MÉTODO DE CLASIFICACIÓN SUPERVISADA SUPPORT VECTOR MACHINE: UNA APLICACIÓN A LA CLASIFICACIÓN AUTOMÁTICA DE TEXTOS.

Supervised Classification Method Support Vector Machine applied to Automatic Text Classification

Barbona Ivana, Beltrán Celina

Cátedra de Estadística, Facultad de Ciencias Agrarias, Universidad Nacional de Rosario

ivanabarbona@gmail.com

Abstract

Support Vector Machine (SVM) is a supervised classification method useful for determining the optimal boundary between two groups that can be linearly separable or not. This method found an hyperplane or set of hyperplane in a space of dimensionality that can become infinite. Then, using an inverse transformation the border between these two groups in the original space is obtained.

In 2 categories clasification, the hyperplane that has the maximum distance or margin with his closest points is sought. The elements belonging to a category are on one side of the hyperplane while cases belonging to the other category are on the other side.

This paper presents an application of SVM method to classify a set of texts. The classification criterion used was the genre to which the text (Scientific / No Scientific) belongs. The characterization of the text is based on the frequency distribution of the morpho-syntactic categories. The final results are percentages of misclassification in a grid with SVM method varying the penalty constant C and other parameters within several kernel considered. The best performance was obtained for SVM with linear kernel and C = 0.1 and 0.2 (19.33%)

Keywords: Support Vector Machine, Learning Machine, Supervised Classification Methods, Text Classification.

Resumen

Support Vector Machine (SVM) es un método de clasificación supervisada que permite determinar la frontera óptima entre dos grupos que pueden ser linealmente separables o no. Mediante la utilización de vectores soporte se encuentra un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad que puede llegar a ser infinita. Luego, mediante una trasformación inversa se obtiene la frontera 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. Los puntos pertenecientes a una categoría estarán a un lado del hiperplano mientras que los casos que pertenezcan a la otra categoría estarán al otro lado.

En este trabajo se realiza una aplicación del método SVM para clasificar un conjunto de textos. El criterio de clasificación utilizado fue 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 resultados finales representan porcentajes de mala clasificación en una grilla para el método SVM variando la constante de penalización C y otros parámetros dentro de varios kernel considerados. El mejor desempeño se obtuvo para SVM con kernel lineal y C= 0.1 y 0.2 (19.33%)

Palabras claves: Support Vector Machine, Learning Machine, Métodos de Clasificación Supervisada, Clasificación de Textos.

1. INTRODUCCION

Las Máquinas de Vectores de Sporte (Support Vector Machines o SVM) son un conjunto de algoritmos de aprendizaje supervisado que se utilizan para resolver problemas de clasificación y regresión.

Determinan la frontera óptima entre dos grupos que pueden ser linealmente separables o no. Para esto, se basan en un algoritmo que encuentra un hiperplano en un espacio de dimensión que puede llegar a ser infinita

El objetivo del trabajo es utilizar SVM y evaluar su desempeño mediante la aplicación a un problema de clasificación automática de textos Científicos y No Científicos.

2. MATEROAL Y MÉTODOS

Se cuenta con una base de datos que consiste en 150 textos clasificados en “CIENTÍFICOS “y “NO CIENTÍFICOS” y 12 variables que representan características propias de cada texto de las cuales se decide utilizar las siguientes:

- ✓ Género al que pertenece el texto
- ✓ Proporción de adjetivos
- ✓ Proporción de adverbios
- ✓ Proporción de clícticos
- ✓ Proporción de copulativos
- ✓ Proporción de determinantes
- ✓ Proporción de sustantivos
- ✓ Proporción de preposiciones
- ✓ Proporción de verbos

Support Vector Machines:

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 cómo 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.

En este trabajo se prueban los siguientes kernels:

- Lineal (n=1)
- Polinomio de segundo grado (n=2)
- Radial Basis Function (RBF)

Se aplica SVP variando la constate de penalización C (que impone una cota máxima al coeficiente α_i).

También se trabaja con distintos valores del parámetro γ del kernel RBF.

3. RESULTADOS

Se evalúa el desempeño de cada algoritmo mediante el % de mala clasificación (%MC) calculado como:

$$\%\text{MC} = \frac{Total\,de\,textos\,mal\,clasificados}{\text{Total de textos}} x100$$

Se observa tanto en la Tabla 1 como en la Figura 1 lo siguiente:

- El menor %MC se da en el caso de kernel lineal con C=0.1 y 0.2 (19.33% en ambos).
- El peor %MC se observa para kernel RBF con parámetro γ=0.3 y C ≥ 1.5.

- El desempeño del kernel RBF fue peor en comparación con el Lineal y el de grado 2 en todos los casos.
- El kernel lineal presenta mejor desempeño que el resto de los kernels evaluados para todos los valores de C aplicados.

Se observa tanto en la Tabla 1 como en la Figura 1 lo siguiente:

- El menor %MC se da en el caso de kernel lineal con C=0.1 y 0.2 (19.33% en ambos).
- El peor %MC se observa para kernel RBF con parámetro γ=0.3 y C $\geq$ 1.5.
- El desempeño del kernel RBF fue peor en comparación con el Lineal y el de grado 2 en todos los casos.
- El kernel lineal presenta mejor desempeño que el resto de los kernels evaluados para todos los valores de C aplicados.

Tabla 1: %MC para las variantes de SVM aplicadas

		SVM							
		Kernel: Lineal	**Kernel: Pol. grado 2**	**Kernel: RBF**					
				γ = 0	**γ = 0.1**	**γ = 0.2**	**γ = 0.3**	**γ = 0.5**	**γ = 1**
C	**0.1**	19.33	*21.33*	*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*
	0.3	20.67	*24.00*	*40.00*	*40.00*	*40.00*	*40.00*	*40.00*	*40.00*
	0.4	*20.67*	*24.67*	*40.00*	*40.00*	*40.00*	*40.00*	*40.00*	*40.00*
	0.5	*20.67*	*24.67*	*40.00*	*40.00*	*40.00*	*40.00*	*40.00*	*40.00*
	0.6	*20.00*	*25.33*	*39.33*	*39.33*	*40.00*	*40.00*	*40.00*	*40.00*
	0.7	*20.00*	*26.00*	*39.33*	*38.67*	*40.00*	*40.00*	*40.00*	*40.00*
	0.8	*20.00*	*26.00*	*38.00*	*38.00*	*40.00*	*40.00*	*40.00*	*40.00*
	0.9	*20.00*	*27.33*	*38.67*	*36.67*	*39.33*	*40.00*	*40.00*	*40.00*
	1	*20.00*	*26.00*	*36.00*	*31.33*	*40.00*	*40.67*	*40.00*	*40.00*
	1.5	*20.00*	*26.67*	*32.67*	*31.33*	*40.67*	*41.33*	*40.67*	*40.00*
	2	*20.00*	*26.67*	*33.33*	*32.00*	*40.67*	*41.33*	*40.67*	*40.00*
	2.5	*20.00*	*30.00*	*33.33*	*32.67*	*40.67*	*41.33*	*40.67*	*40.00*
	3	*20.00*	*29.33*	*33.33*	*32.67*	*40.67*	*41.33*	*40.67*	*40.00*
	3.5	*20.00*	*29.33*	*34.00*	*33.33*	*40.67*	*41.33*	*40.67*	*40.00*
	5	*20.00*	*28.00*	*34.67*	*34.67*	*40.67*	*41.33*	*40.67*	*40.00*
	10	*20.67*	*26.67*	*34.00*	*34.00*	*40.67*	*41.33*	*40.67*	*40.00*
	20	*20.67*	*27.33*	*34.00*	*34.00*	*40.67*	*41.33*	*40.67*	*40.00*

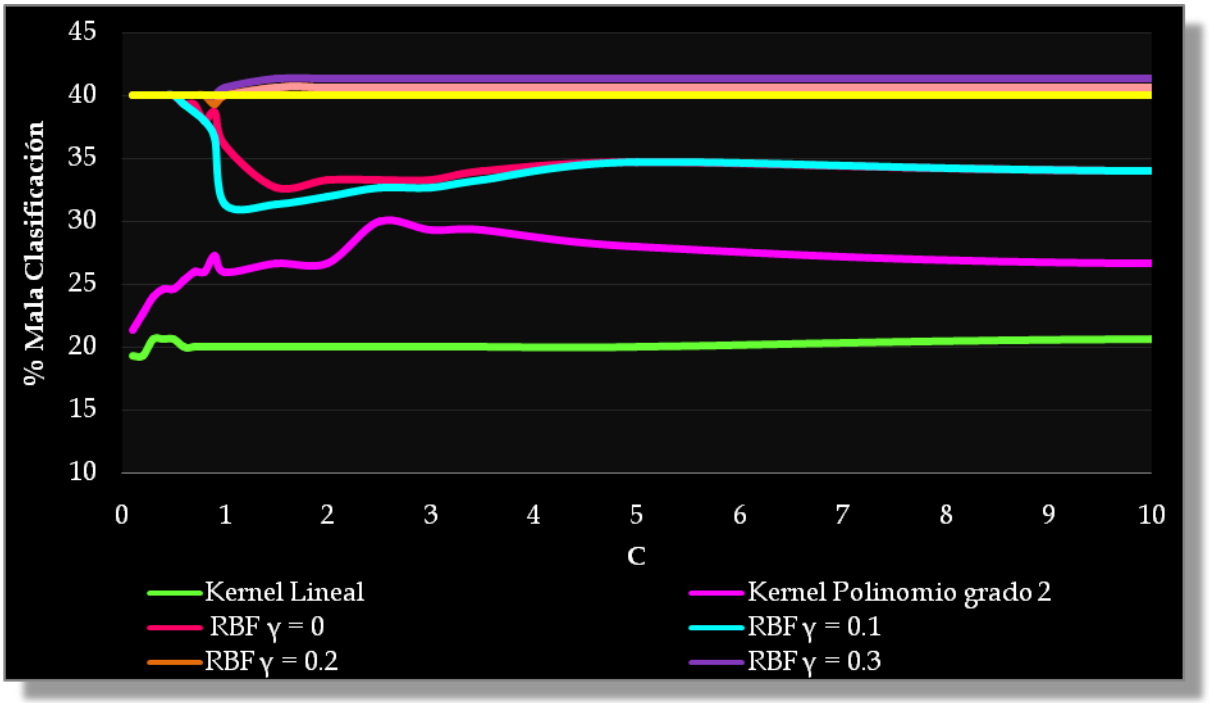

Figura 1: Desempeño según el %MC para cada variante de SVM aplicada

- El kernel lineal, además de tener un buen desempeño, presenta resultados estables en cuanto al %MC a través de los distintos valores de la constante de penalización C.

4. CONCLUSIONES

- ✓ El kernel lineal presenta el mejor desempeño que el resto de los kernels para cualquier valor de C, lo que indicaría que la frontera de clasificación de textos CIENTÍFICOS y NO CIENTÍFICOS el espacio original de las variables es lineal.
- ✓ Considerando el caso de kernel RBF con $\gamma=0$ y $\gamma=0.1$ se observa variabilidad en el %MC para valores de $C \leq 5$ (Figura 1). Esto puede deberse a que el parámetro C está relacionado con el costo de clasificar mal y al aumentar C, aumenta el número de vectores soporte (penalizando más por clasificar mal). Por lo tanto esto puede llevar a un sobreajuste. Lo ideal es elegir un valor de C que permita un equilibrio entre la mala clasificación y la simpleza del modelo.
- ✓ 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.

Referencias

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

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

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

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