International Journal For Technological Research In Engineering
Volume 4, Issue 8, April-2017 *ISSN (Online): 2347 - 4718*

SUPERVISED CLUSTERING FOR CLASSIFICATION WITH FEATURE SPACE HETEROGENEITY

R.Pavani[1], K.Mounasree[2], B.Kiranmai[3], J.V.Kirankumari[4]

Abstract: Feature space heterogeneity often exists in many real world data sets so that some features are different importance for classification over different subsets. Moreover, the pattern of feature space heterogeneity might dynamically change over time as more and more data are accumulated. In this paper, we develop an incremental classification algorithm, Supervised Clustering for Classification with Feature Space Heterogeneity (SCCFSH), to address this problem. In our approach, supervised clustering is implemented to obtain a number of clusters such that samples in each cluster are from the same class. After the removal of outliers, relevance of features in each cluster is calculated based on their variations in this cluster. The feature relevance is incorporated into distance calculation for classification. The main advantage of SCCFSH lies in the fact that it is capable of solving a classification problem with feature space heterogeneity in an incremental way, which is favourable for online classification tasks with continuously changing data. Experimental results on a series of data sets and application to a database marketing problem show the efficiency and effectiveness of the proposed approach.
Key words: Classification, Clustering, Heterogeneity, Feature Relevance.

I. INTRODUCTION

In classification problems, feature space heterogeneity is the phenomenon that a data set consists of some heterogeneous subsets, and the optimal features for classification are distinct over different subsets. The challenge of this problem is that we do not know how many heterogeneous subsets exist in the data set or which subset each sample belongs to. In the last decade, the problem of feature space heterogeneity in data has been addressed under different names, such as local feature relevance, case-specific feature weights, relevance in context, feature space and class heterogeneity, and attribute instability.

Feature space heterogeneity exists widely in various application fields of classification techniques, such as marketing, customs inspection decision, credit scoring, and medical diagnosis. For example, in marketing, a major concern of the market managers is to develop and implement efficient marketing programs by fully utilizing the customer databases and identifying the households that are most likely to be interested in the marketing programs. The above process can be formulated as a classification problem, in which the Features (attributes) are characteristics of the households such as demographic, psychographic, and behavioural information, and the target variable is whether a household responds to the marketing messages.

II. LITERATURE REVIEW

The phenomena that relevant features for classification vary across the data set have been observed by many researchers and practitioners. Until recently, a number of classification methods have been developed, which can be divided into two categories. In the first category, one of the best known methods is "bagging" [18]. In this approach, k subsets are generated by randomly sampling from the original set of samples. Consequently, relevant features might be different in the obtained subsets. Based on this approach, Puuronen. Proposed a Meta-Level Classification (MLC) method, which can be used to deal with the problem of feature space heterogeneity. MLC first divides the training sample set into some subsets and obtains the component classifiers based on these subsets. In the application phase, testing samples are put into the training sample set, and MLC dynamically selects the optimal component classifier for a testing sample by comparing the performance of different classifiers in its neighbourhood. Different from the method of sample partitioning, the Random Subspace Method (RSM) [20] divides the whole feature set into a number of feature subsets and constructs different classifiers based on the whole training samples with different feature subsets obtained. Feature space heterogeneity in testing samples is considered through synthesizing (usually by voting) the application results of all classifiers. These methods deal with the problem of feature space heterogeneity by firstly dividing sample set or feature set into different subsets in a random way and then training component classifiers in the subsets. These component classifiers are then combined for classification, mostly by major voting or selecting the optimal one. A major problem of these methods lies in the random set (sample set or feature set) partitioning, which may result in seriously biased component classifiers due to the feature redundancy and irrelevance in some subsets, especially for high dimensional data sets. In the second category, modified lazy learning methods are applied to classification problems with feature space Mathematical Problems in Engineering 3 heterogeneity. Friedman addresses the problem of feature space heterogeneity by investigating the variability of feature relevance in different data subsets. In his method, the local relevance of features in each subset is measured by the estimated reduction in classification error. Hastie and Tibshirani developed an adaptive form of nearest neighbour classification method for dealing with feature space heterogeneity. In their approach, distance metric for each sample is adaptively calculated in an iterative process using local discriminative information of features. Therefore, different relevant features are taken into account for classification in different subsets. Although both works report favourable results on their local approaches

compared to global ones, both of them are computationally expensive. Aredes and Vidal propose a locally weighted lazy learning approach for better classification accuracy. In their method, different samples would have different feature weights obtained by approximately minimizing the Leaving-One-Out (LOO) classification error of the given training set. However, the computational complexity of this method is high because of the gradient descent algorithm employed to search for the optimal weights. In spite of the fact that many researchers have been carried out for dealing with feature space heterogeneity in classification, we have not found any for incremental learning among them.

Researches on incremental classification are mainly focused on statistical methods, neural networks, and evolutionary algorithm. Instance-based learning, especially K-nearest neighbour (K-NN) learning, is a widely used nonparametric incremental classification approach where training or learning does not take place until a query is made. In contrast to complex learning algorithms such as neural networks or support vector machines, K- NN learning does not require a complex function fitting process or model training procedure. Thus, it is easy to do incremental learning. Nevertheless, once a query point with unknown class label is presented, conventional K-NN learning traverses the whole data set to find the K nearest neighbours of the query point. Therefore, the computational time and requirement of computer storage space of K-NN are not scalable to large amounts of data.

To solve this problem, Li and Ye propose a data mining algorithm based on supervised clustering to learn data patterns and use these patterns for classification. This algorithm enables a scalable and incremental learning of patterns from data with both numeric and nominal variables. However, it calculates the feature relevance by using squared correlation coefficient between predictor variables and target variable over the entire data set, regardless of the possible heterogeneity that exists in the feature space.

The Proposed Approach the Supervised Clustering for Classification with Feature Space Heterogeneity (SCCFSH) proposed in this paper is based on the ECCAS. However, SCCFSH differs significantly from ECCAS in that it takes feature space heterogeneity into consideration. SCCFSH first divides the data set into a number of subsets in a supervised way and then explores the feature relevance in each subset obtained. The main steps of SCCFSH include grid-based supervised clustering, supervised grouping of clusters, removal of outliers, calculation of feature relevance in each cluster, and distance based classification.

III. PROPOSED WORK

Proposed Approach is to develop a Supervised Clustering for Classification with Feature Space Heterogeneity (SCCFSH) to address this problem and consists of four main steps: grid-based supervised clustering, supervised hierarchical grouping of clusters, feature relevance evaluation in each cluster, and weighted distance calculation for classification.

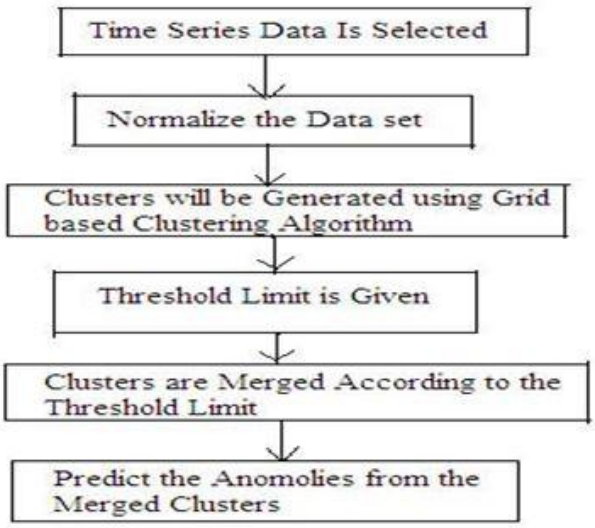

Fig 1: Flowchart of proposed work

- Grid-based supervised clustering: The grid-based supervised clustering procedure first divides the -dimensional space of samples into grid cells and then generates clusters within the grid cells
- Supervised hierarchical grouping of clusters: Supervised grouping of clusters plays an important role in the proposed SCCFSH. If some underlying clusters cover the area of several grid cells, grid-based supervised clustering would divide these large clusters into small clusters. Therefore, refinement of the clustering results is needed.
- Feature relevance evaluation in each cluster: After the above three steps, we could obtain a number of clusters, and samples in each cluster are with the same class label. Vucetic and Obradovic [29] argue that some features with the same values may result in quite different outputs (class labels) in different regions. Therefore, spatial characteristics of samples should be explored for better classification performance. Motivated by this statement, we investigate the feature relevance in the obtained clusters that represent different spatial distributions of samples.
- Weighted distance calculation for classification

Algorithm 1: Main procedure of supervised grouping of clusters
(INPUT: Clusters C1, C2... Ck obtained by grid-based supervised clustering.
OUTPUT: R = {New 1, New 2... CNew m}. (1) Set R =, R0 = {C1, C2... C3} and t=1;
(2) Among all possible pairs of clusters (Or, Os) in Rt−1 find the one, say (Oi, Oj), such that
d (Oi, Oj) = min r,s {d (Or, Os)};
(3) IF Label (Oi) = Label (Oj) THEN set Oq = Oi Oj and produce the new clustering
R = (Rt−1 − {Oi, Oj}) Oq; ELSE Rt = Rt−1 − {Oi, Oj},
R = R Oi Oj; Set t = t + 1;

(4) Repeat steps 2-3 until Rt = Rt−1, return R.

Algorithm 2: Main steps of SCCFSH
INPUT: Training data set S1 = {X1,X2,...,XN} with label set L = {y1, y2,...,yN}; Testing data set S2 = {XN+1,XN+2,...,XN+M} with unknown labels.
OUTPUT: Predicted labels ŷ N+1, ŷ N+2... ŷ N+M for S2.
(1) Apply the grid-based supervised clustering to S1 and L;
(2) Apply the algorithm shown in Algorithm 1 to Clusters C1, C2... Ck obtained in step 1;
(3) Calculate relevance of each feature Rk (d), (d = 1, 2 . . . p) in CNew 1, CNew 2... CNew m obtained in step 2;
(4) LOOP for each data point Xj, (j = N + 1, N + 2. . . N + M) in S2: (i). Calculate the distance between Xj , and CNew i , (i = 1, 2 . . . m) obtained in step 2, by using the distance metric defined in (6); (ii). Set y j̃ as the label of the nearest cluster ;
(5) Output the predicted labels y^ N+1, ŷ N+2... Y^ N+M for S2.

Fig 2. Visualising Raw Data set

Fig 3. Normalized Data set

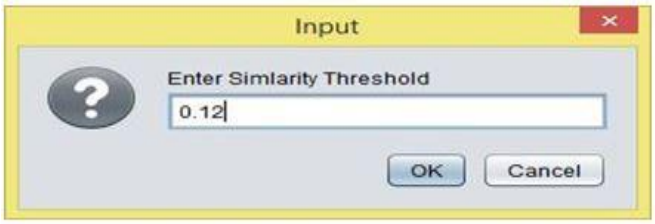

Fig 4: Selecting Threshold Limit

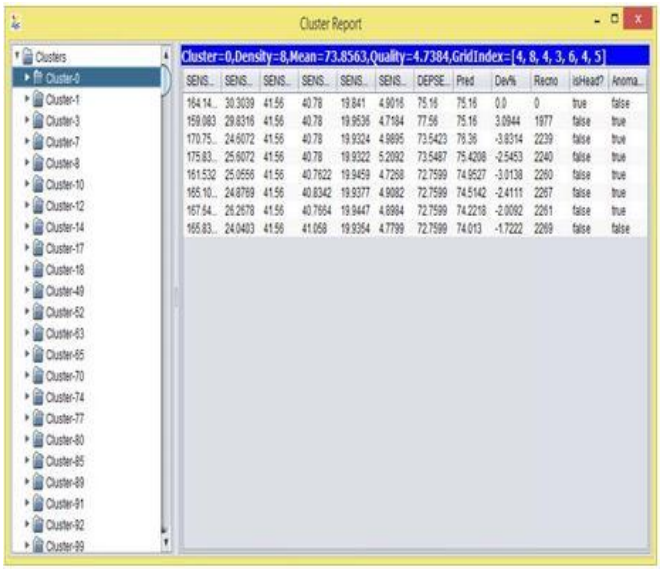

Fig 5: Retrieving Cluster Report after Merging

IV. CONCLUSION

In this paper, we develop a Supervised Clustering for Classification with Feature Space Heterogeneity (SCCFSH) to address the problem in a scalable and incremental way. In this study, we considered Time series dataset which can be normalized and then by applying Grid based supervised clustering method clusters will be generated. Based on the threshold value clusters will be merged. In spite of the fact that we only consider classification problems. By this an employee can easily identify the anomalies and generate cluster reports.

REFERENCES

[1] Y. Liu, Y. Liu, and K. C. C. Chan, "Dimensionality reduction for heterogeneous dataset in rushes editing," Pattern Recognition, vol. 42, no. 2, pp. 229–242, 2009.

[2] J.-T. Wong and Y.-S. Chung, "Analyzing heterogeneous accident data from the perspective of accident occurrence," Accident Analysis and Prevention, vol. 40, no. 1, pp. 357–367, 2008. Mathematical Problems in Engineering 9

[3] X. Li and N. Ye, "A supervised clustering and classification algorithm for mining data with mixed variables," IEEE Transactions on Systems, Man, and Cybernetics A, vol. 36, no. 2, pp. 396–406, 2006.

[4] Scrypnyk and T. K. Ho, "Feature selection and training set sampling for ensemble learning on heterogeneous data," Tech. Rep., DIMACS, 2003.

[5] M. K. Lim and S. Y. Sohn, "Cluster-based dynamic scoring model," Expert Systems with Applications, vol. 32, no. 2, pp. 427– 431, 2007.

[6] R. Paredes and E. Vidal, "Learning weighted metrics to minimize nearest-neighbour classification error," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 7, pp. 1100–1110, 2006.

[7] S.-U. Guan and S. Li, "Incremental learning with respect to new incoming input attributes," Neural Processing Letters, vol. 14, no. 3, pp. 241–260, 2001.

[8] S.-U. Guan and F. Zhu, "An incremental approach to genetic algorithms-based classification," IEEE Transactions on Systems, Man, and Cybernetics B, vol. 35, no. 2, pp. 227–239, 2005.

