An Analysis on the Performance of a Classification based Outlier Detection System using Feature Selection

Kurian M.J.
Research Scholar
Research and Development Centre
Bharathiar University ,Coimbatore

Gladston Raj S., PhD
Head of Department of CS
Govt. College, Nedumangadu
Trivandrum, Kerala, India

ABSTRACT

Outlier detection can be viewed as a classification problem if a training data set with class labels is available. Generally, in a typical medical dataset such as a cancer data set, if there are samples available with class information, then it is possible to apply a classification based outlier detection method. The general idea of classification-based outlier detection method is to train a classification model that can distinguish normal data from outliers [7]. Previous work had implemented and evaluated using the three classifications based outlier detection algorithms and found that the k-neighborhood algorithm was capable of identifying and classifying the outliers better than the other two compared algorithm in terms of accuracy, f-score, Sensitivity/Recall, error rate. Further, the cpu time of the k-neighborhood algorithm also minimum [23]. In this work, the performance of outlier detection using feature selection algorithms are evaluated but the results clearly shows that the impact of feature selection algorithm on the cancer dataset is very low and does not improve the overall classification performance.

Keywords

Outlier, Gini Index , Information Gain, Chi-Square

1. INTRODUCTION

In statistics, there are number of validation methods for outlier detection. In data mining concept, outliers are treated as “meaningful input signals”. The aim of this work is to study the classification algorithms of data mining for outlier detection in cancer datasets based on the unique characteristics of the objects

1.1 Outliers in Data

An outlier is an abnormal observation in the sense that the data for unusual observations are removed from the mass of the data or most deviated from other values in a random sample from a population. We should try to understand why they appeared before the elimination of these points.

1.2 Outlier Detection in High-Dimensional Data

Most of the cases are aimed to detect outliers in high dimensional data. As dimensionality increases, error or noise rate also increases. The outlier detection method faces the challenges like data subspaces, data sparsity and scalability. So Aggrawaland Yu developed the sparsity coefficient-based subspace outlier detection method[5] kriegel ,Schubert , and Zimek Proposed angle-based outlier detection [13].

1.3 Feature Selection Algorithms

Feature selection methods select a subset of the attributes or properties or the original variables. Filter and wrapper are the main two strategies

1.4 Problem Specification

Even though the training data set is heavily biased or unbalanced, we can model classification based outlier detection and any new or fresh data can be classified accordingly

Support Vector Machines, Fuzzy Art Neural Networks can be used to solve such one class classification problem. Sometimes, the model can detect new outliers that may not appear close to any outlier objects in the training set. This occurs as long as such new outliers fall outside the decision boundary of the normal class. Most difficult task is to obtain high-quality training data. Moreover, the problem becomes much difficult if the data is a multi-dimensional one

By assigning variables to different groups, the model can reduce the size of the data set and then can improve the efficiency of outlier detection. Second idea is to eliminate some variables by using the data reduction methods such as Principal components and factor analysis.

In this work, some of the popular classification algorithms for outlier detection in multi-dimensional cancer data set are evaluated and so, proposed dimensionality reduction and feature selection methods for measuring the training performance and accuracy testing issues in classification based outlier detection method.

2. MODELING CLASSIFICATION BASED OUTLIER DETECTION SYSTEM

2.1 Outlier Detection methods

2.1.1 Supervised, Semi-Supervised and Unsupervised Method

Supervised approach develops a predicative model for normal as well as outlier classes and new model instance is compared against it. But in semi-supervised outlier detection mode, training data set is available for normal data set. In a unsupervised outlier detection mode, the training data is not available, the data instances which are frequent are treated as normal and others are outliers.

2.1.2 Statistical Methods, Proximity-Based methods

Statistical methods can be used on the assumption of the data normality. Data objects that do not follow these methods are treated as outliers. In the proximity based method, the closeness of outlier object to its nearest objects significantly different from the closeness of the object to the most of other objects in the data set. So the varies tests like Grubb's test, Rosner's test and Dixon's test for outlier detection can be used on the normally distributed data.

2.1.3 Classification based Methods

If a training data set with class labels is available then outlier detection is treated as a classification problem. Once the classification model is constructed, the outlier detection process is very fast one. It only needs to compare the objects against the model learned from the training data.

2.2 Dimensionality Reduction Algorithm

The number of variables that are used to describe an object is the dimensionality of that object. The search for features in deep relation between variables is known as data exploration It is necessary to reduce the number of dimension using dimensionality reduction methods

The dimensionality reduction is the process of a search for a small set of features to describe a large set of observed dimensions. Since the small set is much faster than large one, it decreases the computational processing time

2.2.1 Reasons for Dimensionality Reduction

- Some features of cancer data may be irrelevant
- To visualize high dimensional cancer data on a low dimensional space data
- "Intrinsic" dimensionality of the cancer data may be smaller than the number of features
- In some cases, data analysis such as regression or classification can be more accurately in reduced space than in the original space.

2.3 Model of Classification Based Outlier Detection with Feature Selection

In this work, we have evaluated three feature selection based dimensionality reduction techniques, 1.Chi Square, 2.Information Gain and 3.Gini Index

2.3.1 Chi Square

The Chi-square measures the degree of dependence of feature of class The feature and the class are considered dependent if chi-square is greater than the critical value determined by degrees of freedom.. Such features are selected.

$$\chi^2(f) = \sum_{v \in V} \sum_{i=1}^{m} \frac{(A_i(f=v) - E_i(f=v))^2}{E_i(f=v)}$$

where *Ai(f = v)* is the number of instances in class *ci* with *f = v* and *Ei(f = v)* is the expected value of *Ai(f = v*), calculated as *P(f = v)P(ci)N*.

2.3.2 Information Gain

It is a measure of how an attribute is for predicting the class of each of the training data. Information gain is a measure of reduction in uncertainty once the value of an attribute is known.

The information gain of a feature *f* is defined as

$$\begin{aligned} G(f) \;&=\; -\sum_{i=1}^{m} P(c_i) \log P(c_i) \\ &+ \sum_{v \in V} \sum_{i=1}^{m} P(f=v) P(c_i|f=v) \log P(c_i|f=v) \end{aligned} ,$$

where $\{c_i\}_{i=1}^{m}$ denotes the set of classes, *v*∈V is the set of possible values for feature *f* [5].

2.3.3 Gini Index

The Gini coefficient or Index is a measure of resource inequality in a population developed by the Italian statistician Corrado Gini and published in his 1912 paper "Variabilità e mutabilità". It can be used to measure any form of uneven distribution. Index varies form 0 to 1, zero means no inequality (no uncertainty) and 1 means maximum possible inequality (maximum uncertainty) The Gini coefficient is often calculated with the more practical Brown Formula

$$G = |1 - \sum_{k=1}^{n} (X_k - X_{k-1})(Y_k + Y_{k-1})|$$

Where, Gini coefficient and X_k: :cumulated portion of one variable for k= 0 to 1 with $X_0 = 0$, $X_n = 1$.Y_k: cumulated proportion of the target variable, for k = 0,...,n, with $Y_0 = 0$, $Y_n = 1$

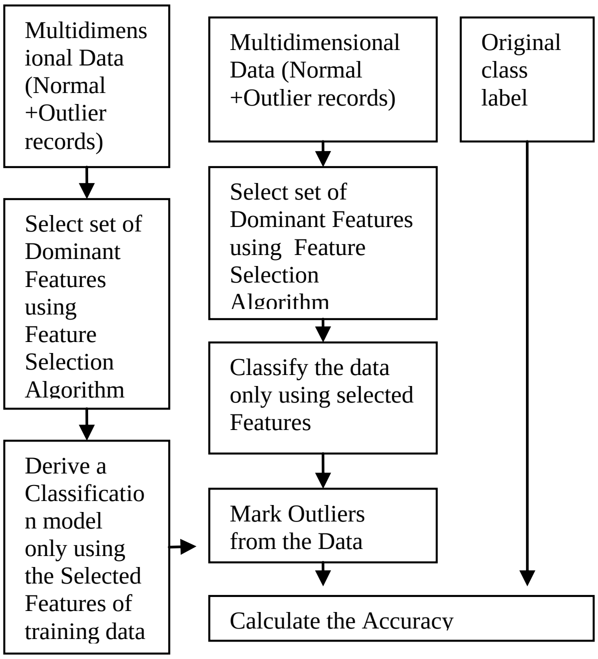
Multidimensional Data (Normal +Outlier records)
Select set of Dominant Features using Feature Selection Algorithm
Derive a Classification model only using the Selected Features of training data
Multidimensional Data (Normal +Outlier records)
Select set of Dominant Features using Feature Selection Algorithm
Classify the data only using selected Features
Mark Outliers from the Data
Original class label
Calculate the Accuracy

Fig 1: Classification Based Outlier System

2.4 The Used Classification Algorithms

2.4.1 C4.5 Classifier

C4.5 is a tree pruning algorithm in Decision tree-based approach and creates a tree model by using values of only one attribute at a time.

2.4.2 Decision Table Classifier

Decision table is a hierarchical breakdown of the data, with two attributes at each level of the hierarchy and the most important attributes for classifying the data.

2.4.3 K-Nearest Neighbors Classifier

The "nearest" measurement refers to the Euclidean distance between two instances. For example, the Euclidean distance between t_i and t_j is

$D(t_i, t_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$ where p is number of attributes.

3. THE EVALUATION

The performance of the classification algorithms under evaluation were tested with the breast cancer database called "Wisconsin Breast Cancer Database"

3.1 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 Table1 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 being the most analistic. The class is distributed with 444 (65.0%) benign samples and 239 (35.0%) malignant samples (Tan et al 2003).

Summary of the WBCD dataset: Table 1

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%)

3.2 Metrics Used For Evaluation

In order to evaluate the performance of algorithms under consideration with a suitable metric, we used Rand Index and Run Time as two measures

3.2.1 Total Run Time

The total run time is the sum of time required for training and testing. Here we just only mention the time taken for training and compare the CPU times only

3.3 The metrics and Validation Method Used for Performance Evaluation

Performance of Classifiers depends on the properties of the data to be classified and is measured with metrics Sensitivity, Specificity, Accuracy, Precision, F_Score, and Error Rate.

3.3.1 Confusion Matrix

A confusion matrix is used to reveal the type of classification errors a classifier makes. The advantage of using this matrix is that it not only tells how many got misclassified but also what misclassification occurred.

Figure 1: A confusion matrix.

Predicated class		
+	-	Actual Class
TP	FN	+
FP	TN	-

True Positives (TP)- positive tuples correctly classified. False Negative (FN)- positive tuples incorrectly labeled as negative. False Positive (FP) – negative tuples incorrectly labeled as positive. True Negative (TN) – negative tuples correctly classified

3.3.2 The Metrics

Sensitivity =TP/ (TP + FN) , known as true positive rate.

Specificity =TN/(TN + FP) , known as true negative rate.

Accuracy = (TP + TN)/ (TP + FP + TN + FN)

Positive Predictive Value =Precision = TP/(TP +FP)

F_Score = 2 *(precision * Recall)/(Precision + Recall)

Error Rate = (FP +FN)/ (TP + FP + TN + FN)

Time: The *cpu time also considered as the metric to evaluated the speed of the algorithm*

3.4 Validation Methods

The K-fold cross validation method is used for validating the performance with respect to different metrics.

3.4.1 K-fold Cross-Validation

In K-fold cross validation, the available data is randomly divided into k disjoint subsets of approximately equal size. One of the subsets is then used as the test set and the remaining k-1 sets are used for building the classifier. The accuracy is estimated using the test set. It is repeated K times so that each subset is used as a test subset once.

In the first iteration, to obtain first model, subsets $s_2, \ldots, s_k$ collectively serve as the training set l, which is tested on s_1; the second iteration is trained in subsets $s_1, s_3, \ldots, s_k$ and tested on s_2; and so no.

3.5 About the Implementation

With Matlab version 7.4.0 (R2007a),we have developed the proposed outlier detection software . To implement this outlier detection software, we used the Mex and Java interface of matlab . We used the standard weaka implementation of the classification algorithms and only passed the default parameters while invoking the classifier algorithms.

4. THE RESULTS AND DISCSSION

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 nalignant nature of that case

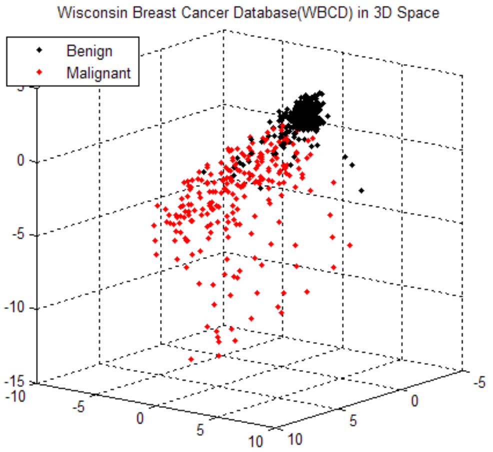

Figure 1: The Plot of WBDC Data Clearly Showing the Benign Cluster and Malignant Outliers

The flowing 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 1 - The Performance of Classification Algorithms with different Feature Selection Algorithms used for Outlier Detection

Algorithm	Precision %	F-Score %	Sensitivity %	Specificity %	Accuracy %	Error Rate %
C4.5 Classifier	**96.18**	**95.82**	**95.58**	**92.60**	**94.53**	**5.47**
Chi-square + C4.5 Classifier	96.20	95.75	95.40	92.92	94.53	5.47
Information Gain + C4.5 Classifier	96.64	96.34	96.12	93.65	95.25	4.75
Gini Index + C4.5	96.18	96.59	97.10	92.72	95.53	4.47
Decision Table	**96.12**	**96.19**	**96.35**	**92.51**	**95.03**	**4.97**
Chi-square + Decision Table	96.20	95.71	95.34	92.75	94.43	5.57
Information Gain + Decision Table	96.33	96.07	95.92	92.99	94.91	5.09
Gini Index + Decision Table	96.14	96.42	96.77	92.36	95.28	4.72
k-Neighbourhood	**96.07**	**96.66**	**97.31**	**92.23**	**95.57**	**4.43**
Chi-square + k-Neighbourhood	96.44	95.75	95.20	93.34	94.56	5.44
Information Gain + k-Neighbourhood	96.49	96.16	95.94	93.27	95.00	5.00
Gini Index + k-Neighbourhood	96.02	96.55	97.16	92.28	95.47	4.53

4.1 The Effect of Feature Selection Algorithms

In this experiments, a test is conducted for outlier detection performance with different number of features. But found that if the number of features less than that of the original number of features, then there was no improvement in performance.

For example, by examining the results of the above table and can understand it very clearly. In the above table, only the first 5 selected features are used for classification.

4.2 The performance without Feature Selection

The following bar charts are showing the performance of the algorithms without using any feature selection algorithm. 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 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 k neighborhood classifier performed well. It means, k neighborhood classifier is capable of marking outliers correctly better than other two algorithms.

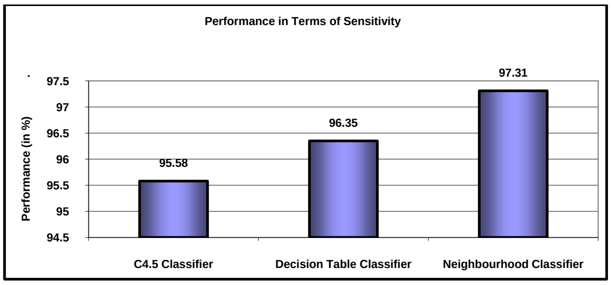

Figure 2: The Sensitivity/Recall Chart

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 k neighborhood classifier performed well. It means, k neighborhood classifier is capable of marking normal as well as the outliers correctly better than other two algorithms.

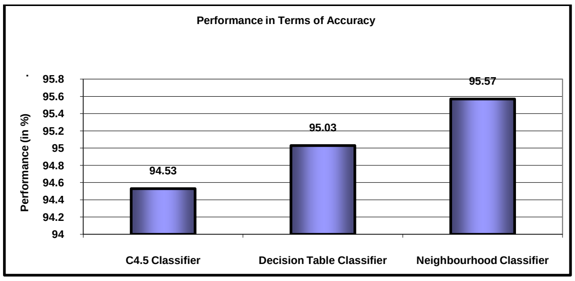

Figure 3: The Accuracy Chart

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 k neighborhood classifier performed well. It means, k neighborhood classifier is capable of marking normal as well as the outliers correctly better than other two algorithms.

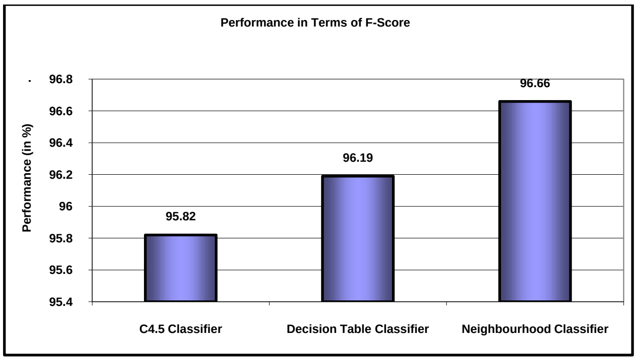

Figure 4: The F-Score Chart

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 k neighborhood classifier performed well. It means, the lower value of error rate signifies that the k neighborhood classifier is making less error while identifying the malignant as well as outlier data.

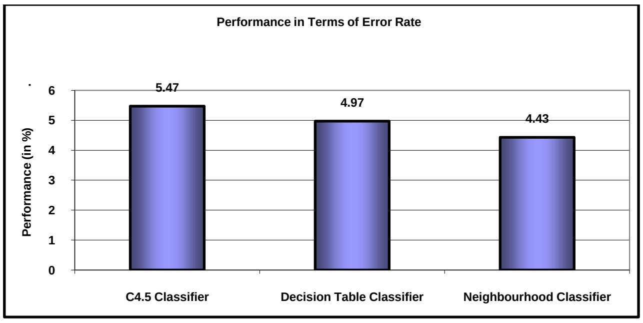

Figure 5: The Error Rate Chart

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 k neighborhood classifier performed poor. It doesn't mean that it is performing poor – it means, it is performing good in identifying the outliers by missing some normal records.

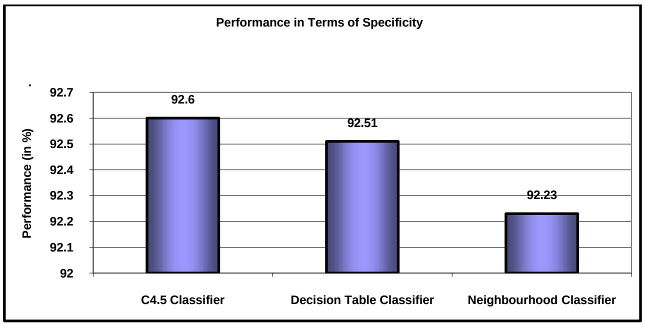

Figure 6: The Specificity Chart

The following bar chart shows the performance of the algorithm in terms of precision.

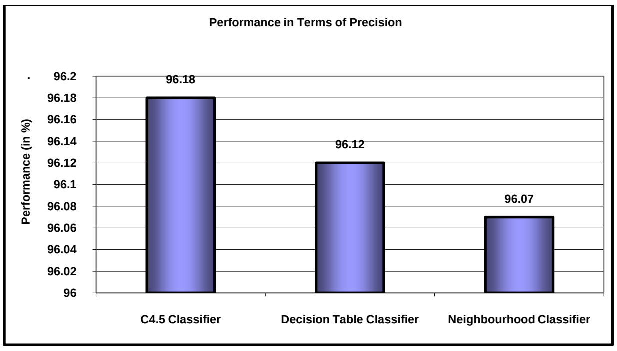

Figure 7: The Precision Chart

The following bar chart shows the performance of the algorithm in terms of cpu time. In this case, cup time measures the time taken for the classifier to classify the entire dataset. It is measured in seconds. As shown in the graph, with respect to cpu time, the k neighborhood classifier performed very good.

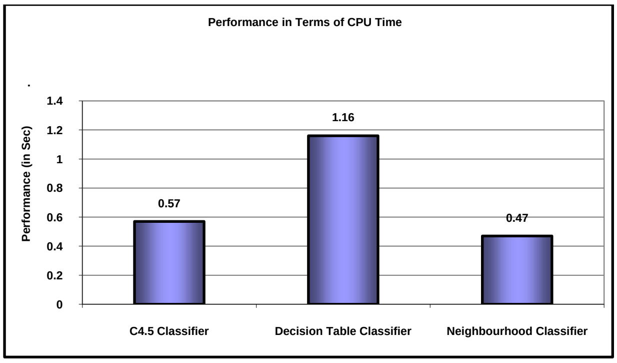

Figure 8: The CPU Time Chart

5. CONCLUSION

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

In this work, the performance of outlier detection using feature selection algorithms are evaluated and the results clearly shows that the impact of feature selection algorithm on the cancer dataset is very low and does not improve the overall classification performance.

In this case, three classifier algorithms without any feature selection were evaluated and found that the k-neighborhood algorithm was capable of identifying and classifying the outliers better than the other two compared algorithm in terms of accuracy, f-score, Sensitivity/Recall, error rate. Further, the cpu time of the k-neighborhood algorithm also minimum. So found that k-neighborhood algorithm performed very well for detecting outliers in cancer data.

The excellent outlier detection performance of the k-neighborhood algorithm and other algorithms shows that a classification algorithm will be capable of accurately identifying/ classifying the multidimensional outlier data in its subspace. Previous works on general clusters and classification of multi-dimensional data shows the ways to get better performance while dealing with sub-spaces using feature or dimensionality reduction techniques. In future works, we may address the ways to improve the performance of the outlier detection using feature selection and dimensionality reduction techniques.

Further, this work may address the possibility of improving k-neighborhood algorithm using a good distance metric or good neighborhood relationship function. Future works may address these issues and improve the performance of the outlier detection in cancer data.

6. ACKNOWLEDGMENTS

This study can be considered as a work under the guidance of Holy Spirit. I would like to thank all those who participated in discussions 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 at Govt. College Nedumangad, Kerala, India

7. REFERENCES

[1] DASH, M., & LIU, H (1997) Feature selection for classification. Intelligent Data Analysis, 131- 156.

[2] R Kohavi, G John, Wrappers for feature subset selection. Artif Intell J Spec Issue Relevance97(1–2), 273–324 (1997)

[3] 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

[4] 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.

[5] YU, L. & LIU, H. (2003) Feature Selection for High Dimensional Data: A Fast Correlation-Based Filter Solution. Proceedings of the Twentieth International Conference on Machine Leaning (ICML-03). Washington, D.C .

[6] Y. Lu and J. Han, "Cancer classification using gene expression data," Information Systems, vol. 28, pp. 243-268, 2003.

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

[8] Huan Liu, Lei Yu (2005) Toward Integrating Feature Selection Algorithms for Classification and Clustering, IEEE Transactions On Knowledge and Data Engineering, VOL. 17, NO. 4, April 2005

[9] Ella Bingham, Aristides Gionis, Niina Haiminen, Heli Hiisil¨a, Heikki Mannila, Evimaria Terzi, ”Segmentation and dimensionality reduction”. 2006 SIAM Conference on Data Mining, pp. 372-383.

[10] Guyon, S Gunn, M Nikravesh, L Zadeh, Feature Extraction, Foundations and Applications (Springer, Berlin, 2006)

[11] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," Bioinformatics, vol. 23, pp. 2507-2517, 2007.

[12] 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

[13] Y. Song, J. Huang, D. Zhou, H. Zha, and C. Giles, "Iknn: Informative k-nearest neighbor pattern classification," Knowledge Discovery in Databases: PKDD 2007, pp. 248- 264, 2007

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

[15] 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

[16] 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.

[17] Binita Kumari (2012) "Feature Subset Selection in large Dimensionality using Correlation based GA-SVM" International Journal of Computer Applications Vol.45. No.6. pp 5-8 May 2012.

[18] 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.

[19] L. Rutkowski, L. Pietruczuk, P. Duda, and M. Jaworski, "Decision Trees for Mining Data Streams Based on the McDiarmid's Bound," IEEE Trans. Knowledge and Data Eng., vol. 25, no. 6, pp. 1272- 1279, 2013.

[20] Ammu P.K and Preeja V (2013) " Review on Feature Selection Techniques of DNA Microarray Data" International Journal of Computer Applications ,Vol. 61 No. 12 , pp 39-44 January 2013.

[21] E.T. Venkatesh and A. Kalyana Saravanan (2013)" New Scheme to identify Intrusion Outliers by Machine learning Technique " International Journal of Computer Applications , Vol. 84. No.13. pp 13 -16 Dec. 2013

[22] T.Ediwin Prabakaran and S.Venkata Lakshmi (2014) " Application of K-Nearest Neighbour Classification Method for Intrusion Detection in Network Data" International Journal of Computer Application ,Vol.97- No.7 , pp 34- 37 ,July 2014

[23] Kurian M.J and Gladston Raj S,(2015) " Outlier Detection in Multidimensional Cancer Data Using Classification Based approach " International Journal of Applied Engineering Research ,Vol.10, No.79,pp. 342-348 , 2015.