International Journal of Computing and Applications | **ISSN: 0973-5704**
Volume 13, Number 1, (January-June 2018), pp 201-209
© Serials Publications, New Delhi (India)

IMPROVING THE PERFORMANCE OF A HYBRID CLASSIFIER BASED OUTLIER DETECTION SYSTEM USING DIMENSIONALITY REDUCTION TECHNIQUES

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

[1] *Research Scholar, Research and Development Centre, Bharathiar University, Coimbatore.*
Email: kurianmj@yahoo.com
[2] *HOD of Computer Science, Government College, Nedumengadu, Kerala.*
Email: gladston@rediffmail.com

Abstract: In data mining, outlier detection can be treated as classification problem with the availability of class based sample data. Classification based outlier identification method cab be applied to the samples, available with class information, of medical cancer dataset. The core idea of the method is to train a classification model that can distinguish normal data from outliers. Initially, derive KNN, C4.5 and DT classification models using training data and classify test data using these models. Secondly, derive two hybrid classifiers KNN-C4.5 and KNN-DT using highest precision from KNN and highest recall from C4.5 and DT respectively. This work tried to improve the performance of hybrid classification algorithms using dimensionality reduction techniques such as Principal Component Analysis and Locality Preserving Projection and evaluated the performance of outlier detection. The results clearly show that the impact of such hybridizing and dimensionality reduction significantly improved the overall classification performance to a considerable level.

Keywords: Outlier, Hybrid classifier, Data mining, Decision Table, C4.5, KNN KNN-DT, KNN-C4.5, PCA, LPP.

1. INTRODUCTION

In Data Mining, Outliers are meaningful items rather than disturbance. In some applications outlier represent unique characteristics of the objects. This work concentrates on the performance of the algorithms in outlier detection using dimensionality reductions techniques. The whole data is examined for the strange items that are far moved away from the data set .These observations are known as outliers.

Outlier Detection in High-Dimensional Data

In high dimensional data set, some attributes may be irrelevant .But by using feature selection approaches such as filler and wrapper, it has to find the subset of the original attributes.

Problem Specification

The identification of outlier can be viewed as classification problem which can lead to the invention of unpredicted awareness in the medical field. The general idea is to develop a model based on classification, which can distinguish normal data from abnormal [7].

In medical cancer dataset, the available number of malignant/outlier samples are less than that of the normal/benign and it causes an inaccurate classifier model. Many solutions like factor analysis and principle component methods were suggested to improve the efficiency of the algorithm with the elimination of variables. The algorithms for diminution of dimensionality and selection of attribute can be used to give up the problems raised in performance evaluation and accuracy testing of classification algorithm for detection of outliers in cancer data set.

To enhance the efficiency of algorithm, propose a hybrid classification approach using KNN and decision tree.

2. HYBRID CLASSIFICATION MODELING

A. *Methods of Outlier Identification*

The popular methods of outlier detection are supervised, semi supervised, unsupervised proximity-based. The Grubb's test detects one outlier at a time in a univariate data. The Rosener test is a sequential procedure for identifying maximum of 10 outliers. So there is a more sophisticated and speedy method such as classification based outlier detection, which heavily depends of the quality and availability of training data set.

B. *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, if we combine the class labels provided by these two classifier, then the resultant class label will be much accurate than the two.

C. *The Used Classification Algorithms to Create Hybrid Classifier*

(a) **Decision Table Classifier:** A predictive modeling tool having a hierarchical data breakdown with two attributes at each stage and thus identifies the best attributes for the categorization of data.

(b) **K-Nearest Neighbors Classifier:** 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 Euclidean distance between two states a_i and a_j is

$$D(a_i, a_j) = \sqrt{\sum_{k=1}^{n} (y_{ik} - y_{jk})^2}$$

where, n denoted as the number of properties in each data instance and the value of k must be more than the classes in the problem.

(c) **C4.5 Classifier:** It is an efficient decisional algorithm, which creates a model of tree by considering one attribute at a time. Initially, the algorithm arranges the dataset on the attribute's value. After that it looks for regions that contain only one class and mark it as leaf. Again, the algorithm selects another attribute and repeat the branching process until it produces all leaves. The attribute selection is based on the calculation of the information gain for each attribute by subtracting the entropy of the attribute form the entire dataset entropy.

Gain Ratio (attribute, dataset)

$$= \frac{\text{Gain (attribute, dataset)}}{H(p(\text{range}_1), \ldots, p(\text{range}_n))}$$

with

H(p(range$_1$), ..., p(range$_n$))

$$= \sum_{i=1}^{n} p(\text{range}_i) \log\left(\frac{1}{p(\text{range}_i)}\right)$$

D. *The Dimensionality Diminution Algorithms*

The dimensionality diminution is the search for a subset of attributes, number of variables, to describe the original dimension.

(a) **Principal Component Analysis:** Principal Component Analysis is used to leaving out the data which is of the least important to the information stored in the data. It compresses an N- dimensional vector to M-dimensional vector, where M<N.

(b) **LPP (Locality Preserving Projection):** LPP is a classical linear technique which projects the data along the directions of maximal variance by calculating the optimal linear approximations to the Eigen functions of the Laplace Beltrami operator on the main fold.

The Outline of the Improved Hybrid classification Model

- Reduce the dimension of input data
- Classify the reduced dimension data using algorithm 1 and find the classification labels A1.
- Classify the reduced dimension data using algorithm 2 and find the classification labels A2.
- Let A1 be the set of class labels Provided by algorithm 1 which is capable of identifying benign 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.
- Let A2 be the set of class labels Provided by algorithm 2 which is capable of identifying malignant records with greater accuracy.
- 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 improved hybrid outlier detection system that is going to construct and test in this work.

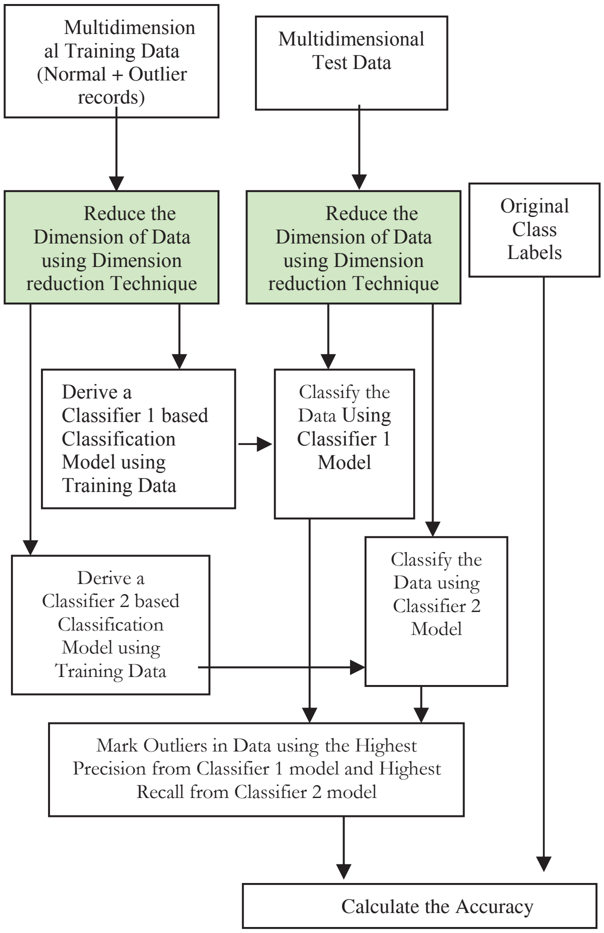
Multidimensional Training Data (Normal + Outlier records)
Multidimensional Test Data
Reduce the Dimension of Data using Dimension reduction Technique
Reduce the Dimension of Data using Dimension reduction Technique
Original Class Labels
Derive a Classifier 1 based Classification Model using Training Data
Classify the Data Using Classifier 1 Model
Derive a Classifier 2 based Classification Model using Training Data
Classify the Data using Classifier 2 Model
Mark Outliers in Data using the Highest Precision from Classifier 1 model and Highest Recall from Classifier 2 model
Calculate the Accuracy

Figure 1: The Improved Hybrid Outlier Detection System

3. THE EVALUATION

The "Wisconsin Breast Cancer Database " is used for the performance evaluation of the algorithms

Breast Cancer Dataset

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 has 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 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 benign and malignant samples of 444 (65.0%) and 239 (35.0%) respectively (Tan et. al., 2003).

Evaluation Metrics

Performance of a classification algorithms are evaluated with metrics and use two measures such as Rand Index and Run Time.

Table 1
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%)

(a) **Total Run Time:** The total time taken for training and testing is called total run time. Here, compare the CPU times only. In the following tables concentrate only on the time taken for training because e the time taken for training is very much higher than the time required for testing the network with same number of records.

The Metrics and Validation Method Used for Performance Analysis

The Performance of the selected algorithms are depend on data's characteristics, which is measured with specificity, accuracy, sensitivity, error rate, f-score and precision of metrics.

(A) Confusion Matrix

A Confusion matrix shows the type of classification error a classifier produced. This matrix tells how many got misclassified and what level of misclassification occurred.

Figure 2
A confusion matrix

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

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.

$$\text{Accuracy} = (w + z)/(w + x + y + z)$$

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:

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

Error Rate

The equation for the error rate is as follows:

$$\text{Error rate} = (y + x)/(w + x + y + z)$$

Time

The Speed of the algorithm is evaluated using a metric, which is treated as the cpu time.

(C) Validation Methods

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 one 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.

4. 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 invoking the algorithms with default parameters. 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. Outliers are the red points which are deviating from black cluster signifies the malginant nature.

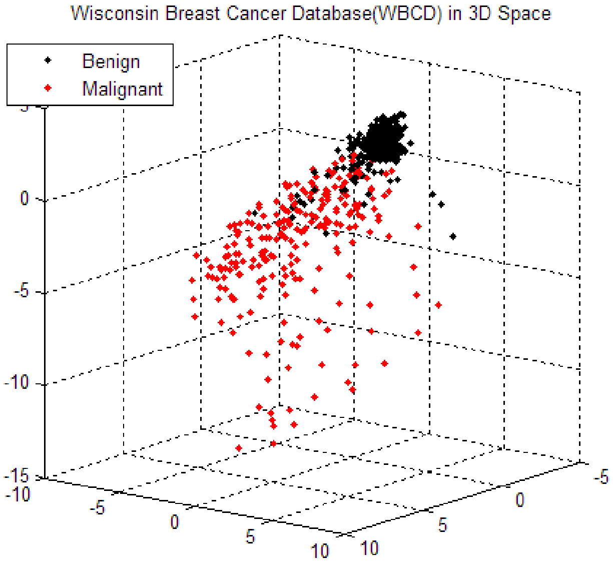

Figure 3: Benign Cluster and Malignant Outliers

The performance of the algorithm with respect to different metrics is shown below. A 10-fold validation is taken in each trail and each value is an average of 10 trails. So, each table cell value is the average of 100 separate runs with different training and testing data sets.

Table 1
The Performance of Outlier Detection with different Feature Dimensionality Reduction Algorithms and Classification Algorithms

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
PCA+ kNN-C4.5	99.49	97.88	96.38	99.00	97.31	2.69
LPP+ kNN-C4.5	99.41	97.87	96.43	98.92	97.28	2.72
KNN-DT	98.33	96.75	95.31	96.87	95.85	4.15
PCA+ KNN-DT	99.75	97.89	96.15	99.53	97.37	2.63
LPP+ KNN-DT	99.53	97.80	96.17	99.10	97.21	2.79

This experiment concentrate on the e examination of the real improvement in performance only due to the hybrid classification. Here, only two classification algorithms are selected to make a hybrid since one is providing better sensitivity and other is providing better specificity. So, it is only interested in evaluating the improvement in performance.

The performance of the algorithm in terms of Accuracy is shown below. With respect to accuracy, the proposed PCA+KNN-DT hybrid and PCA+kNN-C4.5 hybrid algorithms are performed well.

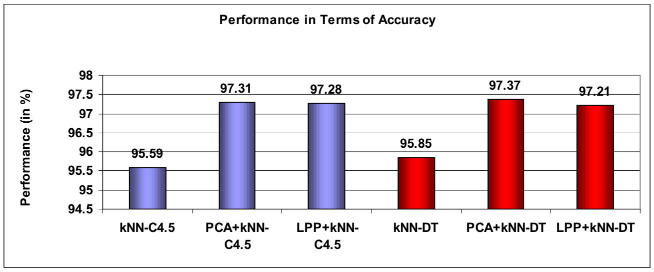

Figure 4: The Accuracy Chart

The performance of the algorithm in terms of *f*-score is shown below. The *f*-score measures the capability of the algorithms to correctly identify the normal as well as outliers in the data. With respect to *f*-score, the proposed PCA+KNN-DT hybrid algorithm and proposed PCA+kNN-C4.5 hybrid algorithm are performed well.

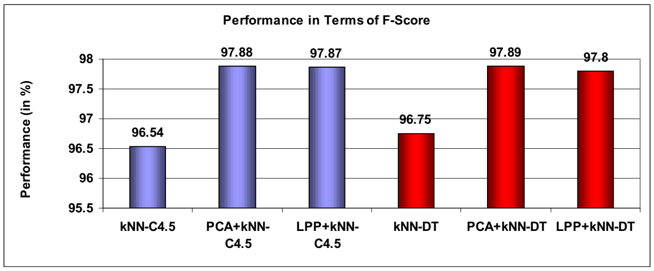

Figure 5: The F-Score Chart

The Precision performance of algorithm is shown below. 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 PCA+KNN-DT and PCA+kNN-C4.5 hybrid algorithms performed well in aspect of precision.

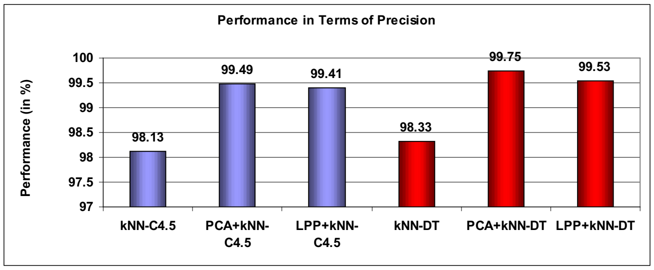

Figure 6: The Precision Chart

Normally, error rate is the measure of how much the algorithm wrongly identifies both the normal as well as outliers in the data. The graph given below reveals that, with respect to error rate, the proposed hybrid algorithms such as PCA+KNN-DT and PCA+KNN-C4.5 are performed well.

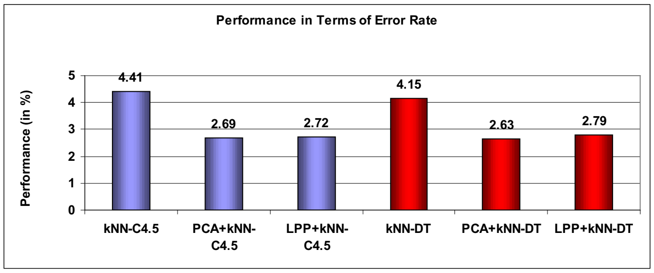

Figure 7: The Error Rate Chart

The performance of the algorithm in terms of specificity is shown below. 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 PCA+kNN-DT hybrid algorithm, PCA+kNN-C4.5 hybrid algorithm performed well.

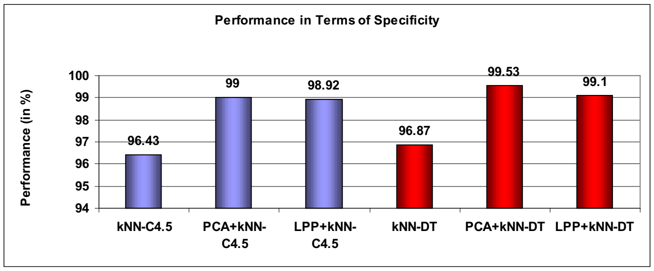

Figure 8: The Specificity Chart

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, there was no improvement in performance due to the application of dimensionality reduction.

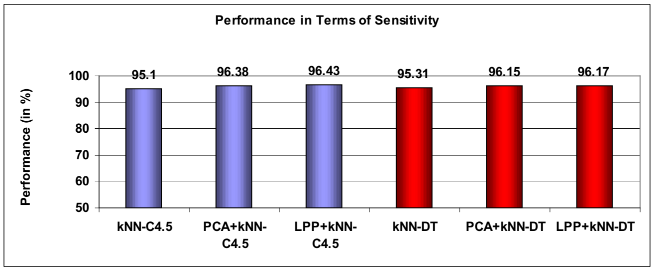

Figure 9: The Sensitivity/Recall Chart

The comparative study of previous work with this work is shown in the table.

Table 2
The Comparison with Resent Works

S.No.	*Classifiers*	*Classification accuracy*
1	SVM-RBF kernel[9]	96.84%
2	SVM[10]	96.99%
3	CART with feature selection (Chi-square)[11]	94.56%
4	C4.5 [12]	94.74%
5	Hybrid Approach[14]	95.96%
6	Linear Discreet Analysis[15]	96.8%
7	Neuron-Fuzzy[16]	95.06%
8	Supervised Fuzzy Clustering [17]	95.57%
9	SMO+J48+NB+Ibk[8]	97.28%
10	*Proposed PCA+C4.5	97.29%
11	*Proposed PCA+DT	97.31%
12	*Proposed PCA+KNN	95.85%
13	#Proposed PCA+kNN-C4.5	97.31%
14	#Proposed LPP+kNN-C4.5	97.28%
15	#Proposed PCA+KNN-DT	97.37%
16	#Proposed LPP+KNN-DT	97.21%

*The First Five Principal Components were used in this proposed classification models.

#The First Three Principal Components were used in this proposed hybrid classification models.

As shown in the following bar chart, the following five of implemented hybrid algorithms competed all other previous methods in terms of accuracy.

1. Proposed PCA+C4.5 Algorithm
2. Proposed PCA+DT Algorithm
3. Proposed PCA+kNN-C4.5 Hybrid Algorithm
4. Proposed LPP+kNN-C4.5 Hybrid Algorithm
5. Proposed LPP+KNN-DT Hybrid Algorithm

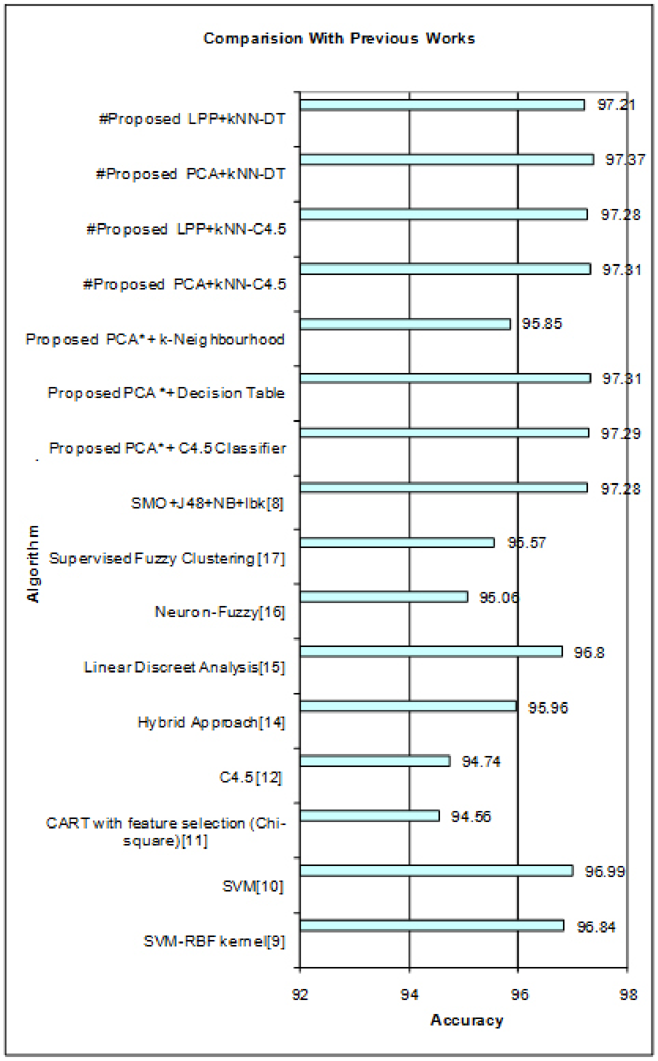

Figure 10: Comparison of Accuracy

5. CONCLUSION

The hybrid classification based outlier detection algorithms are implemented under Matlab and improved their performance using dimensionality reduction techniques and also evaluated its performance using different metrics. The significant and comparable results are obtained at this experiment. The table and graphs in the previous section shows the overall results.

In this work, the performance of two hybrid classification algorithms using KNN, C4.5 and

Decision table hybrid classifier for outlier detection is evaluated and the result clearly shows that the impact of dimensionality reduction with hybrid classification 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.

Acknowledgement

This study can be considered as a work under the guidance of Holy Spirit, I would like to thank all those who have 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 and 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 and 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 and 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.

[31] Kurian M.J and Dr. Gladston Raj S, "An Analysis on the Performance of a K-Nearest-Neighbor Classification Based Outlier Detection System using Feature Selection and Dimensionality Reduction Techniques" International Journal of Scientific Inventions and Innovations Vol. 1. No. 1, PP. 1-7 July 2016.