The 9th Pacific-Rim Conference on Multimedia 2008 (PCM 2008), Tainan, Taiwan

Local Separability Assessment: a Novel Feature Selection Method for Multimedia Applications

Kun Tao[1,2], Shou-Xun Lin[1], Yong-Dong Zhang[1], and Sheng Tang[1]

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
[2] Graduate School of the Chinese Academy of Sciences, Beijing, 100049, China
{ ktao, sxlin, zhyd, ts }@ict.ac.cn

Abstract. Feature selection technology can help to reduce feature redundancy and improve classification performance. Most general feature selection methods do not perform well on high-dimension large-scale data sets of multimedia applications. In this paper we propose a novel feature selection method named Local Separability Assessment. We try to measure the separation level of samples in subregions of feature space, and integrate them for evaluating the separability of features. Our method has favorable performance on large-scale continuous data sets, and requires no priori hypothesis on data distribution. The experiments on various applications have proved its excellence.

Keywords: Local Separability Assessment, feature selection.

1 Introduction

The performance of a feature selection method mainly depends on its metric function. According to the metric function, feature selection methods can be broadly divided into Filter methods and Wrapper methods [1]. Considering the calculation amount, Filters are more suitable to deal with large-scale multimedia data sets [2]. There are some classical metrics which have been widely discussed [1, 3]: Distance measures are intuitive and computable, but their performances decrease on high-dimension data or data with complex prior distributions. Most measures based on Consistency, information or dependence are more suitable to deal with discrete values. Although using discretization algorithms can be helpful, they have to face the problem of combination explosion or sharp increase of computational complexity. Relief-based methods have been proved to be effective. But when the number of instances increases, the calculation amount also becomes a bottleneck. Most of above feature selection methods are not suitable for multimedia applications.

Furthermore, in applications of multi-modal information fusion, another problem is proposed: we need to evaluate and compare the separabilities of two feature sets which are extracted independently. Here the metrics based on correlation of features become helpless, and other metrics also have to face the problem of calculation amount and asymmetrical feature dimensionalities.

To overcome above difficulties, we need some new ideas: The global separability is positively correlated with local separabilities. If classes are separable and few

instances are interleaved, they are also separable in most local subregions of appropriate size. So we propose a Local Separability Assessment (LSA) method, which simplifies the problem of solving global separability to solving local separabilities. The LSA can deal with large-scale data sets effectively and efficiently.

2 The Local Separability Assessment Method

2.1 The Metric Function

The LSA metric function is calculated based on the separabilities of subregions. First, the feature space should be separated into subregions of appropriate size. A density-based K-means method is used to accomplish this job [4]. Each cluster is regarded as a subregion of the feature space. The subregions with too little samples are marked as invalid. We calculate the subregion's Fisher's discriminant ratios on every dimension:

$$f_{cd} = \begin{cases} 1 & \text{samples in c are all positive or negative} \\ (\mu_{cd}^{+} - \mu_{cd}^{-})^2 \big/ [(\sigma_{cd}^{+})^2 + (\sigma_{cd}^{-})^2] & \text{else} \end{cases} \quad (1)$$

Where c denotes a subregion, d denotes a feature dimension, μ and σ is the mean and variance, + and – denote the sample classes. Then we define S_c as the separability metric of a subregion, which is the average of f_{cd} of all dimensions. The weighted mean of S_c values of all valid subregions is named as S_{LSA}:

$$S_{LSA} = \sum_{\forall c \in C_{valid}} S_c N_c \Big/ \sum_{\forall c \in C_{valid}} N_c \quad (2)$$

Where N_c is the sample number of subregion c, C_{valid} is the set of all valid subregions. The resulting S_{LSA} is the separability metric of our method.

2.2 The Search Strategy

For the task of selecting feature subsets, it's inefficient to calculating the separability metrics of all subsets. We use a heuristic backward search strategy, and the system diagram is shown in Fig. 1.

During the iteration process of searching, we remove one dimension of feature at each step. It's costly and unnecessary to compute the metrics of all (D-1)-dimension

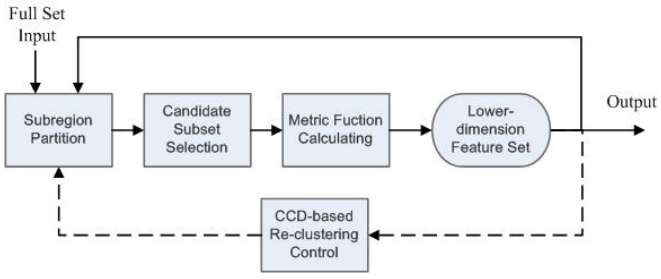

Fig. 1. System diagram of subset selection

subsets. So we use a heuristic method to find some candidate subsets: For the current D-dimension feature set, we calculate the mean centers of positive samples and negative samples respectively in each subregion. The line vector connecting them is considered to the normal vector of a cutting plane. We Calculate average of normalized normal vectors, the components of which can indicate the feature separabilities of corresponding dimensions. Removing the m dimensions with lowest component values respectively, m candidate (D-1)-dimension subsets are formed up. The subset with highest LSA separability will be selected as the result of current step.

We should also try to avoid frequent clustering. We can inherit the cluster partition of a D-dimension feature set when selecting its subset. With the decrease of feature dimensionality, the sample points are projected to a new space. Thus, some clusters separated in original space overlap with each other in new space, and their centers become closer and closer. Here we define the Cluster Center Dispersion (CCD) as:

$$CCD = \sum_{\forall c} \min_{\forall c' \neq c} \{\text{Distance}(Center_c, Center_{c'})\} \tag{3}$$

Where $Center_c$ is the center point of cluster c. Using the cluster partition of D-dimension feature set, we can compute the CCD of the subsets in lower-dimension space. If the CCD in new space is some percent lower than the original CCD, we regard that it's necessary to afresh the cluster partition and the clustering algorithm will be implemented on new feature subset.

3 Experiments

3.1 Experiments on Classical Data Sets

We first validate our method on feature subset selection problems. Many different data sets can be obtained from the UCI Machine Learning Repository [5]. We select two well-known data sets for our experiments: Sonar and Breast Cancer Wisconsin (wdbc). Using our LSA method, Relief-F and Information Gain separately, we selected the best subsets of some different dimensionalities. All data sets are evaluated by LIBSVM and 10-fold Cross-Validation, and the accuracy changes between full-featured sets and their subsets are illustrated in Fig. 2. We can see that the performance of our method is better than that of Relief-F or Information Gain.

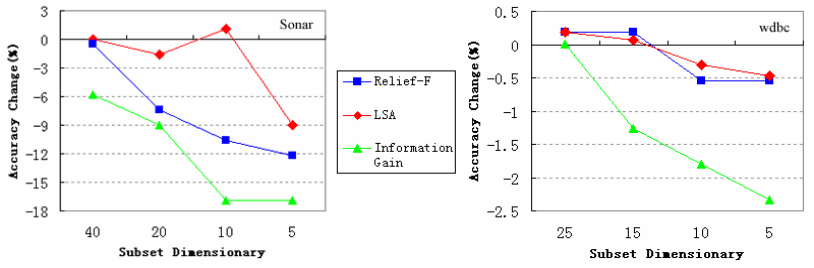

Fig. 2. Experiments results of Sonar and wdbc

3.2 Experiments on Multimedia Data Set

Sometimes we need evaluate the feature sets before training classifiers or fusion. Here we use the TRECVID 2007 dataset [6] for experiments of feature evaluation. 20 concepts are selected here, and 6 different image features are extracted for concept detection. We use the test precision of SVM classifiers as the ground truth. A rank list of features can be made based on test precision. Comparing it with the rank list based on separability metric, we can get to know whether a separability metric is helpful. A measurement named Translation Error Rate (TER) [7] is used to measure the differences of two rank lists. We also calculate the separability metric based on Relief-F for contrasting. For each concept, we make the rank list of 6 features based on LSA and Relief respectively. For all 20 concepts, the mean value of TERs based on LSA is 3.70, while the mean value based on Relief is 4.15. The comparison result shows that the LSA method can give a better assessment of feature separability.

4 Conclusions

Our LSA method is developed for the feature selection problem of high-dimension large-scale multimedia data sets, and tries to satisfy the requirements of both computational complexity and efficiency. Experiments show that our method not only works well on subset selection problems, but also show excellent performance on feature evaluation of multimedia data. For future work we plan to extend our method for multi-class problems, and study its potentiality on discrete data sets.

Acknowledgments. This work was supported in part by the National Basic Research Program of China (973 Program, 2007CB311100), the National High Technology and Research Development Program of China (863 Program, 2007AA01Z416) and the National Nature Science Foundation of China (60873165)

References

1. H. Liu and H. Motoda, "Feature Selection for knowledge Discovery and Data Mining", Kluwer Academic Publishers, 1998.
2. I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection", Journal of Machine Learning Research, 2003, Vol. 3, pp. 1157-1182.
3. C. Yun, D. Shin, H. Jo, J. Yang, and S. Kim, "An Experimental Study on Feature Subset Selection Methods", IEEE International Conference on Computer and Information Technology, 2007, pp. 77-82.
4. Y. X. Xie, Y. N. Hu, etc., "An efficient indexing algorithm of clustering supporting QBE image retrieval for large image database", Mini-Micro Systems, 2001, vol. 10, pp. 1229-1233. (in Chinese).
5. UCI Machine Learning Repository, http://www.ics.uci.edu/~mlearn/MLRepository.html
6. TREC Video Retrieval Evaluation, http://www-nlpir.nist.gov/projects/trecvid/
7. R. Schwartz, B. Dorr, R. Schwartz, L. Micciulla and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation", Proceedings of Association for Machine Translation in the Americas, 2006.