♡ A mis padres
♡ A mi hermano
♡ A mis abuelitos
♡ A mi novio

RESUMEN

Para analizar las propiedades de objetos celestes, los astrónomos utilizan sus espectros. En un espectro se encuentra contenida toda la información de la composición química del objeto. Un espectro está formado por tres componentes: el continuo, las líneas de absorción y las líneas de emisión.

La cantidad de información disponible que puede analizar un astrónomo para sus investigaciones, se ha vuelto demasiado grande para realizar un análisis manual de ella, por lo tanto, se han comenzado a emplear métodos de aprendizaje automático para analizar de forma automática los parámetros en espectros. Utilizando estos métodos los astrónomos pueden evaluar grandes cantidades de información en un lapso de tiempo aceptable.

Desde una perspectiva teórica, es muy importante identificar los principales periodos de formación estelar en las galaxias, esto con el fin de entender su proceso de evolución, así como su formación. Pero esta tarea presenta algunos problemas, debidos principalmente a que los espectros observacionales normalmente tienen niveles de ruido desconocidos y además presentan otras características difíciles de modelar.

El enfoque más común que se ha utilizado para determinar los parámetros de espectros, consiste en ajustar un espectro sintético con parámetros conocidos a uno observacional, asumiendo que éste tiene los parámetros del espectro con que mejor se ajuste. Este enfoque puede conducir a resultados erróneos, debido a distorsiones producidas por ruido o por la lejanía de las galaxias, también a efectos de discretización, pero principalmente a la presencia de líneas de emisión en espectros reales, las cuales no se toman en cuenta para la creación de los modelos de síntesis de evolución estelar.

Para evitar los problemas anteriores, en esta tesis se utilizaron sólo cinco líneas del espectro, las cuales contienen la información relevante asociada con las edades de las principales poblaciones estelares. Cada una de estas líneas se codifica mediante los anchos en doce regiones, obteniendo así, el comportamiento general de su perfil. Utilizando esta representación del espectro se entrenaron clasificadores para determinar las poblaciones estelares que lo integran.

Para resolver el problema se usaron dos enfoques principales de aprendizaje automático. El primero consistió en evaluar tres tipos de ensambles usando algoritmos tradicionales del aprendizaje automático. Mediante los resultados experimentales obtenidos se encontraron ciertas regularidades en los datos, lo que condujo a la creación del segundo enfoque. Donde el algoritmo propuesto tiene forma de árbol, pero se hace una predicción parcial en cada nodo y se decide a partir del vector final de predicciones, aceptando la que presente el menor error entre el espectro original y el espectro reconstruido mediante los parámetros predichos.

Los resultados experimentales mostraron que el segundo enfoque obtuvo mejores ajustes de los perfiles de las líneas, así como mejores espectros reconstruidos. El tiempo de predicción no sobrepasó en ningún caso los 2.5 segundos por espectro, además la precisión obtenida se puede calificar como aceptable en el caso de espectros observacionales, ya que el error total es solamente 0.8% peor que el que produce una búsqueda exhaustiva.

ABSTRACT

To analyze the properties of celestial objects, astronomers use spectra. In a spectrum, is encoded all the information about the chemical composition of the object. A spectrum is a plot of flux against wavelength and it is formed for three components: the continuum, absorption lines, and emission lines. The continuum is a smooth spectrum generated for solid objects and dense gases that emit radiation in a wide range of wavelength; the absorption and emission lines represents different atomic interactions in specific wavelengths of the light.

For astronomers, the amount of data that they need to analyze has become too large to be treated manually, therefore, a variety of machine learning methods have started to be used to perform automatic determination of parameters in spectra. Using those methods, astronomers can evaluate large amounts of information in an acceptable period of time.

From a theoretical point of view, in order to understand galaxy formation and evolution, it is very important to identify the principal periods of star formation, but the task of determining the ages of stellar populations in galactic spectra is not trivial, since spectra are normally noisy and present other hard to model features.

The most common approach that has been used, consists of fitting a spectrum to a synthetic model. This approach often yields erroneous results when is applied to observational spectra, mainly because of the presence of emission lines, which are not taken into consideration in standard synthetic models. Those lines are considerably stronger than absorption lines and therefore they define the shape of the spectrum; also there is the problem of certain distortions such as noise and discretization effects.

The approach we propose consists of fitting just a few lines of the spectrum that contain the information related to the age of the stellar populations. It has the additional advantage that it does not use regions with strong emission lines.

The main contribution of this work is an efficient machine learning method, insensitive to emission lines and robust in the presence of noise, to determine the age of stellar populations in observational galactic spectra.

It is a novel method for constructing ensembles where the decision in the prediction is done subdividing the data and organizing them in a tree structure, obtaining a good approximation when the instance reaches a leaf node or in the path behind. Experimental results show that this method yields a better accuracy than the traditional ensembles, and the total error in observational spectra is only 0.8% worst than the exhaustive search.

AGRADECIMIENTOS

Agradezco al Dr. Olac Fuentes por la confianza y apoyo que me brindó durante el desarrollo de la tesis, así como por todos los valiosos conocimientos que compartió conmigo a lo largo de estos dos años, su disposición para resolver mis dudas y el tiempo dedicado a componer mis errores.
Gracias a los doctores Roberto y Elena Terlevich por plantear el problema astrofísico y por ayudarme con su experiencia, así como al M.C. Juan Pablo Papaqui, por colaborar conmigo en las etapas iniciales de esta tesis.
Quiero agradecer a los doctores Ariel Carrasco, Jesús González y Carlos Reyes, por sus acertados comentarios y el tiempo dedicado para la lectura y corrección de esta tesis. También agradezco a todas las personas de ciencias computacionales que de alguna manera me apoyaron durante la maestría, en especial al Dr. Gustavo Rodriguez, Dr. Aurelio López, Dr. Luis Villaseñor y M.C. Thamar Solorio.
Un agradecimiento muy especial, es para el Dr. Jesús Favela, quien me mostró que el horizonte es más grande cuando realmente queremos verlo y por contestar hasta mis cartas más triviales, así como para el M.C. Aarón Govea por su amistad y su confianza.
Agradezco a las instituciones, sin las cuales este trabajo y mis estudios de maestría no hubieran sido posibles, es decir, al Instituto Nacional de Astrofísica Óptica y Electrónica (INAOE), así como al Consejo Nacional de Ciencia y Tecnología (CONACyT), a éste último por el apoyo brindado a través de la beca crédito número 171595.

También deseo expresar mi más sincero agradecimiento:
⋆ A mis amigos en Guadalajara: Ana, Cristy, Karla, José y José Enrique; por hacerme sentir siempre tan cerca.
⋆ A mis compañeros de la maestría, a los del grupo de aprendizaje automático y sobre todo a los del cubo de la suerte: Toño, Carlos y Dana, quienes hicieron de las noches de desvelo algo especial, casi de película.
⋆ A toda mi inmensa familia por su apoyo constante, en particular a mis abuelitos: Otilia, Miguel†, Feliciana† y Salvador, así como a mis primos Rafa, Yunuén, Mario, Argel y Grecia.
⋆ A mi hermano Trelew por su cariño, por compartir conmigo sus más profundos pensamientos y por confiar en mí de manera incondicional; a mi cuñada Sofía que ya es una nueva pieza en el rompecabezas de mi vida y a mi sobrino(a) por ser una chispita que enciende las llamas de la ilusión.
⋆ A mi novio Luis David, por compartir mis alegrías y contener mis lágrimas, por demostrarme que el mundo es algo mejor cuando hay personas buenas en él; en fin, por llegar a mi vida y cambiarla para siempre.

Mi agradecimiento más profundo es para mis padres, por ser ejemplo constante, por enseñarme que *"cuando te caes te levantas"*, por inculcar en mí el deseo insaciable por el conocimiento y sobre todo por su amor sin límites.

Indice

Capítulo 1

INTRODUCCIÓN

En la actualidad, las distintas disciplinas del conocimiento ya no se encuentran aisladas, por el contrario, éstas interactúan entre sí para contribuir de forma integral a la ciencia. En este sentido, la astronomía y las ciencias computacionales no son la excepción.

En los últimos años, el desarrollo científico y tecnológico ha conducido a una explosión en la cantidad de datos y mediciones. La astronomía es un claro ejemplo de esto, donde la gran cantidad de datos disponibles hacen que el trabajo necesario para hacer análisis efectivos, sea tan grande, que un análisis manual no resulta ni cercanamente práctico, ya que las conclusiones cualitativas se obtienen sólo analizando una cantidad enorme de datos, y en ocasiones, más de una vez. Aquí es donde intervienen principalmente las ciencias computacionales, las cuales están contribuyendo activamente en la tarea de efectuar análisis automáticos de grandes cantidades de datos astronómicos.

1.1 Determinación automática de parámetros astrofísicos

Las herramientas computacionales se han convertido en una parte fundamental en el apoyo del análisis e interpretación de datos astronómicos. En estas tareas, el aprendizaje automático ha jugado un papel relevante, permitiendo que los sistemas de análisis cuenten con la capacidad de generalización, la cual es muy valiosa cuando la cantidad de datos disponibles es muy grande y sólo es posible analizar manualmente una cantidad reducida de los mismos.

El análisis automático de datos astronómicos se ha desarrollado con cierto éxito, este tipo de análisis varía desde la clasificación morfológica de galaxias (ver [3, 12, 30]) hasta la determinación de parámetros estelares, donde el enfoque predominante consiste en ajustar un espectro sintético con parámetros estelares conocidos, a otro del cual se quiere conocer los suyos; para lograr esto se han empleado diversos métodos de aprendizaje automático, tales como redes neuronales artificiales (ver [9, 10, 15, 19, 40, 47, 48, 49]), regresión lineal localmente ponderada (ver [16]), y algoritmos genéticos (ver [35]), entre otros.

Este enfoque puede producir resultados erróneos cuando se intenta aplicar a ciertos espectros observacionales, debido principalmente a la presencia de fuertes líneas de emisión (ver figura 2.2), las cuales no se toman en cuenta para la elaboración de modelos de espectros

sintéticos.

Las líneas de emisión pueden ser mucho más fuertes que las líneas de absorción, o incluso estar en una magnitud mayor que el resto del espectro, dominando así su perfil; lo que ocasiona que al estandarizar el espectro sea muy diferente de lo que sería sin esas líneas y por lo tanto se ajuste a un espectro con características muy diferentes a las que en realidad presenta.

Las líneas de emisión no son el único problema que presentan los espectros observacionales, también está el ruido provocado por efectos atmosféricos o incluso por errores en el espectrógrafo, así como las distorsiones causadas por mediciones deficientes, entre otros factores.

Muchas de las diferencias que existen entre un espectro sintético y uno observacional se pueden modelar, con la excepción principal de las líneas de emisión; siendo ésta una alteración muy importante, y esta ha sido una limitante para que los sistemas de análisis automático sean capaces de usarse en datos reales, aunque paradójicamente se hayan construido con este fin.

1.2 Determinación automática de poblaciones estelares

Las estrellas son los constituyentes fundamentales de las galaxias, las estrellas evolucionan y por lo tanto las galaxias también. La evolución de las galaxias queda representada por el tipo de poblaciones estelares que las forman.

Una población estelar es un conjunto de estrellas que presentan grandes similitudes cinemáticas, químicas y edades. En términos generales, existen 2 tipos de poblaciones estelares tipo I y tipo II. Las poblaciones de tipo I están compuestas mayormente por estrellas muy jóvenes, aunque también hay estrellas viejas tienen presencia alta de elementos pesados, y su tasa de formación estelar alta. Las poblaciones tipo II están formadas por estrellas viejas, pocas estrellas jóvenes, pocos elementos pesados y su tasa de formación estelar es baja. A partir de estos dos tipos de poblaciones estelares con características tan diferentes, es posible establecer una población intermedia, la cual presente un equilibrio entre estrellas muy jóvenes y estrellas muy viejas, contando mayormente con estrellas de mediana edad. Por lo tanto para cuestiones prácticas una galaxia estaría formada sólo por tres principales poblaciones estelares.

Las comparaciones entre las observaciones y los modelos de poblaciones estelares en evolución, ayudan a entender el proceso de la formación de las estrellas. Se pueden observar galaxias con formación estelar activa, en muchas de ellas se pueden encontrar muestras de un proceso violento de formación de estrellas de gran masa, este fenómeno se llama 'starburst' o brote estelar. Las huellas de estos fenómenos se pueden encontrar en el espectro galáctico, así como el tiempo transcurrido desde su ocurrencia.

Determinar estos fenómenos a partir del espectro es muy importante para entender la formación y evolución de las galaxias, pero esta tarea no es trivial, ya que la información no está codificada de manera explícita, sino que hay que encontrar los patrones mediante los cuales sea posible identificar dichas propiedades, además los espectros galácticos están

normalmente contaminados por ruido u otros factores externos.

La determinación automática de las edades de poblaciones estelares aplicada a espectros observacionales, es un problema de investigación nuevo debido a la reciente disponibilidad de los datos, así como a la falta de soluciones eficientes; por lo tanto no se cuenta con trabajos que presenten resultados significativos para su comparación.

Sin embargo, existen dos enfoques que no usan aprendizaje automático para resolver este problema, los dos se basan en el ajuste de combinaciones de espectros sintéticos, pero presentan serias limitaciones y no es posible extenderlos.

- El primero consiste en ajustar el continuo de los espectros a otro sintético, pero este enfoque no toma en cuenta características que aportan indicios importantes acerca de la edad de las poblaciones, como son los perfiles de ciertas líneas de absorción.
- En el segundo enfoque, se hace la determinación de la edad por medio del ajuste de las líneas de la serie de Balmer y la línea del calcio (ver [44]).

En estos dos enfoques, se guardan en una base de datos todas las posibles combinaciones de espectros con ciertos parámetros, estas combinaciones se comparan contra los espectros de los que se quiere determinar sus parámetros, al final, se asignan los parámetros del espectro más parecido.

Las principales desventajas de estos trabajos radican mayormente en el tiempo que consumen, el cual es muy grande comparado contra el que puede producir un método de aprendizaje automático. Otra desventaja es que estos métodos no pueden generalizar a ejemplos no vistos, es decir, si una combinación no está en la base de datos, no es posible determinar una buena predicción.

Con un espacio de búsqueda tan grande como el que presentan las posibles soluciones para un espectro, los métodos de búsqueda exhaustiva no se pueden extender sin que el tiempo de búsqueda crezca exponencialmente.

Tomando en cuenta lo anterior, se puede establecer que hasta el momento no se ha desarrollado un método eficiente para determinar las edades de las principales poblaciones estelares de una galaxia real.

1.3 Objetivos

De acuerdo con lo anterior, los objetivos planteados para esta tesis son los siguientes:

Objetivo General.

Desarrollar un sistema de análisis de espectros, que sea capaz de determinar las edades de las tres principales poblaciones estelares de una galaxia, así como la proporción en la que cada población está presente.

Objetivos Particulares.

El sistema deberá ser:

- Rápido. El tiempo de predicción por espectro deberá ser pequeño. Entendiendo por *pequeño*, un tiempo no mayor a 10 días para analizar 100000 espectros.
- Preciso. El error de ajuste por espectro, deberá ser comparable al producido por una búsqueda exhaustiva.
- Capaz de manejar datos observacionales. El sistema se probará con un conjunto no clasificado de datos reales, su ajuste deberá ser tan bueno como el producido por una búsqueda exhaustiva y será capaz de manejar de manera eficaz el problema de las líneas de emisión.
- Insensible a efectos de ruido y atenuación.

1.4 Metodología de la solución propuesta

En este trabajo se propone un método de aprendizaje automático para determinar, a través del espectro, las principales poblaciones estelares que forman una galaxia. En la sección 7.2 se compara el desempeño obtenido mediante métodos de aprendizaje automático contra el de una búsqueda exhaustiva.

El método propuesto está pensado para ser insensible a efectos de ruido y atenuación (ver la sección 2.2.1.1), además de que será capaz de manejar datos observacionales y encontrar una buena predicción de sus parámetros. Para esto, el ajuste principal se realizará únicamente en los perfiles de ciertas líneas importantes y no en todo el espectro, por lo tanto se evitarán los problemas principales de las líneas de emisión.

La metodología general que se usó para la elaboración del trabajo contenido en esta tesis es la siguiente (ver figura 1.1):

1. Identificar la región del espectro que codifica la mayor cantidad de información acerca de las poblaciones estelares que forman una galaxia. Esta región deberá estar libre de la presencia de grandes líneas de emisión.

2. Codificar la región seleccionada del espectro, de tal manera que se reduzca la dimensión del espectro y se preserve la información relevante sobre las poblaciones estelares que lo integran.

3. Usar los atributos encontrados para entrenar métodos de aprendizaje automático que sean capaces de identificar las edades y proporciones de las poblaciones estelares.

4. Reconstruir los espectros predichos y compararlos con los espectros originales, para determinar la eficacia de los clasificadores; en el caso de espectros simulados, comparar los parámetros conocidos.

5. Hacer las estadísticas necesarias para encontrar patrones de comportamiento en los datos para mejorar los clasificadores.

6. Iterar este proceso a partir del punto 2, hasta encontrar una solución satisfactoria.

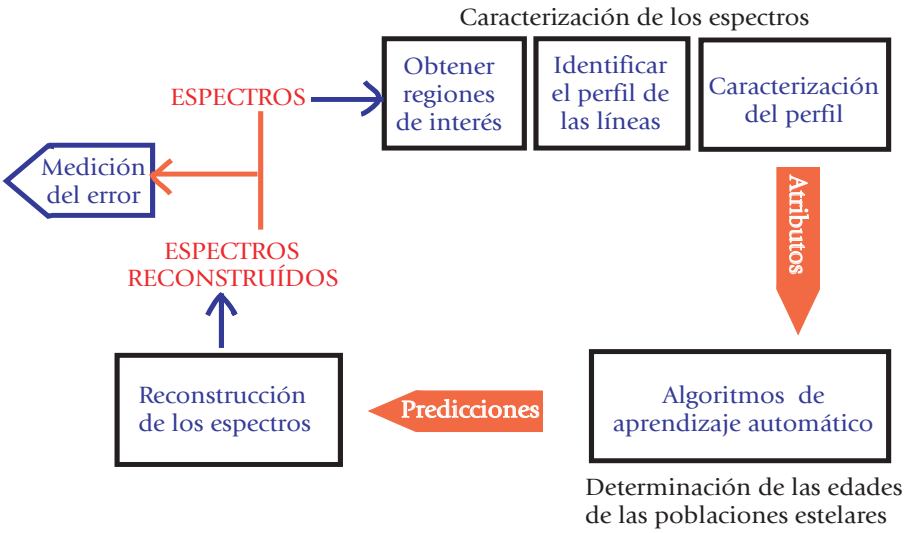


Figura 1.1: Esquema general de la metodología empleada

A partir de esta metodología, se usaron dos enfoques principales de aprendizaje automático para resolver el problema de clasificación; el primero consistió en evaluar tres tipos de ensambles, uno de redes neuronales artificiales, otro de regresión lineal localmente ponderada y el último una combinación de los dos anteriores. Mediante los resultados experimentales obtenidos en esta etapa se encontraron ciertas regularidades en los datos, lo que condujo a la creación del segundo enfoque, donde el algoritmo propuesto tiene forma de árbol, pero la decisión de la predicción no se hace únicamente en los nodos hoja, sino que se van haciendo predicciones parciales en cada nodo y se decide a partir del vector final de predicciones, aceptando la que presente el menor error entre el espectro original y el espectro reconstruido usando dicha predicción. Este método usa internamente otros algoritmos de aprendizaje automático supervisado y no supervisado.

Los resultados experimentales mostraron que el segundo enfoque obtuvo mejores ajustes de los perfiles de las líneas, así como mejores espectros reconstruidos. El tiempo de predicción no sobrepasó en ningún caso los 2.5 segundos por espectro, además la precisión obtenida se puede calificar como aceptable en el caso de espectros observacionales, ya que el error total es solamente 0.8% peor que el que produce una búsqueda exhaustiva.

1.5 Organización de la tesis

La estructura del resto de la tesis es como sigue:

En el **capítulo 2** se revisan los conceptos astrofísicos básicos y se presentan los datos usados, así como la manera de generarlos y sus propiedades.

En el **capítulo 3** se muestran los métodos de aprendizaje automático que se emplearon a lo largo del trabajo. Este capítulo se divide en cinco secciones principales. La primera sección está dedicada a presentar algunos conceptos fundamentales del aprendizaje automático. En la

segunda sección se habla sobre el aprendizaje no supervisado y sus principales características. En la tercera se trata el tema del aprendizaje supervisado con especial énfasis en redes neuronales artificiales y árboles de decisión. Por último se presentan los conceptos fundamentales sobre ensambles de clasificadores y la manera de hacerlos, así como una revisión de las métricas de error usadas en este trabajo.

El **capítulo 4** muestra la manera de obtener los atributos asociados a cada espectro, para con ellos, entrenar los clasificadores correspondientes. Se presenta el proceso usado para definir el perfil de ciertas líneas espectrales, donde se encuentra gran parte de la información relevante para identificar la edad de las poblaciones estelares; también se analizan dos maneras de caracterizar el perfil, una por medio de análisis de componentes principales y otra por medio de la extracción de anchos de las líneas.

En el **capítulo 5** se presenta el primer acercamiento a la solución general del problema por medio de técnicas de aprendizaje automático; aquí se usan ensambles de decisión compuestos por algoritmos que efectúan regresión múltiple. También se presentan los experimentos realizados usando datos simulados.

El **capítulo 6** comienza con un análisis sobre el comportamiento observado en los experimentos realizados con datos sintéticos y da pie para la creación del ensamble de decisión jerárquica, el cual presenta una estructura de árbol, con nodos que realizan clasificación y obtienen una partición de los datos para expandir el árbol. En este capítulo se presentan los experimentos realizados con datos reales y los problemas encontrados para realizar esta tarea.

En el **capítulo 7** se muestran los resultados obtenidos mediante los experimentos definidos en capítulos anteriores, se hace una comparación en cuanto a la eficiencia y precisión de los algoritmos, y se muestran algunos ejemplos de los ajustes realizados.

Por último, el **capítulo 8** está dedicado a presentar las conclusiones obtenidas durante la elaboración de este trabajo, así cómo las cuestiones que quedaron abiertas a nuevas investigaciones.

Capítulo 2

ANTECEDENTES

2.1 Conceptos previos

La luz no es otra cosa que radiación electromagnética. La fuente más importante de esta radiación es el movimiento acelerado de los electrones que forman parte de los átomos. La luz visible constituye una pequeña porción de la amplia familia de ondas electromagnéticas. Los diferentes nombres que se les atribuyen sólo son producto de una clasificación histórica, ya que solamente se diferencian en la longitud de onda y frecuencia. Sin embargo, esta diferencia es crucial a la hora de establecer sus propiedades, en particular, las que tienen que ver con la emisión y absorción de la radiación electromagnética.

La luz de las estrellas y galaxias, se descompone en su gama de colores, el resultado de esa descomposición se llama espectro.

Una galaxia es una agrupación de un gran número de estrellas y materia interestelar, cuya organización y mantenimiento como un todo tiene por causa las interacciones gravitacionales entre sus componentes.

De acuerdo a su morfología, las galaxias, en términos generales, se clasifican en: espirales, elípticas e irregulares. En las galaxias espirales se puede observar una estructura bien definida, estas galaxias están compuestas por bulbo, disco y halo.

El bulbo central, el cual se corresponde con el núcleo de la galaxia, generalmente es de forma esférica y tiene mayor concentración de estrellas amarillas más viejas. El disco está formado por una serie de brazos formados por gas polvo y en su mayoría por estrellas azules más jóvenes, las cuales giran alrededor del bulbo central. El halo es la parte exterior de la galaxia, éste envuelve a las dos estructuras anteriores, tiene menor cantidad de materia y en él se encuentran cúmulos de cientos de miles de estrellas. Las galaxias están formadas por cierta cantidad de poblaciones estelares.

Se llaman poblaciones estelares a los diferentes grupos de estrellas que forman la galaxia, estos grupos están compuestos por estrellas con similares características cinemáticas, edades y composiciones químicas.

Existen dos tipos de poblaciones estelares: tipo I y tipo II. Las de tipo I se encuentran únicamente en los discos galácticos y tienen una presencia relativamente alta de elementos pesados, además en las regiones donde hay presencia de este tipo de poblaciones la tasa de formación estelar es relativamente alta, debido a que en estas regiones es elevada la presencia de gas y polvo interestelar. Las poblaciones tipo I poseen mayormente estrellas jóvenes pero también presencia de viejas (ver [37]).

Por otro lado las poblaciones tipo II se ubican en los Bulbos y Halos galácticos formando los cúmulos globulares que son mayormente estrellas viejas y por lo tanto la presencia de elementos pesados es baja, entonces la tasa de formación estelar también es baja.

La manera más efectiva de estudiar las propiedades de los cuerpos celestes, específicamente, de una galaxia, es a través de la espectroscopía. La espectroscopía es una rama de la química que estudia las líneas espectrales emanadas desde moléculas y átomos diferentes, incluyendo la posición e intensidad de las líneas de emisión y absorción; esta disciplina es muy importante para determinar la composición de las estrellas, nubes interestelares o galaxias, a través de su espectro.

Se llama espectro a una representación parcial de la luminosidad en función de la longitud de onda, es decir, un espectro es la relación entre la intensidad de radiación y la longitud de onda emitida por un cuerpo celeste. La luminosidad o flujo, es la energía total emitida por un objeto por segundo.

La longitud de onda λ, es la distancia de cresta a cresta o de valle a valle de una onda electromagnética o de otro tipo. Las longitudes de onda están relacionadas con la frecuencia F: cuanto más larga la longitud de onda, más baja es la frecuencia. La longitud de onda y la frecuencia están relacionadas mediante la siguiente fórmula:

$$F = \frac{c}{\lambda} \tag{2.1}$$

Donde c es la velocidad de la luz.

La unidad de medida utilizada en astronomía para representar la longitud de onda de los espectros estelares son los Angstroms (abreviados por $\mathring{A}$) los cuales representan 1×10^{-13} de un metro. La resolución de un espectro se refiere a la distancia entre un punto muestreado y el siguiente, el rango espectral especifica la longitud de onda que cubre el espectro.

En un espectro estelar, se pueden observar ciertas líneas, las cuales están relacionadas con transiciones electrónicas de los átomos que forman la estrella, pudiendo ser líneas correspondientes a emisión o a absorción de energía en el tránsito electrónico. De este modo, cuando la radiación del interior de la estrella fluye hacia el exterior a través de la superficie estelar, los átomos de la región superficial absorben parte del flujo, y esto produce rayas negras de absorción en el espectro.

La forma de espectro más sencilla, llamada espectro continuo, es la emitida por un cuerpo sólido o líquido que puede ser llevado hasta altas temperaturas. Estos espectros no presentan

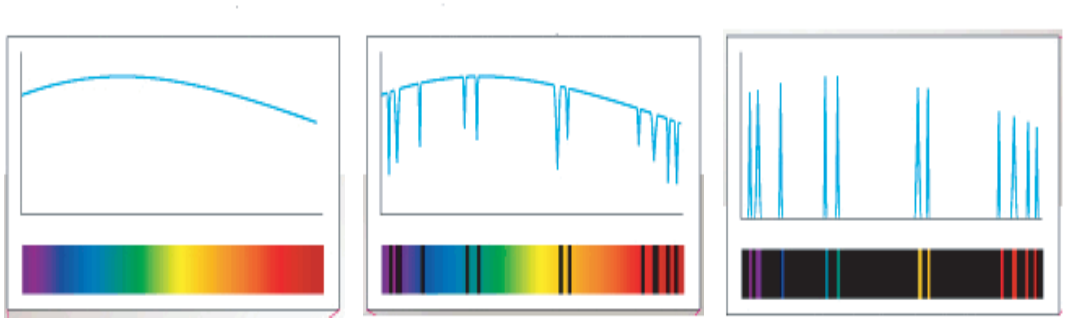
líneas porque contienen luz de todos los colores. Los espectros continuos sólo se pueden analizar con métodos espectrofotométricos.

Detallando lo anterior, se pueden identificar tres tipos de espectros:

1. Espectro continuo: aquí aparecen continuamente toda la variedad de colores, desde el rojo hasta el azul. Un espectro continuo lo producen los sólidos incandescentes, los líquidos y los gases que se encuentran sometidos a una gran presión (ver figura 2.1(a)).

2. Espectros de líneas de absorción: aquellos en que faltan en forma discreta, sobre un continuo, ciertos colores. El espectro de absorción o de rayas oscuras, se produce cuando una fuente emite radiación continua que pasa a través de un gas más frío (ver figura 2.1(b)).

3. Espectros de líneas de emisión: en los que aparecen en forma discreta solamente ciertos colores. El espectro de emisión o de rayas brillantes, se produce por gases enrarecidos en condiciones de temperatura muy alta. Cada gas tiene su espectro característico (ver figura 2.1(c)).



(a) Espectro continuo (b) Espectro en absorción (c) Espectro en emisión

Figura 2.1: Tipos de espectros

2.2 Espectros galácticos

Una galaxia está compuesta por estrellas de diferentes edades. La evolución galáctica suele ser diferente en términos específicos, de galaxia en galaxia; pero en términos generales, la evolución de una galaxia se da en ciertas fases. En este trabajo nos enfocaremos en galaxias que presentan tres fases principales de formación estelar, donde la primera corresponde al origen de la galaxia y está representada por una población vieja, la segunda y tercera fases corresponden a periodos de formación estelar masiva, estos periodos se representan con las poblaciones intermedia y joven respectivamente. Se asume que también existe una formación constante de estrellas durante el periodo de vida de la galaxia, pero que esta formación constante no influye de manera determinante en el espectro galáctico resultante.

Un espectro galáctico con periodos de formación estelar, se puede simular combinando distintos modelos sintéticos de evolución estelar, donde cada modelo representa cada uno de los diferentes brotes masivos de estrellas.

2.2.1 Propiedades particulares

Ninguna galaxia es exactamente igual a otra, muchas veces, galaxias que tienen muchas similitudes también presentan serias diferencias. El análisis de espectros galácticos, se hace más difícil debido precisamente a las grandes diferencias que pueden presentar ciertos tipos de galaxias; por ejemplo, las galaxias con núcleos con brotes de formación estelar (ver [27, 37]), que son regiones de las galaxias con líneas de emisión donde se están produciendo de manera súbita y durante periodos de tiempo relativamente cortos, nuevas estrellas; así como las llamadas galaxias activas, las cuales son muy luminosas y compactas. Esta luminosidad es producida por extraños procesos en el centro de las mismas. Existen varios tipos de galaxias activas.

Un tipo de galaxias activas, son las galaxias Seyfert. Estas galaxias tienen un núcleo muy brillante y compacto que produce un espectro continuo con fuertes líneas de emisión, pero con una débil emisión de ondas de radio. Algunas Seyfert tienen líneas de emisión estrechas y brillan más en el infrarrojo.

Las radio galaxias, son otro tipo de galaxias activas, ellas emiten grandes cantidades de energía en forma de ondas de radio. Tienen un componente central de estrellas, con chorros de energía que salen de él y se extiende por miles de años luz.

Los cuásares, objetos celestes muy antiguos y lejanos, son cientos o miles de veces más luminosos que una galaxia común, aunque toda esa energía se forma en un volumen de espacio muy pequeño. Los cuásares cuya luminosidad varía en meses, tienen un tamaño similar al sistema solar. Cada cuásar se encuentra adentro de una galaxia, sin embargo la galaxia anfitriona es muy difícil de observar por la luminosidad del objeto.

Algunos teóricos, consideran a todas las galaxias activas como instancias particulares de un mismo tipo de objetos y establecen que las diferencias entre sus características sólo se deben a la observación desde la tierra. Hasta el momento, no existe un solo enfoque aceptado para definir estas galaxias, ni ha sido posible analizar cantidades significativas de las mismas. Por lo tanto, la necesidad de determinar sus parámetros de manera automática, se hace aun más importante.

2.2.1.1 Características difíciles de modelar

Un modelo sintético de evolución estelar, no considera todos los posibles efectos a los que podría estar sujeto un objeto celeste, existen diversos factores que hacen que un modelo sintético no sea tan parecido al objeto real que supuestamente está modelando. A continuación se explican algunos de estos factores.

Líneas de emisión. Las líneas de emisión no están consideradas, ni en los modelos de síntesis estelar más avanzados; estas líneas pueden aparecer en diferentes posiciones dependiendo de la composición química del objeto que se analice. Por lo tanto, no existe una ley general que sea capaz de determinar en qué lugar aparecerá una linea de emisión en un objeto determinado, ni qué tanto se afecta una línea de absorción que presenta en su interior una línea de emisión.

Para analizar galaxias reales, es muy importante tener un método que sea capaz de lidiar con las líneas de emisión. Como se explicó anteriormente, existen muchos tipos de galaxias que presentan fuertes líneas de emisión, las cuales tienen propiedades particulares muy importantes para la astronomía.

Atenuación. Las ondas provenientes de un cuerpo celeste sufren una atenuación a lo largo de su camino hasta la tierra, esta atenuación está relacionada con la distancia a la que se encuentra el objeto celeste del observador. En general, la intensidad de radiación de una fuente luminosa va disminuyendo en función de $(1/d)^2$, donde d es la distancia a la que se encuentra la fuente. Esta ley de atenuación de la luminosidad aparente de una fuente de luz se conoce como ley del inverso del cuadrado.

Considerando que un espectro se representa como la relación que existe entre la intensidad de la radiación de un objeto celeste contra la longitud de onda, resulta lógico, de acuerdo a la ley del inverso del cuadrado, que cuanto más lejano esté un objeto, más débil será su espectro.

Ruido. Se considera ruido a cualquier perturbación no deseada en el espectro, este tipo de perturbaciones no proporcionan información alguna. En todos los espectros observacionales existe ruido en mayor o menor grado, pero la distribución del ruido en un espectro no es uniforme ya que éste puede ser ocasionado por diversos factores, las causas del ruido se pueden clasificar en dos categorías principales: externas e internas.

Los tipos externos pueden ser causados por efectos atmosféricos, aunque este tipo de ruido tiende a disminuir en longitudes de onda pequeñas; también puede ser originado por fuentes externas a la atmósfera, éste se genera principalmente en la Vía Láctea. Otro tipo de ruido externo es el ruido industrial que es causado por el hombre, principalmente en zonas urbanas.

El ruido interno es aquel que se produce en el interior de los dispositivos, en este caso, de los espectrógrafos. A este ruido también se le conoce como ruido blanco y es causado por el movimiento aleatorio de los electrones libres dentro de un conductor, este tipo de ruido aumenta con la temperatura.

Resumiendo lo anterior, el ruido causado por efectos externos y el ruido de lectura del espectrógrafo, pueden causar serias distorsiones en un espectro observacional, de tal forma que el espectro obtenido sea ligera o severamente distinto del espectro limpio asociado con algún cuerpo celeste.

2.2.2 Regiones de interés

Una parte fundamental del planteamiento del problema, consiste en determinar las edades de poblaciones estelares en espectros observacionales, esta tarea no es trivial, porque como ya se mencionó en las secciones 2.2.1 y 2.2.1.1, los espectros reales presentan entre otras características difíciles de modelar, grandes líneas de emisión; estas líneas aparecen en diversos lugares del espectro, muchas incluso aparecen adentro de líneas de absorción. La longitud de onda donde aparecen y su comportamiento, no es predecible de manera general. Por lo tanto, si se intentara hacer un ajuste del espectro completo, estas líneas pueden ocasionar que el perfil espectral se modifique significativamente, impidiendo la correcta determinación de sus parámetros.

Un enfoque alterno al ajuste del espectro completo, consiste en ajustar sólo ciertas regiones del mismo, la restricción más importante establece, que la información necesaria para determinar los parámetros que se buscan debe de estar contenida de alguna manera en las sub-regiones que se van a ajustar.

En este caso particular, se buscan regiones que codifican la información relativa a la edad de las poblaciones estelares más importantes que conforman el espectro. Según la investigación de ciertos astrónomos (ver [17]), gran parte de la información de la edad de una galaxia está presente en la serie del hidrógeno de Balmer ($[3743\mathring{A}, 3757\mathring{A}], [3763\mathring{A}, 3781\mathring{A}], [3785\mathring{A}, 3812\mathring{A}], [3818\mathring{A}, 3854\mathring{A}]$), y se puede determinar por el tamaño de los anchos y los perfiles de estas líneas.

Todas las líneas de Balmer tienen comportamientos similares; durante los primeros 4 millones de años de una galaxia, los anchos equivalentes de las líneas no cambian significativamente, después de eso, los anchos crecen continuamente hasta los 500 millones de años, después de este periodo, comienzan a decrecer. El comportamiento en la fuerza de las líneas tiende a ser similar, ya que esta aumenta hasta los 500 millones de años, después de eso permanece constante. Gracias a este comportamiento, las líneas de la serie de Balmer son muy útiles para determinar la edad de poblaciones menores a 500 millones de años, es decir, de las poblaciones jóvenes e intermedias principalmente.

Otra línea cuyo perfil se ve modificado por los efectos de la edad es la línea de calcio, la cual es más útil para reconocer poblaciones viejas, esta línea se encuentra entre los $3909\mathring{A}$ y los $3958\mathring{A}$.

En la figura 2.2 se pueden apreciar las líneas que forman la serie de Balmer y la línea de calcio, en un modelo sintético y en uno observacional.

Las principales ventajas de determinar la edad de las poblaciones estelares usando como referente sólo unas cuantas líneas son las siguientes:

- La cantidad de datos se reduce dramáticamente.
- Es posible manejar datos reales que presentan grandes líneas de emisión.
- Se pueden ajustar espectros cuyo rango espectral sea reducido o incluso incompleto.

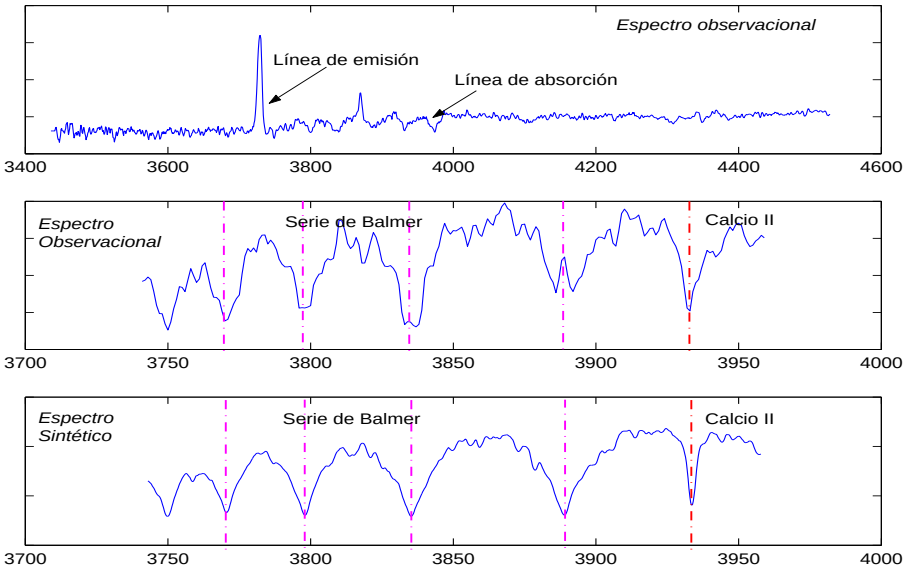


Figura 2.2: Regiones de interés: 1.-Espectro observacional, 2,3.-Espectro observacional y espectro sintético en la región de interés $[3743\mathring{A} - 3958\mathring{A}]$ mostrando las líneas de la serie de Balmer y la línea K de calcio

Las desventajas:

- Se pierde la información acerca del comportamiento global del espectro.
- No es posible ajustar datos donde el muestreo sea mayor a $2\mathring{A}$ por que la información importante se pierde entre un punto y otro.

2.3 Espectros simulados

Como ya se ha mencionado anteriormente, en astronomía es deseable poder analizar grandes cantidades de objetos celestes, pero no siempre es posible que astrónomos expertos determinen los parámetros astrofísicos de dichos objetos o de sus espectros. Por otra parte, tampoco es práctico hacer el análisis de espectros observados comparándolos contra otros espectros reales, en primer lugar porque no se dispone de la cantidad suficiente de espectros analizados, ni se tiene la certeza de que el análisis y los parámetros que representan a un espectro real sean los correctos. Por estas y otras razones, los astrofísicos han desarrollado modelos sintéticos de estrellas y galaxias. Estos modelos representan objetos celestes con ciertas características particulares y hacen que el estudio de otros objetos similares sea más fácil y exacto.

Usando espectros simulados, los astrónomos conocen de antemano los parámetros físicos de cierto objeto (sintético) y pueden comparar el comportamiento y propiedades específicas del mismo contra las propiedades que pudiera presentar algún otro objeto observacional; de esta manera, determinar sus posibles características, sin necesidad de contar con otro tipo de información especializada. Esto es, los espectros sintéticos sirven para hacer predicciones partiendo del objeto observacional y comparando con el simulado. También se pueden hacer

predicciones en el sentido inverso, es decir, se formula una teoría y se predicen parámetros que podría tener un objeto en condiciones particulares, entonces se hace la simulación con los parámetros obtenidos y se comparan las predicciones con la simulación del mismo.

La manera tradicional de hacer un espectro simulado es, mediante la combinación de modelos sintéticos que representan las propiedades físicas que se espera estén presentes en el espectro simulado resultante, también se le puede añadir ruido u otros defectos para volverlo tan realista cómo sea posible.

2.3.1 Modelos sintéticos

Los modelos sintéticos son aproximaciones a espectros con ciertas características, y sirven cómo base para formar los espectros simulados, estos modelos se hacen basándose en las leyes físicas que regulan la formación y evolución estelar.

Los modelos sintéticos ($\overrightarrow{m}_{edadx}$) de evolución estelar, usados en este trabajo, corresponden a un subconjunto de los modelos de síntesis espectral creados por Creviño et al. [7].

Estos modelos representan galaxias donde la formación estelar ocurre de dos maneras: formación constante de estrellas durante todo el periodo de tiempo que abarca el espectro y formación estelar abrupta en cierto periodo de tiempo, la edad de esta última formación estelar es la que determina la edad que representa el modelo.

En total, se usaron 14 espectros de alta resolución muestreados a $0.3\mathring{A}$ en un rango espectral de $3000\mathring{A}$ a $6900\mathring{A}$. Estos modelos están divididos en tres grupos de acuerdo a su edad:

- P_{joven}: Una población muy joven con edades entre 3 y 80 millones de años, la cual representa la componente de formación estelar, cinco espectros.
- P_{inter}: Una población intermedia con edades entre 100 y 800 millones de años, cuatro espectros.
- P_{vieja}: Una población muy vieja con edades entre $1,000$ y $9,999$ millones de años, cinco espectros.

2.3.2 Generación de espectros simulados

El espectro simulado, $\overrightarrow{S}sim$, es una combinación de tres modelos sintéticos que representa una galaxia con tres principales poblaciones estelares de distintas edades.

Cada espectro simulado está constituido por dos vectores, el primero, representa la longitud de onda y el segundo es la intensidad del flujo que incide en esa longitud de onda por segundo.

La manera de generar cada espectro simulado consiste en sumar tres modelos sintéticos de diferentes edades, donde cada uno de ellos ha sido multiplicado por una cierta proporción; de tal manera que la suma de las tres proporciones sea igual a 1. De esta forma, el espectro simulado resultante contiene la información de tres poblaciones estelares y sus respectivas proporciones. La generación de espectros simulados se hace como sigue. Definamos:

$M_{modelos}$ la matriz que contiene los 14 modelos sintéticos:

$$M_{modelos} \in \mathrm{R}^{14\mathrm{x}717} = \begin{pmatrix} \overrightarrow{m}_{edad1} \\ \overrightarrow{m}_{edad2} \\ \vdots \\ \overrightarrow{m}_{edad14} \end{pmatrix} \tag{2.2}$$

$\overrightarrow{p_1} \in \mathrm{R}^{1\mathrm{x}5}$, $\overrightarrow{p_2} \in \mathrm{R}^{1\mathrm{x}4}$ y $\overrightarrow{p_3} \in \mathrm{R}^{1\mathrm{x}5}$ son vectores que representan las proporciones de las poblaciones estelares: P_{joven}, P_{inter} y P_{vieja} respectivamente, cada uno con sólo un elemento mayor a cero, que representa una edad específica del espectro simulado, donde:

$$\sum_{i=1}^{3} \sum \overrightarrow{p_i} = 1 \tag{2.3}$$

$\overrightarrow{edades}$ es el vector que resulta de concatenar $\overrightarrow{p_1}$, $\overrightarrow{p_2}$ y $\overrightarrow{p_3}$. Éste contiene la proporción de edades, con sólo tres elementos mayores a cero:

$$\overrightarrow{edades} \in \mathrm{R}^{1\mathrm{x}14} = [\overrightarrow{p_1}, \overrightarrow{p_2}, \overrightarrow{p_3}] \tag{2.4}$$

Entonces $\overrightarrow{S}sim$ el espectro simulado se construye cómo sigue:

$$\overrightarrow{S}sim = smooth(\overrightarrow{edades} * M_{modelos}) + \overrightarrow{S}_{ruido} \tag{2.5}$$

Donde $\overrightarrow{S}_{ruido} \in \mathrm{R}^{1\mathrm{x}717}$ es un vector aleatorio del mismo tamaño que $\overrightarrow{S}sim$ el cual representa ruido gaussiano. La función *smooth* se usa para atenuar el espectro, simulando la resolución deficiente que pueden presentar las observaciones.

Debido a que todos los espectros simulados se generaron en la misma longitud de onda, por simplicidad cada uno se representa por un sólo vector que contiene la intensidad del flujo. Cada espectro galáctico simulado se define por $\overrightarrow{edades}$. La tarea consiste en predecir los elementos diferentes de cero que mejor ajusten el perfil de las líneas y además produzcan un espectro resultante tan parecido al observacional como sea posible. En la figura 2.3 se muestra un espectro simulado mediante la combinación de tres modelos sintéticos.

Se decidió generar los espectros simulados usando 3 poblaciones, debido principalmente a que como se simula un espectro por medio de una combinación lineal, y a que algunos modelos sintéticos son casi combinaciones lineales unos de otros, más de 3 espectros hacen que sea muy difícil diferenciar una edad de otra. Además, para efectos prácticos, un astrónomo sólo necesita conocer las dos o tres poblaciones estelares dominantes en una galaxia.

Otro punto que se definió fue el número de modelos sintéticos usados. En este trabajo se usaron 14 modelos con edades espaciadas lo más uniformemente posible. Se utilizaron únicamente 14 de los 44 disponibles, porque con 14 se alcanza a cubrir el rango de las edades más significativas y además, porque los modelos de edades cercanas presentan diferencias mínimas, las cuales podrían parecer incluso ruido en un espectro normalizado.

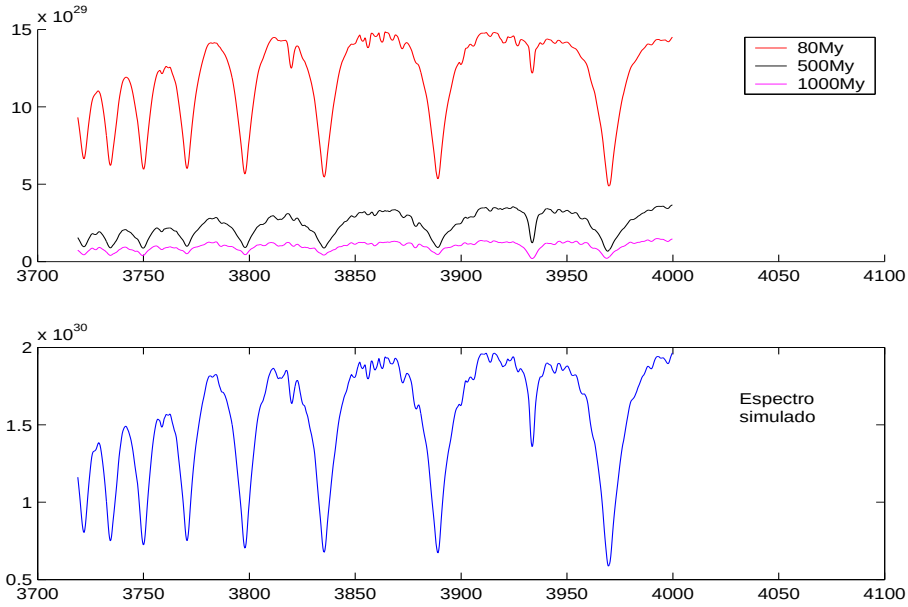


Figura 2.3: Generación de un espectro simulado, mediante la combinación de 3 modelos sintéticos

2.3.2.1 Efectos adicionales

Con el objetivo de volver a los espectros simulados, tan realistas como fuera posible, se les añadieron ciertos efectos. A continuación se explican.

Atenuación. La función *smooth* (ver la sección 2.3.2) realiza una convolución por medio del método de ventaneo gaussiano, con el fin de suavizar el espectro, para simular la atenuación que podría estar presente debido tanto a la distancia lejana a la que se encuentra, así cómo a una deficiente resolución del espectrógrafo usado para obtenerlo.

smooth usa un filtro gaussiano $g(t)$ con desviación estándar $\sigma = 0.70$ y tamaño de la ventana $5 \leq \mu \leq 11$ para aplicar una convolución al vector que contiene el espectro $f(t)$.

$$\begin{aligned} g(t) &= e^{\frac{1}{2}\left(\frac{t}{\sigma}\right)^2} \\ -\mu/2 &\leq t \leq \mu/2 \end{aligned} \tag{2.6}$$

En términos estrictos, la convolución se hace de la siguiente manera:

$$h(t) \equiv \int_{-\infty}^{\infty} f(u)g(t-u)du \tag{2.7}$$

Pero como se tiene una secuencia de puntos, es decir, una serie discreta de valores, la convolución se realiza usando:

$$h_i = \sum_{j=0}^{m} f_i g_{i-j} \tag{2.8}$$

Donde m es el número total de puntos de la función f y h_i es el resultado de la convolución

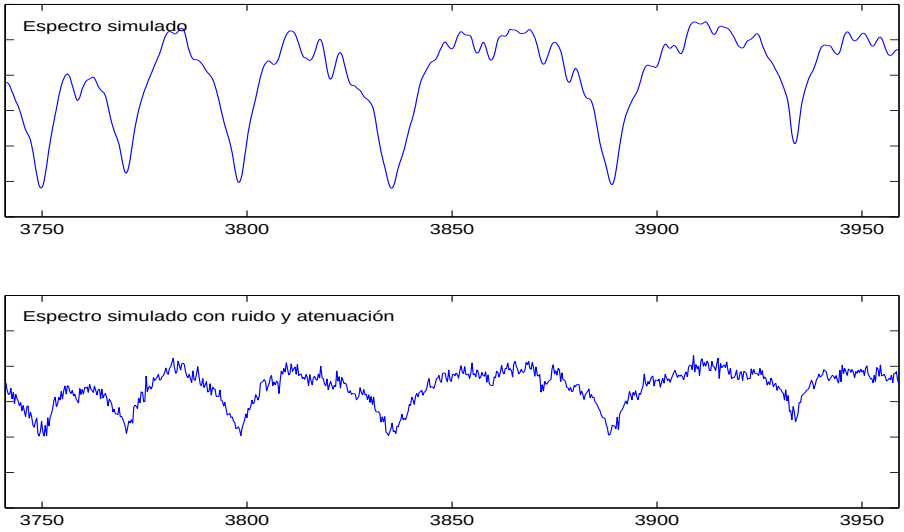
en el punto i. Este proceso produce un espectro con sus perfiles suavizados.

Ruido gaussiano. El ruido gaussiano se simula, sumando un vector aleatorio con media 0 y desviación estándar 0.1, del mismo tamaño que el espectro simulado, de esta manera, los puntos originales del espectro quedan desplazados de su valor original en un 10% aproximadamente.

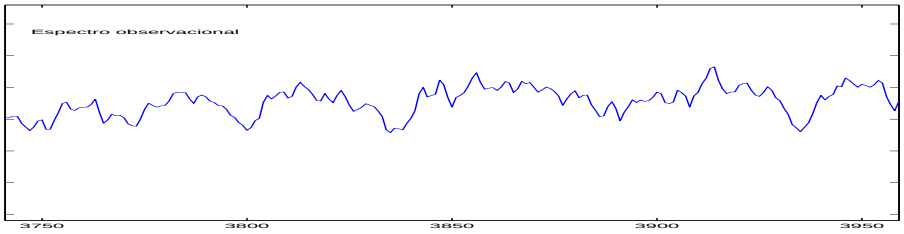
$$\overrightarrow{S}_{ruido} \in \mathrm{R}^{1\times m} = random(1, m) * 0.1 \tag{2.9}$$

Donde $random(1, m)$ es una función que produce un vector aleatorio, con media 0 y desviación estándar 1, de tamaño $1 \times m$.

Un espectro simulado, el cual ha sido suavizado con la función *smooth* y se le ha aplicado un 12% de ruido gaussiano, se puede ver en la figura 2.4, comparado contra un espectro limpio.



(a) Espectro simulado limpio y espectro con ruido y atenuación



(b) Espectro observacional en la misma región

Figura 2.4: Comparación de un espectro simulado contra uno observacional

2.3.2.2 Consideraciones

Actualmente, los modelos sintéticos de evolución estelar están en un proceso de desarrollo, por lo tanto, no se pueden considerar perfectos, por el contrario, con el fin de contar con modelos generales, muchas veces no toman en cuenta factores importantes que están presentes en objetos observacionales. Sin embargo, es posible simular algunos efectos que permiten que un

espectro sintético llegue a tener características tan realistas como las encontradas en espectros observacionales.

En la figura 2.4 se puede apreciar cómo se transforma un espectro simulado limpio después de añadirle efectos que simulan ruido y atenuación. El espectro simulado transformado tiene mayor parecido con un espectro observacional, aunque a diferente resolución, que con el espectro que le dio origen.

Es válido ajustar espectros reales con sintéticos aunque el estado de la ciencia en cuanto a modelos sintéticos no esté en el nivel más deseable, ya que estos esfuerzos hacen posible que los astrónomos puedan llegar a encontrar nuevos datos que permitan perfeccionar los modelos existentes.

Capítulo 3

MÉTODOS DE APRENDIZAJE AUTOMÁTICO

3.1 Introducción al aprendizaje automático

El aprendizaje automático, particularmente el aprendizaje automático inductivo, se dedica a crear, de manera automática, modelos generales a partir de conjuntos de datos específicos. Así, a partir de los modelos creados, es posible extraer el conocimiento que se encuentra de manera implícita en los datos, de tal forma que se pueda hacer una evaluación y predicción sobre datos nuevos.

Las técnicas de aprendizaje automático se han empleado frecuentemente en problemas donde es común que existan grandes cantidades de datos, los cuales no tienen modelos teóricos satisfactorios asociados, y que además no resulta práctico analizarlos de forma manual.

Existen dos conceptos importantes que sirven para identificar la eficacia de un algoritmo de aprendizaje automático

Generalización. La generalización es la habilidad que tiene un clasificador de asignar una predicción correcta a instancias nuevas.

Sobreajuste. El sobreajuste sucede cuando un clasificador aprende a asignar predicciones correctas únicamente a los ejemplos con los que fue entrenado y por lo tanto, presenta un grado de generalización pobre.

En el aprendizaje automático se pueden identificar dos tipos de problemas principales, estos son: problemas de clasificación y problemas de regresión. A continuación se explican brevemente.

Clasificación. Un problema de clasificación, es aquel donde las instancias están organizadas en cierto número de grupos o clases, y se busca predecir a qué grupo corresponde una serie de instancias nuevas. En éste sentido, un algoritmo entrenado como clasificador tendrá un conjunto discreto, cerrado, de posibles salidas en la predicción.

Aunque el problema a resolver en éste trabajo, se plantea más naturalmente como de regresión múltiple, también se puede visualizar como dos problemas separados, uno de regresión y otro de clasificación, donde se tienen tres sub problemas de clasificación, y cada uno de ellos corresponde a un conjunto de edades. Éste enfoque se usó en los experimentos que se describen en la sección 5.2.1.

Regresión. Cuando el problema consiste en aproximar una función, se dice que se tiene un problema de regresión. En éste caso la salida de un algoritmo que se entrene para predecir una función, tendrá como salida un número real para cada instancia. Si se trata de un problema de regresión multiple, la salida será un vector de números reales.

En éste trabajo, la manera más eficaz de manejar la tarea a resolver, es como un problema de regresión múltiple, en la cual es necesario obtener la predicción de 14 vectores de números reales, donde sólo tres de ellos tendrán valores mayores a cero simultáneamente.

3.2 Aprendizaje no supervisado

El aprendizaje no supervisado es una de las ramas del aprendizaje automático, donde el algoritmo por sí mismo es capaz de encontrar patrones y de identificar cuales instancias están más cercanas a cada tipo de patrón.

En el aprendizaje no supervisado, no es necesario que las instancias en el entrenamiento estén etiquetadas, es decir, no es necesario que ya tengan una clase o un grupo asignado, los algoritmos se encargan de asignarles la clase. Generalmente las clases se forman de acuerdo a las distancias que existen entre las instancias, colocando aquellas que tengan una distancia pequeña en la misma clase.

Algunos de los algoritmos de aprendizaje no supervisado más populares son los mapas auto-organizados o de Kohonen [25], k-means y fuzzy k-means, entre otros (ver [53]).

Los métodos de aprendizaje no supervisado se utilizan mayormente y con buen éxito en la clasificación de textos.

3.2.1 K-means

Uno de los algoritmos de aprendizaje no supervisado, más usado en diversas aplicaciones, es K-means; éste es un método de agrupamiento no jerárquico (ver [2, 4, 31]), el cual, a partir de un conjunto de instancias no clasificadas, puede encontrar los k grupos donde mejor se organizan todos los ejemplos.

Inicialmente toma k instancias $(c_1, c_2, \ldots c_k)$ del conjunto de datos X, donde k es el número fijo de grupos requeridos; a estas instancias se les llama prototipos, y cada una de ellas se asume como el centroide inicial de cada uno de los k grupos.

El objetivo principal de k-means es obtener k grupos disjuntos, donde los elementos de cada grupo, sean tan parecidos entre si, y tan diferentes a los elementos de los otros grupos como sea posible. Para lograr esto, el algoritmo intenta minimizar el error dado por E:

$$E = \sum_{j=1}^{k} \sum_{x_i \in C_j} \mid x_i - c_j \mid^2 \tag{3.1}$$

para

$$x_i \in X$$
$$i = \{1, \ldots, n\}$$

Donde C_j es el $j - esimo$ grupo, c_j es su centroide y n es el número total de instancias x_i en el conjunto de datos.

A continuación, se examina cada instancia en el conjunto de datos y se asigna a un grupo dependiendo de la distancia mínima que presente a los centroides.

La posición de los centroides se vuelve a calcular después de que se ha hecho el agrupamiento para todos los elementos, usando:

$$c_j = \sum_{x_i \in Cj} \frac{x_i}{|C_j|} \tag{3.2}$$

Este proceso se repite hasta que el error E no cambie significativamente, o hasta que la pertenencia de los grupos se mantenga muy similar.

La elección de un k adecuado depende del problema, y repercute significativamente en los resultados que produce el algoritmo. Generalmente el usuario hace pruebas con diferentes valores, hasta que se obtienen los resultados deseados; aunque también existen maneras automáticas de encontrar un número óptimo de grupos, por ejemplo en el trabajo presentado por Davies y Bouldin [11], donde se promedian las medidas de similitud entre cada grupo y su grupo más cercano para obtener un conjunto de índices relacionados a los diferentes valores de k.

3.3 Aprendizaje supervisado

El aprendizaje supervisado es aquel donde el clasificador cuenta con una guía externa sobre el número de patrones a identificar, así como ejemplos que presentan dichos patrones. Es decir, en el aprendizaje supervisado, los algoritmos aprenden a identificar las diferencias que presentan instancias de diferentes clases y a reconocer las similitudes de ejemplos en la misma clase.

Los algoritmos típicos del aprendizaje automático supervisado son: las redes neuronales artificiales, algoritmos genéticos, regresión lineal localmente ponderada, árboles de decisión y métodos estadísticos, entre otros. En las siguientes subsecciones se explicará brevemente el funcionamiento de algunos de estos algoritmos.

3.3.1 Regresión lineal localmente ponderada (LWR)

El método de Regresión Lineal Localmente Ponderada o (LWR) por sus siglas en inglés, es un algoritmo perteneciente a la familia de métodos de aprendizaje automático basados en instancias, los cuales, simplemente almacenan los ejemplos de entrenamiento hasta que es necesario clasificar un ejemplo nuevo, es entonces cuando construyen un modelo individual para cada nuevo ejemplo, los puntos se ponderan de acuerdo a su proximidad al punto de interés y se realiza una regresión sobre los más cercanos para determinar la clasificación final.

Para ajustar la función $f(x)$ en el punto explícito x_q, LWR construye una aproximación local a f en una región que rodea a x_q. Esta aproximación se usa para calcular el valor de $\widehat{f}(x_q)$ que es la salida estimada para el punto x_q; no es necesario almacenar la descripción de $\widehat{f}$, ya que para un nuevo ejemplo a ser clasificado, se construirá una nueva descripción $\widehat{f}$ que esté definida por el entorno local del nuevo punto.

Regresión lineal localmente ponderada, ajusta la función objetivo f cercana a x_q usando una función lineal de la forma:

$$\widehat{f}(x) = w_0 + w_1 a_1(x) + ... + w_n a_n(x) \tag{3.3}$$

Donde $a_i(x)$ denota el valor del $i - esimo$ atributo de la instancia x. En este caso, el objetivo es realizar una aproximación local de la función objetivo; esto se puede lograr de tres maneras:

1. Minimizando el error cuadrático sólo en k vecinos, donde kNN son los k vecinos más cercanos de x_q:
$$E_1(x_q) \equiv \frac{1}{2} \sum_{x \in kNN} (f(x) - \widehat{f}(x))^2 \tag{3.4}$$

2. Minimizar el error cuadrático en el conjunto total de ejemplos de entrenamiento D, ponderando el error de cada ejemplo de entrenamiento usando una función K de acuerdo a su distancia a x_q:
$$E_2(x_q) \equiv \frac{1}{2} \sum_{x \in D} (f(x) - \widehat{f}(x))^2 \mathrm{K}(d(x_q, x)) \tag{3.5}$$

3. Combinando 1 y 2
$$E_3(x_q) \equiv \frac{1}{2} \sum_{x \in kNN} (f(x) - \widehat{f}(x))^2 \mathrm{K}(d(x_q, x)) \tag{3.6}$$

LWR es un algoritmo muy eficaz en ciertos problemas, y presenta la ventaja intrínseca de los algoritmos de aprendizaje basados en instancias, de que puede modificarse continuamente el conjunto de entrenamiento sin impactar su eficiencia. Una de sus desventajas principales radica en que, si los elementos del conjunto de entrenamiento son demasiado parecidos entre sí, este algoritmo podría no hacer los mejores modelos locales de los datos.

3.3.2 Redes neuronales artificiales (ANN)

Las redes neuronales artificiales, son métodos del aprendizaje automático que constituyen un enfoque robusto para aproximar funciones discretas o continuas, es decir, para realizar clasificación o regresión de manera eficiente (ver [29, 36]).

Las redes neuronales se emplean frecuentemente en problemas donde los datos contienen ruido, donde se tolera que el tiempo de entrenamiento sea grande y en problemas donde no sea importante que un humano interprete el conocimiento almacenado en el modelo.

Se dice que las redes neuronales están inspiradas parcialmente por los modelos biológicos de adquisición del conocimiento, ya que se puede establecer una analogía con el sistema neurológico, donde a grandes rasgos, se tiene un conjunto de neuronas interconectadas que actúan de cierta manera dependiendo de los estímulos que reciban, y las redes neuronales artificiales están constituidas por un conjunto de unidades interconectadas, donde cada unidad recibe como entrada, un conjunto de valores reales y genera un solo valor real como salida, el comportamiento de cada unidad se modifica de tal forma que el comportamiento grupal produzca el tipo de salida que mejor represente cierto patrón de entrada.

Se asemejan al comportamiento cerebral, en dos aspectos principales:

- El conocimiento de la red se adquiere a través de un proceso de aprendizaje.
- Los pesos que interconectan las neuronas, conocidos como pesos sinápticos, se usan para almacenar el conocimiento adquirido.

Entonces, una red neuronal artificial, es un modelo matemático compuesto por un gran número de elementos procesadores altamente interconectados, los cuales trabajan de manera conjunta para resolver problemas específicos. Se pueden describir como un grafo dirigido (ver [51]), en donde cada nodo i ejecuta una función de transferencia usando:

$$y_i = f_i \left(\sum_{j=1}^{m} w_{ij} x_j - \theta_i \right) \tag{3.7}$$

Donde y_i es la salida para el nodo i, x_j es la $j - esima$ entrada del nodo, w_{ij} es el peso de la conexión entre los nodos i j y θ_i es el umbral definido para ese nodo.

3.3.2.1 Redes neuronales feedforward-backpropagation

Estas redes permiten un poder de representación muy amplio, ya que incluyen funciones que no son linealmente separables y alcanzan buenos resultados con funciones muy complejas. Estas redes están compuestas por unidades interconectadas que reciben como entrada un vector de valores reales y regresan como salida un valor real, típicamente entre cero y uno.

El poder de representación de las redes neuronales consiste en la combinación lineal que realizan usando los valores de sus entradas con ciertos valores llamados pesos que se corresponden directamente con cada una de las unidades, estos pesos se van modificando con el

entrenamiento y sirven para determinar la contribución que representa cada combinación de unidades adyacentes para lograr cierto resultado.

Las redes neuronales feedforward, o de alimentación hacia adelante, se llaman así porque los valores que entran a la red recorren un camino secuencial desde la primera capa hasta la de salida, donde ayudan a producir una predicción determinada. Se dice que estas redes están entrenadas usando el algoritmo de backpropagation o de retropropagación del error, debido a que la corrección de pesos se hace comparando la predicción obtenida para una instancia en la capa de salida contra el valor correcto, modificando los pesos de las conexiones pertinentes hasta la capa de entrada.

La unidad que conforma típicamente a las redes feedforward se llama unidad sigmoidal, debido a que generalmente utiliza una función sigmoidal o logística para obtener su salida. En nuestro caso, la función utilizada es:

$$y_i = \sigma(net) = \frac{1}{1 + e^{-net}} \tag{3.8}$$

donde

$$net = \overrightarrow{w} \cdot \overrightarrow{x} \tag{3.9}$$

$\overrightarrow{w}$ representa el vector de pesos de las conexiones, y $\overrightarrow{x}$ representa los valores de las entradas. La unidad sigmoidal calcula una combinación lineal de sus entradas con los pesos de sus respectivas conexiones y después le aplica un umbral al resultado. La salida de cada unidad es un número real entre 0 y 1.

El algoritmo de backpropagation. El algoritmo de backpropagation sirve para aprender los pesos en una red multicapa, dada una red con un número fijo de unidades e interconexiones. Este algoritmo utiliza el gradiente descendente para minimizar el error cuadrático entre los valores objetivo de la red y los que produce. En este caso, el error se define como:

$$E(\overrightarrow{w}) \equiv \frac{1}{2} \sum_{d \in D} \sum_{k \in out} (t_{kd} - o_{kd})^2 \tag{3.10}$$

Donde *out* es el conjunto de unidades de salida de la red, t_{kd} y o_{kd} son los valores objetivo y la predicción que produce la red respectivamente, asociados con la $k - esima$ unidad de salida para el ejemplo d.

El algoritmo de backpropagation tiene que buscar en un gran espacio de hipótesis para definir los valores de los pesos en cada unidad de la red. Esto trae consigo la desventaja principal de que el algoritmo de backpropagation converge en la mayoría de los casos sólo a mínimos locales; en la práctica, aunque esto resulta cierto, el algoritmo de backpropagation obtiene muy buenos resultados para problemas del mundo real. El algoritmo de backpropagation se detalla en la tabla 3.1.

Este algoritmo comienza construyendo una red con el número deseado de unidades de entrada, ocultas y de salida. Dada esta estructura de la red, el ciclo principal del algoritmo

itera, calculando el error para cada ejemplo de entrenamiento, calcula el gradiente con respecto al error de ese ejemplo y actualiza los pesos de la red.

El criterio de paro del algoritmo se puede establecer tras un número fijo de iteraciones, también puede ser cuando el error de los ejemplos de entrenamiento sea menor a un cierto umbral.

```
Crear una red feedforward con n_in unidades de entrada,
n_hid unidades ocultas y n_out unidades de salida.

Inicializar todos los pesos de la red a numeros aleatorios pequeños
Hasta que se alcance la condición de paro, hacer:
    Para cada < x, t > en el conjunto de entrenamiento, hacer:
        Ingresar la instancia x a la red
        Calcular la salida o_u para cada unidad u en la red.
        Para cada unidad de salida k en la red, calcular su error δ_k
          δ_k ← o_k(1 − o_k)(t_k − o_k)
        Para cada unidad oculta h calcular su error δ_h
          δ_h ← o_h(1 − o_h) Σ_{k∈outputs} w_kh δ_k
        Actualizar el peso w_ji en la red
          w_ji ← w_ji + △w_ji
          donde
          △w_ji = ησ_j x_ji
    termina
termina
```

x_{ji} denota la entrada del nodo i a la unidad j, w_{ji} es su peso correspondiente y η es la tasa de aprendizaje.

Tabla 3.1: Algoritmo de backpropagation

Las redes neuronales feedforward pueden representar diversos tipos de funciones con un error muy pequeño, usando por lo menos dos capas de unidades, entre las funciones que pueden ajustar están las funciones booleanas, funciones continuas; funciones arbitrarias también se pueden ajustar usando por lo menos tres capas de unidades.

3.3.3 Árboles de decisión

El aprendizaje con árboles de decisión es un método para aproximar funciones donde la relación importante entre los valores de los atributos y la información de la clase está representada como un árbol (ver [28, 32, 33, 34, 39]). En una árbol de decisión los ejemplos se distribuyen desde la raíz hasta algún nodo hoja. Cada nodo especifica una prueba sobre algún o algunos atributos del ejemplo, el resultado de esta prueba determina la ruta que seguirá hacia el siguiente nivel; los nodos hoja determinan la clasificación del ejemplo.

Esta familia de algoritmos, asume que la pertenencia a una clase se puede identificar a partir de ciertas combinaciones de los atributos de la instancia. Entonces, mediante una estructura de árbol es posible de encontrar esas combinaciones en la mayoría de los casos, de tal forma que el conocimiento codificado de esta manera puede ser analizado por un especialista,

el cual puede validar o rechazar las combinaciones de atributos obtenidas.

Existe una gran variedad de árboles de decisión, están aquellos cuya forma de evaluar instancias sólo permite valores enteros o nominales en un rango finito tanto en los atributos como en la función objetivo; también existen otros que permiten valores reales para los atributos, pero resuelven una tarea de clasificación; por último están los árboles que efectúan regresión. Brevemente explicaremos algunos de ellos, centrando la atención en el tipo de árboles que permiten valores reales.

Los árboles que manejan una cantidad grande de clases y atributos o los que manejan valores reales, tienen algunos problemas intrínsecos a su naturaleza, y estos problemas tienen relación directa con los criterios que se usan para su crecimiento:

1. Hasta qué punto deben de seguir creciendo para conseguir la precisión deseada.
2. Qué tanto deben de sacrificar la precisión de sus predicciones para mantener la eficiencia computacional.

Es a causa de estos problemas que generalmente se adopta el criterio de *Occam's Razor*, que en este caso consiste en seleccionar como mejor opción al árbol más pequeño que sea capaz de minimizar el error asociado con los datos de entrenamiento.

3.3.3.1 El algoritmo ID3

El algoritmo ID3 es el más básico para formar árboles de decisión, este algoritmo parte de un conjunto de datos de entrenamiento; mediante una evaluación de los atributos, determina cual de ellos es el que permite hacer una mejor discriminación entre las clases, es decir, cual de ellos presenta mayor ganancia de información, esto lo determina mediante una medida que se llama entropía, y está dada por:

$$Entropia(X) \equiv \sum_{i=1}^{c} -p_i log_2 p_i \tag{3.11}$$

Donde X es el conjunto de instancias y c es el número de clases en que se divide el conjunto de ejemplos, p_i es la proporción de X que pertenece a la clase i.

Una vez que se determina la entropía de X, se procede a identificar la eficacia de cada atributo para clasificar los datos de entrenamiento. Esto se hace mediante:

$$Ganancia(X, a_j) \equiv Entropia(X) - \sum_{v \in Valores_{aj}} \frac{|X_v|}{|X|} Entropia(X_v) \tag{3.12}$$

Donde a_j representa al atributo j que describe los datos, $Valores_{aj}$ es el conjunto de posibles valores que puede asumir el atributo a_j. X_v es el subconjunto de X donde el atributo a_j presenta el valor v.

Mediante la medida de la ganancia de información, se puede seleccionar al atributo que sea el más apto para discriminar las clases, en el nodo actual se asigna una prueba para

determinar los valores del atributo elegido, extendiendo tantas ramas como posibles valores pueda presentar el atributo en cuestión. Este proceso se repite hasta que todos los ejemplos queden clasificados de manera adecuada, o hasta que el error en la clasificación sea muy pequeño.

Id3 es un algoritmo efectivo para identificar instancias cuyos atributos están dados por un conjunto finito de valores nominales o enteros y resuelve un problema de clasificación.

3.3.3.2 Árboles de discriminación lineal

Los algoritmos de aprendizaje basados en árboles han demostrado ser métodos eficientes para trabajos de clasificación, pero este tipo de tareas casi nunca tienen sólo atributos nominales o enteros para representar a las instancias. Por esta razón, los algoritmos de árboles han evolucionado, de tal forma que son capaces de manejar números reales como atributos, uno de los casos más populares es el C4.5, pero también están aquellos que además, no limitan la forma de discriminar en los nodos mediante únicamente un atributo, éste es el caso de los árboles de discriminación lineal (ver[22, 45, 46]).

Los árboles de discriminación lineal, son un tipo de árbol que puede manejar atributos reales y que realizan un trabajo de clasificación, donde cada uno de sus nodos efectúa un proceso de clasificación, el cual se va extendiendo por sus ramas, hasta que las instancias llegan a un nodo hoja el cual define la clasificación final. Existen tres categorías principales:

- Una función lineal, donde se hace la discriminación por medio de un umbral; divide el nodo en dos partes.
- Una máquina de discriminación lineal, que divide el nodo en cuantas clases existan.
- Entropía difusa para discriminar, generalmente parten el nodo en dos clases pero se pueden extender.

Función lineal. El tipo más sencillo de árbol de discriminación lineal es aquel cuyos nodos están formados por una combinación lineal de los atributos que definen a las instancias o por alguna función lógica o algebraica sobre algunos de estos atributos. Los nodos se ramifican cuando mucho en dos, donde la pertenencia a cada una de estas ramas se define por un umbral que se aplica al resultado producido por la función de discriminación usada. Este tipo de árbol convierte un problema de multi-clasificación en varios problemas de clasificación binaria.

Máquinas de discriminación lineal. El segundo tipo de árbol de discriminación, entrena máquinas de discriminación lineal en cada nodo. Una máquina de discriminación lineal o LMD por sus siglas en inglés, es una función de discriminación $g_i(X)$ que tiene la forma de $W_i^T X$, donde W es una matriz de coeficientes ajustables que efectúa una combinación lineal sobre X que es la matriz de instancias.

Con este método se asigna la pertenencia de una instancia $\overrightarrow{x} \in X$ a la clase i si y sólo si:

$$\begin{gathered}(\forall i : i \neq j) \\ g_i(\overrightarrow{x}) > g_j(\overrightarrow{x})\end{gathered} \tag{3.13}$$

Si no existe un sólo máximo la clasificación queda como indefinida. Cuando se presentan ciertas distribuciones mal comportadas de instancias en un nodo, éste puede producir una sola clasificación para todas las instancias, si es así, ese nodo se convierte en un nodo hoja, el cual asigna la clasificación más frecuente.

Entropía difusa. En este caso cada nodo consiste de un vector de coeficientes $\overrightarrow{w}$. Entonces, el problema consiste en realizar una optimización, la cual pretende encontrar el $\overrightarrow{w}$ que produzca una mejor partición.

La manera de hacer la discriminación en cada nodo, es a través del concepto de entropía difusa, definida por la función de discriminación:

$$g(\overrightarrow{x}) = \frac{1}{1 + e^{-\overrightarrow{w}\,\overrightarrow{x}}} \tag{3.14}$$

Donde una instancia $\overrightarrow{x}$ puede pertenecer a más de una clase de manera simultánea pero en diferente grado. El objetivo de este tipo de árboles no es minimizar el error de las particiones de manera inmediata, sino producir el árbol más pequeño cuyo error total sea mínimo.

3.3.3.3 Árboles de regresión

Los algoritmos de árbol que se explicaron anteriormente realizan tareas de clasificación. Pero también existe una categoría muy especial de árboles cuya tarea no consiste en asignar una clase determinada a cada instancia, sino que se usan para aproximar una función, éstos son los árboles de regresión (ver [5, 6, 24, 34, 42, 43]). A continuación se explicará brevemente el funcionamiento de estos árboles, así como sus características particulares.

Existen básicamente dos tipos de árboles de regresión, el tipo más simple, consiste en nodos que separan el conjunto de datos en un número determinado de ramas; en los nodos hoja, a las instancias que lleguen hasta ellos se les asigna el promedio que se obtuvo de las instancias que formaron ese nodo. Este tipo de árboles no es muy eficaz, ya que no pueden encontrar los valores correctos de muchas funciones simples. El otro tipo de árboles de regresión usa este tipo de nodos únicamente para expandir el árbol, pero en los nodos hoja usa modelos de regresión lineal que permiten asignar un valor más acertado a instancias que aproximan funciones más complejas. Dos algoritmos muy comunes, aunque no muy eficientes son M5 y RETIS.

El principal problema de los árboles de regresión consiste en que es muy difícil extenderlos hacia problemas complejos, o que manejen un número grande de atributos. Para tratar este problema, se crearon algunos algoritmos como EACH, RISE, HTL, que integran métodos de aprendizaje basados en instancias para los nodos hoja, con métodos de partición que forman estructuras de árbol.

3.3.3.4 Algoritmos con estructura de árbol

El concepto tradicional de un árbol de decisión se ha ido modificando, de tal manera que se ha combinado el uso de diversos algoritmos de aprendizaje automático con la útil estructura de árbol, la cual proporciona las ventajas del aprendizaje modular.

Los algoritmos, donde el aprendizaje se produce codificando ciertos aspectos relevantes de las instancias, en una estructura de árbol han tenido éxito en diversos problemas; a continuación se describen tres ejemplos.

El algoritmo de multi-clasificación jerárquico propuesto en [26], donde un problema de c clases se descompone recursivamente en $c-1$ problemas de clasificación binaria, codificados en meta clases; al final, cada hoja del árbol le corresponde a una de las c clases originales. Este algoritmo realiza extracción de características reduciendo el espacio, proyectándolo en un espacio de características más pequeño, al momento que construye cada nivel del árbol.

El principal problema de este enfoque consiste en que la complejidad computacional puede llegar a ser inmanejable para una cantidad mediana o grande de clases.

Un algoritmo para modelado acústico propuesto en [52], consiste en organizar modelos ocultos de Markov en una estructura de árbol. Inicia a partir de tres estados, las salidas relacionadas a esos estados se clonan para inicializar el siguiente nivel del árbol, donde se hace un agrupamiento de los nodos que pertenecen al mismo estado. El número de nodos y de niveles se sigue incrementando de esta manera hasta que se alcanzan los resultados deseados.

El algoritmo de mezcla jerárquica de expertos propuesto en [23], aquí el aprendizaje se trata como un problema de probabilidad máxima, en este algoritmo, un problema grande se divide en problemas más pequeños, cuyas soluciones parciales se combinan para obtener la solución final del problema.

La manera en que funciona este algoritmo es estableciendo una serie de funciones lineales en los nodos superiores de un árbol, estas funciones se encargan de establecer la ruta que seguirán las instancias para llegar a los nodos hoja. En los nodos hoja están los llamados expertos, los cuales también son funciones lineales, pero que determinan la solución para un subproblema. El resultado final se puede dar como una distribución de probabilidades.

3.4 Validación de los algoritmos

En esta sección se explican algunas de las técnicas usadas en esta tesis para establecer métricas de evaluación de los algoritmos. Primero se explican dos maneras de comprobar la capacidad de generalización de los algoritmos de aprendizaje, como son la validación cruzada y la validación por medio de un conjunto de prueba. Por último, se presentan las métricas usadas para cuantificar los errores que se muestran en los resultados.

3.4.1 Validación cruzada

La validación cruzada es un método efectivo para probar la validez de los algoritmos, que consiste en realizar una división aleatoria de tamaño k del conjunto total de ejemplos, dejando $k-1$ grupos de instancias para el entrenamiento y usando el grupo restante para prueba, este proceso se repite k veces de tal forma que al final se tiene una forma confiable para determinar la capacidad de generalización de los algoritmos empleados, evitando comúnmente combinaciones que presenten comportamientos particularmente malos para los algoritmos. En la tabla 3.2 se explica el procedimiento de la validación cruzada.

`Dividir` C_{datos} `en` k `subconjuntos,`
$g_1 \bigcup g_2 \bigcup \ldots \bigcup g_k = C_{datos}$
`Para` $i = 1$ `hasta` k `hacer`
$C_{entrena} = C_{datos} \setminus g_i$
$C_{prueba} = g_i$
`Entrena el modelo usando los ejemplos en` $C_{entrena}$
`Evalúa el modelo para los ejemplos de` C_{prueba}
`Guarda la predicción`
`Termina`
`Evalúa la precisión de la predicción`

Tabla 3.2: Algoritmo de validación cruzada

Normalmente se usa $k = 10$ pero esto puede variar de acuerdo a la cantidad total de ejemplos o al tiempo necesario para entrenar un modelo.

3.4.2 Conjunto de prueba

Otra forma de validar los algoritmos de aprendizaje automático, es a través de un conjunto de prueba, este conjunto está formado por instancias que no se han presentado a los algoritmos en la fase de entrenamiento, es decir, son instancias completamente nuevas para los clasificadores. Las instancias pueden tener algunas características similares a las de otras en el conjunto de entrenamiento, pero en términos generales deben de ser diferentes.

Es preferible que las instancias de este conjunto estén bien distribuidas sobre el espacio de representación, para asegurar que se obtenga una buena apreciación de la capacidad de generalización de los algoritmos; aunque esta característica también debe de tomarse en cuenta para el conjunto de entrenamiento, ya que con este enfoque es más fácil producir clasificadores con sobreajuste.

Esta es una buena opción cuando el proceso de entrenamiento es muy largo y es muy ineficiente entrenar diez clasificadores, es particularmente útil en el caso de que se puedan tener ejemplos cubriendo todo el espacio de representación y se busca obtener estadísticas de manera rápida.

3.4.3 Error medio cuadrático

De acuerdo con la teoría de Gauss de los errores, que supone que estos se producen por causas aleatorias, se toma como la mejor estimación del error, el llamado error medio cuadrático (RMS) definido por:

$$RMS = \sqrt{\frac{\sum_{i=1}^{n} \mid pred_i - y_i \mid}{n}} \tag{3.15}$$

Donde $pred_i$ es la predicción para la instancia i, y y_i es el valor correcto, también de la instancia i, n es el número total de instancias.

Este error también sirve como una métrica probabilística para determinar el desempeño de un clasificador con instancias no observadas anteriormente, de esta manera, los mejores resultados serán aquellos que minimicen el error medio cuadrático.

Para evaluar la precisión de un ajuste, en los experimentos que se explicarán en las siguientes secciones, se calcula el error medio cuadrático entre los puntos de flujo del espectro original, contra los puntos de flujo de un espectro reconstruido mediante la predicción obtenida usando los clasificadores.

Capítulo 4

CARACTERIZACIÓN DE LOS ESPECTROS

Para resolver el problema que se plantea en esta tesis, es muy importante tomar en cuenta las características de los datos, así como las restricciones que se indican para su manejo.

Con la finalidad de cumplir algunos de los objetivos establecidos, es necesario que los datos tengan un cierto preprocesamiento. Este preprocesamiento vuelve más difícil la fase de clasificación, pero también presenta muchas ventajas, por ejemplo puede eliminar efectos externos de los espectros, como los que se explicaron en la sección 2.2.1.1, esta característica en particular resulta especialmente útil si se quiere trabajar con espectros reales.

La caracterización de los espectros consiste en encontrar una manera eficiente de codificar cada espectro, manteniendo la información relevante, pero reduciendo la dimensionalidad de los datos tanto como sea posible.

Debido a que los dos tipos de datos que se usan en la tesis, presentan serias diferencias, es necesario analizarlos y encontrar la manera de identificar qué propiedades comparten y qué estrategia de caracterización puede ser útil para ambos tipos. A continuación se presentan las características relevantes de los dos tipos de espectros y se detallan los procesos de caracterización a los que se sometieron estos datos.

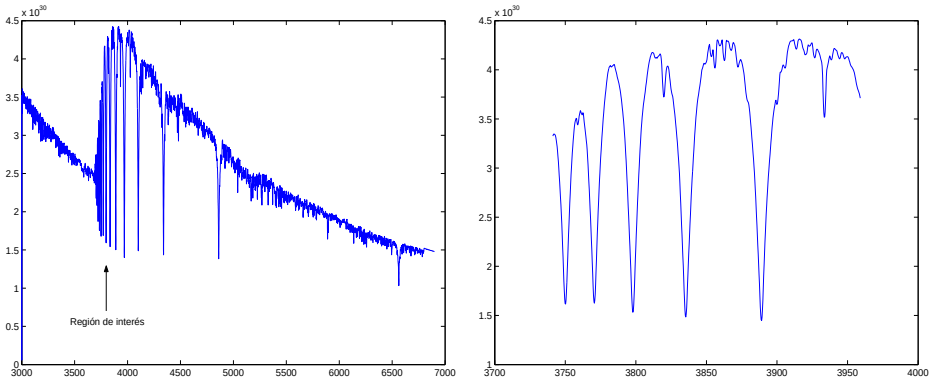
4.1 Espectros

La identificación de las poblaciones estelares que conforman una galaxia a partir de su espectro, se hace en dos tipos de datos. Primero en espectros simulados, usando estos datos cuyos parámetros están plenamente identificados, se entrenan los clasificadores correspondientes, además, con este tipo de datos se puede medir la validez de los clasificadores entrenados. Una vez que los clasificadores tienen un grado de validez aceptable en datos simulados, se pueden usar para determinar las poblaciones estelares en espectros observacionales, aunque sus parámetros no se conozcan.

4.1.1 Simulados

Se generaron 10800 datos simulados mediante el procedimiento que se explicó en la sección 2.3.2, se usaron 7200 para entrenamiento y 3600 para prueba, estos espectros tienen un rango espectral de 3000 $\AA$ a 6900 $\AA$ y una resolución de 0.3$\AA$, la región de interés tiene 717 puntos de flujo.

Dos ejemplos de los datos sintéticos se pueden ver en la figura 4.1, donde el primer recuadro muestra un espectro en todo el rango espectral que se maneja en este trabajo, y la segunda imagen es un espectro simulado en la región de interés, mostrando que los perfiles de sus líneas están muy bien definidos.



(a) Ejemplo de un espectro sintético (b) Región de interés

Figura 4.1: Espectros sintéticos

Estos datos se crearon a partir de espectros de alta resolución, donde ligeras variaciones en el perfil de las líneas son indicadores del tipo de poblaciones estelares que integran al espectro.

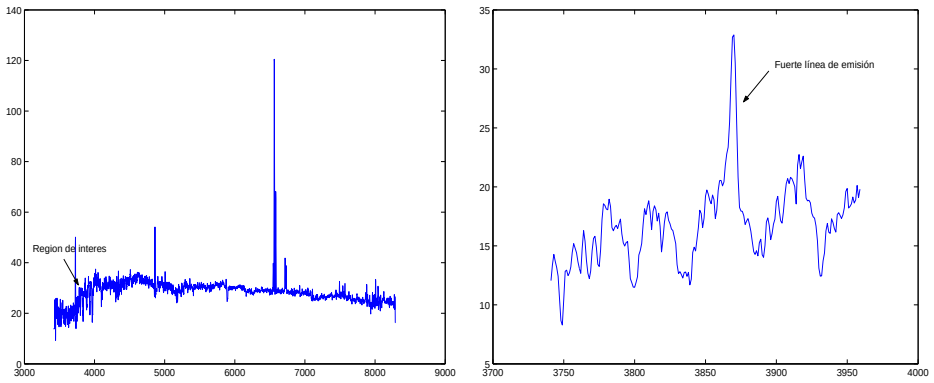
Los perfiles de las líneas están muy bien definidos a pesar del ruido y la atenuación, además todos presentan curvas casi simétricas. La mayor parte de las asimetrías que presentan están en la parte superior de las líneas, estas asimetrías también ayudan al proceso de identificación de edades.

4.1.2 Reales

Para este trabajo se usaron 1155 espectros observacionales con una resolución de 1$\AA$ y un rango espectral de 3435$\AA$ a 8315$\AA$. En estos espectros se observa una gran cantidad de líneas de emisión, los lugares donde aparecen estas líneas no son regulares, además también hay líneas de emisión adentro de líneas de absorción.

En la figura 4.2 se puede ver que las fuertes líneas de emisión modifican el perfil del espectro, haciéndolo lucir muy diferente a los espectros sintéticos, también se aprecia que los perfiles de las líneas no están tan definidos como en el caso sintético.

La resolución que presentan estos espectros es muy pobre en relación a la de los espectros simulados, por lo tanto es más difícil encontrar las sutiles variaciones del perfil que identifican a las poblaciones; además los perfiles de las líneas no están bien definidos, aunque de esta falta



(a) Ejemplo de un espectro observacional (b) Región de interés

Figura 4.2: Espectros observacionales

de definición no se puede asumir la presencia de gran cantidad de ruido, ni es posible suavizar el espectro mediante algún filtro gaussiano ya que esas asimetrías pueden representar mucha información y el comportamiento del ruido en los datos observacionales no es necesariamente gaussiano.

Ninguno de los ejemplos observacionales está clasificado, por lo tanto el error en la predicción sólo puede determinarse mediante el error entre el espectro original y el reconstruido.

4.2 Enfoque

El objetivo del proceso de caracterización del espectro consiste en reducir la dimensionalidad de los datos, así como evitar las fuertes líneas de emisión que no aportan información relevante sobre la edad de las poblaciones de la galaxia y que generalmente están presentes en espectros observacionales. Este proceso también es útil para hacer que los algoritmos de regresión, que se explicarán más adelante, no tengan que lidiar con problemas tales como la atenuación del espectro o el ruido.

En algunos casos el preprocesamiento de los espectros puede causar que la identificación de las edades sea más difícil, pero se decidió aplicarlo por las ventajas adicionales que trae consigo; sobre todo para el manejo de datos observacionales.

Cómo ya se explicó anteriormente, existen muchos problemas en tratar de ajustar un espectro completo a una plantilla, por lo tanto, el enfoque que se sigue en este trabajo es ajustar únicamente las cinco líneas donde se codifica la mayor parte de la información sobre la edad de sus poblaciones (ver la sección 2.2.2). Mediante el ajuste únicamente en esta región, es posible evitar los lugares donde pueden aparecer grandes líneas de emisión, así como contar con una porción mucho más manejable del espectro. El procedimiento está dividido en dos fases:

1. Identificación del perfil de las líneas

2. Caracterización del espectro.

En la primera etapa se extrae la forma del perfil de cada línea, esto con el fin de limpiar los datos tanto de ruido como de otras líneas cercanas; en la segunda etapa, se buscan atributos relevantes que sean capaces de codificar la información relevante pero que a su vez permitan reducir el tamaño de los datos. A continuación se explican las dos etapas y los procedimientos que se siguieron en cada una de ellas.

4.3 Identificación del perfil de las líneas

La identificación del perfil de las líneas, es un procedimiento que se realiza para limpiar los datos, resaltando la información relevante y omitiendo líneas espectrales que aparecen en ciertas regiones donde sólo contaminan el perfil de líneas importantes.

Este procedimiento es muy útil para obtener perfiles más definidos, sobre todo en datos reales, donde los perfiles son muy asimétricos y presentan serias irregularidades, producidas principalmente por contaminación de otras líneas y ruido. El resultado de la identificación del perfil en la primera línea de la serie de Balmer se puede observar en la figura 4.3, en el recuadro de la izquierda se muestra el alineamiento y en el de la derecha se muestra la identificación del perfil.

- Desventajas:
    1. Pierde información.
    2. Hace que el proceso de clasificación sea más difícil.
- Ventajas:
    1. Reduce el impacto del ruido en los perfiles.
    2. Obtiene perfiles de líneas con forma definida.
    3. Hace que los datos sean insensibles a otros efectos que los contaminan, como los que se explicaron en la sección 2.2.1.1.
    4. Es posible comparar espectros de diferentes resoluciones.
    5. Elimina líneas que no proporcionan información relevante.
    6. La cantidad de información que pierde es poca.

El proceso de identificar el perfil de las líneas se realiza mediante dos pasos: el primero sirve para estandarizar los datos, mientras que en el segundo se encuentran los puntos que definen el perfil de las líneas; estos pasos se detallan en las siguientes subsecciones.

4.3.1 Alineamiento y normalización

La normalización del espectro es un paso necesario, debido a que los modelos sintéticos y los espectros observacionales se encuentran a escalas diferentes, sin este paso, no es posible realizar un ajuste entre estos dos conjuntos.

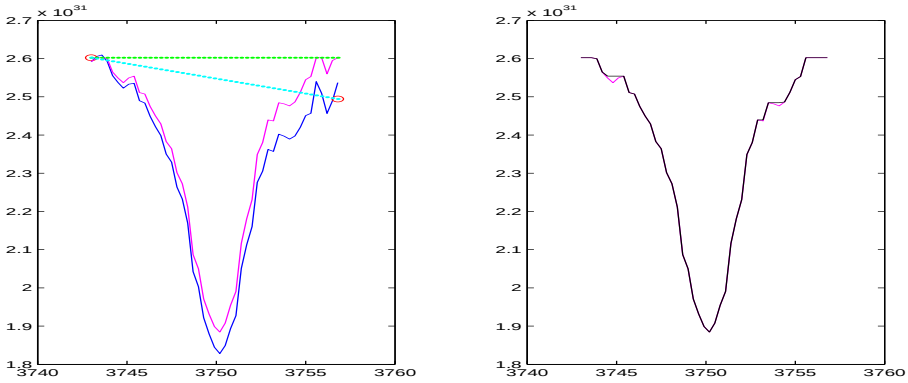


Figura 4.3: Proceso de identificación del perfil de una línea

La normalización consiste en modificar la escala de los datos, de tal forma que el mínimo sea -1 y el máximo sea 1. Esto se logra usando:

$$\overrightarrow{s}_{norm} = \frac{2(\overrightarrow{s} - min)}{max - min} - 1 \tag{4.1}$$

Donde $\overrightarrow{s}_{norm}$ es el espectro normalizado, min y max son los valores mínimo y máximo de los flujos del espectro $\overrightarrow{s}$ respectivamente.

El proceso de normalización se realiza en dos etapas: en la primera se normalizan las cuatro líneas correspondientes a la serie de Balmer y en la segunda se normaliza la línea del calcio, de esta manera, no se guarda ninguna relación entre la fuerza de las líneas que identifican poblaciones jóvenes con la fuerza de la línea que identifica poblaciones viejas, sino únicamente sus perfiles.

El alineamiento, se realiza para cada línea. Se encuentran los puntos donde la absorción es mínima, al principio y al final de la línea, se ajusta una función lineal a esos puntos y se suma al resto de la línea.

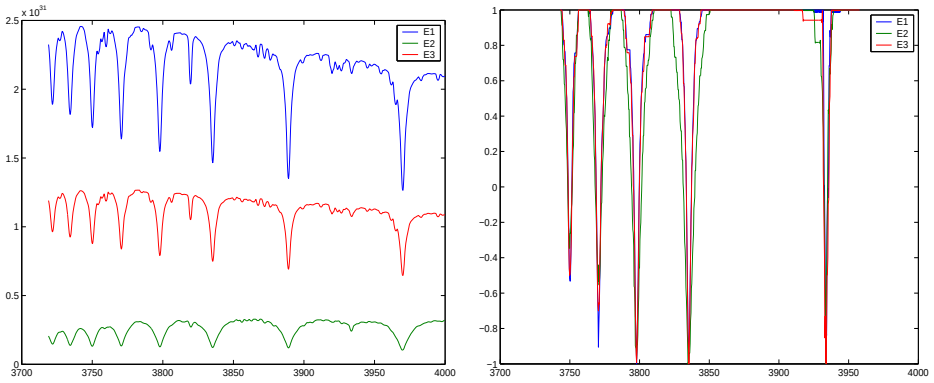
Se dice que este proceso vuelve más difícil la clasificación de los espectros porque un espectro alineado y normalizado con una cierta combinación de parámetros puede ser muy similar a un espectro con parámetros muy diferentes pero con un nivel de ruido mayor. Este efecto se puede ver en la figura 4.4.

Este método de alineamiento y normalización de las líneas de Balmer y Calcio, ya ha sido utilizado con anterioridad por [44].

4.3.2 Identificación de perfiles

La identificación de perfiles consiste en definir cuales son los puntos relevantes que contribuyen a la forma de la línea en cuestión, eliminando puntos que forman parte de otras líneas cercanas.

Este proceso se realiza de la siguiente manera: primero se identifica el punto de máxima absorción (pma) de la línea que se quiere caracterizar, cuyos valores están dados en $\overrightarrow{p} \in \mathrm{R}^{1\times n}$, donde n es el número total de puntos que forman la línea. Una vez, identificado el pma se debe de encontrar $\overrightarrow{j}$ y $\overrightarrow{k}$ donde:



(a) Espectros sintéticos (b) Espectros alineados y normalizados

	$\overrightarrow{p_1}$					$\overrightarrow{p_2}$				$\overrightarrow{p_3}$				
E1	**0.2**	0	0	0	0	0	**0.4**	0	0	0	0	0	**0.4**	0
E2	0	0	0	**0.4**	0	0	**0.3**	0	0	0	**0.3**	0	0	0
E3	**0.1**	0	0	0	0	**0.1**	0	0	0	**0.8**	0	0	0	0

Figura 4.4: Comparación de espectros limpios sin preprocesar contra los mismos espectros alineados y normalizados. Los valores que generaron cada espectro se presentan en la tabla.

$$\begin{aligned} \overrightarrow{j} &= \overrightarrow{p}[i] < \overrightarrow{p}[i-1] \\ \forall i: &\quad 2 \leq i \leq pma \end{aligned} \tag{4.2}$$

$$\begin{aligned} \overrightarrow{k} &= \overrightarrow{p}[i] > \overrightarrow{p}[i-1] \\ \forall i: &\quad pma+1 \leq i \leq n \end{aligned} \tag{4.3}$$

Después de haber encontrado todos los puntos que presentan un comportamiento monótono creciente desde el punto de máxima absorción hacia ambos lados, se define un vector $\overrightarrow{v} \in \mathrm{R}^{1\times n}$ el cual está muestreado uniformemente $\overrightarrow{v}[i+1] = \overrightarrow{v}[i] + 0.3$ y se realiza una interpolación cúbica, usando splines, entre $\overrightarrow{p}[\overrightarrow{j} \bigcup \overrightarrow{k}]$ y $\overrightarrow{v}$, de tal forma que se obtiene un vector $\overrightarrow{p}_{nuevo}$ que representa el perfil limpio de la linea de interés. Este vector tiene valores en todo el rango espectral deseado con una resolución específica, en este caso fue cada $0.3\mathring{A}$, por lo tanto es posible comparar espectros con diferentes resoluciones.

Este proceso excluye otras líneas cercanas a la línea de interés, las cuales sólo añaden ruido o información irrelevante al momento de hacer el ajuste, también tiene la ventaja de que permite manejar espectros de diferentes resoluciones, preferentemente entre $0.2\mathring{A}$ y $1\mathring{A}$, sin la necesidad de algún procedimiento adicional.

4.4 Caracterización del espectro

Una vez identificado el perfil de las líneas, es necesario realizar su caracterización, esto por dos razones principales:

1. Reducir la dimensión de los datos para trabajar con una cantidad mínima de atribu-

tos por espectro, lo que permite crear clasificadores más eficientes, tanto en tiempo de entrenamiento, como en tiempo de predicción.

2. Obtener características generales asociadas a cada espectro. Esta es la razón más importante, debido a que permite reducir en gran medida los efectos de ruido. Es decir, que si existen puntos contaminados por grandes cantidades de ruido, mediante la caracterización del espectro, estos puntos no se toman en cuenta tal y como están, sino junto con los puntos adyacentes, como un todo, al que cada punto hace una contribución.

En este trabajo se siguieron dos enfoques principales para la caracterización de los espectros, el primero por medio del análisis de componentes principales; habiéndose encontrado serios problemas en este enfoque, se decidió hacer la caracterización por medio de la extracción de anchos en el perfil de las líneas. A continuación se explican ambas formas.

4.4.1 Caracterización por medio del análisis de componentes principales

El análisis de componentes principales, permite realizar una transformación lineal que mapea los datos desde un espacio dimensional determinado a un espacio más pequeño. Esta reducción implica cierta pérdida de información, pero es posible preservar tanta información como sea posible, minimizando $\|X - \widehat{X}\|$, donde X son los datos originales y $\widehat{X}$ son los datos reconstruidos mediante k componentes principales.

4.4.1.1 Preliminares

El método de PCA, proyecta los datos en la dirección hacia donde existe la mayor variación de los datos; esa dirección se determina por medio de los eigenvectores de la matriz de covarianza correspondientes a los eigenvalores más grandes. La magnitud de cada eigenvalor corresponde a la varianza de los datos en la dirección de su respectivo eigenvector.

La manera de obtener los k mejores componentes principales es como sigue:

Suponiendo que X es la matriz compuesta por $\overrightarrow{x_i}$ $i = 1, \ldots n$ vectores, cada uno compuesto por m valores; entonces calcular $\overline{x}$, con:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} \overrightarrow{x_i} \tag{4.4}$$

A partir de $\overline{x}$ se sustrae la media $\Phi_i = \overrightarrow{x_i} - \overline{x}$ y se forma la matriz $A = [\Phi_1, \Phi_2, \ldots, \Phi_n]$ $A \in \mathrm{R}^{m \times n}$, después se calcula la matriz de covarianza $C \in \mathrm{R}^{m \times m}$:

$$C = \frac{1}{n} \sum_{r=1}^{n} \Phi_r \Phi_r^T = AA^T \tag{4.5}$$

y sus correspondientes eigenvalores $\lambda_1 > \lambda_2 > \ldots > \lambda_m$ y eigenvectores $u_1, u_2, \ldots, u_m$. Como C es una matriz simétrica, entonces cualquier vector se puede escribir como una combinación lineal de los eigenvectores:

$$\overrightarrow{x} - \overline{x} = \sum_{i=1}^{m} b_i u_i \tag{4.6}$$

La reducción de la dimensionalidad de los datos, se hace entonces considerando sólo los k eigenvalores más grandes, donde $k \ll m$

$$\overrightarrow{x} - \overline{x} = \sum_{i=1}^{k} b_i u_i \tag{4.7}$$

Por lo tanto, los datos se pueden representar por sólo k atributos. En el caso de los espectros, el 99% de la varianza está contenida en los primeros 2 eigenvectores, pero se observó que con 30 componentes principales se obtenían reconstrucciones más fieles, por lo tanto se decidió hacer los experimentos con 30 pc.

4.4.1.2 Consideraciones

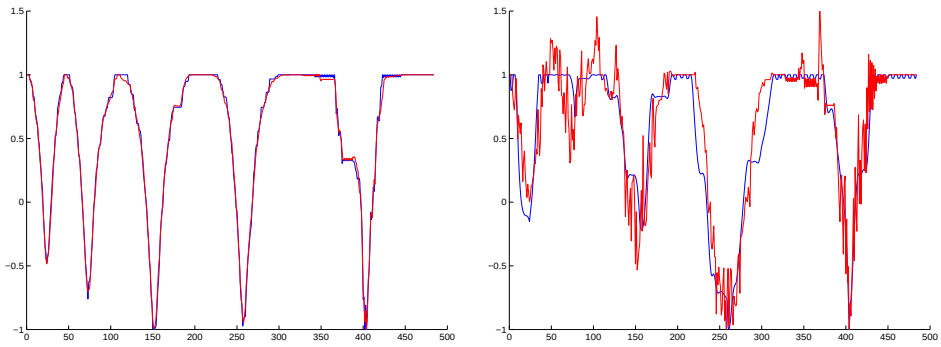
Cómo ya se explicó anteriormente, es deseable y aún más, necesaria, la reducción de la dimensionalidad de los datos; debido a las características que presenta el método de análisis de componentes principales, se consideró como una opción obvia. Primero se aplicó al espectro completo, pero tratándose de datos sintéticos producidos todos de una manera similar, donde ciertas regiones de los mismos se crearon como combinaciones lineales, el número de componentes principales obtenidos fue muy pequeño; por lo tanto, se decidió aplicarlo únicamente a la región de interés del espectro previamente preprocesado, es decir, al perfil de sus líneas.

La caracterización del espectro por medio de componentes principales, resultó muy exitosa en el caso de los espectros simulados, con tan sólo 30 pc o menos fue posible reconstruir exitosamente cada espectro, obteniendo una reconstrucción tan similar al original, que el error producido es despreciable. Desafortunadamente, este comportamiento no se conservó al momento de caracterizar espectros reales, cuyos perfiles no están tan bien definidos como en el caso de los espectros sintéticos.

La reconstrucción de espectros observacionales, produce un espectro con una gran cantidad de ruido, pero este no es el mayor problema, la principal desventaja consiste en que, en muchos casos, la reconstrucción no respeta la forma de ciertas líneas, incluyendo su magnitud y la relación de una región a otra.

En la figura 4.5, se puede ver cómo un espectro simulado se puede recuperar casi a la perfección usando 30 componentes principales, pero al intentar recuperar un espectro observacional, éste no se puede recuperar de manera satisfactoria.

Aunque la técnica de PCA's muestra un muy buen desempeño para caracterizar espectros sintéticos, cuando los espectros reales se proyectan en el espacio generado por los anteriores, los valores de su proyección son muy diferentes a las proyecciones sintéticas. Por lo tanto, un clasificador entrenado para reconocer las sutiles variaciones en los espectros sintéticos no será capaz de generalizar a partir de valores tan diferentes cómo los obtenidos por las proyecciones de los datos reales. Las medias de los valores producidos por 30 pc para datos reales y sintéticos,



(a) Espectro simulado (b) Espectro observacional

Figura 4.5: Reconstrucción de espectros, usando 30 componentes principales

así como sus desviaciones estándar se muestran en la figura 4.8.

4.4.2 Caracterización por medio de anchos

La caracterización del espectro por medio de anchos consiste en encontrar el promedio del ancho de una línea en cierta región, este proceso se repite para las cinco líneas de la región de interés, en doce regiones para cada una (ver la figura 4.7).

En la figura 4.6 se presenta la región superior de las dos últimas líneas de Balmer en espectros con parámetros de edad distintos, donde se puede apreciar que los anchos de las líneas son diferentes, estas diferencias se presentan como combinaciones de anchos en todas las líneas, lo cual permite encontrar comportamientos de los perfiles relevantes a cada tipo de población estelar. Los espectros que aparecen en esa figura son los mismos que se muestran en la figura 4.4.

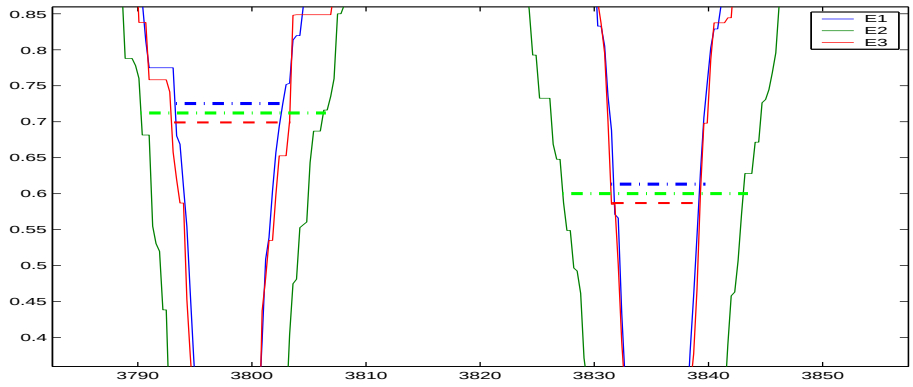
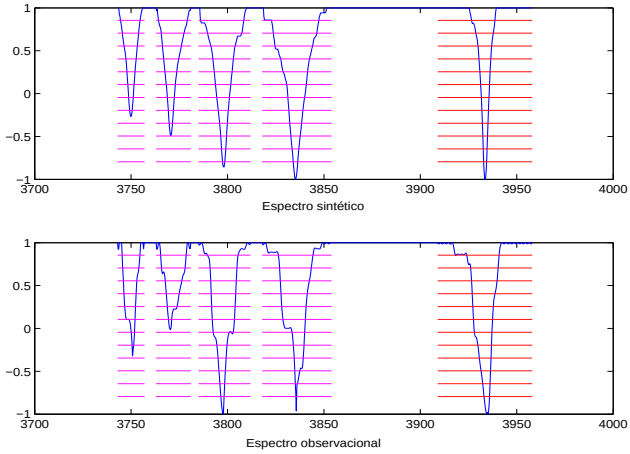


Figura 4.6: Los anchos de las líneas presentan algunas diferencias que permiten identificar los tipos de espectros

4.4.2.1 Extracción de anchos

La fase de extracción de anchos es una parte clave del proceso de caracterización del perfil de las líneas, en esta etapa se elimina la mayoría de los efectos producidos por el ruido, logrando

que estos efectos no repercutan de forma decisiva en el momento de la clasificación. También se logra reducir la dimensionalidad de los datos en un 90%, y aunque modifica la representación de la línea, mantiene su información relevante.



(a) Extracción de anchos en las 5 regiones de interés

(b) Extracción de anchos en una línea

Figura 4.7: Extracción de anchos

Para realizar este proceso, se define un conjunto de ns intervalos $INTER$ igualmente espaciados, desde -1 a 1. Para cada $inter_s \in INTER$, $1 \leq s \leq ns$, se encuentra el comportamiento de la línea hacia la izquierda y hacia la derecha desde el punto de máxima absorción, $\overrightarrow{p}_l$ y $\overrightarrow{p}_r$ respectivamente, entonces, el ancho promedio aw_s en el intervalo s, se obtiene:

$$\begin{gathered} aw_s = \frac{\sum \overrightarrow{p}_r[i]}{ni_r} - \frac{\sum \overrightarrow{p}_l[j]}{ni_l} \\ \forall i : \overrightarrow{p}_r[i] \in inter_s, \forall j : \overrightarrow{p}_l[j] \in inter_s \end{gathered} \tag{4.8}$$

Donde ns_l es el número de puntos pertenecientes al intervalo s en el lado izquierdo del punto de máxima absorción y ns_r el correspondiente al lado derecho.

El impacto del ruido se reduce porque después de este proceso, las líneas se caracterizan por el comportamiento global que presentan en cierto intervalo y no por puntos individuales que podrían estar contaminados con ruido.

4.4.2.2 Consideraciones

Las características obtenidas mediante la extracción de anchos, permiten considerar las variaciones que existen en los perfiles de las líneas pertenecientes a espectros con diferentes tipos de poblaciones estelares. Además estas características se mantienen similares tanto en los datos sintéticos como en los reales, permitiendo así que un clasificador entrenado con los anchos de datos sintéticos sea capaz de generalizar y obtener predicciones aceptables a partir de datos reales.

En la tabla 4.1 se puede apreciar la diferencia que existe entre los valores encontrados en datos sintéticos y los obtenidos de datos reales. En la figura 4.8 se muestra esta diferencia

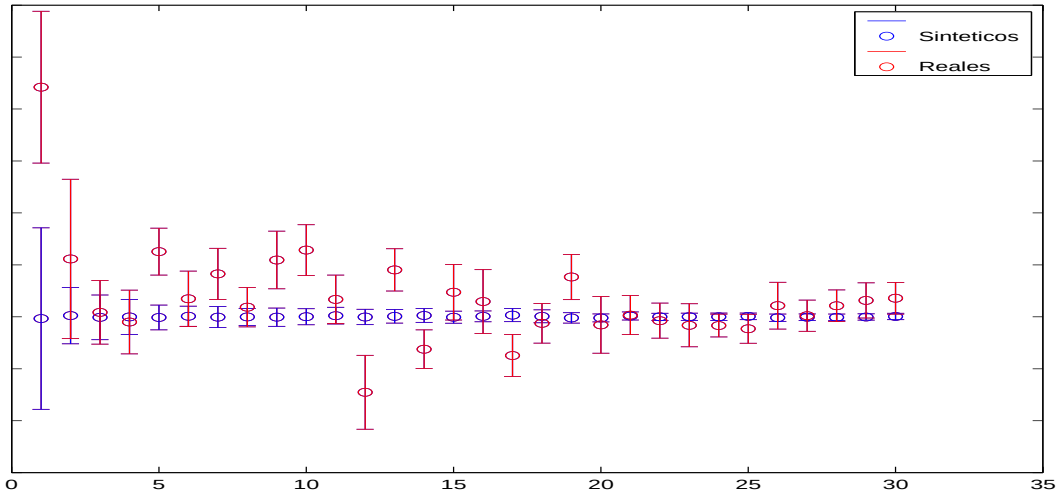
de manera gráfica, presentando las medias y las desviaciones estándar de las características producidas por el método de PCA y por el método de extracción de anchos. Se puede ver que en el caso de las características producidas por PCA, en muchas ocasiones los rangos de los datos reales y de los sintéticos son totalmente disjuntos, mientras que en los anchos se tiene gran similitud. Esta aseveración cobra más importancia considerando que los componentes principales que presentan mayor discordancia están entre los primeros 15, siendo estos los más relevantes.

30 PC	60 ANCHOS
0.408	0.193

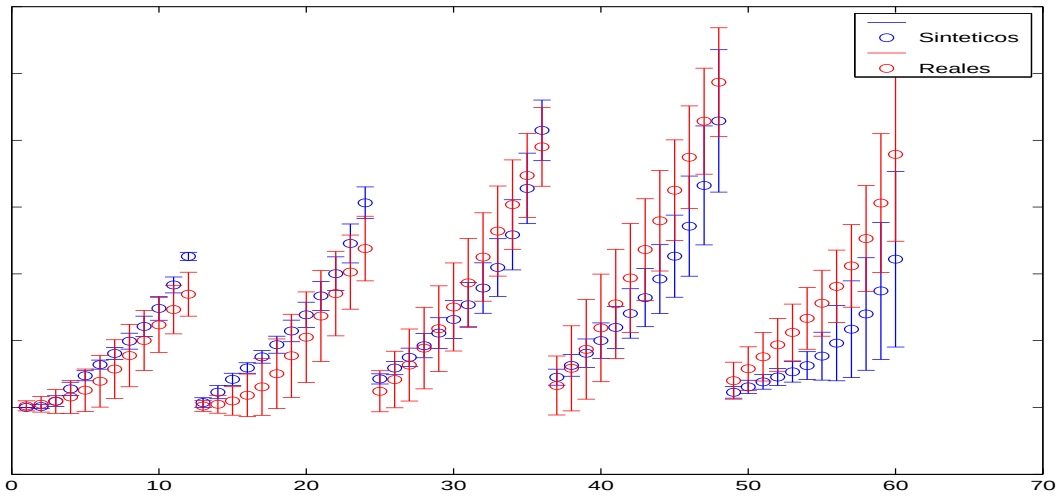
Tabla 4.1: Promedio de la diferencia de las medias de datos sintéticos contra datos reales producidos por 30 pc y 60 Anchos; los valores de las medias están normalizadas de 0 a 1

A partir de los argumentos presentados anteriormente, se concluyó que la mejor opción para caracterizar el perfil de las líneas espectrales es por medio de sus anchos; extrayendo 12 anchos por línea en 5 líneas. Este número se estableció empíricamente para obtener regiones con suficiente concentración de información. Si se estableciera un número mayor de regiones, los anchos estarían determinados por muy pocos puntos o en algunos casos ninguno; mientras que una menor cantidad de regiones, no obtiene las variaciones importantes del perfil de las líneas.

En las siguientes secciones, cada espectro estará definido por 60 atributos.



(a) 30 componentes principales



(b) 60 anchos

Figura 4.8: Medias y desviaciones estándar producidas por 30 componentes principales y 60 anchos; se comparan los valores producidos por datos reales y los correspondientes de los datos sintéticos

Capítulo 5

TRES ENSAMBLES Y EXPERIMENTOS CON DATOS SIMULADOS

5.1 Ensambles de clasificadores

Un ensamble es un grupo de clasificadores entrenados de manera independiente, donde la predicción obtenida por cada uno de ellos se combina mediante votación o por promedio. Usualmente el resultado obtenido por el ensamble es más preciso que los resultados individuales de cada uno de los clasificadores que lo conforman, siempre y cuando los errores no estén correlacionados (ver [8, 13, 18, 21, 29, 38, 50]) y el error individual sea menor de 0.5.

Los ensambles son útiles para solucionar posibles problemas de sobreajuste o de mínimos locales que pudieran tener los clasificadores individuales (ver [41]), también sirven para reducir el impacto del ruido presente en los ejemplos.

Existen diversas maneras de hacer un ensamble, pero se pueden dividir en dos categorías de acuerdo a la forma en que se construyen: ensambles homogéneos y ensambles heterogéneos. A continuación se explican brevemente.

5.1.1 Ensambles homogéneos

Los ensambles homogéneos se construyen usando el mismo clasificador de base, es decir, usan el mismo algoritmo de entrenamiento. El hecho de utilizar el mismo algoritmo en todos los clasificadores podría causar que los errores de los distintos clasificadores estuvieran correlacionados, por lo tanto se recurre a la manipulación de ciertos factores, de tal manera que se puedan obtener clasificadores distintos, aún cuando el algoritmo base sea el mismo.

Una forma común de hacer ensambles homogéneos consiste en manipular el conjunto de entrenamiento, a la manera de Boosting y Bagging (ver [13, 50]), donde a partir de un conjunto inicial, se crean multiconjuntos de datos, favoreciendo la ocurrencia de datos que no son bien clasificados. Otra forma consiste en realizar un proceso iterativo de clasificación, donde en

ciertas etapas, con el fin de facilitar la clasificación se forman metaclases, las cuales agrupan algunas de las clases originales; desagrupándolas posteriormente para obtener la clasificación original. También es posible formar clasificadores independientes modificando el conjunto de atributos, es decir, usando subconjuntos diferentes de atributos para formar cada clasificador. En este trabajo, se usa el último enfoque para formar los ensambles a los que se denominará 'tradicionales'.

5.1.2 Ensambles heterogéneos

Los ensambles heterogéneos están formados por un conjunto de clasificadores que han sido entrenados utilizando diferentes algoritmos base, lo cual permite usar las ventajas intrínsecas de cada tipo de algoritmo y obtener clasificadores cuyos errores no están correlacionados.

Las principales ventajas de este tipo de ensambles radican en la obtención de clasificadores muy diferentes entre si, de tal forma que se pueda tener una certeza en cuanto a que los errores individuales no están correlacionados. Por lo tanto el posible sobreajuste en algún clasificador se compensa en las predicciones de los demás clasificadores. Esta característica puede funcionar muy bien en ciertos casos, pero también puede ser una de las principales desventajas de este tipo de ensambles; la cual sucede cuando se hace regresión y las predicciones obtenidas por cada tipo de algoritmo son muy diferentes entre si, provocando que al promediarlas se obtenga una predicción alejada las predicciones individuales, pudiendo ser mucho peor la predicción del ensamble que las de los clasificadores individuales.

5.2 Experimentos con datos simulados

El primer enfoque para resolver el problema de la identificación automática de poblaciones estelares en espectros galácticos, consistió en obtener la predicción a partir de ensambles de clasificadores, donde se entrenaron clasificadores independientes tomando como atributos subconjuntos distintos de anchos de las cinco líneas de la región de interés del espectro.

Se hicieron pruebas con dos ensambles homogéneos y uno heterogéneo, tomando como algoritmos de base las redes neuronales artificiales de tipo feedforward-backpropagation y regresión lineal localmente ponderada.

5.2.1 Ensamble homogéneo de redes neuronales artificiales

Un ensamble homogéneo de redes neuronales, está compuesto por un conjunto de clasificadores cuyo algoritmo de clasificación es una red neuronal artificial. Cada red debe de ser entrenada independientemente y se espera que produzcan predicciones diferentes una de otra. La manera más fácil de lograr esto, es entrenar cada red con un conjunto diferente de datos y/o de atributos.

Las redes neuronales son aptas para problemas tanto de regresión o de clasificación, por lo tanto sus salidas se combinan de acuerdo al problema que se quiera resolver.

En este primer enfoque para resolver el problema, se usaron dos tipos de ensamble, en el primero, se abordó el problema como dos sub-problemas, uno de clasificación y otro de regresión. La tarea de clasificación consistió en encontrar las tres edades de las poblaciones estelares, donde se contó con tres conjuntos de clases $e_1 = e_3 = \{1, 2, 3, 4, 5\}$ y $e_2 = \{1, 2, 3, 4\}$. En este caso, se entrenaron 5 clasificadores de redes neuronales con diferentes conjuntos de datos y diferentes subconjuntos de atributos, en este caso se utilizaron 50 de los 60 atributos para entrenar cada clasificador. Cada red neuronal feedforward, estaba compuesta por 50 neuronas en la capa de entrada, es decir, los 50 atributos del subconjunto relacionado con esa red; 42 neuronas ocultas y 14 neuronas de salida representando las 14 edades posibles, el valor mayor de los primeros 5 elementos representaba la clase del conjunto e_1, el valor mayor desde el elemento 6 al 9 representaba la clase del conjunto e_2 y de igual manera, el valor mayor desde el elemento 10 al 14 representaba la clase del conjunto e_3. Como se puede apreciar, en este experimento importaban únicamente las posiciones de los tres números mayores y no propiamente su valor. Para el problema de regresión se usaron 5 clasificadores entrenados con las mismas restricciones que en el problema anterior, pero la arquitectura de la red neuronal fue: 50 neuronas de entrada, 42 ocultas y 3 en la capa de salida, los tres valores de la predicción, representaban la proporción en la que cada población estaba presente en los espectros; siendo el primer valor el correspondiente a la población joven, el segundo a la intermedia y el tercero a la vieja.

La segunda manera para llegar a una solución por medio de un ensamble homogéneo de redes neuronales, consistió en tratar el problema como uno de regresión múltiple, donde la predicción estuviera formada por un vector de 14 elementos. Donde la posición de sus elementos representaba una de las 14 edades correspondiente a esa posición y el valor de cada elemento estaba relacionado con la proporción de la edad asociada.

En este caso, la arquitectura de la red consistió de 50 neuronas en la capa de entrada, 42 neuronas en la capa intermedia y 14 neuronas en la capa de salida.

	Error medio cuadrático	Tiempo total de entrenamiento	Tiempo de predicción por instancia
Enfoque 1	0.0428	376 min.	1.3 seg.
Enfoque 2	0.0432	200 min.	0.8 seg.

Tabla 5.1: Desempeño del ensamble de redes neuronales para datos sintéticos

Los resultados obtenidos en datos sintéticos, mostraron que la precisión fue muy similar, siendo ligeramente superior cuando la predicción se obtiene de dos ensambles. El tiempo requerido por los ensambles separados fue mayor, por lo tanto en futuras pruebas se decidió utilizar el enfoque del ensamble único de regresión múltiple para los 14 parámetros.

Tomando en cuenta lo anterior, el problema computacional se traduce en: encontrar un vector de $\overrightarrow{predicciones} \in R^{1\times 14}$, con las mismas restricciones que el vector de $\overrightarrow{edades}$ de la sección 2.3.2, tal que al multiplicarlo por la matriz de modelos $M_{modelos}$ produzca un espectro reconstruido donde se minimice el ajuste en el perfil de las 5 líneas de interés contra las líneas

del espectro original, y además, ese espectro reconstruido sea tan similar al original, como el que determinaría una búsqueda exhaustiva. En el resto de los experimentos, éste será el problema que se pretenda resolver.

5.2.2 Ensamble homogéneo de regresión lineal localmente ponderada

El ensamble homogéneo de regresión lineal, consistió también de cinco clasificadores entrenados de manera independiente, pero usando regresión lineal localmente ponderada como algoritmo base. Para obtener clasificadores diferentes, además de variar el conjunto de datos usado en el entrenamiento, también se varió la cantidad de vecinos empleados para construir el modelo (se usaron valores desde 100 hasta 500 vecinos). Un esquema general de la estructura de este ensamble se muestra en la figura 5.1.

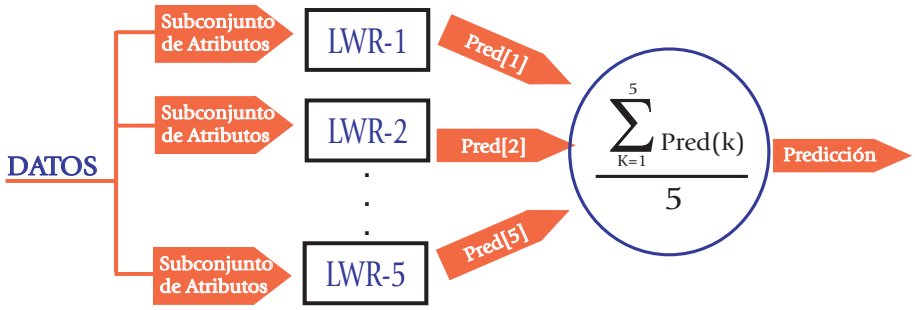


Figura 5.1: Ensamble homogéneo de regresión lineal localmente ponderada LWR-E

Error medio cuadrático	Tiempo total de entrenamiento	Tiempo de predicción por instancia
0.043	0 min.	1.5 seg.

Tabla 5.2: Desempeño del ensamble de LWR para datos sintéticos

Los resultados que se obtuvieron usando este ensamble, muestran un tiempo de entrenamiento mínimo, esto se explica porque el algoritmo de base LWR es un método de aprendizaje basado en instancias (ver la sección 3.3.1), por lo tanto, el entrenamiento consiste únicamente en almacenar los datos para crear los modelos en tiempo de predicción.

5.2.3 Ensamble heterogéneo combinando redes neuronales y regresión lineal localmente ponderada

En este caso se combinaron por promedio, los resultados obtenidos por 2 clasificadores de redes neuronales artificiales y 2 clasificadores que usan regresión lineal localmente ponderada. Cada clasificador se entrenó usando conjuntos de datos diferentes, así cómo subconjuntos de atributos distintos; los resultados presentados, son los obtenidos usando validación cruzada en diez corridas.

El tiempo de entrenamiento y el de predicción, es comparable con el producido por los experimentos con ensambles homogéneos, pero la precisión fue peor en un 2% que la obtenida

en los experimentos previos.

Error medio cuadrático	Tiempo total de entrenamiento	Tiempo de predicción por instancia
0.044	80 min.	1.1 seg.

Tabla 5.3: Desempeño del ensamble heterogéneo de ANN y LWR para datos sintéticos

Aunque generalmente, en un ensamble heterogéneo favorece la diversidad de los clasificadores para conseguir predicciones más precisas. En este caso las redes neuronales producen una predicción muy diferente a la obtenida mediante regresión lineal, por lo tanto al promediar las predicciones, se llega a un resultado más alejado de la realidad incluso que el producido por los clasificadores individuales.

Mediante estos experimentos, se concluyó que un ensamble heterogéneo tradicional no es una buena opción para el tipo de datos que se tienen para este problema particular, pero el comportamiento tan diferente de los clasificadores se consideró un hallazgo importante para mejorar el desempeño de los ensambles posteriores.

5.3 Consideraciones adicionales

Los resultados obtenidos al analizar los datos sintéticos fueron aceptables, especialmente en el caso de los ensambles homogéneos. Aunque el ensamble heterogéneo no produjo resultados tan buenos como los ensambles anteriores, ese experimento ayudó a identificar el comportamiento particular que presentan los algoritmos empleados, encontrando algunas regularidades en cuanto a qué tipo de ejemplos son los que tienden a clasificarse mejor y cuales producen la mayoría de los problemas.

Los experimentos que se explicaron en este capítulo también se aplicaron a los datos reales, los ajustes de las líneas presentaron un error entre 0.323 y 0.357. Pero aún así, estos ajustes no se consideran buenos, ya que un buen ajuste en el perfil de las líneas no se reflejó en una correcta reconstrucción del espectro, donde se obtuvieron errores desde 0.411 hasta 0.467.

Se concluyó que los perfiles poco definidos de los espectros observacionales, así como altos niveles de ruido fueron algunas causas del mal desempeño de los ensambles. Pero también se observó que los clasificadores presentaron un comportamiento que se puede definir como de sobreajuste. En casi todos los casos sintéticos se logró obtener un espectro reconstruido muy similar al original, pero esta habilidad no se conservó tratándose de datos reales, determinando así, que los ensambles no pudieron adaptar el comportamiento general de los espectros para obtener buenas predicciones.

Capítulo 6

EL ENSAMBLE DE DECISIÓN JERÁRQUICA Y EXPERIMENTOS CON DATOS REALES

6.1 Motivación

El comportamiento que se pudo observar a partir de los datos simulados con que se entrenó a los ensambles, muestra que el desempeño de los clasificadores tiene una regularidad, y es que los datos se pueden agrupar en 2 o 3 grupos bien definidos, siendo en todos los casos uno de estos grupos, el que presenta un resultado notablemente inferior en cuanto a su predicción. Este fenómeno se puede apreciar en la figura 6.1, donde, en el recuadro de la derecha se muestran los espectros agrupados usando k-means, aquí los ejemplos rojos se clasificaron mucho mejor que los ejemplos azules. En el recuadro de la derecha se muestra la distribución de los ejemplos para los diferentes valores de la población intermedia.

A partir de estas observaciones, se concluyó que cuando se entrena un clasificador con esos datos, éste tiende a aprender mejor los patrones que presenta el grupo mayor, descuidando el proceso de aprendizaje que presenta el grupo menor. Por lo tanto, se decidió usar un enfoque, donde el aprendizaje hiciera uso de las regularidades de los datos, clasificando en la primera fase los espectros con patrones regulares que son fáciles de identificar y dejando para fases posteriores clasificadores especializados en los datos que presentan problemas.

De esta manera, se planteó la necesidad de crear un algoritmo que fuera capaz de tomar en cuenta estos requerimientos, desarrollando así, el ensamble de decisión jerárquica (HDE). El cual aprende patrones de manera modular, y va tomando las decisiones a través de una ruta en el árbol resultante (ver figura 6.2).

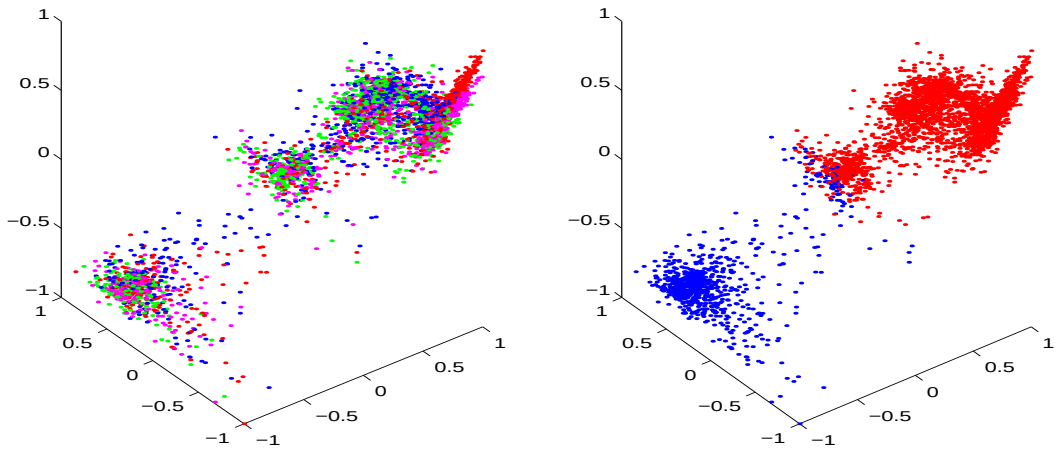


Figura 6.1: Distribución de los espectros para la población intermedia y usando k-means para agruparlos.

6.2 El Ensamble de decisión jerárquica (HDE)

El ensamble de decisión jerárquica presenta una estructura de árbol (ver figura 6.2), como tal, se va formando un nodo en cada paso, el cual se expande hacia un nivel inferior por medio de sus hijos o se define como nodo hoja, es decir, sin descendientes.

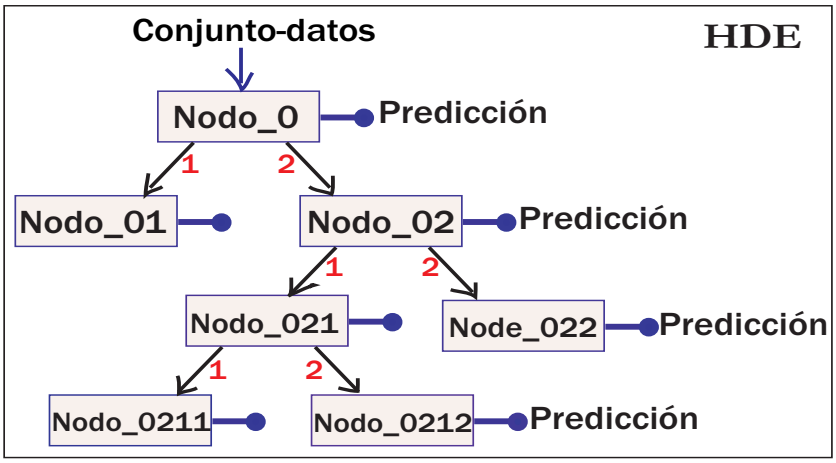


Figura 6.2: Estructura de árbol del Ensamble de Decisión Jerárquica (HDE)

El proceso de entrenamiento se realiza para cada nodo (ver figura 6.3), y es como sigue:

- Se tiene como entrada un conjunto de instancias, el cual se separa en conjunto de entrenamiento y conjunto de prueba.
- Se eligen ciertos atributos que presumiblemente ayudan a discriminar el comportamiento de los datos. Usando el conjunto correspondiente se entrena un clasificador, después se prueba su capacidad de generalización mediante el conjunto de prueba.

- Usando la predicción obtenida se reconstruyen los espectros correspondientes, estos se comparan con los espectros del conjunto original.
- Todos aquellos espectros cuya diferencia con los reconstruidos sea menor a un cierto umbral, se separan del conjunto de datos; mientras que los espectros reconstruidos cuya diferencia sea mayor se agregan al conjunto de datos.
- Si el nuevo conjunto es lo suficientemente grande, se divide en dos grupos, cada uno de los cuales pasa como entrada para un nodo hijo respectivamente.
- Este proceso se repite hasta que el árbol alcanza una longitud determinada, o hasta que no hay nodos que expander.

El pseudocódigo del entrenamiento se muestra en la tabla 6.1.

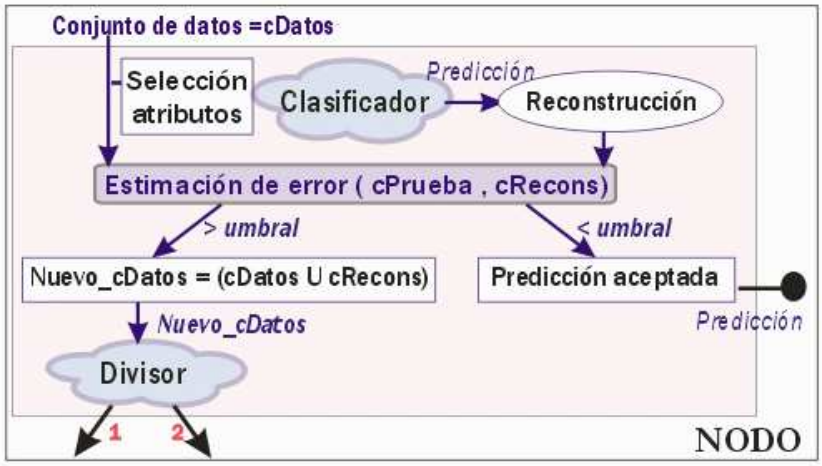


Figura 6.3: Nodo interno del HDE

El proceso de predicción se hace a partir del modelo generado en el entrenamiento. Al árbol se le da como entrada el conjunto de datos de los cuales se quiere predecir sus parámetros. Cada instancia recorre alguna de las rutas del árbol, en cada nodo, su núcleo clasificador obtiene cierta predicción, con la cual se reconstruye el espectro correspondiente y se calcula el error, si este error es menor a un umbral, se acepta la predicción generada hasta ese punto de la ruta, si no, el núcleo divisor determina la ruta que deberá seguir dicha instancia. Las predicciones y errores que genera cada uno de los nodos en la ruta de la instancia se guardan en una memoria temporal. La decisión final se puede hacer mediante tres criterios:

1. Elegir la predicción con la cual se obtiene el error mínimo.
2. Combinar las predicciones que producen los errores mínimos.
3. Formar una distribución de probabilidad usando todas las predicciones y sus errores.

1. Definir las funciones:
   - f: parámetros → datos originales
   - features: datos originales → espacio de características
   - RMS: calcula el error RMS entre 2 matrices
   - classify: Obtiene la clasificación para cada instancia en $test_{set}$ usando el clasificador entrenado
   - clustering: Agrupa $data_{set}$ en n grupos
2. Definir los parámetros:
   - $data_{set}$: es el conjunto de características
   - threshold: umbral para aceptar una predicción
   - n: número de grupos
   - id: etiqueta del nodo

HDE$(data_{set}, threshold, n, id)$
`inicia`
  `dividir` C_{datos} `en` $C_{entrena}$ `y` C_{prueba}
  `entrenar el clasificador usando` $C_{entrena}$
  $pred =$ **classify**$(clasificador, C_{prueba})$
  `reconstruir` **f**$(pred)$
  $error =$ **RMS**$(\mathbf{f}(C_{datos}), \mathbf{f}(pred))$
  $I_{menos} = error < umbral$
  $I_{mas} = error \geq umbral$
  $C_{datos} = C_{datos} \bigcup C_{datos}[I_{mas}]$
  $C_{datos} = C_{datos} \setminus$ **features**$(\mathbf{f}(pred[I_{menos}]))$
  `si` $size(C_{datos}) > 0$ `entonces`
    $[grupo_1, \ldots, grupo_n] =$ **clustering**(C_{datos}, n)
    `para` $i = 1$ `hasta` n `hacer`
      HDE$(grupo_i, umbral, n, id + i)$
    `termina`
  `termina`
`termina`

Tabla 6.1: Algoritmo HDE

En muchas ocasiones es preferible tener una predicción estadística de los parámetros que se quieren determinar, pero en este trabajo se optó por hacer la predicción mediante un vector único, de esta manera se puede hacer una comparación directa con los otros métodos.

A continuación se explica con mayor detalle cada parte del funcionamiento de los nodos internos que forman el árbol.

6.2.1 Selección de atributos

El problema de seleccionar un subconjunto de atributos que sean los mejores para clasificar un conjunto de instancias en ciertas clases, ha sido ampliamente estudiado [20].

Considerando un conjunto X de n instancias formado por m atributos, este conjunto se puede agrupar en c clases o grupos disjuntos, se pueden seleccionar m_f atributos, con $m_f < m$ los cuales ayuden a discriminar mejor qué elementos pertenecen a cada grupo. Esto se hace mediante el discriminante de fisher (ver [14]) donde los atributos seleccionados mantienen una mínima dispersión dentro de los grupos y una separación máxima entre las medias de los mismos.

Sea $X_j \in \mathrm{R}^{n_j \times m}$ la matriz de instancias que pertenecen a la clase j, donde $\overrightarrow{\mu}_j$ es su matriz de medias correspondiente y $\overline{\mu}$ es la matriz de medias totales, n_j es el número total de instancias en la clase j y c es el número total de clases .

Se define $E \in \mathrm{R}^{m \times m}$ como la variación entre las clases

$$E = \sum_{j=i}^{c} n_j \cdot (\overrightarrow{\mu}_j - \overline{\mu})(\overrightarrow{\mu}_j - \overline{\mu})' \tag{6.1}$$

También se define $I \in \mathrm{R}^{m \times m}$ como la variación al interior de las clases

$$I = \sum_{j=1}^{c} \sum (X_j - \overrightarrow{\mu}_j)(X_j - \overrightarrow{\mu}_j)' \tag{6.2}$$

Entonces, el cociente que se tiene que maximizar está dado por: $F = \frac{E}{I}$.

En esta etapa, se calculan las variaciones intra y entre clases, se favorecen aquellos atributos que presentan una variación entre clases grande y variación intra clases pequeña. Los atributos se ordenan de acuerdo a la ponderación que reciban de cumplir lo anterior, y se seleccionan los m_f mejores atributos.

Se decidió aplicar esta fase en cada nodo, debido a que los atributos que mejor clasifican en cada nivel, varían dependiendo del tipo de ejemplos que se tengan.

6.2.2 Núcleo clasificador

El núcleo clasificador se encarga de realizar la predicción de los parámetros de las instancias, de este núcleo depende la capacidad de generalización del nodo.

Para realizar la predicción, tiene que generar y entrenar un modelo de aprendizaje automático, este modelo se guarda para realizar las futuras predicciones, es decir, cada núcleo

clasificador tiene asociado un modelo de aprendizaje junto con los parámetros que se requieran para su correcto funcionamiento.

El entrenamiento del nodo se realiza como sigue:

1. Dividir el conjunto en prueba y entrenamiento
2. Entrenar un clasificador
3. Guardar el modelo asociado con el clasificador
4. Probar la capacidad de generalización usando el modelo y el conjunto de prueba.

El núcleo clasificador es independiente del algoritmo de aprendizaje automático que se utilice, sólo es necesario guardar el modelo generado en el entrenamiento.

De acuerdo al problema en cuestión, este núcleo puede usar cualquier algoritmo de aprendizaje automático, sólo depende de la tarea que vaya a realizar, ya sea regresión, regresión múltiple o clasificación.

En este trabajo, se usaron redes neuronales artificiales y regresión lineal localmente ponderada, se eligieron estos algoritmos porque se realizó regresión múltiple, y los anteriores son algoritmos que se desempeñan bien en este tipo de problemas.

En tiempo de predicción, el núcleo clasificador sólo se utiliza para generar la predicción de las instancias mediante el modelo que se almacenó en cada nodo.

6.2.3 Estimación del error y regeneración del conjunto de datos

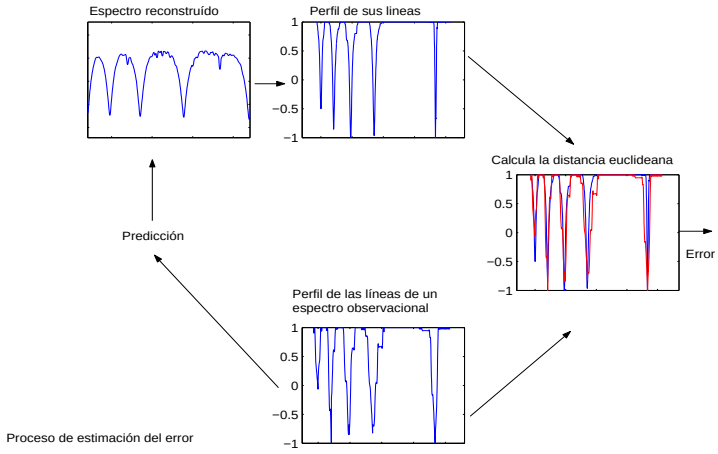
La estimación del error de las predicciones es una fase crucial en la generación del árbol, ya que de esto depende qué nuevos patrones se van a aprender en los niveles inferiores.

En esta etapa, se toman las predicciones hechas por el núcleo clasificador, con ellas se reconstruyen los espectros ideales que se forman con esos parámetro. Después se caracterizan, y la caracterización de los espectros reconstruidos se compara con la caracterización de los espectros originales mediante la distancia euclidiana (ver figura 6.4)

Se calcula la distancia en cada par de puntos entre los espectros caracterizados, y se calculan dos conjuntos de indices, donde I_{mas} representa a los elementos cuya distancia promedio sea mayor a un umbral predefinido, e I_{menos} representa a los elementos cuyo error sea menor que el umbral.

Con estos conjuntos de indices se cuenta con una métrica de la eficacia del nodo, éste será más eficaz en la medida en que I_{menos} sea mayor. Esta métrica de eficacia no siempre beneficia al comportamiento global del árbol; como se explicará en la sección 6.3.3, a veces es necesario sacrificar la eficacia individual de ciertos nodos para incrementar la eficacia global del árbol.

La regeneración del conjunto de datos, consiste en eliminar del conjunto de datos, a todos aquellos espectros cuya clasificación se ha aceptado en el nivel actual y además, en tiempo de entrenamiento sirve para agregar nuevos espectros al conjunto de entrenamiento, que son los que presentan más problemas para el clasificador.



(a) Estimación del error en la reconstrucción del espectro

(b) El error se calcula entre cada par de puntos

Figura 6.4: Estimación del error

Este paso ayuda a evitar el sobreajuste, el cual podría ser causado por entrenar un clasificador usando un número muy reducido de datos, también es útil para enfocar a los clasificadores siguientes sólo en los ejemplos que resultan difíciles de clasificar, así como en proveer al conjunto de datos de espectros que son los principales causantes de errores en ese nivel.

El proceso es como sigue:

1. Quitar del conjunto de datos los ejemplos del conjunto de prueba cuyo índice sea elemento de I_{menos}

2. Agregar al conjunto de datos, los ejemplos reconstruidos cuyo índice esté en I_{mas}

El nuevo conjunto de datos estará conformado sólo por los ejemplos que presentan más problemas para los clasificadores anteriores.

6.2.4 Núcleo divisor

El núcleo divisor, representa la prueba sobre atributos que determina cómo se va a establecer la división de instancias en cada nodo de un árbol.

Este núcleo determina la forma que tendrá el árbol. En el HDE, a diferencia de los algoritmos tradicionales, la manera de formar el discriminante es a través de un algoritmo de aprendizaje no supervisado. El número de ramas que se forman en cada nodo, no depende del número de clases del problema, ni de ninguna otra característica intrínseca de los datos. El número de ramas es un parámetro determinado por el usuario, pero es recomendable usar un número pequeño para no desgastar al conjunto de datos, además de obtener árboles más pequeños.

Como se mencionó anteriormente, el núcleo divisor es independiente del algoritmo, el único requisito es que este algoritmo sea capaz de formar grupos a partir de datos no etiquetados. Se

decidió usar este criterio para expandir el nodo, porque un algoritmo con estas características puede encontrar las regularidades existentes en los datos, siendo estas mismas regularidades y diferencias las que ayudan a formar clasificadores más especializados en el siguiente nivel.

En tiempo de predicción, la función de este núcleo es indicarle a la instancia en cuestión que ruta deberá seguir hacia el siguiente nivel.

En todos los experimentos realizados con HDE, se utilizó k-means como algoritmo de agrupamiento en el núcleo divisor, y se estableció a 2 el número de grupos. Cada uno de los dos grupos que se forman, pasan al siguiente nivel para generar dos nuevos nodos.

6.2.5 Poda en HDE

Como en la mayoría de los algoritmos con estructura de árbol es posible hacer poda en ciertas ramas. Se aplica la poda a una rama cuando en uno de sus nodos el núcleo divisor produce dos grupos con una cantidad muy desigual de elementos, donde uno de los grupos tiene muy pocos elementos para realizar un buen proceso de aprendizaje. En este caso existen dos opciones:

1. Truncar la rama y convertir el nodo en un nodo hoja.
2. Eliminar un nivel de la rama, esto se hace ignorando el nodo en cuestión y reemplazándolo por el nodo hijo que se forma a partir de la mayor cantidad de elementos.

En este trabajo se presentan resultados usando el segundo criterio de poda, reemplazando un nodo con desempeño pobre por uno que puede ser mejor. Se seleccionó este enfoque debido a que la longitud de las ramas del árbol no repercute de manera negativa en la decisión final, ya que ésta se va tomando a lo largo de toda la ruta y no sólo por los nodos hoja. Por lo tanto una rama larga sólo puede ayudar a mejorar la predicción.

6.2.6 Comparación con otros algoritmos

La naturaleza general del HDE no corresponde directamente con ninguno de los algoritmos que se explicaron en la sección 3.3, pero de manera particular, si presenta similitudes con varios de estos algoritmos. A continuación se hace una comparación con algunos métodos.

El ensamble de decisión jerárquica combina la potencia de los ensambles con las ventajas del aprendizaje modular de los árboles de regresión.

Ensambles de clasificadores. El HDE se puede comparar con los ensambles de clasificadores, en cuanto a que usa diversos clasificadores entrenados de manera independiente y cuyas predicciones se pueden combinar para obtener una predicción única.

Árboles de regresión. El HDE puede realizar tareas tanto de regresión como de clasificación, dependiendo del algoritmo que se elija para formar el núcleo clasificador. Además de que los algoritmos usados por HDE permiten realizar tareas de regresión multiple, lo cual no se ha implementado en los árboles tradicionales de regresión.

Algoritmos con estructura de árbol. La mejor categoría en la que se puede definir el HDE es precisamente en la de los algoritmos con estructura de árbol, aunque también presenta diferencias significativas en cuanto a la implementación común de estos métodos.

La principal diferencia consiste en que, todos los algoritmos revisados que presentan estructura de árbol, hacen la predicción únicamente en los nodos hoja, mientras que el HDE va haciendo predicciones durante todo el recorrido de la instancia por el árbol, y al final elige la que proporciona mejores resultados, aunque también puede determinar una solución mediante la combinación de las mejores predicciones a manera de ensamble.

6.3 Experimentos con datos reales

En todos los experimentos realizados, se creó cada HDE con 5400 espectros iniciales, a los cuales se les agregaron efectos de ruido y atenuación; en los subsiguientes niveles del árbol, los espectros que se generaron para realizar la medición del error y modificar el conjunto final de datos estaban limpios. Además, la profundidad del árbol se limitó a 5, para mantener igualdad de condiciones con los experimentos anteriores y para mantener la eficiencia.

El núcleo divisor en los tres casos, estuvo formado por k-means como algoritmo base. A continuación se presentan los experimentos realizados con este algoritmo, así como las variaciones que se implementaron.

6.3.1 HDE con núcleo clasificador de redes neuronales

El ensamble de decisión jerárquica con núcleo de redes neuronales artificiales, mostró un buen desempeño en cuanto a precisión, los resultados fueron mejores a los obtenidos mediante las técnicas tradicionales de ensambles

Error medio cuadrático	Tiempo total de entrenamiento	Tiempo de predicción por instancia
0.306	480 min.	$[1.60 \pm 0.4]$ seg.

Tabla 6.2: Desempeño del HDE con núcleo clasificador de redes neuronales para datos reales

6.3.2 HDE con núcleo clasificador de regresión lineal localmente ponderada

Se observó que este algoritmo, usando núcleos de regresión lineal, tiende a producir árboles mayores, tanto en anchura, como en profundidad. Por los resultados obtenidos, se encontró que el HDE con regresión lineal locamente ponderada presentó más dificultades llegar a una buena solución con sólo 5 niveles que el HDE-ANN. Esta característica se utilizó para mejorar el desempeño del HDE en el siguiente experimento.

Error medio cuadrático	Tiempo total de entrenamiento	Tiempo de predicción por instancia
0.298	115 min.	$[2.05 \pm 0.4]$ seg.

Tabla 6.3: Desempeño del HDE con núcleo clasificador de regresión lineal para datos reales

6.3.3 Ensamble heterogéneo de decisión jerárquica

Este ensamble se diseñó para aprovechar las características que presentan los clasificadores de regresión lineal en las primeras etapas, generando una mayor cantidad de ejemplos que se agregan al conjunto de datos en fases tempranas del entrenamiento, y obteniendo mejor precisión en la clasificación en las últimas fases mediante redes neuronales artificiales. Es decir, se entrenaron clasificadores con núcleos de regresión lineal en los primeros niveles del árbol y clasificadores con núcleos de redes neuronales en los últimos niveles. Este esquema se puede ver en la figura 6.5.

Error medio cuadrático	Tiempo total de entrenamiento	Tiempo de predicción por instancia
0.291	355 min.	$[1.75 \pm 0.5]$ seg.

Tabla 6.4: Desempeño del HDE heterogéneo

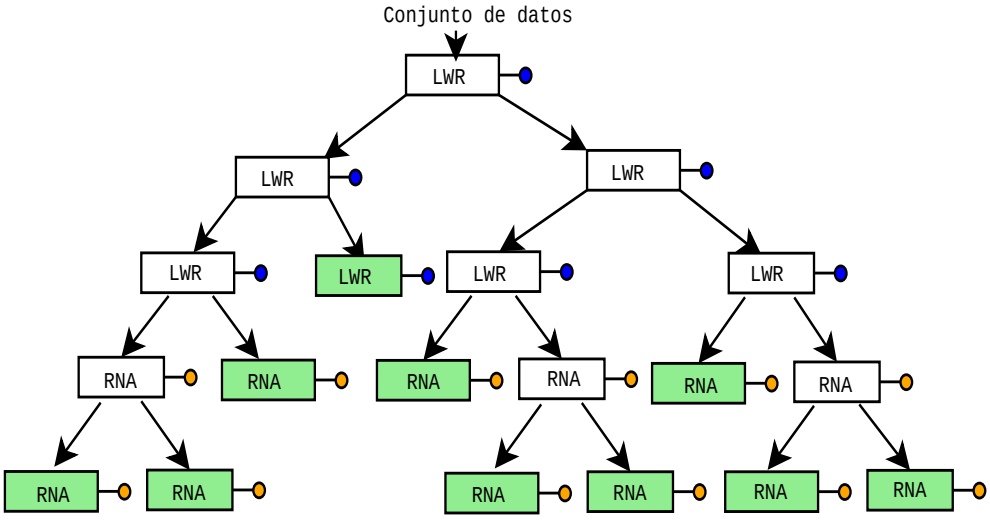


Figura 6.5: Ejemplo de un árbol producido por el ensamble heterogéneo de decisión jerárquica

El ensamble heterogéneo de decisión jerárquica hace uso del comportamiento que se identificó en la sección 5.2.3, encontrando que las predicciones obtenidas por un tipo de clasificador y el otro son muy diferentes entre sí. Los árboles de redes neuronales tienden a ser más pequeños y no muy anchos, mientras que los de regresión lineal son anchos y largos. Por lo tanto, los clasificadores de regresión lineal generan más ejemplos para el conjunto de entrenamiento que las redes neuronales, así se producen árboles más anchos que cuando se usan sólo clasificadores de redes neuronales. En este ensamble, los primeros nodos proveen de una cantidad mayor de

ejemplos relevantes a los niveles inferiores.

6.4 Consideraciones adicionales

En esta sección se mostró un algoritmo con estructura de árbol, pero que va haciendo predicciones a lo largo de toda la ruta que pudiera seguir una instancia. Al final la decisión se puede tomar de diversas maneras, pero en este trabajo, esta decisión se deja a un sólo clasificador, el tiempo de entrenamiento con una cantidad relativamente grande de datos es algo mayor que en otros métodos, pero no intratable, mientras que la capacidad de generalización que alcanza, incluso mediante uno solo de sus clasificadores especializados, es mayor que la que alcanzan otros métodos combinando las predicciones de varios clasificadores.

El desempeño que mostró el HDE usando datos reales es muy superior al obtenido en experimentos anteriores, ya que produce mejores ajustes de los perfiles y además también produce mejores espectros reconstruidos.

En el último experimento se aprovechó la capacidad de HDE de integrar diferentes algoritmos en cada nodo. Así, se utilizaron dos algoritmos que se comportan de manera diferente con los datos, produciendo un clasificador más robusto y con muy buena capacidad de generalización. El HDE heterogéneo obtuvo resultados muy diferentes a los producidos por el ensamble heterogéneo de la sección 5.2.3, donde se usaron los mismos algoritmos, pero el resultado en ese caso fue incluso peor que el de los clasificadores individuales.

Capítulo 7

RESULTADOS

7.1 Experimentos

Para determinar la edad de las 3 principales poblaciones estelares en espectros simulados y observacionales se utilizaron diversos métodos combinados en ensambles, estos métodos se describieron en los capítulos 5 y 6, para efecto de simplicidad, en este capítulo se usa la siguiente notación:

ANN-E Ensamble homogéneo de redes neuronales artificiales

LWR-E Ensamble homogéneo de regresión lineal localmente ponderada

ANN-LWR Ensamble heterogéneo de redes neuronales artificiales y regresión lineal localmente ponderada

HDE-ANN Ensamble de decisión jerárquica con núcleo de redes neuronales artificiales

HDE-LWR Ensamble de decisión jerárquica con núcleo de regresión lineal localmente ponderada

HET-HDE Ensamble heterogéneo de decisión jerárquica

ORIGINAL Parámetros originales

EXHAUSTIVA Determinación de parámetros por medio de una búsqueda exhaustiva [1]

En casi todos los experimentos, las pruebas se realizaron con 3600 espectros simulados y 1155 espectros reales. Se decidió utilizar 3600 espectros simulados tanto para prueba, cómo para entrenar cada clasificador individual, debido a que con este número es posible cubrir de manera aceptable el espacio de la solución. Las características particulares de estos datos se resumen en la tabla 7.1. Únicamente en los resultados mostrados en la tabla 7.2 se utilizó un número diferente de ejemplos.

[1] Debido a la gran cantidad de tiempo que consume, se probó únicamente para 161 datos sintéticos y 100 reales.

Característica	Simulados	Reales
Número	3600	1155
Resolución	$0.3\mathring{A}$	$1\mathring{A}$
Rango espectral	[3435-8315]$\mathring{A}$	[3000-6900]$\mathring{A}$
Líneas de emisión	No	Si
Ruido	10%	No determinado

Tabla 7.1: Resumen de las características de los datos de prueba

7.2 Comparación de resultados

Los resultados de los experimentos realizados se presentan en un orden lógico, primero se hace la comparación de los métodos de aprendizaje automático contra un enfoque de búsqueda exhaustiva, a través de estos resultados se concluye la utilidad del enfoque de aprendizaje automático. Una vez justificado este enfoque, se procede a comparar los resultados de eficiencia y precisión de los diferentes algoritmos de aprendizaje que se utilizaron en este trabajo.

En la tabla 7.2 se presentan los resultados obtenidos de promediar los métodos de aprendizaje automático y los producidos por una búsqueda semi-exhaustiva. Se dice que es semi-exhaustiva porque el espacio de búsqueda es demasiado grande para recorrerlo completamente[2]. En esta tabla se puede apreciar que el tiempo consumido por la búsqueda es muy grande comparado con los métodos de aprendizaje automático, además, debido a la incapacidad para recorrer todo el espacio, no es posible encontrar los parámetros óptimos, a pesar de la gran cantidad de tiempo consumido.

Métricas de desempeño	Búsqueda exhaustiva	Aprendizaje automático	Parámetros originales
Tiempo de predicción p/i	1830 seg.	1.46 seg.	-
Error en datos sintéticos	0.0542	0.0562	0.0538
Error en datos reales	0.291	0.308	-
Recuperación de parámetros en datos sintéticos	**Búsqueda exhaustiva**	**Aprendizaje automático**	**Parámetros originales**
Error 0 en $\overrightarrow{p}_1$	95%	95%	100%
Error 1 en $\overrightarrow{p}_1$	4%	4%	0%
Error 0 en $\overrightarrow{p}_2$	79%	77%	100%
Error 1 en $\overrightarrow{p}_2$	13%	9%	0%
Error 0 en $\overrightarrow{p}_3$	54%	51%	100%
Error 1 en $\overrightarrow{p}_3$	13%	15%	0%

Tabla 7.2: Comparación del comportamiento de la búsqueda exhaustiva contra un promedio de los métodos de aprendizaje automático para la misma cantidad de datos.

En este experimento, si se hubiera querido hacer la predicción de los 3600 espectros del conjunto de prueba de los datos sintéticos, una búsqueda exhaustiva hubiera tomado 76 días y 3 horas aproximadamente, mientras que con el método de aprendizaje automático más

[2]Los resultados presentados son de 161 datos sintéticos y 100 datos reales, tanto en los métodos de aprendizaje automático cómo en la búsqueda exhaustiva.

tardado, en el peor de los casos, sólo toma alrededor de 2 horas.

Debido a que el análisis de espectros se realiza típicamente sobre conjuntos desde 1000 hasta 100000 espectros, es necesario que el método empleado sea rápido, por lo que la búsqueda exhaustiva queda totalmente descartada, mientras que la capacidad de generalización y eficiencia de los algoritmos de aprendizaje automático se vuelven características muy apreciadas para resolver el problema en cuestión. Consolidando al aprendizaje automático como la mejor opción para determinar las poblaciones estelares de galaxias observacionales a partir de su espectro.

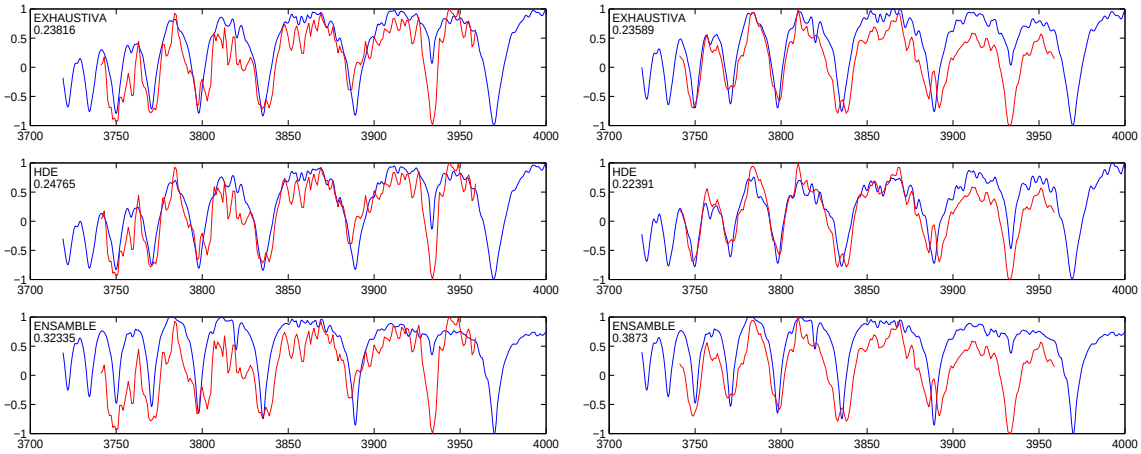
En la tabla 7.2 también se muestra la eficacia de los métodos para recuperar los parámetros originales para las tres poblaciones. Siendo $Error0$ el porcentaje de instancias que predijeron la misma edad que los parámetros originales y $Error1$ el porcentaje de instancias cuya predicción se encuentra a una edad de distancia del parámetro correcto. Se puede apreciar que los porcentajes son similares en los métodos de aprendizaje automático y en la búsqueda exhaustiva y que las edades viejas son más difíciles de predecir que las jóvenes. Esto se debe a que los niveles de ruido pueden volver a un espectro con ciertos parámetros, más parecido a un espectro limpio con parámetros diferentes.

La comparación con respecto a la exactitud de los parámetros fue factible hacerla en los espectros simulados debido a que se conocen de antemano las edades originales. En el caso de espectros observacionales no es posible hacerla, ya que estos espectros no están analizados. Tampoco es posible afirmar que los parámetros que encuentra la búsqueda exhaustiva son más correctos que los que encuentran los métodos de aprendizaje automático debido a que, como se ve en la figura 7.1(b), en algunos casos el error que produce la búsqueda exhaustiva es mayor al obtenido con algún método de aprendizaje automático. Aunque se tienen algunas diferencias, los parámetros obtenidos mediante la búsqueda exhaustiva y los del aprendizaje automático son similares, esto se puede observar en la figura 7.1.

Se dice que el desempeño de los métodos de aprendizaje automático es comparable, en cuanto a precisión, al de la búsqueda exhaustiva, porque el error en datos simulados aumenta solamente 3.6%, de igual manera, en datos reales este error aumenta en 5.8%. En términos reales, las diferencias de estos errores representan una pérdida en la precisión total del 0.1% y del 0.8% respectivamente. Estas cantidades son muy pequeñas, considerando que un espectro puede tener ruido del 10% o más. En cuanto a la recuperación de los parámetros, la diferencia de errores está entre el 2 y 3%, cantidades que también se consideran pequeñas.

A continuación se comparan los resultados obtenidos por los experimentos definidos en la sección anterior, se hace la comparación en cuanto a la eficiencia en tiempo de entrenamiento y de predicción (ver tabla 7.3), así como de la precisión obtenida por cada algoritmo (ver tabla 7.4).

El tiempo de entrenamiento requerido por el ensamble de regresión lineal es 0 debido a que usa un método de aprendizaje basado en instancias, donde el entrenamiento consiste únicamente en almacenar los ejemplos para utilizarlos posteriormente en la etapa de predicción, en la cual se puede ver un ligero aumento del tiempo consumido con respecto a los



(a) Ejemplo 1

(b) Ejemplo 2

p1	EXH	0	0	0	0.10	0
	HDE	0	0	0	0.38	0
	ENS	0	0	0	0	0.171
p2	EXH	0	0.10	0	0	
	HDE	0	0.13	0	0	
	ENS	0	0.40	0	0	
p3	EXH	0.79	0	0	0	0
	HDE	0.82	0	0	0	0
	ENS	0	0	0.42	0	0

(c) Parámetros encontrados para el ejemplo 1

p1	EXH	0	0	0	0	0.10
	HDE	0	0	0	0	0.21
	ENS	0	0	0	0	0.32
p2	EXH	0	0.30	0	0	
	HDE	0	0.33	0	0	
	ENS	0	0.42	0	0	
p3	EXH	0.59	0	0	0	0
	HDE	0.45	0	0	0	0
	ENS	0	0.24	0	0	0

(d) Parámetros encontrados para el ejemplo 2

Figura 7.1: Comparación de parámetros y ajustes entre la búsqueda exhaustiva, el HDE heterogéneo y el mejor ensamble tradicional aplicados a dos espectros reales

Algoritmo	Tiempo total de entrenamiento	Tiempo de predicción por instancia
ANN-E	200 min.	**0.8** seg.
LWR-E	**0** min.	1.5 seg.
ANN-LWR	80 min.	1.1 seg.
HDE-ANN	480 min.	**1.60** seg.
HDE-LWR	**115** min.	2.05 seg.
HET-HDE	355 min.	1.75 seg.

Tabla 7.3: Comparación de la eficiencia de los algoritmos

otros ensambles simples.

En el caso de los HDE, el tiempo de entrenamiento depende del número de clasificadores que sea necesario producir. Como la profundidad del árbol se restringió a 5, en el peor de los casos un HDE debería entrenar 62 clasificadores. En los experimentos realizados en este trabajo, en promedio cada HDE consistió de 30 nodos, requiriendo una mayor cantidad para HDE's formados con núcleos de regresión lineal (55 aproximadamente), que para los formados con núcleos de redes neuronales (11 aproximadamente). En cuanto al tiempo de predicción, cada instancia debía recorrer a lo más 5 clasificadores; pero al tiempo consumido en la predicción, se añade el tiempo necesario para determinar la ruta mediante k-means.

Aunque la precisión de los algoritmos en los datos sintéticos es muy similar, existen dife-

Algoritmo	RMS-Error en Datos Simulados	RMS-Error en Datos Reales	
	Ajuste de Perfiles	*Ajuste de Perfiles*	*Reconstrucción del espectro*
ANN-E	0.0451	0.348	**0.411**
LWR-E	**0.0435**	**0.323**	0.425
ANN-LWR	0.0447	0.357	0.467
HDE-ANN	0.0435	0.306	0.347
HDE-LWR	0.0437	0.299	0.413
HET-HDE	**0.0431**	**0.291**	**0.321**
ORIGINAL	**0.0414**	-	-

Tabla 7.4: Comparación de la precisión de los algoritmos

rencias significativas en los datos reales. Donde el HDE-HET superó en 10% al mejor ensamble tradicional en ajuste de perfiles y en 28% en cuanto al ajuste de espectros reconstruidos. Estas diferencias representan una ganancia del 2% y 4% en el error total del espectro respectivamente. Esta ganancia es muy importante, teniendo en cuenta que la búsqueda exhaustiva superó en 0.1% y 0.8% al promedio de clasificadores automáticos.

Una comparación visual de los ajustes producidos por los métodos de aprendizaje automático se puede apreciar en las figuras B.1, B.2 y B.3.

7.3 Consideraciones adicionales

Aunque el tiempo requerido para una búsqueda exhaustiva es constante y el tiempo consumido por el ensamble de decisión jerárquico está determinado en relación al número de nodos que debe de recorrer cada instancia en una ruta determinada, el tamaño del árbol está tan limitado, que es prácticamente despreciable comparado contra el tiempo que tarda la búsqueda exhaustiva.

Comparando los resultados obtenidos mediante la búsqueda exhaustiva contra los de aprendizaje automático, y considerando el tamaño de los conjuntos que típicamente se analizan en la astronomía, se concluyó la utilidad de los métodos de aprendizaje automático.

Los ensambles tradicionales presentan una precisión similar que el HDE en cuanto a los datos sintéticos pero el tiempo requerido es menor en los primeros. Esta aseveración podría concluir en que es preferible usar los ensambles simples, pero considerando los resultados en los datos reales, se puede observar que los métodos tradicionales no presentaron un buen grado de generalización, por otra parte, la diferencia en cuanto al tiempo es muy pequeña para ser tomada en cuenta como factor decisivo.

La generalización obtenida en datos reales, comparando no sólo el ajuste en el perfil de las líneas, sino el espectro reconstruido es mucho mejor en el caso de los resultados producidos en el ensamble de decisión jerárquica que en cualquier caso de los ensambles tradicionales.

El ensamble heterogéneo de decisión jerárquica produce menor error que todos los demás experimentos, y el ajuste que produce del espectro observacional se considera aceptable para determinar sus parámetros de edad.

Capítulo 8

CONCLUSIONES Y TRABAJO FUTURO

8.1 Consideraciones finales

La solución que se plantea en esta tesis, para el problema de encontrar las tres principales poblaciones estelares en galaxias reales, se dividió en dos etapas: caracterización de los espectros e identificación de poblaciones. A continuación se presentan algunos puntos importantes al respecto.

En la primera etapa de la solución, se decidió ajustar únicamente una región del espectro, en esta región se encuentra gran parte de la información relevante a las edades de sus poblaciones estelares. Una vez seleccionada la región de interés e identificado el perfil de las líneas, se usaron dos métodos para realizar la reducción de la dimension de los datos y encontrar características relevantes:

1. Análisis de componentes principales
2. Extracción de anchos en los perfiles de las líneas.

El método de PCA reconstruyó casi a la perfección los espectros sintéticos, pero al momento de recuperar espectros reales, produjo espectros con gran cantidad de ruido. Sin embargo, éste no fue el mayor problema, la principal desventaja consistió en que, en muchos casos, la reconstrucción no respeta la forma de las líneas principales, incluyendo su magnitud y la relación de una región a otra. Por lo tanto se concluyó que este método no es apropiado para manejar espectros observacionales.

Por otra parte, las características obtenidas mediante la extracción de anchos, permitieron considerar las variaciones existentes en los perfiles de las líneas de espectros con diferentes tipos de poblaciones estelares. Además estas características se mantuvieron similares tanto en los datos sintéticos como en los reales, permitiendo así que un clasificador entrenado con los anchos de datos sintéticos fuera capaz de generalizar y obtener predicciones aceptables a partir de datos reales.

En la segunda etapa de la solución, se usaron diversos algoritmos de aprendizaje para determinar las edades de las poblaciones estelares. En esta etapa se siguieron también dos enfoques:

1. Ensamble de clasificadores construido de la manera tradicional, donde las predicciones se obtienen por votación.

2. Ensamble de decisión jerárquica, que es un algoritmo de aprendizaje con forma de árbol, el cual utiliza otros algoritmos de aprendizaje supervisado y no supervisado, y que hace uso de las regularidades de los datos para entrenar clasificadores especializados, capaces de generalizar bien en ejemplos con características particulares.

En el primer enfoque, los resultados obtenidos al analizar datos sintéticos fueron aceptables, casi en todos los casos se logró obtener un espectro reconstruido muy similar al original, pero esta habilidad no se conservó tratándose de datos reales, ya que un buen ajuste en el perfil de las líneas no se reflejó en una buena reconstrucción del espectro.

Con el segundo enfoque, los resultados con datos sintéticos fueron similares a los ensambles del enfoque 2, sólo que el tiempo de entrenamiento es mayor. Sin embargo, aún con una cantidad relativamente grande de datos, no llega a ser intratable; mientras que la capacidad de generalización que alcanza, incluso mediante uno solo de sus clasificadores especializados, es mayor que la de los otros métodos combinando las predicciones de varios clasificadores. El desempeño que mostró el HDE usando datos reales es muy superior al obtenido en experimentos anteriores, ya que produjo mejores ajustes de los perfiles y además también mejores espectros reconstruidos.

Para comparar los resultados en cuanto a precisión, se usó una costosa búsqueda semi exhaustiva en el espacio de la solución. Presumiblemente esta búsqueda sería capaz de encontrar los parámetros que representan de manera óptima al espectro asociado. La precisión obtenida con este método fue casi tan buena como si se hubieran utilizado los parámetros originales, pero el tiempo requerido sobrepasó en gran mediada los límites de la eficiencia.

Los métodos de aprendizaje automático produjeron predicciones con una precisión muy similar a la de la búsqueda exhaustiva pero en un tiempo mucho menor, haciendo factible el análisis de grandes cantidades de datos en un tiempo razonable.

8.2 Conclusiones

Es posible determinar de manera aceptable, tanto la edad como la proporción de las tres principales poblaciones estelares que forman una galaxia. En esta tesis, el problema se resolvió de la siguiente manera:

- Caracterizando una región del espectro: extrayendo los anchos y perfiles de las líneas de la serie de Balmer y la línea K de calcio.

- Analizando los anchos extraídos mediante técnicas de aprendizaje automático.

Los resultados obtenidos en espectros simulados y observacionales, son únicamente 0.8% peores a los que podría obtener un astrónomo experto, o en este caso una búsqueda exhaustiva.

La exactitud en las predicciones de datos reales, depende de los modelos de síntesis de evolución estelar empleados para entrenar los clasificadores.

Los algoritmos propuestos pueden utilizarse para analizar grandes colecciones de espectros observacionales. Considerando lo siguiente:

- Los espectros pueden tener cualquier resolución entre 0.2 $\mathring{A}$ y 1$\mathring{A}$.
- La única región del espectro que es indispensable está entre 3750$\mathring{A}$ hasta 3960$\mathring{A}$
- Los efectos de ruido y atenuación no impactan decisivamente en la precisión de las predicciones.
- Es posible analizar espectros de galaxias que presentan grandes líneas de emisión.
- El tiempo consumido por los algoritmos propuestos es de aproximadamente 2 segundos por espectro. Es decir, que para analizar los 100,000 espectros del SDSS nuestros métodos tardarían aproximadamente dos días y medio, mientras que una búsqueda exhaustiva tardaría alrededor de 342 años.

Para resolver el problema planteado en esta tesis, se creó un método que combina diversos algoritmos de aprendizaje automático, el cual es capaz de aprender a clasificar instancias valiéndose de las regularidades en sus atributos. Este algoritmo se puede extender a otros problemas en dominios diferentes al de datos astronómicos.

8.3 Trabajo futuro

Como trabajo futuro, se plantea analizar los datos para encontrar otras características relevantes, así como otra manera de codificar los espectros. También se observa la posibilidad de incluir en la región de interés otras líneas, como son, algunas líneas de helio.

Otro aspecto que se trabajará a futuro, consiste en ampliar la forma de representar la solución del problema, obteniendo una distribución gaussiana de incidencias de las poblaciones en lugar de únicamente tres valores continuos.

Se planea también, utilizar otros algoritmos de aprendizaje supervisado y no supervisado para mejorar el desempeño del HDE, así como adaptarlo a la resolución de problemas en dominios diferentes.

Apéndice A

NOTACIÓN

Notación	Interpretación
$\overrightarrow{x}$	El vector x
X	La matriz X
$\overrightarrow{x}[i]$	El elemento i del vector x
$X[i,j]$	El elemento en la fila i, columna j de la matriz X
$f(a)$	La función f con el argumento a
$x \leftarrow y$	A x se le asigna y
$[\overrightarrow{x}, \overrightarrow{y}]$	Concatenación horizontal de $\overrightarrow{x}$ e $\overrightarrow{y}$
$[\overrightarrow{x}; \overrightarrow{y}]$	Concatenación vertical de $\overrightarrow{x}$ e $\overrightarrow{y}$

Tabla A.1: Notación

Abreviaturas	Significado
$\mathring{A}$	Angstroms $= 1 \times 10^{-13}$ de un metro
ANN	Redes Neuronales Artificiales
HDE	Ensamble de Decisión Jerárquica
LWR	Regresión Lineal Localmente Ponderada
PCA	Análisis de componentes principales
PC	Componentes principales
p/i	Por instancia
RMS	Error Medio Cuadrático

Tabla A.2: Abreviaturas

Apéndice B

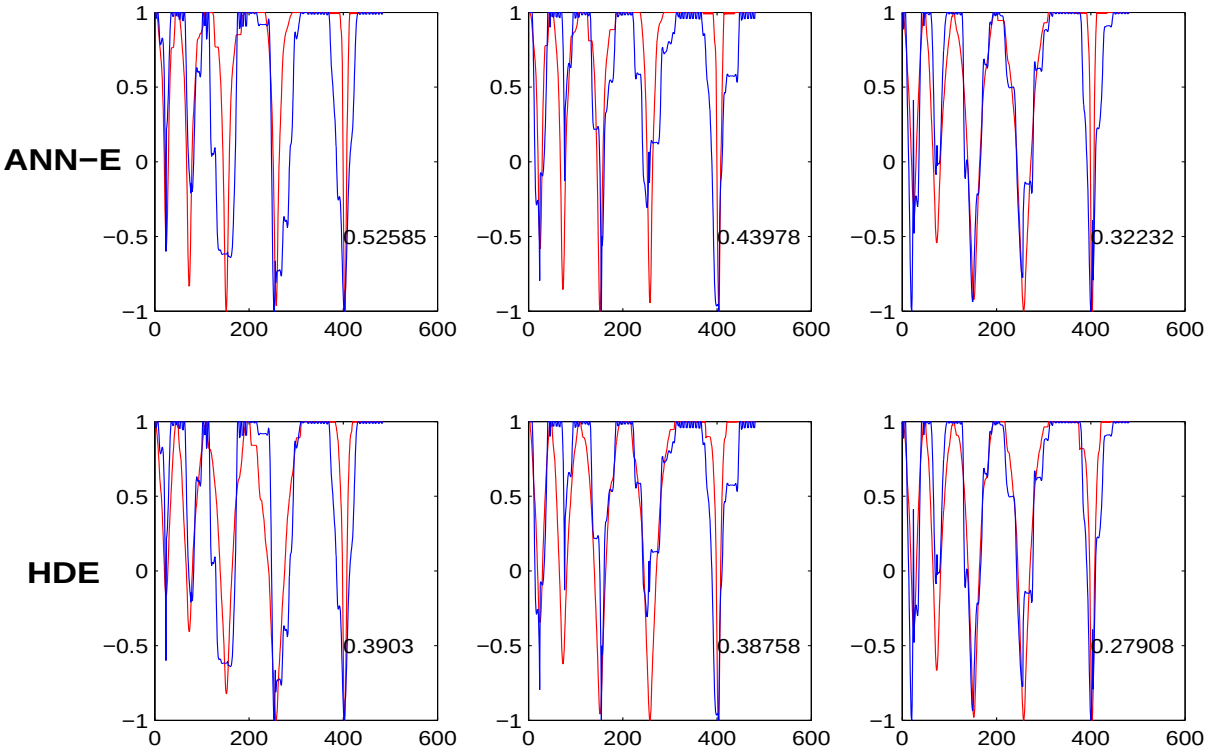
COMPARACIÓN DE LOS MÉTODOS DE APRENDIZAJE AUTOMÁTICO EN EL AJUSTE DE ESPECTROS REALES

En las figuras B.1, B.2 y B.3, se muestra la comparación de 3 distintos espectros, donde cada columna está relacionada con un espectro y cada fila está asociada con el desempeño de un algoritmo.

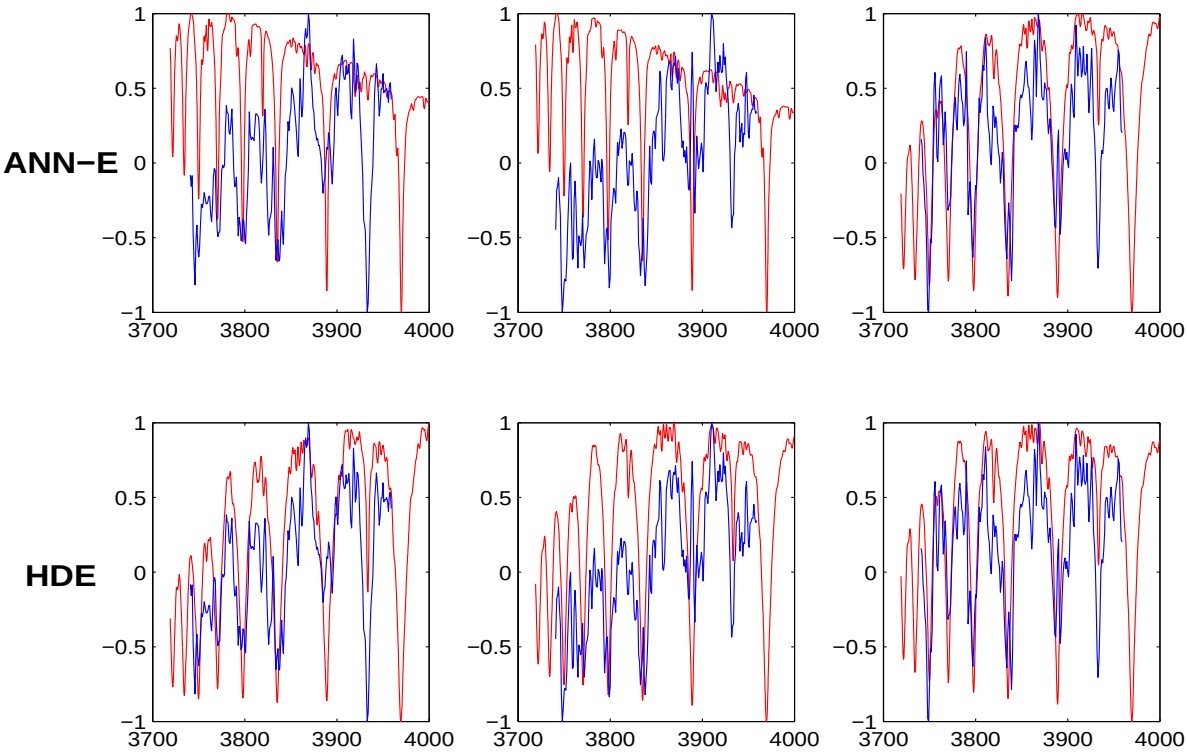
En la parte (a) se comparan los ajustes de perfiles en las líneas mostrando el error RMS en la región inferior de cada recuadro. En la parte (b) se comparan los espectros reconstruidos a partir de las predicciones.

En cada figura se presentan los mismos 3 espectros, en azul el espectro observacional y en rojo el espectro asociado a las predicciones de cada modelo para ese espectro en particular.

Los espectros observacionales varían de figura en figura con el fin de obtener muestras representativas del desempeño de los clasificadores; en ningún momento se presentan únicamente los mejores resultados.

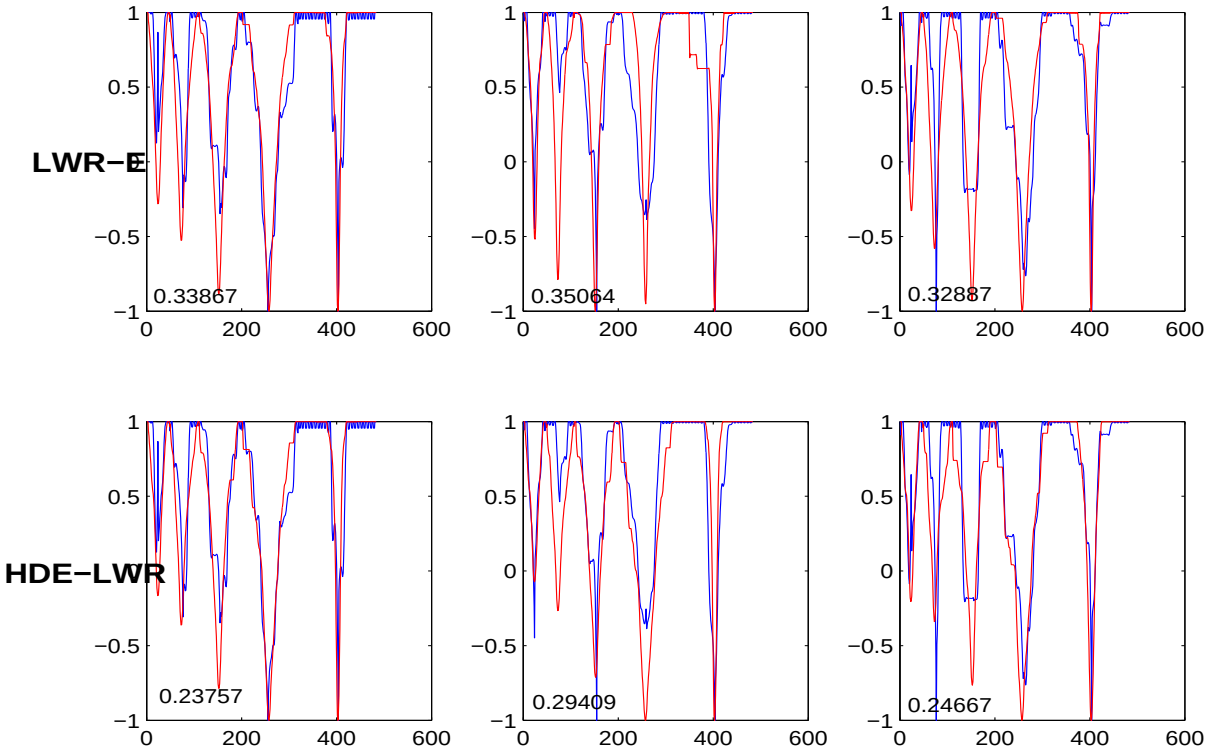


(a) Ajuste de los perfiles de las líneas

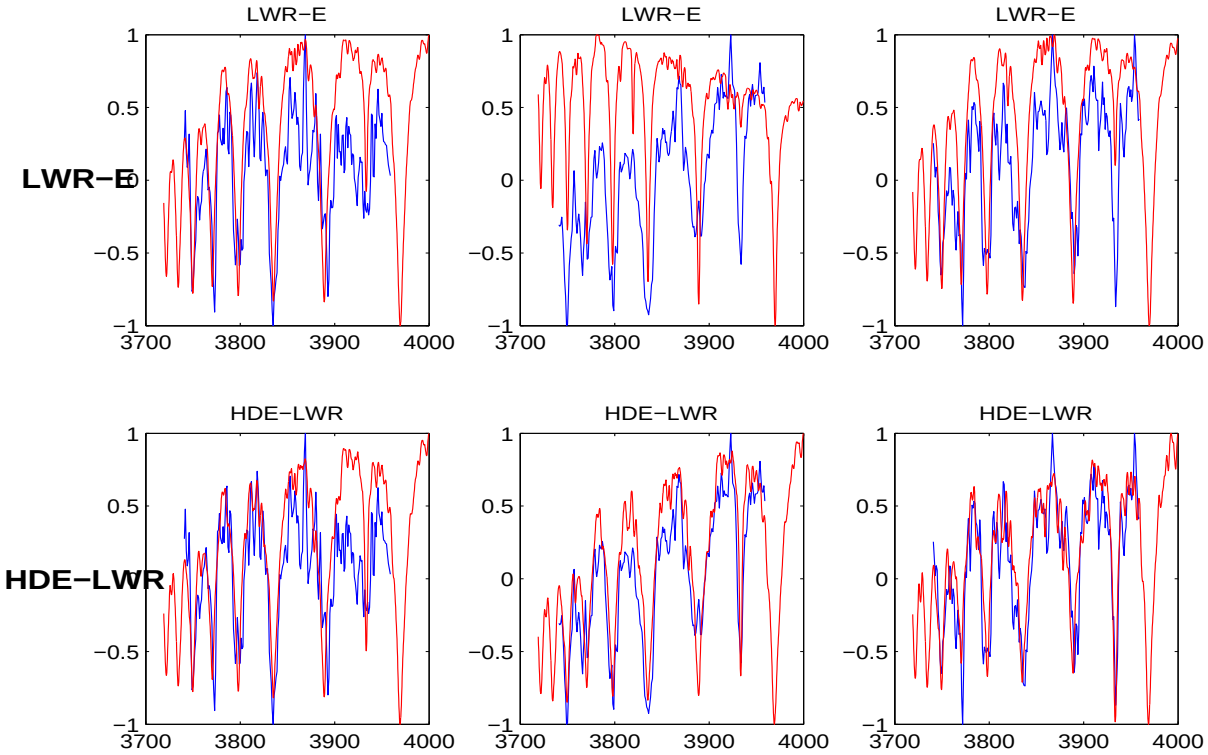


(b) Ajuste del espectro

Figura B.1: Comparación de ajustes entre ANN-E y HDE-ANN

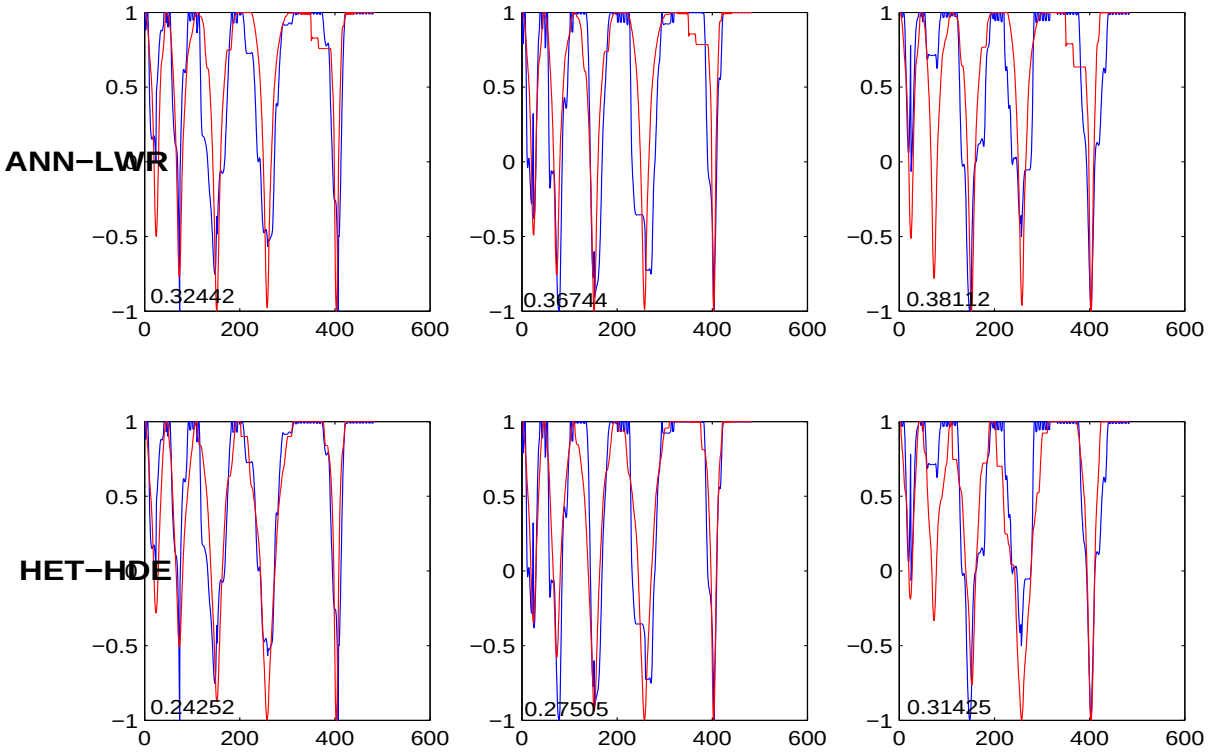


(a) Ajuste de los perfiles de las líneas

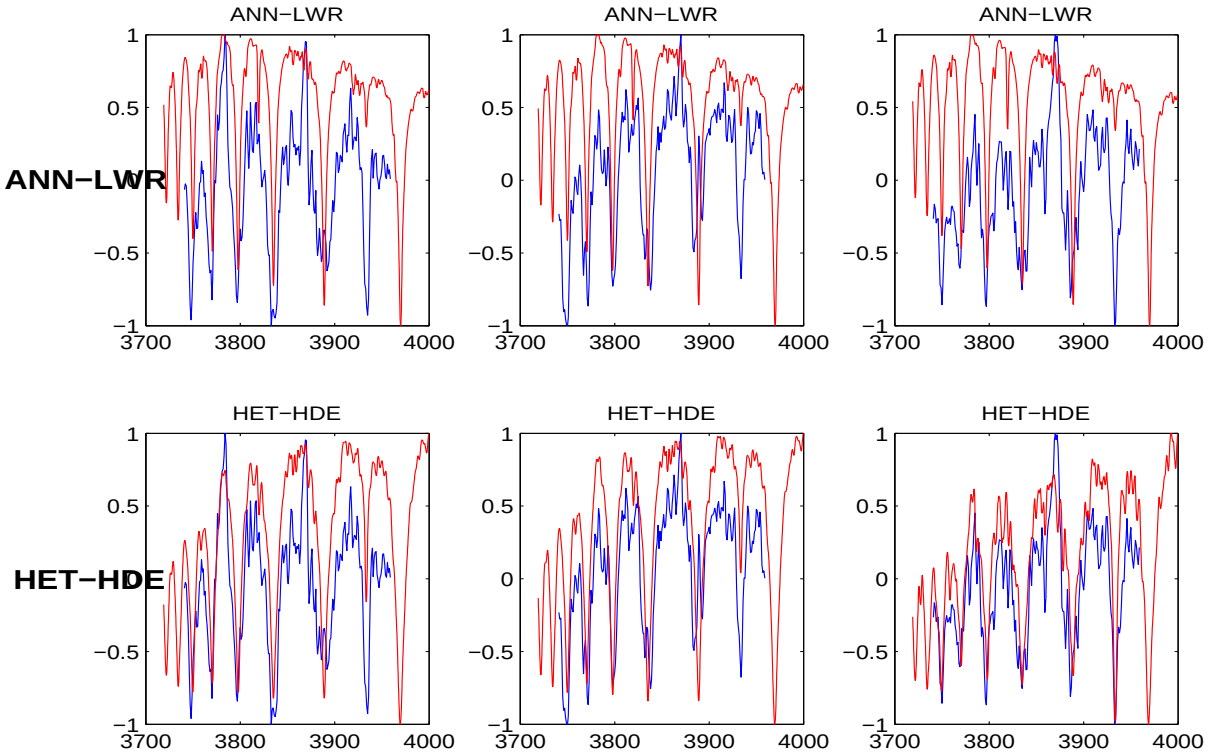


(b) Ajuste del espectro

Figura B.2: Comparación de ajustes entre LWR-E y HDE-LWR



(a) Ajuste de los perfiles de las líneas



(b) Ajuste del espectro

Figura B.3: Comparación de ajustes entre ANN-LWR y HET-HDE

Indice de figuras

Indice de tablas

Bibliografía

[1] Abazajian, K. et al.: The first data release of the Sloan Digital Sky Survey. The Astronomical Journal, **126** (2003) 2081–2086

[2] Alsabti, K., Ranka, S. and Singh, V.: An efficient Kmeans clustering algorithm. In First Workshop on High-Performance Data Mining, (1998)

[3] Bazell, D., Aha, W. D.: Ensembles of classifiers for morphological galaxy classification. The Astrophisical Journal, **548** (2001) 219–223

[4] Bottou, L. and Bengio, Y.: Convergence properties of the K-means algorithm. In Advances in Neural Information Processing Systems, MIT Press, **7** (1995)

[5] Ceci M., Appice A., Malerba D.: Comparing Simplification Methods for Model Trees with Regression and Splitting Nodes. in N. Zong, Z.W. Ras, S. Tsumoto, E. Suzuki (Eds.), Foundations of Intelligent Systems, Lecture Notes in Artificial Intelligence, **2871** (2003) 49-5-6

[6] Ceci M., Appice A., Malerba D.: Simplification Methods for Model Trees with Regression and Splitting Nodes. Series Lecture Notes in Artificial Intelligence, **2734** (2003) 434–445.

[7] Cerviño et al.: Evolutionary synthesis models for young star forming regions that compute the Stellar Energy Dispersion and Distribution. A&A, **363** (2000) 970

[8] Chawla, N. V., et al.: Learning Ensembles from bites: A scalable and accurate approach. Journal of Machine Learning Research, **5** (2004) 421–451

[9] Coryn A. L., Bailer-Jones, Irwin, M., and Von-Hippel, T.: Physical parametrization of stellar spectra: The neural network approach. Monthly Notices of the Royal Astronomical Society, **292** (1997) 157

[10] Coryn A. L., Bailer-Jones, Irwin, M., and Von-Hippel, T.: Automated Classification of Stellar Spectra II. Monthly Notices of the Royal Astronomical Society, (1998)

[11] Davies D.L., Bouldin D.W.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2) 224-227, (1979)

[12] De la Calleja J., Fuentes O.: Automated Classification of Galaxy Images, Proceedings of the Eight International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES), (2004)

[13] Dietterich, T. G.: An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Unpublished manuscript, (1998)

[14] Fisher, R.: The use of multiple measures in taxonomic problems. Ann. Eugenics, **7** (1936) 179–188

[15] Folkes S. R., Lahav O., Maddox S. J.: An Artificial Neural Network Approach to Classification of Galaxy Spectra. Mon. Not. R. Astron., (1996)

[16] Fuentes, O. and Gulati, R. K.: Prediction of Stellar Atmospheric Parameters using Neural Networks and Instance-Based Learning. Experimental Astronomy, **12** 1 (2001) 21–31

[17] González-Delgado R. M., Leitherer C., Heckman T. M.: Synthetic spectra of H Balmer and He I absorption lines. II. Evolutionary synthesis models for starburst and poststarburst galaxies. The astrophysical Journal Supplement Series, **125** (1999) 489–509

[18] Grossman, D., Williams, T.: Machine learning ensembles: An empirical study and novel approach. Unpublished manuscript, (2000)

[19] Gulati,R. K., Gupta, R., Gothoskar, P., and Khobragade, R.: Ultraviolet stellar spectral classification using multilevel tree neural networks. Vistas in Astronomy, Neural Networks in Astronomy, **3** (1993) 38–293

[20] Guyon I., Elisseeff A.: An introduction to variable and feature selection, Journal of Machine Learning Research **3** (2000) 1157–1182

[21] Hen-Hu, Y. Hwang, J. N.: Neural Network Signal Processing. CRC Press, 2001

[22] John G.H.: Robust Linear Discriminant Trees. Artificial inteligence and Statistics, (1994)

[23] Jordan M. I., Jacobs R. A.: Hierarchical mixture of experts and the EM algorithm. Neural Computation, **6**(1994) 181–214

[24] Karalic A.: Employing linear regression in regression tree leaves. In Proceedings of ECAI-92, (1992)

[25] Kohonen T.: Self Organization and Associative Memory. Springer-Verlag, **3** (1989)

[26] Kumar S., Ghosh J. and Crawford M.: A Hierarchical Multiclassifier System for Hyperspectral Data Analysis. Lecture Notes in Computer Science, **1857** (2000) 270

[27] Leitherer C.: Star Formation in Starburst and Active Galactic Nuclei. The Central kpc of Starburst and AGN, **30** (2001)

[28] Melia M., Jordan M. I.: Learning with Mixtures of Trees Journal of Machine Learning Research, (2000) 1–48

[29] Mitchell T. M.: Machine Learning. McGraw-Hill, (1997)

[30] Naim A., Lahav O., Sodré L., Storrie-Lombardi M. C.: Automated morphological classification of APM galaxies by supervised artificial neural networks. Mon. Not. R. Astron. Soc. **275** (1995) 567–590

[31] Ng, A. Y., Jordan, M. I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems, MIT Press, **14** (2002) 849–856

[32] Quinlan, J. R.: Induction of Decision Trees. Machine Learning, **1** (1986) 81–106

[33] Quinlan J.R.: Simplifying decision trees. Interational Journal of Man-Machine Studies, **27** (1987) 221–334

[34] Quinlan J.R.: Learning with continuous classes. In Adams and Sterling, editor, Proceedings of AI'92, (1992) 343–348

[35] Ramirez, F., Fuentes, O. and Gulati, R. K.: Prediction of Stellar Atmospheric Parameters Using Instance-based Machine Learning and Evolutionary Algorithms, Experimental Astronomy, **12** 3 (2001) 163–178

[36] Rumelhart D. E., Widrow B., Lehr M.: The basic ideas in meural networks. Communications of the ACM, 37 **3** (1994) 87–92

[37] Serote-Roos M., Boisson C., Joly M., Ward M. J.: Stellar populations in active galactic nuclei -I. The observations Mon. Not. R. Astron. Soc. **301** (1998) 1–14

[38] Sharkey, A. J. C.: Combining Artificial Neural Nets. Springer-Verlag, (1999)

[39] Shih Y. S., Tsai H. W.: Variable selection bias in regression trees with constant fits Computational Statistics and Data Analysis, **45** (2004) 595–607

[40] Snider, S., et al.: Three-dimensional Spectral Classification of low-metallicity stars using artificial neural networks. The Astrophisical Journal, **562** (2001) 528–548

[41] Sollich P, Krogh A.: Learning with ensembles: how over-fitting can be useful. Advances in Neural Information Processing Systems, Cambridge, MA: MIT Press, (1996) 190–196

[42] Torgo L.: Functional Models for Regression Tree Leaves. In Proceedings of the International Conference on Machine Learning, (1997)

[43] Torgo L.: Computationally Efficient Linear Regression Trees. Classification, Clustering and Data Analysis: Recent Advances and Applications, (2002)

[44] Torres-Papaqui J.P, Terlevich R: Determination of nuclear SB ages in Seyfert galaxies.

[45] Utgoff P.E., Brodley C.E. Linear Machine Decision Trees. University of Massachusetts, COINS Technical Report 91-10, (1991)

[46] Utgoff P.E., Brodley C.E. Multivariate Decision Trees. University of Massachusetts, COINS Technical Report 92-82, (1992)

[47] Weaver, W.: Spectral Classification of unresolved Binary stars with Artificial Neural Networks. The Astrophysical Journal, **541** (2000) 298–305

[48] Weaver, W., Torres-Dodgen A.: Accurate two-dimensional classification of stellar spectra with Artificial Neural Networks. The Astrophysical Journal, **487** (1997) 847–857

[49] Weaver, W., Torres-Dodgen, A. V.: Neural Network classification of the near-infrared spectra of A-type stars . The Astrophysical Journal, **446** (1995) 300–317

[50] Webb G. I., Zheng Z.: Multi-Strategy Ensemble Learning: Reducing Error by Combining Ensemble Learning Techniques. IEEE Transactions on Knowledge and Data Engineering,

[51] Yao, X.: Evolving artificial neural networks. Proceedings of the IEEE, **87** 9, 1423 – 1447 (1999)

[52] Young, S.J., Odell, J.J., and Woodland, P. C.(1994), Tree-based state tying for High Accuracy Acoustic Modelling. In Proceedings ARPA Workshop on Human Language Technology, **18** 307–312 (1994)

[53] Zhang, T., Ramakrishnan, R., Livny, M.: Birch: A new data clustering algorithm and its applications. Data Mining and Knowledge Discovery, **2** (1997) 141–182