Comparativa de Aproximaciones a SVM Semisupervisado Multiclase para Clasificación de Páginas Web

A Comparison of Approaches to Semi-supervised Multiclass SVM for Web Page Classification

Arkaitz Zubiaga, Víctor Fresno
Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
C/Juan del Rosal, 16, E-28040 Madrid
{azubiaga, vfresno}@lsi.uned.es

Resumen: En este artículo se realiza un estudio sobre clasificación semisupervisada multiclase de páginas web mediante SVM. Se propone tanto la combinación de clasificadores binarios semisupervisados como la de clasificadores multiclase supervisados. Los experimentos demuestran un gran rendimiento para los primeros, siendo en algunas ocasiones una mejor opción la no utilización de documentos no etiquetados en la fase de aprendizaje, y basarse directamente en algoritmos supervisados.
Palabras clave: SVM, multiclase, semisupervisado, clasificación de páginas web

Abstract: In this paper we present a study for semi-supervised multiclass web page classification using SVM. We propose not only combining binary semi-supervised classifiers, but also multiclass supervised ones. Our experiments show great performance for the latter method, where ignoring unlabeled documents could be better for some cases, using only labeled documents for the learning task, directly based on supervised algorithms.
Keywords: SVM, multiclass, semi-supervised, web page classification

1. Introducción

El número de documentos web está creciendo rápidamente en los últimos años, lo que hace que su organización resulte cada vez más complicada. Es por ello que, actualmente, la clasificación de páginas web se ha convertido en una tarea necesaria.

La clasificación de páginas web puede definirse como la tarea que organiza una serie de documentos web etiquetándolos con sus correspondientes categorías prefijadas. Aunque se han realizado múltiples estudios para clasificación de textos, sobre todo en la rama de noticias, su aplicación a las páginas web está aún por profundizar (Qi y Davison, 2007).

En este trabajo se pone el foco en la clasificación de páginas web enmarcada dentro del paradigma del aprendizaje automático (Mitchell, 1997). Los problemas de clasificación se pueden dividir en diferentes tipos. Por una parte, la clasificación puede ser *binaria*, donde únicamente existen dos categorías posibles para cada documento, o puede ser *multiclase*, donde se dispone de tres o más categorías; y por otra, el sistema de aprendizaje con el que se alimenta el clasificador puede ser *supervisado*, donde todos los documentos de entrenamiento están previamente etiquetados, o *semisupervisado*, donde se aprende con una colección de entrenamiento que está compuesta por algunos documentos etiquetados y muchos no etiquetados.

Se han aplicado diferentes algoritmos al problema de la clasificación de textos (Sebastiani, 2002). Ante esta problemática, las máquinas de vectores de soporte (SVM, Support Vector Machines (Joachims, 1998)) se perfilan como una buena alternativa. Teniendo en cuenta que la clasificación de páginas web es, generalmente, un problema multiclase, y que el número de documentos etiquetados del que se dispone, comparado con las dimensiones de la Web, es muy reducido, el

problema se convierte en multiclase y semisupervisado.

Existen diversos estudios referentes tanto a SVM multiclase como a SVM semisupervisado, pero apenas se ha investigado en la unión de ambos casos. Este artículo propone y evalúa diferentes aproximaciones para la implementación de un método de SVM multiclase y semisupervisado.

En la sección 2 se explican los avances obtenidos en los últimos años en la clasificación mediante SVM, tanto para aprendizaje semisupervisado como para taxonomías multiclase, además de presentar alternativas para clasificación semisupervisada multiclase. En la sección 3, se muestran los detalles de la experimentación realizada, para seguir en la sección 4 con el análisis de los resultados de ésta. En la sección 5, para finalizar, se exponen las conclusiones deducidas tras el proceso.

2. Clasificación con SVM

En la última década, SVM se ha convertido en una de las técnicas más utilizadas para clasificación, debido a los buenos resultados que se han obtenido. Esta técnica se basa en la representación de los documentos en un modelo espacio vectorial, y asume que los documentos de cada clase se agrupan en regiones separables del espacio de representación; en base a ello, trata de buscar un hiperplano que separe ambas clases, el cual maximice la distancia entre los documentos de cada clase y el propio hiperplano, lo que se denomina margen. Este hiperplano se define mediante la siguiente función:

$$f(x) = w \cdot x + b$$

Esto supondría tener en cuenta todos los valores posibles para la función, para después quedarse con los que maximicen los márgenes. Esto resulta muy difícil de optimizar, por lo que se utiliza la siguiente función de optimización equivalente:

$$\text{mín}\, \frac{1}{2}||w||^2 + C\sum_{i=1}^{l} \xi_i^d$$

$$\text{Sujeto a:} \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0$$

donde C es el parámetro de penalización y ξ_i es la distancia entre el hiperplano y el documento i.

De esta manera únicamente se resuelven problemas linealmente separables, por lo que en muchos casos se requiere de la utilización de una función de kernel para la redimensión del espacio; de este modo, el nuevo espacio obtenido sí que resulta linealmente separable. Posteriormente, la redimensión se deshace, de modo que el hiperplano encontrado será transformado al espacio original, constituyendo la función de clasificación.

Es importante destacar que esta función únicamente puede resover problemas binarios.

2.1. SVM multiclase

Debido a la naturaleza dicotómica de SVM, surgió la necesidad de implementar nuevos métodos que pudieran resolver problemas multiclase. Como aproximación directa, (Weston y Watkins, 1999) proponen una modificación de la función de optimización que tiene en cuenta todas las clases:

$$\text{mín}\, \frac{1}{2}\sum_{m=1}^{n} ||w_m||^2 + C\sum_{i=1}^{l}\sum_{m \neq y_i} \xi_i^m$$

Sujeto a:

$$w_{y_i} \cdot x_i + b_{y_i} \geq w_m \cdot x_i + b_m + 2 - \xi_i^m, \xi_i^m \geq 0$$

Otras técnicas para la aproximación a SVM multiclase de k clases se han basado en la combinación de clasificadores binarios (Hsu y Lin, 2002):

- *one-against-all* construye k clasificadores que definen otros tantos hiperplanos que separan la clase i de los *k-1* restantes, donde a cada nuevo documento se le asigna la clase cuyo clasificador maximice el margen:

$$\hat{C}_i = \arg \max_{i=1,\ldots,k} (w_i x + b_i)$$

- *one-against-one* construye $\frac{k(k-1)}{2}$ clasificadores, uno para cada par de clases posible. Posteriormente, clasifica cada nuevo documento mediante todos los clasificadores, donde se añade un voto a la clase ganadora para cada caso, resultando la que más votos sume la clase propuesta.

2.2. Aprendizaje semisupervisado para SVM (S^3VM)

Las técnicas de aprendizaje semisupervisado se diferencian en que, además de los do-

cumentos previamente etiquetados, se utilizan documentos no etiquetados para la fase de entrenamiento (Joachims, 1999). Las SVM semisupervisadas se conocen también por sus iniciales S^{3}VM. En el caso de SVM, su adaptación a aprendizaje semisupervisado supone un gran coste computacional, ya que la función resultante no es convexa, para lo que es mucho más complicado obtener el mínimo. Para relajar la computación de esta función, se suelen utilizar técnicas de optimización convexa (Xu et al., 2007), donde la obtención del mínimo para la función resultante es mucho más sencilla. No obstante, casi todo el trabajo existente en la literatura relativa a este aspecto ha sido para clasificaciones binarias.

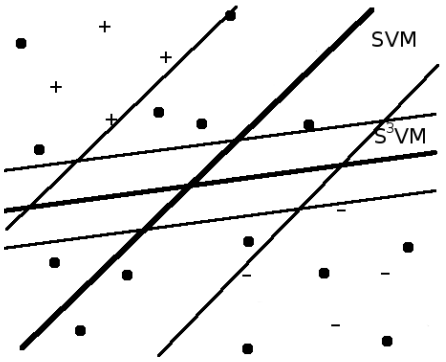

Figura 1: SVM vs S^{3}VM, donde +/- representan documentos etiquetados, y los puntos, no etiquetados

2.3. S^{3}VM multiclase

En los problemas donde la taxonomía dispone de más de dos categorías y el número de documentos previamente etiquetados es muy pequeño, se precisa la combinación de las dos técnicas anteriormente expuestas, lo que supone un método de S^{3}VM. Los problemas reales de clasificación de páginas web suelen cumplir con estas características, ya que el número de categorías suele ser mayor que dos, y la pequeña colección de documentos etiquetados de la que se dispone normalmente implica la necesidad de utilizar documentos no clasificados en la fase de entrenamiento.

La única aproximación a S^{3}VM multiclase encontrada en la literatura ha sido propuesta por (Yajima y Kuo, 2006), con una técnica que traslada la función multiclase directa al entorno semisupervisado. La función de optimización resultante es la siguiente:

$$\text{mín}\ \frac{1}{2}\sum_{i=1}^{h} \beta^{i^T} K^{-1} \beta^i$$

$$+C\sum_{j=1}^{l}\sum_{i\neq y_j} \text{máx}\{0, 1-(\beta_j^{y_j}-\beta_j^i)\}^2$$

donde β representa el producto entre un vector de variables y una matriz de kernel definidas por el autor.

Esta función de optimización, sin embargo, puede resultar muy costosa, debido a la cantidad de variables que se deben tener en cuenta en el proceso de minimización de la misma, lo que hace interesante el problema de encontrar otros enfoques a S^{3}VM multiclase.

2.3.1. Alternativas para S^{3}VM multiclase

Ante la escasez de propuestas para la implementación de métodos de S^{3}VM multiclase, nuestro objetivo es el de proponer y comparar diversas técnicas aplicables a este entorno, basándose en las ya utilizadas para problemas supervisados multiclase y semisupervisados binarios.

- *2 steps*: Hemos denominado así a la técnica que se basa en la aproximación supervisada multiclase explicada en la sección 2.1. Este método trabaja, en el primer paso, sobre la colección de entrenamiento, aprendiendo con los documentos etiquetados y prediciendo los no etiquetados; a posteriori, se etiquetan estos últimos según las predicciones obtenidas. Como segundo paso, se realiza la clasificación habitual para este método, ya que ahora la colección se ha convertido en supervisada, con todos los ejemplos de entrenamiento etiquetados.
- *one-against-all* y *one-against-one* son propuestas basadas en la combinación de clasificadores binarios semisupervisados, que aunque se han utilizado en colecciones supervisadas, nunca han sido experimentadas a colecciones con documentos no etiquetados. Sin embargo, el enfoque *one-against-one* plantea un problema intrínseco de ruido en la fase de entrenamiento con los documentos no etiquetados, ya que cada clasificador para un par de categorías únicamente debe ser alimentado por documentos que le correspondan, y el problema radica en la

imposibilidad de excluir aquellos ejemplos no etiquetados que no deberían incluirse.

- *all-against-all*: Aparte de los dos anteriores, en este trabajo se presenta una nueva propuesta de combinación de clasificadores binarios, que hemos denominado *all-against-all*, y que podría ser utilizada tanto para aprendizaje supervisado como semisupervisado. En ella se definen $2^{n-1}-1$ clasificadores, correspondientes a todos los enfrentamientos posibles entre las clases, teniendo en cuenta que todas las clases deben caer en uno u otro lado. Por ejemplo, para un problema de cuatro clases, se generarán los clasificadores *1 vs 2-3-4*, *1-2 vs 3-4*, *1-2-3 vs 4*, *1-3 vs 2-4*, *1-4 vs 2-3*, *1-2-4 vs 3* y *1-3-4 vs 2*. Cada nuevo documento recibido en la fase de clasificación se someterá a cada uno de los clasificadores generados, sumando, como voto, el valor del margen obtenido en cada caso para las clases en el lado positivo. Una vez realizado esto, se asignará la clase para la que mayor votación ha obtenido cada documento. Esta aproximación puede ser muy costosa para grandes taxonomías, ya que el número de clasificadores aumentaría de forma considerable, pero se podría esperar un buen rendimiento para cantidades de clases reducidas.

3. Diseño de la experimentación

Para la realización de la experimentación se ha procedido a la implementación de los algoritmos descritos en el apartado anterior, y su ejecución sobre las colecciones de datos escogidas. Todos los documentos de las colecciones utilizadas están etiquetados, por lo que cada una de ellas se ha dividido en una colección de entrenamiento, que sirve para que el clasificador aprenda, y otra de test, que sirva para que el sistema cree las predicciones y se pueda evaluar su rendimiento. A continuación se explican con más detalle las características de la experimentación llevada a cabo.

3.1. Colecciones de datos

Para esta experimentación se han utilizado colecciones de páginas web que ya han sido probadas anteriormente en problemas de clasificación:

- *BankSearch* (Sinka y Corne, 2002), compuesta por 10.000 páginas web sobre 10 clases, de muy diversos temas: bancos comerciales, construcción, agencias aseguradoras, java, C, visual basic, astronomía, biología, fútbol y motociclismo. 4.000 ejemplos han sido asignados a la colección de entrenamiento, y los 6.000 restantes a la de test.

- *WebKB* [1], formada por 4.518 documentos extraídos de 4 sitios universitarios y clasificados sobre 7 clases (estudiante, facultad, personal, departamento, curso, proyecto y miscelanea). La clase miscelanea se ha eliminado de la colección debido a la ambigüedad, resultando 6 categorías. De todos los ejemplos que componen la colección, 2.000 se han asignado al entrenamiento y 2.518 al de test.

- *Yahoo! Science* (Tan et al., 2002), que tiene 788 documentos científicos, clasificados sobre 6 ámbitos diferentes de la ciencia (agricultura, biología, ciencias terrestres, matemáticas, química y otros). Se han definido 200 documentos para el entrenamiento, y 588 para el test.

Desde la colección de entrenamiento, para cada caso, se han creado diferentes versiones, entre las que varía el número de documentos etiquetados, dejando el resto como no etiquetados, pudiendo probar así las diferentes aproximaciones semisupervisadas.

Para la representación vectorial de los documentos que componen cada colección, se ha basado en los valores tf-idf de los unitérminos encontrados en los textos, excluyendo las de mayor y menor frecuencia. Los unitérminos resultantes han sido los que han definido las dimensiones del espacio vectorial.

3.2. Implementación de los métodos

Para la implementación de los diferentes métodos de clasificación descritos en la sección 2.3.1, se requieren un clasificador semisupervisado binario y otro supervisado multiclase, para después combinarlos. Para el primero, se ha escogido SVMlight[2], y para el segundo, su derivado SVMmulticlass.

[1] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/

[2] http://svmlight.joachims.org

Basándose en ambos algoritmos, se han implementado los correspondientes métodos para el comportamiento *2 steps* supervisado y las técnicas *one-against-all*, *one-against-one* y *all-against-all* semisupervisadas.

Finalmente, además de los algoritmos comentados, se ha simplificado el algoritmo *2 steps* a un solo paso, *1 step*, donde utilizando únicamente un clasificador supervisado multiclase se entrena con los ejemplos etiquetados y se predicen los ejemplos de test, ignorando por tanto los ejemplos no etiquetados. Este método sirve para evaluar la aportación de los documentos no etiquetados en el aprendizaje.

3.3. Medidas de evaluación

La medida de evaluación escogida para el rendimiento de los algoritmos propuestos ha sido el "acierto" (accuracy), que es la que se viene utilizando en el área de la clasificación de textos. El acierto mide el porcentaje que comprende el número de predicciones correctas sobre el total de documentos testeados.

Se han considerado de la misma manera los aciertos sobre cualquiera de las clases, sin que ninguna de ellas tenga una mayor importancia respecto a las demás, por lo que no existe ponderación alguna en la evaluación.

4. Análisis de los resultados

En las figuras 2, 3 y 4 se muestran los resultados de la clasificación, en función del tamaño de la muestra etiquetada. Para cada una de estas muestras, se realizaron 9 ejecuciones, y se obtuvo la media de todas ellas, que es la que se representa en las gráficas.

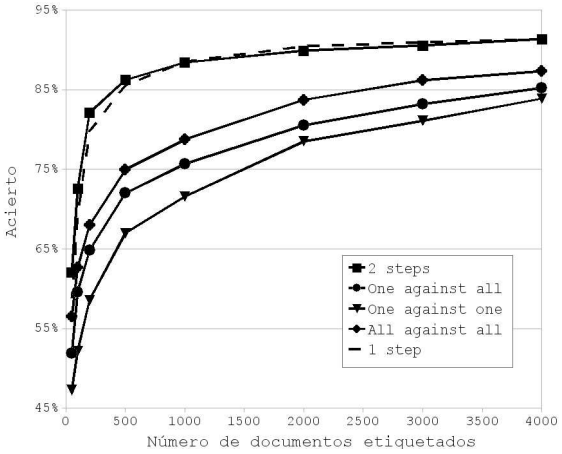

Figura 2: BankSearch

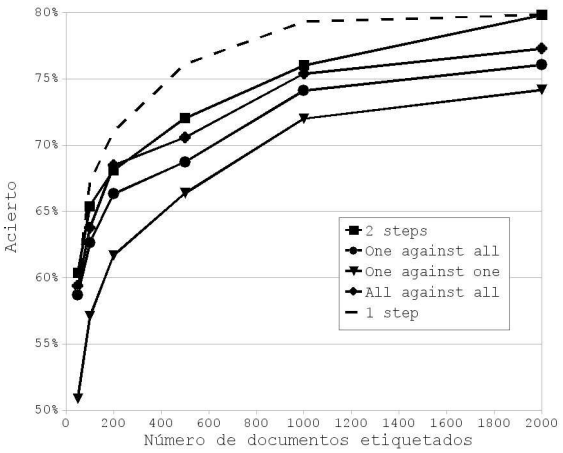

Figura 3: WebKB

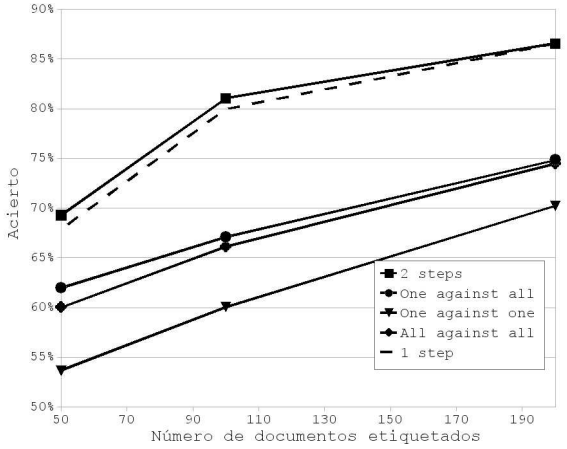

Figura 4: Yahoo! Science

Los resultados obtenidos pueden resumirse en las siguientes ideas:

- En todos los casos el mejor comportamiento se obtiene para uno de los algoritmos basados en clasificadores multiclase supervisados, bien sea el *1step* o el *2 steps*; incluso en los casos con menos documentos etiquetados, destacan sobre los basados en clasificadores semisupervisados binarios.
- De las tres técnicas semisupervisadas comparadas, destaca la propuesta *all-against-all*, ligeramente superior a *one-against-all*, y muy superior a *one-against-one*. Este último método, de hecho, demuestra que el ruido que se había previsto sí que existe, y que la calidad de los resultados obtenidos es baja.
- El método *1step*, que ignora los documentos no etiquetados para la fase de

aprendizaje, muestra unos resultados similares a los de *2 steps* para las colecciones *BankSearch* y *Yahoo! Science*, pero notablemente superiores para *WebKB*, donde las clases son más homogeneas. En este caso es donde mejor resulta ignorar los documentos no etiquetados, mediante el método *1step*.

- Para todas las colecciones, según se aumenta el número de documentos etiquetados, se mantiene el ranking obtenido por los algoritmos.

5. Conclusiones

En este trabajo se ha realizado un estudio comparativo de clasificación multiclase semisupervisada de páginas web mediante SVM. Se han introducido dos nuevas técnicas para S^3VM multiclase, *2 steps* y *all-against-all*, siendo los que mejores resultados han ofrecido. Además, se han aplicado por primera vez las técnicas *one-against-all* y *one-against-one* sobre clasificación semisupervisada, con unos resultados considerables para el primero de ellos, pero inferiores para el segundo.

Por otro lado, la no inclusión de documentos no etiquetados en la fase de aprendizaje, aplicada mediante la técnica *1step*, ha mostrado que en algunas ocasiones puede afectar de forma positiva. Ignorar los documentos no etiquetados para aprender ha resultado mejor cuando las clases son más heterogeneas. Queda pendiente, por tanto, su aplicación sobre otras colecciones más extensas y con más diversos temas.

Entre los algoritmos que combinan clasificadores binarios, *all-against-all* ha demostrado la mayor efectividad, aunque el gran número de clasificadores a considerar hace que su coste computacional aumente, por lo que su mejora en cuanto a eficiencia resultaría un interesante avance.

Como trabajo futuro, quedan por comparar los resultados respecto al algoritmo semisupervisado multiclase nativo.

Bibliografía

C.-H. Hsu y C.-J. Lin. 2002. *A Comparison of Methods for Multiclass Support Vector Machines.* IEEE Transactions on Neural Networks.

T. Joachims. 1998. *Text Categorization with Support Vector Machines: Learning with many Relevant Features.* Proceedings of ECML98, 10th European Conference on Machine Learning.

T. Joachims. 1999. *Transductive Inference for Text Classification Using Support Vector Machines.* Proceedings of ICML99, 16th International Conference on Machine Learning.

T. Mitchell. 1997. *Machine Learning.* McGraw Hill.

X. Qi y B.D. Davison. 2007. *Web Page Classification: Features and Algorithms.* Informe Técnico LU-CSE-07-010.

F. Sebastiani. 2002. *Machine Learning in Automated Text Categorization* ACM Computing Surveys, pp. 1-47.

M.P. Sinka y D.W. Corne. 2002. *A New Benchmark Dataset for Web Document Clustering.* Soft Computing Systems.

C.M. Tan, Y.F. Wang y C.D. Lee. 2002. *The Use of Bigrams to Enhance Text Categorization.* Information Processing and Management.

J. Weston y C. Watkins. 1999. *Multi-class Support Vector Machines.* Proceedings of ESAAN, the European Symposium on Artificial Neural Networks.

Z. Xu, R. Jin, J. Zhu, I. King y M. R. Lyu. 2007. *Efficient Convex Optimization for Transductive Support Vector Machine.* Advances in Neural Information Processing Systems.

Y. Yajima y T.-F. Kuo. 2006. *Optimization Approaches for Semi-Supervised Multiclass Classification.* Proceedings of ICDMW'06, the 6th International Conference on Data Mining.