Selection Of The Best Classifier From Different Datasets Using WEKA
Ranjita kumari Dash
Assistant Professor, Institute Of Technical Education and Research, SOA University

Abstract

In today’s world large amount of data is available in science, industry, business and many other areas. These data can provide valuable information which can be used by management for making important decisions. By using data mining we can find valuable information. Data mining is the popular topic among researchers. There is a lot of work that cannot be explored till now. But, this paper focuses on the fundamental concept of the Data mining that is Classification Techniques. In this paper, Naive Bays, Functions, Lazy, Meta, Nested dichotomies, Rules and Trees classifiers are used for the classification of data set. The performance of these classifiers analyzed with the help of correctly classified instances, incorrectly classified instances and time taken to build the model and the result can be shown statistical as well as graphically. WEKA data mining tool is used for this purpose. WEKA stands for Waikato Environment for Knowledge Analysis. Three datasets are used on which different classifiers are applied to check which classifier is giving the best result, where different measurements are taken. 71 different classifiers are applied on this dataset. The dataset is in ARFF format. 10 fold cross validation is used to provide better accuracy. Finally the classification technique which provides the best result will be suggested. The result shows that no single algorithm always performed the best for each dataset.

KEY TERM’S

Bays Net, J48, Mean Absolute Error, Naive Bays, Root Mean-Squared Error

1. Introduction

Data mining is the process of extracting patterns from data [10, 11]. It is seen as an increasingly important tool by modern business to transform data as the technology advances and the need for efficient data analysis is required. Data mining involves the use of data analysis tools to discover previously unknown, valid patterns and relationships in large data set. It is currently used in a wide range of areas like marketing, surveillance, fraud detection, and scientific discovery etc.

In this paper we process a cancer dataset and use different classification methods to learn from the test data set.

Classification is a basic task in the data analysis that requires the construction of a classifier, that is, a function that assigns a class label to instances described by a set of attributes. It is one of the important applications of data mining. This technique predicts categorical class labels. In this paper, we are giving the comparison of various classification techniques using WEKA. Our aim is to investigate the performance of different classification methods using WEKA. Classification of data is very typical task in data mining. There are large number of classifiers that are used to classify the data such as Bayes, function, lazy learners, Meta, rule based and Decision tree etc. The goal of classification is to correctly predict the value.

For Breast cancer, there is a substantial amount of research with machine learning algorithm [1]. Machine learning covers such a broad range of processes that it is difficult to define precisely [6]. Young women being diagnosed in their teens, twenties and thirties. Even if the percentage is very low compared to that of older women aged 40 years and older [7, 8, 9]. 1% of all diagnosed breast cancers are in men. We report the case of a 34-year-old woman affected by
2. WEKA

WEKA stands for Waikato Environment for Knowledge Analysis. WEKA is created by researchers at the University of Waikato in New Zealand. WEKA was first implemented in its modern form in 1997. The GNU General Public License (GPL) is used here. The figure of WEKA is shown in the figure. The software is written in the Java™ language and contains a GUI for interacting with data files. For working of WEKA, we do not need the deep knowledge of data mining for which WEKA a very popular data is mining tool. WEKA also provides the graphical user interface of the user and provides many facilities. In this paper, we are giving the comparison of various classification techniques using WEKA. WEKA is a state-of-the-art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. The data file normally used by WEKA is in ARFF file format. ARFF stands for Attribute Relation File Format, which consists of special tags to indicate differentiating in the data file. WEKA implements algorithms for data pre-processing, classification, regression and clustering and association rules. It also includes visualization tools. It has a set of panels, each of which can be used to perform a certain task. The new machine learning schemes can also be developed with this package. WEKA is open source software issued under General Public License. The algorithms are applied directly to a dataset. The main features of WEKA includes

- 49 data pre-processing tools
- 76 classification/regression algorithms
- 8 clustering algorithms
- 15 attribute/subset evaluators + 10 search algorithms for feature selection.
- 3 algorithms for finding association rules
- 3 graphical user interfaces

3. METHODS

This section describes the classification methods used in this paper. We discuss each method and explain how the method has been used in our experiment. For this Breast cancer dataset we have taken eight methods Bayes, Functions, Lazy, Meta, Misc, Nested dichotomies, Rules and Trees classifiers for the classification of data set.

3.1. NAIVE BAYES CLASSIFIER

Bayes methods are also used as one of the classification solutions in data mining. In our work we use six main Bayesian methods namely AODE, AODEsr, Naive Bayes, Bayesian net, Naive Bayes simple and Naive Bayes updateable, that are implemented in WEKA software for classification. Naive Bayes is an extension of Bayes theorem in that it assumes independence of attributes. This assumption is not strictly correct when considering classification based on text extraction from a document as there are relationships between the words that accumulate into concepts. Problems of this kind, called problems of supervised classification, are ubiquitous. Naive Bayes sometimes also called as idiot's Bayes, simple Bayes and independence Bayes. This is important for several reasons.

It is easy to construct without any need for complicated iterative parameter estimation schemes. This means it may be readily applied to huge datasets. It is robust, easy to interpret, and often does surprisingly well though it may not be the best classifier in any particular application.

3.2. FUNCTION CLASSIFIER

Function classifier uses the concept of neural network and regression. Here two examples from neural network and regression will be taken for discussing the scenario[2]. A multilayer perceptron is a free forward artificial neural network model that maps sets of input data onto a set of appropriate output. It is a modification of the standard linear perceptron in that it uses three or more layers of neurons with nonlinear activation functions and it is more powerful than the perceptron in that it can distinguish data that is not linearly separable or separable by a hyperplane[4]. A multilayer perceptron has distinctive characteristics. The model of each neuron in the network includes a non linear activation function. The network contains one or more layers of hidden neurons that are not part of the input or output of the network. These hidden
neurons enable the network to learn complex tasks by extracting progressively more meaningful features from the input patterns. The network exhibits a high degree of connectivity determined by the network. A change in the connectivity of the network requires a change in the population of synaptic connections on their weights[5].

3.3. RULES CLASSIFIER

Association rules are used to find interesting correction relationship among all the attributes. They may predict more than one conclusion. The number of records an association rule can predict correctly is called coverage. Support is defined as coverage divided by total number of records[5]. Accuracy is the number of records that is predicted correctly expressed as a percentage of all instances that are applied to the methods of this algorithm are Conjunctive Rule, Decision table,DTNB,JRip,NNge,Oner,Rider and Zero. Rules are easier to understand than large trees. One root is created for each path from the root to the leaf. Each attribute value pair along a path forms a conjunction. The leaf holds the class prediction. Rules are mutually exclusive. These are learned one at a time. Each time a rule is learned, the tuples are removed. decision tree induction has been studied in details in both areas of pattern recognition and machine learning [13, 14]. This synthesizes the experience gained by people working in the area of machine learning and describes a computer program called ID3.

3.4. LAZY CLASSIFIER

When making a classification or prediction, lazy learners can be computationally expensive. They require efficient storage techniques and well suited to implementation on parallel hardware. They offer little explanation or insight into the structure of the data. Lazy learners however, naturally support incremental learning. They are able to model complex decision spaces having hyper polygonal shapes that may not be as easily describable by other learning algorithms. The methods of this algorithm are IBI, IBK, K- Star, LBK and LWL.

3.5. META CLASSIFIER

Meta classifier includes a wide range of classifier. When the attributes have a large number of values because the time and space complexities depend not only on the number of attributes, but also on the number of values for each attribute.

3.6. DECISION TREES

With the help of figures we are showing the working of various algorithms used in WEKA. We are showing also advantages and disadvantages of each algorithm. Every algorithm has their own importance and we use them on the behaviour of the data. Deep knowledge of algorithms is not required for working in WEKA. This is the main reason WEKA is more...
suitable tool for data mining applications. This paper shows only the clustering operations in the WEKA, we will try to make a complete reference paper of WEKA.

Table for best algorithms:

<table>
<thead>
<tr>
<th>Name of algorithm</th>
<th>Correctly classified instance</th>
<th>Incorrectly classified instances</th>
<th>Time taken to build the model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesnet</td>
<td>72.028</td>
<td>27.972</td>
<td>0.03</td>
</tr>
<tr>
<td>Simple logistic</td>
<td>75.1748</td>
<td>24.8252</td>
<td>1.44</td>
</tr>
<tr>
<td>K-Star</td>
<td>73.4266</td>
<td>26.5734</td>
<td>0</td>
</tr>
<tr>
<td>Filtered classifier</td>
<td>75.5245</td>
<td>24.4755</td>
<td>0</td>
</tr>
<tr>
<td>Ordinal classifier</td>
<td>75.5245</td>
<td>24.4755</td>
<td>0.01</td>
</tr>
<tr>
<td>Misc</td>
<td>69.9301</td>
<td>30.0699</td>
<td>0</td>
</tr>
<tr>
<td>Decision Table</td>
<td>73.4266</td>
<td>26.5734</td>
<td>0.5</td>
</tr>
<tr>
<td>J48</td>
<td>75.5245</td>
<td>24.4755</td>
<td>0.01</td>
</tr>
</tbody>
</table>

4.2 Comparison between LUNG dataset, HEART dataset, DIABETES DATASET
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Correctly Classified Instances in %</th>
<th>Incorrectly Classified Instances in %</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Time taken to build model in seconds (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multilayer Perceptron</td>
<td>100</td>
<td>0</td>
<td>0.4</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>Multiclass Classifier</td>
<td>77.2135</td>
<td>22.765</td>
<td>0.3</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>SPegasos</td>
<td>77.7344</td>
<td>22.2656</td>
<td>0.3</td>
<td>0.19</td>
<td></td>
</tr>
</tbody>
</table>

Table no -2(lung dataset,heart dataset,diabetes dataset)
5. References


