Performance Comparison for C4.5 and K-NN Techniques on Diabetics in Heart Problem

Viswanathan K 1, Dr. Mayilvahanam P 2, R. Christy PushpaLeela 3
Technical Architect 1, Head 2, Assistant Professor 3
Cognizant Technology Solutions, Chennai, India 1, Vels University Chennai, India 2, Women’s Christian College, Chennai, India 3

Abstract:
The aim of this research paper is to study and discuss the various classification algorithms applied on different kinds of datasets in terms of medical and compare the dataset performance. The classification algorithms with maximum accuracies on various kinds of medical datasets are taken for performance analysis. The result of the performance analysis shows the most frequently used algorithms on particular medical dataset. This research paper also discusses comparison of C4.5 and K-NN. This research paper summarizes various reviews and technical articles which focus on the current research on Medical diagnosis.

Keywords: Classification, Medical Diagnosis, C4.5, K-NN, Dataset and UCI, WEKA tool, Data-preprocessing.

I. INTRODUCTION

Data mining is the process of discovering actionable information from large sets of data. Data mining, the extraction of hidden predictive information from large databases. Data mining is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining uses mathematical analysis to apply the patterns and movements that exist in data. These patterns cannot be discovered by traditional data exploration because the relationships are too complex or because there is too much data. This process can be defined by using the following six basic steps:
1. Problem Definition
2. Data prepare
3. Exploring data
4. Apply the model
5. Walk around & authenticating models
6. Deploying & updating models

Data-Preprocessing:

Is a data mining technique that involves transforming raw data into a reasonable format. Real-world data is frequently incomplete, inconsistent and lacking in certain actions and is likely to contain many errors. Data preprocessing is a supported method of resolving these issues called data-preprocessing.

Fig 1: Data Mining – Process Flow

II. DATA-CLASSIFICATION

Data classification is one of the data mining methodologies used to forecast and classify the predetermined data for the specific class. There are different classifications methods and some of the methods are below.
1. Support Vector Machine (for linear and nonlinear data-SVM)
2. K-nearest neighbor classifier (KNN)
3. C4.5
4. Bayesian classification (Statistical classifier)
5. Decision tree
6. Rule based classification (IF THEN Rule)
7. Classification - Association Rule
2.1 SVM—Support Vector Machine

SVM is a classification method. SVM supports data mining and test mining. It was developed during the year of 1992. SVM Generally is capable to delivered high performance in terms of classification accuracy compare to other classification.

2.2 C4.5

C4.5 algorithm is a greedy algorithm and was developed by Ross Quinlan and its used for the induction of decision trees. C4.5 is often referred to as a statistical classifier. The decision tree algorithm C4.5 is developed from ID3 in the following ways.

2.3 K-NN

It is the nearest neighbour algorithm. The k-nearest neighbour’s algorithm is a technique for classifying objects based on the next training data in the feature space. The algorithm operates on a set of d-dimensional vectors, $D = \{ x_i | i = 1 \ldots N \}$, where $x_i \in k d$ denotes the $i^{th}$ data point. Selection k points in k d as the initial k cluster representatives. Algorithm iterates between two steps till junction.

Step 1: Data Assignment each data point is assign to its adjoining centroid, with ties broken arbitrarily. This results in a partitioning of the data.

Step 2: Relocation of “means”. Each group representative is relocating to the center (mean) of all data points assign to it. If the data points come with a possibility measure (Weights), then the relocation is to the expectations (weighted mean) of the data partitions.

“Kernelize” k-means though margins between clusters are still linear in the embedded high-dimensional space, they can become nonlinear when projected back to the original space, thus allowing kernel k-means to deal with more complex clusters.

The k-medoid algorithm is similar to k-means except that the centroids have to belong to the data set being clustered. Fuzzy c-means is also similar, except that it computes fuzzy membership functions for each clusters rather than a hard one.

III. Heart Disease

Diagnosis of heart disease is a significant and tedious task in medicine. Heart disease occurs when the arteries which normally provide oxygen and blood to the heart choked completely. Various types of heart diseases are,

- Heart failure
- Cardiomyopathy
- Coronary heart disease
- Hypertensive heart disease
- Inflammatory heart disease

Heart disease Common risk factors are below

- Age
- Gender
- High blood pressure
- Smoking Habits
- Abnormal blood lipids
- Fatness
- Diabetes
- Family antiquity
- No Physical task

IV. Data Set (Heart Patients Data Set)

The first step is the pre-processing to remove the inconsistent data. Apply the algorithm is used to classify the data.

<table>
<thead>
<tr>
<th>Sl no</th>
<th>Attribute</th>
<th>Data-Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>Numeric</td>
</tr>
<tr>
<td>2</td>
<td>Gender</td>
<td>Nominal</td>
</tr>
<tr>
<td>3</td>
<td>BP Systolic</td>
<td>Nominal</td>
</tr>
<tr>
<td>4</td>
<td>Diabetic</td>
<td>Numeric</td>
</tr>
<tr>
<td>5</td>
<td>BP Dialic</td>
<td>Numeric</td>
</tr>
<tr>
<td>6</td>
<td>Height</td>
<td>Numeric</td>
</tr>
<tr>
<td>7</td>
<td>Weight</td>
<td>Numeric</td>
</tr>
<tr>
<td>8</td>
<td>BMI</td>
<td>Numeric</td>
</tr>
<tr>
<td>9</td>
<td>Hypertension</td>
<td>Nominal</td>
</tr>
<tr>
<td>10</td>
<td>Rural</td>
<td>Nominal</td>
</tr>
<tr>
<td>11</td>
<td>Urban</td>
<td>Nominal</td>
</tr>
<tr>
<td>12</td>
<td>Disease</td>
<td>Nominal</td>
</tr>
</tbody>
</table>

Table 1: Data Set

V. WEKA TOOL

WEKA is a data mining tool. It provides the facility to classify the data through various algorithms. WEKA stands for Waikato Environment for Knowledge Learning and it was developed by the University of Waikato, New Zealand Software.
VI. Classification Matrix

The basic phenomenon used to classify the Heart disease classification using classifier is its performance and accuracy. The performance of a chosen classifier is validated based on error rate and computation time. The classification accuracy is predicted in terms of Sensitivity and Specificity. It compares the actual values in the test dataset with the predicted values in the trained model. In this example, the test dataset contained 208 patients with heart disease and 246 patients without heart disease.

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Classified as a Healthy(0)</th>
<th>Classified as a not Healthy(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Healthy (0)</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>Actual not Healthy (1)</td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

**Table 2: confusion matrix**

Where
- TP - True Positive
- TN - true Negative
- FP - False Positive
- FN - false Negative
- CT - Computing Time

For measuring accuracy rate and Error Rate, the following mathematical model is used.

**Accuracy** = \( \frac{TP + TN}{TP + FP + TN + FN} \)

**CERe** = \( \frac{FP + FN}{TP + FP + TN + FN} \)

**Performance Comparison (C 4.5 and K-NN)**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SMO</th>
<th>Naïve Bayes</th>
<th>C4.5</th>
<th>J48</th>
<th>KNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart Problem</td>
<td>77.34</td>
<td>76.3</td>
<td><strong>84.5</strong></td>
<td>73.8</td>
<td>83.2</td>
</tr>
</tbody>
</table>

**Table 3: Performance Comparison**

**VII. Conclusion**

Main objective of medical data mining algorithm is to get best algorithms that define given data from multiple features. C4.5 is the best classifier for classification and it shows better results in operational. Accuracy provided by C4.5 in binary classification is more reasonable. We have observed that C4.5 accuracy is good.

**VIII. REFERENCES**

[1] International Journal of Computer Science & Information Technology (IJCST) Vol 3, No 4, August 2011 “PERFORMANCE ANALYSIS OF VARIOUS DATAMINING CLASSIFICATION TECHNIQUES ON HEALTHCARE DATA” Shelly Gupta1, Dharminder Kumar2 and Anand Sharma3 IAIM & ACT, Banasthali University, Banasthali, India shelly.gupta24@gmail.com 2Department of CSE, GJUS&T, Hisar, India dr_dk_kumar_02@yahoo.com 3Department of CSE, GJUS&T, Hisar, India andz24@gmail.com.


