DATA MINING OF MEDICAL DATASETS WITH MISSING ATTRIBUTES FROM DIFFERENT SOURCES

by

Sunitha Sajja

Submitted in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in the

Department of Mathematics and Statistics

YOUNGSTOWN STATE UNIVERSITY

December, 2010

DATA MINING OF MEDICAL DATASETS WITH MISSING ATTRIBUTES FROM DIFFERENT SOURCES

Sunitha Sajja

I hereby release this thesis to the public. I understand that this thesis will be made available from the OhioLINK ETD Center and the Maag Library Circulation Desk for public access. I also authorize the University or other individuals to make copies of this thesis as needed for scholarly research.

Signature:

_____

Sunitha Sajja, Student Date

Approvals:

_____

John Sullins, Thesis Advisor Date

_____

Alina Lazar, Committee Member Date

_____

Jamal Tartir, Committee Member Date

_____

Peter J. Kasvinsky, Dean of School of Graduate Studies and Research Date

ABSTRACT

Two major problems in data mining are 1) Dealing with missing values in the datasets used for knowledge discovery, and 2) using one data set as a predictor of other datasets. We explore this problem using four different datasets from the UCI Machine learning repository, from four different sources with different missing values. Each dataset contains 13 attributes and one class attribute which denotes the presence of heart disease and the absence of heart disease. Missing values were replaced in a number of ways; first by using normal mean and mode method, secondly by removing the attributes that contains missing values, thirdly by removing the records that contains more than 60 percent of values missing and filling the remaining missing values. We also experimented with different classification techniques, including Decision tree, Naive Bayes, and MultiLayerPerceptron, using Medical Datasets. Rapid Miner and Weka tools. The consistency of the datasets was found by combining the datasets together and comparing the results of this datasets with the classification error of different datasets. It can be seen from the results that if fewer number of missing values are present, the normal mean and mode method is good. If larger amount of missing values are present than removing instances that contain 60% of missing values and replacing with remaining along with different preprocessing steps works better, and using one dataset as a predictor of other dataset produced moderate accuracy

ACKNOWLEDGEMENTS

First of all I would like to thank the Almighty, for bestowing upon me His grace and guidance, without which, this would have been impossible.

I express my heartfelt thanks to my thesis advisor Dr. John Sullins, for giving me the opportunity, guidance, encouragement and cooperation throughout this research. My special thanks to my thesis committee members Dr. Jamal Tartir, for helping me with all the departmental and logistic support. My deepest gratitude to Dr. Alina Lazar for introducing me to data mining. My sincere thanks to the Mathematics and Statistics department, Youngstown State University and Graduate School for funding me to pursue my Masters.

My greatest thanks to all my friends and especially my parents and sisters, it is your encouragement and love which is very crucial behind whatever I achieved so far.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

1. Introduction

1.1 Background

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Various steps involved in mining the data are data integration, data selection, data cleansing, data transformation, data mining [9].The first step in data mining is Data selection; that is, to only select the data which is useful for data mining. The second step is data cleansing; the data we collected may have errors, missing values; inconsistent data which must be corrected. Even after data cleansing data is not ready for data mining, so the next step is data transformation; that is, aggregation, smoothing, normalization, discretization etc. The final step is the data mining itself; that is, to find interesting patterns within the data. Data mining techniques include classification, regression, clustering, association etc.[1]

The main goal of data mining is classification. Given a collection of records containing a set of attributes, where one of the attribute is class attribute, our goal is to find the model for class attribute as a function of values of other attributes. A training set is used to build the model and a test set is used to classify the data according to that model. To predict the performance of a classifier on new data, we need to assess its error rate on a dataset that played no part in the information of classifier. This independent dataset is called a test set [2]

1.2 Problems

The main focus of this paper is to find the patterns in the data set related to coronary artery disease from UCI heart data sets. According to a 2007 report, nearly 16 million

Americans have coronary artery disease (CAD). In U.S., coronary artery disease is the leading killer of both men and women. Each year, nearly 500,000 people die because of CAD.[8].

Usually a medical dataset consists of a number of tests to be conducted to diagnose a disease. However, most medical datasets have large numbers of missing values because of the tests that are not conducted; many useful attribute values will be missing in medical data set due to the expense of performing tests, attributes that could not be recorded when the data was collected, or attributes ignored by users because of privacy concerns.[6]. Further complicating the problem from a data mining standpoint is that different groups of physicians collect different data; that is, different medical datasets often contain different attributes, making it difficult to use one dataset as a predictor of another.

This paper mainly focuses on two different issues. The first focus of this paper is preprocessing the data mainly dealing with missing attributes, and to find the improved accuracy of the data set after preprocessing steps and to compare the accuracy using different classification algorithms such as decision tree, Naive Bayes and Neural Networks.

The second focus is to train the data using classification techniques and to test the data using different datasets, to verify the results obtained from the trained data. The Cleveland database, collected from Cleveland Clinic Foundation, was used as the training set. The Switzerland, Hungary, and VA Long Beach data sets were used as test sets. All the datasets contain 13 attributes and one class attribute. The Cleveland dataset used for

training contains only 6 missing values, and the three test datasets contain almost 90% missing values.

1.3 Data Sets

The heart data set is collected from UCI repository. Each data set consists of 13 attributes among that 6 are numerical and 8 are categorical attributes and one special attribute.

The following are the four data sets for heart disease:

1. Cleveland

2. Hungary

3. Switzerland

4. VA long beach.

The source and the creator of the datasets:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.

3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D

This datasets contains 76 attributes but only 14 attributes are mostly used in most of the research. The presence value of heart disease is a value in the range of 1,2,3,4, with an absence value 0.

Table 1. All the four Data Sets

Data Sets	Number of Instances
Cleveland	303
Hungarian	294
Switzerland	123
VA Long Beach	200

The 14 attributes that are used are:

1. age- Age in years

2. Sex - (1=Male; 0=Female)

3. Cp -Chest pain type

   Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain,

   Value 4: asymptomatic

4. trestbps-Resting blood pressure (in mm Hg on admission to the hospital)

5. Chol-Serum cholesterol in mg/dl

6. fbs-(Fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

7. restecg-Resting electrocardiographic results

   Value 0: normal

   Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

8. thalach-Maximum heart rate achieved

9. exang- Exercise induced angina (1 = yes; 0 = no)

10.oldpeak- ST depression induced by exercise relative to rest

11.slope- The slope of the peak exercise ST segment

Value 1: up sloping ,Value 2: flat ,Value 3: down sloping

12.ca- Number of major vessels (0-3) colored by flourosopy

13. thal-3 = normal; 6 = fixed defect; 7 = reversable defect

14. The predicted attribute

Diagnosis of heart disease (angiographic disease status)

Value 0: < 50% diameter narrowing

Value 1, 2, 3, 4: > 50% diameter narrowing

The attributes with the largest amount of missing values are slope of the peak exercise, ST segment represented as slope, number of major vessels (0-3) colored by flourosopy represented by ca , normal; fixed defect; reversable defect represented by thal, and serum cholesterol in mg/dl represented by chol.

2. Literature Review

2.1 RapidMiner

RapidMiner, formerly known YALE (Yet Another Learning Environment), is software widely used for machine learning, knowledge discovery and data mining. RapidMiner is being used in both research and also in practical data mining fields.

The Java programming language is used in Rapid Miner, which means it can run in any operating system. RapidMiner can handle many formats of input such as CSV, Arff, SPSS, Xrff, Database example sources, and attributes that are described in XML file format. Different types of attributes that are present are Input, Output, data preprocessing and visualization.

RapidMiner contains more than 500 operators. The nested operator can be described through graphical user interface XML files which are created with RapidMiner. Individual RapidMiner functions can also be called directly from command line. It is used easily to define analytical steps and to generate graphs more effectively. It provides a large collection of data mining algorithms for performing classification. Many visualization tools such as overlapping histogram, 3D scatter plot and tree charts are present.

RapidMiner can handle any type of tasks like classification, clustering, validation, visualization, preprocessing, post processing etc. It also supports many kinds of preprocessing steps such as discretization, Outlier (Detection and removal), Filter, Selection, weighting, Normalization etc are available. All modeling and attribute evaluation methods from Weka are available within RapidMiner.

RapidMiner consists of two views, Design view and Result view. The design view is used to generate the process and run the process. The result view is used to generate the results.

Figure 1. RapidMiner Interface

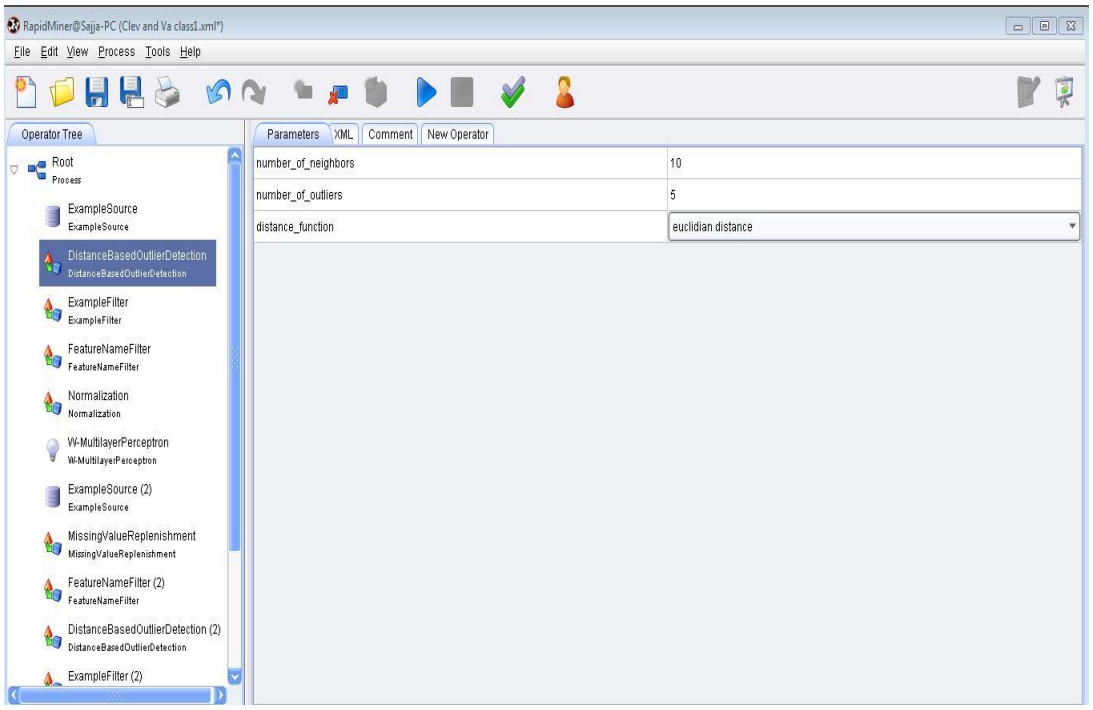
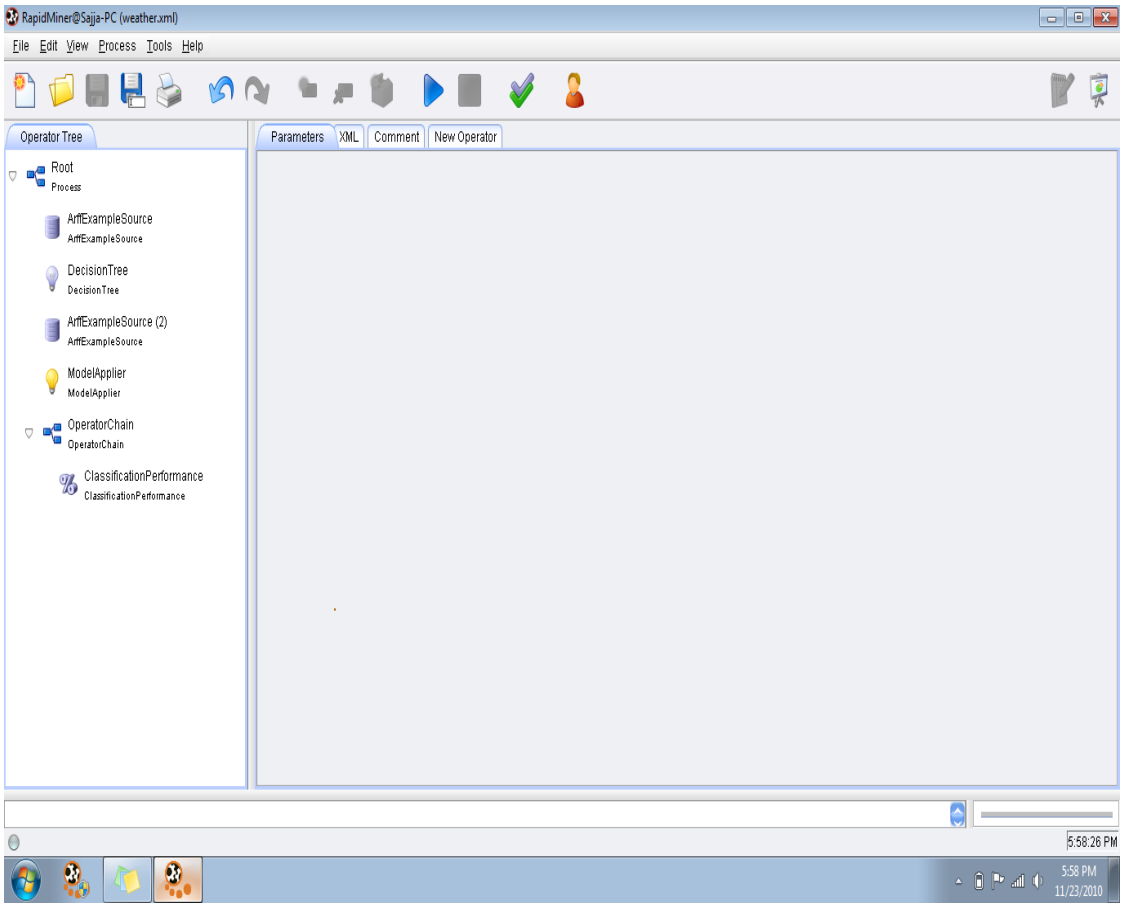


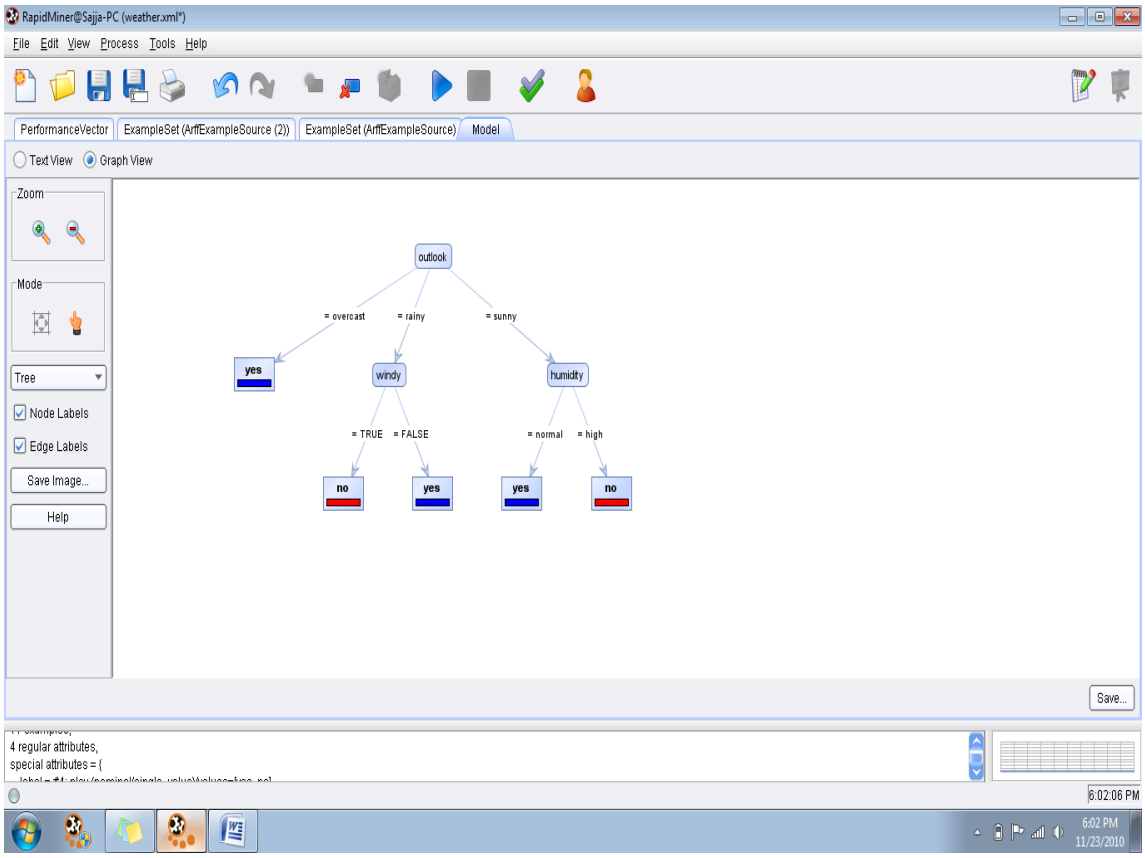
Figure 2. RapidMiner Design View



2.2 Decision Trees

Decision trees are a supervised learning technique commonly used for tasks like classification, clustering and regression. Decision trees are mainly used in the field of finance, engineering, marketing and medicine. Decision tress can handle any type of data that is nominal, numeric and text. They mainly focus on the relationships of the attributes. Input to a decision tree is the set of objects described by the set of properties, and creates output as yes/no decision, or as one of several different classifications.[25].

Figure 3. Decision tree created by RapidMiner



Since decision trees can be represented graphically as tree-like structures they are easier to understand by humans. The root node is the beginning of the tree, and each node is used to evaluate the attributes. At each node, the value of the attribute for the given instance is used to determine which branch to follow to a child node. Classification of instances can be done using a decision tree starting from the root node and continuing until a leaf node is reached. [3]

Decision tree creation involves dividing the training data into root node and leaf node divisions until the entire data set has been analyzed. The data is split until they have the

same set of classification or the splitting cannot be done anymore due to lack of further attributes. [14]

An efficient decision tree is one in which the root node divides the data effectively, and therefore requires fewer nodes. One of the important things is to select an attribute that best splits the data into individual classes. The splitting is done based on the information gain of each attribute. Information gain is based on the concept of entropy, which gives the information required for a decision in bits. Entropy is calculated from

$$\text{Entropy } (P_1, P_2 \ldots P_n) = -\Sigma_i P_i \log_2 P_i$$

Information of data is calculated based on each attribute. Entropy gives how important an attribute is by the information that is given. First the entropy of whole data set is calculated. The split is done based on this attribute. The attribute that can best split the data can be found from this. In the same way this procedure is used until the leaf nodes are reached. [10]

A decision tree is built using a training dataset and these trees can then be used to classify examples in test dataset. Decision trees can also be used to explicitly describe data and also used for decision making, as they produce rules which can be easy to understand and can be read by any user.

Sometimes decision tree learning produces a tree that is too large. If the tree is too large the new samples are poorly generated. Pruning is one of the important steps in decision tree learning that addresses this problem. The size of decision tree can be reduced by pruning (that is, removing) the irrelevant attributes for which the accuracy of decision tree does not get reduced by pruning. Pruning improves the accuracy of the tree for future

instances. The problem of over fitting and also noisy data can be reduced by pruning since the irrelevant attributes created by them are ignored. [27].

2.3 Naive Bayes

Another classifier is Naive Bayes. Naive Bayes operates in two phases, training set and testing set. Naive Bayes is cheap and it can handle around 10,000 attributes. It is also fast and highly scalable model.

Naive Bayes considers attributes as independent of each others in terms of contributing to the class attribute. [7]. For example a fruit may be considered as apple if it round, red, and 4'' in diameter. Although these features depend on each other, Naive Bayes considers these features independently to consider it as apple. [15].

One of the problems with Naive Bayes is that Naive Bayes does not require lot of instances for the possible combination of attributes. Naive Bayes can be used for both binary and multiclass classification problems. Naive Bayes can only handle discrete or discretized attributes. Naive Bayes requires binning. Several discretization methods are present they are Supervised and unsupervised discretization. Supervised discretization method uses class information of training set to select discretization cut point. Unsupervised discretization does not use the class information. [5]

Entire training data set is used for classification and discretization. Unsupervised discretization methods are equal width, Equal frequency and fixed frequency discretization. Error based, Entropy based are supervised discretization methods.[4]. Entropy based discretization uses class information, the entropy is calculated based on the class label then it finds the best split so that the bins are as pure as possible that is the

majority of values in a bin correspond to having the same class label. The split is done based on the maximal information gain.

2.4 Neural Networks

The human brain serves as a model for Neural Networks. Artificial neurons were first proposed in 1943 by Warren McCulloch, a neurophysiologist, and Walter Pitts, an MIT logician. Neural Networks are useful for data mining and decision support applications. They are also useful for pattern recognition or data classification through the learning process.

A neural network contains the neurons and weight building blocks. The strength of the network depends on the interaction between the building blocks. The MultiLayerPerceptron (MLP) Neural Network Model is mostly used, with networks that consist of three layers Input, Hidden and Output. The values of the input layer come from values in a data set. The input neurons send data via synapses to the hidden layer and through output layer through synapses.[19]

The MLP uses supervised technique called back propagation for training. Each layer is fully connected to the succeeding layer. The signal for each neuron is received from the previous layer; each signal is multiplied by a different weight value.[17]. Then the inputs that are weighted are summed and these are passed through the limiting function through this the outputs are scaled through the fixed range of values. Then the output is send to the all the neurons in the next layer. Error at each output is then “back propagated” to the hidden and the inputs, changing the weights based on the derivative of the error with respect to the weights.[21]

MLP training involves adjusting parameters such as the number of neurons to be used in hidden layer. If insufficient number of neurons is used the complex data cannot be modeled and the result would be poor. If more number of neurons are used, it may take long time, it may over fit the data. The network may perform well on the training set but the test set would give poor results on future instances. [20]

3. Methodology

The Cleveland, Switzerland, Hungary and VA Long Beach are the four datasets collected from the UCI Machine Learning Repository for this project. The Cleveland dataset is used for training and Switzerland, Hungary and VA datasets are used for testing. One important problem will be to use the training Cleveland dataset for testing the three datasets. The other important problem is to deal with the missing values in each dataset. All four datasets contains about 13 attributes and one class attribute.

First, all the four datasets are collected in .txt format. The datasets are loaded into RapidMiner using the IO Example Source operator.

3.1. Preprocessing the Data

The first step is to preprocess the Cleveland dataset used for training. This primarily involves dealing with missing values, outliers and feature selection. Different preprocessing steps that are used are

- To fill the missing values
- To deal with outliers
- Attribute selection ,numeric data discretization, Normalization etc

Once preprocessing of Cleveland data is done, the data set is used for training using different algorithms. Then different preprocessing steps are done on the three test data sets.

Missing values in the dataset represent a lot of different things. They may be due to a test that is not conducted or the data that is not available. Missing values in RapidMiner and Weka are usually represented by "?". The quality of the classification of the data could be reduced by the missing values, so filling in the missing values plays an important role in data mining. Different methods are used for dealing with missing values. The most frequently used method is replacing the categorical values with the mode and numerical values with mean. The second method is removing the attributes that contain around 90% of missing data. Attributes that are removed are ca, thal, slope, chol. No change was observed in the classification error when these attributes were removed. The third method that is used is to remove the instances in a data set if it contains 6 or more missing values out of the 13 values that are present, since instances missing too many values would not be good to use due to the consistency of the data. Then the remaining missing values are filled based on the frequency of class attributes.

Outliers are observations that deviate from the original dataset. That is, the instances that are abnormal distances from the other instances in the data. Sometimes they may occur due to some common errors that occur due to data transmission. [26]. In some cases outliers plays a significant role in acquisition of the data. Common methods used to identify outliers are Density based outlier detection, Distance based outlier detection and LOF detection methods.

Distance based outlier detection using a k-nearest neighbor algorithm is used to identify outliers. A density based outlier detection uses density functions like Square distance, Euclidean distance, angle, cosine distance, inverted cosine distance, and the LOFoutlierdetection identifies outliers using minimal upper and lower bounds with a density function.

Feature Selection is mainly used to find features that play an important role in classification, and to remove features with little or no predictive information.[11]. This method is mainly used in data sets with many features .Types of feature selection methods used are Filter method and Wrapper methods. The Filter method selects features independent of the classifier, and Wrapper methods makes use of classifier for feature selection. The filter method select features based on general characteristics of data so the filter methods are much faster than the wrapper method. The wrapper method uses an induction algorithm as a evolution function to select feature subsets.[16]

Four data sets are collected in .txt format. First the Cleveland dataset is read into Rapid miner by using the attribute editor. The (.dat and .aml) files are created for the Cleveland datasets. The average and mode of each attribute is obtained. Then the missing values are identified. The Cleveland dataset contains only 6 missing values. Removing the missing values from the Cleveland dataset would not affect the dataset since less than 2% of the data is missing. So the missing values are removed from the dataset.

The next step would be to deal with the outliers. Distance based outliers methods are used to detect the outliers using the K Nearest Neighbor and Euclidean distance. It has been

identified that depending on the significance of outlier, outlier role is determiner in medical data set, depending on role of the outlier they are removed or not removed**.** **[**13]

The next step would be to select features that play important roles in the data. Feature selection generally improves the accuracy of the classification by basing it on the most relevant features. [12].The Infogainweighting using Filter method and Wrapper method using Forward and Backward selection method are used. First the Information gain of the attributes is obtained. The attributes are selected first by choosing 4 attributes, 5 attributes and so on and also by using 50% of the attributes, 70% of the attributes. By using these different methods we identify the top 10 features that play important roles in the classification, which are selected to build the model. Wrapper method selects features depending on the learning algorithm and the features selected by one algorithm may differ for another algorithm. Forward and backward selections are the two methods used to build a set of features. Forward selection starts only with one subset of attribute and additional attributes are added until there is no performance gain. Backward selection is the opposite of forward selection, as it starts with complete attribute set and attribute are removed from that subset until there is gain in the performance. Decision Tree and MultiLayerPerceptron are used as algorithms to select features using Forward and Backward selection. Then the preprocessed Cleveland data set is saved as a new file.

Once the Cleveland data set used for training is preprocessed it is ready to test with the three testing sets. Before testing the data sets, preprocessing of the three datasets is also done. Three classification algorithms Decision tree, Naive Bayes and MultiLayer Perceptron are used to build a classification model using the Cleveland dataset. Then the model is build using the classification algorithms.

The next important step is preprocessing of test datasets. One important step in preprocessing this datasets is dealing with the missing values, which were much more prevalent in the test datasets. The Switzerland, Hungarian and VA long Beach datasets consist of 50% to 90% of instances with missing values. Because of this, it is not possible to simply remove instances with missing values.

The following missing values methods are used to fill the missing data. We first replaced the categorical values with the mode and numerical values with mean. As the cholesterol attribute has about 99% of missing values in one dataset, it was replaced by normal value based on age and gender. The cholesterol value is replaced by normal level of values that is for Females below age 40 years chol level is 183 mg/dL, from age 40 to 49 years chol level is 119 mg/dL, from 50 years or above chol level is 219 mg/dL. For male below 40 years chol level is 185 mg/dL, age 40 to 49 years chol level is 205 mg/dL, age above 50 years chol level is 208 mg/dL.

By replacing the missing values in all the data sets with this method the Hungarian gave less classification error while the other two datasets still produced high classification error.

Different methods are used to deal with missing in the two data sets that produced the highest classification error. We tested removing the missing attributes that contain around 90% of missing data, ca, thal and slope, chol. Using this method did not affect improve the number of correct predictions, so the third method is used to deal with missing values.

As with the Cleveland dataset, instances in a data set were removed if they contain 6 or more than 6 missing values out of 13 values that are present. Since the ca attribute

contains about 99% of missing values the ca attribute is removed from the datasets. Since ca is a redundant attribute, removing this attributes does not affect the data set. Then the remaining missing values are filled based on the frequency of class attribute. This method, however, did not produce any change in results on the Switzerland data set. There was an increase in number of correct predictions for Hungary data set.

3.2. Building the Model

In order to compare the effectiveness of different classification algorithms, decision tree classification, Naive Bayes, and MultilayerPerceptron are used. First the Cleveland data set is used to build the model using the Decision tree classification algorithm in Rapid miner. Different criteria are used to build the decision tree, the criteria's used are gini index, gain ratio and information gain. Of these, the information gain produced a better decision tree, so the criteria used for attribute selection and also for numerical split for building decision tree is information gain. Simple accuracy is not the best to determine the classifier, so sensitivity and specificity are used instead.

The accuracy on the positive instances is Sensitivity:

Sensitivity = True Positive/ (True Positive + False Negative)

The accuracy on the negative instances is Specificity:

Specificity = True Negative/ (True Negative + False Positive)

Accuracy = True Positive + True Negative/ (True Positive + False Negative+ True Negative + False Positive)

The MultiLayerPerceptron method is used with more than one hidden layer to find the accuracy. One important step in MLP is choosing the number of hidden layers, as RapidMiner allows a choice of the number of hidden layers. First the numbers of hidden layers are chosen as 0, 1, 2 and so on and minimum of one hidden layer is used. The number of hidden layers that are used are 3 hidden layers. In this way, the training set is used to classify the other three datasets. The datasets are tested by replacing the missing values by three different methods.

4. Results

Different experiments are conducted to test how the data sets collected from one source act as a predictor of another data sets collected from different sources, and to compare the accuracy using different algorithm. Since the data set contains missing values three different methods are used to fill the missing values. Then the accuracy is obtained to identify the method that worked better to fill the missing values. Different preprocessing steps are also conducted. The features that play an important role are identified.

The first step is to find percentage of missing values in each data set. It can be noticed from Table 2. Below show that the attribute ca contains 90% to 99% of missing values in all the data sets where the ca value is replaced by using the normal value and also one attribute in the Switzerland data set is missing 100%.

Table 2. Percentage of missing values in each dataset

Attributes	Cleveland	Hungary	Switzerland	VA Long Beach
Age				
Sex				
CP				
Trestbps			2%	28%
Chol		8%	100%	4%
Fbs		3%	61%	4%
Restecg			1%	
Thalach			1%	27%
Exang			1%	27%
Oldpeak			5%	28%
Slope		65%	14%	51%
Ca	1%	99%	96%	99%
Thal		90%	42%	83%

The main goal is to use one data set as a predictor of another data set. Since the Cleveland data set contains less than 2%, for our experiments, we build a model using the Cleveland data set. Models were built from the preprocessed data set using Decision tree, Naive Bayes, and MultiLayerPerceptron (MLP) algorithms.

Table 3 gives the accuracy obtained using three algorithms after building the model in terms of the percentage of correct predictions, in terms of correctly matching one of the five possible values of the classification attribute heart disease, 0, 1, 2, 3and4.The accuracy of correct classification is based both on the True positive and True negative . while the error is calculated based on the number of wrong predictions. It can be observed from the table that the MultiLayerPerceptron worked better building the model with an accuracy of 91.75%, and Decision Tree was second with an accuracy of 81.19%.

Table 3. Accuracy of the Cleveland dataset using classification Algorithms

Algorithm	Accuracy
Decision Tree	81.19%
Naïve Bayes	63.97%
MultiLayerPerceptron	91.75%

Each model built using the Cleveland Data set is used to test the preprocessed datasets Hungary, Switzerland and VA Long Beach. The data from the Table 4 below shows that the highest amount of accuracy is obtained from the Hungary data set, while the others do not perform much better than random selection among the five possible classification values. As seen from the Table 2 we know that the numbers of missing values in the

Hungary data set are less compared to the Switzerland and VA Long Beach. This indicates that the normal method works better for Hungary data set.

Table 4. Accuracy obtained after the model was tested

Training and Test datasets	Decision Tree	Naïve Bayes	MLP
Cleveland and Switzerland	14.75%	31.15%	22.13%
Cleveland and VALongBeach	28.00%	33.50%	30.00%
Cleveland and Hungary	64.97%	65.65%	64.71%

The next important step is using Feature Selection for selecting the features that plays an important role in classification. Information gain weighting is used to select the features that plays appropriate role in the classification of data set. The top 10 attributes are selected to build the model. The following Table 5. gives the Information gain weighting of the attributes.

Table 5. Information gain weights of attributes

Attributes	Weights
Thal	1.0
CP	0.8732776041805517
Ca	0.7895335660685716
Thalach	0.6349096700185458
Oldpeak	0.5578017189627767
Exang	0.5256702757853041
Slope	0.5057053502400372
Age	0.24902728380010647
Sex	0.19151360545536644
Restecg	0.10978766175891774
Trestbps	0.010542392846149206
Chol	0.004312041740539823
Fbs	0.0

After selecting the attributes using the Information gain weighting, the top 10 attributes are used to build the model and the model build is used to test the three data sets to see if the number of correct predictions is improved. Table 6 shows the increase and decrease in the number of correct predictions.

Table 6. Accuracy obtained after information gain weighting of attributes

Data sets	Decision tree	Naïve Bayes	MLP
Switzerland	25.89%	24.11%	22.32%
VA Long Beach	33.16%	38.42%	33.16%
Hungary	64.29%	65.74%	68.03%

Feature selection using the Wrapper method was tested with different algorithms. The Decision tree and MLP algorithms are used to select the attributes, using both forward selection and backward selection methods. Table 7 gives the accuracy obtained by using forward selection method and backward selection methods for all three algorithms.

Table 7. Accuracy of Hungary Data set using Wrapper method

Data sets	Decision tree	Naive Bayes	MLP
Forward Selection	63.38%	65.49%	64.08%
Backward Selection	65.49%	65.49%	64.08%

From Table 4 and Table 6 it can be noted that the model built by the Cleveland data set was better able to classify the Hungary data set than the Switzerland and VA Long Beach data sets. The Switzerland and VA Long Beach algorithms were far worse, with less than 30% of attributes correctly predicted.

In order to improve this, removing instances with at least 6 missing attributes and filling the remaining instances with the instances that are present in the Switzerland and VA Long Beach was done, and the same algorithms and the preprocessing steps used for learning. Table 8 shows the accuracy of the dataset is improved than compared to the first method that is used to fill the missing values. The numbers of correct predictions are improved.

Table 8. Accuracy of the VA Long Beach after replacing the missing values by third method

Algorithm	Accuracy
Decision Tree	56.73%
Naive Bayes	54.44%
MultiLayerPerceptron	46.15%

The next step is to select the features that play an important role in the data set, based on different algorithms. Table 9 shows that the numbers of correct predictions are improved in some cases while the number of correct predictions in some cases remains the same.

Table 9. Feature Selection accuracy of VA Long Beach

Data sets	Decision tree	Naive Bayes	MLP
Information gain weighting	53.45%	55.25%	62.03%
Forward Selection	70.69%	69.32%	59.71%
Backward Selection	62.80%	60.25%	59.75%

5. Conclusion

This study has been conducted to see how feature selection works; dealing with missing values and to see how data set collected from one source can classify data collected from other sources. Studies are conducted both with feature selection and without feature selection. The results showed that selecting the appropriate features in the data set plays an important role in data classification. If the appropriate features are removed from the data set, however it can be seen that the accuracy of the data is reduced.

Three different methods are used to deal with missing values in data sets. The normal method worked better for smaller numbers of missing values after replacing the number the nonrandom missing values of attributes that contains 90% to 100% by normal value of that attribute. On the other hand, if large numbers of nonrandom missing values are present in the data, then removing the instances that contain 60% or more of missing values for each patient record and replacing the remaining values based on the remaining instance worked better compared to the Normal method. The datasets collected from different sources gave moderate accuracy if smaller numbers of missing values are present in the data. If larger numbers of missing values are present in the data, the data collected from one source gives very poor results in classification of the data from the other source.

6. References

[1]. Ansari, A and Ansari, S. (2010). *The concept of Data Mining, Its applications & issues*. Journal of Engineering and Sciences

[2]. Witten, I.H& Frank, E (2005).*Data mining: practical machine learning tools and techniques*. 144-146

[3]**.** Braha, D & Shmilovici, A. (2003).*On the Use of Decision Tree Induction for Discovery of Interactions in a Photolithographic Process*. IEEE transactions on semiconductor manufacturing, vol. 16, no. 4, 644-652

[4]. Kohavi, R and Sahami, M. (1996*). Error-Based and Entropy-Based Discretization of Continuous Features*. KDD-96 Proceedings

[5]. Hussain, F; Limtan, C; Dash, M and Liu, H. (2002). *Discretization: An Enabling Technique.* Data Mining and Knowledge Discovery, 6, 393–423

[6]. Zhang, S; Qin, Z; Ling, C. X and Sheng, S. (2005). "*Missing is Useful": Missing Values in Cost-Sensitive Decision Trees.* IEEE Transactions on knowledge and data engineering, VOL. 17, NO. 12

[7]. Zhang, H. (2004).*The Optimality of Naive Bayes*. American Association for Artificial Intelligence

[8]. The New York times (Health Guide – Coronary Heart Disease)

http://health.nytimes.com/health/guides/disease/coronary-heart-disease/background.html

[9]. Braha, D (Ed). (2002). *Data Mining for Design and Manufacturing*, Springer, 544

[10]. Witten, I.H& Frank, E (2005).*Data mining: practical machine learning tools and*

techniques. 101-105

[11]. Abel,H and Yamaguchi, T. *Constructive Meta-level Feature Selection Method Based on Method Repositories*

[12]. Xiong, W and Wang, C. (2008). *Feature Selection: A Hybrid Approach Based on Self-adaptive Ant Colony and Support Vector Machine*. IEEE International Conference on Computer Science and Software Engineering

[13]. Chen.Y; Dang, X and Peng, H. (2008). *Outlier Detection with the Kernelized Spatial Depth Function*

[14]. Two crows corporation. *Introduction to Data Mining and Knowledge discovery.* Third edition

[15]. Bhargavi, P and Jyothi, S. (2009). *Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils*. IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.8, 117-122

[16]. Zhuo, L ; Zheng ,J ; Wang ,F; Li ,X; Ai ,B ; Qian , J. (2008). *A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine*. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B7

[17]. Witten, I.H& Frank, E (2005).*Data mining: practical machine learning tools and techniques*. 411-413

[18]. Rosenblatt, F. (1958). The Perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 65: 386-408

[19]. Witten, I.H& Frank, E (2005).*Data mining: practical machine learning tools and techniques*. 233-237

[20]. Frank. X, R. (1961). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, Washington DC

[21]. Hecht-Nielsen, R. *Theory of the Backpropagation Neural Network.* I-593--I-605

[22]. Ghannad-Rezaie, M., Soltanian-Zadeh, H., Ying, H., and Dong, M. (2010). *Selection-fusion approach for classification of datasets with missing values*. Pattern Recognition, 43(6), 2340-2350

[23]. Gkirtzou, K; Tsamardinos, I; Tsakalides, P; and Poirazi, P. (2010). *MatureBayes: A probabilistic algorithm for identifying the mature miRNA within novel precursors*. PloS One, 5(8), e11843.

[24]. Sethi, P, and Jain, M. *A comparative feature selection approach for the prediction of healthcare coverage*, 392-403.

[25]. Aitkenhead, M.J. (2008).*A co-evolving decision tree classification method*, Expert Systems with Applications, Volume 34, Issue 1, 18-25.

[26]. Acuña, E and Rodriguez, C. *On detection of outliers and their effect in supervised classification*

[27]. Frank, E. (2000). *Pruning Decision Trees and Lists*.