UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO
CENTRO UNIVERSITARIO UAEM TEXCOCO

"ANÁLISIS DE LAS TÉCNICAS DE SELECCIÓN DE CARACTERÍSTICAS"

T E S I S

QUE PARA OBTENER EL TÍTULO DE
LICENCIADA EN INGENIERIA EN COMPUTACIÓN

PRESENTA:

KARINA CALLEJA CALVARIO

ASESOR:

DR. EN C. JAIR CERVANTES CANALES

REVISORES:

Dr. EN C. FARID GARCÍA LAMONT
Dr. EN I.S. JOSÉ SERGIO RUÍZ CASTILLA

TEXCOCO, ESTADO DE MÉXICO, ABRIL DE 2019

A mis padres y hermanos, pilares fundamentales en mi vida. En reconocimiento a todo el sacrificio puesto para que pudiera estudiar, porque ni con la riqueza más grande del mundo podré pagar todo su apoyo. Es por ellos que soy lo que soy ahora.

Karina Calleja Calvario

Agradecimientos

A mi director de tesis el Dr. Jair Cervantes Canales, con especial y sincera gratitud por su generosidad de apoyo, paciencia, esfuerzo, amabilidad y la siempre disponibilidad que ha brindado a este trabajo. Muchas gracias, por aceptarme parea realizar esta tesis bajo su dirección. Le agradezco el haberme facilitado siempre los medios para llevar a cabo todas las actividades durante el desarrollo y culminación de esta tesis, así como las respuestas a las diferentes inquietudes surgidas.

A mis padres y hermanos, porque con su sacrificio y esfuerzo constante me dieron la posibilidad de llegar a esta instancia. Agradezco la confianza depositada en mí, el apoyo y motivación a lo largo de toda mi formación académica. Hoy veo forjado un anhelo, una ilusión y un deseo: Mi carrera Profesional. Por ustedes la obtuve y con ustedes la comparto.

Abstract

In pattern recognition, features selection is of vital importance, since it reduces the computational costs of data pre-processing and increases the reliability of the classification. In addition to the above, features selection allows obtaining a subset of features with better discriminative power, deleting those features with less discriminative power, and even features that introduce noise into the classifier. Feature selection techniques range from the simplest to very elaborate techniques such as genetic algorithms. In this thesis, an exhaustive analysis of the performance and behavior on different data sets of several techniques of feature selection in the current state of the art is performed. The techniques studied are implemented and compared using several data sets. The experimental results obtained are discussed and compared in this Thesis. The experimental results and the comparative analysis of the techniques will make it possible to make a more specific selection of some technique in terms of complexity and performance when there is a problem of feature selection.

Resumen

En reconocimiento de patrones, la selección de características es de vital importancia, pues reduce los costes computacionales de pre-procesamiento y aumenta la fiabilidad de la clasificación. Aunado a lo anterior, la selección de características permite obtener un subconjunto de características con mejor poder discriminativo, eliminando aquellas con menor poder, e incluso características que introducen ruido en el clasificador. Las técnicas de selección de características van desde las más sencillas hasta técnicas muy elaboradas como Algoritmos genéticos. En esta Tesis, se realiza un análisis exhaustivo del desempeño y comportamiento sobre diferentes conjuntos de datos de varias técnicas de selección de características en el estado del arte actual. Las técnicas estudiadas son implementadas y comparadas utilizando varios conjuntos de datos. Los resultados experimentales obtenidos son discutidos y comparados en esta Tesis. Los resultados experimentales y el análisis comparativo de las técnicas permitirán realizar una selección más propia de alguna técnica en cuanto a complejidad y desempeño cuando se presenta algún problema de selección de características.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

En aprendizaje de máquinas una tarea muy importante es el reconocimiento de patrones (RP). Los algoritmos de RP permiten identificar un objeto a partir de características. Estas características regularmente son obtenidas a través de algoritmos de búsqueda en bases de datos o través de algoritmos de extracción de características en imágenes. Para describir un objeto pueden ser utilizadas una inmensa cantidad de características. Sin embargo, solo algunas son necesarias para identificar perfectamente tal objeto. Cada una de estas características posee una capacidad para identificarlo o clasificarlo con respecto a otros objetos, a esta capacidad se le conoce como poder discriminativo. Pensemos en una manzana para identificarla y clasificarla con respecto a una naranja podría ser necesaria solo la característica color. Sin embargo, clasificar una manzana con respecto a otras variedades de manzanas aumentaría la complejidad del problema y serían necesarias más características, como descriptores de forma, color y aun textura. Las características de color, textura y forma pueden describir perfectamente muchos objetos en imágenes, sin embargo, el número de características que se pueden extraer de una imagen es enorme. Por otro lado, el número de características que define un evento en una base de datos también puede ser enorme.

En los últimos años en las aplicaciones de aprendizaje automático o reconocimiento de patrones, El número de características utilizadas se ha expandido de decenas a cientos de variables o características. Varias tecnicas han sido desarrollados para abordar el problema de reducir las variables irrelevantes y redundantes que son una carga computacional y en muchos casos afectan el redimiento de los clasificadores.

La selección de características es de vital importancia, pues reduce los costes computacionales de pre-procesamiento y aumenta la fiabilidad de la clasificación. Aunado a lo anterior, la selección de características permite obtener un subconjunto de características con mejor poder discriminativo, eliminando aquellas con menor poder, e incluso

características que introducen ruido en el clasificador. Las técnicas de selección de características van desde las más sencillas hasta técnicas muy elaboradas como Algoritmos genéticos.

En esta tesis se estudian las diferentes técnicas de selección de características que existen en el estado del arte. Las técnicas estudiadas son implementadas y comparadas utilizando varios conjuntos de datos. Los resultados experimentales obtenidos son discutidos y comparados en esta Tesis. Los resultados experimentales y el análisis comparativo de las técnicas permitirán realizar una selección más propia de alguna técnica en cuanto a complejidad y desempeño cuando se presenta algún problema de selección de características.

1.1. Problemática

El uso adecuado de características para una buena clasificación ha sido tratado por varios autores **Koppen 2000 Evangelista 2006** y lo llaman, el curso de la dimensionalidad. El curso de la dimensionalidad es un fenomeno que se presenta cuando el número de características asociadas a un conjunto de datos, en lugar de ayudar en el proceso de clasificación afecta el desempeño de un clasificador. Una dimensión grande de características en muchas ocasiones introduce ruido y confunde a los métodos de aprendizaje. Así que un desafío importante en la actualidad es reducir la dimensionalidad de características, esto es fundamental para mejorar el desempeño en la etapa de clasificación.

Un factor importante al momento de reducir características, es eliminar aquellas con un bajo poder discriminante, aquellas que introducen ruido o son poco confiables. Sin embargo, identificar estas características no es una tarea fácil. En esta tesis se analizan diferentes métodos utilizados en el estado del arte para selección de características.

1.2. Justificación

La selección de características es muy importante porque permite disminuir el tamaño del conjunto de datos de entrenamiento y en muchos casos mejora la precisión de clasificación. Un análisis exhaustivo permitirá mejorar el conocimiento sobre el comportamiento de las diferentes técnicas de selección de características.

La selección de características en conjuntos de datos no es un reto fácil debido a que los enfoques diseñados no solo deben retener las características individuales más importantes, sino también rescatar toda la información referente al poder discriminativo

de las características combinadas. Los resultados del análisis serán mostrados y discutidos en detalle. El objetivo de esta Tesis es mostrar un análisis del comportamiento de las diferentes técnicas bajo diferentes conjuntos de datos.

1.3. Objetivos

En esta Tesis mostraremos un análisis del comportamiento de las diferentes técnicas de selección de características bajo diferentes conjuntos de datos. Los siguientes trabajos serán desarrollados:

1. Implementar algoritmos de selección de características más utilizados en la literatura actual.
2. Implementar al menos dos algoritmos de clasificación para probar las técnicas de selección de características
3. Realizar un análisis y discusión de
    - El tiempo reducido al disminuir el número de características
    - La mejora en el desempeño al aumentar la precisión de clasificación

1.4. Estado del arte

La importancia del proceso de selección de características en cualquier problema de clasificación, se pone de manifiesto puesto que permite eliminar las características que puedan inducir a error (características ruidosas), las características que no aporten mayor información (características irrelevantes) o aquellas que incluyen la misma información que otras (características redundantes) **Blum 1997**. Este proceso tiene como ventaja la obtención de una disminución en los tiempos de procesamiento de los datos, menor requerimiento en los espacios donde se almacena la información, menor costo en la obtención de los datos (la definición de características específicas permite desarrollar sensores específicos para obtenerlas) y lo más importante es la selección de un subconjunto de las características originales que aportan la mayor cantidad de información para un problema en particular.

En general, en los procedimientos de selección de características se distinguen cuatro etapas esenciales **Dash 1997**:

1. Procedimiento de Selección: en esta etapa se determina el posible subconjunto de características para realizar la representación del problema

2. Función de Evaluación: en esta etapa se evalúa el subconjunto de características escogidas en el punto anterior.

3. Criterio de Detención: se chequea si el subconjunto seleccionado satisface el criterio de detención de la búsqueda.

4. Procedimiento de Validación: esta etapa se utiliza para verificar la calidad del subconjunto de características que se determinaron.

Los métodos de selección de características se clasifican desde el punto de vista de la manera en que se determina el nuevo subconjunto a evaluar, lo que conduce a 3 clases métodos **Dash 1997**.

1. Métodos Completos. Estos métodos examinan todas las posibles combinaciones de características. Son muy costosos computacionalmente (espacio de búsqueda de orden O(2N) para N características) pero se asegura encontrar el subconjunto óptimo de características. Como ejemplos de estos métodos se puede citar Branch and Bound **Narendra 1977** y Focus **Almuallin 1992**.

2. Métodos Heurísticos. Utilizan una metodología de búsqueda de forma tal que no es necesario evaluar todos los subconjuntos de características. Ello significa una mayor velocidad del método, ya que el espacio de búsqueda es menor que en los métodos anteriores. Estos métodos no aseguran la obtención del mejor subconjunto. A modo de ejemplo es interesante citar en esta categoría los métodos Relief **Kira 1992** y DTM **Cardie 1993**.

3. Métodos Aleatorios. Son aquellos métodos que no tienen una forma específica de definir el subconjunto de características a analizar, sino que utilizan metodologías aleatorias. Con ello se produce una búsqueda probabilística en el espacio de características. El resultado obtenido utilizando este tipo de métodos dependerá del número de intentos, no asegurándose la obtención del óptimo. Pertenece a este grupo los métodos presentados en LVW **Liu 1996** y algunos que utilizan algoritmos genéticos **Vafaie 1994**.

Desde el punto de vista de la función de evaluación, los procedimientos de selección de características se pueden clasificar en 2 categorías **John 1994**.

1. Métodos de filtraje. Estos son métodos donde el procedimiento de selección es realizado en forma independiente a la función de evaluación (clasificación). Se pueden distinguir 4 diferentes medidas: distancia, información, dependencia y consistencia. Como ejemplo de estos métodos tenemos Relief **Kira 1992**, DTM **Cardie 1993**, POEACC **Muciardi 1971** y Focus **Almuallin 1992** respectivamente.

2. Métodos dependientes (wrapped). En estos métodos el algoritmo de selección utiliza como medida la tasa de error del clasificador. Se obtienen generalmente mejores resultados que en el caso anterior, pero trae consigo un costo computacional mucho mayor. En esta categoría se tienen métodos como Oblivon **Langley 1994**.

Preliminares

La selección de características comprende varios pasos y el uso de técnicas especiales para mejorar la selección óptima de características. En el proceso general de la selección de características se ven involucrados diversas técnicas. En esta Sección se muestran varias técnicas de clasificación, así como técnicas para evaluar el desempeño de los clasificadores.

1.5. Técnicas de Clasificación

Con el fin de evaluar la factibilidad de los selectores de características en esta tesis se utilizan varios clasificadores para evaluar y comparar las técnicas de selección de características. Debido a la importancia fundamental de la clasificación y detección en muchas situaciones prácticas, se deben tomar ciertas medidas para seleccionar un modelo. En esta Sección, se muestran diferentes técnicas de clasificación utilizadas en esta tesis para evaluar el desempeño.

Neural networks

Las redes neuronales artificiales (RNA) han recibido gran atención en los últimos años. Estas se han implementado para resolver problemas de predicción y clasificación, áreas en las que tradicionalmente se han utilizado técnicas estadísticas. Sin embargo, en los últimos años han obtenido una inmensa popularidad para resolver problemas de regresión.

Las RNA construyen un modelo de predicción imitando la inteligencia del cerebro humano. De forma similar a como lo hace el cerebro, las RNA identifican regularidades y patrones en los datos, aprenden a partir de la experiencia y proveen resultados que son generalizados a partir del conocimiento obtenido. Una red neuronal es un conjunto de neuronas artificiales interconectadas que utilizan modelos matemáticos para procesar información. Las múltiples conexiones entre las neuronas forman un sistema adaptivo cuyos pesos se actualizan mediante un particular algoritmo de aprendizaje.

Las redes neuronales han sido utilizadas en numerosos campos de aplicación con distintos algoritmos de aprendizaje **Portillo:2009 Jimenez:2013 G:2010 Valverde:2007**.

Elementos de las RNA

En los últimos años, debido al auge de las RNA se han utilizado diferentes RNA para resolver problemas de predicción. Las características esenciales de una RNA son las siguientes:

1. Los elementos básicos de procesamiento (neuronas o nodos)
2. La arquitectura de la red describiendo las conexiones entre nodos
3. El algoritmo de entrenamiento usado para encontrar valores de los parámetros de la red

Una RNA consiste de elementos de procesamiento básico (neuronas) organizadas en capas. Las capas que se encuentran entre la capa de entrada y capa de salida son llamadas capas ocultas. En número de neuronas en la capa de entrada es determinada por la aplicación. La arquitectura o topología de una RNA se refiere al arreglo de las conecciones de la red. Una RNA es especificada mediante:

$$S = \{h(x, \theta), x \in R^m, \theta_i \in \theta\}, \theta \subseteq R^p \tag{1.1}$$

donde $h_\theta(x, \theta)$ es una función no lineal del producto punto de x con θ, es el número de neuronas ocultas, $\theta = \left(\theta_1, \theta_2, ..., \theta_p\right)^T$ es un vector cuyos elementos usualmente son llamados pesos y p es el número de parámetros libres.

Función de costo y algoritmo de aprendizaje

Dado un conjunto de observaciones, la tarea de una RNA consiste en construir un estimador $g(x, \theta)$ de la función desconocida $\varphi(x)$

$$g_\lambda(x, \theta) = f_2\left(\sum_{j=1}^{\lambda} \theta_j^2 f_1\left(\sum_{j=1}^{m} \theta_{ij}^1 x_i + \theta_{m+1,j}^1\right) + \theta_{\lambda+1}^1\right) \tag{1.2}$$

donde $\theta = \left(\theta_1, \theta_2, ..., \theta_p\right)^T$ es el vector de parámetros a ser estimados, f_i es conocida como la función de activación, Las funciones de activación comúnmente utilizadas en RNA son función sigmoidal o tangente hiperbólica.

Los parámetros estimados $\theta^\wedge$ son obtenidos minimizando iterativamente una función de costo

$$J(\theta) = \frac{1}{2n} \sum_{j=1}^{n} \left(y_i - g_\lambda(x_i, \theta)\right)^2 \tag{1.3}$$

El requerimiento básico de cualquier método es la convergencia del algoritmo de entrenamiento a un mínimo local. La función de costo nos da una medida de la precisión con la que el estimador $g_\lambda(x_i, \theta)$ ajusta los datos observados. Para llevar a cabo el proceso de aprendizaje, el algoritmo cambia iterativamente los pesos entre las neuronas minimizando el error cuadrático entre la salida deseada y la obtenida con los pesos actuales. Cada uno de los ejemplos del conjunto de entrenamiento $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ son utilizados para ajustar los pesos en la red. Al ser presentado un ejemplo, la señal es propagada hacia adelante de la red hasta que la salida es obtenida. La salida de la j-ésima unidad oculta es calculada como:

$$g(x, \theta)^h_{nj} = f^h_j \left(net^h_{nj} \right) = \frac{1}{1 + exp\left(-net^h_{nj} \right)} \tag{1.4}$$

donde $net^h_{nj} = \sum \theta^h_{ji} x_{ni}$, θ^h_{j0} es el sesgo cuyo $x_0 = 1$. θ^h_{ji} es el peso de la conexión de la i-ésima neurona de entrada a la j-ésima neurona oculta. f^h_j representa la función de activación de la j-ésima neurona oculta. Por otro lado, la salida de la k-ésima neurona está dado por

$$g(x, \theta)^o_{nk} = f^o_k \left(net^o_{nj} \right) = \frac{1}{1 + exp\left(-net^o_{nk} \right)} \tag{1.5}$$

donde los superíndices h y o se refieren a las cantidades en las capas ocultas y de salida respectivamente. El error entre la salida actual y la salida deseada es calculado para ajustar los pesos mediante $E_n = \frac{1}{2} \sum_{k=1}^{C} \left(y_{nk} - g(x, \theta)^o_{nk} \right)$. El procedimiento de ajuste es obtenido a partir del método de gradiente descendente para reducir la magnitud del error. El procedimiento es primeramente aplicado a los pesos en la capa de salida y retropropagando a través de la red hasta que los pesos en la primera capa han sido ajustados $\Delta\theta^o_{kj} = -\eta \frac{\partial E_n}{\partial \theta^o_{kj}}$ y $\Delta\theta^h_{ji} = -\eta \frac{\partial E_n}{\partial \theta^h_{ji}}$. Este procedimiento es realizado para cada ejemplo en el conjunto de datos hasta que se cumple un criterio de paro. Para un estudio profundo del algoritmo puede referirse a **Rumel:1986 Werbos:1994**.

Support Vector Machines

Las Maquinas de vectores de Soporte (SVM por sus siglas en inglés) son una de las técnicas de clasificación y regresión más utilizadas en los últimos años. Las características clave de las SVMs son el uso de *kernels* al trabajar en conjuntos no-lineales, la ausencia de los mínimos locales, aunado a ello la solución depende de un pequeño subconjunto de datos y el poder discriminativo del modelo obtenido es significativamente

mayor a otras técnicas al optimizar el margen de separabilidad entre clases, estas características permiten a las SVM obtener resultados muy competitivos en comparación con otros clasificadores.

Las SVM fueron desarrolladas por Vapnik y sus colegas en los laboratorios de ATT en 1995 **Vapnik:1998**. En un inicio las SVM fueron diseñadas para resolver problemas de clasificación como reconocimiento de caracteres y reconocimiento facial. Sin embargo, pronto se fueron utilizando no solo para resolver problemas de clasificación sino también para resolver problemas de regresión en variadas implementaciones y múltiples dominios **Vapnik:1998**.

Las SVM son basadas en el principio de minimización del riesgo estructural. El objetivo de las SVM es encontrar una hipótesis $h_\theta(x)$ con una buena habilidad de generalización a partir de un conjunto de entrenamiento. Esta hipótesis es completamente definida a partir de un pequeño subconjunto del conjunto de entrenamiento original. En las siguientes subsecciones se describe formalmente a las SVM.

Teoría de aprendizaje estadístico

La teoría de aprendizaje estadístico es desarrollada con el objetivo de obtener una técnica de aprendizaje con una buena capacidad de generalización. En problemas de clasificación y regresión el objetivo es encontrar una hipótesis (función) a partir de los datos de entrenamiento y usando una máquina de aprendizaje que infiera resultados basados en este conocimiento.

En el caso de aprendizaje supervisado, los datos de entrenamiento son compuestos por pares de entrada y salida. Los vectores de entrada $x \in X \subseteq R^n$ y los puntos de salida $y \in Y \subseteq R$. Los dos subconjuntos X e Y son definidos como espacios de entrada y salida respectivamente. $Y = -1, 1$ o $0, 1$ para problemas de clasificación binaria y $Y = R$ para problemas de regresión.

Es claro que los datos de entrenamiento son generados a partir de una distribución desconocida $P(\mathbf{x}, y)$ definida sobre un conjunto $X \times Y$. Un vector de entrada es trazado desde X con la probabilidad marginal $P(x)$ y su correspondiente punto de salida en Y con la probabilidad condicional $P(\mathbf{x}, y)$.

Después de estas descripciones, el problema de aprendizaje puede ser visualizado como la búsqueda de una apropiada función de aproximación $f : X \longrightarrow Y$ que represente el proceso de obtención de salidas a partir de los vectores de entrada. Esta función puede ser utilizada para generalizar, es decir para producir una salida a partir de un vector de entrada nunca antes visto por el modelo.

Minimización del riesgo empírico

Según Vapnik **Vapnik:1998** el riesgo funcional es definido sobre $X \times Y$ para medir el error promedio obtenido de las salidas real y predicha al utilizar una función de aproximación f. La función de aproximación más adecuada es seleccionada como la función que minimiza este riesgo.

Considerando un conjunto de funciones $F = f(\mathbf{x},\mathbf{w})$ que mapea los puntos de un espacio de entrada $X \subseteq R^n$ dentro de un espacio de salida $Y \subseteq R$ donde w denota los parámetros que definen a f.

Suponiendo que y sea el punto actual de salida correspondiente al vector de entrada $\mathbf{x}$. Ahora si $L(y, f(\mathbf{x},\mathbf{w}))$ mide el error entre el valor actual y y el valor predicho $f(\mathbf{x},\mathbf{w})$ usando la función de predicción f entonces el riesgo esperado es definido como:

$$R(f) = \int L(y, f(\mathbf{x},\mathbf{w}))dP(\mathbf{x}, y) \tag{1.6}$$

donde $P(\mathbf{x}, y)$ es la distribución de probabilidad de los datos de entrenamiento. $L(y, f(\mathbf{x},\mathbf{w}))$ es conocido como la función de perdida y esta puede ser definida como un número de soluciones.

La función de predicción más adecuada es aquella que minimiza el riesgo esperado $R(f)$ y es denotada como f_0. Esta es conocida como la función objetivo. La principal tarea del problema de aprendizaje es ahora encontrar la función objetivo, que es el estimador ideal. Desafortunadamente esto no es posible debido a que probabilidad de distribución $P(\mathbf{x}, y)$ de los datos dados es desconocida y por lo tanto el riesgo esperado no puede ser calculado. Este problema motivo a Vapnik a sugerir el principio de minimización de riesgo empírico (MRE). El concepto de MRE estima el riesgo esperado $R(f)$ usando el conjunto de entrenamiento. Esta aproximación de $R(f)$ es llamada el riesgo empírico. Dado un conjunto de entrenamiento $(\mathbf{x}_i, y_i)$, donde $\mathbf{x}_i \in X \subseteq R^n$, $y_i \in Y \subseteq R(\forall i = 1, 2, ..., n)$ el riesgo empírico es definido como:

$$R_{emp}(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x},\mathbf{w})) \tag{1.7}$$

El riesgo empírico $R_{emp}(f)$ tiene su propio minimizador en F, que puede ser tomado como $\widehat{f}$. El objetivo de MRE es aproximar la función objetivo f_0 mediante $\widehat{f}$. Esto es posible debido a que $R(f)$ converge en $R_{emp}(f)$ cuando el conjunto de entrenamiento n es infinitamente grande.

Dimensión Vapnik-Chervonenkis (VC)

La dimensión Vapnik-Chervonenkis h de una clase de funciones F es definida como el máximo número de puntos que pueden ser exactamente clasificados mediante F.

Matemáticamente: $h = max|X|, X \subseteq R^n$ tal que $\forall b \in -1, 1^{|X|}$, $\exists f \in F$ tal que $\forall x_i \in X(1 \leq i), f(x_i = b_i)$.

La dimensión VC es una medida de la capacidad intrínseca de una clase de funciones F. Burgues [1998] menciona que la dimensión VC de un conjunto de hiperplanos orientados en R^n es $(n+1)$. Esto es, tres puntos etiquetados en 8 diferentes formas pueden ser siempre clasificados por una frontera de decisión orientada en R^2 pero cuatro puntos no. Esto es la dimensión VC del conjunto de líneas orientadas en R^2 es tres. Por ejemplo, el problema XOR no puede ser resuelto utilizando frontera de decisión lineal. Sin embargo, una frontera de decisión cuadrática puede correctamente clasificar los puntos en este problema.

Minimización del Riesgo Estructural MRE

El defecto principal del principio de MRE es que en la práctica siempre tenemos un conjunto finito de observaciones y no puede ser garantizado que minimizando la función de aproximación del riego empírico sobre F también minimizaremos el riesgo esperado. Para enfrentar esta desventaja fue desarrollado el principio de minimización de riesgo estructural por Vapnik y Chervonenkis en 1982. La clave de este principio es que la diferencia entre el riesgo empírico y riesgo esperado puede ser acotada en terminos de la dimensión VC de la clase F de las funciones de aproximación.

Teorema: Sea F una clase de funciones de aproximación de la dimensión h. Entonces para cualquier par $(\mathbf{x}_i, y_i)$, donde $\mathbf{x}_i \in X \subseteq R^n, y_i \in Y \subseteq R(\forall i = 1, 2, ..., n)$ trazadas para cualquier distribución $P(\mathbf{x}, y)$ la siguiente cota se mantiene con probabilidad $1 - \eta(0 \leq \eta \leq 1)$:

$$R(f) \leq R_{emp}(f) + \sqrt{\frac{h\left(ln\left(\frac{2n}{h}\right) + 1\right) - ln\left(\frac{n}{4}\right)}{n}} \tag{1.8}$$

El segundo término del lado derecho se dice que es la confianza VC y $1 - \eta$ se denomina el nivel de confianza. El principio de MRE sugiere que para llevar a cabo una buena generalización es necesario minimizar la combinación del riesgo empírico y la complejidad del espacio de hipótesis. En otras palabras, es necesario seleccionar una hipótesis que tenga una buena relación entre un pequeño error empírico y una pequeña complejidad del modelo.

Support Vector Machines La principal idea de SVM en clasificación binaria es encontrar un hiperplano canónico que maximice la separación entre dos clases de ejemplos de entrenamiento. Consideremos dos conjuntos de puntos que son linealmente separables en R^n los cuales son clasificados dentro de una de dos clases C_1 y C_2 utilizando

hiperplanos lineales. A partir del conjunto infinito de hiperplanos de separación es seleccionado aquel con el máximo margen de separación, que es el que tiene la mejor capacidad de generalización.

Para una mejor precisión así como una mejor capacidad de generalización, el hiperplano que maximiza el margen total es considerado como el óptimo y conocido como el Hiperplano de Margen Máximo. Es claro que para este hiperplano óptimo $d_+ = d_-$. Los datos ya sea de una u otra clase que se encuentran más cercanos al hiperplano son conocidos como vectores de soporte.

Considerando que el conjunto de entrenamiento es compuesto por los pares de entrada salida

$$(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_n, \mathbf{y}_n) \tag{1.9}$$

i.e. $X = \{x_i, y_i\}_{i=1}^n$ donde $x_i \in R^d$ y $y_i \in (+1, -1)$.

El objetivo es clasificar los datos en dos clases encontrando el hiperplano canónico de máximo margen. El espacio de hipótesis es el conjunto de funciones $f(\mathbf{x}, \theta, b) = sgn(\theta^{\mathbf{T}}\mathbf{x} + b)$ donde θ es el vector de pesos, $x \in R^n$ y b es el sesgo. el conjunto de hiperplanos de separación es dado por $\left\{x \in R^n : (\theta^{\mathbf{T}}\mathbf{x} + b) = 0\right\}$, donde $\theta \in R^n, b \in R$. Usando SVM para encontrar el hiperplano de separación máximo se acota el problema a resolver un problema de programación cuadrática (QPP).

SVM para datos linealmente separables

Para datos linealmente separables el problema de optimización cuadrática es dado por

$$\begin{aligned} &\text{mín } J(\theta) = \tfrac{1}{2}\theta^T\theta = \tfrac{1}{2}\left\|\theta\right\|^2 \\ &\text{subject to } y_i(\theta^T\mathbf{x}_i + b); \forall i = 1, 2, ..., n \end{aligned}$$

Para resolver el problema de optimización cuadrática es necesario transformarlo al espacio dual. Entonces los multiplicadores de Lagrange y las condiciones complementarias de Kuhn-Tucker son usadas para encontrar la solución óptima.

Considerando que la solución del problema de Optimización Cuadrática nos genera los multiplicadores de Lagrange optimizados $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)^T$, donde $\alpha_i \geq 0$ y el sesgo óptimo esta dado por b_{opt}. Los vectores de datos para los que $\alpha_i > 0$ son los vectores soporte y se supone que existen en total P_{sv} vectores soporte. Entonces el vector de pesos óptimo puede ser escrito como:

$$\theta_{opt} = \sum_{i=1}^{P_{sv}} \alpha_i y_i \mathbf{x}_i \tag{1.10}$$

La función del hiperplano de decisión óptimo es dada como:

$$y(\mathbf{x}) = sgn\left(\sum_{i=1}^{P_{sv}} y_i \alpha_i (\mathbf{x}^{\mathbf{T}} \mathbf{x}_i) + b_{opt}\right) \tag{1.11}$$

Un dato desconocido es clasificado en una de las dos clases de acuerdo a la ecuación (1.11).

SVM para datos linealmente no separables

En aplicaciones de la vida real los datos de entrenamiento son linealmente no-separables. Como ejemplo podemos tomar el problema de clasificación XOR. En tales casos un Hiperplano de Margen Suave es modelado. El problema de optimización cuadrática en tal caso es dado por

$$\begin{aligned} &\text{mín } J(\theta) = \tfrac{1}{2}\left\|\theta\right\|^2 + C\left(\textstyle\sum_{i=1}^{n}\xi_i\right) \\ &\text{subject to } y_i(\theta^T \mathbf{x}_i + b) \geq 1 - \xi_i; \forall i = 1, 2, ..., n \end{aligned}$$

Aquí las variables flojas ξ_i son introducidas para relajar las condiciones de margen duro y $C > 0$ es la constante de regularización que asigna una penalización a las clasificaciones erróneas. El vector de pesos óptimo y la función de decisión es similar al caso de separabilidad lineal. La única diferencia es que en este caso los multiplicadores de Lagrange son acotados por C i.e. $0 \leq \alpha_i \leq C$ y por los vectores soporte.

Kernels

Las SVM permiten mapear el espacio de los puntos de entrada a un espacio de características altamente dimensional a través de un mapeo no-lineal. El hiperplano de separación no-lineal es construido en este nuevo espacio de características. Este truco permite resolver el problema de los datos de entrenamiento cuando estos no son separables linealmente. Al utilizar una transformación apropiada los datos de entrada pueden ser separables linealmente en el espacio de características.

Debido a esta transformación, los datos linealmente no-separables pueden ser separados por una frontera de decisión π_f en el espacio de características. La frontera de decisión en H puede ser reescrita como:

$$y(\mathbf{x}) = sgn\left(\sum_{i=1}^{P_{sv}} y_i \alpha_i (\varphi(\mathbf{x})^{\mathbf{T}} \varphi \mathbf{x}_i) + b_{opt}\right) \tag{1.12}$$

Si es posible encontrar una función $\mathbf{K}(x_i, x_j) = \varphi(x_i^T)\varphi(x_j)$ entonces esta puede ser directamente utilizada en las ecuaciones de entrenamiento de una SVM, esta función es conocida como la función kernel. Para evitar el computo explícito de del mapeo no-lineal $\varphi(x)$, el kernel asociado debe satisfacer las condiciones de Mercer **Vapnik:1998**. Los kerneles más conocidos que han sido utilizados en las SVM son los siguientes:

1. Kernel lineal $\mathbf{K}(\mathbf{x}, \mathbf{y}) = x^T y$
2. Kernel polinomial $\mathbf{K}(\mathbf{x}, \mathbf{y}) = \left(1 + x^T y\right)^d$
3. Kernel Función radial base (RBF) $\mathbf{K}(\mathbf{x}, \mathbf{y}) = exp\left(\frac{-\|x-y\|^2}{2\sigma^2}\right)$ que en forma simple puede ser escrito como $exp\left(-\gamma \cdot \left\|X^T Y\right\|^2\right)$, $\gamma > 0$.
4. Kernel sigmoidal $\mathbf{K}(\mathbf{x}, \mathbf{y}) = tanh\left(\gamma \cdot X^T Y + r\right)$ que es similar a la función sigmoidal en regresión logística.

1.6. Técnicas de Evaluación de desempeño

A fin de evaluar la exactitud de predicción de un modelo particular o para evaluar y comparar diferentes modelos, se considera el rendimiento relativo en el conjunto de datos de prueba. Debido a la importancia fundamental de la clasificación y detección en muchas situaciones prácticas, se deben tomar ciertas medidas para seleccionar un modelo. Por esta razón, varias métricas de rendimiento se proponen en la literatura. Estas medidas, son reconocidas como un elemento importante en toda gestión de calidad. Para validar nuestros resultados utilizamos varias medidas de desempeño.

En esta Sección se describen algunas métricas de desempeño importantes que son frecuentemente utilizadas.

1.6.1. *Cross-validation*

La validación cruzada (*Cross-validation*), es un método para evaluar y comparar los algoritmos de aprendizaje mediante la división de datos en dos segmentos. El primero se utiliza para entrenar el modelo y el segundo para validar el modelo.

El algoritmo primero divide los datos en k partes iguales. Después realiza k iteraciones de entrenamiento, tomando en cada iteración como conjunto de prueba un

subconjunto diferente y construyendo el modelo con los subconjuntos restantes. La Figura 1.1 muestra un ejemplo con 4 iteraciones. El índice de error estimado es la media de todos los errores obtenidos en cada entrenamiento.

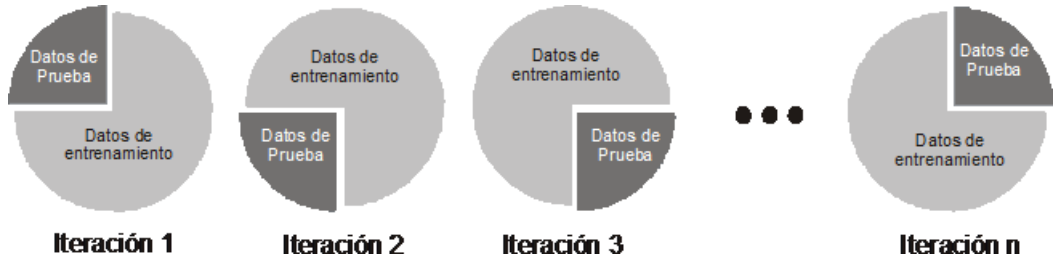

FIGURA 1.1: Validación cruzada de 4 iteraciones

La ventaja de evaluar a partir de k combinaciones de datos de entrenamiento y prueba hace que el método sea más preciso. Sin embargo, en una evaluación con un valor alto en k el proceso es lento al momento de computar. La elección del número de iteraciones depende de la medida del conjunto de datos, aunque lo más común es utilizar la validación cruzada de 10 iteraciones.

1.6.2. F-Measure

F-Measure no es más que la media armónica entre precisión y exhaustividad. La precisión representa el nivel de confianza del clasificado ya que es el porcentaje de datos clasificados correctamente. La exhaustividad representa la cobertura del clasificador, es decir, la cantidad de datos que clasifica frente a los no clasificados y clasificados. Cuando un sistema clasifica todos los datos en una sola categoría, este puede tener una exhaustividad alta, sin embargo, si la clasificación es incorrecta la precisión será baja. Tanto la precisión como exhaustividad están basadas en la matriz de confusión que está formada por cuatro casos:

Verdaderos positivos (TP): Es el caso de los datos positivos que han sido clasificados como positivos, es decir, es el total de datos que han sido clasificados correctamente.

Verdaderos negativos (TN): Es el caso de los casos negativos que ha sido clasificados como negativos, es decir, representan el número de datos clasificados en otra categoría correctamente.

Falsos Positivos (FP): Es el caso de los datos negativos que ha sido clasificados como positivos.

Falsos Negativos (FN): Es el caso de los datos positivos que han sido clasificados como negativos.

Partiendo de estos casos se puede formar la matriz de confusión como se muestra en la Tabla 1.1

CUADRO 1.1: Estructura de la Matriz de Confusión

	Positivos	Negativos
Positivos	TP: Verdaderos Positivos	FN: Falsos Negativos
Negativos	FP: Falsos Positivos	TN: Verdaderos Negativos

La técnica de matrices de confusión, no solo permite conocer el error del modelo predictivo, sino que también muestra el tipo de predicciones correctas e incorrectas cuando se aplica el modelo sobre el conjunto de prueba. Las predicciones correctas estas representadas por la diagonal principal, sin en cambio los elementos ubicados fuera de la diagonal principal, indican los errores de asignación.

Dada la matriz de confusión se pueden obtener la precisión y exhaustividad con las siguientes ecuaciones:

$$precision = \frac{TP}{TP + FP} \tag{1.13}$$

$$exhaustividad = \frac{TP}{TP + FN} \tag{1.14}$$

Para calcular F-Measure de una clase j con otra clase i primero se define la ecuación siguiente:

$$F_{ij} = \frac{2 * presicion(i,j) * exhaustividad(i,j)}{presicion(i,j) + exhaustividad(i,j)} \tag{1.15}$$

entonces F-Measure de un conjunto dado es calculado como sigue:

$$F - Measure = \sum \frac{n_i}{n} \text{máx}(F_{ij}) \tag{1.16}$$

donde n es el número de todo el conjunto de datos y n_i es el número de datos de la clase i.

El rango de los valores calculados esta entre 0 y 1. Un valor F-measure alto indica una mayor calidad de clasificación.

1.6.3. Área ROC

Cuando los errores llevan asociada una perdida que puede cuantificarse, es posible aplicar otra técnica de validación como el análisis de la curva ROC (por sus siglas en inglés, *Receiver Operating Characteristics*). Los gráficos ROC son útiles para visualizar el desempeño de los clasificadores y se utilizan comúnmente en la toma de decisiones médicas, aunque en los últimos años se han utilizado cada vez más en el aprendizaje automático (Tom 2006). El método consiste en un gráfico que ayuda a visualizar la disyuntiva entre la tasa de verdaderos positivos y la tasa de falsos positivos de un clasificador. La tasa de verdaderos positivos se representa en el eje las y, y la tasa de falsos positivos se representa en el eje de las x. La Figura 1.2 muestra un ejemplo de similitud que puede existir entre dos clases.

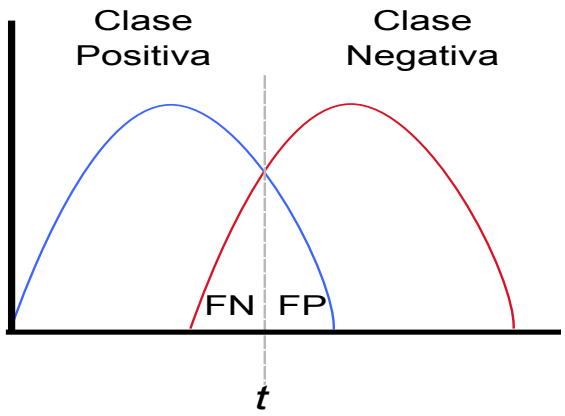

FIGURA 1.2: Representación de similitud entre dos clases. El punto de corte t determina el comportamiento del clasificador.

El comportamiento de pruebas depende del punto de corte t. Si este se desplaza a la derecha (Clase Negativa) disminuye la tasa de falsos positivos y aumenta la tasa de falsos negativos. Inversamente si se desplaza a la izquierda (Clase Positiva) aumenta la tasa de falsos positivos pero disminuye la tasa de falsos negativos. Entonces, para caracterizar el comportamiento entre estas dos clases se utilizan las curvas ROC. Un ejemplo de curva ROC se muestra en la Figura 1.3 .

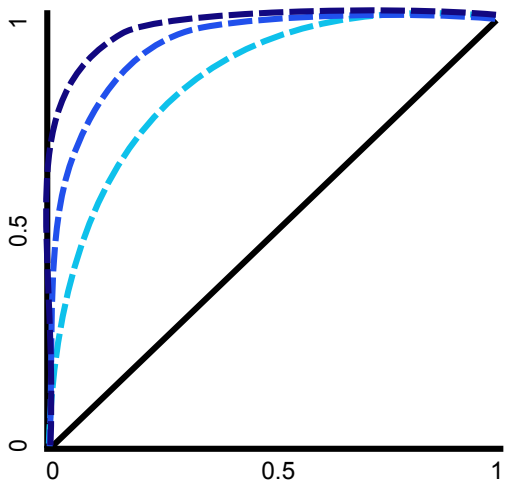

FIGURA 1.3: Representación de tres Curva ROC con distinta área bajo la curva.

Si la prueba fuera perfecta, es decir que no exista solapamiento entre clases, la curva solo tiene un punto (0,1). Sin embargo, si la prueba fuera mala, la curva sería una diagonal de (0,0) a (1,1). La Figura 1.3 muestra un ejemplo de distintos tipos de solapamiento y los tipos de curvas ROC que se generan, en la gráfica mientras más oscuro es el color de la línea más área bajo la curva posee.

El parámetro para evaluar la bondad de la prueba, es el área bajo la curva ROC que toma valores entre 1 (prueba perfecta) y 0.5 (prueba fallida). Esta área puede interpretarse como la probabilidad de que un conjunto de datos ante un clasificador funcione correctamente.

Técnicas de selección de características

La selección de características es un proceso que permite reducir un conjunto de características de acuerdo a un cierto criterio. El criterio utilizado determina los detalles de la evaluación de los subconjuntos de características. En general, el objetivo de las técnicas de selección de características es identificar las más importantes dentro del conjunto de datos y descartar las redundantes o irrelevantes. La selección de características en reconocimiento de patrones es muy importante debido a que esta permite reducir la dimensionalidad del conjunto de datos de entrada, eliminar ruido que sea introducido por características no deseables y en muchos casos mejorar la precisión de clasificación. En esta tesis se muestran las diferentes técnicas de selección de características utilizadas en el actual estado del arte.

Para distinguir entre las características importantes o discriminativas y las que no lo son, es necesario medir la calidad de la característica. Usualmente la evaluación de la métrica trabaja de dos maneras:

1. Calculando el desempeño en términos de eficacia
2. Calculando el desempeño en términos de eficiencia o maximizando la precisión en la predicción

En esta Sección se muestran algunos métodos de selección de características utilizados en esta tesis. Los métodos utilizados en esta tesis son algoritmos genéticos para la selección de características, métodos de filtrado y métodos envolventes.

1.7. Métodos de filtrado

Los métodos de filtrado utilizan técnicas de categorización de variables como criterio principal para seleccionar variables de acuerdo a su importancia. Estos métodos son utilizados por su simplicidad y buen desempeño. La selección de características es independiente de cualquier algoritmo de aprendizaje máquina. Las características son

seleccionadas en base a los resultados de varias pruebas estadísticas como la correlación con la variable de salida.

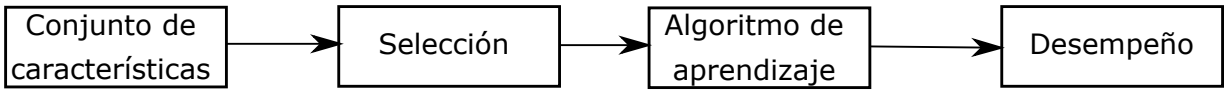

FIGURA 1.4: Pasos más usuales en métodos de selección de características por filtrado

Un criterio de categorización es utilizado para calificar y puntuar las variables y una frontera es utilizada para eliminar variables debajo de esa frontera. Esto permite eliminar las variables menos relevantes. Una propiedad básica de una característica es que contenga información útil de las diferentes clases en los datos. Es decir, que la característica sea útil para discriminar las diferentes clases.

Un punto importante es como evaluar la relevancia de una característica en el conjunto de datos. Varias publicaciones han presentado varias técnicas para medir la relevancia de una variable. Una característica puede considerarse irrelevante si esta presenta una independencia de las etiquetas de clase. Esto es, una variable puede ser condicionalmente independiente de las demás variables, pero no puede ser condicionalmente independiente de las etiquetas de clase. Si esta no tiene influencia en las etiquetas de clase, entonces esta se puede descartar. La correlación entre características desempeña un papel importante en la determinación de características únicas. Para aplicaciones prácticas, la distribución subyacente es desconocida y se mide por la precisión del clasificador. Debido a esto, un subconjunto de características óptimo puede no ser exclusivo porque puede ser posible lograr la misma precisión de clasificador utilizando diferentes conjuntos de características.

Los modelos de filtrado para selección de características usualmente consisten de os pasos:

1. Selección de características usando medidas como información, distancia, dependencia o consistencia con independencia del algoritmo de aprendizaje

2. Aprendizaje y prueba, el algoritmo aprende del entrenamiento de datos con el mejor subconjunto de características obtenido y probado sobre los datos de prueba.

A continuación se enumeran algunas técnicas para evaluar la relevancia de la característica en el conjunto de datos.

1.7.1. Criterio de correlación

Uno de los criterios más simples de selección es el coeficiente de correlación de Pearson, definido como:

$$R(i) = \frac{cov(x_i, Y)}{\sqrt{var(x_i) * var(Y)}} \tag{1.17}$$

donde x_{ij} representa los datos de entrada, $cov()$ es la covarianza y $var()$ es la varianza. Dado un conjunto de pares de ejemplos $[x_i, y_j]$, tal que $x_i \in R^d$ y y_k es la etiqueta de clase $K = 1, ..., Y$. La correlación solo puede detectar dependencias lineales entre la variable y el objetivo.

1.7.2. Información mutua

El criterio de valoración teórica utiliza la medida de dependencia entre dos variables. Para describir la información mutua es necesario describir antes la entropía de Shannon como sigue:

$$H(Y) = -\sum p(y)log(p(y)) \tag{1.18}$$

La ecuación 1.18 representa la incertidumbre (contenido de información) en la salida Y. Suponiendo que se observa una variable X entonces la entropia condicional es dada por:

$$H(Y|X) = -\sum_x \sum_y p(x, y)log(p(y|x)) \tag{1.19}$$

La ecuación 1.19 implica que observando una variable X 4, la incertidumbre en la salida Y es reducida. El decremento en la incertidumbre es dado como:

$$I(Y|X) = H(Y) - H(Y|X) \tag{1.20}$$

Esto nos da la medida de información entre Y y X, esto significa que si X y Y son independientes entonces la medida de información será cero y más grande que cero si estas son dependientes. Esto implica que una variable puede proveer información acerca de otras al proveer la dependencia. Las ecuaciones descritas anteriormente son

dadas para variables discretas, sin embargo pueden ser obtenidas para variables continuas reemplazando las sumas con integrales. La medida de información también puede ser definida como medida de distancia dada por:

$$K(f,g) = \int f(y) log(\frac{f(y)}{g(y)}) \tag{1.21}$$

La medida de K en 1.21 se le denomina la divergencia Kullback-Leibler entre dos densidades las cuales pueden ser usadas como una medida de información. A partir de las ecuaciones anteriores, es necesario conocer la función de densidad de probabilidad de las variables para calcular la medida de información.

Uno de los métodos más simples para calcular la medida de información y seleccionar características es encontrar la medida de información entre cada característica y la etiqueta de clase en la salida. Una vez hecho esto, se categorizan basados en ese valor. Una frontera es asignada para seleccionar d características, donde d es un valor mucho menor que el número de características de entrada. Este es un método muy simple y los resultados pueden ser muy pobres ya que el método no toma en cuenta la medida de información entre características.

En **Cardona 2006** los autores desarrollan un método de categorización basado en información mutua condicional para datos binarios (datos booleanos). Una Tabla de categorizaciones es actualizada y las características mejor categorizadas son seleccionadas usando el criterio de información mutua condicional que es maximizado. La puntuación en cada iteración es calculada utilizando la siguiente ecuación:

$$S(n) = min_{l<k} \hat{I}(Y; X_n | X_{v(l)}) \tag{1.22}$$

El criterio de información mutua 1.22 es actualizado en cada iteración. En la ecuación $S(n)$ representa la puntuación, X_n representa la característica evaluada, $X_{v(l)}$ es el conjunto de características ya seleccionadas. La ecuación iterativamente selecciona características que maximizan la medida de información con la clase y no seleccionan las características similares a las características que ya han sido seleccionadas, esto provee un buen equilibrio entre independencia y poder discriminativo.

Debido a que la distribución de los datos es desconocida, varias técnicas pueden ser utilizadas para evaluar subconjuntos con un clasificador seleccionado.

En **Forman 2003** se consideran doce métricas de selección de características para el problema de clasificación de texto. Todas las características se clasifican utilizando cada métrica y se establece un umbral que seleccionará 100 palabras que luego se aplican al

predictor. En **Javed 2010 Peng 2005** los autores desarrollan un criterio de clasificación basado en densidades de clase para datos binarios. Los autores utilizan un algoritmo de dos etapas menos costoso, una primera etapa con un método de filtro para clasificar las características y un método de envoltura costoso para eliminar las variables más irrelevantes.

El algoritmo RELIEF **Kira 1992 Acuna 2003** es otro enfoque basado en filtros en el que se utiliza un criterio de relevancia de la característica para clasificar las características. Usando un umbral se selecciona un subconjunto de características. El inconveniente del algoritmo RELIEF está en seleccionar un umbral. Los autores en **Acuna 2003** comparan el RELIEF y otros métodos de envoltura para diferentes conjuntos de datos.

Las ventajas de la selección de características son que es computacionalmente ligero y evita el sobreajuste y se ha comprobado que funciona bien para ciertos conjuntos de datos. Los métodos de filtro no se basan en algoritmos de aprendizaje sesgados, lo que equivale a cambiar los datos para adaptarse al algoritmo de aprendizaje. Uno de los inconvenientes de los métodos de selección es que el subconjunto seleccionado podría no ser óptimo, y el subconjunto obtenido podría ser un subconjunto redundante. Algunos métodos de selección, como los criterios de correlación de Pearson y Medida de Información, no discriminan las variables en términos de la correlación con otras variables. Las variables en el subconjunto pueden estar altamente correlacionadas.

Este problema de las variables redundantes y relevantes se trata en **Guyon 2003** con buenos ejemplos En la selección de características, las características importantes son menos informativas por sí mismas pero son informativas cuando se combinan con otras características descartadas **Guyon 2003 Xu2010**. Además, no existe un método ideal para elegir la dimensión del espacio de la característica.

1.8. Métodos de envoltura

A diferencia del filtro en los métodos que utilizan un criterio de relevancia de características, los métodos de envoltura se basan en la clasificación para obtener un subconjunto de características. Los métodos de envoltura utilizan el predictor como una caja negra y el rendimiento del predictor como la función objetivo para evaluar el subconjunto de variables. Se pueden utilizar varios algoritmos de búsqueda para encontrar un subconjunto de variables que maximice la función objetivo que es el rendimiento de clasificación. Los métodos de búsqueda exhaustiva pueden llegar a ser computacionalmente muy caros para conjuntos de datos grandes. Dado que la evaluación de los

subconjuntos es computacionalmente muy alta, se han desarrollado algoritmos heurísticos para encontrar los subconjuntos subóptimos. Por lo tanto, se emplean algoritmos simplificados como la búsqueda secuencial o algoritmos evolutivos como Algoritmo genético (GA)

u Optimización de enjambre de partículas (PSO) que producen resultados óptimos locales y son computacionalmente factibles.

En términos generales, clasificamos los métodos de envoltura en Algoritmos de selección secuencial y Algoritmos de búsqueda heurística. Los algoritmos de selección secuencial comienzan con un conjunto vacío (conjunto completo) y agregan funciones (eliminar características) hasta que se obtenga la función objetivo máxima. Para acelerar la selección, se elige un criterio que aumenta incrementalmente la función objetivo hasta que se alcanza el máximo con el número mínimo de funciones. Los algoritmos de búsqueda heurística evalúan diferentes subconjuntos para optimizar la función objetivo. Se generan diferentes subconjuntos ya sea buscando en un espacio de búsqueda o generando soluciones al problema de optimización.

1.8.1. Algoritmos de selección secuencial

En los métodos de envoltura, se utiliza subconjunto de características y se entrena un modelo con el subconjunto. A partir de las inferencias extraídas del modelo anterior, se agregan o eliminan características de su subconjunto. El problema se reduce esencialmente a un problema de búsqueda. Estos métodos suelen ser computacionalmente muy caros.

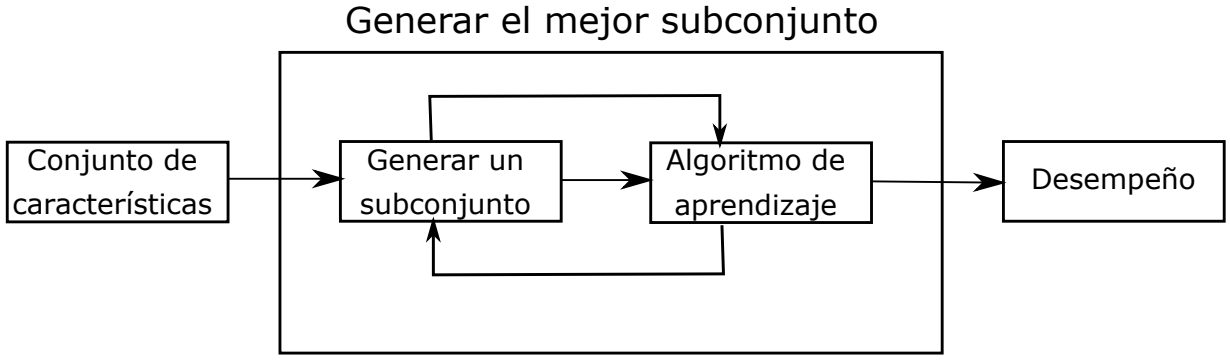

FIGURA 1.5: Pasos más usuales en métodos de selección de características por filtrado

Algunos de los métodos más comunes son los que se enumeran a continuación:

1. Selección hacia adelante: La selección hacia adelante es un método iterativo cuyo algoritmo que inicia sin tener ninguna característica en el modelo. En cada iteración, agrega una característica siempre y cuando proporcione la máxima precisión de clasificación hasta que la adición de una nueva característica no mejore el rendimiento del modelo o hasta que se agregan el número requerido de características. Este algoritmo se le conoce como un algoritmo ingenuo, debido a que la dependencia Entre las características no se tiene en cuenta.

2. Eliminación hacia atrás: en la eliminación hacia atrás, el algoritmo inicia con todas las características y va eliminando la característica menos significativa o que menor aporte tiene en cada iteración, lo que mejora el rendimiento del modelo. Repetimos esto hasta que no se observe ninguna mejora en la eliminación de características.

3. Eliminación de características recursivas: Este es un algoritmo de optimización codicioso que apunta a encontrar el subconjunto de funciones con mejor rendimiento. Crea repetidamente modelos y deja de lado la mejor o la peor característica de rendimiento en cada iteración. Construye el siguiente modelo con las características de la izquierda hasta que se agotan todas las características. Luego clasifica las características según el orden de su eliminación.

1.8.2. Algoritmos de búsqueda heurística

El algoritmo genético (GA) **Goldberg 1989** se puede usar para encontrar el subconjunto de características **Guyon 2003 Nakariyakul 2009 Alexandridis 2005 Rimbaud Yang 1998 Puch 1993** en donde los bits del cromosoma representan si la característica está incluida o no. Se puede encontrar el máximo global para la función objetivo que proporciona el mejor subconjunto subóptimo. Aquí nuevamente la función objetivo es el desempeño del predictor.

En la siguiente sección, analizaremos los métodos integrados que intentan compensar los inconvenientes en los métodos de filtro y envoltura.

Algoritmos Genéticos

Los algoritmos genéticos (AGs), parten de la premisa de emplear la evolución natural como un procedimiento de optimización **Goldberg 1989**. Se caracterizan por representar las soluciones al problema que abordan en forma de cadenas binarias. Esas representaciones binarias les aportan características muy importantes de eficiencia. Sin embargo, es necesario disponer de un método para pasar esa representación binaria al espacio de búsqueda natural de cada problema.

1.8.3. Elementos de un algoritmo genético

Para ejecutar un AG, se requiere de una población de individuos. Cada individuo, es un candidato a ser la solución del problema tratado, o permite llegar a la solución a partir de este.

Cada individuo de la población se representa con una cadena binaria y se denomina *genotipo* del individuo que es análoga al *cromosoma* en el sistema biológico. Cada genotipo representa a puntos x del espacio de búsqueda del problema. A cada punto x se le denomina *fenotipo*. Se usa el término *gen* para referirse a la codificación de una determinada característica del individuo. Cada *gen* puede tomar distintos valores que son llamados *alelos*. Para referirse a una determinada posición de la cadena binaria se usa el término locus. La Tabla 1.2 muestra estas expresiones que se usan comúnmente en la genética y su estructura equivalente en un algoritmo genético:

CUADRO 1.2: Expresiones que se utilizan en la genética con su estructura equivalente en un algoritmo genético

Evolución natural	Algoritmo genético
cromosoma	cadena
genotipo	código de cadena
fenotipo	punto sin codificar
gen	posición de cadena
alelo	valor en una posición determinada
aptitud	valor de la función objetivo

1.8.4. Algoritmo genético básico

En **Araujo 2009** se propuso un algoritmo genético básico, con el objetivo de explicar con claridad el funcionamiento de un AG. El termino básico o simple, es debido a que

en cada una de sus etapas se aplican las elecciones más sencillas posibles. El algoritmo inicia con una población generada aleatoriamente. La función de adaptación, es una función matemática para la que se busca el valor óptimo en un determinado intervalo. El algoritmo entra a un ciclo donde el primer paso es una selección de individuos. Esta selección se realiza de tal manera que solo permanezcan los individuos mejor adaptados. Los individuos a cruzar se eligen de forma consecutiva, ya que se supone que el proceso de selección ha reubicado a los individuos de forma eficiente. Se aplica una mutación aleatoria y se determina el nivel de adaptación de la nueva generación de individuos. El criterio de paro, es un número máximo de generaciones en las que no hubo mejora de aptitudes.

El esquema general de un algoritmo genético básico es el siguiente:

Entrada:	Conjunto de datos de entrada X.
Salida:	Conjunto de los mejores datos aptos para resolver el problema
1:	*Crear población inicial*
2:	*Computar población inicial*
3:	**WHILE** *condición de paro no se cumple* **Do**
4:	*Selección de individuos para la reproducción*
5:	*Cruza de individuos*
6:	*Mutación de individuos*
7:	*Computar la nueva generación* **END**

La estructura se describe con más detalle a continuación:

Población inicial

Los individuos de la población inicial suelen ser cadenas de ceros y unos generados de forma completamente aleatoria. Es decir, se va generando cada gen, con una función que devuelve un cero o un uno con igual probabilidad. Es importante dotar al algoritmo genético de población con suficientemente variedad, para poder explorar todas las zonas del espacio de búsqueda.

Selección de individuos

La idea básica de selección, es utilizar una distribución de probabilidad de selección de una cadena, donde la probabilidad es directamente proporcional a la función de aptitud. Es decir, el proceso de selección debe favorecer la cantidad de copias de los individuos más adaptados. Las técnicas de selección usadas pueden clasificarse en tres

grupos: selección proporcional, selección mediante torneo y selección de estado uniforme. Sin embargo, en este trabajo solo se analizarán algunas técnicas del grupo selección proporcional, para un estudio más a fondo sobre las demás técnicas puede consultar la referencia **Araujo 2009**.

Dos técnicas conocidas dentro de las técnicas de selección proporcional son la ruleta y sobrante estocástico. Estas se describen a continuación.

La Ruleta

Está método ha sido el más comúnmente utilizado desde los inicios de los AGs. El algoritmo presenta el problema de que el individuo menos apto puede ser seleccionado más de una vez. Sin embargo, su popularidad se debe a su simplicidad. El algoritmo de la Ruleta es el siguiente:

- Calcular la suma de valores esperados T.
- Repetir N veces (N es el tamaño de la población)
    - Generar un número aleatorio r entre 0.0 y T
    - Ciclar a través de los individuos de la población sumando los valores esperados hasta que la suma sea mayor o igual a r.
    - El individuo que haga está suma exceda el límite es el seleccionado.

Sobrante Estocástico

El sobrante estocástico reduce los problemas de la ruleta, pero puede causar convergencia prematura al introducir una mayor precisión de selección. La idea principal es asignar deterministicamente las partes enteras de los valores esperados para cada individuo y luego usar otro esquema para la parte fraccionaria. El algoritmo es el siguiente:

- Asignar de manera determinística el conteo de valores esperados a cada individuo (valores enteros)
- Los valores restantes (sobrantes del redondeo) se usan probabilísticamente para rellenar la población.

Cruza

Este es un método de fusión sobre la información genética de dos individuos. Este proceso provee un mecanismo para heredar características a su descendencia donde intervienen ambos padres.

La forma más simple del operador de cruce es el cruce mono punto, que consiste en seleccionar una única posición en la cadena de ambos padres e intercambiar las partes divididas por dicha posición. La Figura 1.6 muestra un ejemplo de cruce.

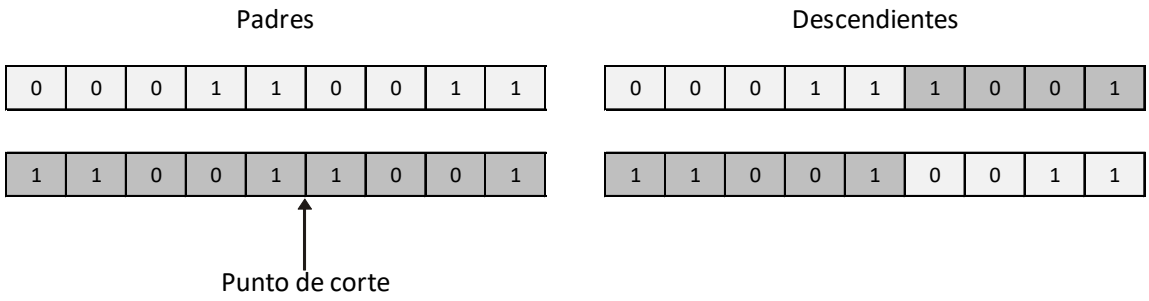

FIGURA 1.6: Cruza de dos cadenas binarias y sus descendientes correspondientes

Mutación

La mutación es un proceso donde el material genético puede ser alterado en forma aleatoria, debidamente a un error en la reproducción o la deformación de genes. A diferencia de la genética humana, la probabilidad en un algoritmo genético es mayor. De hecho en un algoritmo genético, la mutación es una forma de evitar caer en mínimos locales.

La forma más sencilla de mutación consiste en cambiar el valor de una de las posiciones de la cadena. Si el valor es cero pasa a uno, y si es uno pasa a cero. La Figura 1.7 muestra un ejemplo:

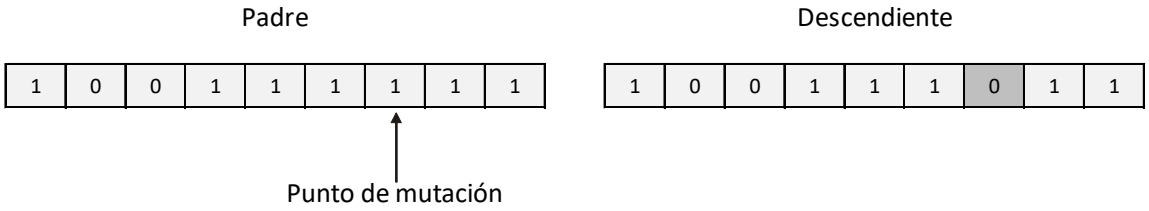

FIGURA 1.7: Mutación de una cadena binaria

Condición de paro

Es necesario especificar las condiciones en las que el algoritmo deja de evolucionar y se presenta la mejor solución encontrada. La condición de paro más sencilla, se

presenta al detectar que la mayor parte de la población ha convergido a una forma similar, careciendo de la suficiente diversidad para que tenga sentido continuar con la evolución.

El principal inconveniente de los métodos de envoltura es la cantidad de cálculos necesarios para obtener el subconjunto de características. Para cada evaluación de subconjunto, el predictor crea un nuevo modelo, es decir, el predictor se entrena para cada subconjunto y se prueba para obtener la precisión del clasificador. Si el número de muestras es grande, la mayor parte de la ejecución del algoritmo se gasta en el entrenamiento del predictor. En algunos algoritmos, como la selección de características usando GA, el mismo subconjunto de características puede evaluarse varias veces, ya que las precisiones del clasificador para los subconjuntos evaluados no se almacenan para su recuperación futura. Otro inconveniente de utilizar el rendimiento del clasificador como función objetivo es que los clasificadores son propensos a sobre entrenamiento **Kohavi 1997**. El sobre entrenamiento se produce si el modelo de clasificador aprende demasiado bien los datos y proporciona una capacidad de generalización deficiente. El clasificador puede introducir sesgo y aumenta el error de clasificación.

El uso de la precisión de la clasificación en la selección de subconjuntos puede resultar en un subconjunto de características incorrectas con alta precisión pero una generalización deficiente. Para evitar esto, se puede utilizar un conjunto de pruebas de reserva para guiar la precisión de la predicción de la búsqueda **Kohavi 1997**.

Metodología

En este Capítulo se muestran la metodología llevada a cabo en los experimentos. En los experimentos llevados a cabo se trabajaron con diferentes técnicas selección de características. la Figura 1.8 muestra la metodología seguida en el desarrollo de los experimentos llevados a cabo en esta tesis. Primero, se seleccionan algunos conjuntos de datos con diferentes dimensionalidades y tamaños, se normalizan los conjuntos de datos, se aplican algunas técnicas de selección de características con los parámetros acordes y se prueba el desempeño. Cada uno de los procesos se detalla en este Capítulo.

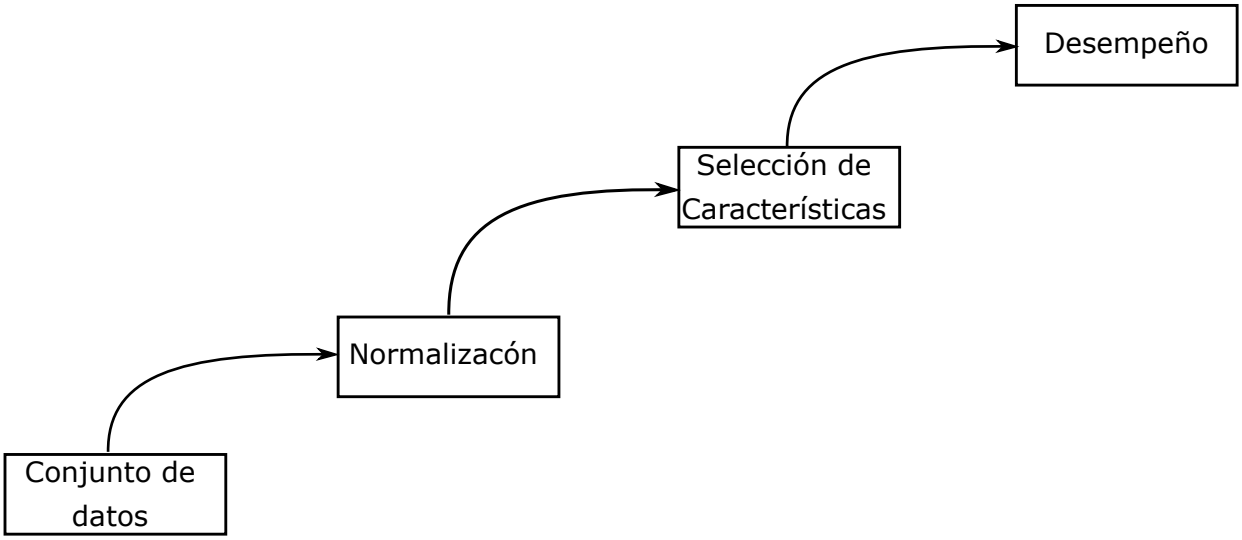

FIGURA 1.8: Pasos más usuales en métodos de selección de características por filtrado

1.9. Conjuntos de datos

Para validar de manera justa el rendimiento de los diferentes métodos de selección de características, se utilizaron 17 conjuntos de datos de referencia en experimentos de simulación. Estos conjuntos de datos están disponibles en el repositorio de aprendizaje automático de la UCI (Blake et al 1998), y la mayoría de ellos se utilizan con frecuencia

en publicaciones. La Tabla 1.3 muestra los conjuntos de datos con los que se realizaron los experimentos, así como el tamaño de cada conjunto de datos, sus atributos y numero de clases.

La documentación completa de la información original se puede obtener en el sitio web de la UCI. Es importante tener en cuenta que estos conjuntos de datos tienen clases nominales y difieren mucho en el tamaño de la muestra y en el número de características (de 4 a 785).

CUADRO 1.3: Conjuntos de datos utilizados en los experimentos

Conjunto	Tamaño	Atributos	Clases
Anneal	898	39	6
Audiology	226	70	24
Autos	202	26	7
Balance	625	5	3
Breast Cancer	286	10	2
Car	1210	7	4
Colic	368	23	2
Convex	8000	785	2
Pima Diabetes	768	9	2
German	700	21	2
Glass	214	10	7
Hay	373	10	4
Hepatitis	155	20	2
Hypothyroid	3772	30	4
Ionosphere	351	35	2
Iris	150	4	3
Kr-vs-Kp	3196	37	2
Labor	57	17	2
Letter	20000	17	26
Lymphography	148	19	4

1.10. Normalización de datos

En reconocimiento de patrones, la normalización puede ser realizada por diferentes razones. En los casos más simples, la normalización de las calificaciones significa ajustar los valores medidos en diferentes escalas a una escala teóricamente común, a menudo antes de promediar. En casos más complicados, la normalización puede referirse a ajustes más sofisticados donde la intención es alinear todas las distribuciones de probabilidad de los valores ajustados.

La normalización se refiere a la creación de versiones de estadísticas modificadas, donde la intención es que estos valores normalizados permitan la comparación de valores normalizados correspondientes para diferentes conjuntos de datos de una manera que elimine los efectos de ciertas influencias generales.

En los experimentos llevados a cabo, cada conjunto de datos utilizado fue normalizado dentro de un determinado rango de valores para que los atributos sean comparables en algún sentido. Debido a que los parámetros calculados se encuentran en rangos distintos, es necesario realizar una transformación de los datos. Para ello se implementó la normalización puntuación estándar, si la media de la población y la desviación estándar son conocidas la puntuación estándar de un vector x es calculada como:

$$z = \frac{x - \mu}{\sigma} \tag{1.23}$$

donde μ es la media de la población y σ es la desviación estándar de la población

El valor absoluto de z representa la distancia entre la puntuación bruta y la media poblacional en unidades de la desviación estándar. z es negativo cuando el puntaje bruto está por debajo de la media, positivo cuando está arriba de la media.

Calcular z usando esta fórmula requiere la media de la población y la desviación estándar de la población, no la media de la muestra o la desviación de la muestra. Pero conocer la media real y la desviación estándar de una población a menudo no es realista, excepto en casos como las pruebas estandarizadas, donde se mide a toda la población.

Cuando se desconoce la media de la población y la desviación estándar de la población, la puntuación estándar se puede calcular utilizando la media de la muestra y la desviación estándar de la muestra como estimaciones de los valores de la población.

En estos casos, la puntuación z es

$$z = \frac{x - \bar{x}}{S} \tag{1.24}$$

donde $\bar{x}$ es la media de la muestra y S es la desviación estándar de la muestra.

Hay dos ventajas principales de tener un conjunto de datos normalizado:

1. Mayor consistencia. La información se almacena en un solo lugar y en un solo lugar, lo que reduce la posibilidad de datos inconsistentes.

2. Mapeo de objeto a datos más fácil. Los conjuntos de datos normalizados en general están más cercanos conceptualmente a los conjuntos orientados a objetos

porque estos promueven una alta cohesión y un acoplamiento flexible entre clases dando como resultado soluciones similares.

La principal desventaja de la normalización es un rendimiento más lento.

1.11. Selección de características

La selección de características no sólo implica la reducción de dimensionalidad, sino también la elección de atributos en función de su utilidad para el análisis. A fin de simplificar el proceso de selección de características es recomendable ordenar el espacio de características con base en su capacidad para describir el objeto, de este modo una vez ordenado se debe determinar cuántas de estas características serán consideradas en el proceso de clasificación.

El algoritmo general del método utilizado en los experimentos realizados se muestra a continuación:

Require: S- Datos con características $X, |X| = n$
J- Medida de evaluación a ser optimizada
GS- Operador
Ensure: Solución- Subconjunto de características c $L := puntoinicial(X)$
Solución:= Mejor L de acuerdo a J
repeat
1: $L :=$ Estrategia de busqueda $(L, GS(J), X)$;
2: $X' :=$ Mejor L de acuerdo a J
3: if $J(X') \geq J(Solucion)$ or $(J(X') = J(Solucion) and |X'| < |Solucion|)$ then $Solucion := X'$;
Stop(J,L)

Algoritmo 1: Algoritmo general de selección de características)

1.11.1. Técnica de filtrado

Existen enormes esfuerzos de investigación en el desarrollo de medidas de desempeño indirectas, basadas principalmente en las cuatro medidas de evaluación (información, distancia, dependencia y consistencia), para seleccionar las características. Este modelo se llama el modelo de filtro. El nombre "filtro"procede de filtrar las características no deseadas antes de aprender. Se utilizan heurísticas basadas en las características generales de los datos para evaluar la bondad de los subconjuntos de características.

Las razones que influyen en el uso de los filtros son aquellas relacionadas con la eliminación de ruido, la simplificación de datos y el aumento del rendimiento. Los métodos de filtrado pueden manejar datos de alta dimensión y proporcionan subconjuntos generales de características que pueden ser útiles para cualquier tipo de proceso de aprendizaje;

Un modelo de filtro consta de dos etapas

1. Una primera etapa que utiliza medidas como información, distancia, dependencia o consistencia, con independencia del algoritmo de aprendizaje;

2. Una segunda etapa de aprendizaje y prueba, el algoritmo aprende de los datos de entrenamiento con el mejor subconjunto de funciones obtenido y probado sobre los datos de prueba.

El modelo de filtro tiene varias propiedades:

1. Medir la incertidumbre, las distancias, la dependencia o la consistencia suele ser más barato que medir la precisión de un proceso de aprendizaje. Por lo tanto, los métodos de filtro son generalmente más rápidos.

2. no se basa en un sesgo de aprendizaje en particular, de tal manera que las características seleccionadas se pueden usar para aprender diferentes modelos.

3. puede manejar datos de mayor tamaño, debido a la simplicidad y la baja complejidad de las medidas de evaluación.

En los experimentos realizados se utilizó correlación como método de filtraje.

1.11.2. Técnica envolvente

La forma más simple de selección de características consiste en utilizar un clasificador como método evaluador para decidir la inserción o eliminación de una determinada característica en el subconjunto, utilizando cualquier métrica para el desempeño predictivo. El objetivo es sencillo; para lograr la mayor precisión predictiva posible se seleccionan las características que logran esto para un algoritmo de aprendizaje fijo. Este modelo es el llamado modelo de envoltura.

En los experimentos llevados a cabo se seleccionó un algoritmo de búsqueda hacia adelante comenzando con un conjunto vacío.

1.11.3. Técnica envolvente basada en algoritmo genético

El algoritmo genético (GA) se puede usar para encontrar el subconjunto de características en donde los bits del cromosoma representan si la característica está incluida o no. Se puede encontrar el máximo global para la función objetivo que proporciona el mejor subconjunto subóptimo. Aquí nuevamente la función objetivo es el desempeño del predictor.

En los experimentos llevados a cabo se utilizó un algoritmo genético básico con las siguientes características:

1. Los mejores N individuos se eligen del grupo de padres y descendientes, es decir, una mejor descendencia reemplaza a los padres menos aptos.

2. Se utiliza un operador de cruza media uniforme altamente disruptivo (HUX) que cruza exactamente la mitad de los alelos no coincidentes, en donde los bits a cruzar se seleccionan al azar.

3. Durante el paso de reproducción, cada miembro de la población principal se selecciona al azar sin reemplazo y se empareja para el apareamiento. No se cruzan todas las parejas, pero antes de aparearse, se calcula la distancia de Hamming entre los padres y, si la mitad de esta distancia no supera un umbral d, no están acopladas. El umbral generalmente se inicializa a L = 4, donde L es la longitud del cromosoma. Si no se obtiene descendencia en la generación, el umbral se reduce en uno. Debido a estos criterios de apareamiento de apareamiento de solo padres diversos, la población converge a medida que el umbral disminuye.

4. Si no se genera descendencia y el umbral se reduce a cero, se introduce una mutación cataclísmica para crear una nueva población. La mejor persona en la población padre actual se toma como la plantilla para crear la nueva población. El resto N- 1 individuos se obtienen cambiando aleatoriamente un porcentaje (35-40 %) de bits de la plantilla. La mutación regular después del paso de cruce se omite cada vez y la mutación mencionada se realiza si es necesario.

Resultados experimentales

En esta sección se realiza un análisis de la influencia de los parámetros en el funcionamiento de cada uno de los métodos utilizados en nuestro proyecto, así como una comparativa global de los resultados obtenidos en cada uno de los métodos para diferentes galerías de prueba.

Antes de comenzar con el análisis de datos, conviene establecer una forma de comparar los resultados que se obtengan de la ejecución de los métodos. Supongamos que se realiza la consulta de una imagen sobre una base de datos, produciéndose varias imágenes como resultado. En una aplicación de esta clase, hay dos tipos de errores que nos interesa diferenciar:

1. Falsos positivos. Se trata de resultados que han sido devueltos incorrectamente, esto es, que se han reconocido como pertenecientes a la categoría consultada pero que no lo son.

2. Falsos negativos. Son imágenes que no son reconocidas como pertenecientes a la categoría consultada, o lo son con un bajo nivel de aceptación, cuando en realidad deberán ser aceptadas.

Para representar la variación de estos radios para diferentes ajustes de las técnicas de comparación, usaremos una métrica de comparación que se denomina área bajo la curva AUC-ROC (Receiver Operating Characteristic) que representan la tasa de falsos positivos frente a la tasa de falsos negativos. En estas curvas, se ha de maximizar el área que cada curva deja por encima, minimizando así el error global cometido. También es interesante conocer en qué punto la curva intersecciona con la recta $f(x) = x$. Esta intersección significa que, para ese ajuste de parámetros concreto, el número de falsos positivos y negativos se iguala, lo que puede ser interesante. Se utilizará como medida de bondad alternativa el área sobre la curva ROC.

1.12. Reducción de dimensionalidad

La tabla 1.4 muestra los resultados en la disminución de la dimensionalidad de cada unos de los conjuntos de datos al utilizar las diferentes técnicas de selección de características. La primera y segunda columna en la Tabla describen el conjunto de datos y la dimensión original del conjunto de datos. La tercera columna nos da la dimensionalidad del conjunto de datos al utilizar un algoritmo genético para seleccionar características.

CUADRO 1.4: Conjuntos de datos utilizados en los experimentos

Conjunto	Atributos	GA	Filtro	Envolvente
Anneal	39	15	19	5
Audiology	70	34	7	5
Autos	26	8	6	3
Balance	5	3	3	4
Breast Cancer	10	6	6	3
Car	7	6	3	3
Colic	23	10	4	4
Convex	785	35	8	4
Pima Diabetes	9	5	4	5
German	21	7	4	12
Glass	10	9	6	5
Hay	10	7	7	3
Hepatitis	20	10	11	11
Hypothyroid	30	7	3	3
Ionosphere	35	14	15	4
Iris	4	3	3	3
Kr-vs-Kp	37	10	3	4
Labor	17	8	5	3
Letter	17	12	10	4
Lymphography	19	11	11	3

1.13. Tiempo de entrenamiento

En esta Sección se muestran los resultados de tiempos de entrenamiento obtenidos en los conjuntos de datos utilizados. La Tabla 1.5 muestra los resultados en tiempo de entrenamiento. Las columnas de la Tabla muestran los tiempos de entrenamiento obtenidos utilizando Árboles de decisión (Tiempo-DT), Redes Bayesianas (Tiempo-Bayes), redes neuronales (Tiempo-NN) y *Support Vector Machines* (Tiempo-SMO) utilizando el algoritmo de Optimización Mínima Secuencial (SMO). En la Tabla se muestran los resultados utilizados con varios conjuntos de datos: El conjunto de datos entero (Com),

el conjunto de datos al utilizar la selección de características basada en filtro (Fil), Selección de características utilizando técnicas de envoltura (Env) y utilizando un algoritmo genético para selección de características (GA).

CUADRO 1.5: Tiempos de entrenamiento sobre los distintos conjuntos de datos

Conjunto de datos	Tiempo-DT				Tiempo-Bayes				Tiempo-NN				Tiempo-SMO			
	Com	Fil	Envo	GA	Com	Fil	Env	GA	Com	Fil	Env	GA	Com	Fil	Env	GA
Anneal	0.02	0.02	0.01	0.06	0.14	0.02	0.01	0.03	39.33	16.25	0.39	10.39	1.34	0.45	0.27	0.58
Audiology	0.03	0.01	0.02	0.07	0.05	0.01	0.01	0.04	14.36	2.88	2.44	20.92	1.64	1.41	3.77	1.73
Autos	0.03	0.05	0.01	0.01	0.05	0.01	0.01	0.05	8.87	0.46	0.34	0.89	0.22	0.24	0.11	0.08
Balance	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.75	0.41	0.73	0.56	0.49	0.45	1.6	0.47
Breast Cancer	0.01	0.02	0.01	0.01	0.02	0.02	0.01	0.03	4.71	1.74	0.2	2.59	0.09	0.03	0.07	0.12
Car	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	6.48	2.82	1.42	5.71	1.13	0.55	0.5	1.1
Colic	0.02	0.01	0.01	0.01	0.04	0.02	0.01	0.03	9.76	0.19	0.56	1.79	0.19	0.16	0.04	0.09
Convex	84.09	0.22	0.03	1.56	4.48	0.06	0.03	0.21	8367.38	6330.92	3401.41	937.33	5812.77	3880.13	187.57	30.51
Pima Diabetes	0.02	0.11	0.03	0.03	0.03	0.02	0.01	0.01	0.83	0.45	0.51	0.66	0.82	0.8	0.43	0.02
German	0.03	0.01	0.03	0.01	0.02	0.02	0.02	0.02	23.64	0.45	4.97	2.91	0.38	0.27	0.39	0.48
Glass	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.02	0.63	0.66	0.41	0.52	0.09	0.86	0.31	0.14
Hay	0.04	0.03	0.01	0.03	0.03	0.02	0.01	0.02	0.73	0.58	0.33	0.58	29	35.79	0.16	0.19
Hepatitis	0.01	0.01	0.05	0.01	0.03	0.04	0.03	0.11	0.59	0.34	0.28	0.26	0.03	0.46	0.02	0.01
Hypothyroid	0.09	0.03	0.02	0.06	0.04	0.04	0.08	0.03	44.51	3.26	2.75	5.87	3.25	1.7	1.87	35.46
Ionosphere	0.07	0.05	0.01	0.02	0.05	0.04	0.01	0.01	3.13	0.8	0.18	0.72	32.99	0.01	0.01	17.71
Iris	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.2	0.12	0.11	0.12	0.02	0.02	0.01	0.04
Kr-vs-Kp	0.05	0.01	0.01	0.01	0.04	0.01	0.01	0.01	42.06	1.32	1.65	5.12	2	0.02	0.12	27.42
Labor	0.01	0.01	0.01	0.01	0.03	0.02	0.03	0.01	0.35	0.06	0.03	0.16	0.02	0.01	0.04	0.01
Letter	3.22	3.87	0.07	4.23	0.54	0.36	0.05	0.42	423.64	396.91	157.14	469.23	14.82	5.28	2.25	7.77
Lymphography	0.01	0.01	0.01	0.01	0.01	0.01	0.03	0.02	1.92	0.73	0.12	0.69	0.26	0.03	0.02	0.06

En los resultados se observa que la reducción de dimensionalidad no tiene importancia en términos de tiempo cuando los conjuntos de datos son pequeños. En conjuntos de datos grandes, la selección de características y reducción de la dimensionalidad puede ayudar a reducir los tiempos de entrenamiento y en algunos casos mejorar la precisión de clasificación. Sin embargo, en los conjuntos de datos utilizados no es posible apreciar una ventaja significativa. Esto puede deberse a dos factores:

1. Los conjuntos de datos son muy pequeños y no es notable o significativa la diferencia en los tiempos de entrenamiento con los diferentes conjuntos.

2. Los conjuntos de datos utilizados probablemente ya han sido sometidos a una selección de características preliminar. Debido a que son conjuntos de datos clave para la clasificación y no específicamente para extracción de características.

1.14. Métricas de desempeño

En esta Sección se muestran los resultados de desempeño obtenidos en los conjuntos de datos utilizados. Las Tablas 1.6, 1.7 y 1.8 muestran los resultados obtenidos en

Precisión, F-Measure y AUC-ROC respectivamente. En las columnas de las Tablas se muestran los resultados obtenidos utilizando Árboles de decisión (Tiempo-DT), Redes Bayesianas (Tiempo-Bayes), redes neuronales (Tiempo-NN) y *Support Vector Machines* (Tiempo-SMO) utilizando el algoritmo de Optimización Mínima Secuencial (SMO). En las Tablas se muestran los resultados utilizados con varios conjuntos de datos: El conjunto de datos entero (Com), el conjunto de datos al utilizar la selección de características basada en filtro (Fil), Selección de características utilizando técnicas de envoltura (Env) y utilizando un algoritmo genético para selección de características (GA).

CUADRO 1.6: Desempeños obtenidos utilizando la métrica *Accuracy*

Accuracy																
Conjunto de datos	Decisión trees				Bayesian net				NN				SMO			
	Com	Fil	Env	GA	Com	Fil	Env	GA	Com	Fil	Env	GA	Com	Fil	Env	GA
Anneal	98.44	97.88	76.16	98.21	96.21	96.32	75.87	97.55	98.99	97.77	75.72	98.88	**99.10**	99.33	76.39	**99.10**
Audiology	77.87	69.46	25.22	76.10	76.10	67.69	25.22	74.33	81.41	69.02	25.22	**77.87**	83.18	69.02	25.22	**77.91**
Autos	81.95	76.58	34.14	77.56	68.29	69.26	34.14	69.75	80	68.39	33.17	73.65	**83.19**	67.80	34.14	**81.67**
Balance	76.64	63.52	62.24	73.28	72.32	63.52	62.24	74.72	90.72	62.56	61.12	77.28	**95.85**	63.52	63.52	**77.6**
Breast Cancer	75.52	73.07	70.27	73.07	72.028	73.42	70.27	73.42	76.68	71.67	70.27	71.67	77.79	74.41	70.27	75.87
Car	89.58	76.36	69.66	89.58	84.54	75.86	69.66	84.12	96.03	76.36	69.66	96.03	98.76	76.61	69.66	98.76
Colic	85.32	81.52	71.46	85.32	81.25	81.52	71.46	83.42	80.43	81.52	71.46	79.89	85.05	81.52	71.46	84.51
Convex	63.56	53.7	49.67	61.6	50.13	50.13	50.13	50.13	79.27	50.13	49.96	72.38	80.12	53.65	52.19	79.52
Pima Diabetes	73.82	74.6	73.82	74.86	74.34	74.21	73.56	75.52	75.18	76.43	75.13	75.52	75.39	76.43	75.49	76.82
German	75.14	71.14	69.57	76.57	75.42	70	71.28	75.28	74.71	70.14	65.85	73.57	75.85	70	72.42	76
Glass	66.82	65.88	66.82	68.69	70.56	68.22	65.42	70.56	67.75	68.22	68.22	65.88	71.03	71.02	55.14	70.73
Hay	67.87	64.07	51.74	64.07	61.66	61.12	50.40	61.12	59.78	64.87	45.30	64.87	68.56	64.94	50.40	68.08
Hepatitis	83.87	81.29	81.93	82.58	83.22	85.80	83.87	86.45	80	83.22	81.93	82.58	85.16	85.51	84.51	84.51
Hypothyroid	99.57	96.63	96.63	97.48	98.59	96.60	96.60	97.69	92.55	93.08	93.08	96.28	99.6	92.28	92.28	94.93
Ionosphere	91.453	90.59	74.92	90.59	89.45	90.59	74.92	90.02	91.16	93.44	74.92	90.88	92.03	87.74	74.92	92.12
Iris	96	96	71.33	96	92.66	94.66	63.33	94.66	97.33	95.33	74	95.33	97.82	96	78.66	97.82
Kr-vs-Kp	99.43	66.05	90.42	94.14	87.92	66.05	90.42	89.48	99.34	64.64	90.42	94.33	99.81	66.05	90.42	94.14
Labor	73.68	80.70	78.94	77.193	87.71	89.47	75.43	88.21	85.96	89.47	78.94	85.96	96.49	88.36	78.94	95.13
Letter	87.98	87.35	56.39	88.36	74.36	73.145	56.39	74.67	82.09	75.89	56.39	78.08	88.05	75.31	56.39	87.58
Lymphography	77.02	77.02	56.75	74.32	85.81	82.10	56.75	81.75	84.45	81.75	55.40	81.08	86.48	85.81	56.75	83.10

Los resultados obtenidos muestran que aún cuando los tiempos de entrenamiento obtenidos con redes Bayesianas y Árboles de decisión son muy buenos, los desempeños de estos algoritmos no son competitivos en comparación con los obtenidos con las redes neuronales o SVM. En cuanto a los resultados obtenidos con la métrica *Accuracy*, es importante puntualizar que los resultados obtenidos con esta métrica son importantes cuando los conjuntos de datos son balanceados, sin embargo no nos dan ninguna medida de la sensitividad y especificidad del clasificador.

La Tabla 1.7 muestra los resultados obtenidos con *F-measure*, que es otra métrica de desempeño utilizada en aprendizaje máquina. Cuando construye un modelo para un problema de clasificación, casi siempre se obtiene la precisión de ese modelo como el número de predicciones correctas de todas las predicciones realizadas. Esta precisión se podría interpretar como la exactitud de la clasificación.

CUADRO 1.7: Desempeños obtenidos utilizando la métrica *F-measure*

F-Measure																
Conjunto de datos	Decisión trees				Bayesian net				NN				SMO			
	Com	Fil	Env	GA	Com	Fil	Env	GA	Com	Fil	Env	GA	Com	Fil	Env	GA
Anneal	0.984	0.979	0.659	0.982	0.964	0.965	0.663	0.975	0.99	0.975	0.663	0.989	0.991	0.993	0.663	0.991
Audiology	0.754	0.652	0.102	0.737	0.719	0.641	0.102	0.713	0.794	0.654	0.102	0.754	0.805	0.663	0.102	0.763
Autos	0.822	0.766	0.189	0.777	0.684	0.691	0.189	0.692	0.801	0.668	0.22	0.735	0.835	0.672	0.24	0.817
Balance	0.749	0.606	0.703	0.595	0.694	0.606	0.717	0.595	0.911	0.599	0.741	0.586	0.961	0.606	0.744	0.953
Breast Cancer	0.713	0.685	0.58	0.685	0.711	0.727	0.58	0.727	0.647	0.705	0.58	0.705	0.718	0.65	0.58	0.721
Car	0.894	0.75	0.572	0.894	0.839	0.741	0.572	0.839	0.961	0.749	0.572	0.961	0.988	0.754	0.572	0.988
Colic	0.85	0.817	0.67	0.85	0.813	0.817	0.67	0.833	0.805	0.817	0.67	0.799	0.848	0.817	0.67	0.842
Convex	0.633	0.505	0.335	0.453	0.335	0.335	0.335	0.335	0.785	0.342	0.412	0.65	0.80	0.505	0.518	0.790
Pima Diabetes	0.736	0.742	0.731	0.743	0.742	0.738	0.729	0.75	0.751	0.762	0.739	0.753	0.768	0.752	0.768	0.757
German	0.746	0.591	0.659	0.762	0.744	0.605	0.678	0.739	0.744	0.635	0.649	0.73	0.748	0.605	0.657	0.749
Glass	0.668	0.636	0.649	0.687	0.686	0.655	0.624	0.686	0.659	0.655	0.652	0.64	0.692	0.701	0.502	0.672
Hay	0.676	0.638	0.495	0.638	0.613	0.61	0.48	0.61	0.596	0.646	0.399	0.646	0.682	0.630	0.426	0.680
Hepatitis	0.825	0.797	0.79	0.808	0.837	0.862	0.843	0.867	0.803	0.79	0.815	0.818	0.849	0.843	0.831	0.834
Hypothyroid	0.995	0.969	0.969	0.975	0.986	0.968	0.968	0.978	0.926	0.907	0.907	0.961	0.997	0.886	0.886	0.941
Ionosphere	0.913	0.905	0.703	0.905	0.894	0.905	0.703	0.899	0.909	0.933	0.703	0.908	0.922	0.873	0.703	0.927
Iris	0.96	0.96	0.708	0.96	0.927	0.947	0.615	0.947	0.973	0.953	0.737	0.953	0.976	0.96	0.78	0.976
Kr-vs-Kp	0.994	0.66	0.904	0.941	0.879	0.660	0.904	0.894	0.993	0.646	0.904	0.943	0.998	0.660	0.904	0.941
Labor	0.736	0.807	0.789	0.771	0.877	0.894	0.754	0.882	0.859	0.894	0.789	0.859	0.964	0.883	0.789	0.951
Letter	0.879	0.873	0.563	0.883	0.743	0.731	0.563	0.746	0.820	0.758	0.563	0.780	0.880	0.753	0.563	0.875
Lymphography	0.770	0.770	0.567	0.743	0.858	0.821	0.567	0.817	0.844	0.817	0.554	0.810	0.864	0.858	0.567	0.831

Sin embargo, la precisión de la clasificación por sí sola no suele ser suficiente información para tomar esta decisión. Para ello es necesario, además de evaluar la solidez del modelo mediante el uso de validación cruzada sobre datos no vistos utilizar otras métricas de desempeño. Una de esas métricas es *F-measure*.

La métrica *F-measure*, tambien conocida como *F-score* se utiliza para medir la precisión de una prueba y equilibra el uso de la precisión y *recall* para hacerlo. La métrica *F-measure* puede proporcionar una medida más realista del rendimiento de una prueba utilizando tanto la precisión como recall.

Los resultados de la Tabla 1.7 muestran que los resultados obtenidos con el algoritmo genético superan a los resultados obtenidos con la técnica de filtrado. Esto nos permite argumentar que el algoritmo genético afecta menos el equilibrio del clasificador.

Por otro lado, el área bajo la curva muestra el rendimiento de un modelo de clasificación en todos los umbrales de clasificación. La curva ROC se desarrolló para medir la eficacia en la detección de objetos enemigos en campos de batalla mediante pantallas de radar. El análisis ROC se aplicó posteriormente en diferentes campos de investigación como medicina, radiología, psicología y otras áreas durante varias décadas y recientemente ha encontrado aplicación en áreas como aprendizaje automático y minería de datos. El área bajo la curva es definida por dos parámetros: Tasa de verdaderos positivos (TPR) y Tasa de falsos positivos (FPR). Una curva ROC representa TPR frente a

CUADRO 1.8: Desempeños obtenidos utilizando la métrica *AUC-ROC*

ROC Area																
Conjunto de datos	Decisión trees				Bayesian net				NN				SMO			
	Com	Fil	Env	GA	Com	Fil	Env	GA	Com	Fil	Env	GA	Com	Fil	Env	GA
Anneal	0.995	0.991	0.497	0.994	0.992	0.992	0.497	0.988	0.99	0.992	0.497	0.985	0.99	0.995	0.505	0.991
Audiology	0.927	0.907	0.418	0.919	0.956	0.944	0.423	0.959	0.942	0.913	0.443	0.961	0.978	0.919	0.447	0.973
Autos	0.922	0.917	0.473	0.912	0.899	0.909	0.473	0.907	0.929	0.817	0.485	0.881	0.941	0.832	0.519	0.927
Balance	0.811	0.630	0.775	0.618	0.886	0.63	0.815	0.622	0.977	0.692	0.877	0.692	0.987	0.662	0.792	0.981
Breast Cancer	0.584	0.607	0.483	0.607	0.698	0.700	0.507	0.7	0.623	0.619	0.491	0.619	0.655	0.564	0.500	0.625
Car	0.972	0.896	0.495	0.972	0.969	0.894	0.595	0.969	0.984	0.899	0.606	0.984	0.999	0.849	0.606	0.999
Colic	0.81	0.78	0.678	0.815	0.843	0.781	0.673	0.865	0.857	0.825	0.692	0.85	0.827	0.814	0.628	0.819
Convex	0.636	0.512	0.5	0.61	0.5	0.5	0.5	0.5	0.790	0.52	0.61	0.72	0.801	0.532	0.545	0.802
Pima Diabetes	0.751	0.745	0.721	0.791	0.806	0.800	0.771	0.802	0.793	0.806	0.801	0.809	0.802	0.781	0.800	0.799
German	0.712	0.494	0.606	0.75	0.774	0.696	0.675	0.773	0.734	0.711	0.614	0.726	0.671	0.504	0.55	0.671
Glass	0.807	0.786	0.793	0.815	0.876	0.835	0.805	0.876	0.847	0.818	0.828	0.833	0.882	0.854	0.742	0.878
Hay	0.809	0.786	0.697	0.786	0.833	0.831	0.703	0.831	0.794	0.823	0.684	0.823	0.819	0.774	0.681	0.806
Hepatitis	0.708	0.69	0.568	0.628	0.882	0.875	0.865	0.874	0.823	0.568	0.795	0.826	0.756	0.863	0.694	0.706
Hypothyroid	0.993	0.982	0.982	0.987	0.997	0.982	0.982	0.994	0.873	0.827	0.827	0.979	0.996	0.5	0.5	0.745
Ionosphere	0.892	0.89	0.614	0.894	0.948	0.964	0.614	0.954	0.915	0.955	0.632	0.928	0.902	0.84	0.691	0.915
Iris	0.968	0.968	0.803	0.968	0.98	0.982	0.775	0.982	0.998	0.995	0.849	0.995	0.998	0.968	0.914	0.998
Kr-vs-Kp	0.999	0.666	0.931	0.965	0.952	0.666	0.951	0.933	0.999	0.676	0.938	0.981	0.999	0.676	0.901	0.998
Labor	0.695	0.78	0.684	0.735	0.974	0.883	0.7	0.941	0.923	0.961	0.786	0.95	0.973	0.832	0.769	0.963
Letter	0.954	0.956	0.747	0.957	0.979	0.978	0.747	0.980	0.954	0.930	0.747	0.938	0.98	0.965	0.747	0.972
Lymphography	0.785	0.778	0.516	0.775	0.916	0.912	0.509	0.922	0.92	0.915	0.525	0.912	0.869	0.869	0.523	0.846

FPR en diferentes umbrales de clasificación. Reducir el umbral de clasificación clasifica más elementos como positivos, por lo que aumentarán tanto los falsos positivos como los verdaderos positivos.

La Tabla 1.8 muestra los resultados obtenidos con la métrica AUC-ROC. La Tabla muestra una vez más que los resultados menos afectados al reducir la dimensionalidad de los conjuntos son aquellos donde se ha utilizado el algoritmo genético para seleccionar características, esto en comparación con el conjunto de datos completo. Aunque es importante aclarar que en algunos casos el método de filtrado presenta mejores resultados que el algoritmo genético. Es importante mencionar que la técnica de envolvente es la que peores resultados presenta en todos los conjuntos de datos.

Conclusiones

En esta tesis se han realizado diferentes pruebas con diferentes algoritmos en el estado del arte para la selección de características. En los experimentos llevados a cabo se realizaron pruebas con diferentes técnicas de selección (Filtrado, Envolvente y algoritmo genético) de características y técnicas de clasificación (árboles de decisión, redes Bayesianas, redes neuronales y SVM), aunado a ello los resultados se compararon con tres métricas de desempeño diferentes y se peresentan los resultados. En el Capítulo 4 es mostrada la metodología llevada a cabo para la selección de características.

De los resultados obtenidos y mostrados en el Capítulo anterior, se puede argumentar que los métodos de selección de características envolvente son la técnica que más afecta al desempeño de los clasificadores. Por otro lado, la técnica que menos afecta al desempeño del clasificador es el algoritmo genético. Sin embargo, durante los experimentos realizados se pudo observar que es necesario realizar más pruebas utilizando conjuntos de datos sin procesar. Esto podría darnos una mejor idea del comportamiento de las técnicas de selección de características.

Sobre el tiempo de entrenamiento, en las Tablas se puede apreciar que el uso de técnicas de selección de características reduce significativamente el tiempo de entrenamiento, sobre todo es más palpable cuando se utilizan redes neuronales y SVM.

Por otro lado, es muy importante recalcar que la reducción de características que se obtuvo no es tan importante en algunos casos. Esto podría deberse a que la mayoría de los conjuntos de datos posiblemente ya fueron procesados.

Bibliografía

[1] Mario Koppen. The Curse of Dimensionality. (held on the internet), September 4-18 2000. 5th Online World Conference on Soft Computing in Industrial Applications (WSC5).

[2] Evangelista, Paul F., Mark J. Embrechts, and Boleslaw K. Szymanski. "Taming the curse of dimensionality in kernels and novelty detection.İn Applied soft computing technologies: The challenge of complexity, pp. 425-438. Springer Berlin Heidelberg, 2006.

[3] Blum A. L. and Langley P., "Selection of relevant features and examples in machine learning" Artificial Intelligence, Vol. 97, Nos. 1-2, 1997, pp. 245-271.

[4] Dash M. and Liu H., "Feature selection for classification" Intelligent Data Analysis, Vol. 1, 1997, pp. 131-156.

[5] Narendra P. M. and Fukunaga K., "A branch and bound algorithm for feature selection". IEEE Transactions on Computers, Vol. 26, September 1977, pp. 917-922.

[6] Almuallin H. and Dietterich T. G., "Learning with many irrelevant features". Proceedings of Ninth National Conference on Artificial Intelligence, MIT Press, Cambridge, Massachusetts, 1992, pp. 547-552.

[7] Kira K. and Rendell L. A., "The feature selection problem: Traditional methods and a new algorithm". Proceedings of Ninth National Conference on Artificial Intelligence, MIT Press, Cambridge, Massachusetts, 1992, pp. 129-134.

[8] Cardie C., "Using decision trees to improve case-based learning". Proceedings of Tenth International Conference on Machine Learning, Morgan Kaufmann Prublishers, University of Massachusetts, Amherst, June 1993, pp. 25-32.

[9] Liu H. and Setiono R., “Feature selection and classification – A probabilistic wrapper approach”. Procedings of Ninth International Conference on Industrial and Engineering Applications of AI and ES, Fukuoka, Japan, June 1996, pp. 419-424.

[10] Vafaie, H. and Imam, I.F., "Feature selection methods: Genetic algorithm vs. greedy-like search,"Proceedings of the 3rd International Fuzzy Systems and Intelligent Control Conference, Louisville, KY, March 1994.

[11] John G. H., Kohavi R. and Pfleger P., “Irrelevant features and the subset selection problem”. Proceedings of the Eleventh International Conference on Machine Learning. New Brunswick, Morgan Kaufmann, 1994, pp. 121-129.

[12] Mucciardi A. N. And Gose E.E., “A comparison of seven techniques for choosing subsets of pattern recognition”, IEEE Transactions on Computers, Vol. 20, September 1971, pp.1023-1031.

[13] Langley P. and Sage S., “Oblibious decision trees and abstract cases”. Working Notes of the AAAI94 Workshop on Case-Based Reasoning, Seattle, WA: AAAI Press, 1994, pp.113-117.

[14] Forman G. An extensive empirical study of feature selection metrics for text classification. J Mach Learn Res 2003;3:1289–306.

[15] Yang Y, Perdersen JO. A comparative study on feature selection in text categorization. International conference on machine learning; 1997.

[16] Javed K, Babri HA, Saeed M. Feature selection based on class-dependent densities for high-dimensional binary data. IEEE Trans Knowl Data Eng 2010;24.

[17] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27.

[18] Fleuret F. Fast binary feature selection with conditional mutual information. Mach Learn Res 2004;5:1531–55.

[19] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.

[20] Kira K, Rendell LA. The feature selection problem: traditional methods and a new algorithm. In: Proceedings of tenth national conference on artificial intelligence; 1992. p. 129–34.

[21] Acuna E, Coaquira F, Gonzalez M. A comparison of feature selection procedures for classifier based on kernel density estimation. Proc Comput Commun Control Technol 2003;1:468–72.

[22] Xu Z, King I, Lyu MR-T, Jin R. Discriminative semi-supervised feature selection via manifold regularization. IEEE Trans Neural Networks 2010;21.

[23] Goldberg D. Genetic algorithms in search, optimization and machine learning. Addison-Wesley; 1989.

[24] Sun Y, Babbs C, Delp E. A comparison of feature selection methods for the detection of breast cancers in mammograms: adaptive sequential floating search vs. genetic algorithm. Conf proc IEEE eng med biol soc, vol. 6.

[25] Nakariyakul S, Casasent DP. An improvement on floating search algorithms for feature subset selection. Pattern Recog 2009;42:1932–40.

[26] Alexandridis A, Patrinos P, Sarimveis H, Tsekouras G. A two-stage evolutionary algorithm for variable selection in the development of rbf neural network models. Chemomet Intell Lab Syst 2005;75:149–62.

[27] Jouan-Rimbaud D, Massart DL, Leardi R, Noord OED. Genetic algorithms as a tool for wavenumber selection in multivariate calibration. Anal Chem 67.

[28] Yang J, Honavar V. Feature subset selection using a genetic algorithm. IEEE Intell Syst Appl 1998;13:44–9.

[29] Puch W, Goodman E, Pei M, Chia-Shun L, Hovland P, Enbody R. Further research on feature selection and classification using genetic

algorithm. In International conference on genetic algorithm; 1993. p. 557–64.

[30] Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997;97:273–324.

[31] Cesar Augusto Cardona y David Velazquez. Selección de características relevantes usando información mutua. Dyna, Vol. 73, Num 149, Pags. 149-163.