Revista Colombiana de Estadística
Junio 2011, volumen 34, no. 1, pp. 15 a 37

Donde se muestran algunos resultados de atribución de autor en torno a la obra cervantina

Wherein are Shown some Results of Autorship Attribution to Cervantes' Work

Freddy López[a]

Departamento de Matemáticas, Instituto Venezolano de Investigaciones Científicas, Estado Miranda, Venezuela

Resumen

En este artículo se aplican algunos métodos de clasificación a un conjunto de textos con el objetivo de estudiar la probabilidad que el libro *Novela de la tía fingida* haya sido escrita por Miguel de Cervantes. Esta novela se le ha atribuido históricamente, pero existen algunas posiciones encontradas al respecto. Los métodos usados en este artículo contemplan: regresión logística, regresión logística aditiva, análisis discriminante lineal, cuadrático, regularizado, de mezclas y flexible, árboles de clasificación, método de los k-ésimos vecinos más cercanos, método de Bayes ingenuo y máquinas de soporte vectorial.

Los métodos fueron calibrados y aplicados utilizando un corpus de autores contemporáneos a Cervantes (Lope de Vega, Jerónimo de Pasamonte, Alonso Fernández de Avellaneda, Mateo Alemán y Francisco de Quevedo) junto con más de cuarenta variables, principalmente palabras y signos de puntuación, medidas sobre muestras de los textos escritos por estos autores.

Con respecto a estos métodos, la mayoría clasifica la obra como cervantina; sin embargo, es recomendable ampliar el corpus utilizado para el estudio e incluir más autores para la comparación.

***Palabras clave*:** análisis discriminante, árboles de clasificación, máquinas de aprendizaje, regla de Bayes, regresión logística, validación cruzada.

Abstract

In this paper, some classification methods are applied to a set of texts with the aim of studying the probability that the book *Novela de la tía fingida* has been written by Miguel de Cervantes. This novel has been historically attributed to him but there are some encountered positions about this. The methods used in this paper range from: logistic regression, additive logistic

[a]Estudiante de postgrado. E-mail: freddy.vate01@gmail.com

regression, linear, quadratic, regularized, mixture and flexible discriminant analysis, classification tree, k-nearest neighbour, Naive Bayes method and support vector machines.

Methods were trained and applied using a corpus of authors contemporary to Cervantes as Lope de Vega, Jerónimo de Pasamonte, Alonso Fernández de Avellaneda, Mateo Alemán, and Francisco de Quevedo and more than forty variables, mainly words and punctuation marks, measured over written texts by these authors.

Respect to these methods, most of them classify the novel as another Cervantes' work; however, is our recommendation to include more texts from these authors and more authors.

***Key words*:** Bayes rule, Classification tree, Cross validation, Discriminant analysis, Logistic regression, Machine learning.

1. Introducción

El problema de atribución de autor hace referencia a la asignación de un autor a un texto cuya autoría es desconocida (anónima) y es un problema que puede abordarse de diversas maneras, como por ejemplo, atribuir un texto a determinado autor por el lugar donde fue encontrado, por semejanzas y giros del lenguaje propios de un autor, por el estilo, por el tema tratado, por el metro y ritmo (en textos poéticos), etc.

La atribución *cuantitativa* consiste en realizar mediciones al texto anónimo, y compararlas con textos de autores de la época y asignar la autoría del texto a aquel autor al que esté estadísticamente más cercano.

Los problemas de atribución pueden pensarse como un problema de clasificación bajo el supuesto de que se conocen con certeza los posibles autores del texto. Muchas veces no se puede estar plenamente seguro de que los autores postulados sean efectivamente posibles autores, y es un problema que puede no tener solución.

Si se tiene un grupo de candidatos a autores que sea confiable, entonces es razonable aplicar técnicas estadísticas (multivariantes), reducción de dimensiones y técnicas de clasificación. En general, las variables que se utilizan en este tipo de trabajo son el resultado del conteo de palabras más frecuentes (ver sección 2).

En la literatura reciente se encuentra que Jockers, Witten & Criddle (2008), comparan dos métodos de clasificación para determinar la autoría del *Libro del Mormón* entre un grupo de posibles autores (Salomón Spalding, Sidney Rigdon y Oliver Cowdery). Al final, el estudio concluye que el libro es autoría de Rigdon y Spalding.

Hoover (2002) investiga el análisis de conglomerados dentro de los textos de un mismo autor y logra clasificar correctamente todas las secciones de todas las novelas investigadas. Cada una de estas secciones consta de 2500 palabras. Lamentablemente, reporta que cuando el número de palabras más frecuentes se hace muy grande, la clasificación puede fallar. Luego propone utilizar el orden de aparición de las palabras más frecuentes dentro de la sección estudiada.

Binongo (2003) utiliza los dos primeros componentes principales para la distinción de las características de los dos autores que estudia: L.F. Baum y R.P. Thompson. El problema abordado por Binongo fue la autoría del libro *The Royal Book of Oz* (el libro número 15 de la saga; ver Baum 2001), y contó con esos dos autores potenciales. En su trabajo, el primer componente separó claramente a los dos autores y parte de la validación de sus resultados la logró incluyendo, de forma ilustrativa (Lebart, Morineau & Warwick 1984) otros trabajos de Baum. El trabajo concluye gratamente con la inclusión del libro de Martin Gardner *Visitors from Oz* (Gardner 1998) y notando que la forma de escribir de Gardner está más próxima al estilo de Thompson que de Baum, creador original de la historia.

Koppel, Schler & Argamon (2009) dan un resumen sobre las técnicas estadísticas empleadas hasta hoy y aborda especialmente el tema de las máquinas de soporte vectorial. Proponen un método denominado *desenmascaramiento*, cuya idea principal consiste en remover, por etapas, las variables que son más útiles al momento de separar un texto de un autor y de otro; de esta forma, la precisión en la clasificación se deteriorará conforme se vayan removiendo variables. La idea es que cuando un texto pertenece a un autor, sus características propias permanecerán en él a pesar del tema o género tratado. Cuando las variables que mejor separan han sido removidas, los textos de un mismo autor serán prácticamente indistinguibles.

Dos excelentes trabajos de recopilación y fuentes bibliográficas son el trabajo de Joula (2006) y el trabajo de Grieve (2007). Este último compara las distintas variables frecuentemente utilizadas en la atribución de autor.

Este trabajo está organizado como sigue: La sección 2 describe las variables a ser utilizadas, cómo se trabajan y cuáles serán los textos a utilizar; la sección 3 lista los métodos que se emplearán dando una cortísima descripción de cada uno; la sección 4 presenta los resultados de entrenar y aplicar los métodos con los textos bajo examen, incluyendo la aplicación a la *Novela de la tía fingida*, atribuida a Cervantes y, por último, se presentan algunas conclusiones.

2. Procesamiento de los textos

2.1. Las variables

El procedimiento inicial consistió en tomar un libro y considerar sus divisiones naturales, los capítulos, como individuos. En algunos casos no fue posible dado que el tamaño del capítulo no era razonable o el libro no presentaba divisiones. Estos puntos serán abordados más adelante en la sección 4. Así, a estos individuos les serán medidas las siguientes variables:

1. **Conteo de las palabras más utilizadas.** Una palabra es una cadena de caracteres delimitada por un signo de puntuación o espacio. Estas se buscan de la siguiente manera: se colapsan todos textos en estudio, se contabilizan todas las palabras y se ordenan. Las palabras más frecuentes son utilizadas entonces como variables (en términos de estadística multivariante).

2. **Conteo de las frases de dos y tres palabras más utilizadas.** Estas frases de dos y tres palabras se encuentran de la misma manera como se describe en el ítem anterior, con la diferencia de que no son palabras aisladas. Estas no pueden ser muchas porque no se repiten mucho en todos los textos.

3. **Conteo de signos de puntuación más utilizados.** La cantidad de signos de puntuación utilizados en una novela es amplia. Para utilizar los signos de puntuación como variables o características, son necesarias aquellas que más se repiten.

4. **Medida de riqueza verbal.** Existen muchas medidas de riqueza verbal (un listado de ellas puede ser consultado en Grieve 2007). Una de ellas, la que es utilizada aquí, consiste en el número de palabras distintas divididas entre el número palabras utilizadas en el texto. Una pequeña introducción de esta medida puede revisarse en Bird, Klein & Loper (2009, p. 8).

Así, para todos los textos utilizados (ver sección 2.4),

```
"que",    "de",     "y",     "la",    "a",      "en",   "el",  "no",
"con",    "los",    "se",    "por",   "lo",     "las",  "su",  "le",
"me",     "como",   "del",   "un",    "si",     "mi",   "es",  "yo",
"para",   "al",     "una",   "dijo",  "porque", "ni"
```

son las 30 palabras más repetidas de un total de 47.560 palabras, mientras que

```
",", ".", ";", "-", ":", "?", "£", "ą", "!"
```

son los signos de puntuación que más se repiten, de un total de 28 signos de puntuación.

Además de ellos,

```
"de la"  ,  "lo que",  "que no"
```

son las 3 frases de dos palabras más usadas, y

```
"de lo que"
```

es la frase de tres palabras más repetida.

Estas características son contadas en cada capítulo y guardadas en un registro, junto a la medida de riqueza verbal (`RT`). Frecuentemente, a estas características (signos de puntuación y palabras) se les denomina *tokens*. Es importante observar que los *tokens* anteriores no se encuentran necesariamente en todos los textos utilizados. En efecto, los *tokens* definidos arriba son los que más se repiten en todos los textos a ser utilizados y no hay garantía que cada uno de ellos se encuentre en cada capítulo.

Además, se cuentan cuántos *tokens* tiene cada fragmento, y luego se divide por él cada una de las mediciones anteriores; de esta manera, se tienen registros estandarizados por el número de *tokens* que tiene cada fragmento a ser estudiado.

En la siguiente sección se hará una pequeña demostración de cómo se trabaja con cada capítulo.

2.2. Ejemplo

Supóngase que es de interés obtener las medidas anteriormente descritas al siguiente fragmento del prólogo a las *Novelas Ejemplares* de Cervantes.

> A esto se aplicó mi ingenio, por aquí me lleva mi inclinación, y más, que me doy a entender, y es así, que yo soy el primero que he novelado en lengua castellana, que las muchas novelas que en ella andan impresas todas son traducidas de lenguas estranjeras (sic), y éstas son mías propias, no imitadas ni hurtadas: mi ingenio las engendró, y las parió mi pluma, y van creciendo en los brazos de la estampa. Tras ellas, si la vida no me deja, te ofrezco los *Trabajos de Persiles*, libro que se atreve a competir con Heliodoro, si ya por atrevido no sale con las manos en la cabeza; y primero verás, y con brevedad dilatadas, las hazañas de don Quijote y donaires de Sancho Panza, y luego las *Semanas del jardín*.

Así, este fragmento, contiene seis veces la palabra `que`, cinco veces la palabra `de`, nueve veces la palabra `y`, etc. Nótese que no contiene las frases `lo que`, `que no`, `de lo que` mientras la frase `de la` aparece una vez. Estos valores serán divididos por la cantidad de *tokens* que haya en el texto.

El fragmento contiene además diecisiete comas (`,`), dos puntos (`.`), un punto y coma (`;`), un dos puntos (`:`) y no hay signos de admiración ni interrogación.

Este fragmento muestra 87 *tokens* distintos de un total de 153, así, su riqueza verbal, según se ha definido, es $\frac{87}{153} = 0,568627451$, y la proporción de la palabra `que`, por ejemplo, es $\frac{6}{153} = 0,03921569$.

De esta forma, para este fragmento, se cuenta con los siguientes valores observados para las variables estudiadas:

que	de	y	la	a
0,039215686	0,032679739	0,058823529	0,019607843	0,019607843
en	el	no	con	los
0,026143791	0,006535948	0,019607843	0,019607843	0,013071895
se	por	lo	las	su
0,013071895	0,013071895	0,000000000	0,039215686	0,000000000
le	me	como	del	un
0,000000000	0,019607843	0,000000000	0,006535948	0,000000000
si	mi	es	yo	para
0,013071895	0,026143791	0,006535948	0,006535948	0,000000000
al	una	dijo	porque	ni
0,000000000	0,000000000	0,000000000	0,000000000	0,006535948
s.coma	s.punto	s.puntoycoma	s.guión	s.dospuntos
0,111111111	0,013071895	0,006535948	0,000000000	0,006535948
s.intcerrado	s.intabierto	s.exclabierto	s.exclcerrado	RT
0,000000000	0,000000000	0,000000000	0,000000000	0,568627451
de_la	lo_que	que_no	de_lo_que	
0,006535948	0,000000000	0,000000000	0,000000000	

2.3. Reducción de dimensiones

Una vez se cuente con todas estas medidas para cada uno de los textos en estudio, se realizará un análisis de componentes principales (Jolliffe 2002) y se retendrán algunos componentes. Se compararán los resultados de clasificación con el uso de las variables originales y los componentes retenidos.

2.4. Textos a ser procesados

Se consideran textos de Cervantes así como de otros autores de la época. A continuación se listarán los autores y libros utilizados:

- De Cervantes: Las dos partes de Don Quijote: *El ingenioso hidalgo don Quijote de la Mancha* y *El ingenioso caballero don Quijote de la Mancha*; sus novelas ejemplares: *La gitanilla*, *El amante liberal*, *Rinconete y Cortadillo*, *La española inglesa*, *El licenciado Vidriera*, *La fuerza de la sangre*, *El celoso extremeño*, *La ilustre fregona*, *Las dos doncellas*, *La señora Cornelia*, *El casamiento engañoso* y *Los trabajos de Persiles y Segismunda*.
- De Lope de Vega: *La Dorotea* y *Novelas a Marcia Leonarda*.
- De Jerónimo de Pasamonte: *Vida y trabajos de Jerónimo de Pasamonte*.
- De Alonso Fernández de Avellaneda: *Segundo tomo del Ingenioso Hidalgo don Quijote de la Mancha*.
- De Mateo Alemán y de Enero: Los dos libros de Guzmán de Alfarache: *Primera parte de Guzmán de Alfarache*, *Segunda parte de la vida de Guzmán de Alfarache, atalaya de la vida humana*.
- Francisco de Quevedo y Villegas: El Buscón: *Historia de la vida del Buscón, llamado Don Pablos, ejemplo de vagabundos y espejo de tacaños*.

3. Métodos de clasificación

En esta sección se hará una muy corta explicación de los métodos que se emplearán aquí. Para un estudio en profundidad de cada técnica se recomienda revisar la bibliografía recomendada en cada aparte. En lo sucesivo, Y representará una variable dicotómica, tomando valor 1 cuando el texto evaluado pertenece a Cervantes y 0 en caso contrario. $\mathbf{x}$ representa, por su parte, las variables predictoras (que están en función de las variables descritas en la sección 2) con las que se busca explicar Y.

3.1. Regresión logística

La regresión logística es un tipo de regresión en donde la variable respuesta es categórica. En el caso que nos ocupa, es específicamente dicotómica. Siguiendo a Hosmer & Lemeshow (2000), la esperanza condicional que el resultado

(`1=Cervantes, 0=otro autor`) esté presente se denotará por $E(Y = 1 \mid \mathbf{x}) = \pi(\mathbf{x})$, donde $E(\cdot)$ es el operador esperanza. La forma específica del modelo de regresión logística es

$$\pi(\mathbf{x}) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}} \tag{1}$$

que, luego de aplicar la 'transformación logit', toma la forma

$$g(\mathbf{x}) = \ln\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{2}$$

donde los parámetros a ser estimados son $\beta_0, \ldots, \beta_p$.

3.2. Regresión logística aditiva

Cuando cada término lineal es reemplazado por una función suavizada más general, digamos f_j, se obtiene la denominada regresión logística aditiva, cuya forma es

$$g(\mathbf{x}) = \beta_0 + f_1(x_1) + \cdots + f_p(x_p)$$

Para trabajar con esta ecuación se debe minimizar

$$\sum_{i=1}^{n}(y_i - \alpha - \sum_{j=1}^{p} f_j(x_{ij}))^2 + \sum_{j=1}^{p} \lambda_j \int f_j''(t_j)^2 dt_j \tag{3}$$

donde λ_j son parámetros que deben ser calibrados y, además, se puede demostrar que la minimización de (3) es un modelo aditivo de splines cúbicos. La forma más popular de maximizar esta ecuación y hallar a las funciones f_j, es conocida como 'algoritmo de *backfitting*'. Consiste básicamente en fijar un valor para α (por ejemplo el promedio de los valores de Y) y luego aplicar un suavizado de splines cúbicos a $\{y_i + \widehat{\alpha} + \sum_{k\neq j} \widehat{f}_k(x_{ik})\}$, $i = 1, \ldots, n$, para obtener un nuevo valor de $\widehat{f}_j$ (para mayores detalles consultar Hastie, Tibshirani & Friedman 2009).

3.3. Análisis discriminante lineal y cuadrático

Asúmase que se tienen dos poblaciones normales con distintos vectores de medias $\boldsymbol{\mu}_0$ y $\boldsymbol{\mu}_1$ e igual matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$. Asúmase que la de dimensión de esas poblaciones es p.

La función discriminante es la combinación lineal de las p variables que forman el conjunto de datos tal que se maximice la distancia entre los dos grupos de vectores de medias.

La combinación lineal es de la forma $z = \mathbf{a}'\mathbf{y}$, donde el vector de parámetros a estimar es $\mathbf{a}$. Se puede demostrar $\mathbf{a}$ está en función de $\mathbf{S}_i$, n_i, $\mathbf{S}_{\text{man}} = \frac{(n_0-1)\mathbf{S}_0+(n_1-1)\mathbf{S}_1}{n_0+n_1-2}$, los estimadores de la matriz de covarianza de cada población, los tamaños de muestra y la matriz mancomunada de varianzas y covarianzas con $\overline{\mathbf{y}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{y}_{ij}$, para $i = 0, 1$. Se utiliza $\mathbf{S}_{\text{man}}$ bajo el supuesto de que las dos

poblaciones involucradas tengan iguales matrices de varianzas y covarianzas. En este caso, una nueva observación, $\mathbf{x}_0$, se clasificará en una de las poblaciones de acuerdo con la cercanía de $\mathbf{a}'\mathbf{x}_0$ al punto medio $m = \frac{1}{2}(\overline{x}_0 + \overline{x}_1)$, donde $\overline{x}_i = \mathbf{a}\overline{\mathbf{y}}_i$ (ver Rencher (2002), capítulo 8 y Johnson & Wichern (1998), capítulo 11).

Por otra parte, si la igualdad en las matrices de varianzas y covarianzas no se puede sostener, se puede usar la regla de asignar $\mathbf{x}_0$ a la población que haga máximo el valor $Q_i(\mathbf{x}_0) = \log p_i - \frac{1}{2}\log|\mathbf{S}_i| - \frac{1}{2}(\mathbf{x}_0 - \overline{\mathbf{y}})^t\mathbf{S}_i^{-1}(\mathbf{x}_0 - \overline{\mathbf{y}})$, $i = 0, 1$, donde p_i son las probabilidades previas de cada grupo. Este cómputo, $Q_i(\mathbf{x}_0)$, es conocido como *puntaje de discriminación cuadrático*.

En general, los términos p_i son desconocidos y se suele trabajar asignándoles valores proporcionales al número de individuos que presenta cada población según su aparición en el conjunto de datos. En este trabajo se adoptará este enfoque (algunos autores, sin embargo, sugieren el uso de p_i completamente equilibrados. Ver por ejemplo Johnson & Wichern 1998, pp. 670-672).

3.4. Análisis discriminante regularizado

El análisis discriminante regularizado busca un equilibrio entre el análisis discriminante lineal y el análisis discriminante cuadrático; obligando a la matriz de covarianza muestral de cada población, $\mathbf{S}_k$, $k = 0, 1$, acercarse a la matriz mancomunada de covarianza $\mathbf{S}_{\text{man}}$, en un intento de reducir el sesgo en la estimación de los autovalores (Jolliffe 2002, p. 207). La regularización tiene la forma

$$\mathbf{S}_k(\alpha) = \alpha\mathbf{S}_k + (1 - \alpha)\mathbf{S}_{\text{man}}$$

con $k = 0, 1$ y $\alpha \in [0, 1]$; en la práctica se suele hallar el valor de α por validación cruzada.

3.5. Análisis discriminante de mezclas

Supóngase que los datos pueden ser expresados como una distribución de mezclas. Este supuesto puede establecerse cuando los grupos no son homogéneos o se sospeche que no lo son. Entonces, un modelo de mezclas normal para la k-ésima clase tiene una densidad expresada por

$$P(\mathbf{x} \mid Y = k) = \sum_{r=1}^{n_k} \pi_{kr}\phi(\mathbf{x}; \mu_{kr}, \mathbf{\Sigma})$$

donde la notación $\phi(\mathbf{x}; \mathbf{m}, \mathbf{s})$ representa la densidad normal de la variable $\mathbf{x}$, de media $\mathbf{m}$ y matriz de varianzas y covarianzas $\mathbf{s}$, los valores π_{kr} suman 1, n_k es el tamaño de cada población y $k \in \{0, 1\}$, las dos poblaciones. Adicionalmente, se asume que cada uno de los grupos tiene la misma matriz de varianzas y covarianzas $\mathbf{\Sigma}$. Ahora, dado el modelo normal anterior para cada clase, las probabilidades de clase posterior están dadas por

$$P(Y = k \mid \mathbf{x} = x) = \frac{\sum_{r=1}^{n_k} \pi_{kr}\phi(\mathbf{x}; \mu_{kr}, \Sigma)p_k}{\sum_{l=1}^{K}\sum_{r=1}^{n_l} \pi_{lr}\phi(\mathbf{x}; \mu_{lr}, \Sigma)p_l}$$

donde p_k son las probabilidades previas de cada clase, tal que $p_0 + p_1 = 1$. Estas últimas pueden ser vistas como la proporción de elementos pertenecientes a cada clase.

Los parámetros de los modelos normales se encuentran maximizando el logaritmo de la función de verosimilitud conjunta sobre $P(Y, \mathbf{x})$ gracias al algoritmo `EM` (Dempster, Laird & Rubin 1977). Este método y su forma de cómputo puede consultarse en detalle en Hastie et al. (2009, ver sección 12.7).

3.6. Análisis discriminante flexible

Supóngase que se cuenta con datos de la forma $(y_j, \mathbf{x}_j) = (y_j, x_1, \ldots, x_p)$, $j = 1, \ldots, n$, donde y_j puede tomar valores en $\{0, 1\}$ y x_j, $j = 1, \ldots, n$, es un conjunto de variables métricas.

Entonces se define una función $\theta : \{0, 1\} \longmapsto \mathbb{R}$ que asigna puntajes reales a las categorías de la variable respuesta, de manera que las clases así transformadas sean óptimamente predichas por una regresión lineal cuyas variables predictoras son $x_1, \ldots, x_p$, con parámetros $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$.

De esta forma el problema se reduce a resolver

$$\text{mín}_{\boldsymbol{\beta},\theta} \sum_{j=1}^{n} \left[\theta(y_j) - x_j^t \boldsymbol{\beta}\right]^2$$

Es decir, hallar aquellos valores de $\boldsymbol{\beta}$ y θ para los que la predicción sea mejor en términos del error cuadrado medio. En general, suponiendo que se cuente con K categorías para la variable respuesta, se pueden escoger hasta $L \leq K - 1$ conjuntos de puntajes independientes para las etiquetas de las clases, $\theta_1, \ldots, \theta_L$, con L correspondiendo a una función lineal tal que $\eta_l(\mathbf{x}) = \mathbf{x}^t \boldsymbol{\beta}_l$, $l = 1, \ldots, L$ escogidos para que sean óptimos para la regresión lineal en $\mathbb{R}^p$. Los puntajes $\theta_l(\cdot)$ y los coeficientes $\boldsymbol{\beta}_l$ son escogidos de manera que minimicen el error cuadrado promedio

$$ECM = \frac{1}{n} \sum_{l=1}^{L} \left[\sum_{j=1}^{n} (\theta_l(y_j) - \mathbf{x}_j^t \boldsymbol{\beta}_l)^2 \right]$$

Este tema puede ser consultado en Hastie et al. (2009, sección 12.5).

3.7. Árboles de clasificación

Los árboles de clasificación son básicamente objetos gráficos. Se construyen particionando el espacio de posibles observaciones dentro de subregiones que se corresponden con las *hojas*. Cada observación será clasificada en una hoja del árbol.

Asimismo, la forma de construir el árbol se diferencia de otras técnicas computacionales (más que estadísticas) principalmente en la estrategia de *poda* y la

estrategia al dividir y formar nodos. Algunos métodos y algoritmos son reseñados en Ripley (1996).

En principio, se considera un *atributo* A, el cual puede dividirse con el objetivo de tomar una decisión y la atención se centra entonces en hallar aquel valor que divide el atributo de la mejor manera. Este razonamiento se aplica a los demás atributos del problema.

La estrategia de poda entra en juego debido a que, como en cualquier problema estadístico multivariante, hay variables que no contribuyen al análisis y se hace importante seleccionar aquellas variables que sean determinantes y no tener *ramas* inútiles. Seleccionar aquellas ramas que son de verdadera utilidad y deshacerse de las que no es lo que se conoce como poda. Este trabajo se logra por medio de algoritmos computacionales y existen variantes de ello. El lector interesado puede consultar Ripley (1996).

La poda puede realizarse de dos formas principalmente: de forma manual o de forma automática. La forma manual supone una revisión por parte del investigador de aquella cantidad de hojas que son importantes en la discriminación (esto se logra, en general, por validación cruzada) mientras que la forma automática supone entregarle al software la libertad de encontrar aquel número de ramas adecuado. Ambas estrategias se explican en detalle en Venables & Ripley (2002, pp. 251-256).

3.8. Método de los k-ésimos vecinos más cercanos

Este método consiste en asignar la categoría de un individuo, dependiendo cómo estén distribuidos sus vecinos según las variables que lo caracterizan. Así, supóngase que se desea estudiar si el individuo $\mathbf{x}_j$ pertenece a uno de dos grupos, $y_j = \{0, 1\}$ (el conjunto de datos puede representarse nuevamente como $(y_j, \mathbf{x}_j) = (y_j, x_1, \ldots, x_p)$, $j = 1, \ldots, n$). Entonces se define a $N_k(\mathbf{x}_j)$ como la vecindad a $\mathbf{x}_j$ (según alguna métrica) teniendo en cuenta un entorno que considere solo k vecinos: los k vecinos más cercanos a $\mathbf{x}_j$. Con esta información, se calcula $\widehat{Y}(\mathbf{x}_j)$, que se define como

$$\widehat{Y}(\mathbf{x}_j) = \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x}_j)} y_i$$

Como las observaciones $y_j \in \{0, 1\}$, entonces una nueva observación será asignada a un grupo u otro si $\widehat{Y}(\mathbf{x}_j)$ es mayor o menor que 0,5. La métrica utilizada al definir la vecindad más cercana es comúnmente la distancia euclídea (para otras métricas, revisar pp. 197-198 de Ripley, 1996).

Para la elección del número de vecinos, k, se acostumbra utilizar validación cruzada.

3.9. Método de Bayes ingenuo

El método de Bayes ingenuo asume que, dada una clase $Y = j$, con $j \in \{0, 1\}$, las variables x_k son independientes y, por tanto, $P(\mathbf{x} \mid Y = j) = P_j(\mathbf{x}) =$

$\prod_{k=1}^{p} P_{jk}(x_k)$. Además, asume distribuciones normales para las variables predictoras que son métricas. En nuestro caso, todas las variables predictoras son de este tipo. Luego, trabajando con el Teorema de Bayes, se tiene la regla siguiente

$$P(Y = j \mid \mathbf{x}) = \frac{P_j(\mathbf{x})\pi_j}{\sum_{l=1}^{K} P_l(\mathbf{x})\pi_l} = \frac{\prod_{k=1}^{p} P_{jk}(x_k)\pi_j}{\sum_{l=1}^{K}\prod_{k=1}^{p} P_{jk}(x_k)\pi_l}$$

donde π_j son las probabilidades previas de cada población. Este cálculo sencillo, en general, para grandes volúmenes de datos, reduce significativamente los cálculos.

Una introducción a esta técnica puede ser estudiada en Witten & Frank (2005, pp. 94-97).

3.10. Máquinas de soporte vectorial

Supóngase que se está en el caso en que ningún punto de los dos grupos se superpone; esto es, que cada punto de cada clase está, digamos, *separado* del otro grupo. A esto se le conoce como *caso separable*, y supóngase que se cuenta con datos del tipo $(\mathbf{x}_j, y_j)$, $j = 1, \ldots, n$, con $y_j \in \{-1, 1\}$. Se define un hiperplano por

$$\left\{\mathbf{x} : f(\mathbf{x}) = \mathbf{x}^t\boldsymbol{\beta} + \beta_0 = 0\right\}$$

donde $\|\boldsymbol{\beta}\| = 1$, y la regla de clasificación inducida por $f(\mathbf{x})$ es $G(x) = \text{signo}(\mathbf{x}^t\boldsymbol{\beta} + \beta_0)$. De manera que se busca una línea que divide a los dos grupos. Además, Hastie et al. (2009) muestran que $f(\mathbf{x})$ es la distancia con signo de un punto $\mathbf{x}$ al hiperplano $f(\mathbf{x}) = \mathbf{x}^t\boldsymbol{\beta} + \beta_0 = 0$. Como las categorías son separables es posible hallar una función $f(\mathbf{x}) = \mathbf{x}^t\boldsymbol{\beta} + \beta_0$ con $y_i f(x_i) > 0\ \forall i$ lo que implica que se pueden crear los 'márgenes' más grandes entre los puntos de las clases -1 y 1. Un problema de optimización equivalente es

$$\text{mín}_{\boldsymbol{\beta},\beta_0,\|\beta\|=1}\, M \text{ sujeto a } y_i(\mathbf{x}^t\boldsymbol{\beta} + \beta_0) \geq M, i = 1, \ldots, n$$

donde M es la distancia mínima que existe de cada grupo a la recta $f(\mathbf{x}) = \mathbf{x}^t\boldsymbol{\beta} + \beta_0$; así, la distancia que existe de un grupo a otro es de $2M$.

En el caso que los puntos en las clases no sean separables, una de las formas de abordar el problema, es aún maximizar M pero permitir que algunos puntos estén en el lado incorrecto de los márgenes. Esto se logra incluyendo las variables de holgura $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ en el problema de optimización anterior (Hastie et al. 2009, capítulo 12).

Las soluciones consideradas de esta manera incluyen de entrada las variables originales $(x_1, \ldots, x_p)$ con el objetivo de hallar 'márgenes lineales'. Sin embargo, la mayoría de las veces resulta útil modificar estas variables para obtener en este espacio clases que estén más separadas; esto se logra aplicando diversas transformaciones, digamos $h_l(\cdot)$, a las variables de entrada x_j.

Los métodos más populares para transformar variables son los métodos de expansión de bases, especialmente polinomios. Éste consiste en tomar una variable, digamos x_k y representarla como $x_k = \sum_{m=1}^{L} \beta_l h_l(x_k)$, donde $h_l(\cdot)$ son polinomios.

4. Resultados del entrenamiento y clasificación

4.1. Preliminares. Análisis de componentes principales

Para el caso de los textos listados en la sección 2.4, se cuenta con una matriz de datos de dimensión 393×45, cuyas columnas se corresponden con las 44 variables descritas en la sección 2, junto con una variable respuesta que indica si el texto pertenece a Cervantes o no. El número de filas se corresponde principalmente con el número de capítulos que existen en los libros utilizados en el entrenamiento (ver sección 2.4). Sin embargo, para evitar distorsiones en algunas medidas, se limitaron las dimensiones de los capítulos a no menos de 1000 y a no más de 10000 *tokens*. Esto debido a que algunos capítulos son muy cortos (por ejemplo los primeros capítulos del Pasamonte: el primero con apenas 143 *tokens*) o son muy largos, como las novelas ejemplares que no presentan divisiones (por ejemplo La Gitanilla que cuenta con 28006 *tokens*). Cada capítulo de dimensión inferior a 1000 *tokens* fue unido al capítulo que le precedía o sucedía dependiendo de la situación. Cada capítulo de más de 10000 *tokens*, fue dividido en varias partes de 4000 *tokens*, cada una aproximadamente. Nótese que en esta etapa no se esta trabajando aún con la *Novela de la tía fingida*.

Ahora, la forma de las variables estudiadas no presenta en general asimetrías fuertes y presentan patrones como los que se pueden apreciar en la figura1. Los histogramas representan la cantidad de veces que se repitió en el conjunto de datos cada proporción de cada palabra. Así, por ejemplo, se encuentran más de cien fragmentos donde la palabra `que` tuvo una proporción entre 0,045 y 0,05. Tampoco hubo ningún texto donde la proporción de esta palabra estuviera por encima de 0,7.

Asimismo, es importante notar que los datos están siendo considerados como una matriz rectangular sin tomar en cuenta el posible efecto serial que pueda haber entre capítulos o individuos consecutivos. Esto debido a que se está trabajando con varios autores y varios libros. En aquellos casos donde se trabaje con un solo autor con una cantidad suficientemente grande de capítulos es recomenable hacer un estudio previo de la posible correlación serial existente en la secuencia de estos y utilizar las técnicas apropiadas. Nótese que la correlación serial no tiene por qué ser semejante para un mismo autor en diferentes libros. En la figura 2 se muestra la riqueza verbal para las tres series de datos más largas dentro del conjunto de datos en estudio (I Quijote: *El ingenioso hidalgo don Quijote de la Mancha*; II Quijote: *El ingenioso caballero don Quijote de la Mancha* y Persiles: *Los trabajos de Persiles y Segismunda*). Aparentemente no existe un patrón fuerte presente, a excepción, quizá, de la primera parte de don Quijote.

Ahora, a esta matriz de datos se le aplicó la técnica de componentes principales con el objetivo de trabajar con menos cantidad de variables que con el conjunto original (sin embargo, muchas veces es difícil saber qué componente, si lo hay, es el que discrimina mejor. Ver especialmente sección 9.1 de Jolliffe (2002), y las referencias allí citadas).

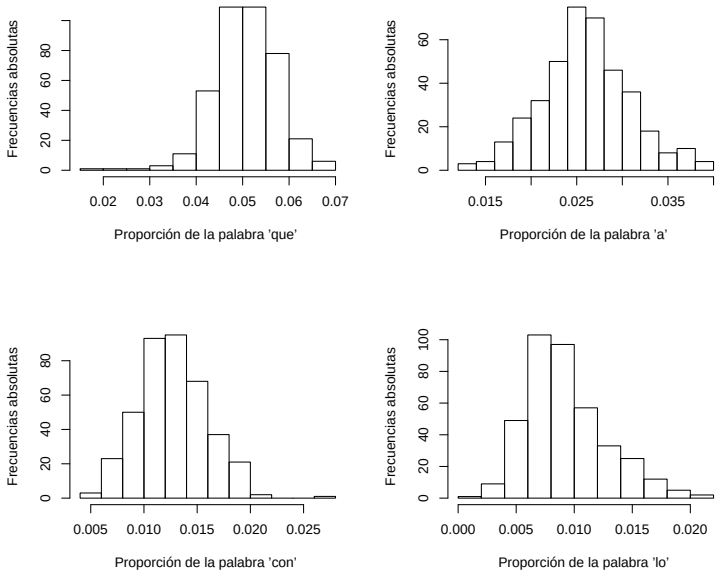

FIGURA 1: Histograma de la proporción de algunas palabras en los textos, a saber: `que, a, con, lo`.

La ejecución del método de componentes principales arroja los autovalores que se observan en la figura 3. Nótese que el criterio clásico de retener aquellos componentes mayores a la unidad se ve satisfecho con al menos 13 componentes. No obstante, en este trabajo se decidió trabajar con aquellos componentes que contribuyeron más en la discriminación según el método de regresión logística univariado de la forma que es sugerida por Hosmer & Lemeshow (2000, cap. 4 y 5). Esto es, se estimaron 44 modelos de la forma

$$g(\mathbf{x}) = \beta_0 + \beta_1 \text{COMPONENTE}_j, j = 1, \ldots, 44$$

donde $g(\mathbf{x})$ es definida como en la ecuación (2). De estos modelos se retuvieron aquellos componentes cuyo β_1 fuera significativo al 15 % puesto que cuando se utilizan valores p tradicionales (como 0,05) frecuentemente se falla al identificar variables que pueden ser importantes. Este procedimiento produjo la retención de 14 componentes. La cantidad de varianza explicada por cada uno de los componentes retenidos puede estudiarse en la tabla 1. Es interesante notar que existen componentes que explican muy poca cantidad de varianza ($<$2 %) pero sin embargo tienen alto poder de clasificación como el componente 38.

Una muestra de los dos primeros componentes principales puede apreciarse en las figuras 4 y 5. Se puede observar cómo estos dos componentes separan perfectamente la obra de Cervantes de Lope de Vega, Jerónimo de Pasamonte, Mateo Alemán y Francisco de Quevedo; mientras que, curiosamente, la separación no queda clara para Alonso Fernández de Avellaneda. Se puede observar asimismo

que el segundo componente en su lado negativo representa al relato autobiográfico por las variables `mi`, `yo` y `me` que allí se agrupan.

También se trabajó con las variables originales o un subconjunto de ellas. Este subconjunto se escogió utilizando el criterio de Akaike en un algoritmo por pasos aplicado a un modelo de regresión logística con todas las variables en estudio (Venables & Ripley 2002, p. 175). Esto es, se comenzó con el modelo inicial (con 44 variables predictoras)

$$g(\mathbf{x}) = \beta_0 + \beta_1 \text{que} + \beta_2 \text{de} + \cdots + \beta_{43} \text{que\_no} + \beta_{44} \text{de\_lo\_que}$$

donde $g(\mathbf{x})$ se define como en la ecuación (2) y utilizando las variables que fueron establecidas en la sección 2.2. En los pasos sucesivos de minimización del AIC se retuvieron en el modelo conjunto las variables siguientes: `que`, `y`, `a`, `el`, `no`, `se`, `por`, `lo`, `las`, `le`, `mi`, `yo`, `para`, `s.punto`, `s.guión` y `que_no`.

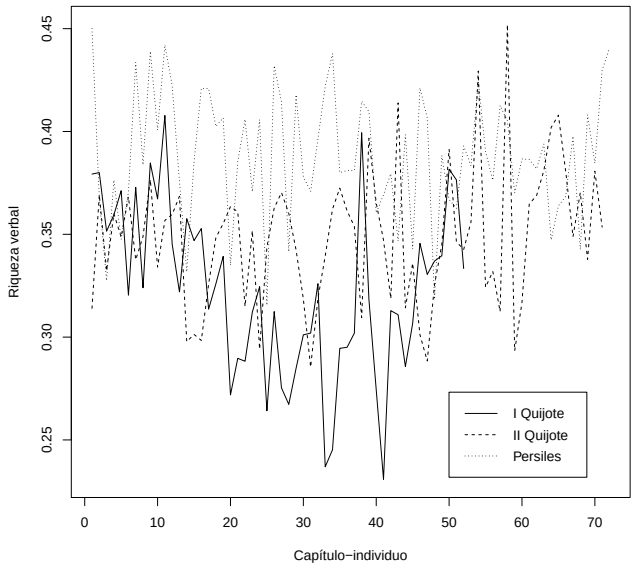

Figura 2: Tres libros más extensos vistos como una serie temporal.

Así, para la mayoría de las técnicas, se trabajó con dos conjuntos de datos principalmente: aquellos derivados de los componentes principales y aquellos que trabajan directamente sobre los datos originales.

4.2. Aplicación de las técnicas

Para el método de los vecinos más cercanos, se utilizaron cien repeticiones para cada número de vecinos para encontrar el número k más adecuado. El resultado puede observarse en la figura 6 y de allí se desprende que se haya escogido como

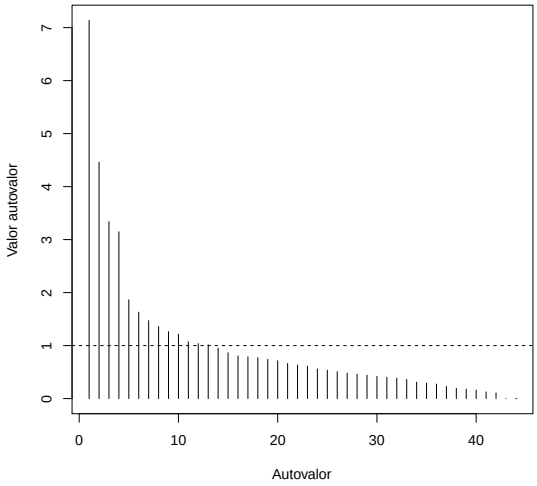

FIGURA 3: Autovalores: Método de componentes principales aplicado a las 44 variables en estudio

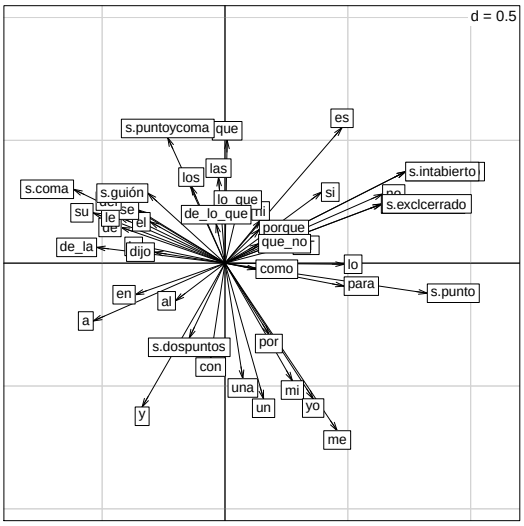

FIGURA 4: Gráfico de los dos primeros componentes principales (variables).

número óptimo 10 vecinos. Se puede notar que el promedio de desaciertos estuvo por debajo del 4 %, siendo un valor considerablemente bueno.

Para tener una idea de la efectividad del clasificador empleado, el conjunto de datos se dividió en dos partes seleccionadas completamente al azar: un 70 % de entrenamiento y 30 % como muestra de prueba. Con el conjunto de datos de entrenamiento se estimó el modelo correspondiente y se probó con el conjunto de

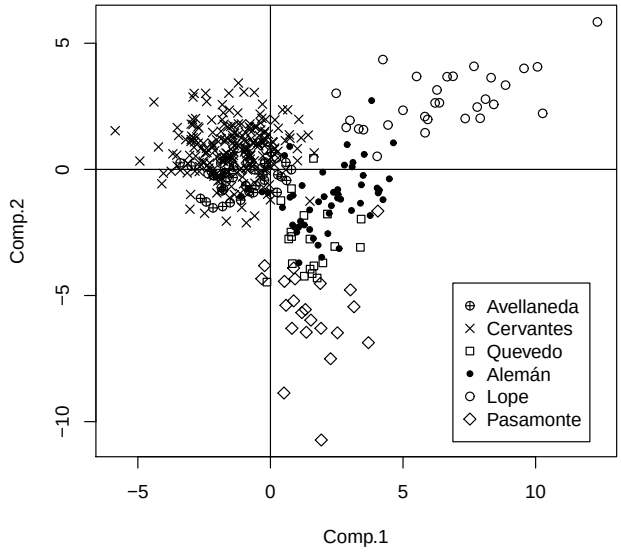

Figura 5: Gráfico de los dos primeros componentes principales (individuos).

Tabla 1: Cantidad de varianza explicada por los componentes retenidos.

No. de componente	1	2	3	4	5	6	7
% de var. explicada	16.22	10.14	7.59	7.16	4.24	3.71	3.35
% de var. acum.	16.22	26.36	33.95	41.10	45.34	49.05	52.40
No. de componente	8	9	12	13	20	31	38
% de var. explicada	3.09	2.87	2.36	2.31	1.63	0.92	0.44
% de var. acum.	55.49	58.36	60.71	63.02	64.65	65.57	66.01

datos para tal fin. El porcentaje de desaciertos fue guardado para dar una idea de lo efectivo que puede ser el método y este procedimiento se repitió 300 veces para cada clasificador (ver figura 7). En esta figura se han puesto, además, en color gris los métodos basados en los datos originales y en blanco los métodos que utilizaron como matriz de entrada algunos de los componentes principales.

Es importante notar que los cuatro primeros métodos con menor error de clasificación (máquinas de soporte vectorial, regresión logística, regresión logística generalizada y análisis discriminante de mezclas) se basan en datos originales y no en derivados del análisis de componentes principales. También es notable que más del 70 % de los métodos ostenten un error promedio de clasificación menor a 10 %, cuestión que pone de manifiesto la alta efectividad de la mayoría de ellos.

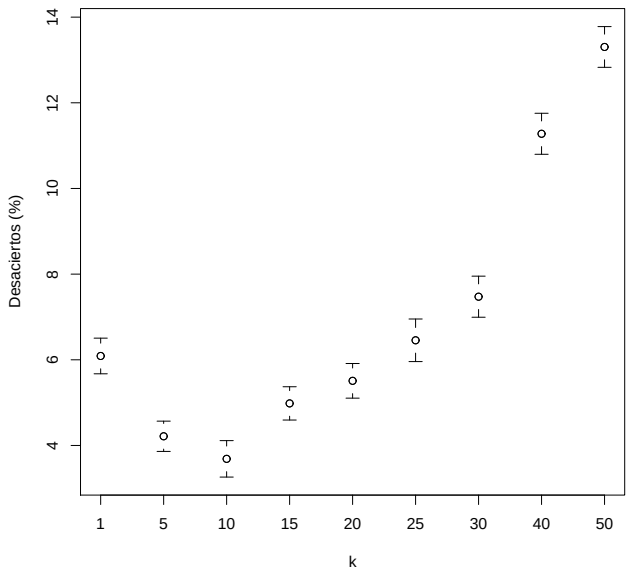

FIGURA 6: Escogencia del número de vecinos en el método de vecinos más cercanos. Nótese que el error de clasificación se hace menor con $k = 10$.

Por otro lado, los métodos basados en árboles de clasificación presentaron las tasas más altas de desaciertos para todas las variaciones ensayadas: todos en torno a 15 %, lo que sugiere la imposibilidad de esta técnica para afinar su calidad luego de cierto punto para este conjunto de datos (ver figura 8 y el detalle de un árbol en la figura 9).

4.3. El caso de la *Novela de la tía fingida*

El caso de la *Novela de la tía fingida* ha enigmado por mucho tiempo a los cervantistas. Hay posiciones encontradas con respecto a la autoría de esta novela. Hablando de los que apoyaban la tesis que Cervantes era su original autor, Andrés Bello, por ejemplo, decía (Aylward 1982, p. 27):

> ...se me acusará de temerario en poner este asunto otra vez en tela de juicio, mayormente después de lo que ha escrito, en modo incisivo i perentorio que acostumbra, don Bartolomé José Gallardo en el número 1ž de *El Criticón*. Pero, después de haber leído cuanto sobre esta materia me ha venido a las manos, que a la verdad no es mucho, no acabo de asegurarme...

y alega posteriormente razones de estilo y lenguaje entre las obras que sabe cervantinas y las que no.

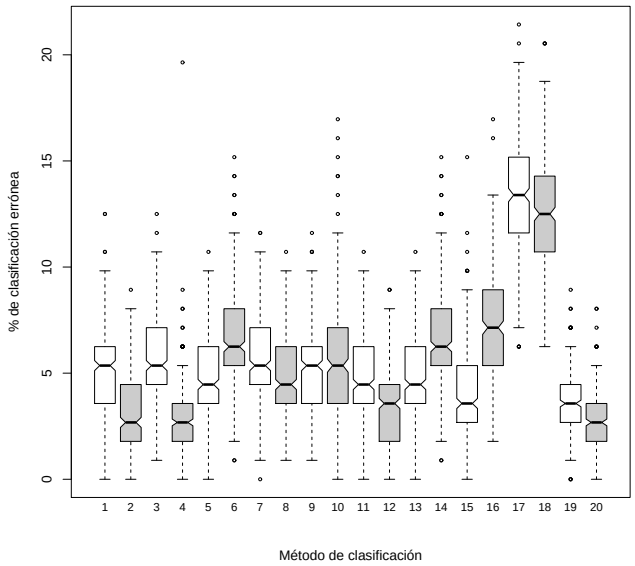

FIGURA 7: Métodos de clasificación: 1 y 2: Regresión logística. 3 y 4: Regresión logística aditiva generalizada. 5 y 6: Análisis discriminante lineal. 7 y 8: Análisis discriminante cuadrático. 9 y 10: Análisis discriminante regularizado. 11 y 12: Análisis discriminante mixto. 13 y 14: Análisis discriminante flexible. 15 y 16: Método de los vecinos más cercanos (10). 17 y 18: Método de Bayes ingenuo. 19 y 20: Máquinas de soporte vectorial (kernel radial). Se comparan los métodos utilizando los componentes principales y las variables originales.

Madrigal (2003), en un trabajo reciente, da razones para creer que el autor de esta novela es el propio Cervantes; y en su opinión, la atribución de una obra podría darse con frases como 'pasando por una calle' (que da inicio a la novela y que es usada sólo por Cervantes en algunas obras citadas por él). Posteriormente encuentra frases en novelas cervantinas y las compara con sus pares en la *Novela de la tía fingida* y concluye así que esta obra es de Cervantes.

Un resumen de la aplicación de las técnicas antes descritas se presenta en las tablas 2 y 3. En ellas se muestran dos columnas con probabilidades. La primera de ellas (con un signo ✠) muestra la probabilidad que la obra sea de Cervantes de acuerdo al método evaluado y curiosamente se obtiene que cerca del 70 % de ellos asignan la novela al grupo de Cervantes, no sin ciertas peculiaridades dignas de mención.

Se observa, por ejemplo, que todos los métodos que utilizan los componentes principales clasifican la novela como cervantina (ver tabla 2). Esto levanta la sospecha que los métodos que utilizan los componentes principales tienden a asignar indiscriminadamente la obra a Cervantes sin ser un resultado justo. La comprobación de esta sospecha es difícil de discutir y de desentrañar puesto que no es sencillo

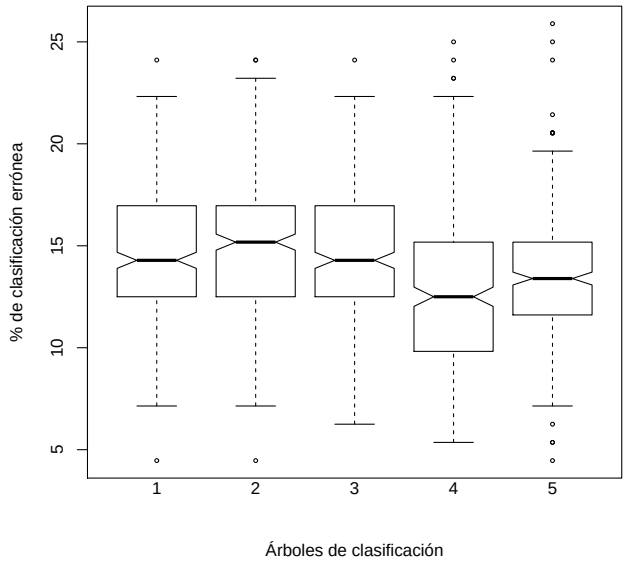

FIGURA 8: Árboles de clasificación: 1: Resultado con poda automática con un subconjunto de los componentes principales. 2: Resultado con poda automática con todos los componentes principales. 3: Resultado con poda por validación cruzada con un subconjunto de los componentes principales. 4: Resultado con poda por validación cruzada con los datos originales. 5: Resultado con poda automática con los datos originales.

saber qué componente pudiera estar causando este efecto (si el efecto, ciertamente, existe). Es probable también que la configuración de los componentes escogidos, conjuntamente, produzca una especie de enmascaramiento. Es notable, además, que las probabilidades que la obra sea una creación cervantina obtenidas con las variables originales, en general, son bastante bajas. La controversia pudiera aún continuar al revisar los resultados de los árboles de clasificación (ver tabla 3) donde la probabilidad más baja (0,82) se corresponde con el uso de un subconjunto de los componentes principales.

La columna de probabilidades de las tablas 2 y 3, cuyo signo es ¶, muestra los resultados de aplicar las extensiones naturales de los métodos expuestos en la sección 3 al caso de varias categorías (donde cada categoría es uno de los autores listados en la sección 2.4).

En el resultado de esta aplicación se confirma la curiosa sospecha de la semejanza entre Alonso Fernández de Avellaneda y Miguel de Cervantes, puesto que únicamente hacen aparición estos dos autores[1].

[1]Los resultados para el análisis discriminante cuadrático y el modelo de regresión multinomial aditivo producen errores de estimación y convergencia. Por estas razones esos resultados particulares no muestran.

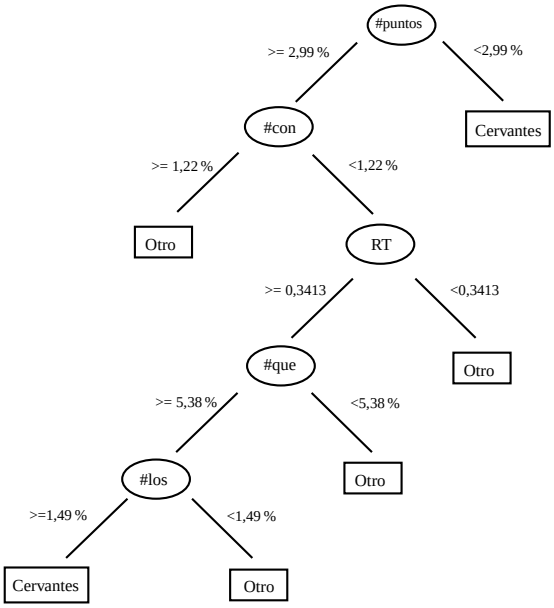

FIGURA 9: Árbol de clasificación: se muestra el árbol que utiliza las variables originales bajo el esquema de poda manual (se corresponde con el método 4 de la figura 8).

Otro dato curioso es el resultado obtenido con el método de las máquinas de soporte vectorial cuando se utilizan todas las variables originales: en el modelo dicotómico se asigna la obra al grupo donde no está Cervantes y al aumentar las clases asigna a la obra como cervantina.

5. Conclusiones

Los métodos de clasificación en general mostraron en el entrenamiento un muy buen desempeño estando la mayoría por debajo del 10 % de desaciertos y aún por debajo del 5 %. El método de Bayes ingenuo presentó una tasa de desacierto superior al 10 % (Yu (2008) obtuvo un resultado donde el método de Bayes ingenuo fue tan competitivo como el método de las máquinas de soporte vectorial).

En ciertos casos, trabajar con algunos componentes principales produjo menor error de clasificación (tal es el caso de las máquinas de soporte vectorial o el análisis discriminante cuadrático); en otros, trabajar con las variables originales produjo mejores resultados (como el método de los vecinos más cercanos o el análisis discriminante lineal) mientras que en otros, como en el análisis discriminante regularizado o el método de Bayes ingenuo, el trabajar con un conjunto de datos u otro no representó mayor diferencia. Esta situación es más palpable cuando se utiliza el método de árboles de clasificación (ver figura 8), donde los resultados fueron prácticamente los mismos para las diferentes estrategias utilizadas. Note

Tabla 2: Resultados de aplicar diferentes métodos de clasificación a la *Novela de la tía fingida*, históricamente atribuida a Cervantes.

Método	Prob.✠	Prob.¶	Método	Prob.✠	Prob.¶
Reg. logística†	0,78	$1{,}00^{C}$	An. Disc. cuadrático†	0,94	$1{,}00^{C}$
Reg. logística‡	0,00	$1{,}00^{A}$	An. Disc. cuadrático‡	0,00	-
Reg. Log. aditiva†	1,00	-	An. Disc. regularizado†	0,80	$0{,}99^{C}$
Reg. Log. aditiva‡	0,00	-	An. Disc. regularizado‡	0,00	$0{,}60^{A}$
An. Disc. lineal†	0,80	$0{,}99^{C}$	10 vecinos más cercanos†*	1,00	$1{,}00^{C}$
An. Disc. lineal‡	0,33	$0{,}60^{A}$	10 vecinos más cercanos◇*	1,00	$1{,}00^{C}$
An. Disc. de mezclas†	0,74	$0{,}99^{C}$	Bayes ingenuo†	0,96	$0{,}90^{C}$
An. Disc. de mezclas‡	0,30	$0{,}86^{A}$	Bayes ingenuo◇	1,00	$1{,}00^{C}$
An. Disc. flexible†	0,80	$0{,}99^{C}$	Máq. de soporte vectorial†	0,99	$0{,}99^{C}$
An. Disc. flexible‡	0,33	$0{,}61^{A}$	Máq. de soporte vectorial◇	0,12	$0{,}71^{C}$

✠: Probabilidades del modelo binario (Cervantes vs Otro).
¶: Probabilidades del modelo que considera cada autor como un grupo.
†: Considerando el subconjunto de los componentes principales establecido en la tabla 1.
‡: Considerando el subconjunto de las variables originales reseñado en la sección 4.1.
◇: Considerando todas las variables originales.
*: Este valor no es una probabilidad. Este método asigna como cervantina a esta novela.
^{A}: Significa que el modelo de múltiple respuesta asigna la obra a Alonso Fernández de Avellaneda.
^{C}: Significa que el modelo de múltiple respuesta asigna la obra a Miguel de Cervantes.

Tabla 3: Resultados de aplicar árboles de clasificación a la *Novela de la tía fingida*, históricamente atribuida a Cervantes.

Método	Prob.✠	Prob.¶
Árboles de clasificación†	0,97	0,92
Árboles de clasificación‡	0,97	0,92
Árboles de clasificación◇	0,82	0,62
Árboles de clasificación♣	0,97	0,97
Árboles de clasificación♠	0,97	0,94

✠: Probabilidades del modelo binario (Cervantes vs Otro). ¶: Probabilidades del modelo que considera cada autor como un grupo (todos asignaron la obra como Cervantina). †: Considerando poda automática con un subconjunto de los componentes principales. ‡: Considerando poda automática con todos los componentes principales. ◇: Considerando poda por validación cruzada con un subconjunto de los componentes principales. ♣: Considerando poda por validación cruzada con los datos originales. ♠: Considerando poda automática con los datos originales.

el lector que se está haciendo referencia a los resultados generales de clasificación (ver figura 7) y no al resultado particular motivo del estudio.

A este respecto resulta, a la luz de los resultados encontrados en este trabajo, extremadamente difícil dar una conclusión concreta. Por ejemplo, si se decidiera considerar solo aquellos clasificadores cuyo error promedio es menor al 5 % (10 en total) se encontraría que 5 de ellos asignan la obra a Cervantes y los restantes 5

a otro autor. Si se decidiera utilizar los clasificadores con un error promedio de menos del 10 %, se encontraría que 10 de ellos clasifican la obra como cervantina y 8 no. Si se consideran todos los métodos, el 68 % de ellos asigna la obra como cervantina.

Por lo pronto, parte de las ampliaciones y mejoras directas que tiene este trabajo comienzan por la extensión del corpus utilizado para autores diferentes a Cervantes, ampliar el panorama de autores considerados, utilizar otras técnicas de clasificación (por ejemplo, Tibshirani, Hastie, Narashimhan & Chu 2003), variar los métodos de calibración empleados (cambiar, por ejemplo, las probabilidades previas a cada grupo; cambiar el punto de corte para los modelos binarios, etc.).

Por último, merece un comentario la sorpresa de la aparente semejanza estilográfica entre Miguel de Cervantes y Alonso Fernández de Avellaneda. Es probable que 'descubrir' que el autor de *La novela de la tía fingida* es estadísticamente muy parecido a Cervantes no dé muchas luces al problema puesto que se ignora qué escritor, *encubriendo su nombre, fingiendo su patria*, estuvo detrás de este pseudónimo.

[Recibido: abril de 2010 — Aceptado: enero de 2011]

Referencias

Aylward, E. T. (1982), *Cervantes: Pioneer and Plagiarist*, Tamesis Books Limited, Londres, UK.

Baum, L. F. (2001), *The Royal Book of Oz*, Dover Publications, New York, States United. Escrito con 'colaboración' de R. Thompson.

Binongo, J. (2003), 'Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution', *Chance* **16**(2), 9–17.

Bird, S., Klein, E. & Loper, E. (2009), *Natural Language Processing with Python*, O'Really, Sebastopol, States United.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Pattern Recognition* **39**, 1–38.

Gardner, M. (1998), *Visitors from Oz: The Wild Adventures of Dorothy, the Scarecrow, and the Tin Woodman*, St Martins Press, New York, States United.

Grieve, J. (2007), 'Quantitative Authorship Attribution: An Evaluation of Techniques', *Literacy and Linguistic Computing* **22**(3), 251–270.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2 edn, Springer, New York, States United.

Hoover, D. L. (2002), 'Multivariate Analysis and Study of Style Variation', *Literacy and Linguistic Computing* **18**(4), 341–360.

Hosmer, D. & Lemeshow, S. (2000), *Applied Logistic Regression*, 2 edn, Wiley, New York, States United.

Jockers, M., Witten, D. & Criddle, C. (2008), 'Reassessing authorship of the Book of Mormon using delta and nearest shrunken centroid classification', *Literacy and Linguistic Computing* **23**(4), 465–491.

Johnson, R. & Wichern, D. (1998), *Applied Multivariate Statistical Analysis*, fourth edn, Prentice Hall, New York, States United.

Jolliffe, I. T. (2002), *Principal Component Analysis*, 2 edn, Springer, New York, States United.

Joula, P. (2006), 'Authorship Attribution', *Foundations and Trends in Information Retrieval* **1**(3), 233–334.

Koppel, M., Schler, J. & Argamon, S. (2009), 'Computational methods in authorship attribution', *Journal of the American Society for Information Science and Technology* **60**(1), 9–26.

Lebart, L., Morineau, A. & Warwick, K. (1984), *Multivariate Descriptive Statistical Analysis*, John Wiley & Sons, New York, States United.

Madrigal, J. L. (2003), 'De cómo y por qué La tía fingida es de Cervantes', *Artifara* (2).

Rencher, A. (2002), *Methods of Multivariate Analysis*, second edn, Wiley, New York, States United.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, UK.

Tibshirani, R., Hastie, T., Narashimhan, B. & Chu, G. (2003), 'Class prediction by nearest shrunken centroids with applications to DNA microarrays', *Statistical Science* **18**(1), 104–117.

Venables, W. N. & Ripley, B. D. (2002), *Modern Applied Statistics with S*, fourth edn, Springer, New York, States United.
*http://www.stats.ox.ac.uk/pub/MASS4

Witten, I. H. & Frank, E. (2005), *Data Mining: Practical Machine Learning Tools and Techniques*, 2 edn, Elsevier, San Francisco, States United.

Yu, B. (2008), 'An evaluation of text classification methods for literacy studies', *Literacy and Linguistic Computing* **23**(3), 327–343.