do, donde todos los documentos de entrenamiento están previamente etiquetados, o *semisupervisado*, donde se aprende con una colección de entrenamiento compuesta por algunos documentos etiquetados y muchos no etiquetados.

En los últimos años, se han aplicado diferentes tipos de algoritmos al problema de la clasificación de textos (Sebastiani, 2002). Para esta tarea, las máquinas de vectores de soporte (SVM, Support Vector Machines (Joachims, 1998)) se han perfilado como una buena alternativa, que ofrecen, entre otras, las siguientes ventajas:

- No se requiere una selección o reducción de términos. En caso de que una clase se distribuya en áreas separadas del espacio vectorial, será la transformación del espacio mediante la función de kernel la que se ocupe de solucionarlo.
- No es necesario realizar un esfuerzo de ajuste de parámetros en el caso de problemas linealmente separables, ya que dispone de su propio método para ello.
- Su transformación a aprendizaje semisupervisado se convierte, generalmente, en un comportamiento transductivo, lo que posibilita el máximo refinamiento en la definición del clasificador.

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 de forma natural en un problema multiclase y semisupervisado. Por ello, y debido a su naturaleza binaria y supervisada, es necesaria una adaptación de la técnica SVM clásica. Existen diversos estudios referentes tanto a SVM multiclase como a SVM semisupervisado, pero apenas se ha investigado en la unión de ambos casos. Frente a una aproximación directa, basada en un problema de optimización complejo, este artículo propone y evalúa diferentes aproximaciones para la implementación de un método de SVM multiclase y semisupervisado, basándose en la combinación de clasificadores.

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. En la sección 3, se presentan las alternativas propuestas en este trabajo para clasificación semisupervisada multiclase. En la sección 4, se muestran los detalles de la experimentación realizada, para seguir en la sección 5 con el análisis de los resultados. En la sección 6, para finalizar, se exponen las conclusiones extraídas tras el proceso.

2. Clasificación con SVM

En la última década, SVM se ha convertido en una de las técnicas más utilizadas para tareas de clasificación, debido a los buenos resultados que se han obtenido. Esta técnica se basa en la representación de los documentos en un modelo de espacio vectorial, donde se 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 cada clase, maximizando la distancia entre los documentos y el propio hiperplano, lo que se denomina margen (ver Figura 1). Este hiperplano se define mediante la siguiente función:

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

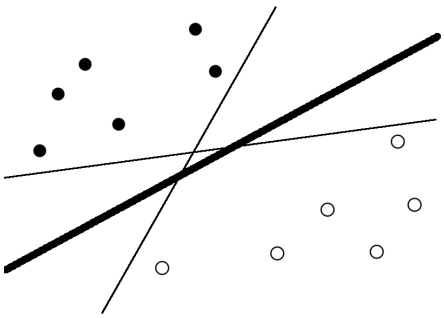

Figura 1: Ejemplo de maximización del margen con SVM, donde la línea más gruesa sería la escogida por el sistema.

La optimización de esta función supondría tener en cuenta todos los valores posibles para w y b, para después quedarse con aquéllos que maximicen los márgenes. Esto resulta muy difícil de optimizar, por lo que en la práctica se utiliza la siguiente función de optimización equivalente (ver Figura 2):

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

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

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

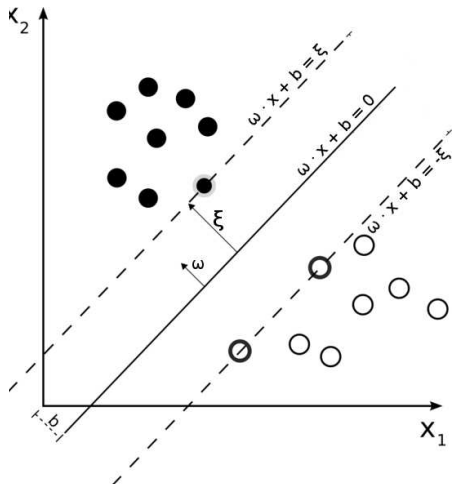

Figura 2: Representación gráfica de la función de clasificación de SVM.

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. Así, el nuevo espacio obtenido resultará 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 resolver problemas binarios y de forma supervisada.

2.1. SVM multiclase

Debido a la naturaleza dicotómica de SVM, surgió la necesidad de implementar nuevos métodos que pudieran resolver problemas multiclase, en los que la taxonomía está compuesta por más de dos clases. Como aproximación directa, (Weston y Watkins, 1999) proponen una modificación de la función de optimización que tiene en cuenta todas las clases, generalizando la función de optimización binaria para el número deseado k de clases:

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

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). Estas técnicas descomponen el problema multiclase en pequeños problemas binarios, aplicando después diferentes funciones de decisión para unirlos. Las técnicas más conocidas para clasificación mediante combinación de problemas binarios son las siguientes:

- *one-against-all* descompone un problema multiclase con k clases en otros tantos problemas binarios, en los cuales cada una de clases se enfrenta al resto. Así, se construyen k clasificadores que definen otros tantos hiperplanos que separan la clase i de los *k-1* restantes. Como función de decisión, a cada nuevo documento se le asigna aquella clase sobre la que su clasificador maximice el margen:

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

- *one-against-one* descompone el problema de k clases en $\frac{k(k-1)}{2}$ problemas binarios, donde se crean todos los posibles enfrentamientos uno a uno entre clases. Así, se obtiene un hiperplano para cada uno de estos problemas binarios. Posteriormente, se somete cada nuevo documento a todos estos clasificadores, y se añade un voto a la clase ganadora para cada caso, resultando como clase propuesta la que más votos suma.

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

Las técnicas de aprendizaje semisupervisado se diferencian en que, además de los documentos previamente etiquetados, se utilizan documentos no etiquetados para la fase de entrenamiento (Joachims, 1999) (ver Figura 3). Así, las predicciones del propio sistema sobre los documentos no etiquetados sirven, a su vez, para seguir alimentando el sistema de aprendizaje.

Las SVM semisupervisadas se conocen también por sus iniciales S^3VM. En el caso de SVM, su adaptación al aprendizaje semisupervisado supone a priori un gran coste

computacional, ya que la función resultante no es convexa, por lo que es mucho más complicada la optimización en busca del mínimo. Para relajar el cálculo 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, por lo que no se ha profundizado en el estudio sobre su aplicación a entornos multiclase.

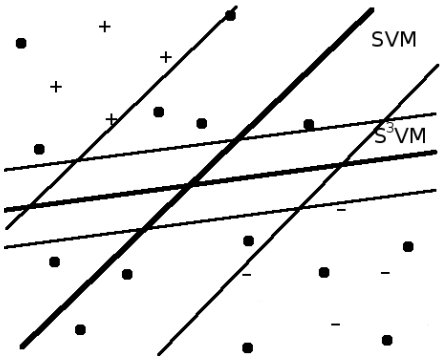

Figura 3: SVM vs S^{3}VM, donde los documentos etiquetados están representados por +/- y los no etiquetados por puntos.

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 características anteriormente expuestas, lo que supone un método de S^{3}VM multiclase. 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.

Actualmente, son pocos los trabajos que se han centrado en la transformación de SVM a semisupervisado y multiclase. Como aproximación directa, se encuentra la propuesta de (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.

Por otro lado, algunos trabajos han empleado otros enfoques para la consecución de una técnica S^{3}VM multiclase. (Qi et al., 2004) utilizan Fuzzy C-Means (FCM) para predecir la clase a la que pertenecen los documentos no etiquetados, tras lo cual utilizan SVM supervisado para aprender con la nueva colección ampliada, y clasifican el resto de documentos. (Xu y Schuurmans, 2005) utilizan una aproximación basada en clustering para la predicción de documentos no etiquetados, para posteriormente entrenar un clasificador SVM. (Chapelle et al., 2006), por último, presentan un método S^{3}VM multiclase basado en Continuation Method, y trasladan las técnicas basadas en combinación de binarios, *one-against-all* y *one-against-one*, al entorno semisupervisado. Aplican estas técnicas sobre colecciones de noticias, para las que obtienen unos resultados muy bajos. No obstante, estas técnicas nunca han sido trasladadas a la clasificación de páginas web.

3. Alternativas propuestas para S^3 VM multiclase

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

En cuanto a la utilización de documentos no etiquetados en fase de aprendizaje para SVM, (Joachims, 1998) presenta un estudio en el que se muestra una gran mejora cuando éstos son considerados para problemas binarios. No obstante, no se ha evaluado su apor-

tación en problemas multiclase, cuando las predicciones sobre un número mayor de clases pueden aumentar el error de forma considerable, perjudicando así a la fase de aprendizaje.

Realizamos dos tipos de propuestas alternativas a la aproximación directa para S^3VM multiclase. Por una parte, proponemos la utilización de técnicas ya empleadas en entornos supervisados, aunque sin un profundo análisis, y basados en la combinación de clasificadores binarios semisupervisados:

- *one-against-all-S^3VM* y *one-against-one-S^3VM* son propuestas basadas en la combinación de clasificadores binarios semisupervisados, vistos en la sección 2.1, que aunque se han utilizado en colecciones supervisadas, apenas han sido aplicadas y estudiadas sobre colecciones con documentos no etiquetados. Cabe destacar que el enfoque *one-against-one-S^3VM* 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 (Chapelle et al., 2006).

Por otra parte, introducimos dos nuevas técnicas para el desarrollo de un sistema de clasificación semisupervisado multiclase basado en SVM:

- *2-steps-SVM*: 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.
- *all-against-all-S^3VM*: Además de las anteriores, en este trabajo se presenta una nueva propuesta de combinación de clasificadores binarios, que hemos denominado *all-against-all-S^3VM*, y que podría ser utilizada tanto para aprendizaje supervisado como para 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 de la clasificación. 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 procede a la fase de predicción, en la que se asignará la clase para la que mayor votación ha obtenido cada documento. Aunque esta aproximación puede ser muy costosa para grandes taxonomías, ya que el número de clasificadores aumentaría de forma exponencial, se podría esperar un buen rendimiento para un número reducido de clases.

4. 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, en el que no se considerarán las categorías de algunos documentos, para así tener una colección semisupervisada,
- 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.

4.1. Colecciones de datos

Para esta experimentación se han utilizado colecciones de páginas web de referencia, que ya han sido utilizadas anteriormente para problemas de clasificación automática:

- *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 han utilizado los valores tf-idf de los unitérminos encontrados en los textos, excluyendo los de mayor y menor frecuencia. Los unitérminos resultantes han sido los que han definido las dimensiones del espacio vectorial.

4.2. Implementación de los métodos

Para la implementación de los diferentes métodos de clasificación descritos en la sección 3, se requiere un clasificador semisupervisado binario y otro supervisado multiclase, para después combinarlos. Para el primer caso, se ha escogido SVMlight[2], y para el segundo, su derivado SVMmulticlass. Basándose en ambos algoritmos, se han implementado los correspondientes métodos para el comportamiento *2-steps-SVM* supervisado y las técnicas *one-against-all-S^3VM*, *one-against-one-S^3VM* y *all-against-all-S^3VM* semisupervisadas.

Finalmente, además de los algoritmos comentados, se ha simplificado el algoritmo *2-steps-SVM* a un solo paso, *1-step-SVM*, 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.

4.3. Medidas de evaluación

La medida de evaluación escogida para el rendimiento de los algoritmos propuestos ha sido el "accuracy", ya que es la que suele utilizarse en el área de la clasificación de textos, sobre todo cuando el problema a tratar es multiclase. El "accuracy" mide el porcentaje 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.

5. Análisis de los resultados

En las figuras 4, 5 y 6 se muestran los resultados obtenidos durante la experimentación con las colecciones *BankSearch*, *WebKB* y *Yahoo! Science*, respectivamente. Estos resultados se presentan en forma de gráfica, en función del tamaño de la muestra etiquetada. Para cada una de las muestras se realizaron 9 ejecuciones. El valor que se representa en las gráficas es la media de todas las ejecuciones realizadas.

Los resultados obtenidos pueden resumirse en los siguientes puntos:

- En todos los casos el mejor comportamiento se obtiene para uno de los algoritmos basados en clasificadores multiclase supervisados, bien sea el *1-step-SVM* o el *2-steps-SVM*; incluso en los casos con menos documentos etiquetados, estos métodos destacan sobre los basados en clasificadores semisupervisados binarios.

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

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

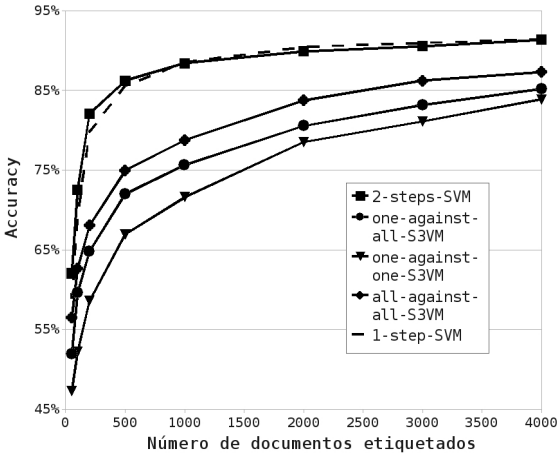

Figura 4: Resultados para BankSearch.

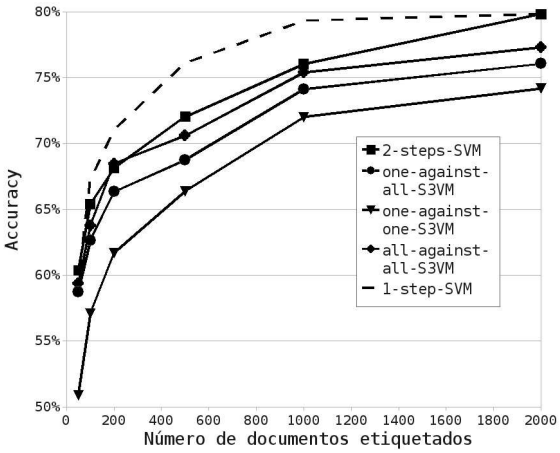

Figura 5: Resultados para WebKB.

- De las tres técnicas semisupervisadas comparadas, destaca la propuesta *all-against-all-S^3 VM* para las colecciones *BankSearch* y *WebKB*, ligeramente superior al de *one-against-all-S^3 VM*, y muy superior al de *one-against-one-S^3 VM*. Únicamente *one-against-all-S^3 VM*, en el caso de la colección *Yahoo! Search*, es algo superior a *all-against-all-S^3 VM*.
- La técnica *one-against-one-S^3 VM* demuestra que el ruido que se había previsto existe, y que, por ello, la calidad de los resultados obtenidos es baja.
- El método *1-step-SVM*, que ignora los documentos no etiquetados para la fase de aprendizaje, muestra unos resultados similares a los de *2-steps-SVM* 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 *1-step-SVM*, un método más sencillo y menos costoso computacionalmente que *2-steps-SVM*.
- Para todas las colecciones, según se aumenta el número de documentos etiquetados, se mantiene el ranking obtenido por los algoritmos.

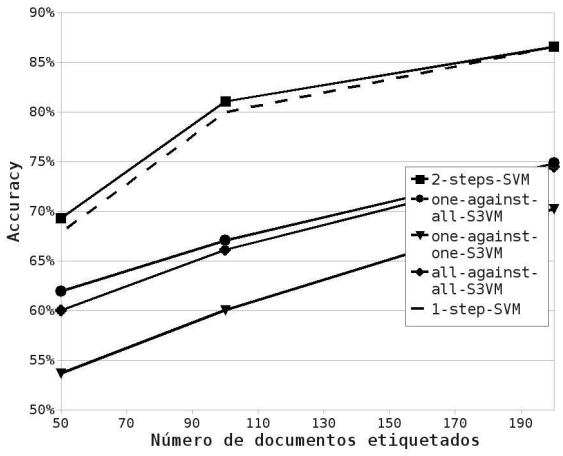

Figura 6: Resultados para Yahoo! Science.

6. 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, que hemos llamado *2-steps-SVM* y *all-against-all-S^3 VM*. El primero, *2-steps-SVM*, ha obtenido los mejores resultados en dos de las tres colecciones. Además, se han aplicado las técnicas *one-against-all-S^3 VM* y *one-against-one-S^3 VM* sobre clasificación semisupervisada, con unos resultados considerables para la primera, pero inferiores para la segunda.

Entre los algoritmos que combinan clasificadores binarios, *all-against-all-S^3 VM* 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.

A su vez, al igual que (Chapelle et al., 2006) muestran en sus resultados sobre colec-

ciones de noticias, los resultados sobre páginas web son también bajos, por lo que se confirma la baja efectividad de *one-against-all-S^3VM* y *one-against-one-S^3VM* para problemas semisupervisados multiclase.

Por otro lado, se ha estudiado la influencia de la no inclusión de documentos no etiquetados en la fase de aprendizaje, aplicada mediante la técnica *1-step-SVM*, y se ha mostrado que en algunas ocasiones puede influir de forma positiva. Ignorar los documentos no etiquetados para aprender ha resultado mejor cuando las clases son más homogéneas. Para las colecciones más heterogéneas, por otro lado, se han obtenido unos resultados parejos tanto considerando como ignorando los documentos no etiquetados. Estos resultados hacen pensar que para un problema multiclase y semisupervisado puede ser más interesante no utilizar datos no etiquetados, ya que los resultados son similares y el coste computacional es menor.

Por último, los resultados obtenidos en este trabajo complementan el estudio presentado por (Joachims, 1999), donde se muestra la superioridad de S^3VM respecto a SVM para problemas binarios. En el caso de un problema multiclase y semisupervisado de páginas web, la inclusión de documentos no etiquetados para problemas multiclase basados en SVM no resulta interesante para las colecciones testeadas, ya que una técnica supervisada obtiene, como mínimo, la misma efectividad para este tipo de entornos.

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

Bibliografía

O. Chapelle, M. Chi y A. Zien 2006. *A Continuation Method for Semi-supervised SVMs.* Proceedings of ICML'06, the 23rd International Conference on Machine Learning.

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.

H.-N. Qi, J.-G. Yang, Y.-W. Zhong y C. Deng 2004. *Multi-class SVM Based Remote Sensing Image Classification and its Semi-supervised Improvement Scheme.* Proceedings of the 3rd ICMLC.

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.

L. Xu y D. Schuurmans. 2005. *Unsupervised and Semi-supervised Multiclass Support Vector Machines* Proceedings of AAAI'05, the 20th National Conference on Artificial Intelligence.

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.