Una revisión de las técnicas de clasificación supervisada en la clasificación automática de textos

A review of supervised classification techniques in automatic text classification

Celina Beltrán; Ivana Barbona
Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, Argentina
beltranc36@yahoo.com.ar

Abstract

The present article is a review which objective is the examination of techniques of multivariate analysis used to classify units. In this paper we compare the performance of the classification methods: Closer Neighbor, Bagging System, Classification Trees, Support Vector Machine, Sequential Minimal Optimization, Logistic Regression, Neural Networks and Discriminant Analysis. For all the methods its functionality and performance is presented describing how it is possible to use them to classify and eventually characterize texts from different genres or disciplines. The classification criterion is the text genre (Scientific / Non-Scientific). The texts characterization is based on the frequency distribution of the morpho-syntactic categories. The texts were classified taking into account simultaneously the measurements made on them. It is considered as a measure for the comparison between methods the misclassification error calculated on a sample of texts not included in the process of construction of the classification rule.
Of the applied methods, Neuronal Networks presents the best performance (3% of poor classification). The next in good performance is the Nearer Neighbor (13% of poor rating) having as main advantages the simplicity of its application and the stability of its behavior. Also acceptable performances were the methods Trees Classification (14% of bad classification) and Discriminant Analysis Quadratic (16,67% of bad classification). It should be noted that, because the groups have different covariance structures, it is expected that the Discriminant Quadratic Analysis will classify better than the Linear Discriminant Analysis (18% of poor classification). On the other hand, it is not possible to know in what way the presence of different covariance structures between the groups affects the remaining methods.
Keywords: Supervised Classification Techniques, Automatic text analysis

Resumen

El presente artículo es una revisión de tema cuyo objetivo es el examen de técnicas de análisis multivariado usadas para clasificar unidades. En este trabajo se compara el desempeño de los métodos de clasificación: Vecino más Cercano, Sistema Bagging, Árboles de Clasificación, Support Vector Machine, Sequential Minimal Optimization, Regresión Logística, Redes Neuronales y Análisis Discriminante. Para todos los métodos se presenta su funcionalidad y desmpeño en la clasificación de textos describiendo cómo es posible utilizarlos para clasificar y eventualmente caracterizar textos de distintos géneros o disciplinas. El criterio de clasificación es el género al que pertenece el texto (Científico / No Científico). La caracterización de los textos está basada en la distribución de frecuencias de las categorías morfo-sintácticas. Los textos se clasificaron teniendo en cuenta simultáneamente las mediciones realizadas sobre ellos. Se considera como medida para la

comparación entre métodos el error de mala clasificación calculada sobre una muestra de textos no incluidos en el proceso de construcción de la regla de clasificación.
De los métodos aplicados, Redes Neuronales presenta el mejor desempeño (3% de mala clasificación). El siguiente en buen desempeño es el del Vecino más Cercano (13% de mala clasificación) teniendo como principales ventajas la simpleza de su aplicación y la estabilidad de su comportamiento. También presentaron desempeños aceptables los métodos Árboles de Clasificación (14% de mala clasificación) y Análisis Discriminante Cuadrático (16,67 % de mala clasificación). Cabe destacar, que debido que los grupos presentan estructuras de covariancias distintas, es de esperar que el Análisis Discriminante Cuadrático clasifique mejor que el Análisis Discriminante Lineal (18% de mala clasificación). Por otro lado, no es posible conocer en de qué manera afecta la presencia de estructuras de covariancias distintas entre los grupos para los métodos restantes.

Palabras clave: Métodos de Clasificación Supervisada, Análisis automático de textos

1. INTRODUCCION

El presente artículo es una revisión de tema cuyo objetivo es el examen de técnicas de análisis multivariado usadas para clasificar unidades. En este trabajo se compara el desempeño de los métodos de clasificación: Vecino más Cercano, Sistema Bagging, Árboles de Clasificación, Support Vector Machine, Sequential Minimal Optimization, Regresión Logística, Redes Neuronales y Análisis Discriminante. Para todos los métodos se presenta su funcionalidad y desmpeño en la clasificación de textos describiendo cómo es posible utilizarlos para clasificar y eventualmente caracterizar textos de distintos géneros o disciplinas. El criterio de clasificación es el género al que pertenece el texto (Científico / No Científico). La caracterización de los textos está basada en la distribución de frecuencias de las categorías morfo-sintácticas. Los textos se clasificaron teniendo en cuenta simultáneamente las mediciones realizadas sobre ellos. Se considera como medida para la comparación entre métodos el error de mala clasificación calculada sobre una muestra de textos no incluidos en el proceso de construcción de la regla de clasificación.

2. MATERIAL Y METODOS

2.1. Diseño de la muestra

El marco muestral para la selección de la muestra de los textos científicos está compuesto por textos académicos, resúmenes de trabajos presentados a congresos y revistas, extraídos de internet pertenecientes a distintas disciplinas. Los textos periodísticos fueron seleccionados de un corpus mayor utilizado por el equipo de investigación INFOSUR. Este corpus se construyó con noticias extraídas de las páginas web de periódicos argentinos (noticias de tipo general, no especializadas en español). La unidad de muestreo fue el texto y la selección de la muestra se llevó a cabo empleando un diseño muestral estratificado.

Luego de obtener las muestras de los estratos, fueron evaluadas y comparadas respecto al número medio de palabras por texto. Se requiere esta evaluación para evitar que la comparación entre los géneros se vea afectada por el tamaño de los textos.

La muestra final para este trabajo quedó conformada de la siguiente manera:

Tabla 1. Conformación de la muestra final

Muestra	Nro. de textos	Cantidad de palabras
Científico	90	14.554
No científico	60	8.080

2.2. Etiquetado: Análisis morfológico de los textos

El software Smorph, analizador y generador morfosintáctico desarrollado en el Groupe de Recherche dans les Industries de la Langue (Universidad Blaise-.Pascal, Clermont II) por Salah Aït-Mokhtar (1998) realiza en una sola etapa la tokenización y el análisis morfológico. A partir de un texto de entrada se obtiene un texto lematizado con las formas correspondientes a cada lema (o a un subconjunto de lemas) con los valores correspondientes. Se trata de una herramienta declarativa, la información que utiliza está separada de la maquinaria algorítmica, en consecuencia, puede adaptarse a distintos usos. Con el mismo software se puede tratar cualquier lengua si se modifica la información lingüística declarada en sus archivos.

El módulo post-smorph **MPS** es un analizador que recibe en entrada una salida Smorph (en formato Prolog) y puede modificar las estructuras de datos recibidos. Ejecuta dos funciones principales: la Recomposición y la Correspondencia, que serán útiles para resolver las ambigüedades que resulten del análisis de Smorph.

La información contenida en estos archivos es la presentada en Beltrán (2009) para implementar el etiquetador.

2.3. Diseño y desarrollo de la base de datos

La información resultante del análisis morfológico se dispuso en una matriz de dimensión: tantas filas como cantidad de objetos lingüísticos tenga el texto y tantas columnas como ocurrencia+lema+valores. Luego, a partir de esta base de datos por palabra, se confeccionó la base de datos por documento que es analizada estadísticamente. La información registrada en esta base corresponde a las siguientes variables:

- CORPUS: Corpus al que pertenece el texto
- TEXTO: Identificador del texto dentro del corpus
- Adj: cantidad de adjetivos del texto
- Adv: cantidad de adverbios del texto
- Cl: cantidad de clíticos del texto
- Cop: cantidad de copulativos del texto
- Det: cantidad de determinantes del texto
- Nom: cantidad de nombres (sustantivos) del texto
- Prep: cantidad de preposiciones del texto
- V: cantidad de verbos del texto
- Otro: cantidad de otras etiquetas del texto
- Total_pal: cantidad total de palabras del texto

2.4. Metodología estadística

Uno de los problemas de los que se ocupan las técnicas estadísticas multivariadas es la clasificación de objetos o unidades en grupos o poblaciones. Dos enfoquen agrupan a las técnicas de clasificación. Uno de ellos es cuando se conocen los grupos o categorías y se pretende ubicar los individuos dentro de estas categorías a partir de los valores de ciertas variables. El segundo enfoque, que no es el utilizado en este trabajo, ocurre cuando no se conocen los grupos de antemano y se pretende establecerlos a partir de los datos observados.
Las técnicas evaluadas en este trabajo tienen por objetivo construir un sistema que permita clasificar unidades en una de las categorías definidas y conocidas previamente en función de las variables relevadas, como así también otras variables que demuestren un aporte significativo en la predicción del grupo de pertenencia.
Luego de la aplicación de cada una de las técnicas definidas en este apartado se debe evaluar la calidad de los resultados, es decir el desempeño para clasificar mediante la validación del mismo. Esto se realiza particionando el conjunto de unidades en dos grupos. Uno es utilizado para la estimación del mismo (entrenamiento del sistema) y el segundo conforma el grupo de validación para la fase de prueba. Se consideró como medida para la comparación entre métodos el error de mala clasificación calculada sobre una muestra de textos no incluidos en el proceso de construcción de la regla de clasificación.
A continuación se presentan algunas de las técnicas multivariadas que tienen por objetivo clasificar unidades en categorías definidas a priori que fueron evaluadas en diferentes aplicaciones por los autores.

2.4.1. Análisis de regresión logística

Esta técnica es un caso particular de los modelos lineales generalizados, modela la probabilidad de que una unidad experimental pertenezca a un grupo en particular considerando información medida o registrada en dicha unidad.
La regresión logística es utilizada en situaciones en las cuales el objetivo es describir la relación entre una variable respuesta categórica, en este caso dicotómica, y un conjunto de variables explicativas que pueden ser tanto categóricas como cuantitativas.
Sea **x** un vector de p variables independientes, esto es, $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$. La probabilidad condicional de que la variable *y* tome el valor 1 (presencia de la característica estudiada), dado valores de las covariables **x** es:

$$P(y = 1/X) = \pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

donde $g(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$

β_0 es la constante del modelo o término independiente
p el número de covariables
βi los coeficientes de las covariables
xi las covariables que forman parte del modelo.
Si alguna de las variables independientes es una variable discreta con k niveles se debe incluir en el modelo como un conjunto de k-1 "variables de diseño" o "variables dummy". El cociente de las probabilidades correspondientes a los dos niveles de la variable respuesta se denomina odds y su expresión es:

$$\frac{P(y=1|X)}{1-P(y=1|X)} = e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}$$

Si se aplica el logaritmo natural, se obtiene el modelo de regresión logística:

$$\log\left(\frac{P(y=1|\,X)}{1-P(y=1|\,X)}\right) = \log(e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

En la expresión anterior el primer término se denomina logit, esto es, el logaritmo de la razón de proporciones de los niveles de la variable respuesta.
Los coeficientes del modelo se estiman por el método de Máxima Verosimilitud y la significación estadística de cada uno de los coeficientes de regresión en el modelo se puede verificar utilizando, entre otros, el test de Wald y el test de razón de verosimilitudes.

Una cuestión importante en este tipo de análisis es determinar si todas las variables consideradas en la función de discriminante contienen información útil y si solamente algunas de ellas son suficientes para diferenciar los grupos. Dado que las variables utilizadas para explicar la respuesta es probable que estén correlacionadas, es posible también que compartan información. Por lo tanto, se puede buscar un subgrupo de variables mediante algún criterio de modo tal que las variables excluidas no contengan ninguna información adicional. Existen varios algoritmos de selección de variables, entre ellos podemos citar:
Método forward: comienza por seleccionar la variable más importante y continúa seleccionando las variables más importantes una por vez, usando un criterio bien definido. Uno de estos criterios involucra el logro de un nivel de significación deseado pre-establecido. El proceso termina cuando ninguna de las variables restantes encuentra el criterio pre-especificado.
Método backward: comienza con el modelo más grande posible. En cada paso descarta la variable menos importante, una por vez, usando un criterio similar a la selección forward. Continúa hasta que ninguna de las variables pueda ser descartada.
Selección paso a paso: combina los dos procedimientos anteriores. En un paso, una variable puede entrar o salir desde la lista de variables importantes, de acuerdo a algún criterio pre-establecido.
En este trabajo se utilizó como evaluación del ajuste del modelo el test de Hosmer-Lemeshow y, dado que el modelo es utilizado para clasificar unidades, se utilizó también la tasa de mala clasificación calculada sobre la muestra independiente excluida en la etapa de estimación.

2.4.2. Árboles de clasificación

Los árboles de clasificación (AC) son una técnica de análisis discriminante no paramétrica que permite predecir la asignación de unidades u objetos a grupos predefinidos en función de un conjunto de variables predictoras. Esto es, dada una variable respuesta categórica, los AC crean una serie de reglas basadas en las variables predictoras que permiten asignar una nueva observación a una de las categorías o grupo.
Es un algoritmo que genera un árbol de decisión en el cual las ramas representan las decisiones y cada una de ellas genera reglas sucesivas que permiten continuar la clasificación. Estas particiones recursivas logran formar grupos homogéneos respecto a la variable respuesta. El árbol determinado puede ser utilizado para clasificar nuevas unidades.
Es un método no-paramétrico de segmentación binaria donde el árbol es construido dividiendo repetidamente los datos. En cada división los datos son partidos en dos grupos mutuamente excluyentes. Comienza el algoritmo con un nodo inicial o raíz a partir del cual se divide en dos sub-grupos o sub-nodos, esto es, se elige una variable y se determina el punto de corte de modo que las unidades pertenecientes a cada nuevo grupo definido sean lo más homogéneas posible. Luego se

aplica el mismo procedimiento de partición a cada sub-nodo por separado. En cada uno de estos nodos se vuelve a repetir el proceso de seleccionar una variable y un punto de corte para dividir la muestra en dos partes homogéneas.
Las divisiones sucesivas se determinan de modo que la heterogeneidad o impureza de los sub-nodos sea menor que la del nodo de la etapa anterior a partir del cual se forman. El objetivo es particionar la respuesta en grupos homogéneos manteniendo el árbol lo más pequeño posible.
La función de impureza es una medida que permite determinar la calidad de un nodo, esta se denota por i(t). Si bien existen varias medidas de impureza utilizadas como criterios de partición, una de las más utilizadas está relacionada con el concepto de entropía

$$i(t) = \sum_{j=1}^{K} p(j/t).\ln\ p(j/t)$$

donde j = 1, ..., k es el número de clases de la variable respuesta categórica y p(j|t) la probabilidad de clasificación correcta para la clase j en el nodo t. El objetivo es buscar la partición que maximiza

$$\Delta i(t) = -\sum_{j=1}^{K} p(j/t).\ln\ p(j/t).$$

De todos los árboles que se pueden definir es necesario elegir el óptimo. El árbol óptimo será aquel árbol de menor complejidad cuya tasa de mala clasificación (TMC) es mínima. Generalmente esta búsqueda se realiza comparando árboles anidados mediante validación cruzada. La validación cruzada consiste, en líneas generales, en sacar de la muestra de aprendizaje o entrenamiento una muestra de prueba, con los datos de la muestra de aprendizaje se calculan los estimadores y el subconjunto excluido es usado para verificar el desempeño de los estimadores obtenidos utilizándolos como "datos nuevos".

2.4.3. Sistema de clasificación BAGGING

La técnica BAGGING (Breiman, 1994) tiene por objetivo combinar distintos clasificadores generados a partir de un mismo conjunto de datos y así lograr una mejora en la predicción de la categoría de pertenencia. Busca mejorar el desempeño de los Árboles de clasificación.

Este procedimiento obtiene muchas muestras de entrenamiento obtenidas, por muestreo Bootstrap, a partir de un único conjunto de datos. Con cada conjunto de entrenamiento obtiene un árbol de clasificación y combina las predicciones de cada uno de ellos para obtener la categoría de pertenencia de una nueva observación. En cada caso se estimó previamente el número de árboles a combinar de modo que el porcentaje de error en la clasificación sea aceptable.

Sea el conjunto de datos E={ $(\mathbf{x_1}, Y_1)\ (\mathbf{x_2}, Y_2)\ (\mathbf{x_3}, Y_3) \ldots (\mathbf{x}_n, Y_n)$ } de tamaño n. A partir de dicho conjunto se generan M muestras mediante el método de Bootstrap, esto es, M muestras aleatorias simples con reposición de tamaño n de E, Ek (k=1,2,…,M) donde cada elemento del conjunto E tiene una probabilidad aproximada de 0.63 de ser seleccionado.

En cada una de las muestras Ek, se obtiene un predictor basado en árboles de clasificación y estos predictores individuales son combinados para obtener una predicción final (predicción Bagging). La predicción por Bagging será la categoría más frecuente hallada en los M predictores individuales.

El algoritmo se resume en los siguientes pasos (Figura 1):

1- Sea E={ $(\mathbf{x_1}, Y_1)\ (\mathbf{x_2}, Y_2)\ (\mathbf{x_3}, Y_3) \ldots (\mathbf{x}_n, Y_n)$ }el conjunto de datos.

2- Se construyen M muestras Bootstrap E_1, E_2,…E_k de tamaño n.

3- Para cada una de ellas se obtiene el predictor g(**x**, E_1), g(**x**, E_2),…, g(**x**, E_M)

4- Se calcula el estimador Bagging mediante

$$gBagg(\mathbf{x}) = \arg\max_{y}(\#\{k : g(\mathbf{x}_1, E_k) = y\}) \text{ para } k=1,2,..,M$$

Figura 1: Esquema del algoritmo de clasificación Bagging

2.4.4. Método del vecino más cercano

La clasificación por el método del vecino más cercano es una de las técnicas no paramétricas de clasificación más utilizadas. La idea en la cual está basado el método es muy simple, para predecir la categoría a la cual pertenece una nueva unidad (clasificar) sólo considera las k unidades del grupo de entrenamiento más cercanas o parecidas a dicha unidad. Este método clasifica a la nueva unidad al grupo al cual pertenece la mayoría de los k vecinos más cercanos del grupo de entrenamiento.

Sea $(\mathbf{x}_1, Y_1)\,(\mathbf{x}_2, Y_2)\,(\mathbf{x}_3, Y_3)\,...(\mathbf{x}_n, Y_n)$ la muestra de entrenamiento, donde la variable Y es la que se refiere a la variable de clasificación y sus niveles corresponden a las distintas categorías a las cuales pertenecen las unidades, y el vector x contiene las covariables utilizadas para asignar la categoría de la variable Y a la cual pertenece la unidad.

La muestra que será utilizada como validación es similar a la de entrenamiento pero sin considerar la variable Y, la cual es conocida pero será utilizada luego de aplicar el sistema para evaluar su desempeño.

Puesto que requiere reconocer las unidades más cercanas a la unidad a clasificar es necesario definir una medida de distancia entre unidades. Esta medida debe ser calculada en función del conjunto de covariables x cuya información se considera relevante para la clasificación. Para variables cuantitativas, como las utilizadas en esta aplicación, algunas de las medidas de distancia usuales son la distancia Euclídea, la distancia de Mahalanobis y otras variantes.

La distancia euclídea entre el punto P y un punto fijo Q con coordenadas P=(x1, x2, ...,xp) y Q=(y1, y2, ...,yp) está dada por

$$d(P,Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2}\,.$$

Una característica de la distancia euclídea es que cada una de las coordenadas contribuye de la misma manera en el cálculo de la distancia. Sin embargo, en muchas situaciones las coordenadas

representan mediciones de diferente magnitud y es deseable que el "peso" de cada coordenada tome en consideración la variabilidad de las mediciones. Esto sugiere distintas definiciones de distancia.

Una distancia "estadística" que tenga en cuenta las distintas variabilidades de las variables se puede construir a partir de las coordenadas estandarizadas, $x_j^* = x_j / \sqrt{S_{jj}}$, para el punto P y $y_j^* = y_j / \sqrt{S_{jj}}$, para el punto Q, con j=1,2,...,p

$$d(P,Q) = \sqrt{\frac{(x_1 - y_1)^2}{S_{11}} + \frac{(x_2 - y_2)^2}{S_{22}} + \ldots + \frac{(x_p - y_p)^2}{S_{pp}}}$$

el "peso" que se le da a la j-ésima coordenada es kj=1/Sjj, para j=1,2,...,p. Si S11=S22=...=Spp, entonces la distancia euclídea es conveniente.

Otro aspecto importante de este método es determinar el valor de k. Si toma un valor grande se corre el riesgo de hacer la clasificación de acuerdo a la mayoría global del conjunto de entrenamiento y no a los parecidos de esta manera se obtendría una predicción constante para toda unidad a clasificar. Por otro lado, si el valor es chico puede perderse exactitud debido a la presencia de ruido en los datos. En esta aplicación el valor de k fue seleccionado buscando minimizar el error de clasificación.

El método del vecino más cercano para clasificar una unidad P se puede describir enunciando unos simples pasos:

1. Se define la distancia entre unidades (puntos) que se va a utilizar
2. Calcular la distancia de P a cada uno de las unidades del conjunto de entrenamiento.
3. Registrar las k unidades más próximas a P.
4. Calcular la frecuencia (cantidad de puntos o unidades), de los k vecinos más cercanos, que pertenecen a cada una de las categorías.
5. Clasificar a la unidad P en la categoría que presente mayor frecuencia.

2.4.5. Redes Neuronales Artificiales: El Perceptrón Multicapa

Las redes neuronales son sistemas pertenecientes a una rama de la inteligencia artificial que emulan al cerebro humano. Requieren un entrenamiento en base a un conocimiento previo del entorno del problema. Una red neuronal es un sistema compuesto por un gran número de elementos básicos, agrupados en capas que se encuentran totalmente interconectadas y que serán entrenadas para reaccionar de una determinada manera a los estímulos de entrada.

Las redes neuronales constituyen naturalmente una técnica de modelización multivariada, es decir, pueden hacer predicciones de dos o más variables simultáneamente. Pueden realizar predicciones tanto de variables continuas como discretas, utilizando las implementaciones apropiadas. En este trabajo son utilizadas para predecir el grupo o categoría de procedencia del texto en función de la distribución porcentual de las categorías morfológicas, información derivada del análisis automático de los mismos.

El Perceptrón Multicapa (MLP, por sus siglas en inglés "Multi-Layer Perceptron") tiene como objetivo la categorización o clasificación de forma supervisada. Para este trabajo se ha utilizado esta red aplicado a la clasificación de textos en dos géneros: Científicos y no Científicos. Utilizando el algoritmo de aprendizaje supervisado Backpropagation, la red aprende la relación entre la proporción de las distintas categorías morfosintácticas y la categoría de pertenencia (género), con el

propósito de lograr clasificar un nuevo texto para el cual se cuenta con el análisis morfológico pero se desconoce su género.

Para realizar la validación del modelo obtenido con los datos del conjunto de entrenamiento, es necesario considerar el error que se comete cuando la red es aplicada sobre un nuevo conjunto de datos, el conjunto de prueba. Esta nueva aplicación brindará como resultado de clasificación la matriz de confusión. La matriz de confusión muestra las predicciones correctas e incorrectas cuando se aplica el modelo sobre el conjunto de prueba y permite comprender en qué sentido se equivoca la red al intentar clasificar los nuevos textos.

2.4.6. Análisis Discriminante

El Análisis Discriminante es una técnica Multivariada exploratoria utilizada para describir si existen diferencias entre *k* grupos de unidades o poblaciones (individuos, objetos, etc.) respecto a un conjunto de *p* variables medidas sobre estas unidades. Mediante éste análisis se obtiene una regla de clasificación basada en una función discriminante que puede ser utilizada con el fin de asignar futuras unidades a una de las *k* poblaciones según sus valores observados.

Existen varios métodos para obtener la función discriminante. En este trabajo se compararán la Función Lineal Discriminante versus la Cuadrática. En el caso del Discriminador Lineal se supone la estructura de covariancias de las *p* variables es la misma para todas las poblaciones. En cambio, para el Discriminador Cuadrático, se supone normalidad multivariada pero estructuras de covariancias distintas para las distintas poblaciones.

Sea ω un individuo que puede provenir de *k* poblaciones $\pi_1, \pi_2, \dots, \pi_k$, con $k \geq 3$. Se quiere encontrar una regla de clasificación para asignar a ω a una de las *k* poblaciones basándose en el vector observado $\boldsymbol{x} = (x_1, x_2, \dots, x_p)'$ de *p* variables para este individuo en particular. Sea $\boldsymbol{\mu}_i$ el vector de medias y $\boldsymbol{\Sigma}_i$ la matriz de variancias y covariancias de las *p* variables en la i-ésima población.

Función Lineal Discriminante.

Se supone matriz de variancias y covariancias $\boldsymbol{\Sigma}$ común para las *k* poblaciones.

Sean $M^2(\boldsymbol{x}, \boldsymbol{\mu}_i) = (\boldsymbol{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)$, con *i=1,…,k* la distancia de Mahalanobis de ω a las poblaciones. Entonces se puede pensar en un criterio de clasificación que asigne a ω a la población más próxima de la siguiente forma:

Si $M^2(\boldsymbol{x}, \boldsymbol{\mu}_i) = mín\{M^2(\boldsymbol{x}, \boldsymbol{\mu}_1), M^2(\boldsymbol{x}, \boldsymbol{\mu}_2), \dots, M^2(\boldsymbol{x}, \boldsymbol{\mu}_k)\}$ se asigna ω aπ_i

Utilizando las funciones lineales discriminantes se obtiene la expresión

$L_{ij}(\boldsymbol{x}) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)$ y trabajando algebraicamente se puede probar que:

Si $L_{ij}(\boldsymbol{x}) > 0$ para todo $j \neq i$, entonces se asigna ω a π_i

Como las funciones $L_{ij}(\boldsymbol{x})$ verifican:

- $L_{ij}(\boldsymbol{x}) = \frac{1}{2}[M^2(\boldsymbol{x}, \boldsymbol{\mu}_j) - M^2(\boldsymbol{x}, \boldsymbol{\mu}_i)]$
- $L_{ij}(\boldsymbol{x}) = -L_{ji}(\boldsymbol{x})$
- $L_{rs}(\boldsymbol{x}) = L_{is}(\boldsymbol{x}) - L_{ir}(\boldsymbol{x})$

Entonces sólo se necesitan *k-1* funciones discriminantes.

Función Cuadrática Discriminante.

Se puede deducir en base a la regla de máxima verosimilitud de la siguiente manera:

Sea la función de densidad de $\boldsymbol{x}$ $f_i(\boldsymbol{x})$ en la i-ésima población π_i.

Si $f_i(\boldsymbol{x}) = máx\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \dots, f_k(\boldsymbol{x})\}$ entonces se asigna ω a π_i

Este criterio se relaciona con las funciones discriminantes $V_{ij}(\boldsymbol{x}) = \ln f_i(\boldsymbol{x}) - \ln f_j(\boldsymbol{x})$
Si se cumple el supuesto de normalidad multivariante y las matrices de covariancias $\boldsymbol{\Sigma_1}, \boldsymbol{\Sigma_2}, \ldots, \boldsymbol{\Sigma_k}$ difieren para las distintas poblaciones entonces se obtiene el siguiente discriminador cuadrático:

$$Q_{ij}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}'\left(\boldsymbol{\Sigma_j^{-1}} - \boldsymbol{\Sigma_i^{-1}}\right)\boldsymbol{x} + \boldsymbol{x}'\left(\boldsymbol{\Sigma_i^{-1}}\boldsymbol{\mu_i} - \boldsymbol{\Sigma_j^{-1}}\boldsymbol{\mu_j}\right) + \frac{1}{2}\boldsymbol{\mu_j'}\boldsymbol{\Sigma_j^{-1}}\boldsymbol{\mu_j} - \frac{1}{2}\boldsymbol{\mu_i'}\boldsymbol{\Sigma_i^{-1}}\boldsymbol{\mu_i} + \frac{1}{2}ln|\boldsymbol{\Sigma_j}| - \frac{1}{2}ln|\boldsymbol{\Sigma_i}|$$

2.4.7. Support Vector Machine

SVP utiliza un algoritmo que se basa en una clase especial de modelo lineal denominado ***hiperplano óptimo de máximo margen.*** Este hiperplano, que pertenece a un espacio de dimensionalidad que puede llegar a ser infinito, es hallado utilizando vectores soporte. Luego, mediante una transformación inversa se obtiene una frontera no necesariamente lineal que separa los grupos en el espacio original.
Los ***vectores soportes*** son las observaciones que están más cerca del hiperplano. Siempre hay como mínimo un vector soporte para cada clase.
La expresión del hiperplano de máximo margen viene dada por:

$$x = b + \sum a_i\, y_i\, (a(i).a)^n$$

dónde y_i es -1 o 1 depende del grupo al que pertenezca la observación; **a**(i) es el vector de valores de atributos correspondientes al i-ésimo vector soporte y **a** otro vector de atributos para una observación; b y α son parámetros calculados por el algoritmo; y *n* se elije según el grado del polinomio kernel con el que se desee trabajar. Algunos de los kernels existentes son Lineal (n=1), Polinomio de segundo grado (n=2), Radial Basis Function (RBF de parámetro γ). Al aplicar el método de SVM hay que tener en cuenta la constante de penalización C que impone una cota máxima al coeficiente α_i

2.4.8. Sequential Minimal Optimization (SMO)

Es un algoritmo que resuelve un problema, que surge en SVM, de optimización de una función cuadrática de varias variables, pero sujetas a una restricción lineal de esas variables.

3. RESULTADOS

En Beltrán 2013 se realizó un análisis exploratorio en el cual se evidencian las características que discriminan los corpus de textos en estudio. En dicho estudio se evidenció que existen diferencias significativas entre los corpus respecto al tamaño de los textos (número de palabras por texto). Esta situación llevó a realizar las sucesivas comparaciones sobre los porcentajes o proporciones de las categorías gramaticales, hallando diferencias significativas ($p<0.05$) para todas las categorías gramáticas excepto la proporción de clíticos y de verbos en los documentos analizados. Asimismo, en un análisis de componentes principales, se dispusieron los textos en el plano de proyección demostrando que los textos procedentes del corpus No Científico presentan un mayor número de adverbios, respecto a las restantes categorías, que los textos Científicos.

La tabla 2 presenta los porcentajes de mala clasificación de las técnicas evaluadas.

Método	**% EMC**
Redes Neuronales	3
Método del Vecino más Cercano	13
Arboles de clasificación	14
Análisis Discriminante Cuadrático	17
Análisis Discriminante Lineal	18
SMO	18
SVM-Kernel Lineal (C=0,1)	19
Regresión Logística	21
SVM-Kernel Polinómio de grado 2 (C=0,1)	21
Sistema de Clasificación Bagging	26
SVM-Kernel Radial Basis Function (C=1 γ=0,1)	31

Tabla 2: Porcentaje de mala clasificación según técnica estadística

Se distinguen varias cuestiones:

- Si bien el modelo de Regresión logística no fue el que mostró un beneficio en términos de precisión y cobertura, devolvió un modelo cuyos coeficientes estimados permitieron la caracterización y descripción de aquellas categorías morfológicas que discriminan los géneros.
- El modelo Perceptrón Multicapa (MLP), estimado para predecir el género de pertenencia de un texto fue la técnica con mejores resultados. Se observó que el porcentaje de clasificaciones incorrectas es muy bajo evidenciando un buen desempeño de la red para discriminar los textos por su género. La arquitectura y características de la red MLP, que brindan mejores resultados y hacen que la red tenga un comportamiento estable por lo que logra la habilidad de generalizar fueron los siguientes:

  Número de capas: 3

  Número de neuronas: 9 en la capa de entrada, 8 en la capa oculta y 2 en la capa de salida

  Atributos: proporciones de categorías morfológicas.

- Las técnicas de Bagging y el Vecino más cercano no evidenciaron ventaja alguna frente a las otras técnicas ya que el desempeño fue bajo. El método del vecino más cercano presentó un error de clasificación y precisión aceptables aunque sin superar a la red.

- También presentó desempeño aceptable el Análisis Discriminante Cuadrático (17 % de mala clasificación). Cabe destacar, que debido que los grupos presentan estructuras de covariancias distintas, es de esperar que el Análisis Discriminante Cuadrático clasifique mejor que el Análisis Discriminante Lineal (18% de mala clasificación). Por otro lado, no es posible conocer en de qué manera afecta la presencia de estructuras de covariancias distintas entre los grupos para los métodos restantes.
- El corpus de textos No Científicos siempre presentó una tasa de error mayor que el grupo de textos Científicos.
- De todas las opciones evaluadas, SVM con kernel lineal y parámetro C=0.1 parece ser la más apropiada para clasificar textos de este tipo.
- El método SMO presentó porcentajes de mala clasificación bajos, del aproximadamente el orden del 18%. No obstante, se observó cierta variabilidad en cuanto a estos porcentajes para diferentes valores del parámetro C, con valores que van del 18% al 40%. Esto estaría indicando cierta inestabilidad del método para clasificar, representando una desventaja.

4. Conclusiones

La técnica de Redes Neuronales es la que presenta una marcada ventaja respecto a los otros métodos en cuanto al porcentaje de mala clasificación. Sin embargo, el método del Vecino más Cercano presenta un buen desempeño adicionando como principales ventajas la simpleza de su aplicación y la estabilidad de su comportamiento.
Una ventaja observada en el método de Árbol de Clasificación es la adaptación para recoger el comportamiento no aditivo de las variables predictoras, de manera que las interacciones se incluyen en forma automática. Sin embargo, en esta técnica se pierde información al tratar a las variables predictoras continuas como variables dicotómicas.
De los métodos de aprendizaje de máquina, el SMO presenta variabilidad en cuanto al %MC para diferentes valores del parámetro C (Figura 1) indicando cierta inestabilidad en el método para clasificar. Mientras que la técnica SVM con kernel lineal es el más estable en su desempeño al variar el valor del parámetro C.
El mejor desempeño del ADC en comparación con el ADL resulta lógico dado que los datos presentan estructuras de covariancias distintas para los grupos.
No es posible saber cómo afectan el cumplimiento (o no) de determinados supuestos en los datos para SMO y SVM, siendo ésto una ventaja del Análisis Discriminante frente a los métodos de aprendizaje de máquina

Referencias

Barbona, I. 2015 Comparación de métodos de clasificación aplicados a textos Científicos y No Científicos. Revista INFOSUR. Grupo INFOSUR. Rosario.

Beltrán, C., Bender, C., Bonino, R., Deco, C., Koza, W., Méndez, B., Moro, Stella Maris. 2008 Recursos informáticos para el tratamiento lingüístico de textos. Ediciones Juglaría. Rosario.

Beltrán, C. 2009 Modelización lingüística y análisis estadístico en el análisis automático de textos. Ediciones Juglaría. Rosario.

Beltrán, C. 2010 Estudio y comparación de distintos tipos de textos académicos: Biometría y Filosofía. Revista de Epistemología y Ciencias Humanas. Grupo IANUS. Rosario.

Beltrán, C. 2010 Análisis discriminante aplicado a textos académicos: Biometría y Filosofía. Revista INFOSUR. Grupo INFOSUR. Rosario.

Beltrán, C. 2011. Aplicación del análisis de regresión logística multinomial en la clasificación de textos académicos: Biometría, Filosofía y Lingüística informática. Revista INFOSUR. Grupo INFOSUR. Rosario.

Bès,Gabriel, Solana, Z y Beltrán, C. 2005 Conocimiento de la lengua y técnicas estadísticas en el análisis lingüístico en Desarrollo, implementación y uso de modelos para el procesamiento automático de textos (ed. Víctor Castel) Facultad de Filosofía y Letras, UNCUYO

Catena, A.; Ramos, M.M; Trujillo, H.M. 2003. ANALISIS MULTIVARIADO. UN MANUAL PARA INVESTIGADORES. Bibiloteca Nueva S.L. España.

Cuadras, C.M. 2008 NUEVOS MÉTODOS DE ANÁLISIS MULTIVARIANTE. CMC Editions. Barcelona, España.

Flórez López, R.; Fernández Fernández, J.M. 2008. LAS REDES NEURONALES ARTIFICIALES. FUNDAMENTOS TEORICOS Y APLICACIONES PRACTICAS. Netbiblio S.L. España.

Johnson R.A. y Wichern D.W. 1992 Applied Multivariate Statistical Análisis. Prentice-Hall International Inc.

Khattre R. y Naik D. (2000) Multivariate Data Reduction and Discriminatio with SAS Software. SAS Institute Inc. Cary, NC. USA

Stokes, M. E., Davis, C.S., Koch, G.G. 1999 Categorical Data Analysis using SAS® System. WA (Wiley-SAS).

Witten, I., Frank, E. 2005. Data Mining. Practical Machine Learning Tools and Techniques. Elsevier.