KNOWLEDGE ENGINEERING: PRINCIPLES AND TECHNIQUES
Proceedings of the International Conference on Knowledge Engineering,
Principles and Techniques, KEPT2009
Cluj-Napoca (Romania), July 2–4, 2009, pp. 61–64

FEATURE SELECTION IN TEXT CATEGORIZATION USING ℓ_1-REGULARIZED SVMS

ZSOLT MINIER[(1)]

Abstract. Text categorization is an important task in the efficient handling of a large volume of documents. An important step in solving this task is the removal of certain features of the text that is not necessary for high precision classification. An interesting and well-founded method of feature selection are embedded methods that work directly on the hypothesis space of the machine learning algorithms used for classification. One such method is ℓ_1-regularization that is used in conjunction with support vector machines. We study the effect of this method on precision of classiffying the 20 Newsgroups document corpus and compare it with the χ^2 statistic feature selection method that is considered one of the best methods for feature selection in text categorization. Our findings show, that the ℓ_1-regularization method performs about the same as the χ^2 statistic method.

1. Introduction

Text categorization is important in information retrieval, the field that lays the theoretical foundations of search engines. As the number of pages published on the internet is growing fast, there arises a need to categorize the pages in order to facilitate further information extraction.

Most modern text categorization systems are based on machine learning algorithms for supervised classification [4], which in turn have their roots in statistics. The basic building blocks of text categorization are text documents and a set of labels. The documents are assigned to possibly more than one label, this can be formalized as a function $f : D \to 2^C$ where D is the set of documents and 2^C is the superset of labels. This function is determined by a predefined set of document and label pairs (x_i, y_i) $(i = 1, \ldots, n)$ that is called the training set. The job of a text categorization system is to predict with high accuracy the label of a document that is not encountered in the training set, that is, to find the best approximation of f based on the available data.

Traditionally text categorization is viewed as the sequential composition of two separate tasks. The first task is to find a representation for text documents that can be efficiently stored. The second task is to use a machine learning algorithm

2000 *Mathematics Subject Classification.* 68P20, 68T30.

Key words and phrases. text categorization, feature selection, support vector machine, linear programming.

that can efficiently learn the representations of documents and has a good predictive performance on new documents.

It has been observed, that it is possible to remove some of the features from the representation of all documents without incurring a performance loss, thus reducing the amount of space necessary for storing the document data [6]. In the remainder of the paper we present a feature selection method based on the ℓ_1-regularized support vector machine.

2. Feature selection by ℓ_1 regularization

A support vector machine is a learning algorithm that is able to infer a decision rule from a set of training data and then by using this rule it is able to predict some properties of previously unseen data [1]. The main advantage of SVMs over similar learning algorithms is their good performance, robustness and relatively good speed.

Let us assume, that the data is given by tuples (x_i, y_i) where $x_i \in \mathbb{R}^d$ and $y_i = \pm 1$. The SVM finds the hyperplane (described by normal vector w and bias b) that separates positive and negative examples with the largest margin. Finding the hyperplane with maximal margin can be shown to be equal to the minimization of both the total loss over the training data and the complexity of the hyperplane that is measured by the norm of its normal vector:

$$(\hat{w}, \hat{b}) = \underset{w,b}{\operatorname{argmin}} \sum_{i=1}^{n} [1 - y_i(x_i' w + b)]_+ + \lambda \|w\|_2^2$$

where λ is called the regularization coefficient. Introducing $\|w\|_2^2$ into the minimization is called regularization because this way the the separating hyperplane is less prone to overfitting the noise in the data. This is achieved by upper bounding the length of its norm, excercising some control on the number of nonzero features. To achieve minimal length, the hyperplane has to disregard some of the less representative features in order to fit the more typical ones well. This minimization problem can be solved using quadratic programming.

In [5] and [2] it is suggested that using a ℓ_1 norm for SVMs results in sparse separating hyperplanes and thus the SVM formulation is slightly modified to:

$$(\hat{w}, \hat{b}) = \underset{w,b}{\operatorname{argmin}} \sum_{i=1}^{n} [1 - y_i(x_i' w + b)]_+ + \lambda \|w\|_1^1$$

This formulation of the SVM can be solved with linear programming. To do this, w has to be expressed with two positive vectors as $w = w^+ - w^-$ so that $|w| = w^+ + w^-$.

Then we can formulate the linear programming solution of the ℓ_1 SVM as:

$$\begin{aligned} &\min \sum_{i=1}^{n} \xi_i + \lambda \sum_{j=0}^{p} (w_j^+ + w_j^-) \\ s.t.\ y_i(b^+ - b^- + x^{'}(w^+ - w^-)) &\geq 1 - \xi_i \quad i \in \{1, \ldots, n\} \\ \xi_i &\geq 0 \quad i \in \{1, \ldots, n\} \\ w_j^+ \geq 0,\ \ w_j^- &\geq 0 \quad j \in \{1, \ldots, d\} \\ b^+ \geq 0,\ \ b^- &\geq 0 \end{aligned}$$

In this method, one can not explicitly set the number of variables one wants to keep, but one has some control over them by setting the appropriate λ, and experiments show that the hyperplanes are indeed very sparse.

3. Experiments

We used the 20 Newsgroups corpus for training and testing. During preprocessing stopwords are removed and stemming is performed, numbers are converted to the token “num” and special characters are deleted.

Transforming the two-class ℓ_1-SVM to multiclass is done by training classifiers for each pair of categories. To make solving that many linear programs easier, 5000 features are pre-selected with the χ^2 statistic method, among which the ℓ_1-SVM has to choose the best ones. Four different sizes were chosen for the training set, having 10, 50, 100, and 300 randomly selected documents for each category. Using the same selections, ℓ_1-SVM, χ^2, and no feature deletion was used for feature selection. The resulting documents were then learned by an ℓ_2-SVM with $\lambda = 1$ using the LIBSVM library [3]. For the studied algorithm, λ was set to be 1/3 of the number of constraints in each linear programming problem (or if there is no solution for that λ, then $\lambda = 1$), and we used the java binding of glpk to solve the LP problems. In the case of the χ^2 statistic, the number of features selected corresponded to the number of features that the ℓ_1-SVM found to be optimal. For every training set size and every feature selection algorithm 10 runs were performed. Mean and standard deviation of performance measures are shown in Table 1.

4. Conclusions

The results show, that the ℓ_1-SVM does induce a sparse model of the data, even if this model is not more efficient for categorization than the much simpler χ^2 statistic method.

It is interesting to note, that using all features, the ℓ_2-SVM achieves better precision than the feature selection methods, this is mostly due to the fact that this corpus has well balanced word distribution, and thus many features contribute to the overall precision of a classifier.

method	#d	#f	mP	mR	mBEP	mF1
ℓ_1-SVM	10	245.50±15.72	38.00±2.20%	37.61±2.13%	37.81±2.13%	37.80±2.13%
χ^2	10	245.50±15.72	40.05±1.63%	40.27±3.10%	40.16±2.22%	40.13±2.22%
full	10	6165.10±508.81	52.42±1.71%	45.58±2.81%	49.00±1.59%	48.70±1.73%
ℓ_1-SVM	50	679.40±18.84	59.61±0.59%	59.06±0.58%	59.34±0.58%	59.23±0.62%
χ^2	50	679.40±18.84	60.66±0.56%	60.23±0.60%	60.44±0.57%	60.44±0.58%
full	50	16697.70±2091.07	69.47±0.80%	61.86±1.92%	65.66±1.11%	65.43±1.21%
ℓ_1-SVM	100	1042.40±18.81	66.75±0.70%	66.09±0.77%	66.42±0.74%	66.42±0.74%
χ^2	100	1042.40±18.81	67.05±0.34%	66.45±0.31%	66.75±0.29%	66.75±0.29%
full	100	23788.30±1387.30	74.72±0.46%	66.53±0.98%	70.63±0.47%	70.38±0.53%
ℓ_1-SVM	300	1909.10±29.83	76.64±1.67%	75.55±0.83%	75.19±1.66%	76.09±1.22%
χ^2	300	1909.10±29.83	75.32±0.24%	74.70±0.35%	75.01±0.28%	75.01±0.29%
full	300	43042.60±1227.71	81.21±0.27%	76.61±1.19%	78.91±0.67%	78.84±0.70%

TABLE 1. Results obtained for the Reuters corpus given in percentage. Notation: #d=number of documents per category, #f=number of selected features, mP=micro-precision, mR=micro-recall, mBEP=micro-breakeven, mF1=micro-F_1

Acknowledgements

We would like to acknowledge the financial support of the grant CNCSIS TD-77/2007 by the Romanian Ministry of Education and Research.

References

[1] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[2] P.S. Bradley and O.L. Mangasarian. Feature selection via concave minimization and support vector machines. *Machine Learning Proceedings of the Fifteenth International Conference (ICML 98)*, pages 82–90, 1998.

[3] Chih-chung Chang and Chih-jen Lin. LIBSVM: a library for support vector machines (version 2.31), September 07 2001.

[4] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[5] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[6] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning*, 1997.

[1] FACULTY OF MATHEMATICS AND COMPUTER SCIENCE, BABEŞ-BOLYAI UNIVERSITY, 400084 CLUJ-NAPOCA, ROMANIA

E-mail address: `minier@cs.ubbcluj.ro`