DATA SELECTION BASED ON FUZZY CLUSTERING

DONGHAI GUAN, WEIWEI YUAN, YOUNG-KOO LEE[*], ANDREY GAVRILOV
AND SUNGYOUNG LEE

Department of Computer Engineering, Kyung Hee University
Suwon, 446-701, Korea

When the number of training data is limited, the performance of supervised learning could be improved if valuable samples are selected for training. In this work, we propose a novel data selection method based on fuzzy clustering. Our method first partitions all the data which need to be classified into clusters. Then training data are selected from each cluster based on their membership degrees. Experimental results show that our proposed fuzzy clustering-based data selection method could effectively improve the performance of learning compared with randomly selecting training samples.

1. Introduction

When designing a supervised learning system, usually we need enough training data. However, in many cases, we have to limit the number of training data. The reason is that only labeled data can be used for training and labeled data are often difficult, expensive, or time consuming to obtain, as they require the efforts of experienced human annotators [1]. In these cases, how to achieve a good classifier as best as possible with a reasonable number of labeled training data is an important issue.

Many approaches have been proposed to solve this issue. These approaches can be divided into two primary topics: semi-supervised learning [2][3][4] and training data selection [5][6][7][8]. Assuming the cost associated with the labeling efforts is uniform for all the samples in a dataset, data selection aims to choose the most valuable samples to label. When labeled data are given, semi-supervised learning aims to utilize unlabeled data to improve learning performance. Our work focuses on data selection.

In this paper, we propose a novel data selection method based on fuzzy clustering. First of all, fuzzy c-means is used for data clustering. Then two parts of samples are selected. The first part includes the samples with high degrees of

---

[*] Corresponding author.

membership in each cluster. Usually these samples are close to the cluster centers. And the other part includes the samples with small difference between the two highest membership degrees. Finally the combination of them is used as training data.

Compared with most existing works in data selection, our proposed method requires less computational effort. Moreover, our method does not rely on the supervised learning algorithms, so it can work with any kind of supervised learning methods, such as neural networks, support vector machines and so on.

To test the effect of our method, we compare the performance of learning based on our method and randomly selecting training samples. Experimental results indicate that our proposed method can achieve better performance.

The structure of the paper is as follows. In section 2, we present our proposed data selection method. Section 3 reports on the experiment results. Section 4 discloses conclusions and future work.

2. Our Data Selection Method

The first step of our data selection method is to partition the samples which need to be classified into clusters by using fuzzy c-means. Then, class membership matrix U is obtained [9][10]. Our method selects training samples through analyzing this membership matrix. The selected samples include the following two parts:

1. The samples with high degrees of membership in each cluster. For a data set D , suppose there are n samples which belong to m clusters. Let u_{ij} denote the element at the ith row and j^{th} column of U . u_{ij} is the membership degree of sample i in cluster j . For each cluster j , if $i_1 = \arg \max_{i=1:n}(u_{ij})$, then sample i_1 is selected firstly because it has the maximum membership degree in cluster j among all the samples. The next sample to be selected is sample i_2 with $i_2 = \arg \max_{i=1:n, i \neq i_1}(u_{ij})$. In turn, other samples are selected using this way. In our current work, the numbers of samples selected from each cluster are usually same or similar.

2. The samples with small difference between the two highest membership degrees. For each sample i , if $j_1 = \arg \max_{j=1:m}(u_{ij})$ and $j_2 = \arg \max_{j=1:m, j \neq j_1}(u_{ij})$, then its two highest membership degrees are

u_{ij_1} and u_{ij_2} . Set $T_i = u_{ij_1} - u_{ij_2}$. If $i_* = \arg\min_{i=1:n}(T_i)$, then sample i_* is selected firstly. The next sample to be selected is sample i_{**} with $i_{**} = \arg\min_{i=1:n,i\neq i_*}(T_i)$. In turn, other samples are selected.

If there are k samples to be selected for training, then one simple method we use is to select $k/2$ samples for each part. In the real applications, we need not follow this combination method exactly. For example, if it is obvious that samples got from part 1 are redundant, then more samples could be extracted from part 2.

3. Experimental Results

Three well-known data sets were used in our experiments. For each data set S , we randomly select some samples as test data S_1 , then training data will be selected from the other part ($S - S_1$). Training data random selection strategy and our strategy are conducted for each data set for comparison. Multi-layer Perception (MLP) with error back propagation (EBP) is used as the supervised classifier in our experiment. The network with one hidden layer is adopted. TANSIG, LOGSIG activation functions are used in the hidden layer and output layer respectively. For fuzzy c-means algorithm, it is configured as follows: fuzzy factor is set to 2, convergence criterion is set to 0.0000001, maximum iteration is set to 100. Euclidean distance is used as the similarity measure. All the experiments in this part are carried out for 500 times. Then the average value of the 500 classification accuracies was reported as result.

The first data set we used is the well-known iris dataset [10]. It contains of four characteristics of iris plants and classifies them into three classes of iris with 50 exemplars in each class. One class is linearly separable from the other two which are not linearly separable from each other. In our experiment, 75 samples are randomly selected as test data. Then training data are selected from the other 75 samples. Different numbers of training data are used, which include 6, 9, 15, 21 and 27. Recall to Section 2, the selected training samples include two parts. The numbers of data assigned to these two parts are given in Table 1. For example, when the number of training data is 6, then there are 3 samples selected from part 1 and part 2 respectively. For part 1, there is one sample selected from each of iris's three classes. Classification error rates on iris are shown in Table 2. Compared with random selection, our method achieves 23.9 percent, 19.1 percent, 11.1 percent, 6.74 percent and 1.40 percent improvement when the number of training data is 6, 9, 15, 21 and 27 respectively.

The second data set is soybean [10], which has 47 instances and 35 attributes. Four classes are represented in the data. 25 instances are randomly selected as test data. Training instances are selected from the other 22 instances. Number of training data includes 8, 12 and 16. Training data distribution for soybean is shown in Table 3. Classification error rates are shown in Table 4. Compared with random selection, 46.7 percent, 26.4 percent and 21.2 percent improvements are achieved at training number 8, 12 and 16 respectively.

The third data set is Dr. William W. Wolberg's Wisconsin Breast Cancer Dataset [10]. Originally this dataset contains 699 samples with 458 samples in the class Benign and 241 samples in the class Malignant. Each sample has 9 input features. There are 16 samples with incomplete features. After filtering out those samples, 683 samples are used in our experiment. 100 samples are randomly selected as test data. Training samples are selected from the other 583 samples. Training data numbers include 6, 10 20 and 30. Training data distribution for this data set is shown in Table 5. Classification error rates are shown in Table 6. Compared with random selection, our method achieves 39.7 percent, 20.8 percent, 7.63 percent and 6.04 percent improvement when the number of training data is 6, 10, 20 and 30 respectively.

Table 1. Iris: training data distribution

Dataset: iris	
Number of Training Data	Our Method
6	3 samples for part 1 (1*3) and 3 samples for part 2
9	6 samples for part 1 (2*3) and 3 samples for part 2
15	9 samples for part 1 (3*3) and 6 samples for part 2
21	12 samples for part 1 (4*3) and 9 samples for part 2
27	15 samples for part 1 (5*3) and 12 samples for part 2

Table 2. Iris: classification error rates

Dataset: iris			
Number of Training Data	random selection	our method	improv
6	0.2770	0.2108	23.9%
9	0.1878	0.1520	19.1%
15	0.1252	0.1113	11.1%
21	0.1009	0.0941	6.74%
27	0.0860	0.0848	1.40%
ave.	0.1554	0.1306	12.4%

Table 3. Sobean: training data distribution

Dataset: soybean	
Number of Training Data	**Our Method**
8	4 samples for part 1 (1*4) and 4 samples for part 2
12	8 samples for part 1 (2*4) and 4 samples for part 2
16	8 samples for part 1 (2*4) and 8 samples for part 2

Table 4. Soybean: classification error rates

Dataset: soybean			
Number of Training Data	**random selection**	**0ur method**	**improv**
8	0.1762	0.0940	46.7%
12	0.0926	0.0681	26.4%
16	0.0548	0.0432	21.2%
ave.	0.1079	0.0684	31.4%

Table5. Breast cancer: training data distribution

Dataset: breast cancer	
Number of Training Data	**Our Method**
6	4 samples for part 1 (2*2) and 2 samples for part 2
10	6 samples for part 1 (3*2) and 4 samples for part 2
20	10 samples for part 1 (5*2) and 10 samples for part 2
30	20 samples for part 1 (10*2) and 10 samples for part2

Table 6. Breast cancer: classification error rates

Dataset: breast cancer			
Number of Training Data	**random selection**	**our method**	**improv**
6	0.1369	0.0826	39.7%
10	0.0910	0.0721	20.8%
20	0.0708	0.0654	7.63%
30	0.0579	0.0544	6.04%
ave.	0.0892	0.0686	18.5%

The results indicate that, compared with randomly selecting training samples:

- For the data set iris, our data selection method achieves 12.4 percent of average improvement.
- For the data set soybean, our data selection method achieves 31.4 percent of average improvement.
- For the data set breast cancer, our data selection method achieves 18.5 percent of average improvement.

4. Conclusions and Future Work

In this work, we propose a novel data selection method based on fuzzy clustering. The experiment results show that it can achieve better classification performance compared with traditional random selection method.

In the future, we will try to find some concrete applications to test the effect of our method.

Acknowledgments

References

1. X. J. Zhu, Semi-Supervised Learning Literature Survey, http://www.cs.wisc.edu/~jerryzhu/research/ssl/semireview.html.
2. S. Basu, Semi-supervised Clustering with Limited Background Knowledge, In Proc. of the 9th AAAI/SIGART Doctoral Consortium, 979-980, (2004).
3. S. Basu, A. Banerjee and R. J. Mooney, Semi-supervised Clustering by Seeding, In Proc. of the 19th International Conference on Machine Learning (ICML), 19-26, (2002).
4. N. Grira, M. Crucianu and N. Boujemma, Unsupervised and Semi-supervised Clustering: a Brief Survey, A Review of Machine Learning Techniques for Processing Multimedia Content, 2004, http://www-rocq.inria.fr/~crucianu/src/BriefSurveyClustering.pdf.
5. D. D. Lewis and W. A. Gale, A sequential algorithm for training text classifiers, In Proc. Of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval, 3-12, (1994).
6. C. Campbell, N. Cristianini and A. Smola, Query learning with large margin classifiers, In Prof. of 17th International Conference on Machine Learning, CA, 111-118, (2000).

7. G. Schohn and D. Cohn, Less is more: Active learning with support vector machines, In Proc. of 17th International Conference on Machine Learning, CA, 839-846, (2000).

8. S. Tong and E. Chang, Support vector machine active learning for image retrieval, In Proc. of the 9th ACM International Conference on Multimedia, Ottawa, 107-118, (2001).

9. J. C. Dunn, A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, In J. of Cybernetics, 32-57, (1973).

10. J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, (1981).

11. http://www.ics.uci.edu/~mlearn/MLRepository.html.