Cost-Sensitive LapSVM for Question Classification

Liwei Yuan[1*]
[1]*School of Information Engineering and Automation*
Kunming University of Science and Technology
Yunnan, Kunming, China
yuanliwei_123@126.com

Ying Chen[2]
[2]*School of Foreign Studies*
Nanjing University of Science and Technology
Jiangsu, Nanjing, China
ychen@njust.edu.cn

Lei Su[1]
[1]*School of Information Engineering and Automation*
Kunming University of Science and Technology
Yunnan, Kunming, China
s28341@hotmail.com

***Abstract*— Community-based Question Answering (CQA) services, such as Baidu Zhidao, have attracted increasing attention over recent years, where the users can voluntarily post the questions and obtain the answers by the other users from the community. However, a large number of labeled data is needed in the process of training the model. Semi-supervised learning that can exploit the abundant unlabeled samples can solve the problem of manually labeled data. In this paper, a semi-supervised question classification method based on Cost-LapSVM is proposed. Traditional LapSVM is a natural out-of-sample extension, which can classify data that becomes available after the training process, without having to retrain the classifier or resort to various heuristics. In many real-world applications, different misclassifications often have different costs. The Cos-LapSVM algorithm is mainly used to solve cost sensitive learning problems.It could utilize the unlabeled samples for semi-supervised learning through the manifold regularization term. During the experiments, the method effectively utilizes a large number of unlabeled question samples and a few of labeled question samples to improve the question classification accuracy.**

Keywords—semi-supervised learning, cost-sensitive lapSVM, question classification

I. INTRODUCTION

CQA (Community Question Answering) is an emerging model of knowledge sharing, which brings the new knowledge-sharing approaches and platforms also the new vitality for Q&A technology [1]. Such as Baidu Zhidao, Sina iAsk, ZhiHu is interactive applications, the users in the community are not only consumers of online information, but also providers of the information, because the community-based QA makes the user share their knowledge and experience together [2]. Question classification, as an important component in Community question answering systems, is the basis of answer extraction. The classification accuracy of questions directly affects the performance of the question answering system [3]. Community-based QA has been attracting many researcher attention and become a research hotspot. The usual task of question classification is to classify the questions by the result that the user desired, such as the type of question aiming for the data, the types for names and so on, to guide the auto QA system to find answers to the questions. Question classification is roughly divided into two-one is rule-based and the other is statistical-based. Because the heuristic question classification rules largely depend on the deep knowledge of languages, it is difficult to manually extract these question classification rules. At present, the research of question classification mainly focuses on the statistical methods and the statistical learning algorithms of the usual text classification mainly include the Naive Bayes, kernel method [4], Snow [5], KNN and SVM, etc. It is worth noting that methods mentioned above are based on the labeled question samples to build model by training classifiers, the acquisition of labeled samples is usually expensive and time-consuming, while the collection of unlabeled samples is relatively easier. Consequently, semi-supervised learning, which learns from a combination of labeled and unlabeled samples for better performance than using the labeled samples alone, has attracted considerable attention.

Roughly speaking, semi-supervised classification approaches can be categorized into four paradigms, that is, generative approaches, semi-supervised large margin approaches, graph-based approaches and disagreement-based approaches [6,7]. Generally, semi-supervised classification methods attempt to exploit the intrinsic data distribution disclosed by the unlabeled samples, and the samples distribution information is generally helpful to construct a better prediction model. To exploit unlabeled data, some assumptions need to be adopted. Graph-based methods are very important branch, where nodes in the graph are the labeled and unlabeled points, and weighted edges reflect the similarities of nodes. The initially assumption of these methods is that all points are located in a low dimensional manifold, and the graph is used to approximate the underlying manifold. By the means, the labels associated with samples can be propagated throughout the graph. By using the graph Laplacian, [8] proposed a novel LapSVM, which can classify samples that becomes available after the training process, without having to retrain the classifier or resort to various heuristics.

In this paper, a new semi-supervised question classification algorithm Cos-LapSVM is proposed [15]. In applications the different misclassifications often have different costs. This classification problem is usually called cost-sensitive learning problem, which aimed to minimize the total misclassification costs. During the question classification experiments on 12 classes, utilizing semi-supervised question classification method Cos-LapSVM can effectively use a large number of unlabeled question samples to improve question classification performance.

The rest of this paper is organized as follows. Section II introduces the method of feature extraction. Section III presents the semi-supervised question classification algorithm Cos-LapSVM. Section IV reports and analyzes the experimental study on the Chinese question classification in CQA. Finally, section V summarizes this paper and proposes the future work.

II. Method of Feature Extraction

A. Word Segmentation

In process of feature extraction, the word segmentation play an important role. The ICTCLAS platform (http://ictclas.nlpir.org/) is used to do word segmentation for Chinese questions and stop words are removed. In this paper when constructing artificial collated fields marked category, a lot of damage entries by word segmentation tool will be found, mainly due to the lack of appropriate rules for the user dictionary.In word segmentation, many categories are exclusive nouns with obvious distinguishing ability, as the process of word segmentation has already led to the destruction of part of speech. To address this issue, we have created a unique user dictionary to solve the destruction of part speech during word segmentation

B. Establishment of Question Classification Feature Space

This paper selects word bag as the question classification feature model, in order to achieve the feature vector of every question, to make word segmentation for every question, and then use TFIDF to proceed feature extraction. However, there is a difference between question classification and general text classification that a question can only separate about ten words or even less. For example, "The air-conditioning brand list" which can only separate three words, i.e., "air-conditioning", "brand" and "list". Therefore, there are many 0s in the feature space, i.e. sparse matrix. In order to solve the problem of sparse matrix, this paper introduces word semantic extension calculation method based on Tongyici Cilin [10] proposed by Liu and Wang[9]. The method achieves ideal results in the keywords extraction. Keywords extraction is the basic work of natural language processing of automatic abstract, text classification and topic detection etc. And it plays a important role. However, there are many texts not able to extract keywords by human effort. To a great extent, Liu and Wang used word semantic extension calculation method based on Tongyici Cilin to solve the problem. In this paper, this method is used to solve the problem of sparse matrix.

Word semantic extension consists of the word similarity and the word relevance. The word similarity algorithm based on coding distance in dictionary is used to calculate the word similarity. Its main idea is to estimate the level of similarity between two words in the synonym structure, and then calculate the similarity between the two words according to the semantic distance between the meanings of the two words. The closer distance, the higher similarity. Word similarity calculation is shown in formula (1)

$$Sim(w_1, w_2) = d * \left(\frac{n-k+1}{n}\right) * \cos\left(n * \frac{\pi}{180}\right) \quad (1)$$

Where the $Sim(w_1, w_2)$ is semantic similarity(0<Sim<1); d is a coefficient, which is up to the branch layer of the corresponding code of two words; n is the number of panel point in branch layer; k is the distance between two branches.

III. Semi-Supervised Question Classification based on Cos-LapSVM

Regularization is a key technology for obtaining smooth decision functions and thus avoiding over-fitting to the training data, which is widely used in machine learning.The Cos-LapSVM is proposed by Qi[15] to solve the cost-sensitive learning problem.

LapSVM [11] is an important semi-supervised algorithm. It is built on two important factors. One is manifold assumption, i.e., similar instances have similar outputs; the other is large margin principle, i.e., the distributions of two different classes have a large margin.

Formally, given a set of training samples $D = \{x_i, y_i\}_{i=1}^{l} \cup \{x_i + j\}_{j=1}^{u}$, where $\{x_i, y_i\}_{i=1}^{l}$ and $\{x_i + j\}_{j=1}^{u}$ are labeled and unlabeled data, respectively. l and u are numbers of labeled and unlabeled sample, respectively. LapSVM then aims to learn a decision function f such that the following functional is minimized, as follows:

$$\arg\min_{f\in H} \frac{1}{l}\sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_H^2 + \frac{\gamma_I}{(u+l)^2}\sum_{i,j=1}^{l+u}\left(f(x_i) - f(x_j)\right)^2 W_{ij} \quad (2)$$

Here, the first term is the classification loss on labeled sample, e.g., hinge loss in SVM that enforces the distributions of two different classes have a large margin; the second term prefers the decision function to be a simple classifier; while the third term enforces that similar instances have similar output according to the similarity weighted matrix W of all training instances. γ_A and γ_I are two parameters trading off these three terms. It has been found that LapSVM is useful for many applications [12, 13]. where The weight matrix W may be defined by k nearest neighbor or graph kernels as follows(3):

$$w_{ij} = \begin{cases} \exp\left(-\|x_i - x_j\|_2^2 / 2\sigma^2\right) \\ 0 \end{cases} \quad (3)$$

Where $\|x_i - x_j\|_2^2$ denotes the Euclidean norm in R_n. So the manifold regularization is defined by

$$\frac{\gamma_I}{(u+l)^2}\sum_{i,j=1}^{l+u}\left(f(x_i) - f(x_j)\right)^2 W_{ij} = f^T L f \quad (4)$$

Where $f(x)=[f(x_1),...,f(x_{l+u})]^T$ $w_{ij} \in w$ $(i, j = 1, \ldots, u + l)$ are the edge weights in the sample adjacency graph, $L = D - W$ is the graph Laplacian, D is the degree matrix of which the diagonal entries $D_{ii}=\sum_{j=1}^{l+u} w_{ij}$ and the others are equal to 0.

It is evident that LapSVM is a cost-blind discriminant analysis method because it does not take any misclassification cost into account. In this paper, we extend LapSVM for cost-sensitive scenarios. Specifically, for each labeled training sample, the misclassification cost is incorporated into the classification loss, i.e.,

$$\arg\min_{f\in H} \frac{1}{l}\sum_{i=1}^{l} c(y_i) V(x_i, y_i, f) + \gamma_A \|f\|_H^2 + \frac{\gamma_I}{(u+l)^2}\sum_{i,j=1}^{l+u}\left(f(x_i) - f(x_j)\right)^2 W_{ij} \quad (5)$$

where $C(y_i)$ corresponds to the misclassification cost for label y_i This leads to our proposed Cos-LapSVM. It can be shown that Cos-LapSVM is a convex optimization whose global optimal solution can be solved efficiently.

TABLE I. COS-LAPSVM ALGORITHM

Algorithm: Cos-LapSVM
Input: the labeled set *L*, the unlabeled set *U*,
Process:
1) Compute the (*l*+*u*) total training sample to construct a data adjacency graph
2) Compute the graph Laplacian $L = D - W$
3) Select kernel function $K(x_i\ x_j)$, Compute the kernel function *K* of (*l*+*u*) training samples
4) Using Standard SVM Algorithm to Solve Quadratic Programming

Output: $f(x)=\sum_{i=1}^{l+u} \alpha_i K(x_i\ x_j)$ the discriminant function

IV. EXPERIMENTS AND RESULT ANALYSIS

The experimental data sets consist of 24,500 Chinese questions collected from baidu zhidao, including 12 classes. The classes respectively are office software, image processing software, multimedia software, browser, system software, Databasecloud computing, embedded, mobile development, Operating system development, network connection, the network uses the Internet industry, and the cloud service. Experiments used 10-folds cross validation, divided the dataset that composed of the 24,500 questions into 10 folds, and each time took a fold as a testset. The testset is 10% of the entire dataset, and the remaining 90% of the data composed training sets by the nine folds. 10 folds are in turn taken as a testset for testing, the whole process above was repeated for ten times and the results were averaged.

A. Comparison of different semi-supervised learning Methods

In order to explore more effective semi-supervised learning algorithm and make full use of unlabeled questions to improve classification accuracy, we compare the different semi-supervised learning algorithms, including SVM, LapSVM and Cos-LapSVM. 3 sets of data are used in this experiment, including 12 classes. They are D1, D2 and D3. These 3 sets of data by the different radios unlabeled data contains 22050 questions. The number of unlabeled examples of D1 is by the radio 90%, the number of unlabeled examples of D2 is by the radio 85% and the number of unlabeled examples of D3 is by the radio 80%. And the feature vector build feature space of 200 dimension. The dataset is shown as in table II.

TABLE II. THE EXPERIMENT DATA SET OF DIFFERENT LEARNING ALGORITHMS.RITHM

Dat Set	UnLabeled Radio	Labeled Data	UnLabeled Data	Test Data
D1	90%	2205	19845	2450
D2	85%	3307	18743	2450
D3	80%	4410	17640	2450

For comparison, the performance of semi-supervised learning method and make full use of unlabeled questions to improve the classfication accuracy, the different semi-supervised learning method as shown in table III.

TABLE III. THE EXPERIMENT DATA OF SEMI-SUPERVISED AND SUPERVISED

DATA SET	SEMI-SUPERVISED LEARNING		SUPERVISED LEARNING
	LAPSVM	COS-LAPSVM	SVM
D1	77.99±0.45	79.45±0.89	76.32±0.71
D2	78.85±0.64	80.34±0.68	76.89±0.45
D3	79.87±0.78	81.45±0.46	77.52±0.27

In table III, the classification accuracy of semi-supervised learning method LapSVM and Cos-LapSVM is slightly higher than supervised learning method SVM. For example the classification accuracy of Cos-LapSVM is 3.13% higher than SVM method during the case that the radio of unlabeled examples is 90%. And the classification accuracy of Cos-LapSVM is 3.45% higher than SVM method during the case that the radio of unlabeled examples is 85%. And the classification accuracy of Cos-LapSVM is higher 3.93% than the LapSVM during the case that the radio of unlabeled examples is 80%. However, the classification accuracy of semi-supervised learning method Cos-LapSVM is much higher than semi-supervised learning method LapSVM. The experiment results indicate the effectiveness of this method. On the other hand, the classification accuracy of most data sets has been improved to a certain extent under different training sample label radio. Comparing Cos-LapSVM with LapSVM algorithm can more effectively and stably mine data structure information, thereby improving learning performance. The classification accuracy of Cos-LapSVM is 1.46% higher than the LapSVM during the case that the radio of unlabeled examples is 90%. And the classification accuracy of Cos-LapSVM is 1.49% higher than the LapSVM during the case that the radio of unlabeled examples is 85%. And the classification accuracy of Cos-LapSVM is higher 1.58% than the LapSVM during the case that the radio of unlabeled examples is 80%.

Therefore, this experiment verifies the effectiveness of the semi supervised learning method Cos-LapSVM compared to other different methods on different radio unlabeled samples.

B. Comparison of different features dimensions

Experimental algorithms exhibit varying performance in different feature dimensions, including both supervised and semi-supervised learning algorithms on data sets. The newly experimental data sets can be obtained from the original data sets with the different dimensionalities of feature space. In the experiment, we use the data sets with different class on unlabeled rate to 70%. The dimensionalities of feature space are set to 200, 500 and 700, respectively. The detailed information of these data sets is shown as table IV.

TABLE IV. THE EXPERIMENT DATA SETS OF DIFFERENT DIMENSIONS.

DATA SET	DIMENSION	UNLABELED RADIO	LABELED DATA	UNLABELED DATA	TEST DATA
D1	200				
D2	500	70%	6615	15435	2450
D3	700				

We still choose supervised learning method SVM, semi-supervised learning method LapSVM and Cos-LapSVM as the classifier. In LapSVM and Cos-LapSVM algorithms, set the regularization parameters γ_A and γ_I values to 0.03120 and l respectively. Table V tabulates the accuracy rates on the data sets under the different dimensionalities.

TABLE V. THE CLASSIFICATION ACCURACY OF DIFFERENT DIMENSIONS

DATA SET	DIMENSION	SEMI-SUPERVISED LEARNING		SUPERVISED LEARNING
		LAPSVM	COS-LAPSVM	SVM
D1	200	80.12±0.24	81.79±0.72	78.58±0.31
D2	500	80.87±0.52	82.28±0.45	79.76±0.63
D3	700	80.53±0.61	81.88±0.62	79.42±0.35

Table V shows that both supervised and semi-supervised learning method achieves the highest classification accuracy rates on D2 with 500-dimensionality in all dimensionalities of feature space. Although the accuracy rates on the data set with 500-dimensionality are all higher than those on the data sets with 200 and 700 dimensionalities, the improvements are not so significant. Compared to D1 with 200-dimensionality, the accuracy rate of LapSVM on D2 with 500-dimensionality increases by 0.75% points, and the accuracy rate of Cos-LapSVM increases by 0.49% points. The accuracy rate of LapSVM on D3 with 500-dimensionality reduce by 0.34% points compared to that on D2, and the accuracy rate of Cos-LapSVM reduce by 0.4% points. The experimental results show that the accuracy rate cannot always improve with the increase of the feature sizes. However, with the growth of the dimension, both the training model and test process need to consume more time.

V. CONCLUSION

In this paper, the Cos-LapSVM semi-supervised classification algorithm is an efficient method which can effectively utilize a large number of unlabeled question samples which are easily obtained to improve the classification accuracy. To improve classification performance, a small amount of labeled question samples is combined with a large number of unlabeled question samples. Experimental results show that this method can significantly improve the accuracy of the question classification. Research on CQA hasachieved certain results, however, as an emerging research topic, some unresolved issues still need to be studied and further explored.

REFERENCES

[1] Zhongfeng Zhang, Qiudan Li, “Studies on Community Question Answering-A survey,” Computer Science, vol. 37, pp.19-23, 2010.

[2] XianLing Mao, XiaoMing Li, “A Survey on question answering system,” Journal of Frontiers of Computer Science and Technology, vol. 6, pp.193-207, 2012.

[3] Yu Zhang, Ting Liu etal, “Modified Bayesian model based question classification,” Chinese Inform Process, vol.19, pp.100-105, 2005.

[4] Taira Jun Suzuki, Sasaki Yutaka， and Maeda Eisaku, “Question classification using HDAG kernel,” ACL Workshop on Mulitilingual Summarization and Question Answering, pp.61-68, 2003.

[5] Xin Li, Roth Dan, “Learning question classifier: The role of semantic information,” Natural Language Engineering,vol.12, pp.229-249, 1998.

[6] Zhi Hua Zhou and Ming Li, “Semi-supervised learning by disagreement,” Knowledge and Information Systems, vol. 24, pp.415－439, 2010.

[7] Xiao Jin Zhu, “Semi-supervised learning literature survey,” Madison:University of Wisconsin, 2008.

[8] M. Belkin, P. Niyogi, V. Sindhwani, Manifold Regularization, “A Geometric Framework for Learning from Labeled and Unlabeled Examples, Journal of Machine Learning Research,” vol. 7, pp.2399－2434, 2006.

[9] Riuyang Liu, Liangfang Wang, “Keyword Extraction Algorithm Combining Semantic Extension Degree and Lexical Chain,” Journal of Computer,vol 40, pp.265-266, 2013.

[10] Jiule Tian , wei Zhao, “Word Similarity Calculation Method based on Tongyici Cilin,” Journal of Jilin University(Information science edition), vol. 28, pp.603-604, 2010.

[11] D. Margineantu, “When Does Imbalanced Data Require More than Cost-Sensitive Learning,” Proc the AAAI 2000 Workshop Learning from Imbalanced Data Sets, pp.47-50 2000.

[12] L. GomezChova, L.G. CampsValls, J. MunozMarı, and J. Calpe, “Semisupervised Image Classification with Laplacian Support Vector Machines,” IEEE Geoscience and Remote Sensing Letters, vol.5, pp.336-340, 2008.

[13] Jiang Wu, YuanBo Diao, MengLong Li, “A Semi-Supervised Learning Based Method: Laplacian Support Vector Machine Used in Diabetes Disease Diagnosis,” Interdisciplinary Sciences Computational Life Sciences, vol. 1, pp.151-155, 2009.

[14] JianSheng Wu and ZhiHua Zhou, “Sequence-Based Prediction of microRNA-Binding Residues in Proteins Using Cost-Sensitive Laplacian Support Vector Machines,” IEEE/ACM transactions on computational biology and bioinformatics, vol.10, 752-759, 2013.

[15] Zhiquan Qi, Yingjie Tian, Yong Shi, Xiaodan Yu, “Cost-Sensitive Support Vector Machine for Semi-Supervised Learning,” International Conference on Computational Science，pp.1684-1689, 2013.