A Comparative Study on Feature Selection Methods of Data Mining Technique

Prepared By

Md. Abdullah Al Mamun
Abu Newaz Md. Niloy
Md. Masum Ahmed

Computer Science and Engineering Department
North Western University
Khulna, Bangladesh

March, 2017

A Comparative Study on Feature selection Methods of Data Mining Technique

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering.

Nagib Mahfuz — Supervisor
Senior Lecturer
Computer Science and Engineering Discipline
North Western University
Khulna.

Md. Mahedi Hasan
Senior Lecturer — Second Examiner
Computer Science and Engineering Discipline
North Western University
Khulna.

Md.Tajul Islam — Head of the discipline
Assistant Professor
Computer Science and Engineering Discipline
North Western University
Khulna.

A Comparative Study on Feature selection Methods of Data Mining Technique

Md. Abdullah Al Mamun

ID No. 20172031011

Abu Newaz Md. Niloy

ID No. 20172031010

Md. Masum Ahmed

ID No. 20172031009

Computer Science and Engineering Discipline

North Western University

Khulna, Bangladesh

March, 2017

Acknowledgement

At first, we are grateful to almighty Allah for giving us strength, patience and intelligence to conduct our thesis properly. We would like to express our cordial regard to our supervisor Md. Inzamam-Ul-Hossain, Senior lecturer of Computer Science and Engineering Discipline, North Western University, Khulna for his valuable suggestions, co-operation and proper guidance. Actually his memorable contribution has inspired us to reach our goal. We are also grateful to our respected teacher's for their suggestions, researcher and weka tool creators for providing us with the necessary tools. At last we would like to express our gratitude to all of ours family members who support and take care of us to perform our thesis work properly.

Abstract

Feature selection methods have become a discernible need in medical data mining. Feature selection methods are used in data preprocessing to achieve a reduced data set that gives the most accurate results. In this paper we made a comparison study among the feature selection methods. Here we compared the main feature selection methods filter, wrapper and embedded method. These methods are explored to find out their use in practice. J48 is used as the method of classification. The weka platform is used to implement the feature selection techniques on medical data sets.

Keywords:

Feature selection, weka, filter, wrapper, embedded, j48, heart disease.

Table of Contents

List of Tables

List of Figure

Chapter 1
INTRODUCTION

1.1 BACKGROUND

In today's chaotic world humans are exposed to more life threatening illnesses. Out of this heart disease has taken the limelight. Heart disease refers to the injury of heart and blood vessels. Some common symptom of this disease are Discomfort, pressure, heaviness, or pain in the chest, arm, or below the breastbone, fullness, indigestion, or choking feeling, Sweating, nausea, vomiting, or dizziness [1]. The extent of the symptoms differs from person to person. The causes of heart disease are heart defects you're born with, Coronary artery disease, high blood pressure, diabetes, smoking, stress, age, sex. Medical diagnosis helps to find the symptoms and causes of this disease and produce information about the different variations of the disease.

Since 1938 when ligature of patent ducts arteriosus was performed; 1944, when created the first systematic pulmonary shunt in a child with cyanotic congenital heart disease; and 1945 when aortic coarctation was repaired the evolution of the treatment with congenital heart disease has been spectacular. This has given birth to a new population of adolescents and adults with congenital heart defects that are more or less repaired but partially never cured that requires specialized cardio logical monitoring. We are facing a problem that does not seem to have any easy solution, especially if we take into account the fact that in recent years the number of patients with congenital heart disease and the number of adults with other forms of heart diseases has gotten greater in size [2].

The introduction of ultrasound techniques, specially 2d echocardiography, radically changed the diagnostic algorithm for congenital heart defects. Thus tomographic image made it possible for the first time to identify noninvasively, anatomical abnormalities that previously required the performance of cardiac catheterization. This improved the diagnostic capability of heart disease in unborn children. Electrocardiogram, Chest X-ray, Echocardiogram, Exercise stress test, Cardiac computerized tomography (CT) or magnetic resonance imaging (MRI), and Cardiac catheterization has made it possible to diagnose heart problems in adults at an early stage making

the treatment procedure much easier. But all this equates to more data, processing which presents a huge problem. This is where data mining is needed.

1.2 Data

In Data Mning techniques, we at first need to consider the data that would be mined. To get an accurate result, data should be correct. But in real world, raw data has missing values. If we do not consider the above situations, we can't get the actual and better result. So, data preprocessing is very important in data mining. The present DM tools have enriched features to visualize data for preprocessing. We can see the relation among the fields to see how they are related with each other and how they would play role in the result. In this project, Heart Diseases dataset is collected from reliable sites UCI Machine Learning Repository [3]. This dataset is created in the University of Oxford in collaboration with the National Centre for voice and speech. Many researchers have used this dataset. There are 270 instances (including attribute names) with no missing values. Here 13 attributes are available. The attribute information for the dataset is shown in table 1.

Table 1: Attribute name and attribute type of the Heart diseases dataset

No. of Attribute	Attribute	Description	Data Type
1.	Age	Age in years	Real
2.	Sex	Male, female	Binary
3.	chest pain type (4 values)	1.Typical angina 2.Atypical angina 3.Non-anginal pain 4.asymptomatic	Nominal

4.	resting blood pressure	Patient resting blood pressure in mm Hg at the time of admission to the hospital.	Real
5.	serum	cholesterol in mg/dl	Real
6.	fasting blood sugar	Boolean measure indicating weather fasting blood sugar is greater than 120 mm/dl	Binary
7.	resting electrocardiographic results (values 0,1,2)	Electrocardiographic results during rest	Nominal
8.	maximum heart rate achieved	maximum heart rate achieved during test	Real
9.	exercise induced angina	Weather exercise induced angina has occurred	Binary
10.	Oldpeak	ST depression induced by exercise relative to rest	Real
11.	Slope	the slope of the peak exercise ST segment	Ordered
12.	number of major vessels	number of major vessels (0-3) colored by flourosopy	Real
13.	Thal	3 = normal; 6 = fixed defect; 7 = reversable defect	Nominal

1.3 Statistics of heart disease

Heart disease, stroke and other cardiovascular disease statistics

Cardiovascular disease is the leading global cause of death, accounting for 17.3 million deaths per year, a number that is expected to grow to more than 23.6 million by 2030 [4]. In 2008, cardiovascular deaths represented 30 percent of all global deaths, with 80 percent of those deaths

taking place in low- and middle-income countries. Nearly 787,000 people in the U.S. died from heart disease, stroke and other cardiovascular diseases in 2011. That's about one of every three deaths in America. About 2,150 Americans die each day from these diseases, one every 40 seconds.

Cardiovascular diseases claim more lives than all forms of cancer combined. About 85.6 million Americans are living with some form of cardiovascular disease or the after-effects of stroke. Direct and indirect costs of cardiovascular diseases and stroke total more than $320.1 billion. That includes health expenditures and lost productivity. Nearly half of all African-American adults have some form of cardiovascular disease, 48 percent of women and 46 percent of men. Heart disease is the No. 1 cause of death in the world and the leading cause of death in the United States, killing over 375,000 Americans a year. Heart disease accounts for 1 in 7 deaths in the U.S. Someone in the U.S. dies from heart disease about once every 90 seconds.

Heart Disease

From 2001 to 2011, the death rate from heart disease has fallen about 39 percent – but the burden and risk factors remain alarmingly high. Heart disease strikes someone in the U.S. about once every 43 seconds. Heart disease is the No. 1 cause of death in the United States, killing over 375,000 people a year. Heart disease is the No. 1 killer of women, taking more lives than all forms of cancer combined. Over 39,000 African-Americans died from heart disease in 2011.Cardiovascular operations and procedures increased about 28 percent from 2000 to 2010, according to federal data, totaling about 7.6 million in 2010.About 735,000 people in the U.S. have heart attacks each year. Of those, about 120,000 die. About 635,000 people in the U.S. have a first-time heart attack each year, and about 300,000 have recurrent heart attacks.

Stroke

In 2010, worldwide prevalence of stroke was 33 million, with 16.9 million people having a first stroke. Stroke was the second-leading global cause of death behind heart disease, accounting for 11.13% of total deaths worldwide. Stroke is the No. 4 cause of death in the United States, killing

nearly 129,000 people a year. Stroke kills someone in the U.S. about once every four minutes .African-Americans have nearly twice the risk for a first-ever stroke than white people, and a much higher death rate from stroke. Over the past 10 years, the death rate from stroke has fallen about 35 percent and the number of stroke deaths has dropped about 21 percent. About 795,000 people have a stroke every year. Someone in the U.S. has a stroke about once every 40 seconds. Stroke causes 1 of every 20 deaths in the U.S. Stroke is a leading cause of disability. Stroke is the leading preventable cause of disability.

Sudden Cardiac Arrest

In 2011, about 326,200 people experienced out-of-hospital cardiac arrests in the United States. Of those treated by emergency medical services, 10.6 percent survived. Of the 19,300 bystander-witnessed out-of-hospital cardiac arrests in 2011, 31.4 percent survived. Each year, about 209,000 people have a cardiac arrest while in the hospital.

Smoking

Worldwide, tobacco smoking (including secondhand smoke) was one of the top three leading risk factors for disease and contributed to an estimated 6.2 million deaths in 2010.16 percent of students grades 9-12 report being current smokers. Among adults,

20 percent of men and 16 percent of women are smokers .Among adults, those most likely to smoke were American Indian or Alaska Native men (26 percent), white men (22 percent), African-American men (21 percent), white women (19 percent), American Indian or Alaska Native women (17 percent), Hispanic men (17 percent), African-American women (15 percent), Asian men (15 percent), Hispanic women (7 percent), Asian women (5 percent).In 2012 there were approximately 6,300 new cigarette smokers every day.

Physical Activity

About one in every three U.S. adults – 31 percent – reports participating in no leisure time physical activity. Among students in grades 9-12, only about 27 percent meet the American

Heart Association recommendation of 60 minutes of exercise every day. More high school boys (36.6%) than girls (17.7%) self-reported having been physically active at least 60 minutes per day on all seven days.

Healthy Diet

Less than 1 percent of U.S. adults meet the American Heart Association's definition for "Ideal Healthy Diet." Essentially no children meet the definition. Of the 5 components of a healthy diet, reducing sodium and increasing whole grains are the biggest challenges. Eating patterns have changed dramatically in recent decades. Research from 1971 to 2004 showed that women consumed an average of 22 percent more calories in that span and men consumed and average of 10 percent more. The average woman eats about 1,900 calories a day and the average man has nearly 2,700, According to the government figures:

Obesity

Most Americans who are older than 20 are overweight or obese, over 159 million U.S. adults – or about 69 percent – are overweight or obese. About 32 percent children are overweight or obese. About 24 million are overweight and about 13 million – 17 percent – are obese. In 2008, an estimated 1.46 billion adults worldwide were overweight or obese. The prevalence of obesity was estimated at 205 million men and 297 million women.

Cholesterol

About 43 percent of Americans have total cholesterol higher of 200 mg/dL or higher. The race and gender breakdown shows 46 percent of Hispanic men, 46 percent of white women, 43 percent of Hispanic women,41 percent of black women, 40 percent of white men, 37 percent of black men.

About 13 percent of Americans have total cholesterol over 240 mg/dL. Nearly one of every three Americans has high levels of LDL cholesterol (the “bad” kind).About 20 percent of Americans have low levels of HDL cholesterol (the “good” kind).

High Blood Pressure

About 80 million U.S. adults have high blood pressure. That’s about 33 percent. About 77 percent of those are using antihypertensive medication, but only 54 of those have their condition controlled. About 69 percent of people who have a first heart attack, 77 percent of people who have a first stroke and 74 percent who have congestive heart failure have blood pressure higher than 140/90 mm Hg. Nearly half of people with high blood pressure (46 percent) do not have it under control. Hypertension is projected to increase about 8 percent between 2013 and 2030.Rates of high blood pressure among African-Americans is among the highest of any population in the world. Here is the U.S. breakdown by race and gender.

46 percent of African-American women have high blood pressure.
45 percent of African-American men have high blood pressure.
33 percent of white men have high blood pressure.
30 percent of white women have high blood pressure.
30 percent of Hispanic men have high blood pressure.
30 percent of Hispanic women have high blood pressure.
In 2000, it was estimated that 972 million adults worldwide had hypertension.

Blood Sugar/Diabetes

The prevalence of diabetes for adults worldwide was estimated to be 6.4 percent in 2010 and is projected to be 7.7 percent in 2030. The total number of people with diabetes is projected to rise from 285 million in 2010 to 439 million in 2030.About 21 million Americans have diagnosed diabetes, almost 9 percent of the adult population, but diabetes rates are growing. In fact, about 35 percent of Americans have pre-diabetes. African-Americans, Hispanics/Latinos and other ethnic minorities bear a disproportionate burden of diabetes in the U.S.

1.3 Data mining and disease detection

Data mining is used to recognize patters in large volumes of data. Various data mining techniques can be applied to the heart diagnosis data to provide an additional source of knowledge to healthcare professionals for making decisions. The massive data produced by diagnostic procedures are hard to deal with, this high dimensional nature of the data has given birth to a wealth of feature selection techniques being used in this field.

Feature selection is the process of extracting relevant feature from the original set of candidate features. Unlike feature extraction feature selection methods are applied to data sets with known features. In this paper we used filer, wrapper and embedded method along with j48 classifier to obtain features of heart disease. We also made a comparison among these methods to obtain the method best suited for disease detection.

Chapter 2
AIMS AND OBJECTIVES

In order to get a desirable outcome out of any work a clear goal needs to be set otherwise there will be no path to follow and the time will be wasted. This chapter provides the aims and objectives of the work that's to be done to successfully complete the project.

2.1 Aims of the project:

The aim of the project is to investigate various feature selection techniques and find out the feature selection technique best suited for disease detection by making a comparison study.

2.2 Objectives of the project:

To achieve the aim mentioned above the following objectives are set and followed for this project.

- The first objective was to get information about data mining and disease detection.
- The second objective was to do some research on the previous work done on feature selection.
- The third objective was to determine the feature selection technique that are widely used and provide the best accuracy and compare them to find the one best suited for disease detection.
- The fourth objective was to find the appropriate tools that can be used to work on the whole project.
- The last objective was to use the tools and techniques to achieve our aim.

2.3 Summary:

In order to make a successful project, one has to have a clear idea of the goal to be achieved and organize the resources and tools that will be vital in reaching the goal. This chapter provides a clear idea of the aim of the project and work that is done. It also shows the objectives that were fulfilled to make this project a success.

Chapter 3
LITERATURE SURVEY

3.1 INTRODUCTION

In the literature review of the comparative study of feature selection techniques for disease detection, a lot of work on disease detection using data mining techniques has been found. Data mining techniques don’t just work for the heart disease data set but also other medical data sets such as blood pressure, diabetes, hepatitis etc. In this chapter the data mining techniques and classifiers relevant to our work are explained.

3.2 Filter method

Filter methods are generally used as a preprocessing step. The selection of features is independent of any machine learning algorithms. Instead, features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable. A statistical test provides a mechanism for making quantitative decisions about a process or processes. The intent is to determine whether there is enough evidence to "reject" a conjecture or hypothesis about the process. The conjecture is called the null hypothesis. For instance the statistical result (contained in the dataset) of (age) an attribute of the heart disease data set shows how it would affect the overall result [5].

Filter methods select features based on a performance measure regardless of the employed data modeling algorithm. Filter method finds the best features for the modeling algorithms to use. Filter methods can rank individual features or evaluate entire feature subsets. Developed measures for feature filtering into can be roughly classified as:

- Information
- Distance
- Consistency
- Similarity
- statistical measures

While there are many filter methods described in literature, a list of common methods is given in Table I, along with the appropriate references that provide details. Not all the filter features can be used for all classes of data mining tasks. Therefore, the filters are also classified depending on the task: classification, regression or clustering.

Univariate feature filters evaluate (and usually rank) a single feature, while multivariate filters evaluate an entire feature subset. Feature subset generation for multivariate filters depends on the search strategy. While there are many search strategies, there are four usual starting points for feature subset generation:

- forward selection
- backward elimination
- bidirectional selection
- heuristic feature subset selection

Forward selection typically starts with an empty feature set and then considers adding one or more features to the set from the supplied data. Backward elimination typically starts with the whole feature set and considers removing one or more features from the set. Bidirectional search starts from both sides - from an empty set and from the whole set, simultaneously considering larger and smaller feature subsets. Heuristic selection generates a starting subset based on a heuristic, and then explores it further.

The most common search strategies that can be used with multivariate filters can be categorized into exponential algorithms, sequential algorithms and randomized algorithms. Exponential algorithms evaluate a number of subsets that grows exponentially with the feature space size. Sequential algorithms add or remove features sequentially (one or few), which may lead to local minima. Random algorithms incorporate randomness into their search procedure, which avoids local minima. Common search strategies are shown in Table 2.

TABLE 2 COMMON FILTER METHODS FOR FEATURE SELECTION [8]

Name	Applicable to task	Study
Information gain	Classification	[6]
Gain ratio	Classification	[7]
Correlation	Regression	[8]
Chi-square	Classification	[7]
Inconsistency criterion	Classification	[9]

Minimum redundancy	Regression	[2]
First correlation based filter(FCBF)	Classification	[8]
Fisher score	Classification	[10]
Relief and ReliefF	Classification, regression	[11]
Spectral feature selection and laplacian score	Classification, clustering	[4]
Feature selection for sparse clustering	Clustering	[12]
Localized Feature selection based on Scatter Separability	Clustering	[13]
Multi-cluster Feature selection(MCFS)	Clustering	[4]
Feature weighting K-means	Clustering	[14]
ReliefC	Clustering	[15]

TABLE 2 SEARCH STRATEGIES FOR FEATURE SELECTION

Algorithm group	Algorithm name
Exponential	Exhaustive search Branch-and-bound
Sequential	Greedy forward selection or backward elimination Best-first Linear forward selection Floating forward or backward selection Beamsearch Race search
Randomized	Random generation Simulated annealing Evolutionary computation algorithms Scatter search

3.2.1 Information Gain

Information gain [13] is the amount of information that's gained by knowing the value of the attribute, which is the entropy of the distribution before the split minus the entropy of the distribution after it. The largest information gain is equivalent to the smallest entropy. So the information gain depends on two things: how much information was available before knowing the attribute value, and how much was available after.

Information gain (IG) is the measure of the amount of information in bits about the class prediction, if the only information available is the presence of a feature and the corresponding class distribution. It measures the expected reduction in entropy (uncertainty associated with a random feature). Given SX the set of training examples, xi the vector of i^{th} variables in this set, |Sxi=v|/|SX|the fraction of examples of the ith variable having value v

$$ig\left(s_{x.x_i}\right) = H(s_x) - \sum_{v=values(x_i)} \frac{|s_{xi=v}|}{|s_x|} \quad H(s_{x\ i=v})\ with\ entropy \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$H(S) = -p + (S)log2\ p + (S) - p - (S)log2\ p - (S) \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

p±(S) is the probability of a training example in the set S to be of the positive/negative class. We discretized continuous features using information theoretic binning.

3.2.2 Relief Attribute Eval

Relief is a feature selection method based on attribute estimation. Relief assigns a grade of relevance to each feature, and those features valued over a user given threshold are selected. Original Relief only handled Boolean concept problems, but extensions have been developed to fix these classification problems (Relief-F) [15] and regression problems (RRelief-F)[17].The extensions differ in the way that neighbor's are searched and in how the performance is evaluated. Relief-F generalizes the behavior of Relief to classification problems. It finds one nearest neighbor of E1 from every class. On these neighbor's, Relief evaluates the relevance of every feature f$\in F$ accumulating it into W[f]. the nearest neighbor from the same class is a hit H, and from different class a miss, M(c) of class c.at the end W[f] is divided by m to gt the average evaluation in [-1,1].

$$w[f] = w[f] - diff(f, E_1, H) + \sum_{c \neq class(E_1)} P(C) \times diff(f, E_{1,} M(C)) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

The diff (f,$E_{1,}E_2$) function calculates the grade in which the value of feature f are different in examples E_1 and E_2 as given in following equation, where value (f, E_1) denotes the value of

feature f on example E_1 and max (f) the maximum value f gets.The distance used considering nearest neighbor is the sum of differences, given by diff function, of all features.

$$diff(f, E_{1,}E_2) = \begin{cases} 0\ if\ value(f, E_1) = value(f, E_2) \\ 1\ otherwise \end{cases} | \frac{|value(f,E_1) - value(f,E_2)|}{\max(f) - \min(f)} \text{.....(4)}$$

In Relief-F, k-nearest neighbors are taken and their contribution is weighted according to their distance to E_1They contribute to positive and negative evaluation of features weightedeach by the diff function on the concept feature.

3.3 Wrapper method

In wrapper methods, a subset of features is used to train a model. Based on the inferences drawn from the previous model, features are added or removed from the subset. The problem is essentially reduced to a search problem. These methods are usually computationally very expensive [8].

Wrapper algorithm considers feature subsets by the quality of the performance on a modeling algorithm, which is taken as a black box evaluator. Therefore, classification is done by evaluating subsets based on the classifier performance while for clustering a wrapper evaluates subsets based on the performance of a clustering algorithm (such as K-means clustering). The evaluation is repeated for each subset, and the subset generation is dependent on the search strategy. Wrapper methods are much slower in finding sufficiently good subsets because they depend on the resource demands of the modeling algorithm. The feature subsets are also biased towards the modeling algorithm on which they were evaluated (even when using cross-validation). Therefore, for a reliable generalization error estimate, it is necessary that both an independent validation sample and another modeling algorithm are used after the final subset is found. On the other hand, it has been empirically proven that wrappers obtain subsets with better performance because the subsets are evaluated using a real modeling algorithm. Practically any combination of search strategy and modeling algorithm can be used as a wrapper, but wrappers are only feasible for greedy search strategies and fast modeling algorithms such as Naïve Bayes, linear SVM, and Extreme Learning Machines.

3.3.1 Correlation-based attributes

CFS evaluation method based on correlation-based attributes reduction is a heuristic algorithm [14]. It can evaluate the 'merit' of the subset of attributes. Its main consideration is the class prediction ability of single attribute and their correlations. The heuristic algorithm is based on the hypothesis below: attributes that belong to quality subset M F are highly correlated to class i C while the attributes themselves are irrelevant to each other. The irrelevant attributes in the subset are hardly related to the classification, so they can be ignored. The redundant attributes can also be eliminated for they are certain to have a correspondence to a high-correlated attribute. The acceptance degree of an attribute is due to its ability to predict the classification in the case library space while other attributes cannot. The evaluation function CFS of the subset is defined as follows

$$M_s = \frac{k\bar{r}_{cf}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \quad \text{.........(5)}$$

Where Ms is the heuristic 'merit' when the subset includes k attributes; $\overline{r_{cf}}$ is the attribute-classification correlative average value (F∈ S) and $\overline{r_{ff}}$attribute correlation average value.

For successive value data, the relativity between attributes can be calculated as follows:

$$r_{xy} = \frac{\sum xy}{n\sigma_x \, \sigma_y} \quad \text{.........(6)}$$

where x σ and y σ denote quadratic mean deviation of the attributes of successive value data.

If one of the two attributes is successive and the other is discrete, the relativity can be calculated as follows

$$r_{xy} = \sum_{i=1}^{k} p(X=x_i) r_{x_{bi}\,Y} \quad \text{.........(7)}$$

For x_{bi}, if x=$x_{i,}$ then $x_{bi=1,}$ else $x_{bi=0.}$

If both attributes are discrete, the relativity can be calculated as follows:

$R_{xy=}{}_{xy=\sum_{I=1}^{k}\sum_{j=1}^{l}p(X=x_{i,}\ y=y_j)rx_{biy_{bj}}}$ ……(8)

According to the above formulations, correlation of the attributes can be calculated no matter it is discrete or successive.

3.3.2 Wrapper Subset Eval

Evaluates attribute sets by using a learning scheme. Cross validation is used to estimate the accuracy of the learning scheme for a set of attributes [20].

The options provided by this attribute evaluator are:

Classifier: Classifier to use for estimating the accuracy of subsets

Folds: Number of x-value folds to use when estimating subset accuracy.

Seed: Seed to use for randomly generating x-value splits.

Threshold: Repeat x-value if standard deviation of mean exceeds this value

3.4 Embedded Method

Feature selection is a part of the learning procedure and specific to given learning machines. Here, optimal feature subset selection is built into the classifier construction, and viewed as a search in the combined space of feature subset and hypotheses. Some examples of this method include, classification trees, random forests, feature selection using weight vector of Support Vector Machines (SVM) and methods based on regularization techniques. These methods are less computationally intensive. The design of embedded feature selection techniques depend on a specific a learning algorithm. [7]

3.4.1 SVM Subset Eval

The Support Vector Machine (SVM) is a classification technique based on statistical learning theor that was applied with great success in many challenging non-linear classification problems and was successfully applied to large data sets. The SVM algorithm finds a hyperplane that optimally splits the training set. The optimal hyperplane can be distinguished by the maximum margin of separation between all training points and the hyperplane [18.19].

SVMAttributeEval evaluates the worth of an attribute by using an SVM classifier. Options provided by the evaluator are:

Attributes To Eliminate PerIteration: Constant rate of attribute elimination

Complexity Parameter: C complexity parameter to pass to the SVM

Epsilon Parameter: P epsilon parameter to pass to the SVM

Filter Type: filtering used by the SVM

Percent Threshold: Threshold below which percent elimination reverts to constant elimination.

Percent to Eliminate PerIteration: Percent rate of attribute elimination.

Tolerance Parameter: T tolerance parameter to pass to the SVM

3.5 **Ranker Search Method**

Ranks attributes by their individual evaluations. Use in conjunction with attribute evaluators [21] Options provided by the search method are:

Generate Ranking: A constant option. Ranker is only capable of generating attribute rankings.

Num To Select: Specify the number of attributes to retain. The default value (-1) indicates that all attributes are to be retained. Use either this option or a threshold to reduce the attribute set.

Start Set: Specify a set of attributes to ignore. When generating the ranking, Ranker will not evaluate the attributes in this list. This is specified as a comma separated list off attribute indexes starting at 1. It can include ranges. Eg. 1,2,5-9,17.

Threshold: Set threshold by which attributes can be discarded. Default value results in no attributes being discarded. Use either this option or num To Select to reduce the attribute set.

3.6 I **IWSS (Incremental wrapper subset selection)**

This attribute selector is specially designed to handle high-dimensional datasets. It first creates a ranking of attributes based on the selected metric, and then it runs an Incremental Wrapper Subset Selection over the ranking (linear complexity) by selecting attributes (using the WrapperSubsetEval class) which improve the performance for a given minimum number of folds out of the folds of the wrapper cross-validation. It contains the theta option which permits to tune an early stopping (sublinear complexity). It contains the replace Selection option, which tests at

each step of the incremental search swapping a selected attribute by the current candidate, this reduces the mean number of selected attributes without decreasing performance but it increases the linear complexity to quadratic.

3.7 **J48 Classifier**

J48 is an open source java implementation of c4.5 algorithm. J48 builds decision trees from a set of training data in the same way as c4.5 and id3.decision trees are built using the concept of information entropy. For every node of the tree j48 algorithm chooses attributes of the data that splits the set of sample data into subsets enriched in one class or the other most effectively. The splitting is done based on the criterion of normalized information gain. The attributes with the highest information gain is chosen for the decision making process. The algorithm then recurs on the smallest sub list [9].

Some base of this algorithm is,

- Handling both continuous and discrete attributes: In order to handle continuous attributes, j48 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it.
- Handling training data with missing attribute values: j48 allows attribute values to be marked as missing. Missing attribute values are simply not used in gain and entropy calculations.
- Pruning trees after creation: j48 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.

Chapter 4
METHODOLOGY

4.1 Introduction

This performance evaluation is done among the common feature selection methods filter, wrapper and embedded. Under the filter method InfoGaniAttributeEval and ReliefAttributeEval are used under the wrapper method CfsAttributeEval and WrapperSubsetEval are used and under the embedded method SvmAttributeEval is used. The same heart disease data set is used to evaluate all the techniques. For classification j48 is used as the common classifier. All the implementation is done on weka3.8.

4.2 Preprocessing

In the field of data mining it is crucial to have a dependable data set. Data needs to be checked properly and carefully to make sure that the data is good for the job because any dirty data can skew the results and give misleading conclusion. So every field of the data should be investigated to find the nature, relation and errors of the data. The heart disease dataset used here has no missing values and is collected from a reliable source. The data is stored in .arff format because weka supports this format. Before starting the modeling process data is checked to verify that there are actually no missing values before modeling. The loaded data is shown in figure 1.

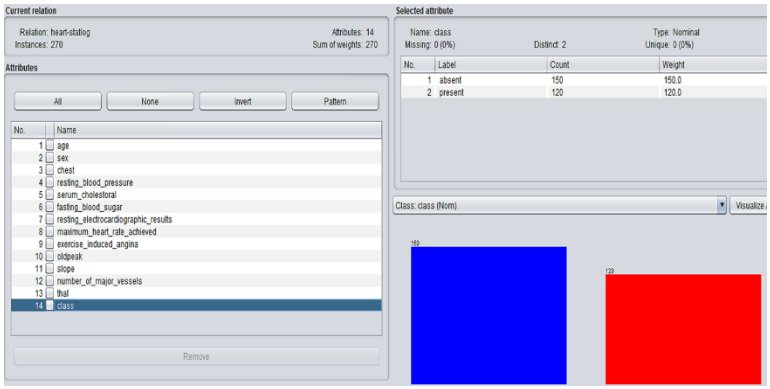

FIGURE 1: Preprocessing of dataset

4.3 Classification and attribute selection

For classification first data is loaded to weka and classification is performed. Classification is performed by going to the classify panel. This panel contains all the classification algorithms (such as j48, ranker) in the weka tool. Here j48 classifier is used for classification. This is done to check the following:

- Accuracy: The degree to which the result of a measurement, calculation, or specification conforms to the correct value or a standard.
- Precision: In pattern recognition and information retrieval binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant.
- Recall: Is the fraction of relevant instances that are retrieved.
- F-measure: F1 score (also F-score or F-measure) is a measure of a test's accuracy.

After classification is performed the attribute selection is done. The default threshold of the attributes used is considered o.5 out of the range (0 to 1)but the default threshold is not often optimal [17]. The process is shown below:

4.4 Filter method

Under the filter method data two techniques are used for the task of performance evaluation InfoGainAttributeEval and ReliefAttributeEval.

4.4.1 InfoGainAttributeEval

At the current state the heart disease data has an accuracy of 76.66% and an average precision: 0.78, recall: 0.76, f-measure: 0.76.

```
Size of the tree :      35


Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         207               76.6667 %
Incorrectly Classified Instances        63               23.3333 %
Kappa statistic                          0.5271
Mean absolute error                      0.274
Root mean squared error                  0.4601
Relative absolute error                 55.4778 %
Root relative squared error             92.5962 %
Total Number of Instances              270

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.793    0.267    0.788      0.793    0.791      0.527    0.744     0.737     absent
                 0.733    0.207    0.739      0.733    0.736      0.527    0.744     0.641     present
Weighted Avg.    0.767    0.240    0.766      0.767    0.767      0.527    0.744     0.694

=== Confusion Matrix ===

   a   b   <-- classified as
 119  31 |   a = absent
  32  88 |   b = present
```

FIGURE 2: Result of InfiGain attribute eval without attribute evaluation

Now attribute section is done on the data. This is done by going to the select attribute panel. This panel contains the attribute evaluators and the search methods. From here the InfoGainAttributeEval is chosen as the attribute evaluator and ranker is used as the search method after selection is performed the results show the attributes of the heart disease data rearranged on the basis of importance. Shown in fig [3]

```
                 exercise_induced_angina
                 number_of_major_vessels
                 thal
                 class
Evaluation mode:     evaluate on all training data


=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 6 class):
        Information Gain Ranking Filter

Ranked attributes:
 0.203   5 thal
 0.19    1 chest
 0.166   4 number_of_major_vessels
 0.13    3 exercise_induced_angina
 0.12    2 maximum_heart_rate_achieved

Selected attributes: 5,1,4,3,2 : 5
```

FIGURE 3: Selected attribute for InfoGainAttribute eval

On the basis of the ranked attributes the source data attributes are altered. This results in a attribute set more related to the task of heart disease detection. Now again classification is done to get an improved accuracy of 81.11% and average precision: 0.81, recall: 0.81, f-measure: 0.81

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         219               81.1111 %
Incorrectly Classified Instances        51               18.8889 %
Kappa statistic                          0.6172
Mean absolute error                      0.2734
Root mean squared error                  0.4107
Relative absolute error                 55.3607 %
Root relative squared error             82.6572 %
Total Number of Instances              270

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.833    0.217    0.828      0.833    0.831      0.617    0.779     0.764     absent
                 0.783    0.167    0.790      0.783    0.787      0.617    0.779     0.672     present
Weighted Avg.    0.811    0.194    0.811      0.811    0.811      0.617    0.779     0.723

=== Confusion Matrix ===

   a   b   <-- classified as
 125  25 |   a = absent
  26  94 |   b = present
```

FIGURE:4 Result of InfiGain attribute Eval with attribute evaluation

4.4.2 ReliefAttributeEval

At this stage state the heart disease data has an accuracy of 76.667% and an average precision: 0.786, recall: 0.767, f-measure: 0.767.

```
Size of the tree :      35


Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         207               76.6667 %
Incorrectly Classified Instances        63               23.3333 %
Kappa statistic                          0.5271
Mean absolute error                      0.274
Root mean squared error                  0.4601
Relative absolute error                 55.4778 %
Root relative squared error             92.5962 %
Total Number of Instances              270

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.793    0.267    0.788      0.793    0.791      0.527    0.744     0.737     absent
                 0.733    0.207    0.739      0.733    0.736      0.527    0.744     0.641     present
Weighted Avg.    0.767    0.240    0.766      0.767    0.767      0.527    0.744     0.694

=== Confusion Matrix ===

   a   b   <-- classified as
 119  31 |   a = absent
  32  88 |   b = present
```

FIGURE:5 Result of Relief attribute eval without attribute evaluation

Now attribute section is done on the data. This is done by going to the select attribute panel and choosing the ReliefAttributeEval as the attribute evaluator and ranker is used as the search method after selection is performed the results show the attributes of the heart disease data rearranged on the basis of importance which are shown in fig [6].

```
Evaluation mode:     evaluate on all training data


=== Attribute Selection on all input data ===

Search Method:
         Best first.
         Start set: no attributes
         Search direction: forward
         Stale search after 5 node expansions
         Total number of subsets evaluated: 17
         Merit of best subset found:     0.316

Attribute Subset Evaluator (supervised, Class (nominal): 6 class):
         CFS Subset Evaluator
         Including locally predictive attributes

Selected attributes: 1,2,3,4,5 : 5
                     chest
                     resting_electrocardiographic_results
                     slope
                     number_of_major_vessels
                     thal
```

FIGURE:6 Selected Attribute for Relief Attribute Eval

Now by removing the irrelevant attribute from the attribute set the dataset gets streamlined for the task and gives a better accurate result which as an accuracy of 82.96% and average precision of 0.83,recall 0.83, f-measure 0.82

```
Size of the tree :      13


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         224               82.963  %
Incorrectly Classified Instances        46               17.037  %
Kappa statistic                          0.6521
Mean absolute error                      0.253
Root mean squared error                  0.3788
Relative absolute error                 51.2208 %
Root relative squared error             76.2229 %
Total Number of Instances              270

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.880    0.233    0.825      0.880    0.852      0.654    0.830     0.808     absent
                 0.767    0.120    0.836      0.767    0.800      0.654    0.830     0.778     present
Weighted Avg.    0.830    0.183    0.830      0.830    0.829      0.654    0.830     0.794

=== Confusion Matrix ===

   a   b   <-- classified as
 132  18 |   a = absent
  28  92 |   b = present
```

FIGURE 7: Result of Relief Attribute Eval with attribute evaluation

4.5 Wrapper Method

Cfs Subset Eval and Wrapper Subset Eval are the two attribute evaluators used under the wrapper method.

4.5.1 CfsSubsetEval:

At his state this heart disease data has an accuracy of 76.66% and an average precision: 0.786, recall: 0.76, f-measure: 0.76.

```
Size of the tree :      35

Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         207               76.6667 %
Incorrectly Classified Instances        63               23.3333 %
Kappa statistic                          0.5271
Mean absolute error                      0.274
Root mean squared error                  0.4601
Relative absolute error                 55.4778 %
Root relative squared error             92.5962 %
Total Number of Instances              270

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.793     0.267     0.788       0.793   0.791      0.527    0.744     0.737     absent
                 0.733     0.207     0.739       0.733   0.736      0.527    0.744     0.641     present
Weighted Avg.    0.767     0.240     0.766       0.767   0.767      0.527    0.744     0.694

=== Confusion Matrix ===

   a   b   <-- classified as
 119  31 |   a = absent
  32  88 |   b = present
```

FIGURE 8: Result cfs subset eval without attribute evaluation

Now attribute section is done on the data. This is done by going to the select attribute panel and choosing the CfsSubsetEval as the attribute evaluator and ranking is used as the search method after selection is performed the results show the attributes of the heart disease data rearranged on the basis of importance. Selected attributes are shown in fig [9].

```
--- Attribute selection on all input data ---

Search Method:
Selected attributes: 12 2 11 8 7 9
Total Search Time in milliseconds: 79.0
Reranking Time in milliseconds: 0.0
Blocks searched: 1
Attribute Evaluator during search:      CFS Subset Evaluator
        Including locally predictive attributes

Options of RerankingSearch:
-method 0 -blockSize 20 -rankingMeasure 0 -search weka.attributeSelection.GreedyStepwise -T -1.7976931348623157E308 -N -1 -num-slots 1
Attribute Subset Evaluator (supervised, Class (nominal): 14 class):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 3,7,8,9,10,12,13 : 7
                     chest
                     resting_electrocardiographic_results
                     maximum_heart_rate_achieved
                     exercise_induced_angina
                     oldpeak
                     number_of_major_vessels
                     thal
```

FIGURE 9: Selected Attribute for cfs subset Eval

Now taking the attributes selected by the attribute evaluator and removing the unimportant attributes the classification is done again on the now more compact data set which gives a better

performance and accuracy than before. These improved results are accuracy: 81.11% average Precision: 0.81, recall: 0.81 and f-measure: 0.81.

```
Size of the tree :      21

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         219               81.1111 %
Incorrectly Classified Instances        51               18.8889 %
Kappa statistic                          0.6153
Mean absolute error                      0.2485
Root mean squared error                  0.4064
Relative absolute error                 50.3256 %
Root relative squared error             81.7825 %
Total Number of Instances              270

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.853     0.242     0.815       0.853   0.834      0.616    0.809     0.780     absent
                 0.758     0.147     0.805       0.758   0.781      0.616    0.809     0.745     present
Weighted Avg.    0.811     0.199     0.811       0.811   0.810      0.616    0.809     0.764

=== Confusion Matrix ===

   a   b   <-- classified as
 128  22 |   a = absent
  29  91 |   b = present
```

FIGURE10: Result of cfs Subset Eval with attribute evaluation

4.5.2 WrapperSubsetEval

Before attribute selection this heart disease data has an accuracy of 76.66% and an average precision: 0.78, recall: 0.76, f-measure: 0.76.

```
Size of the tree :      35

Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         207               76.6667 %
Incorrectly Classified Instances        63               23.3333 %
Kappa statistic                          0.5271
Mean absolute error                      0.274
Root mean squared error                  0.4601
Relative absolute error                 55.4778 %
Root relative squared error             92.5962 %
Total Number of Instances              270

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.793     0.267     0.788       0.793   0.791      0.527    0.744     0.737     absent
                 0.733     0.207     0.739       0.733   0.736      0.527    0.744     0.641     present
Weighted Avg.    0.767     0.240     0.766       0.767   0.767      0.527    0.744     0.694

=== Confusion Matrix ===

   a   b   <-- classified as
 119  31 |   a = absent
  32  88 |   b = present
```

FIGURE 11: Result of WrappersubsetEval without attribute evaluation

Now on the select attribute panel WrapperSubsetEval is chosen as the attribute evaluator and IWSS is chosen as the search method. After the selection process is done it shows the attributes of the heart disease data rearranged on the basis of importance. Selected attributes are shown in fig 12, 13.

```
              number_of_major_vessels
              thal
              class
Evaluation mode:    evaluate on all training data


=== Attribute Selection on all input data ===

Search Method:
Incremental Wrapper Subset Selection (IWSS):
Selected Attributes: 1
Merit of best subset found: 0.5576519916142557
Metric used for creating the ranking: weka.attributeSelection.SymmetricalUncertAttributeEval

Attribute Subset Evaluator (supervised, Class (nominal): 14 class):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.rules.ZeroR
        Scheme options:
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 13 : 1
                     thal
```

FIGURE 12: Selecting process of WrapperSubsetEval

FIGURE 13:Selected Attribute for WrapperSubsetEval

Now on the basis of the attributes selected the unimportant attributes are removed to get an attribute set best for the job and classification is done again to check the improvement compared to the previous attribute set. This attribute set has an accuracy of 81.48% and average precision: 0.81, Recall: 0.81, f-measure: 0.81.

```
Size of the tree :      19


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         220               81.4815 %
Incorrectly Classified Instances        50               18.5185 %
Kappa statistic                          0.6237
Mean absolute error                      0.2583
Root mean squared error                  0.3999
Relative absolute error                 52.3101 %
Root relative squared error             80.4869 %
Total Number of Instances              270

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.847    0.225    0.825      0.847    0.836      0.624    0.807     0.791     absent
                 0.775    0.153    0.802      0.775    0.788      0.624    0.807     0.725     present
Weighted Avg.    0.815    0.193    0.814      0.815    0.814      0.624    0.807     0.762

=== Confusion Matrix ===

   a   b   <-- classified as
 127  23 |   a = absent
  27  93 |   b = present
```

Figure 14:Result of WrapperSubsetEval with attribute evaluation

5.6 Embeded Method:

Here SVM Attribute eval is used as feature selection method.

5.6.1 SVM Attribute Eval

In this process data is loaded into the weka platform through the preprocessing panel. Every attribute checked to make sure there are no missing values and then classification is done by going to the classify panel and selecting the j48 classifier which shows the current accuracy of the provided heart disease data before any attribute selection is performed. At this state the heart disease data has an accuracy of 76.6667% and an average precision of 0.766, recall 0.767 and f-measure 0.767.

```
Size of the tree :      35


Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         207               76.6667 %
Incorrectly Classified Instances        63               23.3333 %
Kappa statistic                          0.5271
Mean absolute error                      0.274
Root mean squared error                  0.4601
Relative absolute error                 55.4778 %
Root relative squared error             92.5962 %
Total Number of Instances              270

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.793    0.267    0.788      0.793    0.791      0.527    0.744     0.737     absent
                 0.733    0.207    0.739      0.733    0.736      0.527    0.744     0.641     present
Weighted Avg.    0.767    0.240    0.766      0.767    0.767      0.527    0.744     0.694

=== Confusion Matrix ===

   a   b   <-- classified as
 119  31 |   a = absent
  32  88 |   b = present
```

FIGURE15: Result of SVMSubsetEval without attribute evaluation

Now attribute selection is performed on the data. This is done by going to the select attribute panel and choosing SVMAttributeEval as the attribute evaluator and ranker as the search the search method which is built in to the evaluator. After evaluation is finished the new rearranged attributes are shown this is done based on the importance of the attributes related to the task.

```
                 oldpeak
                 number_of_major_vessels
                 thal
                 class
Evaluation mode:     evaluate on all training data


=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 6 class):
        SVM feature evaluator

Ranked attributes:
  5   4 number_of_major_vessels
  4   3 oldpeak
  3   1 chest
  2   5 thal
  1   2 maximum_heart_rate_achieved

Selected attributes: 4,3,1,5,2 : 5
```

FIGURE 16: Selected Attributes of SVMSubsetEval

Now taking this attributes into account and removing the unimportant attributes from the evaluation process classification is done again. Which provides an improved accuracy of 81.4815%,average precision of 0.814, recall 0.815, f-measure 0.814.

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         221               81.8519 %
Incorrectly Classified Instances        49               18.1481 %
Kappa statistic                          0.6297
Mean absolute error                      0.2501
Root mean squared error                  0.3949
Relative absolute error                 50.6426 %
Root relative squared error             79.4761 %
Total Number of Instances              270

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.867    0.242    0.818      0.867    0.841      0.631    0.817     0.793     absent
                 0.758    0.133    0.820      0.758    0.788      0.631    0.817     0.747     present
Weighted Avg.    0.819    0.194    0.819      0.819    0.818      0.631    0.817     0.772

=== Confusion Matrix ===

   a   b   <-- classified as
 130  20 |   a = absent
  29  91 |   b = present
```

FIGURE17: Result of SVMSubsetEval with attribute evaluation

Chapter 5
EXPERIMENTAL RESULT

5.1 Introduction

The performance result using weka3.8 are described in this chapter. For performing the experiment intel core i7 quad core @3.2GHz 6th generation processor and 8gb single channel DDR3 primary memory was used. In this chapter the accuracy, precision, recall and F-Measure of the compared attribute evaluator are shown .Heart diseases dataset was used to compare different feature selection methods for the prediction of disease risks. Five attribute evaluator are used to to evaluate their performance. The attribute evaluators are :

- Info Gain Attribute Eval
- ReliefF Attribute Eval
- Cfs Subset Eval
- Wrapper Subset Eval
- Svm subset Eval

Accuracy:

It is a ratio of number of correctly classified instances to the total number of instances. [18]

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ positive + false\ Negative + True\ Negative}$$

Accuracy of attribute selection methods are shown in Table 4, 6, 8, 10, 12.

Precision:

Precision is a proportion of predicted positives which are actual positive. [19]

$$Precision = \frac{TP}{TP+FP}$$

Recall:

Recall is a proportion of positives which are predicted positive

$$Recall = \frac{TP}{TP+FN}$$

F-Measure:

F-Measure is the harmonic mean of precision and recall.

$$F - Measure = \frac{2TP}{2TP + FN + FP}$$

Precision, recall and F-Measure are shown in table 5, 7, 9, 11, 13

5.2 Experiment Result of Filter Method

5.2.1 InfoGainAttributeEval

At first the performance accuracy of InfoGainAttributeEval is shown. It is found that out of the total 270 instances 219 instances have been classified correctly giving an accuracy of 81.11% and 51 instances have been classified incorrectly giving an accuracy of 18.89% shown in table [4]. the precision, recall and f-measure of this evaluator is shown in table 5.

Table[4]: InfoGainAttributeEval accuracy

Partition	Instances	Accuracy
Correct	219	81.11%
Wrong	51	18.88%
Total	270	

Table[5]: precision, recall, f-measure of InfoGainAttributeEval

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Absent	0.833	0.217	0.828	0.833	0.831
Present	0.783	0.167	0.790	0.783	0.787
Weighted Avg.	0.811	0.194	0.811	0.811	0.811

5.2.2ReliefFAttributeEval

At first the performance accuracy of ReliefFAttributeEvalis shown. It is found that out of the total 270 instances 224instances have been classified correctly giving an accuracy of 82.96%and 46instances have been classified incorrectly giving an accuracy of 17.04%shown in table 6. The precision, recall and f-measure of this evaluator is shown in table 7.

Table[6]: ReliefFAttributeEval accuracy

Partition	Instances	Accuracy
Correct	224	82.96%
Wrong	46	17.04%
Total	270	

Table[7]: precision, recall, f-measure of ReliefFAttributeEval

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Absent	0.880	0.233	0.825	0.880	0.852
Present	0.767	0.120	0.836	0.767	0.800
Weighted Avg.	0.830	0.183	0.830	0.830	0.829

5.3 Experimental Result Of Wrapper Method

5.3.1 CfsSubsetEval

At first the performance accuracy of CfsSubsetEval is shown. It is found that out of the total 270 instances 219instances have been classified correctly giving an accuracy of 81.11%and 51instances have been classified incorrectly giving an accuracy of 18.89%shown in table 8. the precision, recall and f-measure of this evaluator is shown in table 9.

Table[8]: Accuracy of cfsAttributeEval

Partition	Instances	Accuracy
Correct	219	81.11%
Wrong	51	18.89%
Total	270	

Table[9]: precision, recall, f-measure of CfsSubsetEval

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Absent	0.853	0.242	0.815	0.853	0.834
Present	0.758	0.147	0.805	0.758	0.781
Weighted Avg.	0.811	0.199	0.811	0.811	0.810

5.3.2 WrapperSubsetEval

Here the performance accuracy of WrapperSubsetEvalis shown. It is found that out of the total 270 instances 220instances have been classified correctly giving an accuracy of 81.48%and 50instances have been classified incorrectly giving an accuracy of 18.52%shown in table 10. the precision, recall and f-measure of this evaluator is shown in table 11.

Table[10]:Accuracy of WrapperSubsetEval

Partition	Instances	Accuracy
Correct	220	81.48%
Wrong	50	18.52%
Total	270	

Table[11]: Precision, recall, f-measure of CfsSubsetEval

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Absent	0.847	0.225	0.825	0.847	0.836
Present	0.775	0.153	0.802	0.775	0.788
Weighted Avg.	0.815	0.193	0.814	0.815	0.814

5.4 Experimental Result of embedded Method

5.4.1 SvmAttributeEval

Here the performance accuracy of SvmAttributeEval is shown. It is found that out of the total 270 instances 221instances have been classified correctly giving an accuracy 81.85% and 49instances have been classified incorrectly giving an accuracy of 18.15%shown in table 12. The precision, recall and f-measure of this evaluator is shown in table 13.

Table[12]: Accuracy of SvmAttributeEval

Partition	Instances	Accuracy
Correct	221	81.85%
Wrong	49	18.15%
Total	270	

Table [13]: precision, recall, f-measure of SvmSubsetEval

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Absent	0.867	0.242	0.818	0.867	0.841
Present	0.758	0.133	0.820	0.758	0.788
Weighted Avg.	0.819	0.194	0.819	0.819	0.818

Table [14]: Comparison of Attribute Evaluation Methods

Heart diseases data set						
Method	Attribute Evaluator + Search method	Classifiers	Accuracy	Precision	Recall	F-Measure
Filter method	InfoGainAttributeEval + Ranker	J48	81.11%	0.811	0.811	0.811
	ReliefFAttributeEval + Ranker		**82.96%**	**0.830**	**0.830**	**0.829**
Wrapper method	CfsSubsetEval + RankerSearch		81.1%	0.811	0.811	0.810
	WrapperSubsetEval + IWSS		81.48%	0.814	0.815	0.814
Embedded method	SVMAttributeEval + Ranker		81.85%	0.819	0.819	0.818

From table 14, it is observed that among the 5 attribute selection methods compared ReliefFAttributeEval gives the best accuracy (82.96%). It has a higher precision of 0.83. Out of the total 270 instances it classified 224 correctly as a result it has the lowest number of misclassified instances table 6. Thus ReliefFAttributeEval is the best attribute evaluation method for heart disease dataset.

5.5 Summary

In this chapter comparison between feature selection methods are shown. It is found that RelieffAttributeEval gives the best accuracy than other feature selection methods. Thus from the analysis, it is observed that ReliefFAttributeEval is the best attribute evaluator. The next chapter is the conclusion chapter which has shortly discus whole report.

Chapter 6
Conclusion and Future work

6.1 Conclusion

Data mining technique is used to mine important information from vast amount of data. Data mining is crucial for data organization. Feature selection is an integral part of data mining. This paper compares some of the feature selection techniques. In literature survey the methods used in this paper as well as the data used are descried. Methodology chapter of this report contains the process by which the feature selection methods are compared. Attribute used for comparison are InfoGainAttribueEval, ReliefFAttributeEval, CfsSubsetEval, WrapperSubsetEval and SVMAttributeEval. For classification process j48 is used as the common classifier. The results of the comparison is shown in the fifth chapter of this report here the accuracy, recall, precision and recall are presented. Here it is found that the ReliefFAttributeEval gives the best accuracy 82.96% with the most number of correctly classified instances. Thus the final outcome of this project is that ReliefFAttributeEval is the best attribute evaluator for the heart disease data set used in this evaluation.

6.2 Future work

In this report we compared feature selection methods to find out which method provides the best accuracy for the heart disease data set. But the methods are not limited to just one data set they can be used to get the optimal accuracy out of any data set. So the future work would be to evaluate the feature selection methods so that they would provide better accuracy for any data set.

REFERENCE

1. http://www.periyaruniversity.ac.in/ijcii/issue/Vol4No2September2014/IJCII-4-2-144.pdf
2. https://archive.ics.uci.edu/ml/datasets/Statlog+(Heart)
3. http://www.revespcardiol.org/en/congenital-heart-disease-present-situation/articulo/13082912/
4. https://www.heart.org/idc/groups/ahamah-public/@wcm/@sop/@smd/documents/downloadable/ucm_470704.pdf
5. http://ieeexplore.ieee.org/abstract/document/7160458/
6. https://archive.ics.uci.edu/ml/datasets/Statlog+(Heart)
7. http://www.periyaruniversity.ac.in/ijcii/issue/Vol4No2September2014/IJCII-4-2-144.pdf
8. http://ieeexplore.ieee.org/abstract/document/7160458/
9. Ross Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.
10. https://en.wikipedia.org/wiki/C4.5_algorithm
11. https://en.wikipedia.org/wiki/Weka_(machine_learning)
12. Mitchell. Machine Learning. McGraw-Hill, New York, 1997.
13. Fayyad and K. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In Proc˙10th Int ˙Conf ˙Machine Learning, pages 194–201, 1993.
14. Hall, M. A.:Correlation-based feature reduction for discrete and numeric class machine learning, Proc. of the 17th International Conference on Machine Learning(2000)
15. Igor Kononenko. Estimating attributes: Analysis and extensions of RELIEF. Inn European Conference on Machine Learning, pages 171–182, 1994
16.] M. RobnikSikonja and I. Kononenko. An adaptation of relief for attribute estimation in
17. regression. In Morgan Kaufmann, editor, Machine Learning: Proceedings of the Fourteenth International Conference, pages 296–304, 1997
18. For more information see: Guyon, I., Weston, J., Barnhill, S., &Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46, 389-422
19. For more information see: Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46, 389-422
20. Page http://www.cs.tufts.edu/~ablumer/weka/doc/weka.attributeSelection.Ranker.html
21. http://weka.sourceforge.net/packageMetaData/IWSS/index.html
22. http://www.ijmlc.org/vol5/517-C002.pdf
23. http://research.ijcaonline.org/volume111/number5/pxc3901189.pdf9accuracy0
24. http://digitalcommons.wku.edu/cgi/viewcontent.cgi?article=1005&context=comp_sci