Clasificación de Prescripciones Médicas en Español

Juan Manuel Rodríguez[1], Enrique Calot[2], Hernán D. Merlino[1,3]

[1] Cátedra de Sistemas de Soporte para Celdas de Producción Flexible.
Departamento Computación Facultad de Ingeniería, Universidad de Buenos Aires.
[2] Laboratorio de Sistemas de Información Avanzados.
Departamento Computación. Facultad de Ingeniería, Universidad de Buenos Aires.
[3] Laboratorio de Investigación y Desarrollo en Arquitecturas Complejas (LIDAC)
Grupo de Investigación en Sistemas de Información (GISI). Universidad Nacional de Lanús.
{jmrodriguez,ecalot,hmerlino}@lsia.fi.uba.ar

Abstract. El siguiente trabajo describe la problemática de la clasificación de textos médicos libres en español. Y propone una solución basada en los algoritmos de clasificación de texto: Naïve Bayes Multinomial (NBM) y Support Vector Machines (SVMs) justificando dichas decisiones y mostrando los resultados obtenidos con ambos métodos.

Keywords: investigación biomédica, clasificación de textos, PLN, Naïve Bayes Multinomial, NBM, Support Vector Machines, SVM

1 Introducción

La investigación biomédica se enfrenta con grandes conjuntos de datos con enormes cantidades de texto libre, es decir texto en forma no estructurada y sin explotar, dificultando las diversas tareas de análisis de dicha información [1]. El análisis de estos conjuntos de datos plantea desafíos únicos por su complejidad, características y relevancia de los mismos. Llevarlos a una forma estructurada permitiría un mejor seguimiento del paciente y/o el hallazgo de información subyacente [2].

Ciertos trabajos demuestran los beneficios potenciales de la estructuración de la información médica, ya sea en la atención, investigación, enseñanza o para el mejor uso de las historias clínicas [3], [4], [5].

En el presente trabajo se evaluaron prescripciones médicas diarias realizadas sobre pacientes internados. Los documentos son textos escritos en un sistema informático, realizados en forma rápida, breve, concisa y con palabras propias del dominio médico. Algunas transcripciones textuales de este tipo de texto se detallan en la tabla 1.

Table 1. Ejemplo de textos medicos a clasificar

Textos médicos
csv y diuresis por turnos
Glibenclamida 5 mg pre edsayuno
sertralina 100 mg vo c / dia

Nótese que la palabra desayuno se encuentra escrita de forma errónea como: "edsayuno". Este tipo de errores constituye una de las dificultades adicionales a sortear para lograr el objetivo principal de este trabajo el cual consiste en la clasificación univoca de un texto dado, de determinadas características, en alguna de las categorías que se detallan en la tabla 2.

Table 2. Categorias en las que se clasificarán los textos médicos descriptos.

Categorías
Dieta
Endovenosa continua
Endovenosa no continua
No endovenosa
No farmacológica

El resultado esperado es un clasificador automático capaz de identificar lo que un médico prescribió a un paciente: si fue una dieta o si fue un fármaco y en este caso si fue un fármaco de administración endovenosa o no. En este último caso deberá identificar si es de administración continua o no.

2 Justificación de la solución escogida

Uno de los algoritmos escogidos para la clasificación de textos fue *Multinomial Naïve Bayes* (MNB), este es uno de los modelos más simples y parte del supuesto de que todos los atributos de los ejemplos (unigramas) son independientes entre sí en el contexto de una clase, esto es llamado "el supuesto de *Naïve Bayes*" [6]. A pesar de que este supuesto es en verdad falso, en la mayoría de las tareas del mundo real *Naïve Bayes* realiza buenas clasificaciones [6].

Una de las razones por las cuales fue seleccionado *Naïve Bayes* como algoritmo de clasificación, es por el trabajo de Banko y Brill [7], en donde se utilizaron cuatro clasificadores distintos: (a) *Winnow* [8], (b) *Perceptron* [9] y (c) *Naïve Bayes*, se observó que estos tienden a converger para grandes volúmenes de datos. Otro de los motivos es que *Naïve Bayes* es un algoritmo rápido, incluso con grandes cantidades de datos [10].

El trabajo se realizó tomando unigramas como características, de aquí en más *features*, del documento a clasificar. La justificación de esta decisión se encuentra en el trabajo de Pang y Lee [11] quienes realizaron una comparación entre distintos *features*: (a) unigramas, (b) bigramas, (c) unigramas y *Parts of speech tags* (POS), (d)

adjetivos, (e) unigramas más usados y (f) unigramas junto a la posición de la palabra en el texto. El resultado se resume en la tabla 3 obtenida de [11].

Table 3. Precisión expresada en porcentaje de los clasificadores: *Naive Bayes* (NB) y *Support Vector Machines* (SVM) utilizando diversos *features:*

Features	Precisión NB	Precisión SVM
unigrama	81.0	**82.9**
Unigrama y bigrama	80.6	82.7
bigrama	77.3	77.1
unigrama+POS *tags*	81.5	81.9
adjetivos	77.0	75.1
2633 unigramas más frecuentes	80.3	81.4
Unigrama y posición en el texto	81.0	81.6

Los resultados obtenidos, como se ve en la tabla son similares, sin embargo la precisión más alta se logró utilizando unigramas como *features* con el clasificador SVM. En el caso de *Naive Bayes* el resultado fue ligeramente mejor cuando se añadieron los *part of speech tags* pero solo en un 0.5 porciento.

Se decidió no utilizar una lista de *stopwords* para filtrar los textos médicos debido a que no son textos ordinarios sino que están muy reducidos y cuentan con muchas abreviaciones. La mayoría están escritos utilizando palabras y abreviaturas propias del dominio medico; una lista de las palabras con alta frecuencia en estos textos podría no coincidir con una lista usual y probada de *stopwords*, por ejemplo, en vez de utilizar la palabra "cada" aquí es más común encontrar: "c/". Por otro lado si bien las listas de *stopwords* realizan un trabajo valioso en sistemas de recuperación de información ya que en búsquedas genéricas no aportan demasiada información adicional, las *stopwords* en sí contribuyen sustancialmente al sentido final de una frase [12].
Finalmente está la salvedad del idioma, los trabajos antes citados fueron todos realizados con textos en inglés. Sin embargo en el trabajo de Tolosa, Peri y Bordignon [13] se demuestra que *Naïve Bayes* es un excelente clasificador de textos tanto en inglés como en español.

El segundo método escogido para realizar la clasificación de textos y poder comparar resultados fue una implementación de *support vector machines* (SVMs) [14] llamada *Sequential Minimal Optimization* (SMO): la cual consiste básicamente en una mejora del algoritmo de entrenamiento de SVMs, que logra que este llegue a ser a ser 1200 veces más rápido para SVMs lineales y 15 veces más rápido para SVMs no lineales [15]. Esta elección se debe a que las SVMs como método de clasificación son muy populares y ampliamente utilizadas [16] debido a su éxito, no solo para clasificar textos sino también para diversos y complejos problemas de clasificación como lo son la detección de tumores [17], la expresión génica [18] o el modelado de células mitóticas [19]. Sang-Bum Kim en [20] menciona que los clasificadores basados en complejos métodos de aprendizaje como los SVMs pertenecen al *state-of-the-art*.

Table 4. Resumen de los algoritmos a utilizar con sus configuraciones y características*:*

algoritmos	*features*	*stopwords*	idioma	características
M Naïve Bayes	unigrama	No	español	texto médico
SMO	unigrama	No	español	texto médico

3 Desarrollo de la solución

Para el desarrollo del trabajo se utilizó un conjunto de entrenamiento de 700 textos por clase (hay 5 clases en total) y 300 textos por clase para el conjunto de prueba. Es decir un total de 5000 textos, de los cuales el 70% se utilizó para entrenar a los algoritmos de clasificación y el 30% restante para validarlos. Como herramienta de trabajo se utilizó Weka [23] en su versión 3.6.11.

Para la preparación de los textos, se utilizó la siguiente configuración (provista por la herramienta):

1. Se identificaron los distintos unigramas dividiendo al texto por los caracteres: tabulador, espacio, punto, coma, punto y coma, dos puntos, comilla simple, comilla doble, paréntesis de cierre y apertura, signo de exclamación y signo de interrogación (\t\s.,;:'"()?!)
2. Se contaron las apariciones de cada unigrama para cada uno de los textos.

3.1 Multinomial Naïve Bayes

Este algoritmo de clasificación busca identificar a la clase que maximice el resultado de la multiplicación entre la probabilidad de una clase dada y las probabilidades individuales de las palabras dada dicha clase, matemáticamente:

$$C_{map} = \operatorname{argmax} P(c_i) \prod_{x \in X} P(w|c_i) \quad (1)$$

Formula 1: Ecuación detrás del algoritmo de clasificación *Naïve Bayes.*

Las probabilidades de las distintas clases (c) junto con las probabilidades de cada una de las palabras (w) de pertenecer a una clase son estimadas en el conjunto de entrenamiento, el cual está constituido por el 70% del total de casos como se menciona en el apartado anterior.

Luego de ejecutar este clasificador sobre un total de 1500 casos se obtuvieron los resultados que se muestran en la tabla a continuación:

Table 4. Resultados *Multinomial Naïve Bayes*

Mediciones	Valor	
Instancias clasificadas correctamente	1284	85.6%
Instancias clasificadas incorrectamente	216	14.4 %
Error absoluto		0.0664
Raíz cuadrada del error absoluto		0.2309
Error relativo		20.7318 %
Raíz cuadrada del error relativo		57.7052 %
Número total de instancias		1500

Se muestra a continuación, la matriz de confusión, la cual es una tabla que indica, clase por clase, las diferencias entre los casos clasificados correctamente (casos positivos) y los casos clasificados de forma erronea (casos negativos), para un conjunto de ejemplos etiquetados [21].

Table 5. Matriz de confusión

A	B	C	D	E	Clases
166	32	27	25	36	A = Dieta
0	272	0	16	9	B = Endovenosa no continua
8	1	303	4	5	C = No farmacológica
4	11	1	288	1	D = Endovenosa continua
4	17	12	3	255	E = No endovenosa

A su vez, con los resultados anteriores se calculó para cada clase, la precisión: fracción de documentos asignados a la clase i que son realmente de la clase i, la exactitud: fracción de documentos en la clase i clasificados correctamente, la medida F (F-Measure) para $\beta=1$ que se define con la siguiente fórmula:

$$\frac{2PE}{P+E} \tag{2}$$

Formula 2: Calculo de la medida de rendimiento *F-Measure*, asumiendo que $\beta=1$

En donde P es la precisión y E la exactitud. El área bajo la curva *Receiver Operating Characteristic* (ROC) que es una de las mejores formas de evaluar el desempeño de un clasificador sobre un conjunto de datos cuando no hay un punto operacional determinado [21]. Por último se calculó el promedio ponderado de cada una de estas medidas para la totalidad de las clases.

Table 6. Detalle de la precisión por clase

Precisión	Exactitud	*F-Measure*	*ROC Area*	Clases
0.912	0.58	0.709	0.823	Dieta
0.817	0.916	0.863	0.959	Endovenosa no continua
0.883	0.944	0.913	0.978	No farmacológica
0.857	0.944	0.899	0.97	Endovenosa continua
0.833	0.876	0.854	0.966	No endovenosa
0.861	0.856	0.85	0.941	Promedio ponderado

3.2 Sequential Minimal Optimization

Sequential Minimal Optimization (SMO) es un algoritmo simple que puede resolver rápidamente los grandes problemas de programación cuadrática (QP) de SVM sin utilizar una matriz de almacenamiento y sin utilizar pasos QP de optimización numérica. SMO descompone al problema QP general en sub-problemas QP, utilizando el teorema de Osuna para asegurar la convergencia [15].

Luego de ejecutar este clasificador sobre los mismos 1500 casos utilizados con MNB se obtuvieron los resultados que se muestran en la siguiente tabla:

Table 7. Resultados SMO

Mediciones	Valor	
Instancias clasificadas correctamente	1278	85.2 %
Instancias clasificadas incorrectamente	222	14.8 %
Error absoluto		0.2487
Raíz cuadrada del error absoluto		0.3295
Error relativo		77.6944 %
Raíz cuadrada del error relativo		82.3543 %
Número total de instancias		1500

Se muestra a continuación la matriz de confusión, para los casos clasificados con el algoritmo SMO:

Table 8. Matriz de confusión

A	B	C	D	E	Clases
174	26	26	20	40	A = Dieta
8	271	0	5	13	B = Endovenosa no continua
16	0	298	2	5	C = No farmacológica
14	8	1	280	2	D = Endovenosa continua
19	6	10	1	255	E = No endovenosa

Finalmente se calcularon los mismos estimadores utilizados para medir el desempeño del clasificador MNB: Precisión, Exactitud, *F-Measure*, *ROC Area*. Se muestran los resultados a continuación separados por clase y el promedio ponderado:

Table 9. Detalle de la precisión por clase

Precisión	*Exactitud*	*F-Measure*	*ROC Area*	Clases
0.753	0.608	0.673	0.776	Dieta
0.871	0.912	0.891	0.96	Endovenosa no continua
0.89	0.928	0.909	0.973	No farmacológica
0.909	0.918	0.914	0.971	Endovenosa continua
0.81	0.876	0.842	0.949	No endovenosa
0.848	0.852	0.848	0.928	Promedio ponderado

4 Conclusiones

Como puede observarse al comparar ambas tablas, los resultados son similares, sin embargo MNB es ligeramente mejor ya que su precisión, su exactitud, su *F-Measure* y el área bajo la curva ROC están más cerca de 1 que los mismos valores para el clasificador SMO. Se muestra un diagrama comparativo donde se observa dicha tendencia gráficamente:

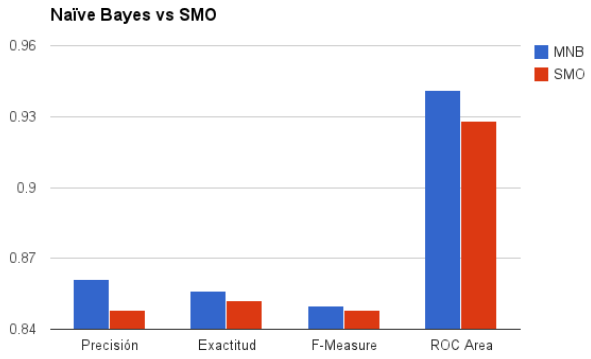

Fig. 1. Tabla comparativa entre los promedios de las medidas de rendimiento utilizadas para evaluar ambos clasificadores.

MNB fue mucho más rápido al momento de construir el modelo que el clasificador SMO, demoró 0.08 segundos contra 13.51 segundos que tardó este último.

Si bien en términos generales una precisión de 0.85 es considerada buena, no es suficiente al tratarse de textos médicos. Sin embargo es un excelente punto de partida que demuestra la viabilidad de la clasificación de textos médicos en español.

5 Futuras líneas de investigación

Los pasos a seguir consistirán en aumentar la precisión y la exactitud hasta llegar un valor cercano 1en ambos casos, de esta forma el algoritmo podría suplir a una persona en la tarea de clasificación. En ese sentido serán tenidos en cuenta trabajos como el de Sang-Bum Kim [20] para mejorar la precisión de MNB o el de Puurula que propone una extensión al modelo MNB llamada *Tied Document Mixture* (TDM) el cual logró, para algunos casos puntuales, una reducción del error promedio de entre un 26% y un 65% [22].

Finalmente se considerarán técnicas de extracción de relaciones semánticas sobre los textos para lograr un detalle más fino de la información a extraer.

Referencias

1. Tolosa, Gabriel Hernán; Peri, Jorge Alberto; Bordignon, Fernando. Experimentos con Métodos de Extracción de la Idea Principal de un Texto sobre una Colección de Noticias Periodísticas en Español. En XI Congreso Argentino de Ciencias de la Computación. 2005.
2. Botsis, Taxiarchis, et al. Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. Journal of the American Medical Informatics Association, 2011, vol. 18, no 5, p. 631-638.
3. González bernaldo de quirós, Fernán; Plazzotta, Fernando; Campos, Fernando; Kaminker, Diego; Martínez, María Florencia, lópez osornio, Alejandro; Seehaus, Alberto; garcía mónaco, Ricardo; Luna, Daniel. Creación de un Sistema de Reportes Estructurados, Codificados y Estándares. AAIM, 2008
4. Thomas, Anil A., et al. Extracting data from electronic medical records: validation of a natural language processing program to assess prostate biopsy results. World journal of urology, 2014, vol. 32, no 1, p. 99-103.
5. Strauss, Justin A., et al. Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm. Journal of the American Medical Informatics Association, 2012.
6. Mccallumzy, Andrew; Nigamy, Kamal. A comparison of event models for Naïve bayes text classification. En AAAI-98 workshop on learning for text categorization. 1998. p. 41-48.
7. Banko, Michele; Brill, Eric. Scaling to very very large corpora for natural language disambiguation. En Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2001. p. 26-33.
8. Golding, Andrew R.; Roth, Dan. A winnow-based approach to context-sensitive spelling correction. Machine learning, 1999, vol. 34, no 1-3, p. 107-130.
9. NG, Hwee Tou; GOH, Wei Boon; LOW, Kok Leong. Feature selection, perceptron learning, and a usability case study for text categorization. En ACM SIGIR Forum. ACM, 1997. p. 67-73.
10. Lowd, Daniel; Domingos, Pedro. Naïve Bayes models for probability estimation. En Proceedings of the 22nd international conference on Machine learning. ACM, 2005. p. 529-536.
11. Pang, Bo; LEE, Lillian; Vaithyanathan, Shivakumar. Thumbs up?: sentiment classification using machine learning techniques. En Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002. p. 79-86.
12. Riloff, Ellen. Little words can make a big difference for text classification. En Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1995. p. 130-136.ç
13. Tolosa, Gabriel Hernán; Peri, Jorge Alberto; Bordignon, Fernando. Experimentos con Métodos de Extracción de la Idea Principal de un Texto sobre una Colección de Noticias Periodísticas en Español. En XI Congreso Argentino de Ciencias de la Computación. 2005.
14. Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features (pp. 137-142). Springer Berlin Heidelberg.
15. Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.
16. Kolesov, Anton, et al. On Multilabel Classification Methods of Incompletely Labeled Biomedical Text Data. Computational and Mathematical Methods in Medicine, 2014, vol. 2014
17. S. M. Hosseini, M. J. Abdi, and M. Rezghi, “A novel weighted support vector machine based on particle swarm optimization for gene selection and tumor classification”,

Computational and Mathematical Methods in Medicine, vol. 2012, Article ID 320698, 7 pages, 2012.
18. S. C. Li, J. Liu, and X. Luo, "Iterative reweighted noninteger norm regularizing svm for gene expression data classification," Computational and Mathematical Methods in Medicine, vol. 2013, Article ID 768404, 10 pages, 2013.
19. Z. Gao, Y. Su,A. Liu,T.Hao, and Z. Yang, "Non negative mixed norm convex optimization for mitotic cell detection in phase contrast microscopy," Computational and Mathematical Methods in Medicine, vol. 2013, Article ID 176272, 10 pages, 2013.
20. Kim, Sang-Bum, et al. Some effective techniques for naive bayes text classification. Knowledge and Data Engineering, IEEE Transactions on, 2006, vol. 18, no 11, p. 1457-1466.
21. Bradley, Andrew P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition, 1997, vol. 30, no 7, p. 1145-1159.
22. Puurula, A., & Myaeng, S. H. (2013, December). Integrated instance-and class-based generative modeling for text classification. In Proceedings of the 18th Australasian Document Computing Symposium (pp. 66-73). ACM.
23. http://www.cs.waikato.ac.nz/ml/weka/, página vigente al 21 de junio de 2014.