# An Analysis on the Performance of a K-Nearest-Neighbor Classification Based Outlier Detection System using Feature Selection and Dimensionality Reduction Techniques

Kurian M. J [1] and Dr. Gladston Raj S [2*]
[1]Research Scholar, Research and Development Centre, Bharathiar University, Coimbatore.
[2]Head of Department of CS, Govt. College, Nedumangadu , Trivandrum, Kerala, India.
kurianmj@yahoo.com, gladston@rediffmail.com

**ABSTRACT -** The general idea of classification-based outlier detection method is to train a classification model that can distinguish normal data from outliers. In the previous work, we have implemented and evaluated three classification 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. In this work, the performance of outlier detection is evaluated using dimensionality reduction algorithms. The results clearly shows that the impact of dimensionality reduction algorithm on the cancer dataset is significantly improved the overall classification performance to a considerable level.

## 1. INTRODUCTION

In Data Mining, Outliers are meaningful input signals, which represent the characteristic of the object. This work aims to study the performance of classification algorithms of data mining for outlier detection using dimensionality reductions. Before the elimination of points, one should study why these points appeared and whether it is likely to continue to appear.

### *1.1. 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, has to find the subset of the original attributes.

### *1.2. Problem Specification*

The identification of outlier can be viewed as classification problem which can lead to the discovery of unexpected knowledge in the medical field. The general idea is to train a classification model that can distinguish normal data from outliers [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 [25, 26]. Many solutions like eliminate variables using factor analysis and principle component methods were suggested to improve the efficiency of the algorithm.

This method proposes to use dimensionality reduction and feature selection algorithms to overcome the training performance and testing accuracy issues in the classification based outlier detection approaches.

## 2. MODELING CLASSIFICATION BASED OUTLIER DETECTION SYSTEM

The popular methods of outlier detection are supervised, semi supervised, unsupervised proximity-based. With the limitation of the Grubb’s test and the Rosener test, there is a need for more sophisticated and speedy method such as classification based outlier detection, which heavily depends of the quality and availability of training data set.

### *2.1. The Used Classification Algorithm*

*K-Nearest Neighbors Classifier*

K-Nearest Neighbors is a method to assign the input instance to the class with the majority of K-Nearest Neighbors by considering the Euclidean distances between two instances

### 2.2. *Feature Selection Technique*

1. Chi Square, 2.Information Gain and 3.Gini Index are used.

### *2.3. Dimensionality Reduction Algorithm*

The famous Algorithms for dimensionality reduction such as Principal Component Analysis, Kernel PCA and LPP (Locality preserving Projection) can be used.

## 3. THE EVALUATION

The performance of the classification algorithms under evaluation were tested with "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 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).

### *3.2. Metrics Used For Evaluation*

Random index and Run time are two measures for evaluating the algorithm under consideration. The total run time is the sum of the times taken for learning and testing and this model concentrate on the time taken for training which is higher than the time taken for testing.

### *3.2.1. Total Run Time*

We calculated the total run time as the sum of time required for training and the time required for testing. Here we compare the CPU times only. Since the time taken for training is the very much higher and the time required for testing the network with same number of records is very in significant, in the following table we just only mention the time taken for training.

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%) |

### *3.3. The Metrics and Validation Method Used for Performance Evaluation*

### *3.3.1. Confusion Matrix*

A Confusion matrix shows the type of classification error a classifier produced. The advantage of using this matrix is that it not only tells us how many got misclassified but also what misclassifications occurred.

| Predicted Class | | |
|---|---|---|
| Positives | Negatives | Actual Class |
| a | b | Positives |
| c | d | Negatives |

Figure 1: A confusion matrix.

The breakdown of a confusion matrix is as follows:

- a is the number of positive examples correctly classified (True Positives –TP)
- b is the number of positive examples misclassified as negative(False Negatives -FN)
- c is the number of negative examples misclassified as positive(False Positives –FP)
- d is the number of negative examples correctly classified(True Negatives –TN).

The performance of the algorithm is measured with metrics Sensitivity, Specificity, Accuracy , Precision ,F-score , Error rate and CPU time.

Sensitivity = TP/ (TP +FN)

Specificity = TN/ (TN +FP)

Error Rate = (T – C) / T, The test data has total of T objects and C of the T objects are correctly classified.

### *3.3.2. Validation Methods*

This work uses the K-fold cross validation method, in which the data set is 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 test set is then used to estimate the accuracy. This is done repeatedly k times so that each subset is used as a test subset once. One of the K-subset is used as a test set and remain K-1 subsets are put together to form a training set. Thus every data point gets a chance to be in a test set exactly once.

In the first iteration , to obtain the first model, subset $x_1,x_2,\ldots,x_k$, collectively serve as a training set , which is tested on $x_{1\,:}$ the second iteration is trained in subsets $x_1,x_3,\ldots,x_k$ and tested on $x_{2:}$ and so on.

*About the Implementation*

The proposed outlier detection software is developed with Matlab version 7.4.0 (R2007a) and uses some of the features of Weaka with Matlab interface code. The Mex and Java interface of matlab is used to implement this outlier detection software. Here, the standard weaka implementation of the classification algorithms is used and only passed the default parameters while invoking the classifier algorithms.

## 4. RESULTS AND DISCUSSION

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.

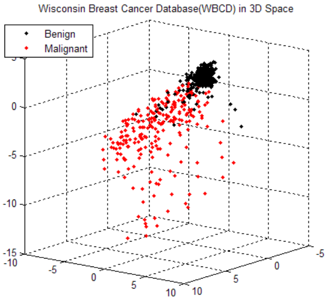

Figure 2: The Plot of WBDC Data Clearly Showing the Benign Cluster and Malignant Outliers

The following table lists the performance of the algorithm with respect to different metrics. In fact, each value is an 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.

With the concept of the number dimension as five, first five features, the above table shows an improvement in the performance of outlier detection with different number of dimensionality reduction algorithms.

### *4.1. The Effect of Feature Selection Algorithms*

This table reveals that there is no improvement in accuracy or reduction in error rate. So it is clear that there is no performance improvement in result by considering less number of features than original one.

Table 2: The Performance of Outlier Detection with different Feature Dimensionality Reduction Algorithms and Classification Algorithms

| Algorithm | Precision % | F-Score % | Sensitivity % | Specificity % | Accuracy % | Error Rate % |
|---|---|---|---|---|---|---|
| **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 |
| PCA + k-Neighbourhood | 96.65 | 96.83 | 97.07 | 93.69 | 95.85 | 4.15 |
| kPCA + k-Neighbourhood | 95.25 | 94.38 | 93.66 | 91.08 | 92.75 | 7.25 |
| LPP + k-Neighbourhood | 96.89 | 97.20 | 97.57 | 94.23 | 96.37 | 3.63 |

The following table shows the comparison of previous results with this work.

Table 3: The Comparison with Recent Works

| Sl No | Classifiers | Classification accuracy |
|---|---|---|
| 1 | CART with feature selection (Chi- square)[11] | 94.56% |
| 2 | C4.5 [12] | 94.74% |
| 3 | Hybrid Approach[14] | 95.96% |
| 4 | Neuron-Fuzzy[16] | 95.06% |
| 5 | Supervised Fuzzy Clustering [17] | 95.57% |
| 6 | Proposed PCA* + k-Neighborhood | 95.85 |

* The First Five Principal Components were used for classification

### *4.2. The Effect of Dimensionality Reduction Algorithms*

In this case, accuracy measures the capability of the algorithms to correctly identify the normal as well as outliers in the data. It means, Proposed PCA+ k-neighborhood and proposed LPP+ k-neighborhood classifiers are capable of marking normal as well as the outliers correctly.

This table 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 table, with respect to f-score, proposed PCA+ k-neighborhood and proposed LPP+ k-neighborhood classifiers performed well.

In this case, error rate measures how much the algorithm wrongly identifies both the normal as well as outliers in the data. The lower value of error rate of proposed PCA+ k-neighborhood and proposed LPP+ k-neighborhood classifiers reveals that classifiers are making less error while identifying the malignant as well as outlier data.

## 5. CONCLUSION

This work is implemented with the classification based outlier detection software under Matlab and evaluated its performance using different metrics and thus arrived at significant and comparable results. The table and graphs in the previous section shows the overall results. In this work, the performance of outlier detection algorithm K-Neighborhood with dimensionality reduction algorithms is evaluated and the result clearly shows that the impact of dimensionality reduction algorithm on the cancer dataset improves the overall classification performance. Also it is clear that there is no significant effect on the result by using the feature selection algorithms. Future works may address these issues and improve the performance of the outlier detection in cancer data with other algorithms.

## ACKNOWLEDGMENT

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] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

[17] 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.

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

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

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

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

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

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

[24] 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.

[25] 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..

[26] 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.

## 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, the Principal Investigator of the Minor Research Project "Outlier Detection In Multidimensional Data", 2010, funded by Universities Grant Commission, India and Convener of the UGC sponsored National Conference in “Recent Trends in Data Mining” organized by B.P.C. College and Computer Society of 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, 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.