Naive Bayes Multinomial para Clasificación de Texto Usando un Esquema de Pesado por Clases

Emmanuel Anguiano-Hernández

Abril 29, 2009

Abstract

Tratando de mejorar el desempeño de un clasificador Naive Bayes Multinomial se modifica la representación de los documentos de prueba cambiando los valores de los atributos por un conteo de apariciones de los mismos atributos en la clase sobre la que se evalua la probabilidad de pertenencia. Se realizan algunos experimentos con un conjunto de noticias y se comparan los resultados contra los obtenidos con el esquema tradicional.

Introducción

Los clasificadores de la familia de Naive Bayes son muy usados en la tarea de Clasificación de Textos debido a que producen resultados comparables con los obtenidos por otros métodos más sofisticados y son relativamente sencillos de implementar. Uno de los aspectos desfavorables de estos clasificadores es que asumen que los términos que aparecen en un documento son todos independientes entre sí lo cuál puede no ser totalmente cierto debido a la naturaleza estructural del lenguaje. Muchos trabajos han tratado de proporcionar información adicional al clasificador para que las consecuencias de la suposición de independencia sea suavizada.

El clasificador Naive Bayes Simple considera la probabilidad de aparición de cada término dada la clase de forma binaria, es decir el término aparece o no y entonces su probabilidad condicional dada la clase es o no considerada. En este sentido, el clasificador Naive Bayes Multinomial suele mejorar el desempeño pues considera el número de apariciones del término para evaluar la contribución de la probabilidad condicional dada la clase con lo que el modelado de cada documento se ajusta mejor a la clase a la que pertenece.

Puede pensarse que si se modifica la representación de los documentos, de manera que el conteo de los términos que aparecen en él se cambia por el número de apariciones del término en la clase cuya probabilidad de pertenencia del documento se está evaluando, se está proporcionando información adicional al clasificador para que la asignación de clase mejore.

Este trabajo trata sobre el desarrollo de algunos experimentos que buscan mejorar el desempeño de un clasificador NBM modificando la representación de los documentos sobre los que se prueba el modelo. El documento está estructurado como sigue: En la sección de Teoria se proporciona información sobre la tarea de Clasificación de Texto, su definición, métodos comunmente usados y forma en que se modifican para mejorar los resultados, también se describe al clasificador NBM y la modificación propuesta así como el conjunto de datos sobre el que se realizan las pruebas. En la sección de Montaje Práctico se describen las condiciones de entrenamiento, prueba y experimentación. En Resultados y Análisis se muestran y comentan los resultados de los experimentos realizados y Finalmente en Conclusiones y Perspectivas se discuten las conclusiones y se proponen modificaciones a los experimentos para buscar mejoras.

Teoría

Clasificación de Texto

La tarea de clasificación de textos consiste en asignar a cada documento de una colección una etiqueta que designa a la clase a la que este documento pertenece. La decisión sobre qué etiqueta debe asignarse a un documento determinado se toma a partir de un modelo construido con una parte de la colección denominada conjunto de entrenamiento. El objetivo de la tarea es construir un modelo que prediga correctamente la clase de un conjunto de documentos, llamado de prueba, que no intervinieron en la construcción del modelo [2].

Cuando se desea mejorar la clasificación para un dominio dado suelen usarse dos enfoques [1]:

- Modificar el modelo del clasificador: ya sea cambiándolo por otro o alterando sus parámetros para adaptarlo a los datos que debe clasificar. Existe una gran cantidad de clasificadores usados en esta tarea como Arboles de Decisión, Maquinas de Vectores de Soporte, k-NN o modelos probabilistas entre los que se encuentran las Redes Bayesianas y Naive Bayes.

- Modificar la representación de los datos: diferentes clasificadores producen mejores o peores modelos segun la representación de los datos que se desea procesar, la elección de una representación adecuada dado un clasificador es un problema por sí mismo. Hay varias maneras de representar un documento antes de procesarlo en un clasificador pero entre las más usuales destaca el modelo vectorial en que cada documento es representado como un vector de dimensión igual al tamaño del vocabulario de la colección y en que el valor de cada atributo corresponde al conteo de apariciones del término correspondiente en el documento aunque también suele usarse una representación binaria (1 si el término aparece en el documento sin importar el número de apariciones, 0 si no es asi) u otros esquemas de pesado.

En este trabajo se emplea un modelo probabilista, particularmente Naive Bayes Multinomial para clasificar documentos de una colección de noticias con una representación vectorial y un esquema de pesado que considera la frecuencia de aparición de cada término del vocabulario en el documento (tf) para el entrenamiento del modelo. Posteriormente la representación de los documentos de prueba se modifica usando un esquema de pesado por clases.

Clasificador Naive Bayes Multinomial

Naive Bayes es uno de los modelos probabilistas más simples y más usados en clasificación de texto porque produce resultados tan buenos como otros modelos más sofisticados.

Se basa en la aplicación de la Regla de Bayes para predecir la probabilidad condicional de que un documento pertenezca a una clase $P(c_i|d_j)$ a partir de la probabilidad de los documentos dada la clase $P(d_j|c_i)$ y la probabilidad a priori de la clase en el conjunto de entrenamiento $P(c_i)$

$$P(c_i|d_j) = \frac{P(c_i)P(d_j|c_i)}{P(d_j)}$$

Dado que la probabilidad de cada documento $P(d_j)$ no aporta información para la clasificación, el término suele omitirse. La probabilidad de un documento dada la clase suele asumirse como la probabilidad conjunta de los términos que aparecen en dichos documentos dada la clase y se calculan como:

$$P(d_j|c_i) = \Pi_{t=1}^{|V|} P(w_t|c_i)$$

Adicionalmente, el modelo Naive Bayes Multinomial considera la frecuencia de aparición de cada término en los documentos x_t en vez de una ocurrencia binaria [1]:

$$P(d_j|c_i) = \Pi_{t=1}^{|V|} P(w_t|c_i)^{x_t}$$

El término $P(w_t|c_i)$ se cálcula a partir del número de apariciones de cada término w_t en una clase c_i pero para evitar el problema de las probabilidades 0 se usa la estimación de Laplace

$$P(w_t|c_i) = \frac{1 + n(w_t, c_i)}{|V| + n(c_i)}$$

Donde $n(w_t, c_i)$ es el número de ocurrencias de w_t en c_i, $vertV|$ es el tamaño del vocabulario y $n(c_i)$ es el conteo total de palabras en c_i. De este modo, la clasificación se hace buscando el argumento que maximiza la función:

$$c^*(d) = argmax_{ci} p(c_i) \Pi_{t=1}^{|V|} P(w_t|c_i)^{x_t}$$

La modificación propuesta para este trabajo consiste en cambiar x_t por $x_t c_i$ la frecuencia de aparición del término en la clase.

Datos

En este trabajo se utiliza la colección R8 de documentos previamente clasificados para entrenar y probar el sistema. R8 es una subcolección de Reuters-21578 una colección de noticias de la agencia Reuters del año 1987 que son usadas como un estándar para evaluar sistemas. Se eligió R8 ya que presenta gran desbalance en el número de noticias que pertenecen a cada una de las clases y por lo tanto es adecuada para probar si un sistema de clasificación es eficiente. En la figura 1 se muestra la distribución de R8.

R8

Clase	Entrenamiento	Prueba
acq	1596	696
crude	253	121
earn	2840	1083
grain	41	10
interest	190	81
money-fx	206	87
ship	108	36
trade	251	75
Total	**5485**	**2189**

Figure 1: Distribución de la colección R8.

Montaje Práctico

Entrenamiento del Modelo

El conjunto de entrenamiento de R8 fue preprocesado para convertir cada noticia a una representación vectorial con 19367 atributos, cada uno correspondiente con una palabra del vocabulario del conjunto. El valor de cada atributo corresponde al conteo de apariciones de dicho término en el documento. En la matriz resultante se calcularon las probabilidades a priori de cada una de las clases $P(c_i)$ y las probabilidades de cada término dada la clase $P(w_t|c_i)$, con ello se construyo una matriz de probabilidades que es el modelo a usarse en la clasificación. También se construyo una matriz con las apariciones de cada término por clase $x_t, c_i = n(w_t, c_i)$ que es usada en la fase experimental de este trabajo.

Prueba del Modelo

El conjunto de prueba también fue trasladado a una representación vectorial con los mismos atributos que el conjunto de entrenamiento y se probó el modelo según la ecuación característica del clasificador Naive Bayes Multinomial evaluando la probabilidad de pertenencia a cada una de las 8 clases. La clase con mayor probabilidad se eligió como la clase de cada uno de los documentos de

prueba y posteriormente se comparó con la etiqueta de clase para evaluar el desempeño del clasificador.

Experimentos

Con el objetivo de probar si el desempeño del clasificador Naive Bayes Multinomial podía mejorarse añadiendo a las instancias de prueba información adicional sobre cada una de las clases a las que podrían pertenecer se experimentó cambiando los valores de x_t para cada documento por $x_t c_i$ el conteo de apariciones de cada término en cada clase en la ecuación de clasificación del modelo, de este modo, al evaluar la probabilidad de pertencia del documento a cada una de las clases se remplazaban los conteos de aparición diferentes de cero de cada término en el documento por el conteo correspondiente al mismo término pero en la clase cuya probabilidad de pertenencia se evalua. Se asignó la etiqueta de la clase que maximizaba la probabilidad de pertenencia y posteriormente se comparó con la etiqueta de la clase para evaluar el desempeño de este montaje experimental. También se comparan los resultados obtenidos con este procedimiento con los obtenidos mediante la evaluación normal con x_t.

Resultados y Análisis

El clasificador Naive Bayes Mutinomial demostró un buen desempeño aun cuando se presentaba un gran desbalance en la distribución de las clases y un elevado número de atributos, en la figura 2 puede verse la matriz de confusión resultante de la clasificación sobre el conjunto de prueba asi como el recuerdo y precisión para cada una de las clases.

	Clase Asignada por NBM									
Clase Real	**acq**	**crude**	**earn**	**grain**	**interest**	**money-fx**	**ship**	**trade**	**Precisión**	**Recuerdo**
acq	**682**	3	8	0	0	1	0	2	0.817746	0.979885
crude	6	**111**	0	0	0	0	3	1	0.720779	0.917355
earn	123	2	**958**	0	0	0	0	0	0.987629	0.88458
grain	0	2	0	**8**	0	0	0	0	1	0.8
interest	8	8	0	0	**63**	2	0	0	0.818182	0.777778
money-fx	3	7	2	0	14	**59**	0	2	0.855072	0.678161
ship	1	4	0	0	0	0	**26**	5	0.896552	0.722222
trade	11	17	2	0	0	7	0	**38**	0.791667	0.506667

Figure 2: Matriz de Confusión para NBM con x_t.

En la tabla puede observarse que las clases sobre las que se obtuvo mejor desempeño son aquellas con mayor número de elementos e interesantemente la clase con menor número de elementos.

Aqui debe mencionarse un detalle importante. El cálculo de probabilidades de pertenencia a cada una de las clases resultó complicado en principio debido al elevado número de atributos. Como consecuencia de esto, las probabilidades

de aparición de cada término dada la clase eran muy bajas y entonces el producto de 19367 números mucho menores que uno desbordaba la capacidad de representación de los números flotantes de doble precisión (recordando que el clasificador fue escrito en c++ pues no pudo usarse el clasificador de weka debido al mismo problema de la alta dimensionalidad) por lo que fue necesario escalar la probabilidad de cada término dada la clase en un factor que empiricamente se fijó en 10000. Como consecuencia de esto, la suma de probabilidades de todos los términos en una clase no sumaba 1 como establece [1] sino 10000. Dado que este escalamiento fue uniforme sobre todos los términos y todas las clases puede suponerse que no afecta los resultados de la clasificación.

Para el montaje experimental remplazando x_t por $x_t c_i$ se obtuvieron resultados desfavorables como se muestra en la figura 3.

	Clase Asignada por NBM									
Clase Real	**acq**	**crude**	**earn**	**grain**	**interest**	**money-fx**	**ship**	**trade**	**Precisión**	**Recuerdo**
acq	**350**	0	346	0	0	0	0	0	0.6917	0.502874
crude	17	**3**	101	0	0	0	0	0	1	0.024793
earn	16	0	**1067**	0	0	0	0	0	0.635497	0.985226
grain	0	0	10	**0**	0	0	0	0	0	0
interest	41	0	40	0	**0**	0	0	0	0	0
money-fx	46	0	40	0	0	**0**	0	1	0	0
ship	19	0	17	0	0	0	**0**	0	0	0
trade	17	0	58	0	0	0	0	**0**	0	0

Figure 3: Matriz de Confusión para NBM con $x_t c_i$.

Puede observarse en la figura 3 que el clasificador asignó prácticamente la totalidad de los documentos de prueba a las clases acq y earn, esto puede explicarse si se toma en cuenta que son las clases mayoritarias tanto en el conjunto de prueba como en el de entrenamiento, por tanto son las clases que mayor cantidad de palabras aportan al vocabulario y en las que también hay la mayor cantidad de apariciones por palabra, por lo mismo las contribuciones de probabilidad de palabras dada la clase $P(w_t|c_i)$ resultaban dominantes para estas clases con lo que las probabilidades de pertenencia a ellas se disparaban en practicamente todos los documentos. Como el clasificador selecciona la probabilidad de pertenencia mayor, asignaba los documentos a estas clases.

Usando como base el argumento anterior se realizó un tercer experimento en el que se impidió al clasificador asignar las clases acq o earn para estimar si la clasificación sobre las clases más o menos balanceadas producia mejores resultados, sin embargo debe mencionarse que siguió usandose la misma información de entrenamiento y únicamente se descartaba la posibilidad de asignar las etiquetas de acq y earn por lo que el clasificador debia elegir entre las 6 restantes. Estas condiciones alteraron las medidas de evaluación pues ahora todos los documentos que pertenencian a acq o earn fueron asignados a otras clases contaminando la precisión mientras que el recuerdo reflejó en mejor medida las posibles mejoras: la clase crude mostró un recuerdo de 0.975207 mientras trade

tuvo $r = 1$.

En las figuras 4 y 5 puede verse la comparativa en las medidas de desempeño para cada uno de los experimentos, *xt* representa la prueba de NBM usando x_t, *xtci* es usando $x_t c_i$ y *xtci-noae* corresponde a cuando el clasificador no tenia permitido asignar a las clases acq ni earn.

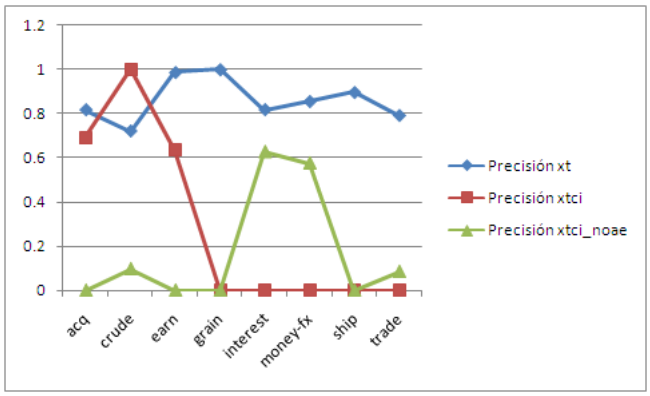

Figure 4: Comparativa de Precisión para los Diferentes Experimentos

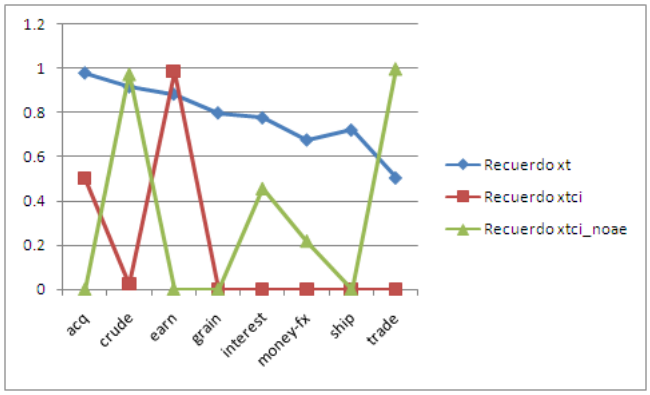

Figure 5: Comparativa de Recuerdo para los Diferentes Experimentos

Conclusiones y Perspectivas

Es posible que debido a la naturaleza desbalanceada del corpus empleado los resultados del experimento de usar el clasificador Naive Bayes Multinomial con un esquema de pesados por clase en que se remplaza el conteo de apariciones

de un término en el documento por el conteo del término en la clase cuya probabilidad de pertenencia se evalua hayan sido desfavorables pues en las medidas de evaluación se observan claros sesgos del clasificador hacia las clases más numerosas. Debido al tiempo de preprocesamiento necesario para usar un nuevo corpus no fue posible probar la modificación de NBM en un corpus balanceado y comparar los resultados para saber si la modificación mejora la clasificación. Mientras tanto de este experimento puede concluirse que dicha modificación a NBM no mejora la clasificación si se usa un corpus tan desbalanceado como R8.

Sin embargo, el clasificador NBM demostró un buen desempeño cuando se usa de la forma tradicional, es decir considerando el número de apariciones de cada término en el documento que se está evaluando, aun cuando los conjuntos de entrenamiento y prueba estaban desbalanceados.

El experimento en que se impidió al clasificador asignar las clases mayoritarias arrojó solo resultados parciales debido a que sería necesario reentrenar el modelo con un conjunto de entrenamiento sin dichas clases para obtener resultados concluyentes.

Posibles modificaciones a los experimentos de este trabajo son:

- Utilizar un corpus balanceado para verificar el desempeño de NBM sin el sesgo en las probabilidades producido por el alto número de documentos que pertenecen a las clases mayoritarias (80 por ciento de los documentos pertenencen a 2 de las 8 clases) y las consecuencias en las probabilidades de aparición de cada término en una clase que se ven viciadas ya que esas dos clases aportan la mayor parte del vocabulario y entonces la mayor parte de las apariciones de cada término.

- Modificar el corpus de entrenamiento existente para balancear las clases. Pues se observó que las clases crude, interest, money-fx y trade están cerca del balance con alrededor de 200 documentos cada una, si las clases mayoritarias son reducidas hasta un número cercano sería posible reentrenar al modelo para clasificar.

Referencias

- [1] Schneider Karl-Michael. *Techniques for Improving the Performance of Naive Bayes for Text Classification.* University of Passau, Department of General Linguistics. Passau, Germany. 2004.

- [2] Álvarez Romero, Juan de Dios. *Clasificación Automática de Textos usando Reducción de Clases Basada en Prototipos* (Tesis de Maestría). INAOEP. México. 2009