Multi-Category Support Vector Machines for Identifying Arabic Topics

Mourad Abbas, Kamel Smaili and Daoud Berkani

CRSTDLA, Speech Processing Lab.
1 rue D. E. Alafghani, Algeria
INRIA-LORIA, Parole team
B.P. 101, Villers les Nancy, France
Polytechnic School, Signal and Communication Lab.
10 rue H. Badi, Algeria
m_abbas04@yahoo.fr, kamel.smaili@loria.fr, dberkani@enp.edu.dz

Abstract. It is known that Support Vector Machines were designed for binary classification. Nevertheless, it would be fruitful to extend this operation to what is called Multi-category classification. That is why Multi-category Support Vector Machines (MSVM) become nowadays the current subject of several serious researches, aiming to achieve high levels of multi-category classification tasks. This technique has been assessed recently in some fields as text categorization, Cancer classification, etc. We should notify that experiments which have been realized until now using MSVM are limited to small data sets, since its computation is more expensive. In this paper we are interested in the use of this method, for the first time in topic identification. The experiments conducted concern topic identification of Arabic language. The corpora are extracted from Alwatan newspaper. Achieved results lead to an improvement of MSVM performance in comparison to the baseline SVM method. Nevertheless, SVM still outperforms MSVM when using larger sizes of the vocabulary.

1 Introduction

The main objective of topic identification is to assign one or several topic labels to a flow of textual data. Labels are chosen from a set of topics fixed a priori. Talking about topics conduct us to clarify the definition of a topic. In [1], each keyword is considered as a topic. Whereas in other works, topics are more sophisticated corresponding to specific subject, for example politics and sports [2]. In our case, we are dealing with six topics: Culture, Religion, Economy, Local news, International news and sports.
Topic identification is used in several areas: to adapt language models for speech recognition and for machine translation, to focus on a specific use for search engines,...etc. In spontaneous speech recognition process the vocabulary has to be as large as possible. Enlarging the vocabulary increases the search space and consequently could reduce system's performance.
A language model is one of the knowledge sources which is used by a speech

Advances in Computational Linguistics.
Research in Computing Science 41, 2009, pp. 217-226

Received 10/11/08
Accepted 15/12/08
Final version 17/02/09

recognition system, in order to find out the best hypotheses respecting linguistic criteria. One way to improve the results of a speech recognition system is to adapt the language model in accordance to the concerned utterance context. The problem of topic adaptation has already been largely addressed. In [3–8], topic information is exploited in different ways, resulting in a significant reduction of the perplexity of the baseline language model and sometimes in an improvement of the word error. Hence, these studies highlight the importance of topic adaptation.
Our objective is to identify one topic among a set of others. For that, six domains have been chosen to realize the related experiments. In this paper we will focus on the use of the MSVM based on the method developed by Guermeur [9–12]. To our knowledge, this is the first time when this method is used to identify topics. Obviously, several studies have been achieved for topic identification by using SVM or a combination of classifiers [13, 7]. The method proposed by Guermeur was initially used to combine protein secondary structure prediction methods. It leads to an enhancement of the prediction by nearly 2%, when compared to the three well-known individually used methods (Gor IV, Sopma and Simpa) [14–16]. In this paper, we will adapt it to our purpose: topic identification. In section 2 we give some information about representation of texts. In sections 3 and 4, a brief description of both SVM and Multi-category SVM are presented. And finally, section 5 describes the experiments conducted for the assessment of this method in the topic identification area.

2 Corpus Representation

Topic identification is based on topic training corpora, which represent the specificities of each topic. Training corpus has to be transformed. Each document d is transformed into a compact vector form. This operation is generally done after the tokenization of the corpus. The dimension of the vector corresponds to the number of distinct words or tokens in the vocabulary. Each entry in the vector represents the weight of each term. For our purpose, after removing the non content words, we calculated both the frequency of each word "Term Frequency", and the documents frequency of a word, which means the number of documents in which the word w occurs at least once [17]. A general vocabulary is constructed using word frequencies. It is based on the use of the Arabic newspaper corpus *Alwatan* which contains many thousands of news articles corresponding to nearly 10 millions words.

3 An Overview of SVM

Support Vector Machine is a supervised technique based on statistical learning theory. It is used for both classification and regression [18]. In classification, it is used to generate a class label from a set of features.
Let us consider a training set described by the couple $(x_i, y_i), i = 1...m$, where $x_i \in R^n$ and $y_i \in \{1, -1\}^m$, SVM requires a solution of the optimizing problem

given by the equation 1 [19, 20]:

$$min_{\omega,b,\xi}\ \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{m}\xi_i \tag{1}$$

subject to

$$y_i(\omega^T\phi(x_i) + b) \geq 1 - \xi_i \qquad \xi \geq 0$$

SVM approach consists in finding a linear separating hyperplane with a maximal margin in a higher dimensional space, in where, training vectors x_i are mapped by the function ϕ **.** C is the penalty parameter of the error which is also called the capacity of the model. Support Vector Machine is based on the so-called structural risk minimization inductive principle. The objective is to minimize an upper bound on the risk with respect to the parameters of the model [11]. This bound is composed of two terms: the empirical risk and the confidence interval. The last one is a function of the model capacity, which can be expressed in terms of different measures, the most common one is known as the VC (Vapnik-Chervonenkis) dimension [11, 21]. In order to minimize the risk in this case, these two terms are jointly minimized.

4 Multi-Categoy SVM

At first, multi-category Support Vector Machines had been realized through the so-called one-versus-rest strategy [22, 23]. After that, more methods have been introduced like the pairwise-coupling decomposition [24, 25] and the so-called k-class SVM proposed by Vapnik in [18].
According to [11] these methods cannot find a satisfactory compromise between training performance and complexity since they are not related to an explicit uniform convergence result, therefore they fail to implement the structural risk minimization principle. The used multi-class classification method which considers all classes at once has been implemented by Guermeur. It is based on the uniform strong law of large numbers.
Being faced to a k-category classification problem, with k superior or equal to 3, an architecture to perform the discriminant analysis is required. the idea is to consider all topics at once, and then many hyperplanes are calculated for the separation of all classes or topics in one step.
Let us consider a set of elements $x = \{x_j\}$ belonging to a subset of R^d. Each element x_j is labeled with the class C_i. i varies from 1 to k. A linear classification can be described as a set of functions f from R^d into R^k. f is then written as follows:

$$f(x) = \alpha x + b \tag{2}$$

$$\alpha = \begin{pmatrix} \alpha_1^T \\ \alpha_2^T \\ \vdots \\ \alpha_k^T \end{pmatrix} \quad and \quad b = \begin{pmatrix} b_1^T \\ b_2^T \\ \vdots \\ b_k^T \end{pmatrix}$$

Moreover, a non linear classification can be realized by introducing a kernel k_e which satisfies Mercer's conditions [26]. $f(x)$ is then given by equation 3:

$$f(x) = \begin{pmatrix} \langle \alpha_1, \phi(x) \rangle \\ \langle \alpha_2, \phi(x) \rangle \\ \vdots \\ \langle \alpha_k, \phi(x) \rangle \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix} \tag{3}$$

Where ϕ is a non linear function. The kernel can be defined by the expression 4:

$$\forall (x_1, x_2) \in R^d \times R^d,\ k_e(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle \tag{4}$$

More theoretical and practical studies about MSVM, can be found in [27, 12, 25, 28].

5 Experiments and Results

In order to realize our experiments we collected several thousands of articles from an on-line Arabic newspaper: Alwatan. The articles that we are interested in belong to the following topics: culture, religion, economy, local news, international news and sports. Nevertheless, some articles can be characterized by more than one topic. We should notify that as some topics are miscellaneous, they need to be subdivided to other subtopics to avoid performance degradations. In addition, the entire corpus need to be processed. Indeed, for the topic identification task, the non content words are not necessary, that is why we eliminated them. The size of the non content words attains 27 % of the entire corpus.
The size of our corpus is about 10 millions of words. It seems to be relatively small compared to corpora of Indo-European languages used in similar experiments. In fact, the size of the corpus extracted from the French newspaper "Le monde" of 4 years, is 80 millions words [13]. Whereas, the size of the corpus extracted from AFP Arabic Newswire of almost 7 years, and released in 2001 by LDC [29, 30] is 76 millions tokens. This gap between the two sizes is justified by the compact form of Arabic words.
In the two forthcoming subsections, we will show performances of the SVM method by realizing the well-known one-versus-rest approach, and compare results to MSVM ones.

5.1 One-versus-rest Approach

The one-versus-rest approach has been widely used to handle the multi-category problem. In our case we trained six one-versus-rest classifiers and assigned a new document d to the topic T_i giving the largest value of $g_i(d)$ for $i = 1, \ldots, 6$, where $g_i(d)$ is the SVM solution from training topic i versus the rest. Training topic i means documents covering the topic T_i. In this experiment, the corpus set dedicated to training contains 4000 documents, while vocabulary size is 3000 and each document contains 150 words. Using these data, we achieved a recall of 81.5% and a precision of 89.16% (see table 1). In this case, MSVM outperforms SVM by 4.33 % in term of recall "Table 5".

Table 1. Performances by using the one-versus-rest method with a vocabulary size 3000

Topics	Recall (%)	Precision (%)
Culture	80	92.31
Religion	53.33	100
Economy	80	92.31
Local news	93.33	70
International news	86.67	92.86
Sports	93.33	87.5
Average	81.11	89.16

Nevertheless, we should point out that SVM leads to better scores when using large corpora and bigger sizes of the vocabulary. Consequently, as it is computationally easier to do one-versus-rest classification using SVM, many related works have been carried out. Indeed, in [31] SVM has been used to identify topics of Arabic texts with a vocabulary size of 43000 words. The resulted Recall and Precision rates are respectively 97.26 % and 98.52%.

5.2 MSVM Experiments

In order to know if the topics number has either a slight or an important influence on results, we preferred starting by identifying three topics: Culture, Religion and Economy. After that we achieved additional experiments by increasing the number of topics to six. Each test document contains a number of words equal to 120. The vocabulary size is 1000 words. We should notify that for all conducted experiments we have attributed 150 documents per topic in the training phase. Tables 2 and 3 show respectively performances of MSVM and SVM methods.

Performance of SVM is largely inferior to MSVM one. In this case the difference, in terms of Recall, is about 20 %. Nevertheless, in forthcoming experiments, we will see that SVM outperforms MSVM when using a vocabulary size equal to 8000 (see table 5).

The identification of three topics using MSVM yields a good performance, 91.66%

Table 2. Performances of MSVM using a vocabulary size 1000, "Identification of three topics".

Topics	Recall (%)	Precision (%)
Culture	90	86
Religion	95	95
Economy	90	90
Average	91.66	90.33

Table 3. Performances of SVM using a vocabulary size 1000, "Identification of three topics".

Topics	Recall (%)	Precision (%)
Culture	80	91.66
Religion	53.33	100
Economy	80	92.33
Average	71.11	94.66

in term of recall. Results decrease when we augment the number of topics to six, with a vocabulary size 1000 "see table 4". Indeed, performances in terms of Recall corresponding to the topics "Economy, International news, Sports, Religion" vary from 80 % to 92%. The other two topics caused an important degradation of the mean result, in fact, their Performances correspond respectively to 66 % and 60 %.
MSVM has a good theoretical background [27], and results should be better than what we achieved. However, many reasons are behind low performances of the aforementioned two topics. As we cited in the previous subsection, the two last ones cover other subtopics and then necessitate to be subdivided. Increasing the training corpus size is also an important factor which contributes to enhance results.

Table 4. Performances of MSVM using a vocabulary size 1000, "Identification of six topics".

Topics	Recall (%)	Precision (%)
Culture	66	60
Religion	92	96
Economy	80	74
Loc. news	60	64
Int. news	82	84
Sports	86	91
Average	77.66	78.16

For that, we conducted more experiments to improve MSVM performance. We used different vocabulary sizes. The size Documents varies from 120 to 150 words and the training corpus is composed of 4000 articles. In table 5 we summarize performances in terms of Recall (R) and Precision (P) resulted from the four realized experiments.

Table 5. Performances of MSVM by using different vocabulary sizes "Identification of six topics".

	Exp1	**Exp2**	**Exp3**	**Exp4**
Vocab. size	1000	2000	3000	8000
R(%) (MSVM)	77.66	80.41	85.55	89.75
P(%) (MSVM)	78.16	82.66	85.44	88.32
R(%) (SVM)	76.50	78.33	81.11	93.45
P(%) (SVM)	80.12	79.56	89.16	90.44

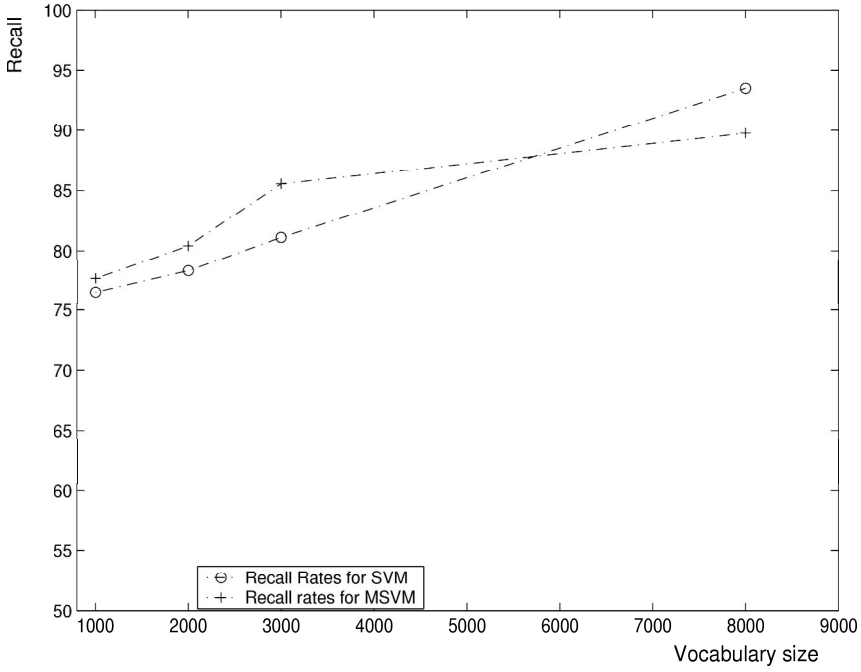

Fig. 1. Recall rates versus vocabulary size, for SVM and MSVM

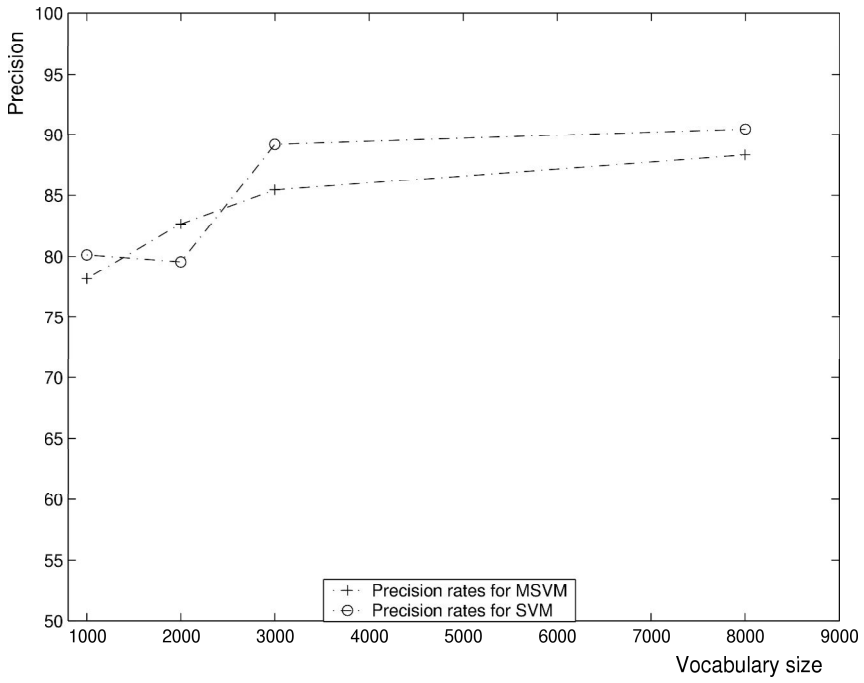

Fig. 2. Precision rates versus vocabulary size, for SVM and MSVM

According to results shown in table 5, and illustrated by figures 1 and 2 it is clear that the vocabulary size improve performances. Indeed, for MSVM we notice that recall varies from 77.66% to 89.75% according to the vocabulary size, while performances of SVM are less than MSVM, except for the size 8000 where SVM gave better results. We can consider MSVM results as an encouraging step in accordance with the computation complexity of the method.

6 Conclusion

In this paper we focused on identifying six topics for Modern Standard Arabic, using a new multi-class classification method based on a uniform convergence result. In fact, this was realized using a method proposed by Guermeur [12]. This is the first use of this method in an important domain like topic identification. For real applications, MSVM is more suitable than SVM, since we do not need to do many binary decisions. In our case we had not to calculate each time the optimal hyperplane to separate two topics. Indeed, when using MSVM, the idea was to consider all topics at once, and then many hyperplanes are calculated for separating all topics in one step.
Due to the computational complexity of the used method we were not able to

use a larger vocabulary. In perspective we aim to improve this result by finding a way to overcome the computation complexity, since experiments showed that increasing the size of training corpora and also that of vocabulary always lead to satisfactory results.

References

1. Seymore, K., Rosenfeld, R.: Using story topics for language model adaptation. In: Proceeding of the European Conference on Speech Communication and Technology, Rhodes, Greece (1997)
2. Yamashita, Y., Tsunekawa, T., Mizoguchi, R.: Topic recognition for news speech based on keyword spotting. In: IEEE International Conference on Spoken Language Processing, Sydney, Australia (1998)
3. Martin, S., Liermann, J., Ney, H.: Adaptive topic-dependent language modelling using word-based varigrams. In: 3rd European Conference on Speech Communication and Technolog, Rhodes, Greece (1997)
4. Mahajan, M., Beeferman, D., Huang, X.: Improved topic-dependent language modeling using information retrieval techniques. In: Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing. (1999)
5. Yang, Y.: An evaluation of statistical approaches to text categorization. Information Retrieval **1** (1999) 69–90
6. Bigi, B., De Mori, R., El-Bèze, M., Spriet, T.: A fuzzy decision strategy for topic identification and dynamic selection of language models. Special Issue on Fuzzy Logic in Signal Processing, Signal Processing Journal **80** (2000)
7. Bigi, B., Brun, A., Haton, J., Smaili, K., Zitouni, I.: Dynamic topic identification: Towards combination of methods. In Recent Advances in Natural Language Processing (RANLP), Tzigov Chark, Bulgary (2001) 255–257
8. Brun, A., Smaïli, K., Haton, J.: Contribution to topic identification by using word similarity. In: International Conference on Spoken Language Processing (ICSLP2002), Denver, USA (2002)
9. Guermeur, Y.: Technical documentation of the multi-class SVM. Technical report, Loria, France (2004)
10. Guermeur, Y.: Combining discriminant models with new multi-class SVMs. Technical Report NeuroCOLT2, 2000-086, Loria, France (2000)
11. Guermeur, Y., Elisseeff, A., Paugam-Moisy, H.: A new multi-class svm based on a uniform convergence result. In: International Joint Conference on Neural Networks (IJCNN00). Volume IV., Come (2000) 183–188
12. Guermeur., Y., Pollastri, G., Elisseeff, A., Zelus, D., Paugam-Moisy, H., Baldi, P.: Combining protein secondary structure prediction models with ensemble methods of optimal complexity. Neurocomputing **56** (2004) 305–327
13. Brun, A.: Détection de thème et adaptation des modèles de langage pour la reconnaissance automatique de la parole. PhD thesis, Henri Poincaré University, Nancy1 (2003)
14. Garnier, J., Gibrat, J.F., Robson, B.: GOR method for predicting protein secondary structure from amino acid sequence. Methods Enzymol **266** (1996) 540–553
15. Geourjon, G., Deleage, G.: SOPMA: significant improvments in protein secondary structure prediction by consensus prediction from multiple alignments. CABIOS **11** (1995) 681–684

16. Levin, J.M.: Exploring the limits of nearest neighbour secondary structure prediction. Protein Eng. **10** (1997) 771–776
17. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: European Conference on Machine Learning (ECML), Chemnitz, Germany (1998) 137–142
18. Vapnik, V.N.: Statistical learning theory. John Wiley & Sons, Inc., N.Y. (1998)
19. Boser, B., Guyon, I., Vapnik, V.: A training algorithm for optimal margin classifiers. In: the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh (1992) 144–152
20. Cortes, C., Vapnik, V.: Support-vector network. Machine Learning **20** (1995) 273–297
21. Vapnik, V., Chervonenkis, A.Y.: On the uniform convergence of relative frequencies of events to their probabilities. Theory of probability and its applications **16** (1971) 264–280
22. Scholkopf, B., Burges, C., Vapnik, V.: Extracting support data for a given task. In: ICKDDM'95, Menlo Park, CA, AAAI Press (1995) 252–257
23. Vapnik, V.: The Nature of Statistical Learning Theory. Spinger, New York (1995)
24. Mayoraz, E., Alpaydin, E.: Support vector machines for multi-class classification. Technical report, IDIAP (1998)
25. Weston, J., Watkins, C.: Multi-class support vector machines. Technical report, Royal Holloway, University of London, Department of Computer Science (1998)
26. Aizerman, M., Braverman, E., Rozonoer, L.: Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control **25** (1964) 821–837
27. Lee, Y., Lin, Y., Wahba, G.: Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. Journal of the American Statistical Association **99 No. 465** (2004) 67–81
28. Bredensteiner, E.J., Bennet, K.P.: Multicategory classification by support vector machines. In: Computational Optimizations and Applications. Volume 12. (1999) 53–79
29. Abdelali, A., Cowie, J.: Regional corpus of modern standard arabic. In: second international conference on Arabic Language Engineering. Volume 1., Algiers, Algeria (2005) 1–12
30. Abdelali, A., Cowie, J., Soliman, H.: Building a modern standard corpus. In: Workshop on Computational Modeling of Lexical Acquisition, The Split Meeting, Split (2005)
31. Abbas, M., Smaili, K.: Comparison of topic identification methods for arabic language. In: Recent Advances in Natural Language Processing RANLP05, Borovets, Bulgaria (2005) 14–17