Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing, 4-5 November 2002

A FUZZY CLUSTERING BASED ALGORITHM FOR FEATURE SELECTION

HAO-JUN SUN [(a)], SHENG-RUI WANG[(b)], ZHEN MEI[(c)]

[(a)]Université de Sherbrooke, Sciences/DMI, Sherbrooke, Qc, Canada J1K2R1
[(b)]School of Computer Science, University of Windsor, Windsor, On, Canada N9B3P4
[(c)]Manifold Data Mining, 501 Alliance Av., Toronto, ON, Canada M6N2J1

E-MAIL: {[(a)]sun, [(b)]wang}@dmi.usherb.ca [(c)] zhen@manifolddatamining.com

Abstract:

This paper deals with a wrapper approach to the problem of feature selection for classification. Based on fuzzy clustering, we develop a new algorithm that operates by testing the error between the cluster structure of the subspace data set and the class structure of the original data set. The true number of clusters in the subspace data set introduces accurate cluster structure information. The classification error rate, based on the difference between the number of clusters in the subspace data set and the number of classes in the original data set, provides a fair evaluation of how well the subset of features represents the original feature set. The experimental results show the advantage of our new algorithm.

Keywords:

Feature selection; Classification; Classification error rate; Fuzzy C-Means clustering.

1 Introduction

The problem of feature selection for classification is defined as follows: Given a set of features, select the subset that performs the best under some classification system. Feature selection can not only reduce the cost of recognition by reducing the number of features that need to be collected, but in many cases it can also provide better classification accuracy due to the effect of finite sample size effect [1]. Using a subset of features can increase the understandability of the acquired knowledge. Feature selection can help data visualization by reducing the number of dimensions.

Many methods are used for feature selection. Dash and Liu summarized these methods[2]. Feature selection involves: generating the subset of features and evaluating them. Three major strategies can be adopted in generating the subset of features: 1. Completes strategy involves examining all possible combinations of features, which becomes too expensive if feature set is large; 2. Heuristic strategy uses certain guideline to control the selection processing; it is simple to implement and produces rapid results[3,4]; 3. Random strategy selects feature randomly (probability approach). Five types of function are often used to evaluate feature subsets: 1. distance measures; 2. information measures; 3. dependence measures; 4. consistency measures; and 5. classification error rate measure.

Considering all of these methods and evaluation functions, the goal of feature selection can also be stated as finding the subset of features which is the most "structurally similar" to the original feature set. The "structural similarity" of two feature sets can be described by the cluster structure of two data sets. Dy and Brodley examined feature selection wrapped around expectation maximization (EM) clustering with order identification[5]. They introduced the clustering algorithm (EM) into the feature selection problem for unsupervised learning. For the classification problem, however, little attention has been paid to the role of clustering methods in feature selection. The difficulty stems from the complexity and inaccuracy of clustering algorithms when the number of clusters is not known.

In this paper, we propose a wrapper approach to feature selection using an efficient clustering technology. The approach is based on the fact that the selected feature subset is "structurally similar" to the original feature set. Based on an efficient clustering algorithm we presented recently[10], we propose here a novel algorithm for feature selection by focusing on the structural similarity in the selection process. We define a classification error rate for evaluating the subset of features. Extensive test results derived by applying the new algorithm to two artificial data sets and an ensemble of real-world data sets are reported.

This paper is organized as follows. In Section 2, we briefly introduce the clustering algorithm, which is based on the model selection strategy, and describe the new feature selection algorithm based on the wrapper approach. Experimental results are given in Section 3. The last section presents our conclusions.

0-7803-7508-4/02/$17.00 ©2002 IEEE

2 Feature Selection

In this section, we briefly introduce the clustering algorithm, which is based on the Fuzzy C-Means (FCM) and the model selection strategy for determining the number of clusters when clustering a given data set. Then, we introduce our new method for feature selection, which is based on cluster structure. Its two main steps are a generation procedure and a result evaluation procedure.

Fuzzy Clustering Algorithm

The FCM algorithm dates back to 1973. Many derivatives have been proposed with modified definitions for the norm and the prototypes for cluster centers[8,7,6]. FCM-based algorithms are the most widely used fuzzy clustering algorithms in practice.

The basic FCM algorithm can be formulated as follows:

$$Minimize\{J_m(U,V;X)=\sum_{k=1}^{n}\sum_{i=1}^{C}u_{ik}^{m}\parallel x_k-v_i\parallel_A^2\}$$

Where n is the total number of data vectors in a given data set and C is the number of clusters; $X=\{x_1,x_2,...x_n\}\subset R^s$ and $V=\{v_1,v_2,...v_n\}\subset R^s$ are the feature data and cluster centers; and $U=(u_{ki})_{n\times C}$ is a fuzzy partition matrix composed of the membership of each feature vector x_k in each cluster i. u_{ki} should satisfy $\sum_{i=1}^{C}u_{ik}=1$ for $k=1,2,...,n$ and $u_{ik}\geq 0$ for all $k=1,2,...n$ and $i=1,2,...C$. The exponent $m>1$ in $J_m(U,V)$ is a parameter, usually called a fuzzifier. The basic FCM algorithm can be found in many text books and papers[8].

In FCM, the number of clusters is a key parameter. In practice, the first job of clustering analysis is to determine the number of clusters. Indeed, determining the number of clusters is one of the most difficult problems in clustering analysis. A simple and useful strategy is model selection[1]. The main idea is to test all of the possible number of clusters, evaluating each result with a validity index and choosing the best as the optimal number of clusters. Based on this idea and FCM, we recently proposed [10,11,14] an efficient algorithm for determining the number of clusters. The algorithm searches from $C_{\min}$, the minimum number of clusters, to $C_{\max}$, maximum number of clusters. In each step, c, it clusters the data to c clusters using FCM, evaluates the result with a validity index, and splits the "worst" cluster into two clusters as the initialization for the next step. Finally, it chooses the optimal number of clusters based on the best validity index value. The FCM-Based Splitting.

Algorithm for determining the number of clusters is as follows:

FBSA Algorithm: (FCM-based Splitting Algorithm)

1. Choose $C_{\min}$ and $C_{\max}$.
2. $C_{\min}$ Initialize cluster centers V.
3. For $c=C_{\min}$ to $C_{\max}$:
   - 3.1 Apply the basic FCM algorithm to update the membership matrix U and the cluster centers V until convergence is obtained.
   - 3.2 Compute a validity value $V_d(c)$.
   - 3.3 Compute a score $S(i)$ for each cluster; split the worst cluster.
4. Compute c_f such that the cluster validity function $V_d(c_f)$ is optimal.
5. Reload the data set, and apply FCM with the optimal cluster number c_f.

There,

$$S(i)=\frac{\sum_{k=1}^{n}u_{ki}}{number\_of\_data\_in\_cluster\_i}$$

is a function that evaluates cluster i to select the "worst" cluster. We have defined a new validity index V_{WSJ} that is much more efficient than the existing ones when dealing with overlapping clusters. For detail see[10,11]. In the proposed algorithm for feature selection, we use V_{WSJ} as the validity function $V_d(c)$.

Feature Selection Procedure

We adopted the heuristic strategy for generating a feature subset. The goal of the procedure is to wrap the feature subset based on the clustering algorithm.Unlike the filter approach, which attempts to assess the merit of features from the data alone, the wrapper approach conducts a search for a good subset using an induction algorithm as part of the evaluation function[9]. The basic idea of our algorithm is to evaluate each subset T_i by a clustering process and then to evaluate a criterion defined by the classification error rate. The Greedy technique will be used in the search procedure. As we know, searching the entire feature subset space will lead to a $O(n^2)$ computation problem. In order to solve the computation problem, we use a multi-steps search process. Each step tests each remaining feature and chooses the best one to add to the selected subset. The newly selected feature is the most ``combinable" with those already selected. In other words, combining the new feature with the existing selected subset should lead to a lower classification error rate and this error should be the lowest among all the errors resulting from combining one non selected feature with the selected subset. The search process stops when adding any of the remaining features to the selected subset would yield an increase in the classification error rate.

FSBC (Feature Selection Based Clustering) Algorithm:

1. Set selected subset *SS* to empty, and $cr = 1$.
2. For any feature f_i, which is not in *SS*,
   - 2.1 Let $T_i = SS \bigcup \{f_i\}$, and create a new subspace data set SP_i using the features of T_i.
   - 2.2 Call FBSA on SP_i to determine the number of clusters and produce clustering results.
   - 2.3 Compute the classification error rate *R(i)* based on the subspace SP_i (defined by T_i).
3. $R(j) = \min_{f_i \in SS} R(i)$ with the lowest classification error rate.
4. IIf $R(j) < cr$ then $cr = R(j)$; $SS = T_j$; goto step 2; else output *SS*, stop.

The classification error rate in the feature selection algorithm is defined as follows: Suppose C is the number of classes in the original data set, S is a subset of features, $SP(S)$ is the subspace data set formed by S, $K(S)$ is the number of clusters in $SP(S)$, and $P(S,k,i)$ is the number of objects in cluster k of $SP(S)$ belonging to the class labeled i in the original data set. According to the majority rule,

$$CP(S,k) = j \mid P(S,k,j) = \max_{1 \le i \le c} \{P(S,k,i)$$

indicates that cluster k is related to class j or the main class label of cluster k is class j. Consequently

$$EC(S,k) = \sum_{i=1, i \neq CP(S,k)}^{n} P(S,k,i)$$

indicates the discrepancy between cluster k and its main class label. The classification error rate of S is defined as

$$R(S) = \sum_{k=1}^{K(S)} EC(S,k) / N$$

where N is the number of objects in the data set.

The value of $R(S)$ shows the accuracy of the representation of the original class information using the data set corresponding to the subset of features. The lower the value of $R(S)$, the better the representation. This means the cluster structure of the subspace data set is "similar" to the class structure of the original data set.

This structural similarity can be interpreted in a more intuitive way. Since the goal of feature selection is to better represent class information, we expect that the selected subset of features leads to a cluster structure with each cluster corresponding as closely as possible to a single class in the original data set. More than one cluster may correspond to a single class. However, the case where one cluster corresponds to multiple classes should be avoided. The classification error rate defined above allows us to grade each subset and distinguish the different cases. Obviously, this evaluation relies on accurate assessment of the structure of the data set corresponding to the subset of features. The use of an efficient clustering algorithm distinguishes our feature selection algorithm from existing ones.

3 Experimental Results

In this section, we will report experimental results on three data sets, of which one comes from the public domain, one is generated using a mixture of Gaussian distributions, and one is a real world data set. The first data set is Corral[12]. This data set has 32 instances. It contains two classes and six Boolean features $(A_0, A_1, B_0, B_1, I, C)$, of which feature I is irrelevant, feature C is correlated to the class label 75\% of the time, and the other four features are relevant to the Boolean target concept: $(A_0 \wedge A_1) \vee (B_0 \wedge B_1)$. In[2], Dash and Liu tested the data using eight different feature selection algorithms. A few of them correctly select the actual subset (A_0, A_1, B_0, B_1), while most produce a subset including C or I. Although the data has not clear class structure, our algorithm results in a final selected feature subset including $\{B_1, B_0, A_1\}$. This result shows that all of the features selected are important, although one important feature is bypassed. Table 1 shows results at each selection step.

Table 1. Selection results for DataSet1

	subset(SS)	(f_i)	R(s)
Step1	Ø	B_1	5/32
Step2	$\{B_1\}$	B_0	5/32
Step3	$\{B_1, B_0\}$	A_1	3/32
Step4	$\{B_1, B_0, A_1\}$	C	5/32

The second data set is generated using a mixture of Gaussian distributions. It contains 250 data points and has ten features $\{x_1, x_2, ..., x_{10}\}$. The first three features are significant. The subspace data set corresponding to the first three features $\{x_1, x_2, x_3\}$ is a mixture of five Gaussian components. The other features are as follows. $x_6 = 2 * x_1, x_7 = 4 * x_2, x_8 = 5 * x_3$ are three relevant features x_4 and x_5 are white-noise uniformly distributed variables. x_9 and x_{10} are "Gaussian noise". They are normal distributions and independent from each other. The class label is based on the first three features. There are 50 data points in each class. Due to the noise and excrescence features, classifying the data set using all features would result in a classification error rate of 178/250. Furthermore, this result does not indicate the class property of the data. By applying our new feature selection algorithm to the data set, the classification error rate is decreased remarkably, reaching 17/250. Table 2 shows the selection results. The selected feature subset is $\{x_2, x_3, x_1\}$.

Table 2. Selection results for DataSet2

	subset(SS)	(f_i)	R(s)
Step1	Ø	x_2	5/32
Step2	$\{x_2\}$	x_3	5/32
Step3	$\{x_2, x_3\}$	x_1	3/32
Step4	$\{x_2, x_3, x_1\}$	x_7	5/32

The third experiment was done on feature sets extracted from an MSTAR small vehicle target/shadow image database. These features include moment, surface, shape, perimeter, Fourier descriptor, complexity, etc. We calculated a total of 20 features for each target and 20 features for the shadow. The feature vectors were previously grouped according to the orientations of the target. (Details about the image segmentation algorithm were presented in[15].) There are 3 classes of targets. The aim of the feature selection algorithm is to find appropriate features to aid in solving the target classification problem. Here we test 11 data sets, each of which contains the observation data from an orientation.

In this problem, we do not know which features are the best for targets. The features may play different roles in different target/ orientation combinations, so different target/ orientation combinations may need different features. Using all features in classification lead to inaccurate classification (average classification error rate is 44.6%) and high time cost.

Table 3 compares the time cost and classification error using selected feature subsets and using all features. Our new feature selection algorithm results in an efficient feature subset for classification of each of the data sets. The number of selected features is between 3 and 6 of the 40 features. This leads to reduced cost in terms of time and memory (85% lower using the selected features than when all features are used). The classification accuracy rises observably. The average classification error rate is down to 23% from 44.6%. Furthermore, the results show that some features are important in practice. For example, f_{16} is selected in most data sets, f_1 and f_{18} are also often selected. This means they are relevant for the targets.

Table 3. Comparison of results for third experiment

	All Feature		Selected Feature		
	Time cost	Error rate	Selected feature	Time cost	Error rate
Data1	21	67/181	$\{f_{16}, f_4, f_1\}$	1.5	28/181
Data2	144	65/165	$\{f_{16}, f_{18}, f_{40}, f_{26}\}$	14	42/165
Data3	103	89/188	$\{f_{36}, f_{18}, f_{16}, f_{27}, f_{39}\}$	15	45/188
Data4	36	63/133	$\{f_{16}, f_{36}, f_{19}, f_{18}\}$	3.6	24/133
Data5	7	30/63	$\{f_{16}, f_1, f_{20}, f_{30}, f_2\}$	0.9	16/63
Data6	4	13/38	$\{f_{19}, f_{15}, f_{37}\}$	0.3	8/38
Data7	6	21/51	$\{f_{16}, f_{25}\}$	0.3	12/51
Data8	16	39/71	$\{f_{16}, f_{25}, f_1\}$	1.3	24/71
Data9	22	47/118	$\{f_{20}, f_{17}, f_{39}, f_{23}, f_5\}$	5	22/118
Data10	53	101/180	$\{f_{16}, f_1, f_{38}, f_{40}, f_{24}\}$	5.2	53/180
Data111	34	107/229	$\{f_{19}, f_5, f_1, f_{25}\}$	3.5	51/229
average	-	0.446	-	-	0.23

4 Conclusions

We have presented a wrapper approach to feature selection using fuzzy clustering, and proposed a new feature selection algorithm (FSBC) based on a clustering method. The particularity of this algorithm can be summarized as follows: 1. the true number of clusters in the subspace data set, for use in determining the cluster structure of the subset of features; 2. the classification error rate when the subspace data set and the original data set contain different numbers of clusters (classes), for use in comparing the cluster structure information of a subspace data set and the class structure information of the original data set. We are currently carrying out an evaluation of the new algorithm, including comparison with existing algorithms and testing on different types of data sets.

References

[1] A. Jain and D. Zongker. Feature Selection: Evaluation, Application, and Small Sample Performance. IEEE Trans. On Pattern Analysis and Machine Intelligence, 19(2):152-157 Feb. 1997.

[2] M. Dash and H. Liu. Feature Selection for Classification. Intelligent Data Analysis, 1(1997):131-156, 1997.

[3] K. Kira and L.A.Rendell. The feature selection problem: Traditional methods and a new algorithm. Proceedings of Ninth National Conference on Artificial Intelligence, 129-134, 1992.

[4] C. Cardie. Using decision trees to improve case-based learning. In Proceedings of Tenth International Conference on Machine Learning, 25-32, 1993.

[5] J. Dy and C. Brodley. Feature Subset Selection and Order Identification for Unsupervised Learning. In Proceedings of the 17th International Conference on Machine Learning, 247-254, 2000.

[6] R. N. Davé and K. Bhaswan. Adaptive Fuzzy c-Shells Clustering and Detection of Ellipses. IEEE T. on Neural Networks, vol. 3(5): 643-662, 1992.

[7] R. Krishnapuram, O. Nasraoui, and H. Frigui. The Fuzzy C Spherical Shells algorithms: a new approach. IEEE Transactions on Neural Networks, 3(5):663-671 Sept. 1992.

[8] J. Bezdek. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.1981.

[9] R. Kohavi and G. John. The Wrapper Approach, in Liu, Huan and Motoda, Hiroshi (eds.): Feature Selection for Knowledge Discovery and Data Mining, Kluwer International Series in Engineering and Computer Science, 1998, Chap. 1.

[10] H. Sun, S. Wang, and Q. Jiang. FCM-based Model Selection Algorithm for Determining the Number of Clusters. Research Report, University of Sherbrooke, No.277 2001.

[11] H. Sun, S. Wang, and Q. Jiang. A new validation index for determining the number of clusters in a data set. In Proceedings of IJCNN, (Washington D.C, USA): 1852-1857, July 2001.

[12] G.H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. Proceedings of the Eleventh International Conference on Machine Learning, 121-129, 1994.

[13] A. K. Jain and R.C. Dubes. Algorithm for Clustering Data. Prentice Hall. Englewood Cliffs, NJ, 1988.

[14] S. Wang, H. Sun, and Q. Jiang. New FCM-based Algorithm for Finding the Number of Clusters. 8th International Conference on Neural Information Processing (ICONIP2001), (Shanghai, China): 564-569, Nov. 2001.

[15] E. Aitnouri, S. Wang and D. Ziou. Segmentation of small vehicle targets in SAR images, Proceedings of Automatic Target Recognition XII, paper 4726-04, SPIE AeroSense, Orlando, Florida, USA, 1-5 April 2002.