Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

IPCM Separability Ratio for Supervised Feature Selection

Wing W. Y. Ng
School of Computer Science and Engineering, South China University of Technology, Guangzhou, China
wingng@ieee.org

Jun Wang
Dept. of Comp. Sci. and Tech., Shenzhen Graduate School, Harbin Institute of Technology, China

Daniel S. Yeung
School of Computer Science and Engineering, South China University of Technology, Guangzhou, China

***Abstract*—Collecting data is very easy now owing to fast computers and ease of Internet access. It raises the problem of the curse of dimensionality to supervised classification problems. In our previous work, an Intra-Prototype / Inter-Class Separability Ratio (IPICSR) model is proposed to select relevant features for semi-supervised classification problems. In this work, a new margin based feature selection model is proposed based on the IPICSR model for supervised classification problems. Owing to the nature of supervised classification problems, a more accurate class separating margin could be found by the classifier. We adopt this advantage in the new Intra-Prototype / Class Margin Separability Ratio (IPCMSR) model. Experimental results are promising when compared to several existing methods using 4 UCI datasets.**

***Keywords*—Supervised Feature Selection, Intra-Prototype / Class Margin, Separability Ratio**

I. Introduction

Selecting relevant features is one of the fundamental problems of dealing with pattern classification problems [2]. However, owing to the ease of data collection, new datasets easily consist of over 1000 features. Some or many of such features may be redundant or irrelevant to the given classification problem. The feature selection problem is to select a subset of *p* features from the original set of *n* features by eliminating irrelevant or redundant features according to a selection criterion and a search strategy. According to the selection criterion, feature selection can be categorized into wrapper, filter and embedded approaches [3-8]. Wrapper feature selection approaches combine both the feature selection and output of the classification system [4]. Most of the wrappers employ the Leave-One-Out searching strategy [23]. In each step, evaluates the training accuracy when one of the features is left out, and then removes the feature yielding the least reduction in training accuracy. The evaluation of feature relevance of filter approaches uses only information from the given dataset, i.e. input features and class labels. There are also some hybrid approaches, such as the feature selection using the localized generalization error model which calculate the relevance based on an estimated generalization error of a feature subset [9, 10]. On the other hand, decision trees employ an embedded approach which embeds the feature selection process into its training algorithm. Feature subsets selected by wrapper and embedded approaches usually yield higher classification accuracies. However, wrapper and embedded approaches are computationally expensive and classifier dependant.

According to the search strategy, filter feature selection methods could be categorized into exhaustive search, branch and bound [11], sequential forward/backward search, floating search [12], and score methods [13, 14]. Searching all possible combinations of *n* features to select the best feature subset is a NP-complete problem. The exhaustive method, therefore, usually is not a reasonable choice for dataset with large number of features. Sequential search method still has a large computational complexity for high dimensional data, but much smaller than that of the exhaustive one. Many score methods have been proposed in the literature to deal with these high dimension data, such as Fisher score [15], Pearson correlation coefficients, Laplacian score [13], and LSDF [14]. Score methods only compute a score for each feature according to a selection criterion, then selects the most discriminative *p* features according to their scores. It is the most efficient feature selection search method. When dealing with high dimensional data. Therefore, the combination of filter and score methods is preferable to large datasets. Many filter methods, such as Fisher Score, LSDF, and Laplacian score (supervised) seek for feature subsets that yield a small intra-class variance and a large inter-class variance.

Current methods made an assumption that all samples in the same class could be effectively represented by a single cluster (or prototype). However, this assumption may not be true in many real world applications [1, 24]. Therefore, we are motivated to propose the Intra-Prototype / Inter-Class Separability Ratio (IPICSR) model for semi-supervised feature selection problems in [1]. In contrast to existing methods, samples in one class are presented by several prototypes and the IPICSR prefers feature subsets yielding small intra-prototype distance and large inter-class distance. In supervised pattern classification problems, the separating

Acknowledgement: This work is supported by a 985 project, South China University of Technology.

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

margin between two classes could be more accurately obtained. More importantly, the separating margin is a more reliable measurement for the separability between two classes when compared to inter-class distance. We will demonstrate this idea later with a toy example. Therefore, we extend the IPICSR model to the Intra-Prototype / Class Margin Separability Ratio (IPCMSR) model in this work for supervised feature selection problems. The IPCMSR model is classifier independent and is scalable to large number of features.

The rest of the paper is organized as follows: The IPCMSR model and its feature selection algorithm are introduced in Section 2. We compared the IPCMSR model with Fisher score, Pearson correlation coefficients, Laplacian score (supervised) and the IPICSR model using 7 UCI datasets and 2 face image recognition problems. The experiment results are discussed in Section 3. We give a conclusion in Section 4.

II. Supervised Feature Selection with IPCMSR

Given a set of training samples: $X = \{x_1, x_2,\ldots, x_m\}$ and $Y = \{y_1, y_2,\ldots, y_m\}$ where X, Y and m denote the set of input vectors of training samples, the set of corresponding class labels and the number of training samples, respectively. Let $P = \{p_1, \mathrm{p}_2,\ldots, \mathrm{p}_m\}$ be corresponding prototype labels for training samples. Let f_{ri} denotes the r^{th} input feature of the i^{th} sample x_i, $i = 1,2,\ldots,m$; $r=1,2,\ldots,d$, where d denotes the total number of features. Then, the IPCMSR model defines the relevance of the r^{th} input feature as:

$$L_r = \frac{\sum\limits_{x_i, x_j \in same\ \ prototype} S_{ij}^{P}\left(f_{ri} - f_{rj}\right)^2}{\sum\limits_{x_i, x_j \in different\ \ classes} S_{ij}^{M}\left(f_{ri} - f_{rj}\right)^2} \tag{1}$$

$$S_{ij}^{P} = \begin{cases} e^{\frac{-\left\|x_i - x_j\right\|^2}{\delta}} & if\ x_i, x_j \in same\ \ prototype \\ 0 & if\ x_i, x_j \in different\ \ prototype \end{cases} \tag{2}$$

$$S_{ij}^{M} = \begin{cases} 0 & if\ x_i, x_j \in same\ class \\ 0 & if\ \left(x_i, x_j \in different\ class\right) \\ & and\ \left(x_i \notin KNNd\left(x_j\right)\ or\ x_j \notin KNNd\left(x_i\right)\right) \\ e^{\frac{-\left\|x_i - x_j\right\|^2}{\delta}} & if\ \left(x_i, x_j \in different\ class\right) \\ & and\ \left(x_i \in KNNd\left(x_j\right)\ and\ x_j \in KNNd\left(x_i\right)\right) \end{cases} \tag{3}$$

where S_{ij} measures the similarity between x_i and x_j, $S_{ij}{}^{P}$, $S_{ij}{}^{M}$, δ and $KNNd(x_i)$ denote the intra-prototypes similarity, class margin similarity, a pre-selected constant and denotes the set of k nearest neighbors of the training sample x_i belonging to different class, respectively. The closer x_i and x_j are, the larger S_{ij} is, and vice versa. Equation (1) yields a smaller value when the r^{th} feature is more relevant. The IPCMSR model favors features which samples within the same prototype located close together while the between class separating margin is large.

The Supervised Feature Selection with IPCSMSR (SFS-IPCMSR) finds prototypes of samples for each class by a standard hierarchical clustering method. The hierarchical clustering method could be replaced by other existing clustering methods and it is transparent to SFS-IPCMSR. Equation (4) is adopted as the distance measurement in the supervised hierarchical clustering algorithm.

$$d_{avg}\left(P_j, P_i\right) = \frac{1}{n_j + n_i} \sum_{x \in P_j} \sum_{x' \in P_i} \left\|x - x'\right\|, j \neq i \tag{4}$$

The hierarchical clustering algorithm outputs a tree structure known as the dendrogram [17] and the topology of dendrogram is a representation of the clustering process. Therefore, the dendrogram could be cut into any given number of clusters. The number of clusters is determined by finding the *knee point* of the curve of distance among clusters versus the number of clusters [18, 24].

Finally, we rank features by Equation (1) and the SFS-IPCMSR returns a list of feature ranking in ascending order of relevance (L_r value). Therefore, user could select features using the ranking list. The algorithm of the SFS-IPCMSR is as follows:

Algorithm: SFS-IPCMSR

1. Input: input features X, target outputs Y, number of nearest neighbors k and the constant δ;
2. Find prototypes information P by supervised hierarchical clustering algorithm;
3. For the r^{th} features, compute L_r using Equations (1);
4. Output: the list of ranking of features in ascending order of the values of L_r.

The number of features in the final feature subset could be decided by user. One may remove features with large L_r values as long as classification results are acceptable. If too many features are removed, the classifier will not be able to generalize the input-output mapping for the given supervised pattern classification problem.

III. Experimental Results

In this section, the SFS-IPCMSR is evaluated on four UCI datasets [19]: *Glass, Ionosphere, Mushroom* and *Sonar*. Table 1 shows the statistical information of these datasets. Twenty training and testing datasets are independently and randomly generated to run the experiment. For each dataset, 50% of samples are used for training and the rest are used as the testing dataset. The training and testing datasets are normalized to [0, 1] by a min-max normalization. The SFS-IPCMSR method is compared with the Laplacian Score (supervised), the Pearson correlation coefficients, the Fisher Score and the SFS-IPICSR in the paper [1]. In the Laplacian Score (supervised), nodes x_i and x_j are connected if they belong to the same class. Prototypes consisting of less than three samples in the supervised hierarchical clustering algorithm are abandoned because they are likely to be outliers. In the Laplacian Score (supervised), the SFS-IPCMSR, and the SFS-IPICSR, the δ value is selected as the average distance between a point and its i^{th} nearest neighbor where $i = (\log(m) + 1)$ [22]. The performances of all algorithms are measured by their testing accuracies on the testing datasets using their selected feature subsets. We employ the nearest neighborhood (1-NN) classifier with Euclidean distance as the classifier in the experiments.

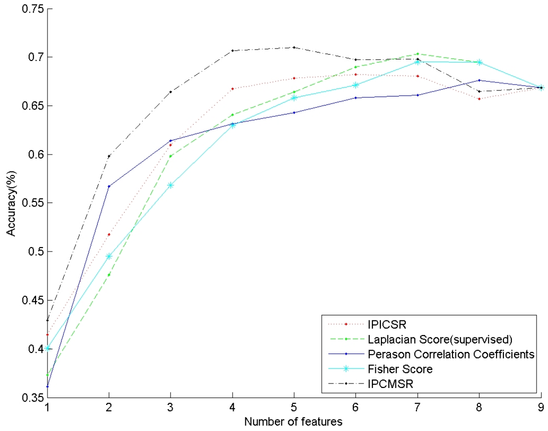

Figure 1. Average testing accuracies vs. different number of selected features for *Glass*

Figures 1 to 4 show plots of testing accuracies versus numbers of selected features for different datasets using different methods. The IPCMSR feature selection method achieves the highest testing accuracies on all datasets. Table 2 shows the averaged testing accuracy over different number of selected features. Table 2 indicates that the Pearson correlation coefficients method performs the worst because it can not capture feature relevance when features are not linearly correlated. On the one hand, the IPCMSR model outperforms Laplacian Score (supervised), Pearson correlation coefficients and Fisher Score methods in all datasets. The IPICSR model performs worse than the IPCMSR model while outperforming other existing feature selection methods. These indicate that the multi-prototypes representation of samples in the same class is a significant improvement to current single-cluster based filter type feature selection methods. The proposed IPCMSR model outperforms the IPICSR model because it is particularly designed for supervised classification problems by making use of class separating margin provided by the supervised training dataset. In supervised classification problems, experimental results support our assumption that features yielding a large class separating margin while a small intra-prototype distance are more relevant.

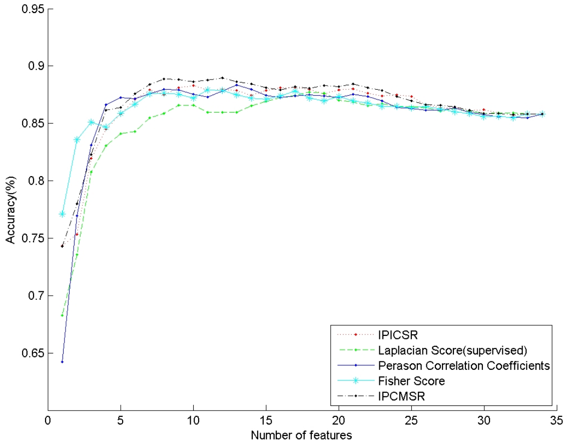

Figure 2. Average testing accuracies vs. different number of selected features for *Ionosphere*

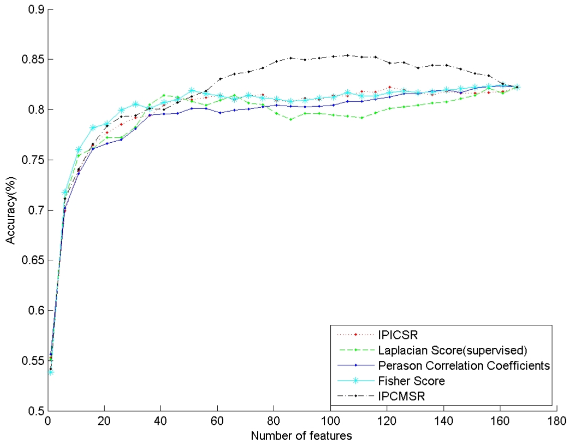

Figure 3. Average testing accuracies vs. different number of selected features for, *Mushroom*

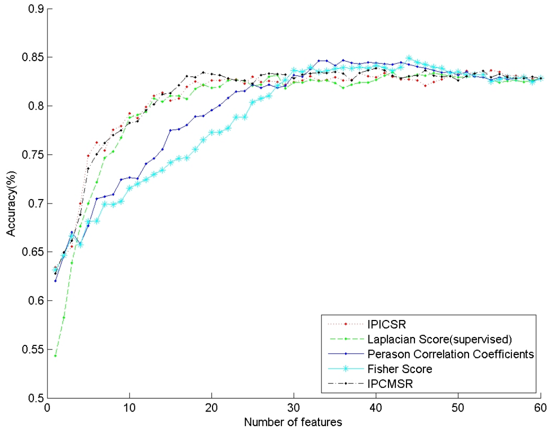

Figure 4. Average testing accuracies vs. different number of selected features for *Sonar*

Datasets	Number of Samples	Dimension	Number of classes
Glass	214	9	4
Ionosphere	351	34	2
Mushroom	476	166	2
Sonar	208	60	2

TABLE I. THE STATISTICS OF GLASS, IONOSPHERE, MUSHROOM, SONAR, VEHICLE, WDBC, AND WINE DATASETS

Datasets	Lap. Score	Corr.	Fisher Score	IPICSR	IPCMSR
Glass	61.21	60.89	60.91	61.95	**64.85**
Ionosphere	85.12	85.86	86.27	86.25	**86.67**
Mushroom	78.96	79.10	79.97	79.64	**81.49**
Sonar	80.11	79.65	78.85	80.83	**81.01**

TABLE II. AVERAGE TESTING ACCURACIES OF DIFFERENT FEATURE SELECTION METHODS

IV. CONCLUSION

We proposed the Intra-Prototypes / Class Margin Separability Ratio (IPCMSR) Model to select relevant features for supervised classification problems. The IPCMSR model prefers features that yield a small intra-prototype distance while yielding a large margin distance between classes. Four UCI datasets are used to examine the performance of the IPCMSR model. Experimental results show that the proposed IPCMSR feature selection method achieves highest testing accuracy than other off-the-shelf filter type feature selection methods. This supports our intuitive idea that real world datasets may consists of more than one prototype in each class of samples. Moreover, the proposed IPCMSR model outperforms the IPICSR model which is designed for semi-supervised learning and ignores class margin information.

One of the important future works is to determine a stopping criterion for determining the number of features in the final feature subset. One of the possible ways may be incorporate the localized generalization error model in [10] to estimate the generalization capability of those intermediate feature subsets and select the one yielding lowest localized generalization error. Further investigation is needed to merge these two models.

REFERENCES

[1] D.S. Yeung, J. Wang, W.W.Y. Ng, "IPIC Separability Ratio for Semi-Supervised Feature Selection", Accepted by International Conference on Machine Learning and Cybernetics, 2009

[2] R. O. Duda, P. E. Hart and D. G. Stork, "Pattern Classification", 2nd edition, Wiley-Interscience, 2000

[3] I. Guyon and A. Elisseeff. "An introduction to variable and feature selection", Journal of Machine Learning Research, vol. 3, pp. 1157-1182, 2003.

[4] R. Kohavi and G. H. John. "Wrappers for feature subset selection", Artificial Intelligence, vol. 97, pp. 273-324, 1997.

[5] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning" Artificial Intelligence, vol. 97, pp. 245–271, 1997.

[6] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," IEEE Trans. on Neural Networks, vol. 5, pp.537-550, 1994.

[7] N. Kwak and C-H. Choi, "Input Feature Selection by Mutual Information Based on Parzen Window", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, pp. 1667-1671, 2002.

[8] H. Peng, F. Long, C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27,, pp. 1226-1238, 2005.

[9] D. S. Yeung, W. W. Y. Ng, D. Wang, E. C. C.Tsang, X-Z. Wang, "Localized Generalization Error Model and Its Application to Architecture Selection for Radial Basis Function Neural Network", IEEE Trans. on Neural Networks, vol. 18, pp. 1294-1305, 2007

[10] W. W. Y. Ng, D. S. Yeung, M. Firth, E.C.C. Tsang and X-Z. Wang, "Feature selection using localized generalization error for supervised classification problems using RBFNN", Pattern Recognition, vol. 41, pp. 3706 – 3719, 2008

[11] B. Yu, and B. Yuan, " A more efficient branch and bound algorithm for feature selection", Pattern Recognition, vol. 26, pp. 883-889, 1993

[12] P. Pudil, J. Novovicova, and J. Kittler, " Floating search methods in feature selection", Pattern Recognition Letters, vol. 15, pp. 1119-1125, 1994

[13] X. He, D. Cai, and P. Niyogi. "Laplacian score for feature selection", Proc. of Advances in Neural Information Processing Systems, 2005.

[14] J. Zhao, K. Lu, X. He, "Locality sensitive semi-supervised feature selection", Neurocomputing, 71, pp. 1842-1849, 2008

[15] C.M. Bishop, "Neural Networks for Pattern Recognition", Oxford University Press, 1995.

[16] F.R.K. Chung, Spectral graph theory, AMS, 1997.

[17] S. Everitt, S. Landau.and M. Leese, "Cluster analysis", London: Hodder, 2001.

[18] Salvador, S., Chan, P. "Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms", Proceedings of the 16th IEEE international conference on tools with AI, pp. 576–584, 2004.

[19] C. Blake, E. Keogh, and C.J. Merz. UCI repository of machine learning databases. [http://www.ics.uci.edu/~mlearn/MLRepository.html], Department of Information and Computer Science, University of California, Irvine, 1998.

[20] F. Samaria, and A. Harter. “Parameterisation of a Stochastic Model for Human Face Identification”, Proceedings of 2nd IEEE Workshop on Applications of Computer Vision, pp. 138 - 142 1994

[21] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman, “From few to many: Illumination cone models for face recognition under variable lighting and pose”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, pp. 643–660, 2001.

[22] U. Von Luxburg, “A Tutorial on Spectral Clustering”, Statistics and Computing vol. 17, pp. 395-416, 2007

[23] J. Bi, K.P. Bennett, M. Embrechts, C.M. Breneman and M. Song, “Dimensionality reduction via sparse support vector machines”, Journal of Machine Learning Resaerch, pp. 1229 – 1243, 2003

[24] D.S. Yeung, D. Wang, W.W.Y. Ng, E.C.C. Tsang and X-Z Wang, “Structured Large Margin Machine: Sensitive to Data Distribution”, Machine Learning, vol. 68, pp. 171 – 200, 2007