IJCTA, 9(10), 2016, pp. 4695-4704

Improving the Performance of a Classification Based Outlier System Using Knn-C4 Hybrid Algorithm

***Kurian M. J **Dr. Gladston Raj S**

Abstract : Outlier detection can be treated as a classification problem with the availability of a training data set with class labels. In a typical medical cancer dataset , it is possible to apply a classification based outlier detection method by the consideration of samples 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]. In the previous work, it was implemented and evaluated some of the classification based outlier detection algorithms and found that some algorithm provided better sensitivity and some algorithm provided better specificity [30]. By combining the better part of this classification algorithms, we design a hybrid classification algorithm. In this work, we proposed a Knn-C4.5 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.

1. INTRODUCTION

Outliers in Data

Outlier is an observation that deviating from others in a random sample, which represents the unique characteristic of the object. This work mainly concentrated on the study of classification algorithms in data mining for the detection of outliers in high dimensional cancer datasets. The distance between objects may be heavily dominated by noise as dimensionality increases[7].

Problem Specification

If a training data set with class labels ,such as "normal" and "outliers", are available , then a classifier can be modeled and any new data can be classified based on that training model[7].The performance of classification algorithm can be improved by proposing a hybrid classification approach using K-NN and C4.5 algorithms.

2. MODELING HYBRID CLASSIFICATION BASED OUTLIER DETECTION SYSTEM

A. Outlier Detection Methods

The popular outlier detection methods are statistical,proximity-based, supervised, semi-supervised andunsupervised.In statistical methods, any data that not following the data normality is treated as outlier. But in proximity-based method, the proximity of outlier object to its neighbors is significantly different from the proximity of objects to most of the objects in the data set [7].

In supervised outlier detection methods, a new data instance is compared against a predicative model build around training data set for normal and outlier classes. But, in semi-supervised, the new data instance which will not

* Research Scholar , Research and Development Centre, Bharathiar University , Coimbatore. kurianmj@yahoo.com

** Head of Department of CS, Govt.College, Nedumangadu , Trivandrum, Kerala, India.gladston@rediffmail.com

satisfy the condition when compared with a model constructed only on normal class is treated as outlier. In unsupervised method,the training data is not available,and so it treats the data instances which are frequent or closely related as normal and other are outliers

Limitations in using Old Methods

The classical method such as Grubb's test is not suitable for Wisconsin Breast Cancer multidimensional dataset because it uses univariate normal data set and most of the points are treated as outliers.

B. Classification Based Methods

In this work we address some of the classification based outlier detection methods.

Advantages of using Classification Based Outlier Detection Model

The main advantage of classification based outlier detection is to train a classification-based model and compare the objects to be examined against the model developed from the training data. So the quality of the method heavily depends on the availability and quality of training data sets.

C. The Used Classification Algorithms

(*a*) C4.5 Classifier

Interactive Dichotomizer-3 is the one of the earliest algorithm for building Decision Trees based on the concept of information gain. C4.5 is an extension of ID-3 in which classification is based on numerical attributes, and missing or noisy data can be incorporated via pruning [13].It is a non backtracking approach such a way that trees are constructed in top-down recursive divide-and-conquer manner.By considering values of one attribute at a time, a decision tree algorithm creates a model.First up all, it arranges dataset on the attribute's value and find theleaves from the dataset as a region that clearly contain only one class. The algorithm select another attribute from the remaining regions and continue the branching until itproduces all leaves .The attribute with highest information gain is selected for partition.

C4.5 avoids the over fitting of training data and uses the Gain Ratio instead of information Gain, as below:

$$\text{Gain Ratio (attribute, set)} = \frac{\text{Gain (attribute, set)}}{\text{H}(p(\text{range}_1),\ldots, p(\text{range}_k))}$$

With $$\text{H}(p(\text{range}_1),\ldots,p(\text{range}_k)) = \sum_{i=1}^{k} p(\text{range}_i)\log\left(\frac{1}{p(\text{range}_i)}\right)$$

The algorithm K-Nearest Neighbors, nearest measurement refers to Euclidean distance between two instances;assign the instance to the class the majority of k-nearest neighbors belong to.The Euclidean distance between t_i and t_j is

$$\text{D}(t_i, t_j) = \sqrt{\sum_{i=1}^{k}(x_{in} - x_{jn})^2},$$

Where, *k* denoted as the number of attributes in each data instance

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 C4.5 is capable of identifying the malignant records in a better manner, then by combine the class labels provided by these two classifier, the resultant class label will be much accurate than the two.

The outline of the proposed Hybrid classification Model

- Classify the data using algorithm 1 and find the classification labels L1
- Classify the data using algorithm 2 and find the classification labels L2
- Let L1 be the set of class labels Provided by algorithm 1 which is capable of identifying benign records with greater accuracy
- L1 = { LB1 , LM1 } where LB1 are the indexes of Benign records and LM1 are the indexes of the Malignant records provided by algorithm 1
- Let L2 be the set of class labels Provided by algorithm 2 which is capable of identifying malignant records with greater accuracy
- L2 = { LB2 , LM2 } where LB2 are the indexes of Benign records and LM2 are the indexes of the Malignant records provided by algorithm 2
- Combine L1 and L2 in such a way to produce L3 = {LB1, LM2}, which will has higher accuracy than both L1 and L2.

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. .

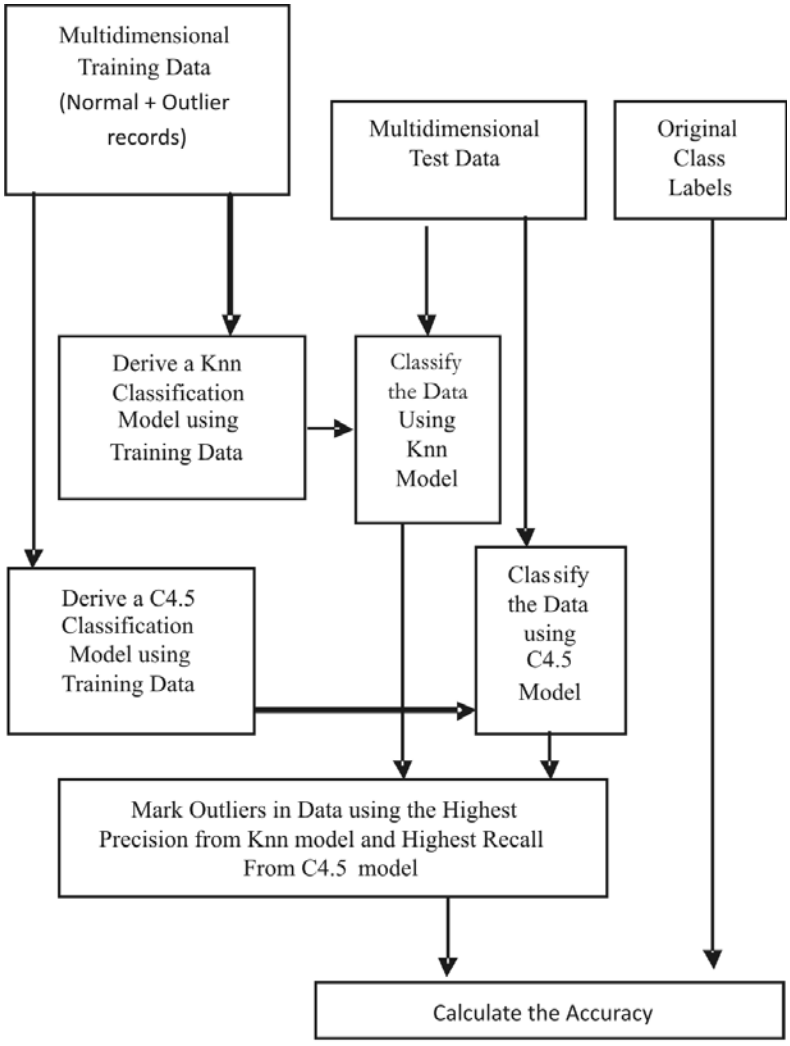

Fig. 1. The Precision and Recall Based Hybrid Outlier Detection System

3. THE EVALUATION

The performance of the classification algorithms under evaluation were tested with "Wisconsin Breast Cancer Database (WBCD)" obtained from the UCI online machine-learning repository at http://www.ics.uci.edu/~mlearn/MLRepository.html. It is summarized in Table 1 and consists of 699 instances taken from fine needle aspirates (FNA) of human breast tissue. The measurements are assigned an integer value between 1 and 10, with 1 being the closest to benign and 10 the most anaplastic. This dataset contains 16 instances with missing attributes' values and so the remaining 683 samples are taken for use [19].

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%)

Metrics Used For Evaluation

To evaluate the classification algorithms under consideration, the two measures are Rand Index and Run Time .

(*a*) Total Run Time

Even though the total run time is the sum of time required for training and the time required for testing ,here we compare the CPU times only. We consider only the time taken for training because the time taken for training is the very much higher than the time required for testing the network with same number of records.

The Metrics and Validation Method Used for Performance Evaluation

The characteristics of the data determine the algorithm's performance and are measured with metrics Sensitivity, Specificity, Accuracy, Precision, F_Score, and Error Rate.

(A) Confusion Matrix

A confusion matrix showsthe performance of classifier and the type of classification errors a classifier makes. The format of a typical confusion matrix is shown below.

Predicted Class		
Positives	Negatives	Actual Class
p	*q*	Positives
r	*s*	Negatives

Fig. 2. A confusion matrix.

The breakdown of a confusion matrix is as follows:

- *p* is the number of positive examples correctly classified (True Positives –TP)
- *q* is the number of positive examples misclassified as negative(False Negatives -FN)
- *r* is the number of negative examples misclassified as positive(False Positives –FP, those do not belong to a class but were allocated to it)
- *s* is the number of negative examples correctly classified(True Negatives –TN ,those belong to a class but were not allocated to it).

(B) The Metrics

Sensitivity & Specificity

$$\text{Sensitivity} = \text{Recall} = p/(p+q) = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = r/(r+s) = \frac{\text{TP}}{\text{TN} + \text{FP}}$$

Accuracy, positive predicative Value, F-Score and Error Rate

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}},$$

$$\text{PPV} = \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

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

$$\text{Errorrate} = \frac{\text{FP} + \text{FN}}{\text{TP} +\_ \text{FP} + \text{TN} + \text{FN}}$$

(C) Validation Methods

The performance of the classifier is evaluated using the *k*-fold cross validation method.. The initial data are randomly partitioned into k mutually exclusive subsets $d_1, d_2, \ldots, d_k$, with approximately equal in size. The training and testing is performed *k* times. In order to obtain the first model, the subsets $d_2, \ldots, d_k$ is selected as the training set, which is tested on d_1; the next iteration is trained in subsets $d_1, d_3, \ldots, d_k$ and tested on d_2; and so no.

4. THE RESULTS AND DISCUSSION

About the Implementation

We have developed the proposed hybrid classification model based on outlier detection software using Mathlab version7.4.0 (R2007a) and incorporated it with the standard fspackage of Matlab.

In the second plot, it 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.

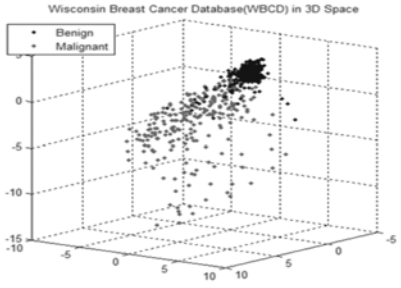

Fig. 3. The Plot of WBDC Data Clearly Showing the Benign Cluster and Malignant Outliers.

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

Algorithm	*Precision %*	*F-Score%*	*Sensitivity %*	*Specificity %*	*Accuracy %*	*Error Rate %*
k-Neighbourhood	96.07	96.66	97.31	92.23	95.57	4.43
C4.5 Classifier	96.18	95.82	95.58	92.60	94.53	5.47
***k*NN-C4.5**	**98.13**	**96.54**	**95.10**	**96.43**	**95.59**	**4.41**

The above 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. 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 another 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. They clearly show 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-C4.5 hybrid algorithm performed well. It means, proposed *k*NN-C4.5 hybrid algorithm is capable of marking normal as well as the outliers correctly better than other two algorithms.

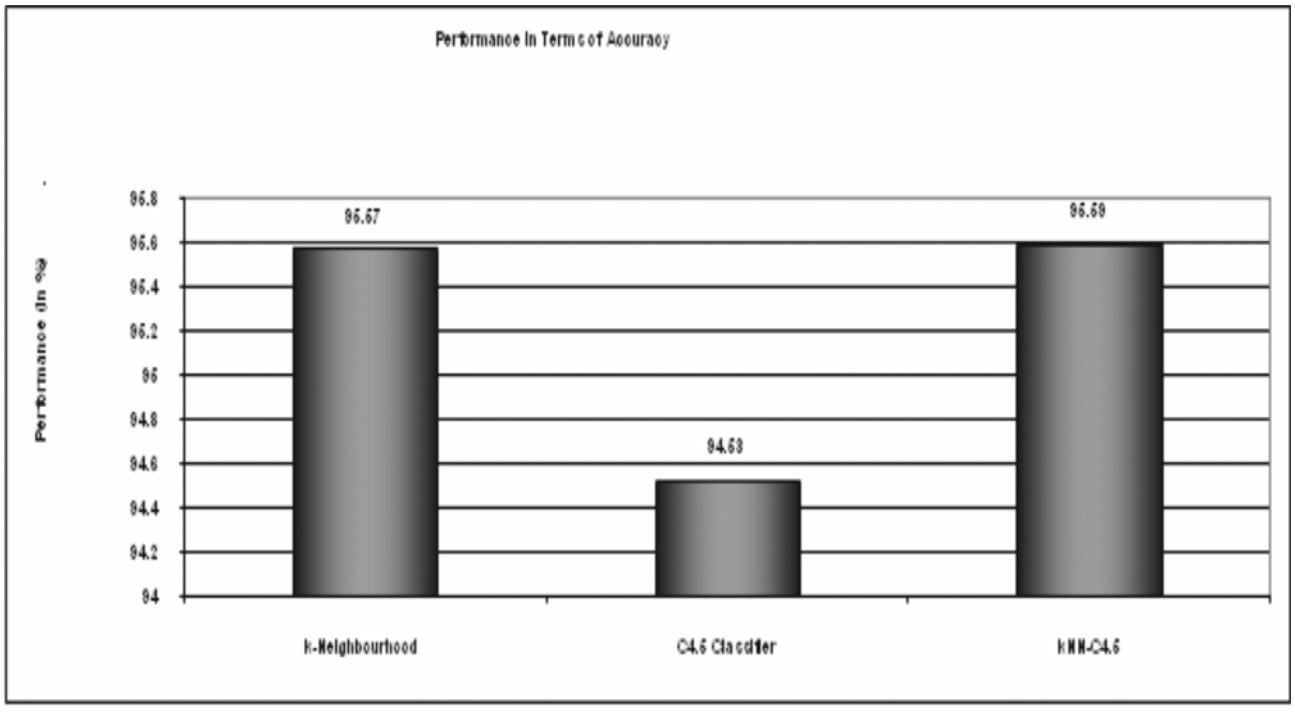

Fig. 4. The Accuracy Chart.

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 *k*NN-C4.5 hybrid algorithm performed well. It means, proposed *k*NN-C4.5 hybrid algorithm is capable of marking normal as well as the outliers correctly better than other two algorithms.

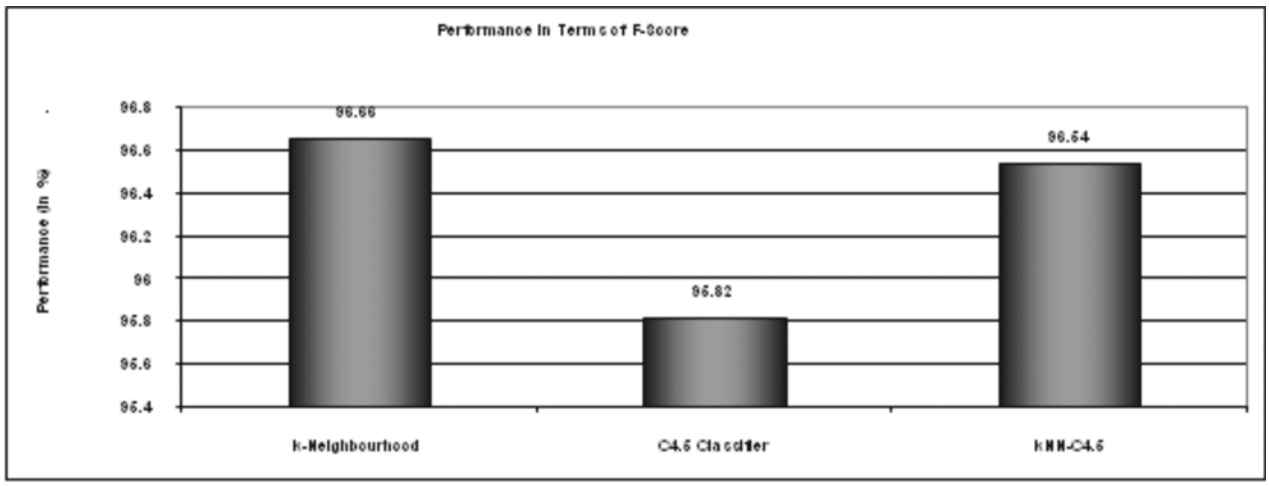

Fig. 5. The F-Score Chart.

The following bar chart 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 *k*NN-C4.5 hybrid algorithm performed well.

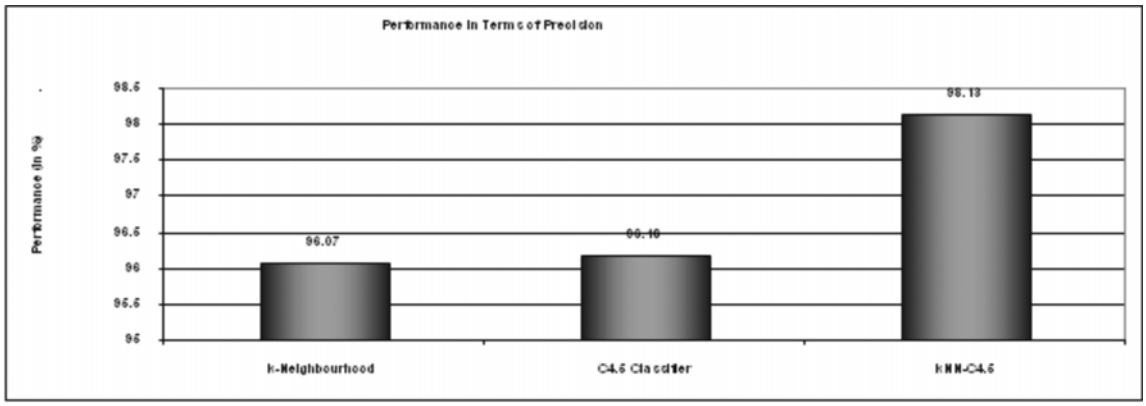

Fig. 6. The Precision Chart.

The following bar chart shows the performance of the algorithm in terms of error rate. In this case, error rate measures how much the algorithm wrongly identifies both the normal as well as outliers in the data. As shown in the graph, with respect to error rate, the proposed *k*NN-C4.5 hybrid algorithm performed well. It means, the lower value of error rate signifies that proposed *k*NN-C4.5 hybridalgorithm is making less error while identifying the malignant as well as outlier data.

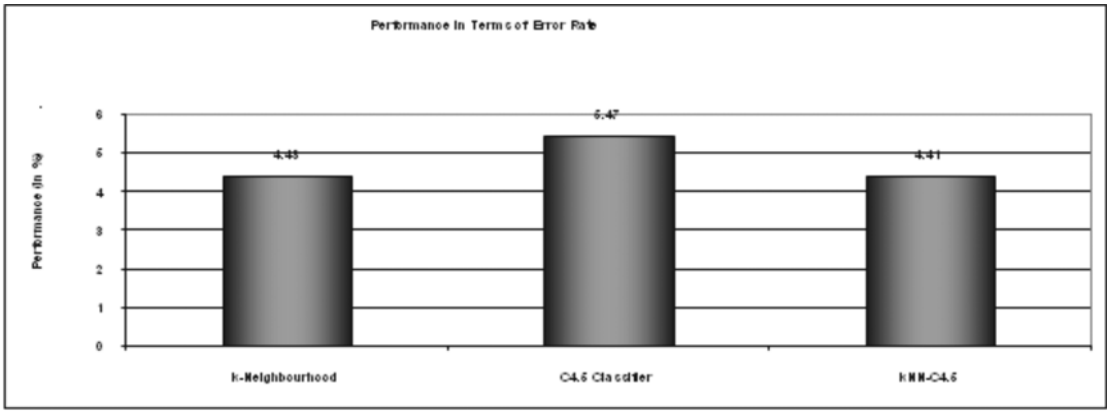

Fig. 7. The Error Rate Chart.

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 *k*NN-C4.5 hybrid algorithm performed well.

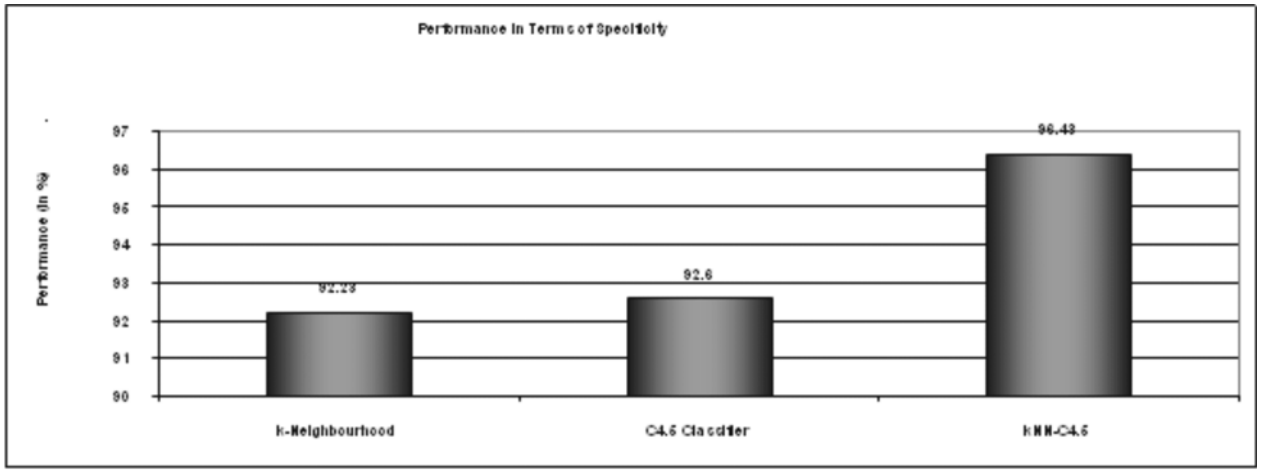

Fig. 8. The Specificity Chart.

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-C4.5 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.

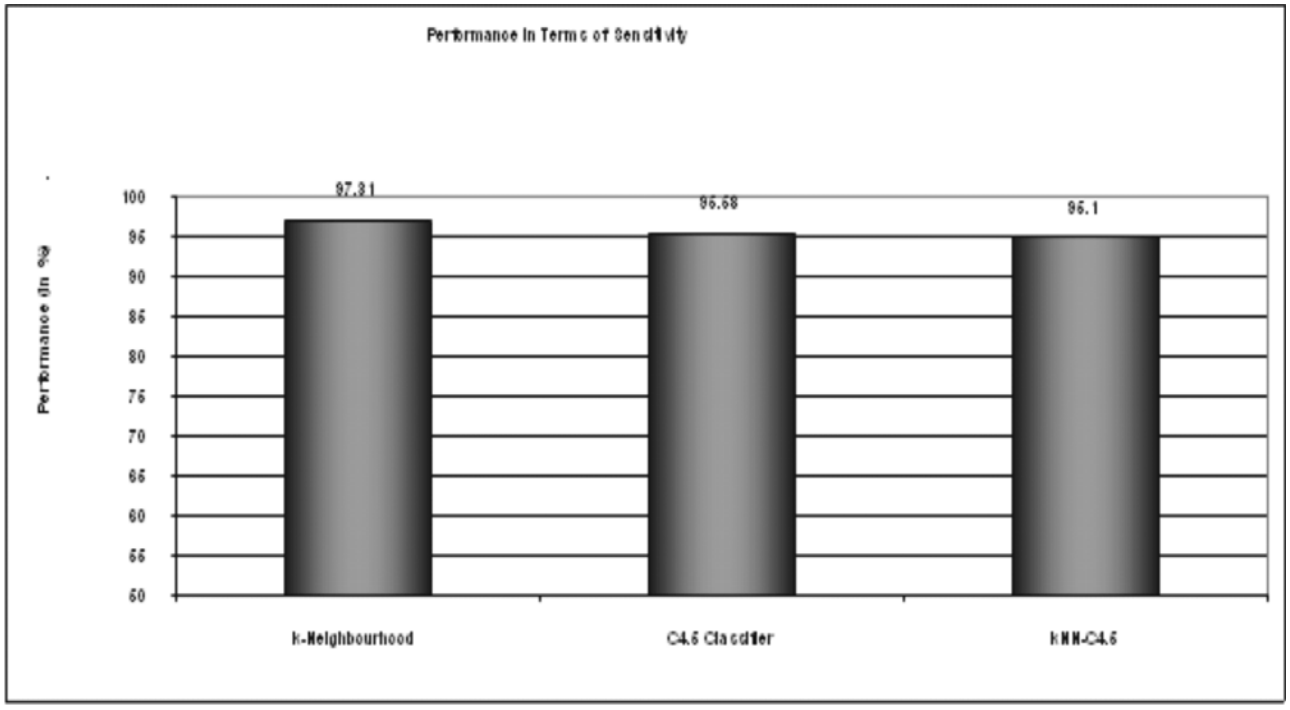

Fig. 9.The Sensitivity/Recall Chart.

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.

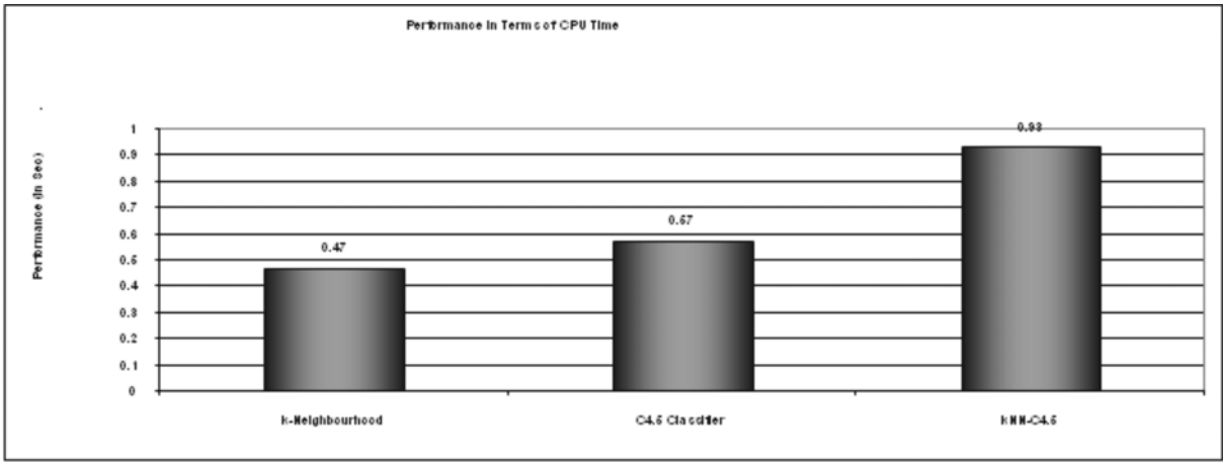

Fig. 10.The CPU Time Chart.

5. CONCLUSION

The hybrid classification based outlier detection algorithm under Matlab is implemented and also evaluated its performance using different metrics. In this work, it has been arrived significant and comparable results. The table and graphs in the previous section shows the overall results [30].

In this work, the performance of knn - C4.5 hybrid classifier for outlier detection is evaluated and the results clearly show that the impact of hybrid technique on the cancer dataset is significantly improving 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.

6. 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

7. 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, IndonesiaJune 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 JianPei, "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. Jones, M.C. Twala.B., Hand, D.T.(2008).Good methods for coping with missing data in decision trees , pattern recongnition letters,29,950-956.
14. 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.
15. 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.
16. 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.
17. T.Joachims, Transductive inference for text classification using support vector machines. Proceedings of international conference machine learning. Slovenia. 1999.
18. J.Abonyi, and F. Szeifert, "Supervised fuzzy clustering for the identification of fuzzy classifiers." Pattern Recognition Letters, vol.14(24), 2195-2207,2003.
19. 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.
20. 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.
21. 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.
22. Chen, Y., Abraham, A., Yang, B.(2006), Feature Selection and Classification using Flexible Neural Tree. Journal of Neurocomputing 70(1-3): 305-313.
23. J. Han and M. Kamber,"Data Mining Concepts and Techniques", Morgan Kauffman Publishers, 2000.
24. Duda, R.O., Hart, P.E.: "Pattern Classification and Scene Analysis", In: Wiley-Interscience Publication, New York (1973)
25. Bishop, C.M.: "Neural Networks for Pattern Recognition". OxfordUniversity Press,New York (1999).
26. Vapnik, V.N., The Nature of Statistical Learning Theory, 1st ed., Springer-Verlag,New York, 1995.
27. Ross Quinlan, (1993) C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA.
28. 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.
29. Jones, M.C. ,Twala,.B, hnd, D.J (2008). Good methods for coping with missing data in decision trees , Patten Recognition Letters ,29(2),950-956.
30. 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
31. 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.