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 an 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 \( A_i(f = v) \) is the number of instances in class \( ci \) with \( f = v \)
and \( E_i(f = v) \) is the expected value of \( A_i(f = v) \), 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

\[ \gamma(f) = \sum_{v \in V} \sum_{i=1}^{m} P(v) \log \frac{P(v)}{P(v|f = v)} \]

\[ = \sum_{v \in V} \sum_{i=1}^{m} P(f = v)P(v|f = v) \log \frac{P(v|f = v)}{P(v)} \]

where \( |c_i| \) denotes the set of classes, \( v \in V \) 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 from 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

\[ G = \left[ 1 - \sum_{k=1}^{n} \left( X_k - X_{k-1}\right) \left( Y_k + Y_{k-1}\right) \right] \]

Where, Gini coefficient and \( X_k \): cumulated portion of one
variable for \( k = 0 \) to \( 1 \) with \( X_0 = 0 \), \( X_1 = 1 \). \( Y_k \): cumulated
proportion of the target variable, for \( k = 0,...,n \), with \( Y_0 = 0 \),
\( Y_n = 1 \)

![Fig 1: Classification Based Outlier System](image)

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

\[ D(t_i, t_j) = \sqrt{\sum_{k=1}^{p} (x_{i,k} - x_{j,k})^2} \]

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 dataset (WBCD): The WBCD dataset 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 being the most anastatic. The class is distributed with 444 (65.0%) benign samples and 239 (35.0%) malignant samples (Tan et al 2003).

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Possible values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clump thickness</td>
<td>Integer 1–10</td>
</tr>
<tr>
<td>Uniformity of cell size</td>
<td>Integer 1–10</td>
</tr>
<tr>
<td>Uniformity of cell shape</td>
<td>Integer 1–10</td>
</tr>
<tr>
<td>Marginal adhesion</td>
<td>Integer 1–10</td>
</tr>
<tr>
<td>Single epithelial cell size</td>
<td>Integer 1–10</td>
</tr>
<tr>
<td>Bare nuclei</td>
<td>Integer 1–10</td>
</tr>
<tr>
<td>Bland chromatin</td>
<td>Integer 1–10</td>
</tr>
<tr>
<td>Normal nuclei</td>
<td>Integer 1–10</td>
</tr>
<tr>
<td>Mitoses</td>
<td>Integer 1–10</td>
</tr>
<tr>
<td>Class</td>
<td>Benign (65.5%), Malignant (34.5%)</td>
</tr>
</tbody>
</table>

Summary of the WBCD dataset: Table 1

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.

<table>
<thead>
<tr>
<th>Predicated class</th>
<th>Actual Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

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, …, s_k collectively serve as the training set s_1, which is tested on s_1; the second iteration is trained in subsets s_1, s_3, …, s_k and tested on s_2; and so no.

3.5 About the Implementation

With Matlab version 7.4.0 (R2007a), we 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 weka implementation of the classification algorithms and only passed the default parameters while invoking the classifier algorithms.

4. THE 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 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

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

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.
The following bar chart shows the performance of the algorithm in terms of cpu time. In this case, cpu 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.

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 show 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


[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.


