Cluster Pattern Matching Using ACO Based Feature Selection for Efficient Data Classification

Sreeja N.K.[1] and A. Sankar[2]

[1] Department of Computer Applications,
Sri Krishna College of Technology, Coimbatore-641042
sreeja.n.krishnan@gmail.com
[2] Department of Mathematics and Computer Applications,
PSG College of Technology, Coimbatore-641004
dras@mca.psgtech.ac.in

Abstract. Cluster Pattern Matching based Classification (CPMC) is a classification technique based on a similarity measure between the training instances and the unknown sample. An Ant Colony Optimization based feature selection is proposed to select the features. According to this approach, the training data set is clustered. The cluster to which the unknown sample belongs is found and each of the selected features of the unknown sample is compared with the corresponding feature of the training instances in the cluster and the class of the unknown sample is predicted based on majority voting of class labels having highest number of matching patterns. A probabilistic approach is used to predict the class label when more than one class label has the same majority. Experimental results demonstrating the efficiency of classification accuracy of CPMC are shown to prove that the proposed approach is better when compared to existing classification techniques.

Keywords: Classification, Pattern matching, Feature selection, Ant Colony Optimization, Clustering.

1 Introduction

Classification is the task of learning from instances which are described by a set of features and a class label. An unknown sample is an instance with a set of features whose class label is to be predicted. The result of learning is a classification model that is capable of accurately predicting the class label of unknown samples. There are several methods in literature which attempts to classify samples based on the patterns in the instances of the training set. One such classification approach is K-Nearest Neighbor (KNN). The drawback of standard KNN classifier is that it does not output meaningful probabilities associated with class prediction [4]. Therefore, higher values of k are considered for classification which provides smoothing that reduces the risk of over-fitting due to noise in the training data [3]. However, choosing higher value of k, leads to misclassification of samples present in the training dataset as shown in section 5.1. A Bayesian solution to this problem was proposed and is known as

V.V. Das and J. Stephen (Eds.): CNC 2012, LNICST 108, pp. 250–254, 2012.

Probabilistic k-Nearest Neighbor (PNN). However, it has been shown that PNN does not offer an improvement in accuracy over the basic KNN [2]. Han et al., [1] proposed a Query Projection Analytical Learning (QPAL) for classification. However, the drawback in this approach is that the training instances with less number of features matching with a query are also considered. An Ant Colony Optimization (ACO) is a swarm intelligence algorithm to solve optimization problems. In this paper, a novel algorithm called Cluster Pattern Matching based Classification (CPMC) using an Ant Colony Optimization (ACO) based approach of feature selection is proposed. Experimental results show that CPMC is efficient for classifying datasets. This paper is organized as follows. Section 2 describes clustering the instances of the training set. Section 3 describes cluster pattern matching based classification. Section 4 describes building a CPMC training model using ACO based feature selection. A comparison with existing methods is described in section 5. Finally, the conclusion is presented in section 6.

2 Clustering the Instances of the Training Set

The instances in the training set were clustered based on a feature chosen arbitrarily termed as cluster feature. The number of clusters was initially set to the number of distinct class labels in the training set. Initially, one instance from each class in the training set was placed in each cluster. The cluster feature value of these instances denotes the mean value of each cluster. Each time an instance was added to the cluster, the difference between the cluster feature value of each instance and the mean value of each cluster was found. The instance is added to the cluster which has a minimum difference value. The mean of the cluster feature value of all instances in the cluster denotes the mean value of the cluster. Also the minimum and the maximum value of the cluster feature value in each cluster were found. This denotes the cluster feature range value in each cluster.

3 Cluster Pattern Matching Based Classification

A novel approach called Cluster Pattern Matching based Classification (CPMC) is proposed to classify the unknown samples. The basic blocks of CPMC algorithm are detailed below.

3.1 Predicting the Class Label of the Unknown Sample

The difference between the mean value of each cluster and the cluster feature value of the unknown sample was found. The unknown sample may either belong to the cluster for which the difference is a minimum or to the cluster whose cluster feature range value contains the cluster feature value of the unknown sample. The similar class labels of the training instances in the cluster to which the unknown sample belongs were grouped and their count was found. This is termed as class label count. The unknown sample whose class label is to be predicted is given by $(x_1, x_2, x_3,\ldots,x_n)$ where $x_1,x_2, x_3,\ldots,x_n$ are the features. An ACO based feature selection method as discussed in section 4 was used to select the features in the training dataset for comparison with the unknown sample. Each of the selected feature value x_i of the instances of the training dataset in the cluster was compared with the corresponding feature value of the unknown sample. The number of features in the training instance whose value matches with the corresponding feature value of the unknown sample was counted and was denoted as feature match count. This was repeated for all training instances in the cluster to which the unknown sample belongs. The training instances in the cluster having the maximum feature match count were grouped. The class label of the unknown sample was predicted as the majority class label of the training instances in the cluster having maximum feature match count. If there were more than one majority class label, the probability of each class label was found by dividing the majority class label count by its corresponding class label count in the cluster. The class label of the unknown sample was predicted as the class label with highest probability.

4 Building a CPMC Training Model Using an Ant Colony Optimization Based Feature Selection

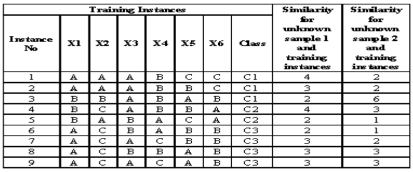
The classification model using CPMC was built by selecting features from the instances of the training dataset. An ACO method is proposed to find optimal subset of features for higher classification accuracy. The ant agent finds the solution by depositing pheromone. The pheromone deposition of the ant agent denotes the features of the instances in the training dataset to be compared with that of the unknown sample. The ant agent has a tabu list denoting the memory. The ant agent has a positive and negative position to deposit pheromone. The random number generated for the positive position must be between 0 and p, where p denotes the number of features. The random number generated for the negative position must be between 0 and q, where q denotes the maximum number of features in the pheromone deposition stored in the tabu list. The features in the pheromone deposition should not be repetitive. To deposit pheromone, the ant agent chooses two random numbers in the positive and negative position. Initially, the random number in the negative position is 0. Depending on the random number in the positive position, the ant agent chooses a group of positions randomly denoting the position of the features in the instances of the training set. The subset of features represents the pheromone deposition of the ant agent. The ant agent computes the energy value by finding the classification accuracy of CPMC for the features denoted by the pheromone deposition using 10 fold cross-validation. The energy value along with the pheromone deposition was stored in the tabu list. If the classification accuracy was less than 99%, the solution is not obtained and the ant agent moves to the next trail by updating the pheromone deposition. To update the pheromone, the ant agent chooses two random numbers in the positive and negative position. Based on the random number present in the positive and negative position, the ant agent chooses a group of positions. The group of positions chosen for negative position denotes the position of the features to be deleted from the pheromone deposition stored in the tabu list. Thus the ant agent updates the pheromone by either adding or deleting a subset of features or both to the features denoted by the pheromone stored in the tabu list. The energy value of the ant agent is evaluated. If the energy value is greater than the previous trail, the tabu list is updated with the pheromone deposition and the energy value. If the energy value is lesser than the previous trail, then the newly added or deleted features are ignored. The process was repeated until the energy value becomes a constant for a series of trails or the classification accuracy was greater than 99%. Once the solution is found, the classification model is built and the feature subset in the pheromone deposition denotes the features that are to be used for comparison by CPMC algorithm to classify unknown samples.

5 Comparison of CPMC with Existing Classification Techniques

To prove the efficiency of CPMC algorithm, an experiment was carried out using 18 classification datasets from the UCI machine learning repository available in the website http://www.ics.uci.edu/~mlearn/databases/.

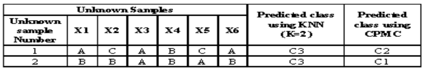
5.1 Comparison of CPMC vs KNN and Probabilistic KNN Approaches

Consider the training instances given in Fig. 1 and the unknown samples for classification given in Fig. 2 for a set of categorical values. With KNN (K=2), the class label of unknown sample 1 is predicted as C3 which has only 3 features similar to that of the unknown sample 1. Similarly, unknown sample 2 is classified as C3 for K=2. However the actual class of the unknown sample 2 is C1 as seen from the training set in Fig. 1 thereby leading to a misclassification.



Training Instances								Similarity for unknown sample 1 and training instances	Similarity for unknown sample 2 and training instances
Instance No	X1	X2	X3	X4	X5	X6	Class		
1	A	A	A	B	C	C	C1	4	2
2	A	A	A	B	B	C	C1	3	2
3	B	B	A	B	A	B	C1	2	6
4	B	C	A	B	B	A	C2	4	3
5	B	A	B	A	C	A	C2	2	1
6	A	C	B	A	B	B	C3	2	1
7	A	C	A	C	B	B	C3	3	2
8	A	C	B	B	A	B	C3	3	3
9	A	C	A	C	A	B	C3	3	3

Fig. 1. Similarity of instances using KNN



Unknown Samples							Predicted class using KNN (K=2)	Predicted class using CPMC
Unknown sample Number	X1	X2	X3	X4	X5	X6		
1	A	C	A	B	C	A	C3	C2
2	B	B	A	B	A	B	C3	C1

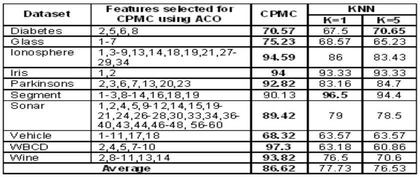
Fig. 2. Unknown samples

According to CPMC algorithm, the unknown sample number 1 is predicted to belong to C2 which has 4 features matching with the unknown sample. Also for unknown sample 2, the class label is predicted as C1. To further show that CPMC is efficient when compared to KNN, a comparison of CPMC with KNN (K=1 and K=5) using a 10 fold cross-validation is shown in Fig. 3.

Fig. 4 show the ten 10 fold cross-validation (performing 10 fold cross-validation 10 times) accuracy of CPMC with probabilistic KNN methods.

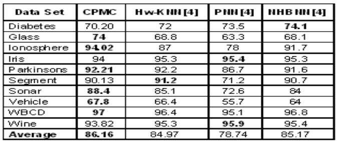
5.2 Comparison of CPMC vs Lazy Learning Approaches

Fig. 5 shows the classification accuracy of CPMC with QPAL and LBR [1].



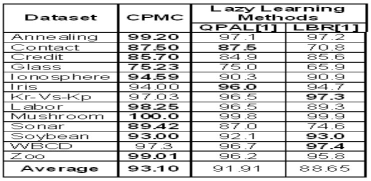
Dataset	Features selected for CPMC using ACO	CPMC	KNN	
			K=1	K=5
Diabetes	2,5,6,8	**70.57**	67.5	**70.65**
Glass	1-7	**75.23**	68.57	65.23
Ionosphere	1,3-9,13,14,18,19,21,27-29,34	**94.59**	86	83.43
Iris	1,2	**94**	93.33	93.33
Parkinsons	2,3,6,7,13,20,23	**92.82**	83.16	84.7
Segment	1-3,8-14,16,18,19	90.13	**96.5**	94.4
Sonar	1,2,4,5,9-12,14,15,19-21,24,26-28,30,33,34,36-40,43,44,46-48, 56-60	**89.42**	79	78.5
Vehicle	1-11,17,18	**68.32**	63.57	63.57
WBCD	2,4,5,7-10	**97.3**	63.18	60.86
Wine	2,8-11,13,14	**93.82**	76.5	70.6
Average		**86.62**	77.73	76.53

Fig. 3. Comparison of CPMC vs KNN using 10 fold cross-validation



Data Set	CPMC	Hw-KNN [4]	PNN [4]	NHBNN [4]
Diabetes	70.20	72	73.5	**74.1**
Glass	**74**	68.8	63.3	68.1
Ionosphere	**94.02**	87	78	91.7
Iris	94	95.3	**95.4**	95.3
Parkinsons	**92.21**	92.2	86.7	91.6
Segment	90.13	**91.2**	71.2	90.7
Sonar	**88.4**	85.1	72.6	84
Vehicle	**67.8**	66.4	55.7	64
WBCD	**97**	96.4	95.1	96.8
Wine	93.82	95.3	**95.9**	95.4
Average	**86.16**	84.97	78.74	85.17

Fig. 4. Comparison of CPMC vs probabilistic approaches



Dataset	CPMC	Lazy Learning Methods	
		QPAL[1]	LBR[1]
Annealing	**99.20**	97.1	97.2
Contact	**87.50**	**87.5**	70.8
Credit	**85.70**	84.9	85.6
Glass	**75.23**	75.0	65.9
Ionosphere	**94.59**	90.3	90.9
Iris	94.00	**96.0**	94.7
Kr-Vs-Kp	97.03	96.5	**97.3**
Labor	**98.25**	96.5	89.3
Mushroom	**100.0**	99.8	99.9
Sonar	**89.42**	87.0	74.6
Soybean	**93.00**	92.1	**93.0**
WBCD	97.3	96.7	**97.4**
Zoo	**99.01**	96.2	95.8
Average	**93.10**	91.91	88.65

Fig. 5. Comparison of 10 fold classification accuracy of CPMC vs Lazy learning approaches

6 Conclusion

An intelligent classification model using a cluster pattern matching based approach using Ant Colony Optimization based feature selection was built to classify data. It was shown that CPMC was better in classifying 14 out of 18 datasets such as annealing, contact, credit, glass, ionosphere, Kr-Vs-Kp, labor, mushroom, parkinsons, sonar, soybean, vehicle, WBCD and zoo when compared to KNN, probabilistic KNN and lazy learning approaches. However for iris and wine datasets the classification accuracy using CPMC was closer to the classification accuracy obtained in other methods.

References

1. Han, Y., Lam, W., Ling, C.X.: Customized classification learning based on query projections. Information Sciences 177, 3557–3573 (2007)
2. Manocha, S., Girolami, M.A.: An empirical analysis of the probabilistic k-nearest neighbour classifier. Pattern Recogn. Lett. 28, 1818–1824 (2007)
3. Ming Leung, K.: k-Nearest Neighbor Algorithm for Classification. Polytechnic University Department of Computer Science / Finance and Risk Engineering (2007)
4. Tomasev, N., Radovanovic, M., Mladenic, D.: A Probabilistic Approach to Nearest-Neighbor Classification: Naive Hubness Bayesian KNN. In: CIKM 2011, Glasgow, Scotland, UK (2011)