Research Journal of Applied Sciences 2 (4): 431-434, 2007

SVM Based Multilevel Classifier Using Ontology

V. Uma and G. Aghila
Department of Computer Science, Engineering and Information Technology
Pondicherry Engineering College, Pondicherry, India

Abstract: A classical multilevel document classification system is vital in many contexts. This study explores the development of a multilevel document classification system based on Support Vector Machine (SVM) using ontology. SVMs have advantage over conventional statistical learning algorithms with features such as high generalization performance, prevention of over fitting, less computational complexity, high accuracy and robustness whereas the support of domain ontology further sharpens the classification by providing accurate required results. In this research SVM is used for implementing high level classification of the document and multi-level classification of the document is provided using Ontology. The comparison graph shows that the developed system based on ontology outperforms the existing system.

Key words: Document classification, support vector machine, Ontology

INTRODUCTION

Text categorization is a conventional classification problem applied to the textual domain. It solves the problem of assigning text content to predefined categories. SVM is a method for supervised learning, applicable to both classification and regression problems. SVM classifiers creates a maximum margin hyper plane that lies in a transformed input space and splits the example classes, while maximizing the distance to the nearest cleanly split examples.

Text Categorization with SVM has been proved to have achieved good performance (Joachims, 1998). In this paper Ontology based document classification has been performed as ontology has unique, hierarchical structure and characteristic of machine reasoning, starting from very primitive terms resulting in a more accurate document classification (Hung *et al*., 2003; Hee *et al*., 2005). Generally, Formal ontology deals with the interconnections of things, with objects and properties, parts and wholes, relations and collectives. Ontology-driven classification is a powerful technique which combines the advantages of modern classification methods with semantic specificity of the ontologies (Hee *et al*., 2005).

In this study the features of SVM are used for the classification of documents at multi-level with the aid of ontology. The classification of documents at multi-level helps in faster retrieval of documents and results in accurate classification.

RELATED WORK

A growing number of statistical classification methods have been applied to text categorization, such as Naïve Bayesian, Bayesian Network, Decision Tree Neural Network, Linear Regression, k-NN and Boosting. However, most machine learning methods over fit the training data when many features are given. Therefore, we need to select features carefully. Support Vector Machines (SVMs) is robust even when the number of features is large. Therefore, SVMs have shown good performance for text categorization (Joachims, 1998).

SVM is a supervised learning algorithm. Training data is given by

$$(x_1,y_1),\ldots,(x_u,y_u), \qquad x_j \in R^n,\ y_j \in \{+1, -1\}$$

Here, x_j is a feature vector of the j th sample; y_j is its class label, positive (+1) or negative (-1). SVM separates positive and negative examples by a hyper plane defined by

$$w \cdot x + b = 0,\ w \in R^n, b \in R \qquad (1)$$

The SVM determines the optimal hyper plane by maximizing the margin. By solving a quadratic programming problem, the decision function $f(x) = \text{sign}(g(x))$ can be derived where

$$g(x) = \left(\sum_{i=1}^{n} \lambda_i y_i (x_i \cdot x) + b\right). \qquad (2)$$

Corresponding Author: V. Uma, Department of Computer Science, Engineering and Information Technology, Pondicherry Engineering College, Pondicherry, India

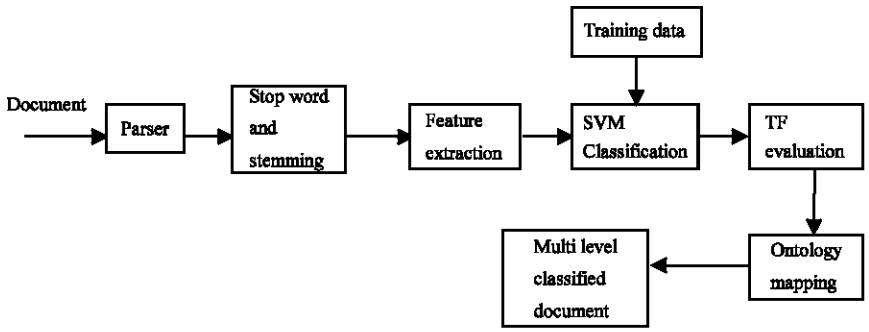

Fig. 1: System architecture

The decision function depends on only support vectors (x_i). Non-linear decision surfaces can be realized by replacing the inner product of (2) with a kernel function K (x. x_i) (Tsutomu *et al*., 2002; Simon and Dapne, 2001; Sato, 2004).

$$g(x) = \left(\sum_{i=1}^{n} \lambda_i y_i K(x_i, x) + b \right). \quad (3)$$

SVMs have high generalization performance independent of dimension of feature vectors (Taku and Yiji, 2000; Fabrizio, 2002). SVMs can carry out their learning with all combinations of given features without increasing computational complexity by introducing the Kernel function. SVMs use a large margin to prevent over fitting (Taku and Yiji, 2000). The experimental results show that SVMs consistently achieve good performance on text categorization tasks, outperforming existing methods substantially and significantly (Joachims, 1998; Yiming and Xin, 1999). SVMs are the most accurate classifier and fastest to train (Susan *et al*., 1998).

In spite of these advantages SVM has the disadvantage of very high training time (Chin *et al*., 2004). The training time for document classification can be considerably reduced by having background knowledge and hence a more comprehensive approach has been developed using the background knowledge available in the ontology. Ontology means information used in a specific domain and relationships defined in relation to the information. The advantages of an ontology-based classification approach over the existing ones, such as hierarchical and probabilistic approach, are that the nature of the relational structure of ontology provides a mechanism to enable machine reasoning (Hee *et al*., 2005).

SYSTEM ARCHITECTURE

In this study, a system that performs text classification at multi-level is explored. The architecture of the system is explained in Fig. 1.

The various modules involved in this system are:

Parser module: Text classification is carried out by transforming documents, which typically are strings of characters, into representations suitable for the learning algorithm and the classification task.

Stop word elimination and Stemming module: Stop word elimination is a process of removing the most frequent word that exists in a web page document. Removing these words will save spaces for storing document contents and reduce time taken during the search process. Stemming is a process of extracting each word from a web page document by reducing it to a possible root word. Porter Stemmer algorithm is used in implementing this module.

Feature extraction unit: Generally there are two types of features: format features and linguistic features (Hu *et al*., 2005). This feature Extraction module considers both format and linguistic features for extracting feature vectors.

SVM classification module: Each distinct word then corresponds to a feature vector. These feature vectors are given as input to the SVM classifier. The classified texts are interpreted and only those words that are classified positive are considered as key words.

Term frequency evaluation module: The term frequency of the generated keywords, in the document is found. A new data base is created in which the term frequency for the generated keywords is stored along with the key word.

Ontology mapping module: The key word that has maximum term frequency is extracted and is mapped on the words of ontology concepts that have been stemmed and sense-tagged. The documents are classified if there exists a positive mapping. Else the keyword with next high term frequency is considered for classification using the

mapping process. Based on the key words the mapping will lead to document classification at multi-levels.

IMPLEMENTATION

The implementation of the system is done in windows platform using JAVA. The SVM classification is done using LIBSVM which can perform multi class classification and has higher accuracy, efficient, fast and mainly provides multi class classification (LIBSVM). Ontology for computer domain is constructed from DMOZ directory. The document classification is done for the documents related to Computer Science. The computer science documents when fed into the system are further classified at various levels of classification. This system can be extended to other domains with minimal input such as knowledge about the domain which is an advantage of the system.

RESULTS

The performance of text categorization using SVM with the aid of domain ontology is measured using the following performance measures.

Precision: Number of correctly identified items as percentage of number of items identified.
Recall: Number of correctly identified items as percentage of the total number of correct items.
F-measure: Weighted Average of precision and recall.

$$\text{F-measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision} + \text{Recall}}$$

The performance of Naïve-Bayes, simple SVM is compared with this system and the graph was plotted for different training set size. The performance of NB classification is the worst, like expected. SVM result has reasonable accuracy, while SVM with the ontology mapping outperforms both. The use of ontology makes the classification more accurate. In this system the training size of the documents does not have much influence on the performance of the classification. This clearly shows that the training size need not be more for high performance which is an advantage (Fig. 2).

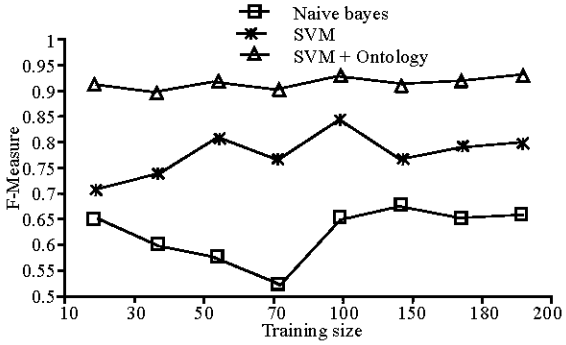

Fig. 2: Comparitive evalution

CONCLUSION

This study suggested an automated document classification method using the support vector machine using the concepts of domain ontology. Our research is distinguished from other studies in the following areas: The key words that are required for classification of the document are classified using SVM. The term frequency of these key words alone is found which reduces the processing time by a considerable amount. The training size of the documents is very less and hence training time is greatly reduced. The feature space dimension is also very less. Multi level classification of the documents is done using domain ontology which is more accurate. All these advantages make this system a promising method for document classification.

REFERENCES

Chih-Ming Chen, Hahn-Ming Lee, Ming-Tyan Kao, 2004. Multi-class SVM with Negative Data Selection for Web Page Classification, Proc. Int. Joint Conf. Neural Networks, IEEE.

Fabrizio Sebastiani, 2002. Machine Learning in Automated Text Categorization, ACM computing surveys, 34: 1.

Hu, Hang Li, Zheng and Meyerzon, 2005. Automatic Extraction of Titles From General Documents Using Machine Learning, JCDL.

Hung-Wu, Tzong-Han Tsai and Wen-Lian Hsu, 2003. Text Categorization using Automatically Acquired Domain Ontology, Proceedings of the sixth international workshop on Information retrieval with Asian languages, ACM.

LIBSVM- A library for Support Vector Machines, www.csie.ntu.edu.tw/~cjlin/libSVM

Mu-Hee Song, Soo-Yeon Lim, Dong-Jin Kang and Sang-Jo Lee, 2005. Automatic Classification of Web Pages based on the Concept of Domain Ontology, Proceedings of the 12th Asia-Pacific Software Engineering Conference (APSEC'05), Proceedings of IEEE.

Sato and Saito, 2004. Extracting Word Sequence Correspondences with SVM International Conference On Computational Linguistics, Proceedings of the 19th international conference on Computational linguistics, Vol. 1.

Simon Tong and Dapne Koller, 2001 Support Vector Machine Active Learning with Applications to Text Classification. Journal of Machine Learning Research.

Susan Dumais, John Platt, David Heckerman and Mehran Sahami, 1998. Inductive Learning Algorithms and Representations for Text Categorization, Proceedings of the seventh international conference on Information and knowledge management, ACM.

Taku Kudoh, Yuji Matsumoto, 2000. Use of Support Vector Learning for Chunk Identification, Proceedings of coNLL and LLL.

Thorsten Joachims, 1998. Text Categorization with Support Vector machines: Learning with Many Relevant Features, Proceedings of ECML, 10th European Conference on Machine Learning.

Tsutomu Hirao, Hideki Isozaki, 2002. Eisaku maeda, Extracting Important sentences with Support Vector Machines, Proceedings of the 19th International Conference on Computational Linguistics-Volume 1, ACM.

Yiming Yang and Xin Liu, 1999. A Reexamination of Text Categorization Methods" 22nd Annual SIGIR, ACM.