An Automated Diagnosis Of Breast Cancer Using Farthest First Clustering And Decision Tree J48 Classifier

R. Delshi Howsalya Devi and P. Deepika

Assistant Professor, K.L.N College of Engineering, Madurai, India.
ME Scholar K.L.N College of Engineering, Madurai, India.

Received 27 May 2016; Accepted 28 June 2016; Available 12 July 2016

ABSTRACT
Breast cancer is one of the most widespread and deadly cancer for women. Early diagnosis and treatment of breast cancer can enhance the outcome of the patients. Due to the difficulties of outlier and skewed data, the prediction of breast cancer survey has presented many challenges in the field of data mining and pattern recognition. To solve these troubles, we have proposed an automated breast cancer diagnosis using data mining algorithms. This approach comprises two main steps: (1) utilization of an outlier filtering approach based on Farthest Clustering to identify and eliminate outlier instances; and (2) Classify the benign or malignant cancer using decision tree J48 classifier. To test the efficacy of the proposed classification model we used the Wisconsin Breast Cancer Dataset (WBCD). In order to assess the capability and effectiveness of the proposed approach, several measurement methods including basic performance (e.g., sensitivity, accuracy, and specificity), AUC and F-measure were utilized. The obtained classification accuracy of 98.4% is a very optimistic result compared to the existing works for the same data set. The result also shows that our approach works well with the breast cancer database and can be a good choice to the well-known machine learning methods.

KEYWORDS: Breast Cancer, Data mining, Farthest First Clustering, Decision Tree, J48, Outlier.

INTRODUCTION
Breast cancer (BC) is the most common invasive cancer in females worldwide. BC is the second leading disease, next to lung cancer, which increases the mortality rate in women [1]. It develops from breast cells, primarily in the milk ducts (ductal carcinoma) or glands (lobular carcinoma) which begin with the formation of a small tumor or a mass [2]. A mass with a smooth, well defined border is non-cancerous (benign). Mass with an irregular border or with speculations may be cancer-ous (malignant) [48]. Even though the causes of BC are not fully and thoroughly understood, every woman needs to be aware of her own chances of developing BC in order to be proactive about the risk reduction strategies and for better management of the disease. Independent studies have identified a number of factors that either increase or decrease the chances of developing breast cancer [3–15].

Data mining has become a popular technology in current research and for medical domain applications [17]. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. The main goal of data mining is to discover new patterns for the users and to interpret the data patterns to provide meaningful and useful information for the users. Data mining is applied to find useful patterns to help in the important tasks of medical diagnosis and treatment. An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.

To Cite This Article: R. Delshi Howsalya Devi and P. Deepika., An Automated Diagnosis Of Breast Cancer Using Farthest First Clustering And Decision Tree J48 Classifier. Advances in Natural and Applied Sciences. 10(10); Pages: 161-166
Unsupervised learning approach is employed to this model. Usually, the result of unsupervised learning is a new explanation or representation of the observation data, which will then lead to improved future responses or decisions. In this paper proposed a automatic diagnosis of breast cancer either benign or malignant using data mining techniques of farthest first clustering and j48 decision tree classification.

The rest of the paper is organized as follows: The study of related researches on the breast cancer is presented in section II. Section III gives brief explanation for the methods used in the outlier detection i.e. farthest first clustering and j48 classifier. The proposed system is presented in section IV. The experimental results and data set are described in section V. Finally, section VI gives the conclusion of the paper.

II. Related Work:

Li, Peng, and Liu (2013) Discussed the quasi formal kernel common locality discriminant analysis for dimensionality reduction and reported a classification accuracy of 97.26%. Zheng, Yoon, and Lam (2014) achieved a classification accuracy of 97.38% with K-means algorithm and support vector machine based model. Ozs en and Ceylan (2014) presented the performance of artificial immune system as a data reduction algorithm. In order to estimate the data reduction performance of artificial immune system, it is evaluated to the fuzzy c-means clustering algorithm. Both data reduction methods are combined with the artificial neural network classifier to get the classification results. They achieved a 97.80% classification accuracy for artificial immune system and artificial neural network combination, whereas a c fuzzy c-means clustering and artificial neural network combination they obtained classification accuracy of 90.04%. Chen (2014) proposed a combined model for breast cancer diagnosis that can work in the absence of labeled training data. Hence, this work explains the feature selection methods in unsupervised learning models. The model combines clustering and feature selection. Their work indicates that selecting a subset of relevant features instead of using all the features in the original data set can improve the interpretability of clustering results. Nguyen, Khosravi, Creighton, and Nahavandi (2015a) obtained a classification accuracy of 97.88% and presented a medical classification model which integrates wavelet transformation and interval type-2 fuzzy logic system. These mechanism are combined together in order increase the dimensionality and uncertainty properly. Interval type-2 fuzzy logic system consists of fuzzy c-means clustering based unsupervised learning and genetic algorithm based parameter tuning. These mechanisms have high computational costs and wavelet transform functions for reducing these computational costs. Nguyen, Khosravi, Creighton, and Nahavandi (2015b) reached a classification accuracy of 97.40% with a medical classification model based on fuzzy standard additive model and genetic algorithm. In this scheme, rule initialization is held by the adaptive vector quantization clustering. A genetic algorithm is used for rule optimization and gradient descent algorithm for parameter tuning. Lee, Anaraki, Ahn, and An (2015) presented a classification approach based on fuzzy-rough feature selection and multi-tree genetic programming for intention pattern recognition using brain signal. The literature survey on the breast cancer diagnosis of this paper differs from the other studies in terms of a number of points. First, the propose classification approach utilizes the fuzzy-rough instance selection method in order to omit useless or erroneous instances based on fuzzy-rough set based concept. Secondly, a recent search approach, i.e. the re-ranking search method is integrated with consistency-based subset evaluation for sinking the number of wrapper estimations in feature selection. Finally, an optimal training set is achieved by fuzzy-rough instance selection and consistency-based feature selection and this set is utilized to build a classification approach based on fuzzy-rough nearest neighbor classifier. Rodger (2014a) proposed a statistical model based on fuzzy nearest neighbor, regression and fuzzy logic for improving energy costs savings in buildings. Rodger (2014b) presented a fuzzy feasibility Bayesian probabilistic estimation model for a supply chain. Pena-Reyes and Sipper (1999) proposed an approach that combined the fuzzy systems and evolutionary algorithms and achieved a classification accuracy of 97.80%. Chou, Lee, Shao, and Chen (2004) presented the artificial neural networks and the multivariate adaptive regression splines and achieved 98.25% classification accuracy. Karabatak and Ince (2009) obtained a classification accuracy of 97.4% with an association rule for dimension reduction and the neural network for performing classification. Shiv Shakti Shrivastava (2013) focused on classification of data mining techniques on breast cancer data with using data mining software. A large amount of medical records are stored in databases. Mining is a relatively new field of researcher whose major objective is to acquire knowledge from large amounts of data. Decision trees are powerful classification algorithms that are become more popular with the growth of data mining in the field of information systems. Deepshree A. Vadeyar (2014) focused on Website can be easily design but to efficient user navigation is not a easy task. The user behavior is keep change and developer view is quite different from what user wants. The way is reorganization of website structure can improve navigation. For reorganization here proposed strategy is farthest first traversal clustering algorithm to perform clustering on two numeric parameters and for finding frequent traversal path of user Apriori algorithm is used. To make reorganization with fewer changes in website structure. E. Venkatesan (2015) focused on compare the four decision tree algorithms in the prediction of the performance accuracy in breast cancer data. Classification is used to classify the elements permitting to the features of the elements through the predefined set of classes. Breast cancer accuracy can be increased by pruning. To reduce the cost of medicine
for patients. To help in improvement clinical studies and analyzing results. To conclude that the classification results of all the four algorithms J48 can show the better performance.

**METHODS AND MATERIALS**

In this section farthest first clustering is used to cluster the data into number of clusters. A small cluster is declared as outliers. The remaining outliers are detected by using LOF. If the value of LOF is greater than threshold means, the point is declared as outliers. Finally the data set is classified either benign or malignant using decision tree J48 algorithm.

**A. Classification Technique:**
Building accurate and efficient classifiers for large databases is one of the essential tasks of data mining and machine learning research. Usually, classification is a preliminary data analysis step for examining a set of cases to see if they can be grouped based on “similarity” to each other. The ultimate reason for doing classification is to increase understanding of the domain or to improve predictions compared to unclassified data. Building effective classification systems is one of the central tasks of data mining. Given a classification and a partial observation, one can always use the classification to make a statistical estimate of the unobserved attribute values and as the departure point for constructing new models, based on user’s domain knowledge.

**B. Decision Tree Classifier: J48:**
Quinlans C4.5 algorithm implements J48 to generate a trimmed C4.5 decision tree. The every aspect of the data is to split into minor subsets to base on a decision[4]. J48 examine the normalized information gain that actually the outcomes the splitting the data by choosing an attribute. In order to make the decision, the attribute utmost standardized information gain is used. The minor subsets are returned by the algorithm. The splitting methods stop if a subset belongs to the same class in all the instances. J48 constructs a decision node using the expected values of the class. J48 decision tree can handle specific characteristics, lost or missing attribute values of the data and differing attribute costs. Here precision can be increased by pruning.

The algorithm describes

**Step 1:** The leaf is labelled with the same class if the instances belong to the same class.

**Step 2:** For every attribute, the potential information will be calculated and the gain in information will be taken from the test on the attribute.

**Step 3:** Finally the best attribute will be selected based on the current selection parameter.

**C. Clustering:**
Clustering is the technique of partitioning of data objects into groups based on similar properties. A Cluster is a gathering of data objects having similar properties. The data objects within the same cluster have tough association between them and pathetic association between the data objects of different clusters. Clustering is an unsupervised learning so no training sample is accessible to partition the data. In this paper we focus on farthest first clustering techniques.

**Farthest First Algorithm:**
Farthest first algorithm same procedure as k-means, this also chooses centroids and allocate the objects in cluster[2]. The max distance and initial seeds are value which is at largest distance to the mean of values. Here cluster task is different, at initial cluster we get link with high Session Count, like at cluster-0 more than in cluster-1, and so on.

\[
\text{Min}\{\text{Max dist(Pi,P1),Max dist(Pi,P2)}\ldots\}
\]

In farthest first it takes point P i then chooses next point P i which is at maximum distance. P i is centroid and p1, p2, ......., pn are points or objects of dataset belongs to cluster. Farthest first actually solves problem of k-centre and it is very efficient for large set of data. In farthest first algorithm we are not finding mean for calculating centroid, it takes centroid arbitrary and distance of one centroid from other is maximum and shows cluster assignment using farthest first. When we performed outlier detection for our dataset we get which objects is outlier.

**D. Local Outlier Factor (LOF):**
We used farthest first method to cluster the data, which help us to figure out the dense regions as well the sparse regions. Than we apply the LOF algorithm over the safe region first, although we call it safe region but to fully utilize the approach of LOF over the entire data and find out the most accurate results we still apply LOF over this region because simply using a clustering cannot figure out the local outliers. Later the same LOF algorithm is applied on the sparse region which is most desirable partition to contain outliers.
E. Pruning:
This important step to the result because of the outliers. All data sets contain a little subset of instances that are not well-defined, and differs from the other once on its neighbourhood. After the complete creation of the tree, that must classify all the instances in the training set, it is pruned. This is to reduce classification errors, caused by specialization in the training set; this is done to make the tree more general. When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of over fitting the data. Such methods typically use statistical measures to remove the least reliable branches, generally resulting in faster classification and an improvement in the ability of the tree to correctly classify independent test data.

IV. Proposed System:
In this paper proposed an automated breast cancer diagnosis using farthest first clustering and decision tree j48 classifier. The proposed algorithm for automated breast cancer diagnosis is outlined in Table 1.

Table 1: Proposed algorithm for automated breast cancer diagnosis.

<table>
<thead>
<tr>
<th>Step 1: Construct Clusters:</th>
<th>Cluster the entire dataset into k cluster using farthest first clustering.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 2: Clusters having fewer numbers of points:</td>
<td>If a cluster contains only fewer numbers of points than the required number of outliers, the radius pruning is avoided for that cluster.</td>
</tr>
<tr>
<td>Step 3: Pruning points inside each cluster:</td>
<td>Calculate distance of each point of a cluster from the radius of the cluster. If the distance of a point is less than the radius of a cluster, the point is pruned.</td>
</tr>
<tr>
<td>Step 4: Detecting outlier points:</td>
<td>We introduce the notion of the local outlier factor LOF, which captures exactly this relative degree of isolation. Then calculate LOF for all the points that are left unpruned in all the clusters. If the outlier factor is greater than threshold then it will declare as outlier otherwise it is not a outlier.</td>
</tr>
<tr>
<td>Step 5: Classification</td>
<td>The preprocessed data set is classified either benign or malignant using decision tree j48 classifier.</td>
</tr>
</tbody>
</table>

4. Experimental Results:
V Experimental Results:
A. Data Set:
The Breast cancer data are collected from the UCI machine learning repository. Data set having 699 instances, 2 classes (benign and malignant type of cancer), and 9 integer-valued attributes is shown in Table 2. All experiments described in this paper were performed using Weka machine learning environment[3].The Weka is an ensemble of tools for data clustering and classification, WEKA version 3.7.13 was utilized as a data mining tool to estimate the performance of the breast cancer diagnosis. This is because the WEKA tool offers a well defined framework for experimenters and developers to create and analyses the various clustering and classification techniques.

Table 2: Breast cancer data set

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>StdDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clump Thickness</td>
<td>1</td>
<td>10</td>
<td>4.418</td>
<td>2.816</td>
</tr>
<tr>
<td>Uniformity of Cell Size</td>
<td>1</td>
<td>10</td>
<td>3.134</td>
<td>3.051</td>
</tr>
<tr>
<td>Uniformity of Cell Shape</td>
<td>1</td>
<td>10</td>
<td>3.207</td>
<td>2.972</td>
</tr>
<tr>
<td>Marginal Adhesion</td>
<td>1</td>
<td>10</td>
<td>2.807</td>
<td>2.855</td>
</tr>
<tr>
<td>Single Epithelial Cell Size</td>
<td>1</td>
<td>10</td>
<td>3.216</td>
<td>2.214</td>
</tr>
<tr>
<td>Bare Nuclei</td>
<td>1</td>
<td>10</td>
<td>3.545</td>
<td>3.644</td>
</tr>
<tr>
<td>Bland Chromatin</td>
<td>1</td>
<td>10</td>
<td>3.438</td>
<td>2.438</td>
</tr>
<tr>
<td>Mitoses</td>
<td>1</td>
<td>10</td>
<td>1.589</td>
<td>1.715</td>
</tr>
<tr>
<td>Class</td>
<td>2</td>
<td>4</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

2 for benign, count-458, weight-458
4 for malignant, count-241, weight-241
After removing outlier from the original data set, the resulting data contains 574 instances and 14 attributes. The detailed description of data after removing outlier from the original data set is outlined in Figure 1.

![Data set after removing Outliers](image)

**Fig. 1:** Data set after removing Outliers

<table>
<thead>
<tr>
<th>Table 3: Classification result of breast cancer diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time taken to build model: 0.08 seconds</td>
</tr>
<tr>
<td>--- Stratified cross-validation ---</td>
</tr>
<tr>
<td>Correctly Classified Instances</td>
</tr>
<tr>
<td>Incorrectly Classified Instances</td>
</tr>
<tr>
<td>Kappa statistic</td>
</tr>
<tr>
<td>Mean absolute error</td>
</tr>
<tr>
<td>Root mean squared error</td>
</tr>
<tr>
<td>Relative absolute error</td>
</tr>
<tr>
<td>Root relative squared error</td>
</tr>
<tr>
<td>Coverage of cases (0.95 level)</td>
</tr>
<tr>
<td>Mean rel. region size (0.95 level)</td>
</tr>
<tr>
<td>Total Number of Instances</td>
</tr>
</tbody>
</table>

From the above table the classification accuracy of breast cancer data set is 98.4%. The experimental results show that the proposed model achieves better accuracy compared to the existing research in the same data set.

**Conclusion:**

Today's clinical databases store detailed information about patient diagnoses, lab test results and details from patient treatments, a virtual gold mine of information for medical researchers. Utilizing data mining techniques with medical treatment data is a virtually unexplored frontier. Studying the extraordinary behavior of outliers helps uncovering the valuable knowledge hidden behind them and aiding the decision makers to improve the health care services. In this paper we proposed an automated diagnosis of breast cancer using farthest first clustering and decision tree J48 classifier. The Wisconsin breast cancer dataset from UCI is used in this research. Our experimental results show the effectiveness of the proposed model. The obtained classification accuracy of 98.4 % is a very optimistic result compared to the existing breast cancer diagnosis for the same data set.

**REFERENCES**


