ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

Novel Method to Construct a New Attributes For Classification with Mixed Dataset Using Cure and Rock Algorithms

Sagunthaladevi. S[1], Dr. Bhupathi Raju Venkata Rama Raju[2]

Research Scholar, Dept of Computer Science, Mahatma Gandhi University, Meghalaya, India[1]

Professor, Dept of Computer Science, IEFT College of Engineering, Villupuram, Tamilnadu, India[2]

ABSTRACT: Classification is a challenging task in data mining technique. The aim of Classification is to group the similar data into number of classifiers. Various classification algorithms have been developed to group data into classifiers. However, these classification algorithms work effectively either on pure numeric data or on pure categorical data, most of them perform poorly on mixed categorical and numerical data types in previous algorithm was used but it is not accurate for large datasets. The quality of a data space representation is one of the most important factors influencing the performance of a data mining algorithm. The attributes defining the data space can be insufficient, making it difficult to discover high quality knowledge. This paper presents a novel technique to solve the attribute construction problem of numeric and nonnumeric values with the help of clustering algorithms in three steps. ROCK and CURE clustering algorithms are included for processing in mixed datasets. These algorithms constructs a new attributes out of the original attributes of the data set and performing an important preprocessing step for the subsequent application of a data mining algorithm. Finally the clustering results on the categorical and numeric dataset are combined as a categorical dataset on which the algorithm is designed for one type of features to handle numeric and nonnumeric feature values.

KEYWORDS: Classification, Clustering, Prediction, CURE algorithm, ROCK algorithm

I. INTRODUCTION

Classification is a supervised learning technique in data mining where training data is given to classifier that builds classification rules. If test data is given to classifier, it will predict the values for unknown classes. Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before called prediction set, which contains the same set of attributes except for the prediction attribute. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how "good" the algorithm is. For example, in a medical database the training set would have relevant patient information recorded previously, where the prediction attribute is whether or not the patient had a heart problem. Assigning an object to a certain class based on its similarity to previous examples of other objects Basically classification is a 2-step process, the first step is supervised learning for the sake of the predefined class label for training data set. Second step is classification accuracy evaluation. Likewise data prediction is also 2-step process. Before the classification and prediction, something should be done beforehand like data cleaning, relevance analysis, data transformation and reduction. The most interesting part, to me, is relevance analysis. Many of the attributes in the data may be redundant. Correlation analysis can be used to identify whether any two given attributes are statistically related. If there is strong correlation between attributes A1 and A2, one of them could be removed from further analysis. For a given data set, its set of attributes defines its data space representation. The quality of a data space representation is one of the most important factors influencing the performance of a data mining algorithm. The attributes defining the data space can be inadequate, making it difficult to discover high-quality knowledge.

However, when the original attributes are individually inadequate, it is often possible to combine them in order to construct new attributes with greater predictive power than the original attributes, facilitating the discovery of knowledge with a high predictive accuracy.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

II. RELATED WORK

Reproducing kernel Hilbert space framework for information theoretic learning was proposed by Xu et al. The framework uses the symmetric nonnegative definite kernel function i.e. cross information potential. Though this framework gives better result than the previous RKHS frameworks, still there is an issue to choose an appropriate kernel function for a particular domain.

Shilton and Palaniswami defined a unified approach to support vector machines. This unified approach is formulated for binary classification and later on extended to one - class classification and regression. Some of the techniques that have been proposed to speed up the training time are sequential minimal optimization, modified sequential minimal optimization, decomposition method and low rank kernel matrix construction method. The classification time of SVM primarily depends on the number of Support Vectors (SVs) involved in the system. So, it is necessary to minimize the number of support vectors that can improve the efficiency and minimize the computation time of the classification process.

Kumar et al. explored a binary classification framework for two stage multiple kernel learning. The distinct advantage of this binary classification framework is that it is easier to leverage research in binary classification and to develop scalable and robust kernel based algorithms. However, kernel methods are processed by operations to the kernel function (such as Gaussian and polynomial kernels) for the data, ignoring both the structure of the input data and the dimensionality problem, and thus cannot always guarantee that the transformed space is useful for classification. The commonly used kernels are the so-called all-function or general purpose ones, such as the Gaussian and polynomial.

Takeda et al. proposed a unified robust classification model that optimizes the existing classification models like SVM, Min-Max probability machine and fisher discriminant analysis. It provides several benefits like well - defined theoretical results extends the existing techniques and clarifies relationships among existing models. Basically, Support vector machines (SVM) are considered as a must try it offers one of the most robust and accurate methods among all well-known algorithms. It has a sound theoretical foundation, requires only a dozen examples for training, and is insensitive to the number of dimensions. In addition, efficient methods for training SVM are also being developed at a fast pace. In a two-class learning task, the aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data. The metric for the concept of the best classification function can be realized geometrically.

Raj Kumar, Dr. Rajesh Verma viewed the pattern classification as an ill - posed problem, it is a prerequisite to develop a unified theoretical framework that classifies and solves the ill posed problems. Recent literature on classification framework has reported better results for binary class datasets alone. For multiclass datasets, there is a lack in accuracy and robustness. So, developing an efficient classification framework for multiclass datasets is still an open research problem.

III. PROPOSED ALGORITHM

Dataset can be classified into two different types, such as numerical dataset, categorical dataset. Numerical dataset values or observations can be measured. Examples: Height, Arm Span and Weight. Scatter plots and line graphs are used to graph numerical data. Categorical Dataset values or observations can be sorted into groups or categories. Examples: Sex, Eye color and Favorite color. Bar charts and pie graphs are used to graph categorical data. The sample data set is small as it is having only five records in it. There are three attributes available in the data set and this data set contains two class labels as A and B. It contains both numerical and categorical attributes. This insufficient data will not lead to a robust classification performance. So this project constructs a new attributes for both numerical and category attribute to improve the performance of the classification.

Table 1: Sample Input Dataset

Patient Id	Disease	BP	Sugar	Class
1	Influenza	120	75	B
2	Malaria	110	80	A
3	Rabies	100	90	B
4	Fever	105	95	A
5	Headache	110	87	B

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

Algorithm
Input – Data Set
Output – New Data Set with more attributes
1. Read Data Set
2. Get the Numerical Attribute Set NA
3. Get the Categorical Attribute Set CA
4. **// Numerical Attribute Construction**
5. For each NA
6. Apply constructive operators, A + B, A*B, A- B, B - A, A/B, B/A, and A2
7. Construct the new numerical attributes
8. **// Categorical Attribute**
9. For each CA
10. Generate the class hierarchy
11. Use level-2 or level-3 value of class hierarchy
12. Construct the new categorical attributes
13. Combine all the attributes

Assuming that a sample set X = {x1,x2,... ,xN}, where each sample xi, i = 1,... ,N, in X has M attributes (means xi =(x1,... ,xM)).

For Numerical Attributes, Gomez and Morales proposed seven constructive operators: A+B, A*B, A-B, B-A, A/B, B/A, and A^2. Each pair of attributes, A and B, will be used to construct the new attributes, named synthetic attributes which are A+B, A*B, A-B, B-A, A/B, B/A, and A^2. Finally, the original attributes are merged, the attributes built up by class possibility, and the attributes created by attribute construction methods as the learning set, the new dataset with extending attributes A, B, A+B, A*B, A-B, B-A, A/B, B/A, and A^2 then form the classification model. For Categorical Attributes, First build class hierearchy of each attribute values. For example disease name is influenza, its level 2(Disease Type) is Virus, and level 3(Specialty) is Infectious disease. So extending new attributes with level2 and level3 attributes for categorical dataset.

Table 2: New Data Set with more attributes

Patient Id	Disease	BP	Sugar	Level 2	BP+Sugar	BP*Sugar	Class
1	Influenza	120	75	Virus	195	9000	B
2	Chickenpox	110	80	Virus	190	8800	A
3	Pneumonia	100	90	Bacteria	190	9000	B
4	Barcoo fever	105	95	Bacteria	200	9975	A
5	Mumps	110	87	Virus	197	9570	B

Now classify this mixed dataset. Classification is a challenging task in data mining technique. The aim of Classification is to group the similar data into number of classifiers. Various classification algorithms have been developed to group data into classifiers. However, these classification algorithms work effectively either on pure numeric data or on pure categorical data, most of them perform poorly on mixed categorical and numerical data types in previous algorithm was used but it is not accurate for large datasets. In this paper propose classify the mixed numeric and categorical data set in efficient manner.

In this work, presents a classification algorithm based on similarity weight and filter method paradigm that works well for data with mixed numeric and categorical features. It proposes a modified description of classifier center to overcome the numeric data only limitation and provide a better characterization of classifiers.

IV. MIXED DATASET CLASSIFICATION

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks. There are different classification

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

algorithms for grouping data to clusters. The algorithms works on different datasets i.e., either on pure Numeric datasets or pure Categorical datasets and few of the algorithms works both mixed numeric and categorical datasets. In this paper a technique is using for mixed datasets splitting it as different datasets and performing suitable algorithms for datasets to form clusters, and combining those split clusters as a categorical datasets and finally applying a suitable clustering algorithm to get the final clusters. Previously they done the same but they used clustering algorithms which doesn't handled outliers perfectly so here CURE and ROCK algorithms are used for Numeric and Categorical datasets.

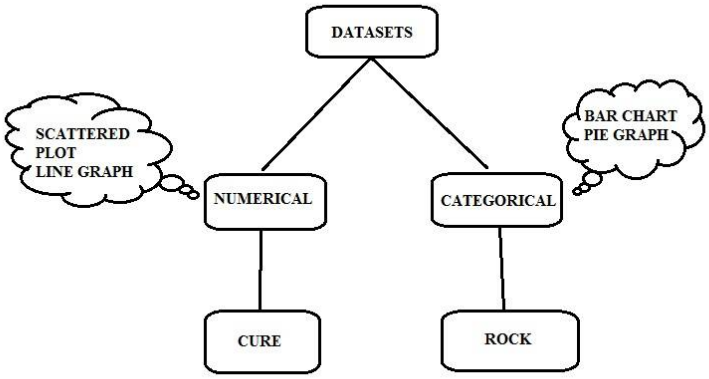

Figure 1: Flow of Work

Algorithm Flow:
Input: The Mixed Dataset DS
Output: Each Data Object Identified
1. Splitting the Dataset into Numeric Dataset (NDS) and Categorical Dataset (CDS).
2. Clustering the Categorical Dataset using Categorical Clustering Algorithm using ROCK.
3. Clustering the Numeric Dataset using Numeric Clustering Algorithm using CURE.
4. Combining the output clusters of CDS and NDS into a Categorical Dataset.
5. Clustering the Combined Categorical Dataset using ROCK Clustering Algorithm to form final classification results.

CURE (Clustering using Representatives):

CURE is also an Hierarchical clustering algorithm used for large datasets that adopts a Middle ground between centroid based mostly and all-points approach. Its belongs to Agglomerative i.e.., bottom-up approach. Here it represents a each cluster by a fixed number of points which are generated by selecting well scattered points from the cluster, then shrink the points towards the cluster centre by a specified faction. The main advantage of CURE is shrinking helps to intense the effects of outliers. Set a target i.e., representative point number c, for each of the clusters select c well scattered points attempting to capture the physical shape and geometry of the cluster. The chosen scattered representative points are then finally shrunk towards the centroid in a fraction of a where 0<= a<=1.

To avoid the problems with non-uniform sized or shaped clusters, CURE employs a hierarchical clustering algorithm that adopts a middle ground between the centroid based and all point extremes. In CURE, a constant number c of well scattered points of a cluster are chosen and they are shrunk towards the centroid of the cluster by a fraction α. The scattered points after shrinking are used as representatives of the cluster. The clusters with the closest pair of representatives are the clusters that are merged at each step of CURE's hierarchical clustering algorithm. This enables CURE to correctly identify the clusters and makes it less sensitive to outliers. Running time is O (n2 log n), making it rather expensive, and space complexity is O (n). The algorithm cannot be directly applied to large databases because of the high runtime complexity. Enhancements address this requirement.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

Random Sampling:

Random sampling supports large data sets. Generally the random sample fits in main memory. The random sampling involves a tradeoff between accuracy and efficiency.

Partitioning:

The basic idea is to partition the sample space into p partitions. Each partition contains n/p elements. The first pass partially clusters each partition until the final number of clusters reduces to n/pq for some constant $q \geq 1$. A second clustering pass on n/q partially clusters partitions. For the second pass only the representative points are stored since the merge procedure only requires representative points of previous clusters before computing the representative points for the merged cluster. Partitioning the input reduces the execution times.

Labeling data on disk:

Given only representative points for k clusters, the remaining data points are also assigned to the clusters. For this a fraction of randomly selected representative points for each of the k clusters is chosen and data point is assigned to the cluster containing the representative point closest to it.

CURE ALGORITHM:

Input: A set of points S
Output: k clusters
Step 1: For every cluster u (each input point), in u.mean and u.rep store the mean of the points in the cluster and a set of c representative points of the cluster (initially c = 1 since each cluster has one data point). Also u.closest stores the cluster closest to u.
Step 2: All the input points are inserted into a k-d tree T
Step 3: Treat each input point as separate cluster, compute u.closest for each u and then insert each cluster into the heap Q. (clusters are arranged in increasing order of distances between u and u.closest).
Step 4: While size (Q) > k
Step 5: Remove the top element of Q (say u) and merge it with its closest cluster u.closest(say v) and compute the new representative points for the merged cluster w.
Step 6: Remove u and v from T and Q.
Step 7: For all the clusters x in Q, update x.closest and relocate x
Step 8: insert w into Q
Step 9: repeat

ROCK (Robust Clustering using linKs):

ROCK is also one of the Hierarchical clustering that deals with concepts of links i.e., the number of common neighbors between two objects for data with non-numeric which is categorical attributes. The formal clustering algorithms for clustering data with Boolean and categorical values use distance functions such as Euclidean and Manhattan distance. However these distance functions doesn't leads to high quality clusters when clustering categorical data. Most of the clustering algorithms follow the same method that is when clustering the similarity between the two points are merged into a single cluster. This one leads to errors where two distinct clusters have a few points or outliers that are close, therefore following the similarity between two points to make a clustering decision would leads that two clusters to be merged. ROCK takes a global approach which is considering the neighborhoods of individual pairs of points.

If two similar points have the same neighborhoods then the two points likely belong to same one and form a cluster and thereby merged. More formally, the two points mi and mj are neighbors i.e., if sim(mi, mj)>= $\hat{I}$, where sim defines the similarity function and $\hat{I}$, considered as the threshold value which is specified by user. The number of links between mi and mj is defined as the number of common neighbors between them. If the links between those points is large then it can said that it is more likely belong to same cluster. By considering neighboring data points in the relationship between the pair of points ROCK is more robust than the standard clustering methods which focuses only on point similarity, A good example of data containing categorical attributes is market basket data here a transaction represents one customer, and each transactions contain set of items purchased by the customer. Use to cluster the customers where those customers with similar buying pattern are in a one cluster. Use for Characterizing different customer groups based on similar buying patterns.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

Targeted Marketing:

Predict buying patterns of new customers based on profile i.e, if he buys bread then he must buy cheese. In market database, where attributes of data points are non-numeric, transactions viewed as records with Boolean attributes corresponding to single item i.e., TRUE if transaction contain item, FALSE if doesn't contain item. Boolean attributes are special case of categorical Attributes. The database consists of set of transactions, each of the transactions contains set of items. According to Jaccard coefficient [60] for sim(Ti,Tj), the similarity between the two transactions Ti and Tj is computed.

A link is defined as the number of common neighbors between two records. Intuitively, it is the number of distinct paths of length 2 that exist between a pair of points. Note that the point is considered as a neighbor itself, there is a link from each neighbor of the "root" point back to itself through the root. Therefore if a point has x neighbors the x^2 links are due to it. Due to the links advantage is it captures the neighborhood related information of the data i.e., a step towards more global and robust solutions. ROCK follows two steps those are Random Sampling, Clustering with links.

Random Sampling:

1) Usually it is a large number of data
2) Enables ROCK to reduce the number of points considered to reduce complexity
3) Clusters generated by the sample points
4) With appropriate sample size, the quality of clustering is not affected

Clustering with links:

It determines the best pairs of clusters to merge at each step of ROCK's hierarchical clustering algorithm. For a pair of clusters Ci and Cj, link [Ci, Cj] stores the number of cross links between the clusters Ci and Cj i.e.,

Algorithm for ROCK Clustering:

Procedure cluster(S,k)//Set on n Sample points and k number of clusters
begin
Step 1: link:= Compute_links(S)
Step 2: for each s E S do
Step 3: q[s]:= build_loacl_heap(link,s)
Step 4: Q=build_global_heap(S,q)
Step 5: while size(Q)>k do{
Step 6: u:= extract_max(Q)
Step 7: v:= max(q[u])
Step 8: delete(Q,v)
Step 9: w:= merge(u,v)
Step 10: for each x E q(u) U a{v} do{
Step 11: link[x,w]:= link[x,u]+link[x,v]
Step 12: delete (q[x].u): delete(q[x], v)
Step 13: insert (q[x],w,g(x,w)); insert(q[w],x,g(x,w)
Step 14: update (Q,x,q[x])
Step 15: }
Step 16: insert (Q, w, q[w])
Step 17: deallocate(q[u]);deallocate (q[v])
Step 18: }
End

This section lists three main Experiments with neat flow of work. First is to apply Constructing a new Attributes for small dataset. The experiment constructs a new attributes for both numerical and categorical attributes for improve the performance of the classification. The Second experiment is apply the CURE clustering algorithm on Numerical Dataset and apply the ROCK clustering algorithm on Categorical Dataset. The third experiment is combining both clustering results as a categorical dataset then applies the ROCK algorithm for final classification. Combined CURE and ROCK algorithms perform well in given mixed data sets and produce expected results.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

V. CONCLUSION AND FUTURE WORK

This paper proposed a novel technique to solve the attribute construction problem with the help of clustering algorithms. First, the original mixed dataset is divided into two sub datasets: the pure categorical dataset and the pure numeric dataset. Next, apply CURE clustering algorithm on Numerical Dataset and apply ROCK clustering algorithm on Categorical Dataset. Finally, the clustering results on the categorical and numeric dataset are combined as a categorical dataset, on which the designed for one type of features to handle numeric and nonnumeric feature values. Our main contribution on this thesis is to provide an algorithm framework for the mixed attributes classification problem, on which existing clustering algorithms can be easily integrated, the capabilities of different kinds of clustering algorithms and characteristics of different types of datasets could be fully exploited. The future work will investigate integrating other alternative classification algorithms into the algorithm framework, to get further insight into this methodology.

REFERENCES

[1] Qasem A. Al-Radaideh, Eman Al Nagi, "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 2, 2012.
[2] Mohd. Mahmood Ali, Mohd. S. Qaseem, Lakshmi Rajamani, A. Govardhan, "Extracting Useful Rules through Improved Decision Tree Induction Using Information Entropy", International Journal of Information Sciences and Techniques (IJIST) Vol.3, No.1, January 2013.
[3] Andreas G.K. Janecek, Wilfried N. Gansterer, "On the Relationship between Feature Selection and Classification Accuracy", JMLR: Workshop and Conference Proceedings 4: 90-105, 2008.
[4] P.Niyogi, F.Girosi, and P.Tomaso, "Incorporating Prior Information in Machine Learning by Creating Virtual Examples," Proc. IEEE, vol. 86, no. 11, pp. 2196-2209, Nov. 1998.
[5] Limère A, Laveren E, and Van Hoof, K. "A classification model for firm growth on the basis of ambitions, external potential and resources by means of decision tree induction", Working Papers 2004 027, University of Antwerp, Faculty of Applied Economics.
[6] Hoi, S. C., Lyu, M. R., and Chang, E. Y. (2006). "Learning the unified kernel machines for classification, In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining", pp. 187-196.
[7] Xu, J. W., Paiva, A. R., Park, I., and Principe, J. C. (2008). A reproducing kernel Hilbert space framework for information-theoretic learning, IEEE Transactions on Signal Processing, Volume 56, Issue 12, pp.5891-5902.
[8] Shilton, A., and Palaniswami, M. (2008). "A Unified Approach to Support Vector Machines", In B. Verma, & M. Blumenstein (Eds.), Pattern Recognition Technologies and Applications: Recent Advances, pp. 299-324.
[9] Kumar, A., Niculescu-Mizil, A., Kavukcuoglu, K., and Daume III, H. (2012). "A binary classification framework for two-stage multiple kernel learning". arXiv preprint arXiv:1206.6428, Appears in Proceedings of the 29th International Conference on Machine Learning.
[10] Takeda, A., Mitsugi, H., and Kanamori, T. (2012). "A unified robust classification model",arXiv preprint arXiv:1206.4599. J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publish, 2001
[11] C. Kim and C.H. Choi, "A Discriminant Analysis Using Composite Features for Classification Problems," Pattern Recognition, vol. 40, no. 11, pp. 2958-2966, 2007
[12] Raj Kumar, Dr. Rajesh Verma, "Classification Algorithms for Data Mining: A Survey", International Journal of Innovations in Engineering and Technology (IJIET), Vol. 1 Issue 2 August 2012, ISSN: 2319 – 1058, pg: 7-14
[13] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.
[14] Data Mining Concepts and Techniques, Third Edition ISBN: 978-0-12-381479-1, Morgan Kaufmann Publishers, 225Wyman Street, MA 02451 (USA), 2012.
[15] Muhammad Husnain Zafar and Muhammad Ilyas, "A Clustering Based Study of Classification Algorithms", International Journal of Database Theory and Application Vol.8, No.1, pp.11-22, 2015.
[16] Yogita Rani, Manju & Harish Rohil, "Comparative Analysis of BIRCH and CURE Hierarchical Clustering Algorithm using WEKA 3.6.9", The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 2, No. 1, January-February 2014.
[17] Mierswa, I, "Evolutionary learning with kernels: a generic solution for large margin problems", In Proceedings of the 8th annual conference on Genetic and evolutionary computation, ACM, New York,pp. 1553-1560, 2006.
[18] Sivaramakrishnan K.R, Karthik K. and Bhattacharyya, "Kernels for Large Margin Time-Series Classification, International Joint Conference on Neural Networks, pp. 2746-2751, 2007.
[19] Hofmann T, Schölkopf B, and Smola A.J, "Kernel Methods in Machine Learning, the Annals of Statistics", Volume 36, pp. 1171-1220, 2008.
[20] Kuo-Ping Wu and Sheng-De Wang, "Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space, Pattern Recognition", Volume 42, Issue 5,pp. 710-717, ISSN 0031-3203, 2009.
[21] Y.Muto and Y.Hamamoto, "Improvement of the Parze n Classifier in Small Training Sample Size Situations," Intelligent Data Analysis, vol. 5, no. 6, pp. 477-490, 2001.
[22] D.C. Li and C.W. Liu, "A Neural Network Weight Determination Model Designed Uniquely for Small Data Set Learning," Expert Systems with Applications, vol. 36, pp. 9853-9858, 2008.
[23] Seema Maitrey, C. K. Jha, Rajat Gupta, Jaiveer Singh, "Enhancement of CURE Clustering Technique in Data Mining", National Conference on Development of Reliable Information Systems, Techniques and Related Issues (DRISTI), Proceedings published in International Journal of Computer Applications® (IJCA), 2012.