Técnicas estadísticas de clasificación. Una aplicación en la clasificación de textos según el género: Textos Científicos y Textos No Científicos

Statistical Techniques in Classification. An Application to Text Classification According to Gender: Scientific – Non scientific

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

Abstract

The problem of unit classification in groups or known population samples offers great interest to Statistics; as a consequence of this several techniques have been developed to fulfill this purpose. This work is aimed to classify scientific and non scientific texts comparing the analysis of classification (CT) and the logistic regression (LR) trees. The scientific texts are abstracts of papers published on journals and conference proceedings coming from different disciplines. The non scientific texts are news report of general interest published on the web page of Argentine newspapers. The information obtained from the morphological analysis of these texts is employed as explanatory variable of the multivariable technique applied in this work. The performance of the techniques was measured by means of the false classification rate (FCR), the precision rate (PR) and the coverage rate (CO) estimated by a sample text not included in the prediction model neither in the tree construction. The classification tree showed a FCR lower than the logistic model while the scientific text samples showed a major precision.

For the CT, the FCR, the PR and the CO resulted in 4%, 84% and 96% for the scientific texts and 28%, 92% and 72% for the non scientific texts, respectively.

For the LR model, the FCR, the PR and the CO resulted in 14%, 83% and 86% for the scientific texts and 26%, 77% and 74% for the non scientific texts, respectively.

Key words: Multivariable logistic regression – Classification Trees – Automatic text analysis– Text classification.

Resumen

El problema de la clasificación de unidades en grupos o poblaciones conocidas es de gran interés en estadística, por esta razón se han desarrollado varias técnicas para cumplir este propósito. Este trabajo se propone la clasificación de textos científicos y no científicos comparando las técnicas de Árboles de Clasificación (AC) y Regresión logística (RL). Los textos científicos corresponden a resúmenes de publicaciones en revistas científicas y actas de congresos de distintas disciplinas y los textos no científicos corresponden a noticias periodísticas de interés general publicadas en páginas web de periódicos argentinos. La información resultante del análisis morfológico de dichos textos es utilizada como variables explicativas en las técnicas multivariadas aplicadas en este trabajo. El

desempeño de las técnicas fue medido con la tasa de mala clasificación (TMC), la precisión (PR) y la cobertura (CO), calculadas sobre una muestra de textos no incluidos en la estimación del modelo y construcción del árbol. El árbol de clasificación presentó una TMC inferior a la del modelo logístico logrando clasificar con mayor precisión los textos científicos.

Para el AC la TMC, PR y CO resultaron 4%, 84% y 96% para los textos científicos y 28%, 92% y 72% para los textos no científicos, respectivamente.

Para el modelo de RL la TMC, PR y CO resultaron 14%, 83% y 86% para los textos científicos y 26%, 77% y 74% para los textos no científicos, respectivamente.

Palabras claves: Regresión logística multivariada, árboles de clasificación, análisis automático de textos, clasificación de textos.

1. INTRODUCCION

Este trabajo se propone la realización del análisis automático de distintos tipos de textos: científicos y no científicos. Los textos científicos corresponden a resúmenes de publicaciones en revistas científicas y actas de congresos de distintas disciplinas y los textos no científicos corresponden a noticias periodísticas de interés general publicadas en páginas web de periódicos argentinos. Se recurre al analizador morfológico Smorph, implementado como etiquetador, para asignar categoría a todas las ocurrencias lingüísticas.

La información resultante del análisis morfológico de dichos textos es utilizada para conformar una base de datos sobre la cual se aplican las técnicas multivariadas de clasificación: Regresión logística y Árboles de clasificación.

El desempeño de las técnicas es evaluado con tres medidas: la tasa de mala clasificación (TMC), la precisión (PR) y la cobertura (CO), calculadas sobre una muestra de textos, muestra de prueba, no incluidos en la estimación del modelo y construcción del árbol.

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 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.

Smorph compila, minimiza y compacta la información lingüística que queda disponible en un archivo binario. Los códigos fuente se dividen en cinco archivos: Códigos ASCII, Rasgos, Terminaciones, Modelos y Entradas.

En el archivo **entradas**, se declaran los ítems léxicos acompañados por el modelo correspondiente. Este indicador de modelo oficia de enlace con el archivo modelos, en el que se especifica la información morfológica y las terminaciones que se requieren en cada ítem.

En el archivo **modelos**, se introduce la información correspondiente a los modelos de flexiones morfológicas. A un conjunto de terminaciones se le asocia el correspondiente conjunto de definiciones morfológicas. En el archivo **terminaciones** es necesario declarar todas las terminaciones que son necesarias para definir los modelos de flexión, se declaran una a continuación de otra, separadas por un punto. Para construir los modelos se recurre a rasgos morfológico- sintácticos. En el archivo **rasgos**, se organizan jerárquicamente las etiquetas. En el archivo de códigos **ASCII** se especifican, entre otros, los caracteres separadores, las equivalencias entre mayúsculas y minúsculas. El archivo "data", contiene los nombres de cada uno de los cinco archivos descriptos anteriormente.

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).

2.3. Diseño y desarrollo de la base de datos

La información que contiene la base de datos es el resultado del análisis de Smorph-Mps almacenada en un archivo de texto. 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. De esta manera se obtuvo una base de datos que posee la información del texto, ocurrencia, lema y etiqueta asignada. Luego, a partir de esta base de datos por palabra (cada unidad o fila es una palabra analizada del texto), 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.

Referido al primer enfoque, una de las técnicas más utilizadas es la Regresión Logística. La regresión logística estima la probabilidad de un suceso en función de un conjunto de variables explicativas. Este modelo expresa matemáticamente la probabilidad de pertenencia a uno de los grupos, de manera que es posible calcularlas y asignar cada unidad al grupo cuya probabilidad de pertenencia estimada sea mayor. Otra técnica muy utilizada son los Árboles de Clasificación que crean una serie de reglas basadas en las variables predictoras que permiten asignar una nueva observación a una de las categorías o grupo.

2.4.1. Árboles de Clasificación

Los árboles de clasificación (AC) se emplean para asignar unidades experimentales a las clases de una variable dependiente a partir de sus mediciones en uno o más predictores. En esta aplicación, los AC se emplean para asignar textos al género al que corresponde: CIENTIFICO – NO CIENTIFICO a partir de la información relevada en el análisis morfológico automático de los textos.

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 (en este caso el género a la que pertenece el texto). El árbol determinado puede ser utilizado para clasificar nuevos textos.

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".

2.4.2. Regresión logística

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

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

β_0 es la constante del modelo o término independiente

p el número de covariables

β_i los coeficientes de las covariables

x_i 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:

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

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

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

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.

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.

3.1. Árboles de Clasificación

Se aplicó la técnica de Árboles de Clasificación para obtener reglas de clasificación que permitan asignar los textos en dos poblaciones, definidas por el género al que pertenecen: CIENTÍFICO y NO CIENTÍFICO. De la misma manera que en el apartado previo, los predictores utilizados corresponden a la distribución de las distintas categorías morfológicas halladas en el análisis automático de los textos (proporción de cada categoría morfológica).

Las variables que mostraron una buena discriminación de los grupos son la proporción de adverbios, nombres, adjetivos, preposición y verbos. El árbol final presenta 10 nodos terminales.

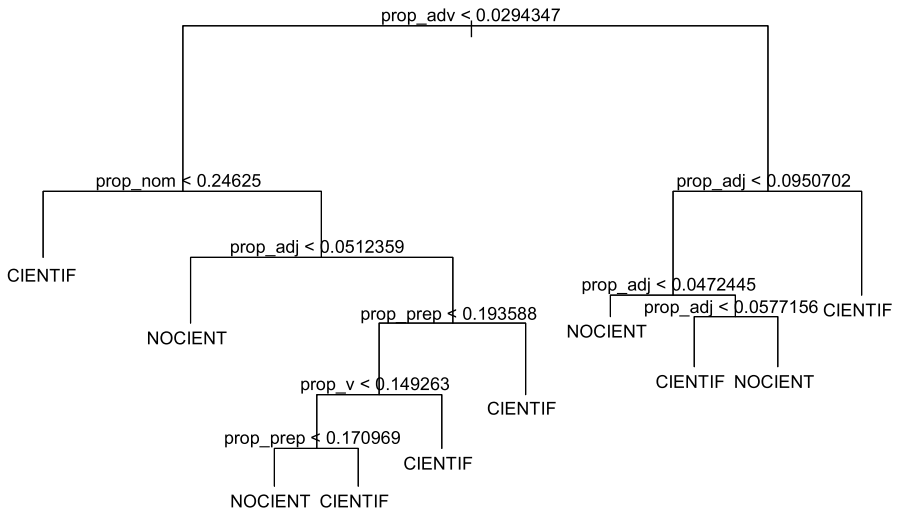

Gráfico 1: Árbol de clasificación

El gráfico 1 muestra el árbol resultante de esta aplicación. La variable regresora más fuertemente asociada con el género es la proporción de adverbios en el texto, categorizada como superiores e inferiores a 0.029 (2.9%). El corpus general queda así dividido en dos grupos, los que presentan más del 2.9% de adverbios y los que no. Luego intervienen en las sucesivas subdivisiones el número de nombres, adjetivos, verbos y preposiciones. Interpretando el árbol resultante, se

encuentran 10 perfiles de textos (que corresponden a los 10 nodos terminales) asociados con una de los dos géneros. Estos son:

- Textos con un porcentaje de adverbios inferior al 2.9% y un porcentaje de nombres menor a 25% son clasificados como textos CIENTÍFICOS.
- Textos con un porcentaje de adverbios inferior al 2.9%, un porcentaje de nombres mayor a 25% y de adjetivos inferior al 5% son clasificados como textos NO CIENTÍFICOS.
- Textos con un porcentaje de adverbios inferior al 2.9%, un porcentaje de nombres mayor a 25%, de adjetivos superior al 5%, de preposición menor al 17% y de verbos menor al 15% son clasificados como textos NO CIENTÍFICOS.
- Textos con un porcentaje de adverbios inferior al 2.9%, un porcentaje de nombres mayor a 25%, de adjetivos superior al 5%, de preposición entre 17% y 19%, y de verbos menor al 15% son clasificados como textos CIENTÍFICOS.
- Textos con un porcentaje de adverbios inferior al 2.9%, un porcentaje de nombres mayor a 25%, de adjetivos superior al 5%, de preposición inferior al 19%, y de verbos mayor al 15% son clasificados como textos CIENTÍFICOS.
- Textos con un porcentaje de adverbios inferior al 2.9%, un porcentaje de nombres mayor a 25%, de adjetivos superior al 5%, de preposición mayor al 19% son clasificados como textos CIENTÍFICOS.
- Textos con un porcentaje de adverbios superior al 2.9% y un porcentaje de adjetivos inferior al 4.7% son clasificados como textos NO CIENTÍFICOS.
- Textos con un porcentaje de adverbios superior al 2.9% y un porcentaje de adjetivos entre 4.7% y 5.7% son clasificados como textos CIENTÍFICOS.
- Textos con un porcentaje de adverbios superior al 2.9% y un porcentaje de adjetivos entre 5.7% y 9.5% son clasificados como textos NO CIENTÍFICOS.
- Textos con un porcentaje de adverbios superior al 2.9% y un porcentaje de adjetivos superior al 9.5% son clasificados como textos CIENTÍFICOS.

El árbol final fue evaluado utilizando la muestra de prueba, no fue utilizada en la construcción del mismo, hallando una tasa de mala clasificación del 14%, siendo 4% para los textos científicos y 28% para los no científicos. Respecto a la precisión y cobertura fueron de 84% y 96% para el género CIENTÍFICO y de 92% y 72% para los textos NO CIENTÍFICOS, respectivamente.

Tabla 2: Tasa de error estimada, Precisión y Cobertura

Medidas de evaluación		
	CIENTIFICO	**NO CIENTIFICO**
Tasa de error	4%	28%
Precisión	84%	92%
Cobertura	96%	72%

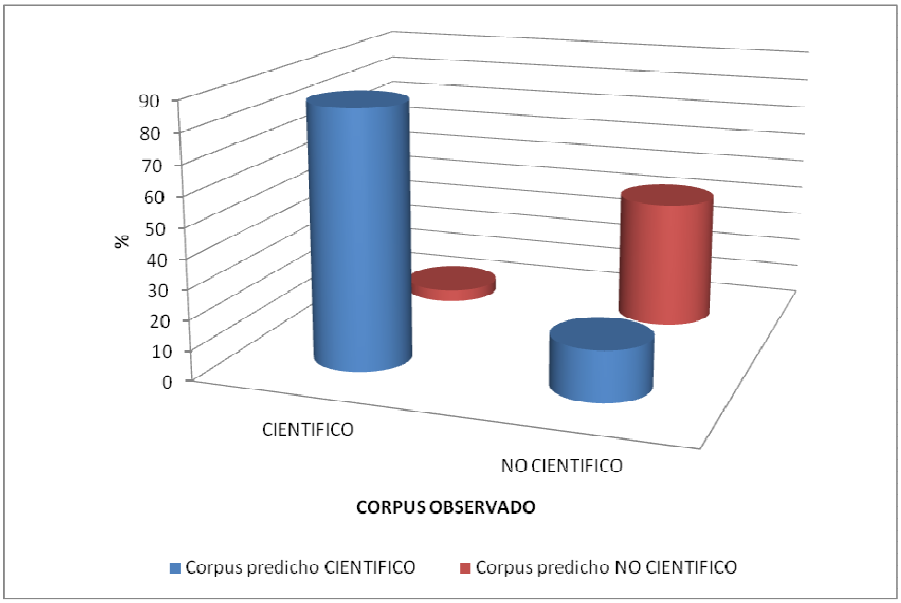

Grafico 2: Clasificación de textos según género mediante Árboles de Clasificación

3.2. Análisis de Regresión Logística

Se realizó un análisis de regresión logística para obtener una regla de clasificación que permita asignar los textos en estas dos poblaciones, definidas por el género al que pertenecen (Científico / No científico), en base a la frecuencia de cada categoría gramatical en el texto.

Para determinar cuáles categorías gramaticales son las responsables de la discriminación se utilizaron los tres algoritmos de selección de variables. En todos los casos se llega al mismo resultado: las variables que logran diferenciar a los dos corpus de textos analizados son el número de adjetivos, adverbios, conjunciones copulativas, determinantes, nombres y preposiciones.

Tabla 3: Coeficientes del modelo de regresión logística

Estimación máximo verosímil					
Coeficiente	**gl**	**Estimador**	**Error estándar**	**Est. Chi-cuadrado**	**Prob. asociada**
Intercepto	1	-0.6868	0.5507	1.5554	0.2123
adjetivos	1	0.1694	0.0562	9.0777	0.0026
adverbios	1	-0.3106	0.0800	15.0862	0.0001
Conj. Cop.	1	0.2769	0.1073	6.6566	0.0099
determinantes	1	0.1216	0.0464	6.8795	0.0087
nombres	1	-0.1995	0.0464	18.5044	<.0001
preposiciones	1	0.1575	0.0544	8.3925	0.0038

Este modelo permite, mediante la utilización de los coeficientes estimados, calcular para cada texto la probabilidad de pertenecer a cada uno de los corpus definidos por el género.

$$P(\in Cien/X) = \pi(x) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_j x_j+...+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_j x_j+...+\beta_p x_p}}$$

$$P(\in No\ Cien/X) = \pi(x) = \frac{1}{1+e^{\beta_0+\beta_1 x_1+...+\beta_j x_j+...+\beta_p x_p}}$$

Con este criterio un texto es asignado al corpus cuya probabilidad es máxima.

La bondad del ajuste se evaluó mediante el test de Hosmer-Lemeshow y la tasa de error de clasificación. Con el modelo de regresión logística obtenido durante la selección de variables se obtuvo una tasa de error global, estimada sobre el corpus de prueba, del 20% y la probabilidad asociada en el test de bondad de ajuste es p=0.9696 evidenciando lo adecuado del modelo. La tabla 4 presenta las medidas de precisión, cobertura y tasa de error para cada género.

Tabla 4: Tasa de error estimada, Precisión y Cobertura

Medidas de evaluación		
	CIENTIFICO	**NO CIENTIFICO**
Tasa de error	14%	26%
Precisión	83%	77%
Cobertura	86%	74%

Tabla 5: Razones de odds estimadas

Razón de odds			
Efecto	**Estimación puntual**	**IC 95%**	
adjetivos	1.185	1.061	1.323
adverbios	0.733	0.627	0.857
Conj. Cop.	1.319	1.069	1.628
determinantes	1.129	1.031	1.237
nombres	0.819	0.748	0.897
preposiciones	1.171	1.052	1.302

Los coeficientes del modelo de regresión logística permiten la interpretación de la misma. La razón de odds para el número de adjetivos es 1.19 lo cual indica que la chance de clasificar a un texto como Científico se incrementa en un 19% al aumentar en número de adjetivos en una unidad. Con respecto al número de adverbios la razón de odds es menor a la unidad por lo tanto si se interpreta el recíproco, 1/0.73=1.36, significa que la chance de clasificar un texto en el corpus No Científico aumenta un 36% al incrementarse en una unidad el número de adverbios. Si analizamos el efecto de las conjunciones copulativas, determinantes y preposiciones, al incrementar en una unidad cada una de estas categorías gramaticales, la chance de clasificar un texto como Científico se incrementa en un 32%, 13% y 17% respectivamente. Al igual que el efecto del número de adverbios, la

probabilidad de clasificar un texto como No Científico se incrementa en un 22% (1/0.82=0.22) al aumentar en una unidad la cantidad de nombres en el texto.

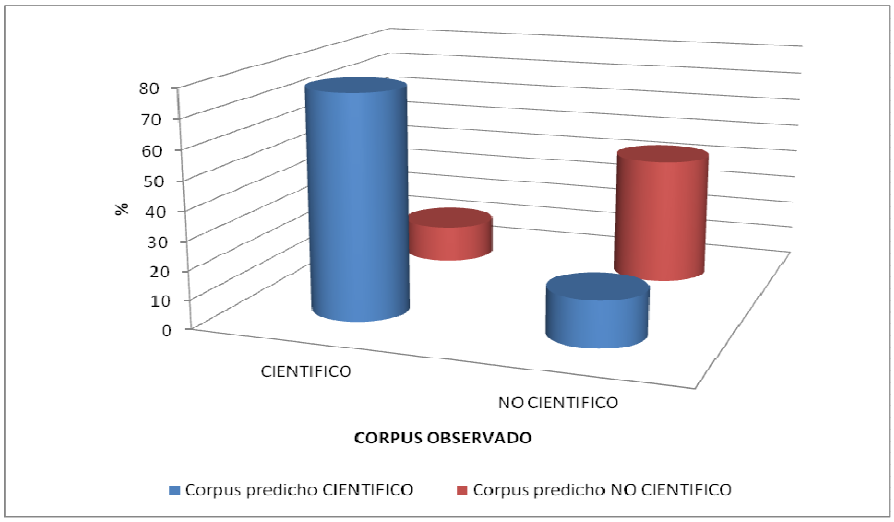

Grafico 3: Clasificación de textos según género mediante Regresión logística

4. CONCLUSIONES

Los resultados del análisis morfológico de los textos se analizaron teniendo en cuenta simultáneamente todas las mediciones realizadas sobre ellos.

El desempeño de las técnicas fue medido con la tasa de mala clasificación (TMC), la precisión (PR) y la cobertura (CO), calculadas sobre una muestra de textos no incluidos en la estimación del modelo y construcción del árbol. El árbol de clasificación presentó una TMC inferior a la del modelo logístico logrando clasificar con mayor precisión los textos científicos. Para el AC la TMC, PR y CO resultaron 4%, 84% y 96% para los textos científicos y 28%, 92% y 72% para los textos no científicos, respectivamente. Para el modelo de RL la TMC, PR y CO resultaron 14%, 83% y 86% para los textos científicos y 26%, 77% y 74% para los textos no científicos, respectivamente.

La diferencia en la tasa de mala clasificación sólo se diferenció en el corpus de textos científicos para el cual con el árbol se obtuvo un 4% de mala clasificación versus un 14% para el modelo de regresión logística.

En ambos tipos de análisis, las diferencias entre los dos tipos de textos están centradas principalmente en el porcentaje de adverbios, adjetivos, nombres y preposiciones presentes. Sin embargo, en el modelo de regresión logística han intervenido otras variables en la discriminación como los determinantes y conjunciones copulativas; mientras que el árbol de clasificación utiliza el porcentaje de verbos, categoría morfológica no utilizada en la regresión.

Una ventaja observada en el árbol de clasificación es la adaptación para recoger el comportamiento no aditivo de las variables predictoras, de manera que las interacciones se incluyen de manera automática. Sin embargo, en esta técnica se pierde información al tratar a las variables predictoras continuas como variables dicotómicas.

Referencias

Aitchison J. 1983. The Statistical Analysis of Compositional Data. Chapman & Hall, London.

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. 2013 Estudio exploratorio para la comparación de distintos tipos de textos: Textos Científicos y Textos No Científicos. Revista de Epistemología y Ciencias Humanas. Grupo IANUS. 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

Cuadras, C.M. 2008 Nuevos métodos de análisis multivariante. CMC Editions. Barcelona, España.

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

Khattre R. y Naik D. 1999 Applied Multivariate Statistics with SAS Software. SAS. Institute Inc. Cary, NC. USA.

Rodrigo Mateos, José Lázaro y Bès, Gabriel G. 2004 Análisis e implementación de clíticos en una herramienta declarativa de tratamiento automático de corpus. En VI Congreso de Lingüística General, Santiago de Compostela.

Solana, Z. Beltrán, C., Bender, C., Bonino, R., Deco, C., Koza, W., Méndez, B., Rodrigo, A., Tramallino, C. 2009 La interlengua de los aprendientes de español como L2. Aportes de la Lingüística Informática. GRUPO INFOSUR- Ediciones Juglaría.