hecho de que todos podemos hablar a más velocidad de la que tecleamos. Sin embargo, mediante el uso de ambos, el teclado y el reconocimiento del habla, nuestro trabajo sería mucho más efectivo.

Este sistema donde está siendo más utilizado es en aplicaciones telefónicas: agencias de viajes, atención al cliente, información etc. La mejora de estos sistemas de reconocimiento del habla ha ido aumentando y su eficacia cada vez es mayor.

Con los importantes avances en el lenguaje natural y las tasas de precisión del habla, los avances en la tecnología de reconocimiento del habla han llevado a una creciente presión para que las empresas construyan experiencias habilitadas por la voz que superen las expectativas de los usuarios. Las mejoras en tándem en la IA, la computación en nube y la ciencia de los datos han permitido que la tecnología como el comando de voz avance a velocidades sin precedentes, cambiando la forma en que las empresas diseñan sus tácticas de servicio al cliente.

8. CONCLUSIONES

El trabajo muestra un aceptable comportamiento de la red neuronal de Kohonen frente a la tarea de reconocimiento de un patrón tan estocástico como lo es la voz. Aunque su tasa de acierto dista aún mucho de las obtenidas mediante otros métodos, existen muchas variables en la red neuronal que pueden ser ajustadas para lograr mejores comportamientos de la misma; entre ellas, se pueden considerar:

- Número de neuronas de la red.
- Dimensión de los patrones y de dichas neuronas.
- Cantidad de locutores y de archivos sonoros utilizados en el entrenamiento.
- Método de codificación o mapeo elegido para la aplicación
- Valores de los coeficientes de aprendizaje y sintonización, y forma de decaimiento en el tiempo.

Todos estos factores interrelacionados contribuirían a mejorar el rendimiento de la red.

La sílaba tiene muchas propiedades que son deseables para la computación vectorial: 1) los modelos basados en sílabas pueden ser conducidos a remover las ramificaciones durante la ejecución y 2) los modelos basados en sílabas son una unidad de organización natural para reducir la computación redundante y define el espacio de búsqueda.

De la misma forma aunque este trabajo no explora los beneficios de la programación paralela, algunas de las conclusiones de este trabajo son aplicables al procesamiento concurrente. A saber, la combinación de información de múltiples cadenas de Markov es una operación obviamente concurrente. El decodificador de dos niveles de Fosler-Lussier puede ser mapeado cuidadosamente en una máquina de procesador múltiple, dado que las probabilidades de diferentes palabras son calculadas independientemente.

Si este es el caso, usando máquinas paralelas y concurrentes puede ser ampliamente ventajosa la investigación del reconocimiento de voz. Asimismo, la combinación de la metodología empleada en el presente trabajo al unirse con la basada en fonemas abre un campo de estudio relevante. Un punto importante que puede incrementar el camino de la investigación en lo que a la inmersión de las sílabas a los sistemas de reconocimiento se refiere, es el hecho de introducir un conjunto de filtros que permitan determinar de manera adecuada las manifestaciones de fonemas de mayor ocurrencia en un corpus de voz que conforman a las sílabas. Además, la particularidad de mejorar el problema de la entonación logrará incrementar el alcance que la sílaba tiene dentro del idioma español.

Finalmente, los trifonemas pueden ser analizados como unidades de reconocimiento y comparar los resultados que se obtengan con los expuestos en este trabajo, procurando establecer una alternativa de utilización de ambas unidades esenciales. Hay idiomas donde la sílaba es una muy buena alternativa, en otros no, tal es el caso del idioma Español.

9. REFERENCIAS Y BIBLIOGRAFÍA

- LOPEZ-COZAR R., MILONE D. H. “A new technique based on augmented language models to improve the performance of spoken dialogue systems”. En: EuroSpeech. 2001, 2001:741-744.
- LOPEZ-COZAR R., RUBIO A. J., BENITEZ M., MILONE D. H. “Restricciones de funcionamiento en tiempo real de un sistema automático de dialogo”. En: Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural, 26, 2000:169-174.
- MALLAT S. G. A Wavelet Tour of signal Processing. 2 da. ed. Academic Press, New York, 1999.
- MARTINEZ C. E., RUFINER H. L., “Acoustic Analysis of Speech for Detection of Laryngeal Pathologies”. En: Proceedings of the Chicago 2000 World Congress IEEE EMBS, Paper No. TH-Aa325-07, Chicago, July 2000.
- MARTINEZ S., “Desarrollo de herramientas computacionales para la detección del nivel de audición y apoyo en el aprendizaje del habla para niños sordos e hipoacúsicos”, Tesina de grado, Bioingeniería, FIUNER, 2001.
- MILONE D. H. “Reconocimiento automático del habla con redes neuronales artificiales”. 2001 . Inédito.
- MILONE D. H., MERELO J. J. “Evolutionary algorithm for speech segmentation”. En: 2002 IEEE World Congress on Computational Intelligence, Hilton Hawaiian Village Hotel, Honolulu, 2002. Paper No. 7270.
- MILONE D. H., RUBIO A. J. “Including prosodic cues in asr systems”. En: Proceedings 5 th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2001) and the 7th International Conference on Information Systems Analysis and Synthesis (ISAS 2001), Orlando, Julio 2001. Paper No. IS0051403.——————. “Prosodic and accentual information for automatic speech recognition”. En: IEEE Trans. on Speech and Audio Processing, 11(4), 2003: 321333.

- MILONE D. H., RUBIO A. J., LOPEZ-COZAR R. "Modelos de lenguaje variantes en el tiempo". Memorias del XXIV Congreso Nacional de Ingeniería Biomédica, Oaxtepec, México, 10-13 de Octubre 2001.
- MILONE D. H., SAEZ J. C., SIMON G., RUFINER H. L. "Self-organizing neural tree networks". En: Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Hong Kong, 3, 1998a:1348-1351.————. "Árboles de redes neuronales autoorganizativas". En: Revista Mexicana de Ingeniería Biomédica, 19(4), 1998b:1326.
- RABINER L., JUANG B. H. Fundamentals of Speech Recognition. Signal Processing Series. Prentice-Hall, 1993.
- RUFINER H. L., GODDARD J., MARTINEZ A. E., MARTINEZ F. M. "Basis pursuit applied to speech signals". En: Proceedings 5th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI 2001) and the 7th International Conference on
- Information Systems Analysis and Synthesis (ISAS 2001), Orlando, July 2001:517520.
- RUFINER H. L. "Comparación entre análisis onditas y Fourier aplicados al reconocimiento automático del habla". Master's thesis, Universidad Autónoma Metropolitana, Diciembre 1996 (Inédito).———. "Modelización biológica, redes neuronales y HMM´s aplicados al reconocimiento automático del habla". Informe de avance, Beca de investigación Conicet, CONICET, 1994.
- RUFINER H. L., CORNEJO J. M., CADENA M., HERRERA E. "Laboratorio de voz". Anales del VIII Congreso de la Asociación Mexicana de Audiología, Foniatría y Comunicación Humana, Veracruz, México, 1997.
- RUFINER H. L., ROCHA L. F., GODDARD J. "Denoising of speech using sparse representations". Proc. of the International Conference on Acoustic, Speech & Signal Processing, 2002:989-992.
- ARONSON L., RUFINER L., FURMANSKY H., ESTIENE P. "Características acústicas de las vocales del español rioplatense". En: Fonoaudiológica, 46 (2), 2000:12-20.
- DELLER J., PROAKIS J., HANSEN J. Discrete Time Processing of Speech Signals. Macmillan Publishing, New York, 1993.
- GAMERO L. G., RUFINER H. L. "Paquetes de onditas evolutivas para clasificación de señales". En: Anales del Ier Congreso Latinoamericano de Ingeniería Biomédica, 1, 1998:784-787.
- GAMERO L. G., PLASTINO A., TORRES M. E. "Wavelet analysis and nonlinear dynamics in a non extensive setting". En: Physica A. 246, 1997:487-509.
- GODDARD J. C., MARTINEZ F. M., MARTINEZ A. E., CORNEJO J. M., RUFINER H. L., ACEVEDO R. C. "Redes neuronales y Árboles de decisión: Un enfoque híbrido". Memorias del Simposium Internacional de Computación organizado por el Instituto Politécnico Nacional, 1995.