Global Journal of Mathematical Sciences: Theory and Practical.
ISSN 0974-3200 Volume 9, Number 1 (2017), pp. 51-64

http://www.irphouse.com

Improving the Performance of a Classification Based Outlier Detection System using KNN-DT Hybrid Algorithm

Kurian M.J [1] and Gladston Raj S [2]

[1] Research Scholar, Research and Development Centre, Bharathiar University, Coimbatore, Tamil Nadu, India.

[2]Head of Department of CS, Govt. College, Nedumangadu, Trivandrum, Kerala, India.

Abstract

In Data mining, outlier detection can be treated as a classification problem with the availability of training data set with class labels. It is possible to apply a classification based outlier detection method if the samples of cancer data set available with class information. The general idea of classification-based outlier detection method is to train a classification model that can distinguish normal data from outliers [7]. The previous work shows that some of the classification based outlier detection algorithms provide better sensitivity and some others provide better specificity. By combining the better part of these classification algorithms, a hybrid classification algorithm can design. This work proposed a KNN-DT hybrid classification algorithm and evaluated the performance of outlier detection. The results clearly show that the impact of such hybridizing significantly improved the overall classification performance to a considerable level.

Keywords: Outlier detection, Hybrid classification, Data mining, Decision table, C4.5, KNN KNN-DT, KNN-C4.5.

I. INTRODUCTION

Outliers are observations deviated from the rest, which represent the unique characteristics of the objects and are very important in applications such as outlier detection in cancer datasets. Due to the increase of the dimensionality, the distance

between objects may be heavily dominated by noise and may not reflect the exact relationship between them. In order to improve the performance of the classification, this work proposes a hybrid classification approach using knn and decision tree algorithms.

II. MODELING HYBRID CLASSIFICATION BASED OUTLIER DETECTION SYSTEM

A. Outlier Detection Methods

A. Supervised, Semi-Supervised, and Unsupervised Methods

In supervised mode, training dataset is available for normal and outlier classes. Even though this approach builds a predicative model, it is difficult to obtain the accurate class labels.

The semi-supervised mode is widely used than supervised because the training dataset is available only for normal class. Outliers are the data instances which do not satisfy this class.

In unsupervised mode is widely used due to unavailability of training dataset. Outliers are the data instances which are not frequent or closely related to each other.

B. Statistical Methods, Proximity-Based Methods

With the assumption of the normalization of data, the Statistical methods are treated as the classical one for the detection of outlier. Proximity-based approaches assume that the proximity of an outlier object to its nearest neighbors significantly deviates from the proximity of the object to most of the other objects in the data set [7].

C. Clustering and Classification Based Methods

Outlier detection methods in clustering are based on the examination of the relation between clusters and objects [7]. But in classification based outlier detection, develop a model which distinguishes normal from outliers.

D. The Model of the Precision and Recall Based Hybrid Outlier Detection System

The main idea of this hybrid classification model is as follows: Some classification algorithms are capable of identifying benign data in a better manner and some algorithms are capable of identifying malignant data (or outlier) in a better manner. So to achieve the high classification accuracy, we propose to combine these two characteristics of two different classification algorithms. For example, if KNN is capable of identifying benign records and DT is capable of identifying the malignant records in a better manner, then the resultant class label will be much accurate than the above one.

The Model of the proposed Hybrid classification

- Classify the data using algorithm 1 and find the classification labels. Let A1 be the set of labels provided by algorithm which is capable of identifying benign records with greater accuracy
- Classify the data using algorithm 2 and find the classification labels . Let A2 be the set of class labels Provided by algorithm which is capable of identifying malignant records with greater accuracy
- A1= { AB1 , AM1 } where AB1 are the indexes of Benign records and AM1 are the indexes of the Malignant records provided by algorithm 1
- A2= { AB2 , AM2 } where AB2 are the indexes of Benign records and AM2 are the indexes of the Malignant records provided by algorithm 2
- Combine A1 and A2 in such a way to produce A3 = {AB1, AM2}, which will has higher accuracy than both A1 and A2.

The following Diagram shows the outline of the precision and recall based hybrid outlier detection system that we are going to construct and test in this work.

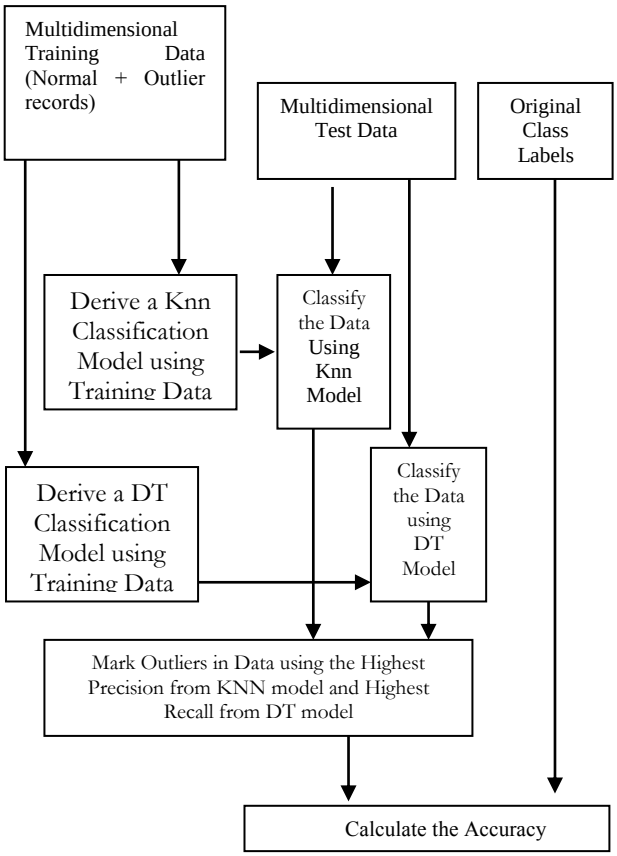

Figure 1: The Precision & Recall Based Hybrid Outlier Detection System

F. The Used Classification Algorithms

a) Decision Table Classifier

A decision table is a predictive modeling tool that performs the hierarchical breakdown of the data, with two attributes at each level of the hierarchy.

b) K-Nearest Neighbors Classifier

The working of KNN is as follows: *Identify the K nearest neighbors to an input instance in the population space and assign the instance to the class the majority of these neighbors belong to*. The "nearest" measurement refers to the Euclidean distance between two instances and calculated with the Euclidean distance between t_i and t_j is

$$D(t_i, t_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

Where, n is the number of attributes in each data instance.

III. THE EVALUATION

The classification algorithm's performance were tested with "Wisconsin Breast Cancer Database "

Breast cancer dataset

Breast cancer dataset (Wisconsin Breast Cancer Database) obtained from the UCI online machine-learning repository at http://www.ics.uci.edu/~mlearn /MLRepository.html

The Wisconsin breast cancer database (WBCD): The WBCD dataset is summarized in Table 1 and consists of 699 instances taken from fine needle aspirates (FNA) of human breast tissue. Each instance consists of nine measurements (without considering the sample's code number), namely clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. The measurements are assigned an integer value between 1 and 10, with 1 being the closest to benign and 10 the most anaplastic. Associated with each sample is its class label, which is either benign or malignant. This dataset contains 16 instances with missing attributes' values. Since many classification algorithms have discarded these data samples, for the ease of comparison, the same way is followed and the remaining 683 samples are taken for use. Therefore, the class is distributed with 444 (65.0%) benign samples and 239 (35.0%) malignant samples (Tan et al 2003).

Table1. Summary of the WBCD dataset

Attribute	Possible values
Clump thickness	Integer 1–10
Uniformity of cell size	Integer 1–10
Uniformity of cell shape	Integer 1–10
Marginal adhesion	Integer 1–10
Single epithelial cell size	Integer 1–10
Bare nuclei	Integer 1–10
Bland chromatin	Integer 1–10
Normal nucleoli	Integer 1–10
Mitoses	Integer 1–10
Class	Benign (65.5%), Malignant (34.5%)

The Metrics and Validation Method Used for Performance Evaluation

The Performance of the selected algorithms are depend on data's characteristics and it is measured with metrics Sensitivity, Specificity, Accuracy, Precision, F_Score, and Error Rate.

A) Confusion Matrix

The type of classification errors a classifier makes can recorded using a *confusion matrix*.

Predicted Class		
Positives	Negatives	Actual Class
w	x	Positives
y	z	Negatives

Figure 2: A confusion matrix.

The entry of a confusion matrix is as follows:

- w (True Positives –TP) is the number of positive examples correctly classified
- x (False Negatives -FN) is the number of positive examples misclassified as negative

- y (False Positives –FP) is the number of negative examples misclassified as positive
- z (True Negatives –TN) is the number of negative examples correctly classified

B) The Metrics

Sensitivity/ Recall

Here, the percentage of sick people who are correctly identified as having the condition and the equation is

$$\text{Sensitivity} = \text{Recall} = w/(w+x) = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

Specificity

Here, the percentage of healthy people who are correctly identified as not having the condition and the equation is

$$\text{Specificity} = z/(y+z) = \frac{\text{TN}}{\text{TN}+\text{FP}}$$

Accuracy

Accuracy is treated as the degree of closeness of measurements of a quantity to its true value.

$$\text{Accuracy} = (w+z)/(w+x+y+z) = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}}$$

Precision/ Positive Predictive Value

The Positive predictive value(PDV,) is calculated using the following equation :

$$\text{PPV} = \text{Precision} = w/(w+y) = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

F_Score

The equation for the f-score is as follows

$$\text{F\_Score} = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Error Rate

The equation for the error rate is as follows

$$\text{Error rate} = (y+x)/(w+x+y+z) = \frac{\text{FP}+\text{FN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}}$$

C) Validation Methods

K-fold cross validation is used in this work for measuring the performance with respect to different metrics.

K-fold Cross-Validation

In this work, the classifier's performance is evaluated by selecting k-fold cross validation as the main metric. The initial data are randomly partitioned into k mutually exclusive subset or folds $f_1, f_2, \ldots, f_k$, each approximately equal in size. The training and testing is performed k times. In the first iteration, fi is tested against the subsets $f_2, \ldots, f_k$, which is collectively serve as the training set in order to obtain a first model; the second iteration is trained in subsets $f_1, f_3, \ldots, f_k$ and tested on f_2; and so no.

IV. THE RESULTS AND DISCUSSION

About the Implementation

The proposed outlier detection software is developed using Matlab version 7.4.0 (R2007a) and decided to use some of the features of Weka . So, the Mex and Java interface of matlab is used to implement this outlier detection software. The standard weka implementation of the classification algorithms is used in this work and only passed the default parameters while invoking the classifier algorithms. The proposed hybrid classification model is developed and the standard fspackage of Matlab is incorporated with it.

In the second plot clearly shows that the benign records are grouped together and form a distinct cluster. The red points that are deviating from the black cluster are the outliers which signifies the malignant nature of that case.

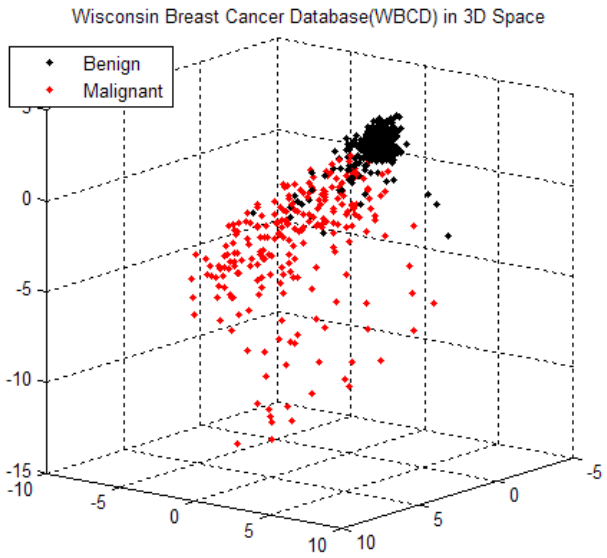

Figure 3: The Plot of WBDC Data Clearly Showing the Benign Cluster and Malignant Outliers

The following table lists the performance of the algorithm with respect to different metrics. In fact, each value is a average of 10 trials. In each trial we did a 10- fold validation. So, each table cell value is the average of 100 separate runs with different training and testing data sets.

Table 2: The Performance of Normal and Proposed Hybrid Classification Algorithm

Algorithm	Precision %	F-Score %	Sensitivity %	Specificity %	Accuracy %	Error Rate %
kNN	96.07	96.66	97.31	92.23	95.57	4.43
Decision Table	96.12	96.19	96.35	92.51	95.03	4.97
C4.5 Classifier	**96.18**	**95.82**	**95.58**	**92.60**	**94.53**	**5.47**
kNN-C4.5	98.13	96.54	95.10	96.43	95.59	4.41
kNN-DT	**98.33**	**96.75**	**95.31**	**96.87**	**95.85**	**4.15**

Even though dimensionality reduction techniques and feature selection techniques will lead to better performance, in our experiments, we didn't use any dimensionality reduction techniques and feature selection techniques. Because, we just want to examine the real improvement in performance only due to the hybrid classification idea. We have selected two classification algorithms to make this hybrid since one is providing better sensitivity and the other is providing better specificity. So we are only interested in evaluating the improvement in performance.

The following bar charts are showing the performance of the algorithms. It clearly shows the difference in performance with respect to different metrics.

The following bar chart shows the performance of the algorithm in terms of Accuracy. In this case, accuracy measures the capability of the algorithms to correctly identify the normal as well as outliers in the data. As shown in the graph, with respect to accuracy, the proposed kNN-DT hybrid algorithm performed well. It means, proposed kNN-DT hybrid algorithm is capable of marking normal as well as the outliers correctly better than other algorithms.

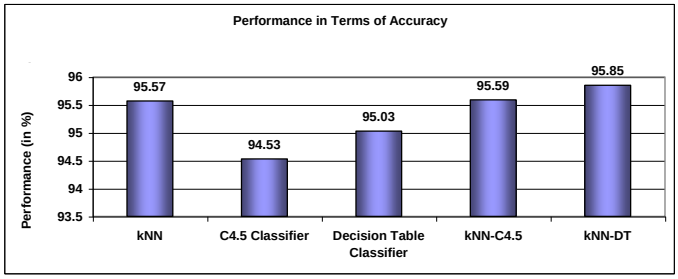

Figure 4: The Accuracy Chart

The following bar chart shows the performance of the algorithm in terms of f-score. In this case, f-score measures the capability of the algorithms to correctly identify the normal as well as outliers in the data. As shown in the graph, with respect to f-score, the proposed kNN-DT hybrid algorithm performed well. It means, proposed kNN-DT hybrid algorithm is capable of marking normal as well as the outliers correctly better than other algorithms.

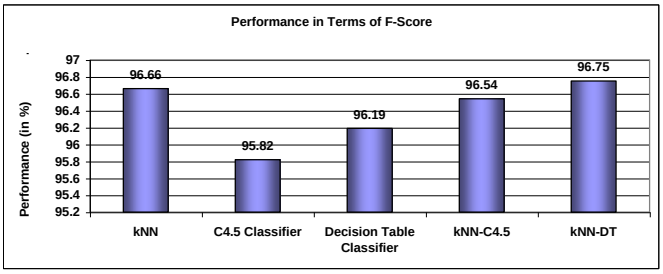

Figure 5: The F-Score Chart

The following bar chart(fig.6) shows the performance of the algorithm in terms of precision. The Positive predictive value(PDV,) or Precision is measures the capability of the algorithms to correctly identify the positives in the data. As shown in the graph, with respect to precision, the proposed kNN-DT hybrid algorithm performed well.

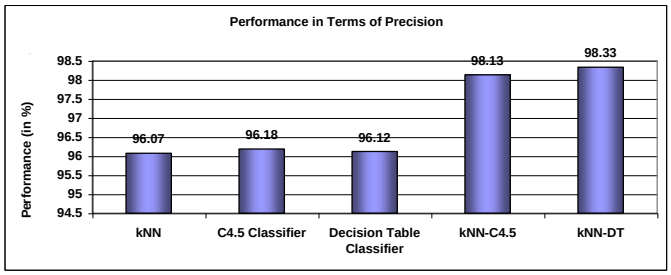

Figure 6: The Precision Chart

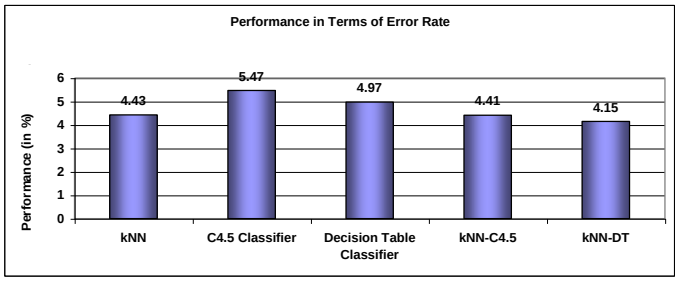

Figure 7: The Error Rate Chart

The above bar chart(fig.7) shows the performance of the algorithm in terms of error rate. In this case, error rate measures how much the algorithm wrongly identify both

the normal as well as outliers in the data. As shown in the graph, with respect to error rate, the proposed kNN-DT hybrid algorithm performed well. It means, the lower value of error rate signifies that proposed kNN-DT hybrid algorithm is making less error while identifying the malignant as well as outlier data.

The following bar chart shows the performance of the algorithm in terms of specificity. In this case, specificity measures the proportion of normal records, that are correctly identified. As shown in the graph, with respect to specificity, the proposed kNN-DT hybrid algorithm performed well.

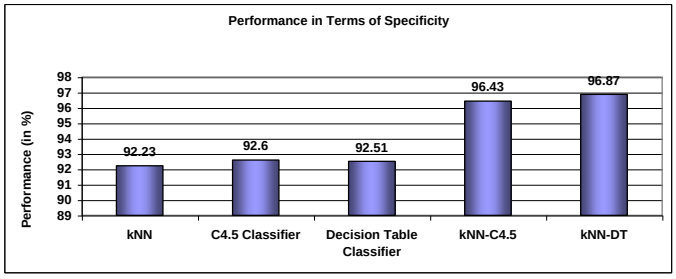

Figure 8: The Specificity Chart

The following bar chart shows the performance of the algorithm in terms of sensitivity or recall. In this case, sensitivity or recall measures the proportion of actual malignant records that are correctly identified as outliers. As shown in the graph, with respect to sensitivity or recall, the proposed kNN-DT hybrid algorithm performed little bit poor. It doesn't mean its overall performance is poor – it means, it is performing good in identifying the outliers by missing some normal records.

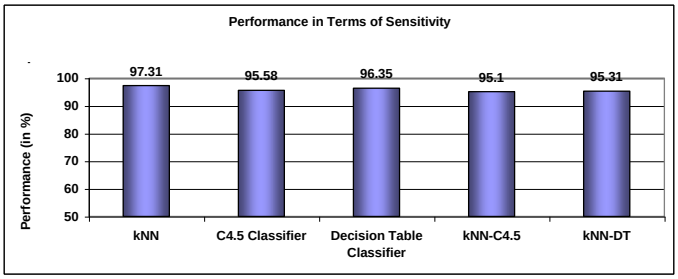

Figure 9: The Sensitivity/Recall Chart

The following bar chart shows the time consumed for the classifier. Even though the proposed hybrid classifier consumed little bit higher time, it provided good improvement with respect to other metrics. So, this slight increase in time can be neglected.

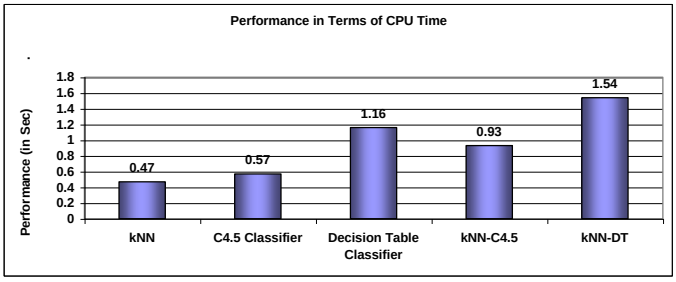

Figure 10: The CPU Time Chart

V. CONCLUSION

We have implemented the hybrid classification based outlier detection algorithm under Matlab and evaluated its performance using different metrics. We have arrived significant and comparable results. The table and graphs in the previous section shows the overall results.

In this work, we evaluated the performance of knn and Decision table hybrid classifier for outlier detection and the results clearly shows that the impact of hybrid technique on the cancer dataset is significantly improve the overall classification performance.

Further, we may address the possibility of improving the classification algorithm using a good distance metric or good neighborhood relationship function and with much suitable hybrid classification. Future works may address these issues and improve the performance of the outlier detection in cancer data.

VI. ACKNOWLEDGEMENT

This study can be considered as a work under the guidance of Holy Spirit , I would like to thank all those who participated in discussion and motivated me while working on this paper. Also , I am greatly thankful to Dr. Gladston Raj S, Head of the Department of Computer Science , Government College Nedumangad ,Kerala. ,India.

REFERENCES

[1] Simon Hawkins, Hongxing He, Graham Williams and Rohan Baxter, "Outlier Detection Using Replicator Neural Networks, DaWaK 2000 Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery Pages 170-180

[2] Graham Williams, Rohan Baxter, Hongxing He, Simon Hawkins and Lifang Gu, "A Comparative Study of RNN for Outlier Detection in Data Mining", ICDM '02 Proceedings of the 2002 IEEE International Conference on Data Mining, Page 709.

[3] Hodge, V.J. and Austin, J. (2004) A survey of outlier detection methodologies. Artificial Intelligence Review, 22 (2). pp. 85-126.

[4] A. Faizah Shaari, B. Azuraliza Abu Bakar, C. Abdul Razak Hamdan, "On New Approach in Mining Outlier" Proceedings of the International Conference on Electrical Engineering and Informatics, Indonesia June 17-19, 2007

[5] Yumin Chen, Duoqian Miao, Hongyun Zhang, "Neighborhood outlier detection", Expert Systems with Applications 37 (2010) 8745-8749, 2010 Elsevier

[6] Xiaochun Wang, Xia Li Wang, D. Mitch Wilkes, "A Minimum Spanning Tree-Inspired Clustering-Based Outlier Detection Technique", Advances in Data Mining. Applications and Theoretical Aspects, Lecture Notes in Computer Science Volume 7377, 2012, pp 209-223

[7] Jiawei Han, Micheline Kamber and Jian Pei, "Data Mining Concepts and Techniques (Third Edition)", Morgan Kaufmann Publishers is an imprint of Elsevier, c 2012 by Elsevier Inc.

[8] Gouda I. Salama, M.B.Abdelhalim, and Magdy Abd-elghany Zeid, Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers, International Journal of Computer and Information Technology (2277 - 0764), Volume 01- Issue 01, September 2012

[9] S. Aruna et al. (2011). Knowledge based analysis of various statistical tools in detecting breast cancer.

[10] Angeline Christobel. Y, Dr. Sivaprakasam (2011). An Empirical Comparison of Data Mining Classification Methods. International Journal of Computer Information Systems,Vol. 3, No. 2, 2011.

[11] D.Lavanya, Dr.K.Usha Rani,..," Analysis of feature selection with classification: Breast cancer datasets",Indian Journal of Computer Science and Engineering (IJCSE),October 2011.

[12] E.Osuna, R.Freund, and F. Girosi, "Training support vector machines: Application to face detection". Proceedings of computer vision and pattern recognition, Puerto Rico pp. 130-136.1997.

[13] Vaibhav Narayan Chunekar, Hemant P. Ambulgekar (2009). Approach of Neural Network to Diagnose Breast Cancer on three different Data Set. 2009 International Conference on Advances in Recent Technologies in Communication and Computing.

[14] D. Lavanya, "Ensemble Decision Tree Classifier for Breast Cancer Data," International Journal of Information Technology Convergence and Services, vol. 2, no. 1, pp. 17-24, Feb. 2012.

[15] B.Ster, and A.Dobnikar, "Neural networks in medical diagnosis: Comparison with other methods." Proceedings of the international conference on engineering applications of neural networks pp. 427-430. 1996.

[16] T.Joachims, Transductive inference for text classification using support vector machines. Proceedings of international conference machine learning. Slovenia. 1999.

[17] J.Abonyi, and F. Szeifert, "Supervised fuzzy clustering for the identification of fuzzy classifiers." Pattern Recognition Letters, vol.14(24), 2195-2207,2003.

[18] Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[19] Street WN, Wolberg WH, Mangasarian OL. Nuclear feature extraction for breast tumor diagnosis. Proceedings IS&T/ SPIE International Symposium on Electronic Imaging 1993; 1905:861-70.

[20] William H. Wolberg, M.D., W. Nick Street, Ph.D., Dennis M. Heisey, Ph.D., Olvi L. Mangasarian, Ph.D. computerized breast cancer diagnosis and prognosis from fine needle aspirates, Western Surgical Association meeting in Palm Desert, California, November 14, 1994.

[21] Chen, Y., Abraham, A., Yang, B.(2006), Feature Selection and Classification using Flexible Neural Tree. Journal of Neurocomputing 70(1-3): 305-313.

[22] J. Han and M. Kamber,"Data Mining Concepts and Techniques", Morgan Kauffman Publishers, 2000.

[23] Duda, R.O., Hart, P.E.: "Pattern Classification and Scene Analysis", In: Wiley-Interscience Publication, New York (1973)

[24] Bishop, C.M.: "Neural Networks for Pattern Recognition". Oxford University Press,New York (1999).

[25] Vapnik, V.N., The Nature of Statistical Learning Theory, 1st ed., Springer-Verlag,New York, 1995.

[26] Ross Quinlan, (1993) C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA.

[27] Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. and Zanasi, A. (1998). Discovering Data Mining: From Concept to Implementation, Upper Saddle River, N.J., Prentice Hall.

[28] Kurian M.J ,Dr. Gladston Raj S. "Outlier Detection in Multidimensional Cancer Data using Classification Based Appoach" International Journal of Advanced Engineering Research(IJAER) Vol. 10 ,No.79 , pp –(342 348) 2015..

[29] Kurian M.J , Dr. Gladston Raj S. " An Analysis on the Performance of a Classification Based Outlier Detection System using Feature Selection" International Journal of Computer Applications (IJCA) Vol.132.No.8. December 2015.

[30] Kurian M.J. ,Dr. Gladston Raj S, " Improving the Performance of a Classification Based Outlier System Using Knn-C4 Hybrid Algorithm " International Journal of Control Theory and Applications (IJCTA) Vol.9 No.10,PP.4695-4704,2016.

AUTHORS PROFILE

Mr. Kurian M.J. received his M.Sc. (Maths), M.C.A., and M.Phil. in computer Science. Now working as Assistant Professor at Baselios Poulose II Catholicos (B. P. C) College, Piravom, Kerala, India. He is the Course Co-ordinator for MSc. Computer Science. and the Principal Investigator of the Minor Research Project "Outlier Detection In Multidimensional Data", 2010, funded by Universities Grant Commission, India. His research interest includes Data Mining and Cyber Plagiarism, and has presented papers in National Seminar on Cyber Criminology organized by Computer Society of India. Currently he is pursuing Ph.D. in Computer Science at Bharathiar University, Coimbatore, Tamilnadu, India.

Dr. Gladston Raj S. received his M.Sc (CS), M.Tech (Image Computing) and PhD in Computer Science from University of Kerala and Completed UGC-NET from University of Kerala and PGDCH (Computer hardware) from MicroCode, He is Now working as Head of the Department of Computer Science at Govt. College Nedumangad, Kerala, India. His area of interest includes Image Processing, Signal Processing, Datamining. He is providing research guidance for Ph.D scholars from different areas of research and has presented several invited talks in this areas of research.