http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 5 (6): 1102-1105, 2006
ISSN 1812-5638
© 2006 Asian Network for Scientific Information

PAMC: Partitioning Around Medoids for Classification

D.K. Swami and R.C. Jain
Department of Computer Applications,
Samrat Ashok Technological Institute, Vidisha (MP)-464001, India

Abstract: Integration of Association Rule Mining and Classification has produced Associative Classification techniques which in many cases have shown better classification accuracy than conventional classifiers. Motivated by this study we explore integration of clustering and classification techniques and propose a clustering based classification approach in this study. Our algorithm is based on Partitioning Around Medoids (PAM) clustering algorithm. Clustering process is unsupervised in general. We develop a classifier model by making use of available class label knowledge of training examples during clustering process. In present study we find accuracy of results is near to other popular classification methods. At present our algorithm can be applied to objects whose features are represented by continuous attributes.

Key words: Data mining, classification, clustering

INTRODUCTION

Large collection of data through modern information systems provides opportunity and challenge for data analysis. Data mining algorithm have well taken up such challenges. Classification has been one of the key data mining tasks. In classification we are given a set of objects with known class labels in the form of records with several fields/attributes. This set of records is called training set. One of the attributes in the record contains class to which object belongs. The objective of classification is to build a model on training set to predict the class of future objects whose class label is not known. For building classifier models myriad techniques have been proposed and while there are in general, better approaches than others, there is no clear winner is term of accuracy, performance and usability given a particular domain of application.

Eager classification approaches are common. Eager classifiers construct a generalization model before receiving new samples to classify. Several popular classification methods such as decision trees, statistical models, neural networks (Lim *et al*., 2000) and classification based on association rules (Swami and Jain, 2005) are eager classification techniques. k Nearest Neighbour (kNN) classification (Cover and Hart, 1967) technique is lazy classification technique in which model learning time is less but prediction time is high as learning of model is needed for predicting class for each of the future objects.

Motivated by accuracy of results of classification methods based on integration of Association Rule Mining and classification techniques, initially in CBA and subsequently in CMAR, CPAR (Yin and Han, 2003) and other similar work, we have made an attempt to integrate clustering and classification. In this study we propose a new classification approach based on clustering method PAM (Kaufman and Rousseau, 1990). Our approach is somewhat similar to kNN classification, a lazy classification approach in which efficiency of classification rapidly decreases as the number of training instances in training set increase. When given a unknown sample, a kNN-classifier searches the feature space for k training samples that are closest to the unknown sample. Closeness between objects is defined in terms of some distance measure. The unknown object is assigned the most common class among its k nearest neighbors. Many attempts have been made to reduce the training instances for instance based classification (Han and Kamber, 2000). Our approach too makes predictions based on distance based closeness of unknown object with objects in the training dataset. But we develop an eager classification approach that over comes this drawback as classification model is created for once, it need not be retrained for prediction of class label for each new future object.

We mention here some study from current literature in which clustering is used for classification. In CBC (Zeng *et al*., 2003) approach some of the unlabeled datasets present in the training data are labeled using clustering for subsequently training classifiers. An

Corresponding Author: D.K. Swami, Department of Computer Applications, Samrat Ashok Technological Institute, Vidisha (MP)-464001, India

approach to improve the accuracy of simple classifiers through a clustering based preprocessing step is presented in (Vilalta *et al*., 2003). The present study differs from these as we do not use clustering as merely a preprocessing step, we develop a classifier based on clustering.

The proposed algorithm works in two steps

- Partition the training dataset into k clusters using PAM approach and build the classifier based upon class knowledge of training objects in each of the clusters.
- Do the class predictions for new objects based upon distance of new object from clusters and class distribution knowledge of training samples in each cluster.

This study makes following contributions

- It proposes a new way to build accurate classifiers. Experimental results show that accuracy of our method is near to accuracy of other popular classification methods.
- Our method works for training dataset where feature space contains numeric attributes, thus eliminating the preprocessing step of discretizing continuous attributes needed in classification approaches like decision tree induction and associative classification.

PARTITIONING AROUND MEDOIDS

In PAM (Kaufman and Rousseau, 1990), k partitions for n objects are formed. Initially randomly k medoids are chosen out of set of objects. Medoid representing a cluster is most centrally located object in the cluster. Each remaining object is clustered with the medoid to which it is the most similar based on the distance between the object and medoid. The strategy then replaces one of the medoids by one of the non medoids as long as the quality of resulting cluster is improving. This quality is estimated using a cost function that measures the average dissimilarity between an object and the medoid of its cluster.

In the iterative process a non-medoid object is randomly chosen for replacement with current medoids. Each replacement causes movement of some objects from one cluster to the other cluster. Each time a reassignment occurs a difference in square error E is contributed to the cost function. Therefore the cost function calculate the difference is square error value if a non-medoid object replaces current medoid. The total cost of swapping is the sum of costs incurred by all non-medoid objects. If the total cost is negative then replacement of medoid with non medoid object is good since the actual square error would be reduced. The process is iterated until good-replacements of medoids are found. In the end k-medoids are returned.

CLUSTER BASED CLASSIFICATION

The proposed algorithm is called PAMC (Partitioning Around Medoids for Classification). It consists of two parts

- PAMC_CWCD (PAMC Cluster Wise Class Distribution)
- PAMC_Classify

Notations used in the explanations of algorithms and pseudo codes presented in Fig. 1-3.

PAMC_CWCD: PAMC_CWCD algorithm is based on PAM approach of clustering. PAM is applied on training set to find best set of k-medoids to represent k clusters. Unlike clustering data, in classification dataset class labels of each of the objects are known. We perform following calculations :

Assume the training data set objects belong to finite number of classes c_1, c_2, c_3,… c_n. Suppose medoids of k clusters are Med_1, Med_2 …Med_k. We compute for each cluster, t_i total number of objects, separate counts of objects for each of the j classes s_{ij}.

Having s_{ij} and t_i, probability that an object in the cluster of medoid Med_i will be in class c_j, can be computed as $p_{ij} = s_{ij}/ t_i$. Knowledge of Med_i and p_{ij} for each of the clusters form the model for classification.

D	Training dataset
Med_i	A medoid object representing a cluster
O_u	object whose class label is unknown
O_i	ith object in training dataset, not selected as medoid
c_i	ith class label in the domain of class labels in training dataset
C_i	ith cluster
t_i	Total number of objects in i^{th} cluster
s_{ij}	Total number of objects bearing the class label c_j in C_i
p_{ij}	Probability that O_u be in C_i and bear the class label c_j

Fig. 1: Notations

```
Algorithm PAMC_CWCD
      input : training data set D
      output : classification model
begin
//k ≥ _number of classes in D
partition training data set D into k partitions
(clusters) using PAM and find Med_i, i = 1..k
for each partition C_i, i = 1 …k
   //total objects in each partition
   Compute t_i
   //compute class distribution
  for each class c_j, j= 1…n
     // total number of objects in each
     //partition C_i and for each class label c_j
     compute s_ij;
    // probability that an object in partition C_i
    // belongs to class label c_j;
     compute p_ij = s_ij / t_i;
   end for;
end for;
 set of  Med_i and p_ij form classifier model;
end .
```

Fig. 2: PAMC_CWCD algorithm

```
Algorithm PAMC_classify
       input :  classification model and unknown O_u
       output : predicted class of O_u

begin
for i = 1 …k
  // distance of unknown object from each medoid
  compute d_i = d (O_u, Med_i)
end for
for each c_j, j=1…n
  for each partition C_i : i=1…k
    compute sum_i = ∑ p_ij / d_i
  end for
end for
// track i for which sum_i is maximum
compute max (sum_i)
// return i^th class label for which sum_i is maximum
return c_i
end.
```

Fig. 3: PAMC_Classify algorithm

The PAMC_CWCD algorithm is presented in Fig. 2.

PAMC_Classify: An unknown object O_u (whose class label is not known) shall be assigned a cluster represented by object Med_h if distance of O_u with Med_h is minimum as compare to its distance with all other medoids i.e.,

$$D_h = d(O_u, Med_h) = \min_i d(O_u, Med_i),\ i=1..k$$

We observe that for unknown object, once a cluster is assigned then majority class present in this cluster can be predicted to be the class of O_u following the kNN approach. We also observe the fact that an unknown object O_u assigned a cluster say C_j shall have more chances of bearing the class label c_i if for cluster C_j the values of s_{ij} is maximum.

We compute

$$Sum_j = \sum pij / D_i,\ i = 1..k,\ i= 1..n$$

Where D_i are distances from medoids Med_i representing ith clusters and give the expression

$$e = \max_j (sum_j)$$

and conclude that the predicted class for O_u is class c_j with max value e.

This prediction is based on the fact that an unknown object's dissimilarity with representative medoids of cluster reduces the probability that object's class label can be predicted by majority class of that cluster. Taking summation improves the result as measures computed using all training data set are used.

The PAMC_Classify algorithm is presented in Fig. 3.

Improving classifier by making use of class knowledge in training data: While applying PAM, non selected object is assigned to a cluster represented by medoid Med_m, to which it is most similar, the measure of dissimilarity being Euclidean distance. At this step we introduce a parameter pfactor ≥1 as penalty factor and make use of class knowledge of training data. If class of a non selected object and a representative medoid is not same, the distance between object and medoid is multiplied by pfactor, this increases the dissimilarity between objects of different classes. This helps improve the accuracy of the classifier. An optimum value of pfactor can be found empirically based upon accuracy of classifier for a training data set.

RESULTS

To evaluate the accuracy of PAMC we chose Diabetes data set from UCI machine learning Repository (Hettich and Bay, 1999). This data set contains 786 records, with 8 attributes and 2 classes. All the eight attributes being numeric, suits to our algorithm as distance metric being Euclidean distance.

Test No.	k	Factor	Accuracy
1	2	1.5	61.71
2	2	2.0	66.42
3	2	2.5	70.28
4	2	3.0	68.85
5	3	1.5	63.14
6	3	2.5	64.57
7	3	3.0	66.00
8	4	1.5	66.42
9	4	2.0	70.28
10	4	2.5	69.71
11	4	3.0	71.71

Fig. 4: Experimental results

For diabetes data set popular classification methods C4.5, CBA, CMAR and CPAR gives 74.2, 74.5, 75.8 and 75.1% accuracy respectively (Yin and Han, 2003). Present results nearing these results that conform our approach.

We used randomly chosen 90% objects from training classifier and remaining 10% for testing the classifier.

Results are presented in Fig. 4.

CONCLUSIONS

In this study we proposed a new clustering based classification approach PAMC: Partitioning Around Medoids for Classification. This method exploits class label knowledge of objects in training dataset during clustering and forms an eager classifier based on PAM clustering approach.

Present study on diabetes database of UCI machine learning database repository show that the accuracy of our method is near to other popular classification method like C4.5 and CMAR. That paves the way for exploring integration of clustering and classification approaches for better results.

REFERENCES

Cover, T. and P. Hart, 1967. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13: 21-27.

Han, J. and M. Kamber, 2000. Data mining: Concepts and Techniques. Morgan and Kaufmann Publishers.

Hettich, S. and S.D. Bay, 1999. The UCI KDD Archive [http://kdd.ics.uci.edu]. Irvine, CA: University of California,

Kaufman, L. and P. Rousseau, 1990. Finding Groups in Data. John Wiley and Sons, New York.

Lim, T.S., W.Y. Loh and Y.S. Shih, 2000. A comparison of prediction accuracy, complexity and training time of thirty-three old and new classification algorithms. Machine Learning, 40: 203-229.

Swami, D.K. and R.C. Jain, 2005. A survey of associative classification algorithms. ADIT. J. Eng., 12: 51-55.

Vilalta, R., M. Achari and C. Eick, 2003. Class Decomposition via Clustering: A new framework for low-variance classifiers. In Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM03), Melbourne, Florida, pp: 673-676.

Yin, X. and J. Han, 2003. CPAR: Classification based on Predictive Association Rules. In Proceeding of the International Conference on Data Mining, SDM. SIAM, pp: 331-335.

Zeng, H.J., X.H. Wang, Z. Chen, H. Lu and W.Y. Ma, 2003. CBC: Clustering Based Text classification requiring minimal labeled data. In Proceedings of ICDM, 2003, pp: 443-450.