Chinese Question Classification with Support Vector Machine

Bo Liu [1+], Zhifeng Hao [2], Xiaowei Yang [2] and Xudong Lin[1]

[1]College of Computer Science & Eng., South China Univ.of Technology, Guangzhou 510640. P.R. China
[2]School of Mathematical Science, South China Univ. of Technology, Guangzhou 510640, P. R. China

Abstract. Question classification plays a crucial important role in the question answering system because categorizing a given question is beneficial to identify an answer in the documents. It is the basic and important module of question answering which task is to assign one or several classes to a given question; the errors of question classification will probably result in the failure of question answering. In recent years, support vector machines have been introduced for solving pattern recognition problems because of their superior performance. In this paper, we will perform the One-against-All Algorithm with support vector machine for Chinese question classification problems. The results show the excellent general performance of the Algorithm.

Keywords: Question Classification; Feature Extraction; Semantic Dependency Relationship; Support Vector Machine; One-against-All Algorithm.

1. Introduction

Question classification (QC) [1] systems play an important role in question answering systems and can be used in a wide range of other domains. The goal of question classification is to accurately assign labels to questions based on expected answer type. With the rapid growth of text available on the Internet, it has become more difficult for users to find special information [2]. The traditional method of querying an Internet search engine often returns thousands of results, containing a ranked list of documents along with their partial snippets [3]. For an average Internet user, it is often time-consuming and laborious to find requested information. Often to find the searched information, a user has to connect to several servers and scan through dozens of documents to locate it [4]. We think that for a human being the most natural and straightforward approach to such a task is to ask a question in a natural language way. The output result should be a correct answer directly. Web-based question answering system (WQAS) can solve this problem. It is a hot topic in information retrieve and information extraction [5].

In recent years, support vector machines (SVMs) have been introduced for solving pattern recognition problems because of their superior performance [6], [7], [8]. The SVMs are developed based on the idea of structural risk minimizations (SRM), which guarantee the good general performance of the method. The standard support vector machines (SVM) [6] were originally designed for binary classifications. Unfortunately, many practical applications consist of multi-classification problems, which are usually converted into binary ones [9].

In the previous work, Vapnik first proposed One-against-All algorithm with discrete decision functions [6] and continuous decision functions [7]. Inoue and Abe [10] presented a fuzzy support vector machine for One-against-All classification, in which the fuzzy membership function is defined instead of the continuous decision function. Later, Abe [11] proved that One-against-All with the continuous decision function is equal to One-against-All with the fuzzy decision function. KreBel [12] converted the $C-$class problem into $C(C-1)/2$ two-class problems, which is called pairwise classification (One-against-One). In this case, the middle unclassifiable region still remained. In order to deal with the middle unclassifiable region, Platt et al. [13] proposed decision-tree-based pairwise classification called Decision Directed Acyclic Graph (DDAG), in which the unclassifiable region is given to one class based on the architecture of the DDAG. Tsujinishi and Abe [14] proposed a fuzzy support vector based on pairwise classification, in which the average membership function or the minimum membership function is defined. To effectively solve the unclassifiable region, Liu et al. [15] proposed the nesting support vector machine for multi- classification, in which, the authors proved the validity of the proposed algorithm for unclassifiable region and give the computational complexity analysis of the method. Later, Liu et al [16] developed this algorithm and give the detail idea of it. Recently, Liu et al. [17] introduce binary tree algorithm based on the kernel fisher discriminante.

In order to improve the accuracy of the multi-classification, Angulo et al. [18] introduced the support vector classification-regression machine for $C-$class classification, in which a new training algorithm with ternary outputs $\{-1,0,+1\}$ is given based on the Vapnik's Support Vector theory. Recently, Hao and Liu [19] proposed twi-map support vector machine algorithm based on the One-against-One approach. Later, Liu et al. [20] introduce Quadratic Map Support Vector Machine Based on One-against-All for Multi-classification, in which, the authors firstly applied the support vector machine into supervised feature extraction.

Question classification plays a crucial important role in question answering systems because categorizing a given question is beneficial to identify an answer in the documents. In this paper, we will perform the multi-classification algorithm with support vector machine to resolve the Chinese question classification. From the [9], we can easily conclude that the One-against-All algorithm is also excellent as the One-against-One approaches; So It will be adopted in the paper. As for feature extraction, we will utilize multi-method, such as keyword extraction, bag of words, head phrase, syntactic features, and semantic features, to extract question features.

The rest of this paper is organized as follows: In Section 2, we review the One-against-All algorithm. Perform experiments will be shown in Section 3, in this Section, we also detail describe Chinese question classification 、 Feature Extraction and the experiments results will be reported. Acknowledgements will be given in Section 4.

2. One-against-All Algorithm

Considering the conventional support vector machine introduced by Vapnik [6], one needs to determine C decision functions for the C-classes. The optimal hyperplane for class i against the remaining classes which has the maximum margin between them is

$$D_i(\mathbf{x}) = w_i^T \Phi(\mathbf{x}) + b_i = 0. \quad (1)$$

where w_i^T is an m-dimensional vector , $\Phi(\mathbf{x})$ a mapping function and b_i is a scalar. After the approach, we will obtain C hyperplanes.

2.1 Discrete Decision Function discriminance

Vapnik [6] first put forward the discrete function discriminance similar with the binary classification. If for the input vector $\mathbf{x}$

$$D_i(\mathbf{x}) > 0 \tag{2}$$

Satisfies for one i, $\mathbf{x}$ is classified into class i. By this formulation, only the sign of the decision function is used, so the decision is discrete.

If (2) is satisfied for plural $i's$, or there is no i which satisfies (2), then $\mathbf{x}$ is unclassifiable as shown in Fig. 1 (the shaded regions are unclassifiable regions).

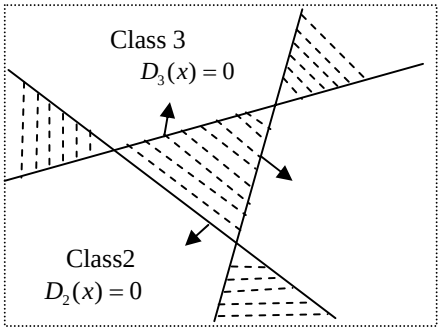


Fig.1. Discrete Decision Function discriminance based on the one-against-all.

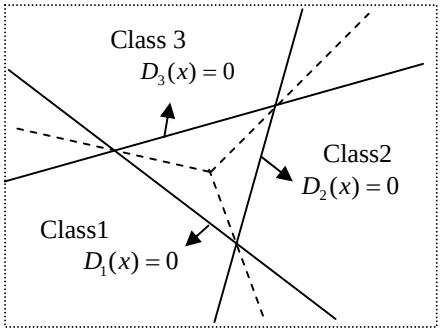


Fig.2. Continuous Decision Function discriminance based on one-against-all.

2.1 Continuous Decision Function discriminance

In order to avoid the unclassifiable regions shown in Fig. 1 and enhance the accuracies of One-against-All, Vapnik [7] proposed the continuous decision function discriminance.

For the input vector $\mathbf{x}$, it will be classified into the class:

$$\underset{i}{\operatorname{argmax}}\ D_i(\mathbf{x}) \tag{3}$$

Since the value of $D_i(\mathbf{x})$ is continuous, by this formulation the decision function is continuous and the unclassifiable regions are resolved as shown in Fig. 2.

3. Performance Evaluation

According to [9], we can obtain that One-against-All algorithm perform just as well as One-against-One scheme. So, in this section, we will utilize the One-against-All algorithm to resolve the Chinese question classification problems.

3.1 Chinese Question Classification

Question classification means putting the questions into several semantic categories [21]. The data set in this real problem consists of the questions corpus which is provided by HIT-IRLab. The question corpus contains 4394 Chinese questions and follows the two-layered question taxonomy, which contains 6 coarse categories and 65 fine categories, as listed in Table 1.

Table 1 The coarse and fine grained question categories

Coarse	Fine	Coarse	Fine
DESC	ABBR, DEFINITION, DEFINITION, MEANING, REASON, OTHER	NUM	AGE, AREA, CODE, COUNT, DISTANCE, FREQUENCY, ORDER, PERCENT, PHONENUMBER, POSTCODE, PRICE, RANGE, SPEED, TELCODE, TEMPERATURE, WEIGHT, LIST, OTHER
HUM	ALIAS, DESCRIPTION, ORGANIZATION, PERSON, LIST, OTHER	OBJ	ANIMAL, CITY, COLOR, CURRENCY, ENTERTAIN, FOOD, INSTRUMENT, LANGUAGE, PLANT, RELIGION, SUBSTANCE, VEHICLE, LIST, OTHER
LOC	ADDRESS, CITY, CONTINENT, COUNTRY, COUNTY, ISLAND, LAKE, MOUNTAIN, OCEAN, PLANET, PROVINCE, RIVER, LIST, OTHER	TIME	DAY, MONTH, RANGE, TIME, YEAR, LIST, OTHER

3.2 Feature Extraction

There are many research results on how to extract features from documents, such as term weighting, co-occurrence of words for topic identification, and keyword extraction using term domain interdependence are all statistical methods [22]. For the problem of question classification, statistical techniques appeared less suitable as a single question is normally very short and hence, does not contains enough words to allow the creation of meaningful statistics.

In the problem, we used the following six steps to generate a fixed–length binary feature vector for each Chinese question.

1. **Chinese Words Segmentation**: Computer automatic words segmentation is a particular research subject in Chinese information processing. It is the foundational work of machine translation，natural language understanding and information retrieval according to the feature of Chinese. In this experiment, we use ICTCLAS (provided by Software Research Lab, Institute

of Computing Tech., Chinese Academy of Sciences) to split the Chinese words, which performs very well in Chinese words segmentation and gives part-of-speech of tagging for each word.

2. **Keyword Extraction**: Keywords are selected from the pre-classified question set provided by HIT-IRLab. After removing stop-words from all question instances in out training set, we add up the frequency of each word, such as noun, verb, adverb and adjective. From each question category we extract keywords that appear in more than 0.1% and less than 98% of the questions in one category.

3. **Bag of Words**: In order to archive high accuracy, we used HowNet (which is an on-line common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents) to group words in a question sentence to generate more meaningful chunks as bag of word.

4. **Head Phrase**: In this experiment, we extract the head phrase of each question sentence as a feature because it plays an important role in defining the type of the question. The head phrase may appear in the beginning, middle or end of the Chinese question. Usually it contains one, two or more of Chinese characters. So it is difficult to locate the head phrase in the Chinese question. Here, we use the syntactic feature to realize it.

5. **Syntactic Features**: Syntactic features are used to represent the syntax of a question [23]. They are so appealing for questions of the same type often have the same syntactic style. That is, they often share a similar structure and vocabulary. Based on part-of-speech of tagging of each word, we try to find out the subject, predicate, object, adjective, adverbial, etc, from the Chinese question.

6. **Semantic Features**: It is possible to achieve reasonable results only using syntactic features; However, as for some questions, such as what questions, are often incorrectly classified when syntactic features are used alone [24], [25]. Especially, Chinese grammar is irregular, obscure and the Chinese semantic features can provide much more information for question classification. HowNet is a powerful natural processing and linguistic tool which is a Chinese lexical database that provides a wealth of semantic information [26]. In the experiment, we use the 76 semantic dependency relationships defined in it.

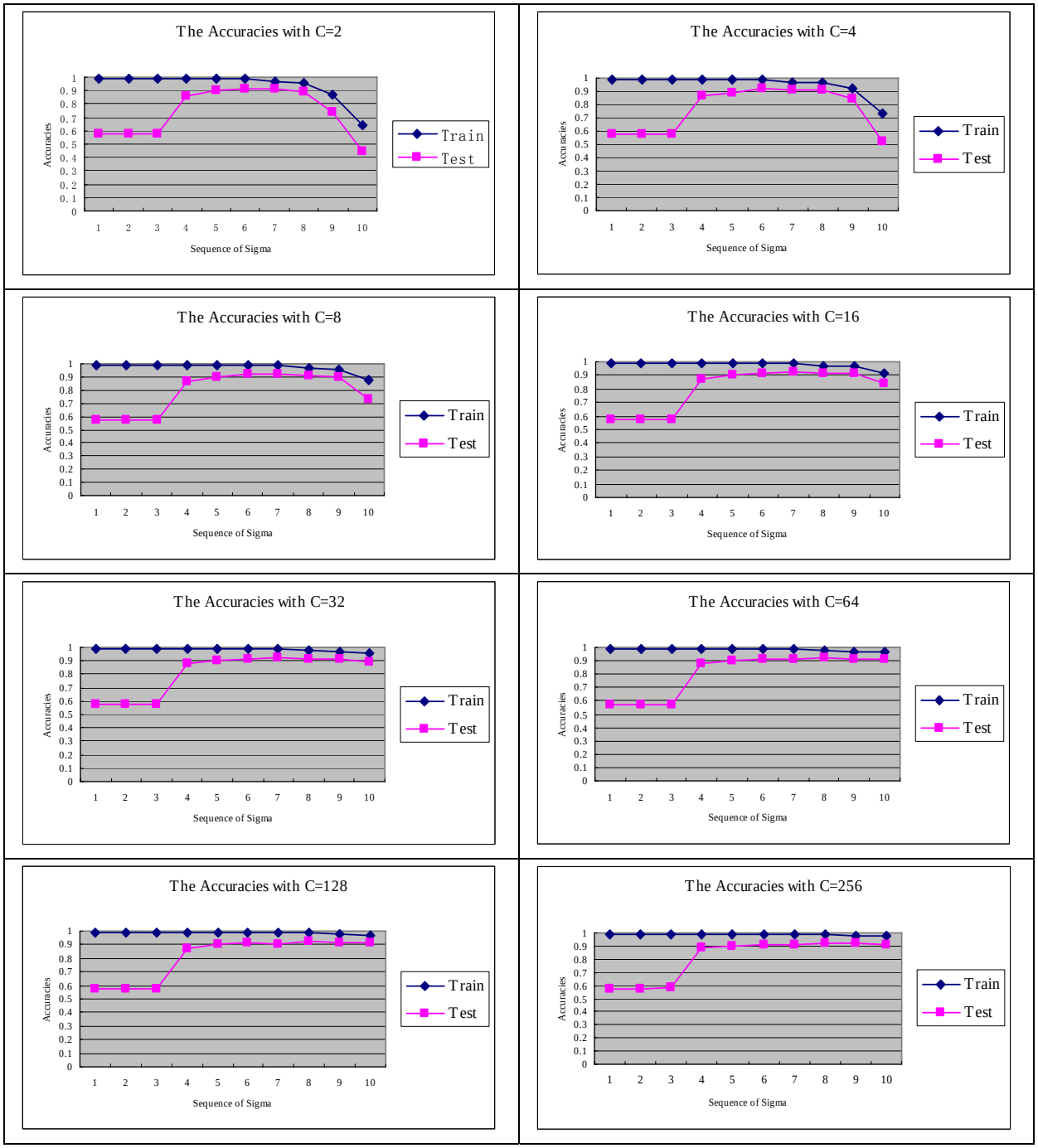
3.3 Simulation Experiment

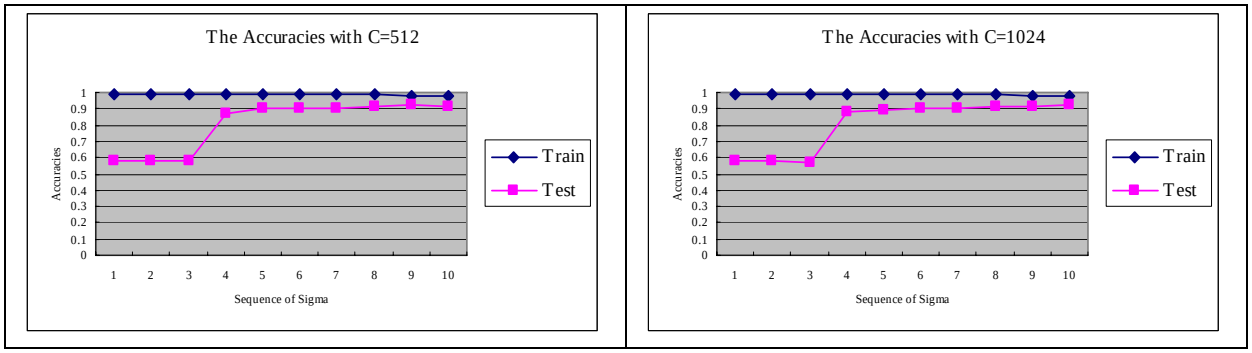
The experiments are run on a PC with a 2.8GHz Pentium IV processor and a maximum of 512MB memory. All the programs are written in C++, using Microsoft's Visual C++ 6.0 compiler. In order to solve the real problems and evaluate the validity of the Nesting algorithm, we split the data set into training set and testing set, selecting randomly 452 (80%) data points for training, and the remained 113 (20%) samples for testing.

There exist so many kernel functions; in the experiment, we select the RBF kernel function which is the most popular kernel function in dealing real problems. As for the hyperparameters, we estimate the generalized accuracy using different kernel parameters σ and cost parameters C : $\sigma = [2^{-3}, 2^{-2}, 2^{-1}, \cdots, 2^{6}]$ and $C = [2^{1}, 2^{4}, 2^{5}, \cdots, 2^{10}]$. By this way, for a given problem we need try 100 combinations to obtain the best hyperparameters.

The results of computation are listed in the Table1, which involves ten figures. Each figure describes the training and testing accuracies of the given problem when fixing the parameter C and let the parameter σ vary in the range of $[2^{-3},2^{-2},2^{-1},\cdots,2^{6}]$.

Table1. The results of Computation





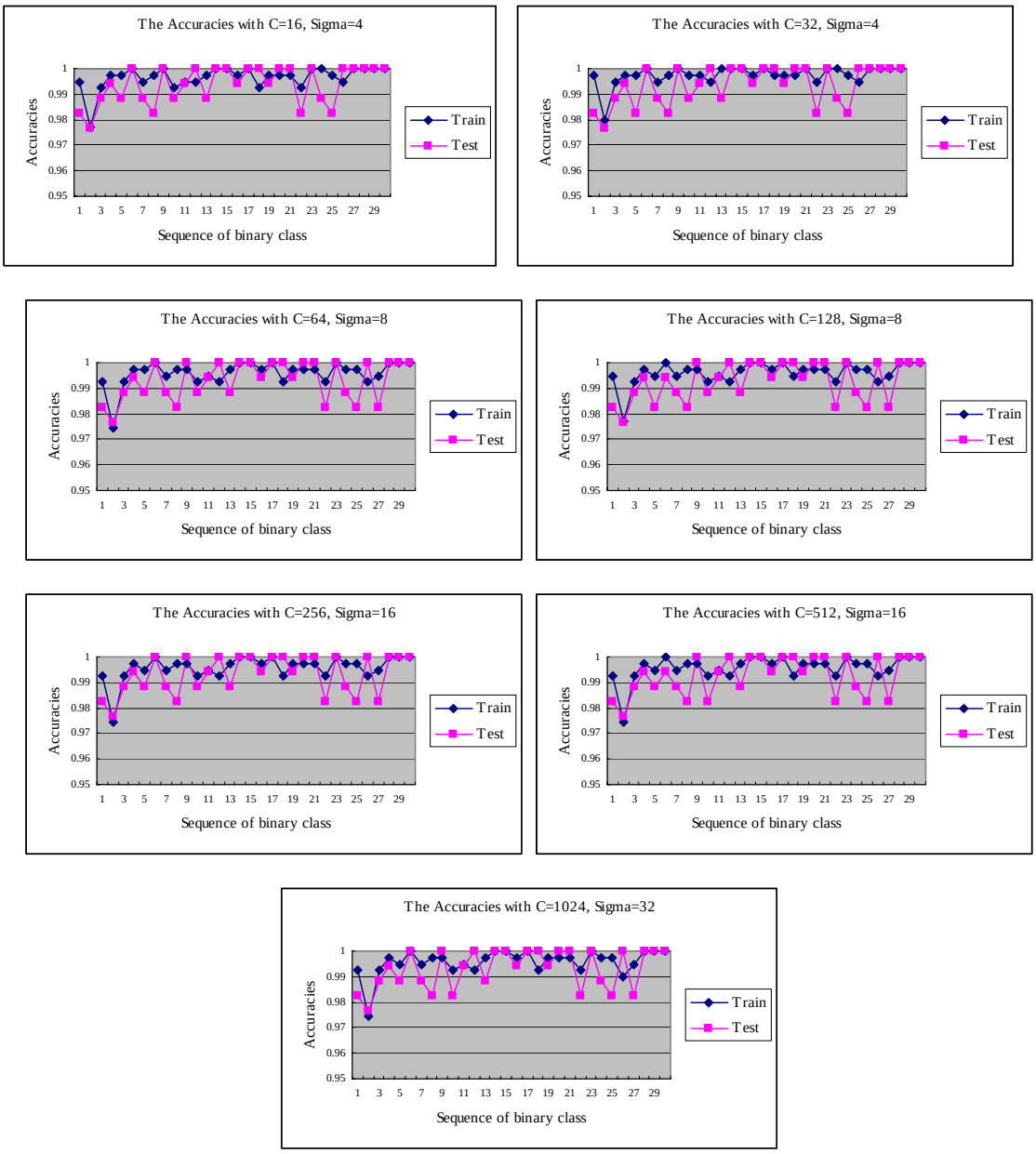
In general, the performance of the algorithm is much more important, we also report the highest testing accuracy and corresponding hyperparameters in the next Table.

Table2. The highest testing accuracy and corresponding hyperparameters

Parameters (C, σ)	Training Accuracy (%)	Testing Accuracy (%)
(16,4)	98.481	92.353
(32,4)	98.481	92.353
(64,8)	98.228	92.353
(128,8)	98.481	92.353
(256,16)	97.975	92.353
(512,16)	97.975	92.353
(1024,32)	97.975	92.353

In the following Table, we also give the training and testing accuracies of binary class of the case with highest testing accuracy. For the experiment is a 30-classes problem, we obtain 30 binary classes in all.

Table3. The training and testing accuracies of binary class of the case with highest testing accuracy



From the results above, we can easily conclude that the training and testing accuracies of the classifier with One-against-All Algorithm is very high and the accuracies of binary class are also excellent. In the support vector machine, we don't need to know the probability distribution of the data samples; In the future, we can utilize audaciously the Algorithm to the Chinese question classification.

4. Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments. This work has been supported by the National Natural Science Foundation of China (10471045, 60433020), Natural Science Foundation of Guangdong Province (970472, 000463, 04020079), Excellent Young Teachers Program of Ministry of Education of China, Fok Ying Tong Education Foundation (91005), Social Science Research Foundation of MOE (2005-241), Key Technology Research and Development Program of Guangdong Province (2005B10101010), Key Technology Research and Development Program of Tianhe District (051G041) and Natural Science Foundation of South China University of Technology (B13-E5050190), open research fund of National Mobile Communications Research Laboratory (N200605).

5. References

[1] B. DM, S. RL and W. RM, "An algorithm that learns what's in a name," Machine Learning, vol. 34, no. 1–3, pp.211–231, 1999.

[2] E. H. Hovy, U. Hermjakob, and D. Ravichandran, "A Question/Answer Typology with Surface Text Patterns," In Proceedings of the Human Language Technology Conference. San Diego, CA. NIST, Gaithersburg, MD. pp. 229-241.

[3] D. METZLER and W. B. CROFT, "Analysis of Statistical Question Classification for Fact-Based Questions," Information Retrieval, vol. 8, pp. 481–504, 2005.

[4] R. Dan, C. Chad and Li Xin, "Question-answering via enhanced understanding of questions," Proceedings of the 11th Text Retrieval Conference, Gaithersburg NIST special Publication, pp. 667-676, 2002.

[5] T. J. Suzuki, S. Yutaka and M. Eisaku, "Question classification using HDAG kernel," AM Workshop on Multilingual Summarization and Question Answering, Sapporo, pp. 61-68, 2003.

[6] V. N. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, London, UK, 1995.

[7] V. N. Vapnik, Statistical Learning Theory. John Wiley & Sons, 1998.

[8] V. N. Vapnik, "An Overview of Statistical Learning Theory," IEEE Trans. Neural Network, vol. 10, no. 5, pp. 988-999, 1999.

[9] R. Rifkin, A. Klautau, "In Defense of One-Vs-All Classification", Journal of Machine Learning Research, vol. 5, pp. 101-141, 2004.

[10] T. Inoue and S. Abe, "Fuzzy support vector machines for pattern classification", Proceedings of International Joint Conference on Neural Networks (IJCNN'01), vol. 2, pp. 1449-1454, July 2001.

[11] S. Abe, "Analysis of Multiclass Support Vector Machines", Proceedings of International Conference on Computational Intelligence for Modelling Control and Automation (CIMCA'2003), pp. 385-396, Vienna, Austria, February 2003.

[12] U. H. G. KreBel, "Pairwise classification and support vector machines", In B. Schölkopf, C. J. Burges, and A. J. Smola, editors, Advances in Kernel Methods: Support Vector Learning, pp. 255-268, The MIT Press, Cambridge, MA, 1999.

[13] J. C. Platt, N. Cristianini, and J. Shawe.-Taylor "Large Margin DAGs for multiclass classification", In S. A. Solla, T. K. Leen, and K. R. Müller, editors, Advances in Neural Information Processing Systems 12, pp. 547-553, The MIT Press 2000.

[14] D. Tsujinishi, S. Abe, "Fuzzy Least Squares Support Vector Machines for Multiclass Problems", Neural Network, vol. 16, pp. 785-792, 2003.

[15] B. Liu, Z. F. Hao and X. W. Yang, "Nesting Support Vector Machine for Multi-Classification," Proceedings of

International Conference on Machine Learning and Cybernetics, Guangzhou, pp. 4220-4225, August 2005.

[16] B. Liu, Z. F. Hao and X. W. Yang, "Nesting Algorithm for Multi-Classification Problems," international journal of Soft Computing, 2006. in press

[17] B. Liu, Z. F. Hao and X. W. Yang, "Binary Tree Support Vector Machine Based on Kernel Fisher Discriminant for Multi-Classification," Proceeding of International Symposium on Neural Networks, Lecture Notes in Computer Science, Springer-Verlag, Berlin Heidelberg New York, 2006. In press

[18] C. Angulo, X. Parra and A. Català, "K-SVCR A Support Vector Machine for Multi-class Classification", Neurocomputing, vol. 55, pp. 57-77, 2003.

[19] Z. F. Hao, B. Liu and X. W. Yang, "Twi-Map Support Vector Machine for Multi-classification Problems," Proceeding of nternational Symposium on Neural Networks, Lecture Notes in Computer Science 3496 Springer-Verlag, Berlin Heidelberg New York pp. 869-874, 2005.

[20] B. Liu, Z. F. Hao and X. W. Yang, "Quadratic Map Support Vector Machine Based on One-against-All for Multi-classification", Dynamics of continuous discrete and impulsive systems-series b-applications & algorithms. in press

[21] H. Kadri, W. Wayne, "Question classification using support vector machines and error correcting code," Proceedings of HLTNACCL 2003. Edmonton, pp. 28-30, 2003.

[22] D. Metzler, W. B. Croft, "Analysis of Statistical Question Classification for Fact-Based Questions. Information Retrieval," vol. 8, pp. 481-504, 2005.

[23] U. Hermjakob, "Parsing and question classification for question answering," ACS-2001 Workshop on Open-Domain Question Answering. Toulouse, pp. 255-262, 2001.

[24] L. Xin and R. Dan, "Learning question classifier," Proceedings of the 19th International Conference on Computational Linguistics, Taipei Morgan Kaufmann Publishers, pp. 556-562, 2002.

[25] L. Xin and R. Dan and S. Kevin, "The role of semantic information in learning question classifiers," Proceedings of the 1st International Joint Conference on Natural Language Processing. Berlin: Spring-Verlag, pp. 451-458, 2004.

[26] M. Q. Lin, J. Z. Li, Z. Y. Wang, and D. J. Lu, "A Statistical Model for Parsing Semantic Dependency Relations in a Chinese Sentence," Chinese Journal of Computers, vol.27 No.12. pp. 1679-1687, 2004.