THOMSON REUTERS
ENDNOTE

[Sagunthaladevi* *et al.*, 6(2): February, 2017]
IC™ Value: 3.00

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

NEW ATTRIBUTE CONSTRUCTION IN MIXED DATASETS USING CLASSIFICATION ALGORITHMS

Sagunthaladevi.S* & Dr.Bhupathi Raju Venkata Rama Raju
Research Scholar, Dept of Computer Science, Mahatma Gandhi University, Meghalaya-793101, India
Professor, Dept of Computer Science, IEFT College of Engineering, Villupuram-605108, Tamilnadu, India

DOI: 10.5281/zenodo.291847

ABSTRACT

Classification is a challenging task in data mining technique. The main aim of Classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then group the similar data into number of classifiers and it assigns items in a collection to target categories or classes. Finally classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes. Various classification algorithms have been developed to group data into classifiers. However, those classification algorithms works effectively either on pure numeric data or on pure categorical data and most of them performs poorly on mixed categorical and numerical data types. Previous classification algorithms do not handled outliers perfectly. To overcome those disadvantages this paper represents NCA and CCA algorithms for Numerical and Categorical datasets to improve the performance of classification. Results of these proposed algorithms are compared with existing ones based on parameters such as accuracy, precision and F-Measures.

KEYWORDS: Classification, Prediction, Mixed Dataset Classification Algorithm (MDCA), Numerical Classifying Algorithm (NCA) and Categorical Classifying Algorithm (CCA).

INTRODUCTION

Classification 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. A classification task begins with a data set in which the class assignments are known. For example, a classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time. In addition to the historical credit rating, the data might track employment history, home ownership or rental, years of residence, number and type of investments, and so on. Credit rating would be the target, the other attributes would be the predictors, and the data for each customer would constitute a case. Classifications are discrete and do not imply order. Continuous floating-point values would indicate a numerical, rather than a categorical target. A predictive model with a numerical target uses a regression algorithm, not a classification algorithm. The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, high credit rating or low credit rating. Multiclass targets have more than two values: for example, low, medium, high, or unknown credit rating. In the model build (training) process, a classification algorithm finds relationships between the values of the predictors and the values of the target. Different classification algorithms use different techniques for finding relationships. These relationships are summarized in a model, which can then be applied to a different data set in which the class assignments are unknown. This paper presents a classification algorithm based on similarity weight and filter method paradigm that works well for data with mixed numeric and categorical features. It proposed a modified description of classifier center to overcome the numeric data only limitation and provide a better characterization of classifiers.

This paper is organized in five sections. Section I give an introduction about the topic. Architecture and Steps for proposed algorithm are discussed in section II. Section III discusses the results of the proposed method and Section IV gives Conclusion.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

METHODOLOGY

New Attributes Construction and their performance evaluation on the basis of attribute construction Time, and numbers of newly constructed attributes are represented in this paper. Research work focuses specifically Mixed Dataset Classification using MDCA Algorithm for both numerical and categorical data. Here, separate algorithms for both numerical (NCA) and categorical (CCA) are introduced for effective outcomes.

MDCA ARCHITECTURE:

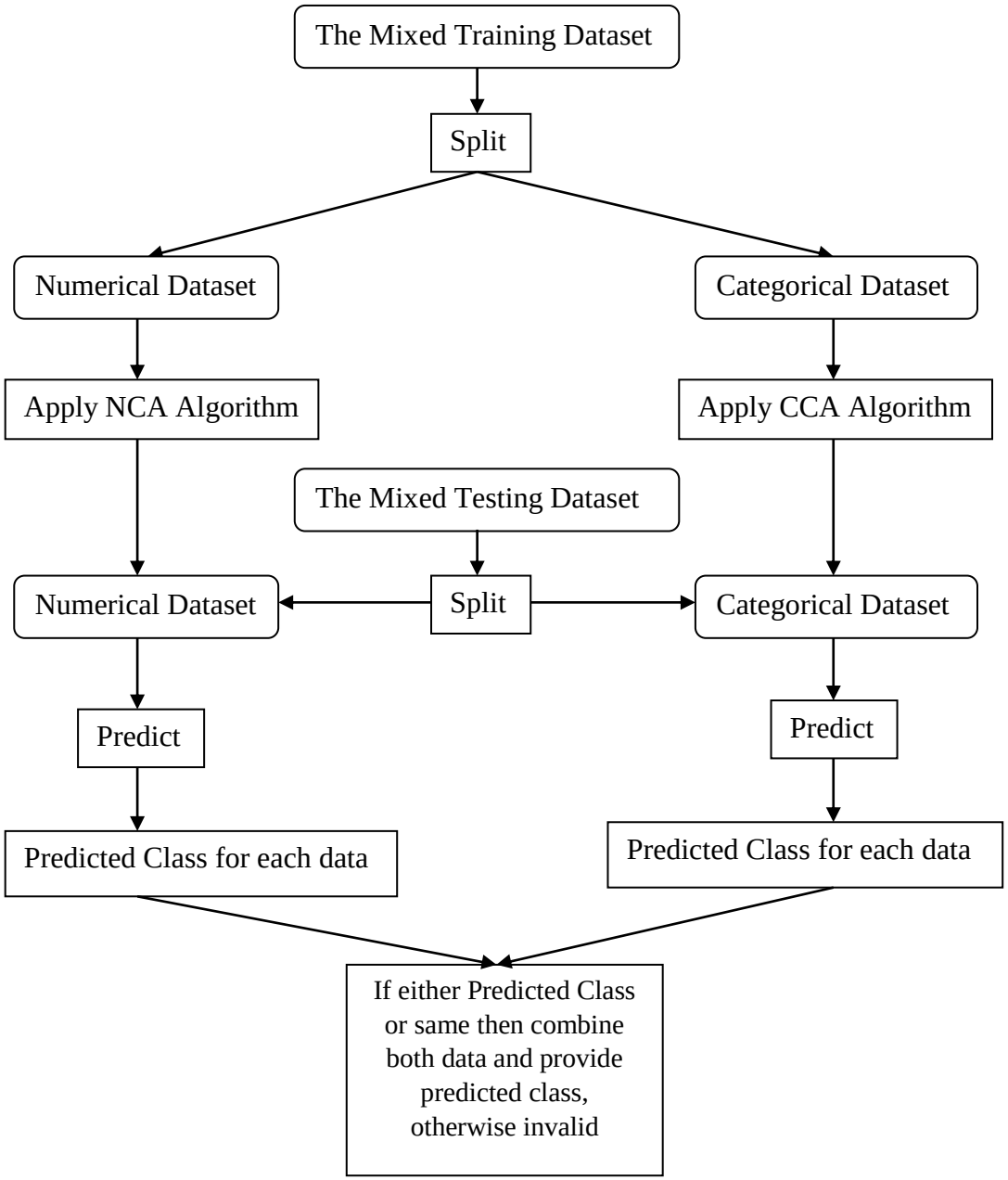

Figure 1: Mixed Dataset Classification Algorithm (MDCA) Architecture

MIXED DATASET CLASSIFICATION ALGORITHM (MDCA):

Algorithm for MDCA is presented below with two different set of algorithms designed for numerical and categorical datasets NCA and CCA respectively.

Input: The Mixed Training Dataset and the Mixed Testing Dataset

Output: Predicted Class for each Mixed Testing Data

1. Splitting the Mixed Training Dataset into Numerical Dataset (NDS) and Categorical Dataset (CDS).

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

2. Classifying the Categorical Dataset using Categorical Classifying Algorithm (CCA).
3. Classifying the Numeric Dataset using Numerical Classifying Algorithm (NCA).
4. Browse the Mixed Testing Dataset.
5. Splitting this testing Dataset into Numerical and Categorical Dataset.
6. Predict the numerical results based on NCA Algorithm.
7. Predict the categorical results based on CCA Algorithm.
8. If both numerical and categorical predicted results are same, combining the numerical and categorical data then provide predicted results. Remains are outliers.

NCA (NUMERICAL CLASSIFYING ALGORITHM):
It is a part of MDCA algorithm used for numerical datasets.
Input: The Numerical Dataset with class attributes (0 or 1)
Output: Classification
Step 1: First set the Rule Set array list is empty.
Step 2: for (each class) {
Step 3: Extract rule based on Dataset, all attributes sets and their values with class value.
Step 4: Add this extracted rule to Rule Set array list.
Step 5: }
Step 6: This extracted Rule Set is used for predict the Testing dataset class values.

CCA (CATEGORICAL CLASSIFYING ALGORITHM):
It is a part of MDCA algorithm used for categorical datasets.
Input: The Categorical Dataset with class attributes (A or B)
Output: Classification
Step 1: First classification attribute (for root node) is selected in the categorical dataset.
Step 2: Followed by compute entropy.
Step 3: For each attribute in dataset, compute Information Gain using this classification attribute.
Step 4: Then select highest gain attribute to be the next node in the tree starting from the root node.
Step 5: Followed by, remove the node attributes, creating reduced dataset.
Step 6: Repeat steps 3 to 5 until all attributes used or same classification value remains for all rows in reduced dataset.

MIXED DATASETS:
Mixed Attribute types dataset contains both numerical and categorical types of attributes. As mixed attribute type datasets are common in real life, clustering and classification techniques for mixed attribute type datasets is required in various informatics fields such as bio informatics, medical informatics, geo informatics, information retrieval, to name a few. These mixed attribute datasets provide challenges in clustering and classification because there exist many attributes in both categorical and numerical forms so mixed attribute type should be considered together for more accurate and meaningful classification.

Table 1: Dataset Characteristics

S.No	Characteristics	Chess	Servo	Teaching Assistant
1	Data Set Characteristics	Multivariate	Multivariate	Multivariate
2	Number of Instances	28056	167	151
3	Area	Game	Computer	N/A
4	Attribute Characteristics	Categorical, Integer	Categorical, Integer	Categorical, Integer
5	Number of Attributes	6	4	5
6	Associated Tasks	Classification	Regression	Classification
7	Missing Values	No	No	No

It is a small dataset, so three mixed small data sets are explained in this section and mentioned in above table namely Chess (King-Rook vs. King) Data Set, Servo Data Set and Teaching Assistant Evaluation Data Set. These three data sets are downloaded from the UCI Machine Learning Repository database for analysis.

MIXED DATASET CLASSIFICATION:
To Evaluate the Performance of mixed dataset classification, three parameters were considered such as
1. Predicted Results

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

2. Classification Time
3. Prediction Time

PREDICTED RESULTS:

For prediction, Mixed Testing Dataset is loaded and it contains Query attributes with class values. First split the testing dataset into numerical and categorical. Then apply NCA & CCA Algorithm and get the predicted class.

CLASSIFICATION TIME:

Before classification, note the current time in Milliseconds (Starting Time). Then execute NCA & CCA Algorithms. After classification, note the current time (Ending Time) once again. For classification time, subtract both Starting and Ending time.

Classification Time = Ending Time – Starting Time

PREDICTION TIME:

Before prediction, note the current time in Milliseconds (Starting Time). Then execute NCA & CCA Algorithms. After prediction, note the current time (Ending Time) once again. For prediction time, subtract both Starting and Ending time.

Prediction Time = Ending Time – Starting Time

NEW ATTRIBUTES CONSTRUCTION:

To Evaluate the Performance of the New Attributes Construction, two parameters were considered, those two parameters are,

1. New Attributes Construction Time
2. Number of new Attributes Constructed.

NEW ATTRIBUTES CONSTRUCTION TIME:

Before new attributes construction, note the current time in Milliseconds (Starting Time). Then construct the new attributes. After new attributes construction, note the current time (Ending Time) once again. For new attributes construction time, subtract both time.

New Attributes Construction Time = Ending Time – Starting Time

NUMBER OF NEW ATTRIBUTES CONSTRUCTED:

Before new attributes construction, note the no of attributes exists (Available attributes before construction). Then construct the new attributes. After new attributes construction, note the no of attributes exists (Available attributes after construction) once again. For newly constructed attributes, subtract both results.

Number of New Attributes Constructed = Available attributes after construction – Available attributes before construction

RESULTS AND DISCUSSION

This chapter documents the results of constructing a new attributes with classification from mixed dataset. This chapter is ordered with two main Experiments. First is to apply Constructing a new Attributes for small dataset. The experiment constructs a new attributes for both numerical and categorical datasets for improve the performance of the classification. In this, two algorithms are elaborated for both kind of datasets i.e., Numerical Classifying Algorithm (NCA) and Categorical Classifying Algorithm (CCA) to extract accurate classification results. The Second experiment is apply the mixed dataset classification algorithm (MDCA) for predict the class value for Mixed Testing Dataset. Prediction contains set of rules for calculating the attributes and those are described in upcoming sessions.

DATASETS CLASSIFICATION RESULTS BASED ON NCA ALGORITHM

Mixed small datasets such as TAE, Servo & Chess datasets characteristics are listed in the below table. Results represent the values in the form of confusion matrix, precision, recall, accuracy and F-measure. Matrix values includes true positive, true negative, false positive and false negative respectively.

Table 2: Various Dataset Classification Results of NCA Algorithm

Dataset	True +ve	True –ve	False +ve	False -ve	Precision	Recall	Accuracy	F-Measure
TAE Dataset	0.53	0.37	0.8	0.2	0.86	0.96	0.90	0.91
Servo Dataset	0.10	0.86	0	0.4	1.0	0.71	0.96	0.83

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Chess Data Set	0.53	0.37	0.9	0.1	0.85	0.98	0.90	0.91

Above table states the results obtained by using CCA algorithm. Three sample datasets are taken from the repository for classification analysis. Major parameters are taken for classification analysis with respect to kinds of datasets taken.

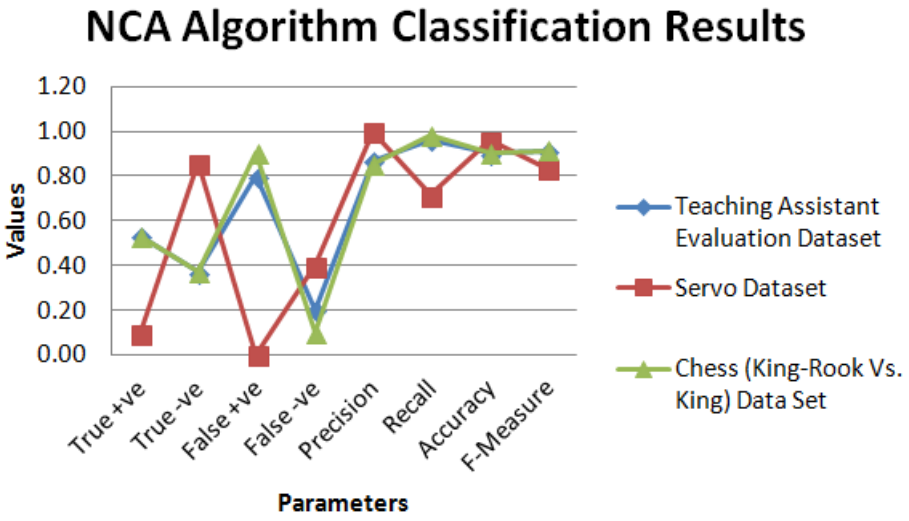

Chart 1: Various Dataset Classification Results of NCA Algorithm

Chart 1 show the NCA algorithm classification results as Teaching Assistant Evaluation dataset and Chess Dataset has high true positive and high F-Measure values. Chess (King-Rook Vs. King) Data Set has high false positive and high recall value and Servo Dataset has high true negative, high false negative, high precision value and high accuracy.

DATASETS CLASSIFICATION BASED ON CCA ALGORITHM:

Mixed small datasets such as TAE, Servo & Chess datasets characteristics are listed in the below table. Results represent the values in the form of confusion matrix, precision, recall, accuracy and F-measure. Matrix values includes true positive, true negative, false positive and false negative respectively.

Table 3: Various Dataset Classification Results of CCA Algorithm

Dataset	True +ve	True –ve	False +ve	False -ve	Precision	Recall	Accuracy	F-Measure
TAE Dataset	0.53	0.38	0.7	0.2	0.88	0.96	0.91	0.92
Servo Dataset	0.11	0.85	0.1	0.3	0.91	0.78	0.96	0.84
Chess Data Set	0.51	0.41	0.5	0.3	0.91	0.94	0.92	0.92

Above table states the results obtained by using CCA algorithm. Three sample datasets are taken from the repository for classification analysis. Major parameters are taken for classification analysis with respect to kinds of datasets taken.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

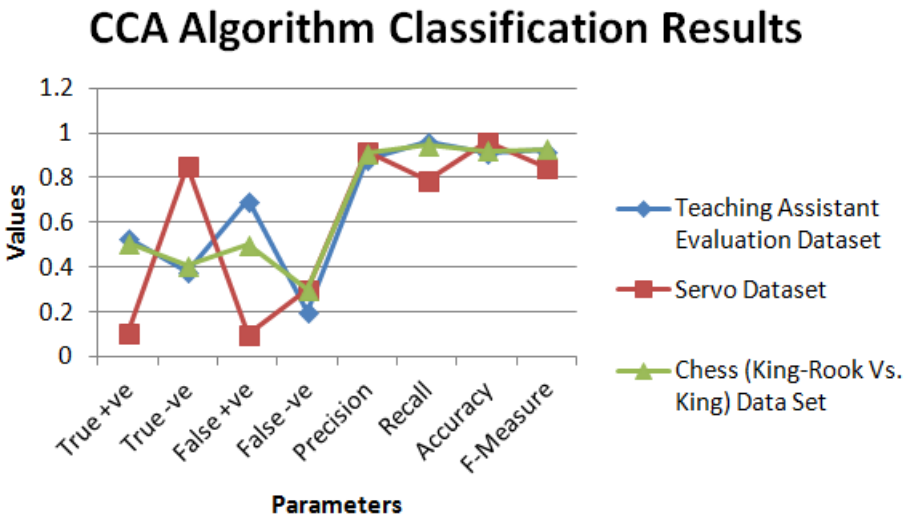

Chart 2: Various Dataset Classification Results of CCA Algorithm

CCA Algorithm Classification results shows that Teaching Assistant Evaluation dataset has high true positive, high false positive and high recall value. Servo Dataset has high true negative and high precision & high accuracy value. Chess (King-Rook vs. King) Data Set has high F-Measure value. Servo & Chess (King-Rook vs. King) Data Set has high false negative value.

ACCURACY USING TEACHING ASSISTANT EVALUATION DATASET

Lists Accuracy using Teaching Assistant Evaluation dataset are tabulated below and its ratios are displayed in chart 3.

S.No	Algorithm	Accuracy
1	J48	41.0596%
2	Decision Table	35.0093%
3	Logistic	41.0596%
4	Multilayer Perceptron	38.4106%
5	Naive Bayes	41.0596%
6	Random Forest	41.0596%
7	VFI	41.0596%
8	ZeroR	34.4371%
9	Genetic Programming	37.7483%
10	NCA Algorithm	42.0596%
11	CCA Algorithm	43.0093%

Table 4: Accuracy Comparison with Existing Algorithms

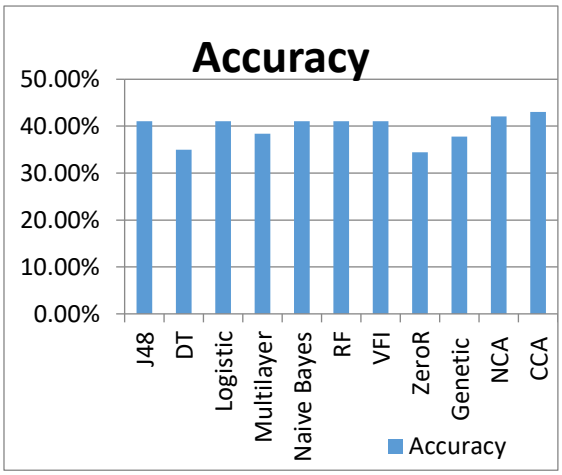

Chart 3: Accuracy comparisons with existing algorithms

Above table lists the accuracies obtained in Teaching Assistant Evaluation dataset. Comparison of existing algorithms with NCA and CCA using TAE dataset shows that proposed algorithms gives more accuracy than existing algorithms. Results state MDCA algorithms produce more accuracy than previous algorithms. Chart 3 displays the results in percentage which are represented in the table 4. It shows few algorithms produce nearly 41 percentages, but proposed algorithms (NCA & CCA) gives 41 and above results in better accuracy classification. Totally the experimental findings shows that the proposed method has better classification accuracy compare to existing methods.

CONCLUSION

Classifying from small mixed dataset is fundamentally difficult, because it having only few attributes in it and also contains both numerical and categorical attributes. This insufficient data will not lead to a robust classification performance. The existing classification algorithms works on different datasets i.e., either on pure Numeric datasets or pure Categorical datasets. But these algorithms are not considered mixed dataset classification. So this work is proposed to construct a new attributes for both numerical and categorical attributes to improve the performance of the classification. To overcome the disadvantages of existing algorithms Mixed Dataset Classification Algorithm (MDCA) is introduced. It split the Mixed Dataset into Numerical and Categorical dataset and then it applies Numerical Classifying Algorithm (NCA) for Numerical dataset and Categorical Classifying Algorithm (CCA) for Categorical dataset. Our Outcome results shows combination of NCA Algorithm with CCA Algorithm produce a better classification algorithm towards precision, recall, and accuracy. Experimental Results point out Compared to other algorithms Numerical Dataset and Categorical Dataset takes minimum time for new attributes construction.

REFERENCES

1. 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
2. Data Mining Concepts and Techniques, Third Edition ISBN: 978-0-12-381479-1, Morgan Kaufmann Publishers, 225Wyman Street, MA 02451 (USA), 2012.
3. Mrs.Sagunthaladevi.S and Dr.Bhupathi Raju Venkata Rama Raju, "Classification Techniques in Data Mining: An Overview", Global Journal of Engineering Science and Researches, Volume *3*, Issue 7, July 2016
4. Mrs.Sagunthaladevi.S and Dr.Bhupathi Raju Venkata Rama Raju, "Novel Method to Construct a New Attributes for Classification with Mixed Dataset Using Cure and Rock Algorithms", International Journal of Innovative Research in Computer and Communication Engineering, Volume 4, Issue 9, September 2016
5. Mrs.Sagunthaladevi.S and Dr.Bhupathi Raju Venkata Rama Raju, "Performance Analysis Of Cure And Rock Algorithms On Constructing A New Attribute With Mixed Datasets", International Journal of Innovative Research in Science, Engineering and Technology, Volume 6, Issue 1, January 2017.
6. 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.
7. 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.
8. 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.
9. Sivaramakrishnan K.R, Karthik K. and Bhattacharyya, "Kernels for Large Margin Time-Series Classification, International Joint Conference on Neural Networks, pp. 2746-2751, 2007.
10. Hofmann T, Schölkopf B, and Smola A.J, "Kernel Methods in Machine Learning, the Annals of Statistics", Volume 36, pp. 1171-1220, 2008.
11. 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.
12. 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.
13. 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.
14. K. Saravanan and S. Sasithra, "Review on classification based on artificial neural networks" International Journal of Ambient Systems and Applications (IJASA) Vol.2, No.4, December 2014.
15. 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.
16. 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.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

17. 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.
18. 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.
19. 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.
20. Hoi, S. C., Lyu, M. R, 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.
21. 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
22. 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.
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.
24. 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
25. 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.