Árboles de clasificación y su comparación con análisis de regresión logística aplicado a la clasificación de textos académicos

Classification trees and a comparison with an analysis of logistic regression applied to the classification of academic texts

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

Abstract

The problem of unit classification into known groups or populations is of great interest in statistics and several techniques have been developed to serve this classification purpose. This work presents a comparison of the classification tree and a logistic tree technique for text classification according to the field to which these texts belong (BIOMETRY and PHYLOSOPHY).

The development of the techniques has been measured with a Wrong Classification Rate (WCR) calculated over a sample of texts not included in the model estimation and tree construction. The classification tree showed a WCR lower than that of the logistic model, and humanistic texts have been classified with more accuracy.

The resulting WCR generated by the classification tree was 10% (17 percentage points within the Biometrics corpus and 3 percentage points within the Phylosophy one). As for the logistic regression model, there was a global result of 20%, and 17 percentage points and 23 percentage points within the Biometrics and Phylosophy corpuses.

Key words: Multivariate logistic regression, classification trees, automatic analysis of texts.

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. En este trabajo se presenta la comparación de la técnica de Árboles de Clasificación y Regresión logística para la clasificación de textos según la disciplina a la que pertenecen (BIOMETRIA y FILOSOFIA).

El desempeño de las técnicas fue medido con la Tasa de Mala Clasificación calculada 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 humanísticos.

La TMC obtenida con el árbol de clasificación fue de 10% (17% dentro del corpus de Biometría y 3% en Filosofía) mientras que con el modelo de regresión logística fie de 20% en forma global y 17% y 23% respectivamente dentro de los corpus de Biometría y Filosofía.

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

1. INTRODUCCION

El problema de la clasificación de unidades en grupos o categorías conocidas es de gran interés en estadística. Esto ha hecho que se desarrollaran diversidad de técnicas para cumplir este propósito. Entre ellas podemos citar al análisis discriminante, la regresión logística y los árboles de clasificación. Si bien el análisis discriminante es una de las técnicas más utilizadas para clasificación, el no cumplimiento del requerimiento de normalidad multivariada hace necesario utilizar técnicas alternativas, como la regresión logística y los árboles de clasificación, que no requieran dicho supuesto.

Este trabajo sigue la línea de investigación iniciada en Beltrán (2010) donde se busca evaluar las técnicas multivariadas aplicadas a la caracterización y clasificación de textos académicos. Este trabajo utiliza al analizador morfológico Smorph, implementado como etiquetador, para asignar una categoría morfológica a cada una de las ocurrencias lingüísticas.

Se utiliza la información resultante del análisis automático de textos académicos provenientes de dos áreas científicas (Biometría y Filosofía) para conformar una base de datos sobre la cual se aplica las técnicas de Árboles de Clasificación y Regresión Logística. Esta aplicación presenta diferencias respecto al análisis discriminante aplicado en trabajos previos.

Mediante la interpretación de los nodos del árbol y de los coeficientes del modelo logístico estimado se busca hallar las características, considerándolas simultáneamente a todas ellas, provenientes del análisis automático de los textos que son más discriminatorias de las áreas científicas de las cuales provienen.

2. MATERIAL Y METODOS

2.1. Diseño de la muestra

El marco muestral para la selección de la muestra es el utilizado en trabajos anteriores. Está compuesto por textos académicos, resúmenes de trabajos presentados a congresos y revistas, extraídos de internet pertenecientes a dos disciplinas: Biometría y Filosofía. La unidad de muestreo fue el texto y la selección de la muestra se llevó a cabo empleando un diseño muestral estratificado con selección proporcional al tamaño, siendo la medida de tamaño el “número de palabras del texto”.

Las muestras de cada área disciplinar se compararon respecto al número medio de palabras por texto. Esta comparación se requiere para evitar que el tamaño de los textos pueda afectar la discriminación entre las disciplinas, aunque la información incluida en los modelos estadísticos corresponde a las proporciones de cada categoría gramatical y no a la frecuencia de ellas.

Se seleccionaron 60 textos de cada disciplina y cada muestra fue particionada aleatoriamente en dos submuestras. En cada corpus, una submuestra se utiliza para el entrenamiento o estimación de los modelos y la otra para su validación.

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.

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.

El módulo post-smorph 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

El resultado del análisis de Smorph-Mps se almacena en un archivo de texto. Esta es la información que contendrá la base de datos.

La información resultante del análisis morfológico se dispone en una matriz de dimensión: tantas filas como cantidad de objetos lingüísticos tenga el texto (palabras) y tantas columnas como ocurrencia+lema+valores. De esta manera se obtiene una matriz con la estructura que se muestra en la tabla 1.

Tabla 1. Fragmento de la matriz de datos obtenida

MUESTRA	TEXTO	OCURRENCIA	LEMA	ETIQUETA
1	1	El	el	det
1	1	problema	problema	nom
1	1	de	de	prep
1	1	las	el	det
1	1	series	serie	nom
1	1	de	de	prep
1	1	tiempo	tiempo	nom
1	1	se	lo	cl
…	…	…	…	…
2	1	Uno	uno	pron
2	1	de	de	prep
2	1	los	el	det
2	1	agentes	agente	nom
2	1	que	que	rel
2	1	ha	haber	aux
2	1	provocado	provocar	v
2	1	una	una	det
2	1	verdadera	verdadera	adj
2	1	transformación	transformación	nom
2	1	en	en	prep
…	…	…	…	…

Abreviaturas:
'adj': adjetivo 'art': artículo 'nom': nombre 'prep': preposición 'v': verbo 'adv': adverbio 'cl': clítico 'aux': auxiliar ´cop': copulativo 'pun': signo de puntuación

Luego, a partir de esta matriz, donde cada fila es una palabra analizada, se confecciona la base de datos por documento que será analizada estadísticamente. Esta nueva matriz, donde cada fila o unidad experimental es el texto, retiene la información de las variables presentadas en la tabla 2 con la estructura definida en la tabla 3.

Tabla 2. Variables de la base de datos por documento

CORPUS	Corpus al que pertenece el texto
TEXTO	Identificador del texto dentro del corpus
Prop_adj	proporción de adjetivos del texto
Prop_adv	proporción de adverbios del texto
Prop_cl	proporción de clíticos del texto
Prop_cop	proporción de copulativos del texto
Prop_det	proporción de determinantes del texto
Prop_nom	proporción de nombres (sustantivos) del texto
Prop_prep	proporción de preposiciones del texto
Prop_v	proporción de verbos del texto

Tabla 3. Fragmento de la base de datos para análisis estadístico

CORPUS	TEXTO	adj	Adv	cl	cop	det	nom	prep	v	OTRO	TOTAL_PAL
1	1	21	4	4	8	30	48	33	17	20	185
1	2	14	0	5	4	14	27	20	9	17	110
1	3	16	5	11	5	28	47	26	18	25	181
…	…	…	…	…	…	…	…	…	…	…	…
2	28	14	2	3	6	30	60	39	16	16	186
2	29	14	0	4	5	24	40	26	12	16	141
2	30	18	5	2	5	35	49	30	19	20	183

2.4. Metodología Estadística

Uno de los problemas que concentran gran interés en estadística es la clasificación de objetos o unidades en grupos o poblaciones.

Es posible distinguir dos enfoques del problema de clasificación:

- El primero 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, para este caso las técnicas más utilizadas son el Análisis Discriminante y la Regresión Logística.
- 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, dentro de estas técnicas se encuentra el Análisis de Clusters.

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 y en la construcción del modelo no hay ningún supuesto en cuanto a la distribución de probabilidad de las variables por lo que puede incluirse cualquier tipo de variable. Este modelo puede considerarse como una fórmula para calcular la probabilidad de pertenencia a uno de los grupos, de manera que se asigna cada unidad al grupo cuya probabilidad de pertenencia estimada sea mayor.

Los árboles de clasificación 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.

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 área disciplinar a la que corresponde: BIOMETRIA-FILOSOFIA 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 la disciplina 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. El proceso termina cuando se hayan clasificado todas las observaciones correctamente en su grupo.

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”.

Los AC son más flexibles que otras técnicas de clasificación porque permiten incorporar predictores medidos virtualmente en cualquier escala: continua, ordinal o mezclas de ambas escalas. Por su condición "no paramétrica", constituyen una opción favorable entre otras técnicas alternativas como el análisis discriminante o la regresión logística.

A modo de ejemplo, supóngase una variable respuesta Y que se pretende discriminar en función de tres predictores X1, X2 y X3. Asumir además que la variable respuesta Y puede asumir dos valores posibles o categorías (SI/NO), las variables explicativas X1 y X2 son cuantitativas continuas y el tercer regresor X3 es una variable categórica nominal que puede asumir sólo dos valores o categorías posibles (a/b). El árbol de la figura 1 representa los resultados de la aplicación de la técnica de AC a este ejemplo.

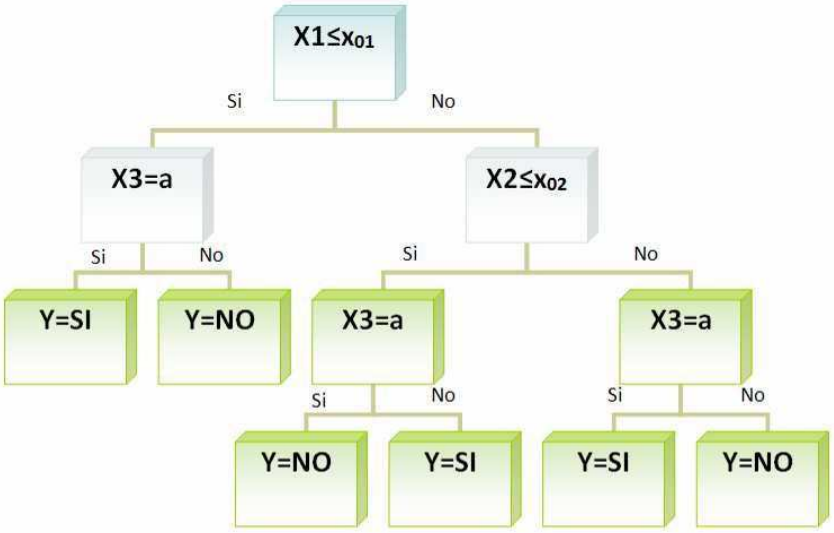

Figura 1: AC

Del árbol de la figura 1, a partir de los seis nodos terminales, se concluye que los regresores X1, X2 y X3 discriminan las categorías de la variable respuesta Y (SI/NO) de la siguiente manera:

- Si X1 es menor o igual que x_{01} y X3=a entonces resulta Y=SI
- Si X1 es menor o igual que x_{01} y X3=b entonces resulta Y=NO
- Si X1 es mayor que x_{01}, X2 es menor o igual a x_{02} y X3=a entonces resulta Y=NO
- Si X1 es mayor que x_{01}, X2 es menor o igual a x_{02} y X3=b entonces resulta Y=SI
- Si X1 es mayor que x_{01}, X2 es mayor a x_{02} y X3=a entonces resulta Y=SI
- Si X1 es mayor que x01, X2 es mayor a x02 y X3=b entonces resulta Y=NO

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

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:

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

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 (en este caso las disciplinas). 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

3.1. Análisis preliminar.

Previamente se hizo referencia a la comparación de los corpus respecto a la cantidad de palabras por texto. La misma se lleva a cabo mediante el test no paramétrico de Wilcoxon para muestras independientes arrojando una probabilidad asociada p=0.796, evidenciando que no existen diferencias significativas entre los corpus respecto al tamaño de los textos.

Comparaciones similares entre los corpus se llevan a cabo para las restantes variables hallando diferencias significativas ($p<0.05$) para la proporción de clíticos, adverbios, determinantes, nombres y preposiciones en los documentos analizados (Tabla 4). La proporción de clíticos, nombres y preposiciones es mayor en los textos de biometría y la proporción de adverbios y determinantes es superior en los textos de filosofía.

Tabla 4. Comparación mediante test de Wilcoxon

Variable	BIOMETRIA	FILOSOFIA	Valor de p
prop_adj	0.08968	0.10098	0.13536
prop_adv	0.01763	0.03136	0.00440
prop_cl	0.02603	0.01572	0.00130
prop_cop	0.02788	0.03376	0.10864
prop_det	0.16037	0.17645	0.01246
prop_nom	0.25971	0.24251	0.04514
prop_prep	0.17819	0.16219	0.02760
prop_v	0.12730	0.12493	0.88246

3.2. Arboles de clasificación

Se aplicó la técnica de árboles de clasificación para obtener reglas de clasificación que permitan asignar los textos en estas dos poblaciones, definidas por el área científica a la que pertenecen: BIOMETRIA y FILOSOFIA. 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, clíticos, preposición y adjetivos. El árbol final presenta 7 nodos terminales.

La figura 2 muestra el árbol resultante de esta aplicación. La variable regresora más fuertemente asociada con el área disciplinar 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. Este primer subgrupo es clasificado como FILOSOFIA y constituye un nodo terminal, mientras que el grupo con menos del 2.9% de adverbios es dividido nuevamente. El predictor más relevante en esta segunda división fue la proporción de clíticos en el texto, categorizada en inferior y superior a 0.011 (1.1%). El subgrupo con un porcentaje de clíticos inferior a 1.1% (que ya tenían definido un porcentaje de adverbios inferior a 2.9%) pertenecen al área de BIOMETRIA y constituyen el segundo nodo terminal. Por otro lado, el grupo con un porcentaje de clíticos superior al 1.1% (y con menos del 2.9% de adverbios) vuelve a subdividirse respecto al mismo predictor, definiendo un nuevo nodo terminal para aquellos que presentan un porcentaje de clíticos superior a 2.8%, clasificándolos en el área de BIOMETRIA. Continuando con la subdivisión del otro grupo, la proporción de preposiciones es el regresor determinante. Textos con un porcentaje de preposiciones superior a 20% (y que ya tenían bajo porcentaje de adverbios y un porcentaje de clíticos entre 1.1% y 2.8%) se clasifican en BIOMETRIA y constituye el cuarto nodo terminal. El resto de los textos se vuelven a separar en dos grupos en función de la proporción de adjetivos. Textos con un porcentaje de preposiciones inferior al 20% (y que ya tenían bajo porcentaje de adverbios y un porcentaje de clíticos entre 1.1% y 2.8%) se dividen teniendo en cuenta los adjetivos. Un porcentaje de adjetivos menor a 7,5% pertenecen a FILOSOFIA y constituyen el quinto nodo terminal. Cuando el porcentaje de adjetivos mayor a 7,5% y menor a 11% pertenecen a BIOMETRIA y constituyen el sexto nodo terminal, mientras que si el porcentaje de adjetivos mayor a 11% pertenecen a FILOSOFIA y constituyen el séptimo nodo terminal.

El árbol final fue evaluado utilizando la muestra que no fue utilizada en la construcción del mismo hallando una tasa de mala clasificación del 10%, siendo 17% para biometría y 3% para filosofía (Figura 3).

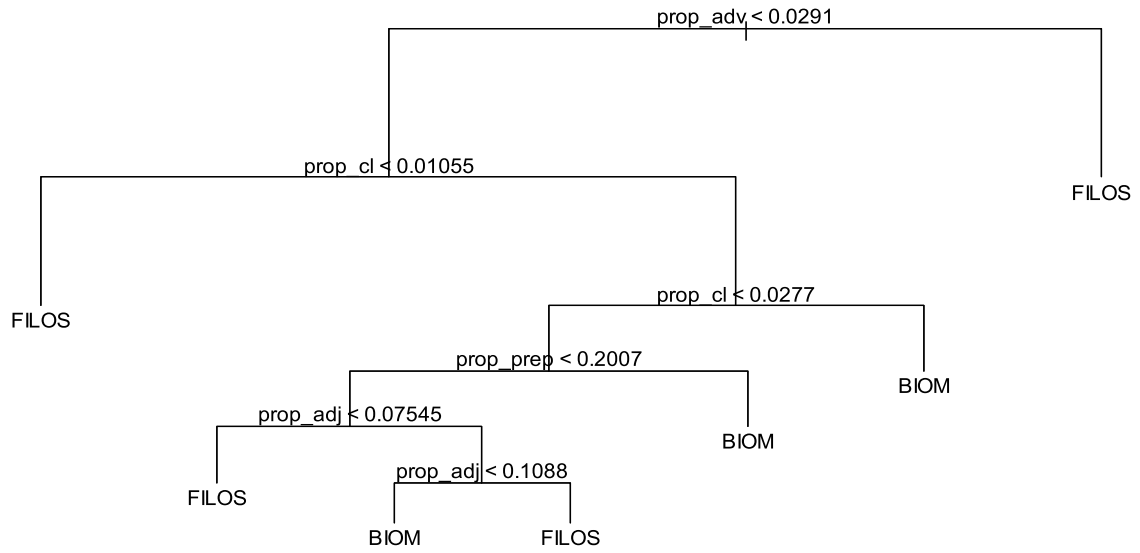

Figura 2: Árbol de clasificación de textos según área disciplinar.

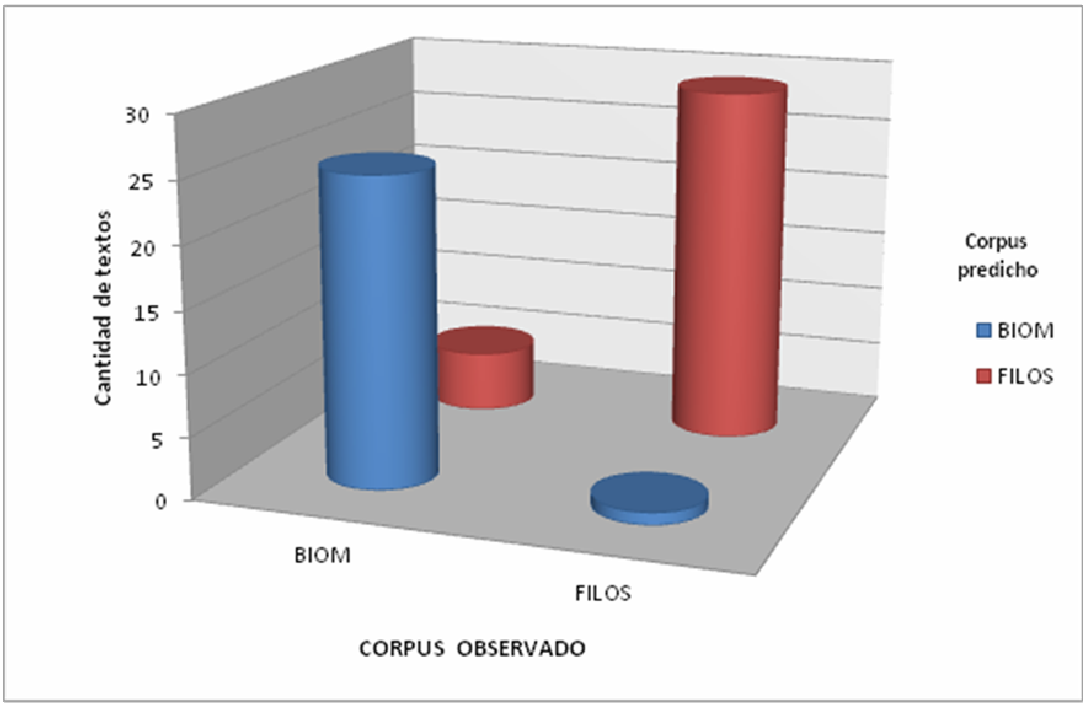

Figura 3: Clasificación de los textos de prueba mediante el árbol final.

3.3. 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 área científica a la que pertenecen, 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 la proporción de adverbios y de clíticos.

El modelo final estimado luego de la selección de variables se muestra en la tabla 5.

Tabla 5: Coeficientes del modelo de regresión logística final

Estimación máximo verosímil					
Coeficiente	**gl**	**Estimador**	**Error estándar**	**Est. Chi-cuadrado**	**Prob. asociada**
Intercepto	1	-0.0362	0.7293	0.0025	0.9604
Prop_adv	1	-78.7323	27.6806	8.0902	0.0045
Prop_cl	1	90.0310	30.0669	8.9662	0.0028

La bondad del ajuste de evalúa 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 obtiene una tasa de error global del 17% mediante validación cruzada y la probabilidad asociada en el test de bondad de ajuste es p=0.2062 evidenciando lo adecuado del modelo. Respecto a la tasa de mala clasificación, ésta resultó de un 20%, siendo 17% dentro del área de biometría y 23% dentro de filosofía (Figura 4).

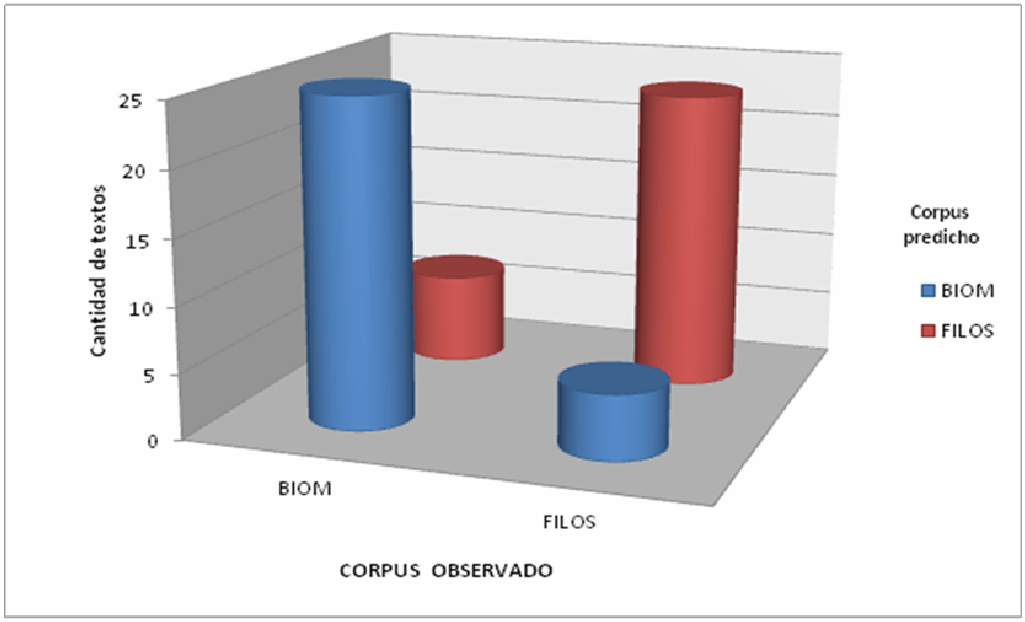

Figura 4: Clasificación de los textos de prueba mediante modelo logístico.

4. CONCLUSIONES

Los resultados del análisis morfológico de los textos se analizaron teniendo en cuenta simultáneamente todas las mediciones realizadas sobre ellos sin aplicar ninguna transformación a las variables.

Si bien el número de unidades utilizadas en el entrenamiento y evaluación no era elevado, el árbol de clasificación obtenido mostró un buen desempeño frente al modelo de regresión logística, 10% y 20% respectivamente. La diferencia en la tasa de mala clasificación sólo se diferenció en el área de Filosofía para la cual con el árbol se obtuvo un 3% de mala clasificación versus un 23% para el modelo de regresión logística.

En ambos tipos de análisis, las diferencias entre los dos tipos de textos está centrada principalmente en el porcentaje de clíticos y de adverbios presentes. Sin embargo, en esta nueva aplicación de los árboles de clasificación han intervenido otras variables en la discriminación como el porcentaje de preposiciones y adjetivos. Estas variables intervienen determinando una interacción entre las variables que no se alcanza a observar en el modelo de regresión logística.

Similares resultados se hallaron en Beltrán (2010) utilizando un análisis discriminante sobre las variables transformadas, debido al requerimiento de distribución Normal de las mismas.

Esta particularidad de los textos analizados de estas disciplinas puede deberse a que, en los textos de biometría/estadística hay más clíticos que en los humanísticos por la frecuencia de expresiones impersonales o pasivas con el clítico “se” del tipo:

“se ajusta un modelo cuadrático”

“se estima la variancia poblacional”

Mientras en los textos de filosofía se manifiesta la presencia de mayor proporción de adverbios.

Respecto a la metodología estadística planteada, se puede afirmar que entre las ventajas de los árboles de clasificación está la adaptación para recoger el comportamiento no aditivo de las variables predictoras, de manera que las interacciones se incluyen de manera automática. Por el

contrario, entre las desventajas se evidencia que las variables predictoras continuas se tratan como variables dicotómicas perdiendo información.

Esta metodología puede ser generalizada a un número mayor de disciplinas. Asimismo, se continuará trabajando en la comparación de las técnicas estadísticas mediante simulación.

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. 2010 Análisis discriminante aplicado a textos académicos: Biometría y Filosofía. 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

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

Hosmer, D.W.; Lemeshow, S. 1989 Applied Logistic Regression. John Wiley & Sons. New York.

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

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

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

Maindonald, J.; Braun, J. 2004. Data Analysis and Graphics Using R.- an example-based approach. Cambridge University Press.

Pogliano, A.M. 2010 “Análisis Estadístico de Datos Aplicados al Estudio de Calidad en Servicios de Traducción”. Tesis Lic. en estadística. Facultad de Cs. Económicas y estadística. UNR.

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.

Verzani, J. 2005 Using R for Introductory Statistics. CHAPMAN & HALL/CRC. Boca Raton London New York Washington, D.C.