Campos de Markov para Mejorar la Clasificación de Textos

Reporte de Proyecto Final

Adriana Gabriela Ramírez de la Rosa
gabrielarr@inaoep.mx

Instituto Nacional de Astrofísica Óptica y Electrónica, 72840 Luis Enrique Erro # 1, Tonantzintla, Puebla, México

Resumen El desempeño de los algoritmos de clasificación en el ámbito de los textos, depende en gran medida de la representación de éstos considerando siempre el área en cuestión; encontrar la mejor representación es una tarea ardua por lo que es deseable que utilizando una sola forma simple de representación de documentos se obtenga una buena clasificación, igual o mejor a una obtenida mediante una representación especializada para cada aplicación. En este trabajo se desarrolla un modelo gráfico probabilista, un Campo de Markov, para mejorar una clasificación de textos. Para ello se utilizan otras clasificaciones y la estructura generada por algún algoritmo de agrupamiento. Los resultados que se obtienen son alentadores.

1. Introducción

La clasificación de textos, es la tarea de asignar un documento de texto libre a una o mas categorías predefinidas basado en su contenido [1]. El buen desempeño de los algoritmos de clasificación en muchas ocasiones esta determinado por la representación de los documento que se van a clasificar; es por ello que se necesita trabajar en encontrar la mejor representación posible según sea la aplicación o el tipo de documentos que se deseen clasificar.

Una de las representaciones mas populares y simples de documentos para su clasificación es utilizar las palabras que estos documentos contienen para así formar un vector por documento, en donde cada componente de dicho vector corresponda a una palabra (o *atributo*) dentro del conjunto de documentos. El peso que se le asigne a cada atributo también es algo en que se tiene que pensar cuidadosamente dependiendo de lo que se quiera obtener. Una vez más uno de los pesos mas simples utilizados en esta tarea es el llamado *peso booleado*, éste consiste en asignar un valor que corresponda a la existencia o no de cada atributo en el documento que se esta representando.

Es deseable que exista una única representación para todos los dominios que proporcione resultados comparables con los mejores obtenidos en tales dominios. Es por ello que en este trabajo se explora la posibilidad de mejorar una clasificación base generada a un conjunto de documentos representados en forma vecto-

rial con pesos binarios, utilizando para ello Campo de Markov y características propias de cada conjunto de documentos a clasificar.

En las siguientes secciones del documento se presentará el trabajo relacionado con este proyecto, la metodología utilizada, cómo se llevó a cabo el desarrollo del mismo; así como los experimentos realizados y los resultados obtenidos. Para finalizar se presentan las conclusiones y trabajo a futuro.

2. Trabajo Relacionado

Los trabajos relacionados se pueden dividir en dos grupos, los que utilizan Campos de Markov para realizar algún tipo de procesamiento o tratamiento a documentos y los trabajos que tratan de mejorar la clasificación mediante diferentes técnicas.

Dentro del primer grupo se encuentra el trabajo de [3], donde se investiga la efectividad de la estructura de una consulta en el lenguaje Japonés mediante la composición y descomposición de palabras y frases; para ellos basan su método en un marco teórico de trabajo usando Campos de Markov. Un trabajo más que cae en el primer grupo corresponde al desarrollado por [5] en el cual utilizan un Campo de Markov para identificar texto dentro de documentos ruidosos. El CAM es usado para modelar la estructura geométrica de los textos impresos, escritos a mano y ruidosos y así rectificar una mala clasificación. En [4] desarrollan un marco de trabajo para modelar dependencias de términos vía Campos de Markov.

Dentro del otro grupo se encuentran una gran variedad de trabajos que van desde la utilización de Redes Neuronales, hasta selección de atributos dependiendo de la colección que se esté trabajando. En este proyecto nos interesan más los trabajos que caen dentro del primer grupo.

3. Metodología y Desarrollo

Para desarrollar este proyecto se siguieron los siguientes pasos:

1. Definir la representación de los documentos
2. Definir un conjunto de datos
   a) Clasificar los conjuntos de datos
3. Definir el Campo de Markov
   a) Definir la vecindad de los documentos
4. Evaluar

A continuación se muestra la metodología descrita arriba.

3.1. Representación de los documentos

Dado un conjunto de datos D con n documentos de texto libre, V el vocabulario del conjunto de datos y $|V| = m$; la representación vectorial de cada D_n esta definida como: $D_n = (v_1, v_2, v_3, ..., v_m)$ donde $\forall v_i \subset V, v_i = 1$ si v_i esta en D_n sino $v_i = 0$.

3.2. Conjuntos de datos

Se utilizaron dos conjuntos de prueba, uno temático *Desastres* y uno no temático *Poetas*[2] de los cuales sólo se utilizaron dos categorías. A continuación se describen brevemente ambos conjuntos.

Cada conjunto de datos con las dos categorías seleccionadas, se dividió en dos subconjuntos, uno para entrenar los modelos de clasificación y un subconjunto más que es el que se trata de clasificar con los modelos generados, y al cual se quiere mejorar la clasificación mediante un Campo de Markov.

Desastres Este conjunto consta de noticias recuperadas de la web organizadas en las siguientes 5 categorias: *Forestal*, *Huracán*, *Inundación*, *Sequia* y *Sismo*.

Para este proyecto sólo se utilizaron *Huracán* e *Inundación*, esta elección se basó en la hipótesis de que textos que hablaran de estas dos temáticas serían muy parecidos por lo que la clasificación inicial sería no muy buena. Sin embargo esta colección de datos obtuvo buenos resultados con los clasificadores utilizados. En el Cuadro 1 se muestran los resultados de cinco clasificadores con este corpus, representado como se describió en la sección anterior.

	Naive Bayes	**K-NN**	**C4.5**	**PART**	**AdaBoost M1**
Precisión	0.9325	0.8645	0.9755	0.829	0.96
Recuerdo	0.9225	0.7895	0.9755	0.866	0.95
Exactitud	92.6 %	80.25 %	97.53 %	86.42 %	95.06 %

Cuadro 1. Resultados de clasificadores con la colección *Desastres*

Poetas El conjunto Poetas es una colección de poemas de cinco poetas contemporaneos: *Efraín Huerta*, *Jaime Sabines*, *Octavio Paz*, *Rosario Castellanos* y *Rubén Bonifaz*.

En este trabajo se usaron los poemas de los poetas *Octavio Paz* y *Rubén Bonifaz*. En el Cuadro 2 se muestran los resultados de los clasificadores para este conjunto de pruebas.

	Naive Bayes	**K-NN**	**C4.5**	**PART**	**AdaBoost M1**
Precisión	0.8715	0.25	0.78	0..78	0.7915
Recuerdo	0.8715	0.5	0.7285	0.7285	0.7855
Exactitud	87.14 %	50 %	72.86 %	72.86 %	78.57 %

Cuadro 2. Resultados de clasificadores con la colección *Poetas*

3.3. Definición del Campo de Markov

Se utilizó un Campo de Markov (CAM) para mejorar la clasificación de un conjunto de prueba dada una clasificación inicial, dado un conjunto de vecindades y un conjunto de clasificaciones consideradas como observaciones. En la definición del CAM es necesario definir un sistema de vecindades y una función de potencial. Éstas se describen a continuación:

Dado un algoritmo de agrupamiento A (para este proyecto *k-Means*), un conjunto de documentos D y los respectivos grupos asignados por A, los vecinos de un documento d_x están definidos por todos los documentos d_i tales que pertenezcan al mismo grupo de d_x.

Para la función de potencial $U_p(x) = \sum_c V_c(x) + \lambda \sum_o V_o(x)$, se definió:

$$V_c(x) = \begin{cases} 0 & \text{si la clase del vecino de } x = \text{la clase de } x \\ 1 & \text{en otro caso} \end{cases}$$

$$V_o(x) = \begin{cases} 0 & \text{si la clase de } x \text{ en la observación } o = \text{la clase de } x \\ 1 & \text{en otro caso} \end{cases}$$

4. Experimentos y Resultados

Para realizar las pruebas fue necesario definir la vecindad de cada conjunto de prueba, así como la configuración inicial o base y las observaciones a utilizar. En el Cuadro 3 resume estas variables para cada conjunto de datos.

Como puede observarse en el Cuadro 3 el clasificador base para la Prueba 1 es K-NN, la cual proporciona un 80.25 % de Exactitud (ver el Cuadro 1), para esta prueba, las observaciones porporcionan una exactitud de entre el 86.42 % y el 97.53 % (ver el Cuatro 1).

La colección *Desastre* al ser temática, proporciona exactitudes de hasta casi el 98 %, por lo que ya es complicado mejorar estos resultados. Por lo tanto para esta prueba, se toma como algoritmo base el resultado del clasificador que proporciones la menor exactitud.

Para la prueba 2, el clasificador base proporciona una exactitud del 50 % y las observaciones proporcionan una exactitud máxima del 87.14 %. El corpus *Poetas* es no temático, por lo que los resultados no son tan buenos que con el corpus *Desastres*.

En la prueba 3, se considera como clasificador base el C4.5, esto porque la clasificación de K-NN es muy mala y se pretende saber si el cambio de la clasificación base influye en los resultados del CAM.

Para generar las vecindades en cada prueba, se utilizó el algoritmo *k-Means* con el número de grupos especificados en el renglón 8 del Cuatro 3, no se analizaron para más grupos a partir del grupo 12, ya que la mayoría de los grupos generados eran de sólo un elemento.

En las columnas de los Cuadros 4, 5, 6 se muestra el número de grupos usados en el conjunto de entrenamiento, el número de iteraciones hechas por el CAM

	Prueba 1	**Prueba 2**	**Prueba 3**
Corpus	*Desastres*	*Poetas*	*Poetas*
Clasificación base	K-NN	K-NN	C4.5
Observación 1	Naive Bayes	Naive Bayes	Naive Bayes
Observación 2	C4.5	C4.5	K-NN
Observación 3	PART	PART	PART
Observación 4	AdaBoost M1	AdaBoost M1	AdaBoost M1
Número de grupos usados	3-11	4-11	4-11

Cuadro 3. Especificaciones para las pruebas

hasta que éste convergió, el valor de λ utilizado en la función de potencial, la exactitud de la clasificación generada por el CAM luego de las iteraciones y por último se muestra la matriz de confusión para tener una idea de cómo queda la clasificación final.

Los resultados obtenidos en la Prueba 1 se muestran en el Cuadro 4, en dicho Cuadro se puede ver que el mejor resultado es obtenido con un $\lambda = 10$ para $4, 6, 7$ y 8 grupos. Para los grupos $9, 10$ y 11 los resultados se mantienen iguales que para los grupos 7 y 8. Cabe recordar que la precisión del clasificador base para esta prueba es de 80.25 %.

En el Cuadro 5 con un precisión base de 50 %, los mejores resultados se observan cuando el valor de $\lambda = 20$, para todas las cantidades diferentes de grupos utilizados. Sin embargo con un $\lambda = 10$ y número de grupos del 6 al 11, también se obtiene el mayor valor.

En el caso de la Prueba 3, el Cuadro 6 un solo resultado que es el mismo para todos los grupos y valores de $\lambda > 10$. Este valor máximo es el mismo obtenido para la Prueba 2, sin embargo la clasificación inicial aqui fue de 72.9 % por lo que la CAM no mejoró, pero tampoco empeoró.

En resumen, para la Prueba 1, el CAM mejora la clasificación de los clasificadores usados tanto en la configuración inicial como en las observaciones. Para las pruebas 2 y 3, no se mejora en relación a todas las observaciones pero el resultado es comparable. La Figura 1 muestra esta tendencia.

5. Conclusiones y Trabajo Futuro

Entre las conclusiones a las que se ha llegado mediante la realización de este proyecto es que sí es posible mejorar la clasificación de textos mediante un Campo de Markov, sin embargo qué tanto mejore, dependerá mucho de las observaciones usadas. Se puede concluir también que no se necesitan muchas iteraciones del campo, pues éste siempre converge entre 2 y 3 iteraciones.

Una conclusión importante es que mientras mas valor se le den a las observaciones, y mientras éstas no sean tan malas, el campo mejorará la clasificación base, si ésta tiene un desempeño menor que las observaciones.

No se exploraron escenarios interesantes como utilizar bases de datos de mayor tamaño, pues las que se utilizaron tienen menos de 100 elementos. También

Número de grupos	Número de iteraciones	λ	Exactitud resultante	Matriz de confusión
3	2	1	53 %	$\begin{array}{\|c\|c\|}\hline 0 & 38 \\ \hline 0 & 43 \\ \hline \end{array}$
	3	10	96.3 %	$\begin{array}{\|c\|c\|}\hline 35 & 3 \\ \hline 0 & 43 \\ \hline \end{array}$
4	2	1	53 %	$\begin{array}{\|c\|c\|}\hline 0 & 38 \\ \hline 0 & 43 \\ \hline \end{array}$
	3	10	**98.8 %**	$\begin{array}{\|c\|c\|}\hline 37 & 1 \\ \hline 0 & 43 \\ \hline \end{array}$
5	2	1	54.3 %	$\begin{array}{\|c\|c\|}\hline 1 & 37 \\ \hline 0 & 43 \\ \hline \end{array}$
	3	10	97.5 %	$\begin{array}{\|c\|c\|}\hline 37 & 1 \\ \hline 0 & 43 \\ \hline \end{array}$
6	2	1	54.3 %	$\begin{array}{\|c\|c\|}\hline 1 & 37 \\ \hline 0 & 43 \\ \hline \end{array}$
	3	10	**98.8 %**	$\begin{array}{\|c\|c\|}\hline 37 & 1 \\ \hline 0 & 43 \\ \hline \end{array}$
7	2	1	56.8 %	$\begin{array}{\|c\|c\|}\hline 3 & 35 \\ \hline 0 & 43 \\ \hline \end{array}$
	3	10	**98.8 %**	$\begin{array}{\|c\|c\|}\hline 37 & 1 \\ \hline 0 & 43 \\ \hline \end{array}$
8	2	1	56.8 %	$\begin{array}{\|c\|c\|}\hline 3 & 35 \\ \hline 0 & 43 \\ \hline \end{array}$
	3	10	**98.8 %**	$\begin{array}{\|c\|c\|}\hline 37 & 1 \\ \hline 0 & 43 \\ \hline \end{array}$

Cuadro 4. Resultados para la Prueba 1

Número de grupos	Número de iteraciones	λ	Exactitud resultante	Matriz de confusión
4	3	10	50 %	$\begin{array}{\|c\|c\|}\hline 33 & 2 \\ \hline 33 & 2 \\ \hline \end{array}$
	2	20	**72.9 %**	$\begin{array}{\|c\|c\|}\hline 33 & 2 \\ \hline 17 & 18 \\ \hline \end{array}$
5	3	10	50 %	$\begin{array}{\|c\|c\|}\hline 33 & 2 \\ \hline 33 & 2 \\ \hline \end{array}$
	2	20	**72.9 %**	$\begin{array}{\|c\|c\|}\hline 33 & 2 \\ \hline 17 & 18 \\ \hline \end{array}$
6	3	10	**72.9 %**	$\begin{array}{\|c\|c\|}\hline 33 & 2 \\ \hline 17 & 18 \\ \hline \end{array}$
	2	20	**72.9 %**	$\begin{array}{\|c\|c\|}\hline 33 & 2 \\ \hline 17 & 18 \\ \hline \end{array}$

Cuadro 5. Resultados para la Prueba 2

Número de grupos	Número de iteraciones	λ	Exactitud resultante	Matriz de confusión
4	3	10	**72.9 %**	$\begin{array}{\|c\|c\|}\hline 34 & 1 \\ \hline 18 & 17 \\ \hline \end{array}$

Cuadro 6. Resultados para la Prueba 3

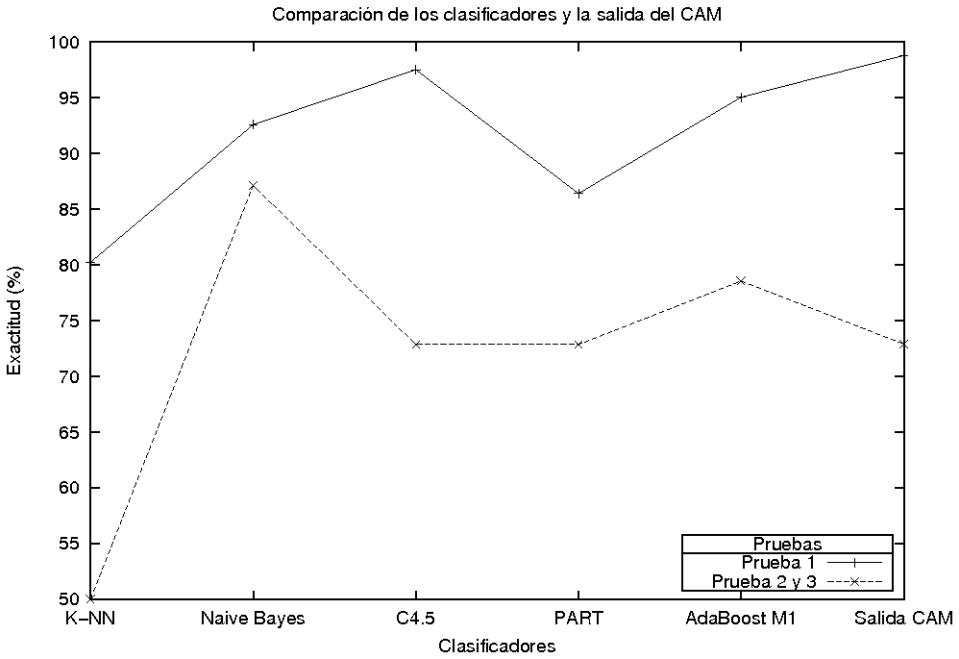

Figura 1. Comparación de resultados

se busca en un trabajo a futuro realizar experimentos tanto para menos como para mayores observaciones.

Referencias

[1] Aas, K., and Eikvil, L. Text categorization: A survey, technical report. Tech. rep., Norwegian Computing Center, 1999.

[2] Coyotl-Morales, R. M., Villaseñor-Pineda, L., y Gómez, M. M., and Rosso, P. Authorship attribution using word sequences. In *11th Iberoamerican Congress on Pattern Recognition, CIARP 200. Lecture Notes in Computer Science* (2006), vol. 4225, Springer.

[3] Eguchi, K., and Croft, W. B. Query structuring with two-stage term dependence in the japanese language. In *Information Retrieval Technology* (2006), Springer, pp. 522–529.

[4] Metzler, D., and Croft, W. B. A markov random field model for term dependencies. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 2005), ACM, pp. 472–479.

[5] Zheng, Y., Li, H., and Doermann, D. Text identification in noisy document images using markov random field. In *In 7th International Conference on Document Analysis and Recognition (ICDAR* (2003), pp. 599–605.