Comparación de métodos de clasificación aplicados a textos Científicos y No Científicos

Comparison among classification methods applied to Scientific and Non Scientific texts.

Ivana Barbona
Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, Argentina.
ivanabarbona@gmail.com

Abstract

The aim of this work is to compare the performance of the classification methods Support Vector Machine, Sequential Minimal Optimization, Logistic Regression, Lineal and Quadratic Discriminant Analysis by applying them to Scientific and Non Scientific Texts. This comparison among classification methods is carried out taking into account the measure of the misclassification error percentage calculated by (total of misclassified texts/total of texts)x100 . The software used for this analysis are Weka and JMP programs.

Keywords: Support Vector Machine - Sequential Minimal Optimization - Logistic Regression – Lineal Discriminant Analysis - Quadratic Discriminant Analysis – Learning Machine.

Resumen

En este trabajo se comparan el desempeño de los métodos de clasificación Support Vector Machine, Sequential Minimal Optimization, Regresión Logística, Análisis Discriminante Lineal y Cuadrático mediante su aplicación a textos Científicos y No Científicos. La comparación entre los métodos se realiza teniendo en cuenta la medida del porcentaje de error de mala clasificación calculado como (total de textos mal clasificados/total de textos)x100. El software utilizado para realizar este análisis son los programas Weka y JMP.

Palabras claves: Support Vector Machine, Sequential Minimal Optimization, Regresión Logística, Análisis Discriminante Lineal, Análisis Discriminante Cuadrático, Métodos de Clasificación, Aprendizaje de Máquina.

1. Introducción

La clasificación automática de textos ha tomado gran importancia en los últimos años debido al aumento de información disponible en formato electrónico. Su objetivo es categorizar documentos dentro de una cantidad fija de categorías predefinidas en función de su contenido.

En el presente trabajo se compara el desempeño de varios métodos de clasificación. Para esto se cuenta con una base de datos de textos clasificados en Científicos y No Científicos a los cuales se les midieron una determinada cantidad de características. La comparación entre los métodos se realiza teniendo en cuenta la medida del porcentaje de error de mala clasificación calculado como (total de textos mal clasificados/total de textos)x100.

Los métodos que se van a utilizar son: Support Vector Machine (SVM), Sequential Minimal Optimization (SMO), Regresión Logística, Análisis Discriminante Lineal (ADL) y Cuadrático (ADC).

2. Material y Métodos.

Los datos a utilizar provienen de un proyecto de investigación que se lleva a cabo en la Facultad de Ciencias Agrarias de la UNR denominado “Modelización Estadística en la Clasificación de Textos: Científicos y No Científicos”. En este proyecto se pretende evaluar el desempeño de las técnicas multivariadas para clasificar/agrupar textos según el género utilizando la frecuencia de aparición de distintas clases de palabras, entre otras características.

La base de datos corresponde a información de 150 textos. Está compuesta por una variable de clasificación GENERO que se refiere al tipo de texto, cuyas categorías son CIENTÍFICO y NO CIENTÍFICO, así como también 12 variables que representan características propias de estos textos.

GENERO	Género 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

Tabla 1: Variables en la base de datos.

Dado que el tamaño en cuanto a la cantidad de palabras es distinto para cada texto, se decide considerar la proporción de cada clase de palabras en lugar de la frecuencia. Además, mediante una selección de variables se descarta del análisis la variable "otro", lo cual permite eliminar la restricción de que la suma de los valores de las variables para cada texto es igual a uno. Restricción que proviene de tener en cuenta la proporción en lugar de la frecuencia, y que podrían causar problemas en la aplicación de algunos métodos.

Por lo tanto, la base definitiva a que se utiliza está compuesta por la variable de clasificación GENERO Y 9 variables que caracterizan a los textos. Éstas son la proporción de adjetivos, adverbios, clícticos, copulativos, determinantes, sustantivos, preposiciones y verbos.

Los Métodos de clasificación utilizados, que van a ser comparados son los siguientes:

2.1. Support Vector Machine:

Es un método de clasificación supervisada lineal que determina la frontera óptima entre dos grupos que pueden ser linealmente separables o no. Encuentra un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad que puede llegar a ser infinita, mediante la utilización de vectores soportes. Los vectores soportes son los datos que caen más próximos al hiperplano. Luego, mediante una trasformación inversa se obtiene la frontera no lineal que separa a esos dos grupos en el espacio original.

En el caso de clasificar en 2 categorías, se busca el hiperplano que tenga la máxima distancia o margen con los puntos más cercanos a él. Luego, los puntos pertenecientes a una categoría estarán a un lado del hiperplano y los casos que se encuentren en la otra categoría estarán al otro lado.

Mediante la Función Kernel se puede proyectar la información a un espacio de características de mayor dimensionalidad, las funciones kernel que se utilizan en este trabajo son:

- Lineal
- Polinomio de segundo grado
- Radial Basis Function.

Se realizó una grilla en la cual se aplicó método de SVM variando la constante de penalización C y otros parámetros dentro de los distintos kernel considerados.

2.2. 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.

2.3. Regresión Logística:

Es un modelo estadístico que sirve para describir la relación entre una variable respuesta categórica y un conjunto de variables explicativas, mediante el uso de la función de enlace logit.

Sea un conjunto de p variables independientes denotadas por el vector $x' = (x_1, x_2, \dots, x_p)$. Y sea la probabilidad condicional de que la variable respuesta Y sea igual a 1 $P(Y = 1/x) = \pi(x)$. El modelo de regresión logística está dado por la ecuación:

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

donde,

$$\pi(x) = \frac{e^{g(x)}}{1 - e^{g(x)}}$$

2.4. 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 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 ≥ 3. Se quiere encontrar una regla de clasificación para asignar a ω a una de las k poblaciones basándose en el vector observado $x = (x_1, x_2, \dots, x_p)'$ de p variables para este individuo en particular. Sea μ_i el vector de medias y Σ_i la matriz de variancias y covariancias de las p variables en la i-ésima población.

2.4.1. Función Lineal Discriminante.

Se supone matriz de variancias y covariancias Σ común para las k poblaciones.

Sean $M^2(x,\mu_i) = (x-\mu_i)'\Sigma^{-1}(x-\mu_i)$, 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 $M^2(x,\mu_i) = \min\{M^2(x,\mu_1), M^2(x,\mu_2), \dots, M^2(x,\mu_k)\}$ se asigna ω a π_i

Utilizando las funciones lineales discriminantes se obtiene la expresión

$$L_{ij}(x) = (\mu_i - \mu_j)'\Sigma^{-1}x - \frac{1}{2}(\mu_i - \mu_j)'\Sigma^{-1}(\mu_i + \mu_j)$$

y trabajando algebraicamente se puede probar que:

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

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

$$L_{ij}(x) = \frac{1}{2}\left[M^2(x,\mu_j) - M^2(x,\mu_i)\right]$$

$$L_{ij}(x) = -L_{ji}(x)$$

$$L_{rs}(x) = L_{is}(x) - L_{ir}(x)$$

Entonces sólo se necesitan k-1 funciones discriminantes.

2.4.2. Función Cuadrática Discriminante.

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

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

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

Este criterio se relaciona con las funciones discriminantes

$$V_{ij}(x) = \ln f_i(x) - \ln f_j(x)$$

Si se cumple el supuesto de normalidad multivariante y las matrices de covariancias $\Sigma_1, \Sigma_2, \ldots, \Sigma_k$ difieren para las distintas poblaciones entonces se obtiene el siguiente discriminador cuadrático:

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

Todos los métodos se implementaron con el programa WEKA, excepto los Análisis DL y DC que fueron aplicados mediante el programa JMP.

Para compararlos se considera la medida del error de mala clasificación calculado como total de textos mal clasificados/total de textos.

3. Resultados

En la tabla 2 se observa una grilla con los resultados del porcentaje de error de mala clasificación para los métodos SVM y SMO variando el kernel y valores de parámetros.

Con respecto al método Support Vector Machine, se consideran distintos valores del parámetro C para los kernel lineal, polinomio de grado 2 y radial basis function. Lo mismo para el parámetro γ del kernel RBF.

		SVM									SMO
		Kernel: Lineal	Kernel: Polinomio grado 2	Kernel: Radial Basis Function.							
				γ = 0.0	γ= 0.1	γ= 0.2	γ= 0.3	γ= 0.5	γ= 1	γ= 5	γ= 10	
C	**0.1**	***19.33***	21.33	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00
	0.2	***19.33***	22.67	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	34.00
	0.3	20.67	24.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	30.00
	0.4	20.67	24.67	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	28.00
	0.5	20.67	24.67	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	25.33
	0.6	20.00	25.33	39.33	39.33	40.00	40.00	40.00	40.00	40.00	40.00	25.33
	0.7	20.00	26.00	39.33	38.67	40.00	40.00	40.00	40.00	40.00	40.00	23.33
	0.8	20.00	26.00	38.00	38.00	40.00	40.00	40.00	40.00	40.00	40.00	21.33
	0.9	20.00	27.33	38.67	36.67	39.33	40.00	40.00	40.00	40.00	40.00	20.67
	1	20.00	26.00	36.00	31.33	40.00	40.67	40.00	40.00	40.00	40.00	18.67
	1.5	20.00	26.67	32.67	31.33	40.67	41.33	40.67	40.00	40.00	40.00	18.67
	2	20.00	26.67	33.33	32.00	40.67	41.33	40.67	40.00	40.00	40.00	19.33
	2.5	20.00	30.00	33.33	32.67	40.67	41.33	40.67	40.00	40.00	40.00	18.67
	3	20.00	29.33	33.33	32.67	40.67	41.33	40.67	40.00	40.00	40.00	18.67
	3.5	20.00	29.33	34.00	33.33	40.67	41.33	40.67	40.00	40.00	40.00	18.67
	5	20.00	28.00	34.67	34.67	40.67	41.33	40.67	40.00	40.00	40.00	***18.00***
	10	20.67	26.67	34.00	34.00	40.67	41.33	40.67	40.00	40.00	40.00	***18.00***
	20	20.67	27.33	34.00	34.00	40.67	41.33	40.67	40.00	40.00	40.00	18.67

Tabla 2: Valores del Error de Mala Clasificación al aplicar SVM y SMO para distintos valores de parámetros.

Método		**Porcentaje de Mala Clasificación**
Regresión Logística		20.67
Análisis Discriminante	**Lineal**	***18***
	Cuadrático	***16.67***

Tabla 3: Resultados de error de mala clasificación al aplicar Análisis Discriminante y Regresión Logística.

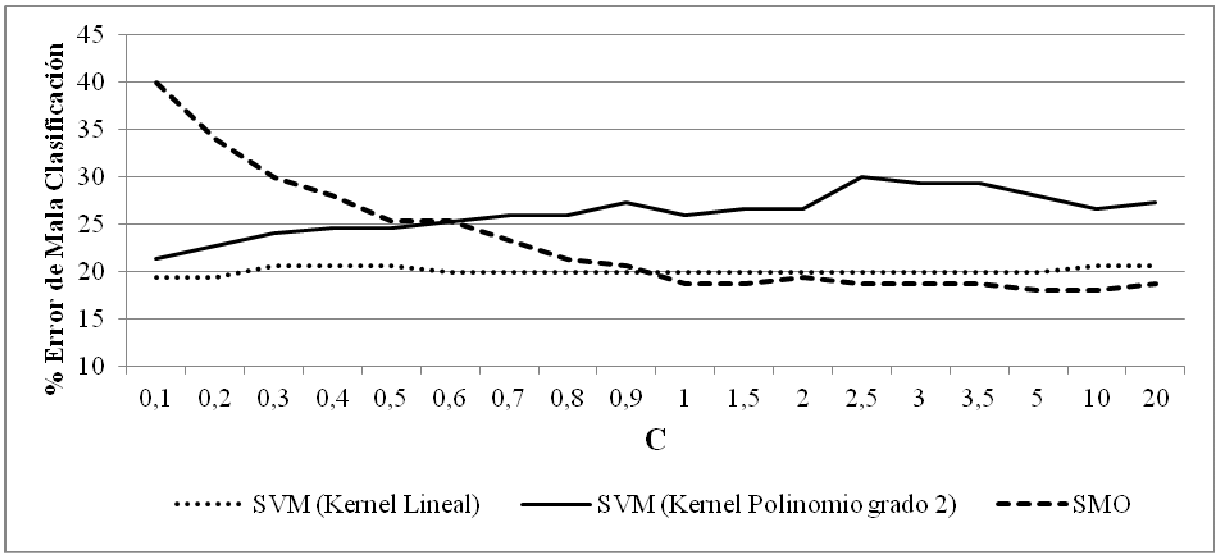

Figura 1: Comparación entre SVM (Kernel Lineal), SVM (Kernel Polinomio grado 2) y SMO.

Se observa que los métodos con porcentajes de mala clasificación menores son el Análisis Discriminante Lineal (18%) y Cuadrático (16.67%), SMO con C=5 y 10 (18% para ambos casos) y SVM con kernel lineal y C= 0.1 y 0.2 (19.33%).

El método SMO presentó porcentajes de mala clasificación bajos, del aproximadamente el orden del 18%. No obstante, se visualiza cierta variabilidad en cuanto a estos porcentajes para diferentes valores del parámetro C (Figura 1), con valores que van del 18% al 40%. Esto estaría indicando cierta inestabilidad del método para clasificar, representando una desventaja. En cambio el método SVM con kernel lineal, obtuvo valores un poco más altos de porcentajes de mala clasificación que SMO, pero al variar el parámetro C, sus resultados fueron más estables, con valores que fueron del 19.33% al 20.67%.

En cuanto al tiempo de ejecución, SVM y SMO no presentaron diferencias. Pero al comprar los métodos observando las demás medidas de error que proporciona WEKA en sus resultados, SVM con Kernel lineal y C= 0.1 o 0.2 resultan mejores.

La tabla 3 muestra los resultados de los errores de mala clasificación para los métodos de Regresión Logística y Análisis Discriminante. Los resultados observados son de esperarse en cuanto al cumplimiento de los supuestos a los cuales están sujetos los ADL y ADC. Dado que los datos con los que se trabajó presentan estructuras de covariancias distintas para los grupos, resulta lógico que el ADC (16.67% de mala clasificación) funcione mejor que el ADL (18% de mala clasificación).

Por otro lado, cabe destacar que no es posible saber cómo afectan las estructuras de covariancias distintas entre los grupos para los métodos SMO y SVM, lo cual representa una ventaja del AD frente métodos de aprendizaje de máquina como el SVM.

4. Conclusiones

De todos los métodos de clasificación considerados, el que presentó el menor porcentaje de mala clasificación fue el ADC (16.67%).

Tanto el ADL como el ADC dieron buenos resultados al clasificar los textos en Científicos y No Científicos, presentando un 18% y 16.67% de mala clasificación respectivamente.

En cuanto a los métodos de aprendizaje de máquina, el que presenta mejores resultados es el SVM con kernel lineal y constante de penalización C= 0.1 o 0.2 (19.33%).

Del resto de los métodos aplicados, el que presenta peores resultados es SVM con kernel RBF, arrojando valores de porcentaje de error de mala clasificación que van del 34% al 40%.

Si bien el método SMO presentó porcentajes bajos de mala clasificación para valores altos de C (18%), no es considerado uno de los mejores debido a la variabilidad que presenta en sus resultados al considerar distintos valores de la constante C, dando indicios de cierta inestabilidad del método para clasificar bien.

Referencias

Beltrán, C. 2010. Análisis discriminante aplicado a textos académicos: Biometría y Filosofía. Revista Infosur. N° 4.

Cherkassky, V., Mulier, F. 2007. Learning From Data. Concepts, Theory, and Methods. John Wiley & Sons.

Cuadras, C. 2012. Nuevos Métodos de Análisis Multivariante. CMC Editions.

Hastie, T., Tibshirani, R., Friedman, J. 2009. The Elements of Statistical Learning. Data Mining, Inference, and Predicion. Springer Series in Statistics.

Hosmer, D., Lemeshow, S., Sturdivant, R. 2013. Applied Logistic Regression. John Wiley & Sons.

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