Application of Algorithm C 4.5-based Attribute Selection with Backward Elimination in Predicting Heart Disease

Ikram Tawary

***Abstract*— World Health Organization (WHO) noted that approximately 12 million people died of cardiovascular disease and about 32 million have heart attacks and strokes each year. Data mining on health have great potential to explore the hidden patterns within a dataset from the domain of health, including heart disease. C4.5 classification algorithm is the most simple, easy diimplemntasikan. However, the algorithm C4.5 still have weaknesses in handling high-dimensional data in. This study will aim to implement the algorithm C4.5 with the selection of attributes so as to reduce the dimensionality of the data, and identify features in the data set with C4.5 algorithm method. From this research, conducted models created with C4.5 algorithm itself already has a fairly good accuracy that is equal to 76.66% with the selection process attributes by C4.5 algorithms, models created can be increased again to 77.40% in the classification of heart disease.**

Keywords: Backforwar Chaining, Expert System, Heart Disease

BACKGROUND

The World Health Organization or the World Health Organization (WHO) headquarters in Geneva, Switzerland, and was established on 7 April 1948. The organization is part of the UN organization dedicated to detecting; preventing and controlling the disease in the world have reported that heart disease is the leading cause of death in high-income countries and low. The World Health Organization has estimated figure of 12 million deaths each year are caused by heart disease. Half of the deaths in the United States and other developed countries are caused by heart

disease. Diagnosis of the disease is regarded as a cumbersome task performed by the doctor to take a decision quickly, accurately and efficiently. In doctors usually diagnose diseases often make decisions based on intuition and experience of the use of the database. This practice is usually inaccurate cause unwanted errors, excessive medical costs as well as affecting the quality of services provided to patients. By him that with the database in data mining are described in this study is expected will be able to help a physician or the entire doctor came out of the difficulty in diagnosing a disease, especially heart disease. Heart disease is the leading cause of death in the world over the last 10 years.

European Public Health Alliance (European Public Health Alliance) reported that heart attacks and other circulatory diseases reached 41% of all deaths in countries Eropa.United Nations Economic And Social Commission For Asia And The Pacific (UNESCAP) reports that a fifth Asian countries, most lives are lost to non-communicable diseases such as heart disease, cancer and diabetes. Australian Bureau of Statistics reported that the heart and circulatory disease is the leading cause of death in Australia amounted to 33.7% of mortality rate. Statistics South Africa reported that heart disease and circulatory third leading cause of death in Africa.

C4.5 algorithms combine evaluation methodology naturally (Anbarasi, Anupriya, & Iyengar, 2010, p. 5372). C4.5 algorithms can also be used to reduce the actual data in order to obtain optimal subset in predicting heart disease. Selection attribute is a process of selecting a subset of the features of the original features. The purpose of the selection of attributes is to identify features in the data set is equally important, then discard other features such as information that is irrelevant and redundant (Maimon & Rokach, 2010, p. 84). With the selection

Manuscript received Jul, 2017.
Ikram Tawary, Program Studi Magister Teknik Informatika, STMIK Amikom Yogyakarta, Jl. Ring Road Utara, Condong Catur, Depok, Kec. Sleman, D.I Yogyakarta – Indonesia, Telp: (0274) 884201, Fax : (0274) 884208.

of attributes can reduce the dimensionality of data, this allows more effective in operation faster than some of the data mining algorithms. On the other hand, the selection of attributes takes up costs are quite expensive and also contrary to the initial asumsis, that all information that is required attributes to achieve maximum accuracy.

This research is expected to provide meaningful input for doctors to diagnose heart disease early on, in order to make the handling early.

Several studies of heart disease prediction algorithms using C.45 among other research topics prediction refinement of heart disease by using a genetic algorithm optimization. The purpose of this study is to predict disease by reducing the number of attributes that exist in order to achieve more accurate results involving 14 attributes early in predicting heart disease using a genetic algorithm to determine the attributes that contribute towards the diagnosis of heart disease. Indirectly reduce the number of required tests taken from patients, from 12 attribute reduced to 6 attribute in the search for genetic using three classifiers such as Naive Bayes, Classification by clustering and Decision Tree. The results of these studies Decision Tree method produces the best value is 99.2%, 96.5% and Naïve Bayes Clasication Via Clustering 88.3%.

In search of genetically produced six attributes from 11 attributes.

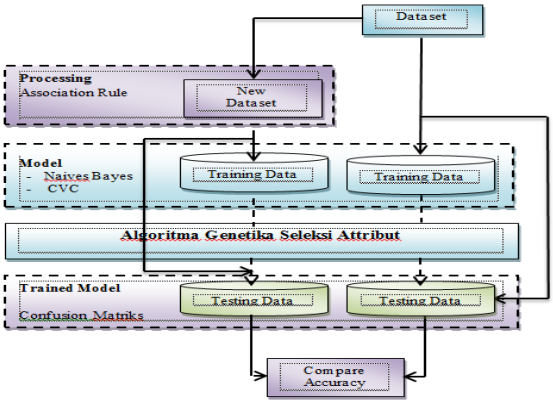

Figure 2.1 The proposed model Anbarasi.

Data Analysis Methode

Based on a review study, this study will use an attribute-based selection algorithm is backward elimination in the prediction of cardiovascular disease compared with C4.5 algorithm by reducing the number of attributes that too much. Previous studies conducted by Dr. Srinivas Raghavendra Rao and Dr. A. Govardhan in 2010 using UCI Heart Disease Data Set with Naives Bayes algorithm produces an accuracy of 84.14% with the original dataset for 14 attributes and add one attribute that is diagnosed as a reference in the classification. Anbarasi, E. Anupriya and N.CH.SNInyegar in 2010 using the UCI Heart Disease Data Set with Naives Bayes algorithm produces an accuracy of 96.5% and 99.2% algotitma Decision Treesebesar the original dataset by 12 attributes and reduced to six attributes in searches genetics using three classification.

Result of Research

Data mining is defined as the process of finding patterns in data. The process is automated (usually) semi-automatic pattern is found to be meaningful in the sense of causing some advantages, usually economical. Data is always used in large size. Data mining is the process of finding meaningful new correlations, patterns and trends by sifting through large amounts of data stored in the repository, using the technology of reasoning patterns and techniques of statistics and mathematics. The term data mining has the nature of a discipline whose primary goal is to discover, explore, or mine the knowledge of the data or information that we have. There are a few major role in data mining, among others:

a. Estimation

The estimation algorithm used is: Linear Regression, Neural Networks, Support Vector Machine. The estimation algorithm is similar to the classification algorithms, but the target variable is in the form of a numerical number and not a categorical (nominal). The model is built from the data with a complete record, which provides the value of a variable as a predictor, and then the estimated value of the target variable is determined based on the value of the predictor variables. The determination of the policy or a value to the process that will be done. Estimates can be made from old data to be processed.

b. Prediction

The same prediction algorithm with estimation algorithm in which the label / target / class of type numeric, the difference is that the data used is the data series of the time (the time series data). The method is suitable for the estimation, namely: Linear Regression, Neural Networks, Support Vector Machine, and others. Predictive nature could produce a class based on various attributes that we provide. Determining the outcome of the ongoing processes. The data used for the prediction derived from the data currently ongoing process. The term predictions are sometimes used also for classification, not only for time series

prediction, because it is able to generate a class based on various attributes that we provide

c. Classification

The algorithm uses the data to target / class / label in the form of categorical value (nominal). Grouping the data there to be in a group of pre-defined group name. The method is suitable for the estimation, namely: Naive Bayes, K-Nearest Neighbor, C4.5, ID3, CART, Linear Discriminant Analysis, and others. For example, if the target / class / label is available, it can be used the nominal value (categorical) as follows: revenues of large, medium and small.

d. Clustering

Clustering is grouping data, the observation and the case into a class that is similar. A cluster (cluster) is a collection of similar data between one another, and have differences when compared with data from other clusters. The method is suitable for the estimation, namely: K-Means, K-Medoids, Self-Organizing Map (SOM), Fuzzy C-Means, and others. The main difference with the classification algorithm clustering is clustering does not have a target / class / labeling, so including unsupervised learning clustering is often used as an early stage in the process of data mining, with the result that formed the cluster will be the input of the next algorithm used.

e. Association

Algorithm association rule (rule association) is an algorithm that finds the attribute. In the business world, often referred to as affinity analysis or market basket analysis. The algorithm association rules depart from the pattern 3JF antecedent, then consequent, "in conjunction with the measurement of support (coverage) and confidence (accuration) are associated in the rules. Association rule algorithms are: A priori algorithm, FP-Growth algorithm, GRI algorithm. Examples on Thursday night, 1000 the customer has to do shopping in supermarkets ABC, where:

1. 200 people buy Rinso.
2. Of the 200 people who buy Rinso, 50 person buying Fanta. Thus the association rules would be "If you buy Rinso, then buy Fanta" with the support of 200/1000 = 20% and confidence 50/200 = 25%.

The reason why women rarely develop heart disease before menopause is not known for certain, but it seems related to the hormones no longer in production during menstruation . Hormone therapy replacement (TPH), which many women do it can prevent attacks that some heart doctors. Because TPH recommend this.

a) Genetic

Your doctor will ask about family history if there is a close family member (parent, brother, sister) of heart disease. If the father had a heart attack before the age of 60 years or 65 years before the mother is exposed to the patient at high risk for heart disease. However, if the parents live to the age when the age 60 to 65 do not have a heart attack, it is not necessary to worry about. It is equally true for brother and sister. Although in a large family, it turns out that a heart attack, probably just a coincidence.

How can heart disease in the family? Most of the answer depends on the genes inherited from the parents which makes us prone to high cholesterol, high blood pressure, or diabetes. In addition similarity family lifestyle also determine, for example, eating the same food and parents smoke, children are usually also smoke. If families tend to develop heart disease, should do a routine examination to the doctor to make sure that does not suffer from high cholesterol.

b) Cholesterol

Previously said, atheroma is a major cause of coronary heart disease, because it arises in fat, especially due to LDL cholesterol (Low-Density Lipoprotein), called plaque, formed in the artery wall. This is what makes it more narrow thus impeding the flow of blood flowing throughout the body, if the plaque is broken then formed blood clots in the affected area of the heart muscle. This can lead to a heart attack. This process generally occurs and damage is more severe in one with high cholesterol levels in the blood. Genetic factors affect a person's cholesterol levels, some families have a gene with high fat levels in the blood. This condition is called hyperlipidemia family or abbreviated HK.

c) Food

Food also plays a major role in determining the level of cholesterol. The more fat - especially animal fat and dairy products are eaten, the higher the cholesterol, and the higher the risk of heart disease. Therefore, reduce the consumption of animal fats in the diet.

d) Framingham Study

Research linking high cholesterol and heart disease is done after World War II in Framingham, a close kotakecil Bostom United States. All residents in check once a year to see if they are affected by the disease jantung.Ternyata there is a close connection with high cholesterol, the higher the blood cholesterol, the higher heart

attack. The study also shows other risk factors, such as smoking, high blood pressure, and diabetes. Various risk factors that have been ensured after observation for nearly 40 years, since the study was started. Until now, the study is still ongoing.

e) Smoke

Smoking is closely linked to risk of heart disease. The chemicals in cigarette smoke are absorbed into the bloodstream from the lungs and circulate throughout the body and affects every cell of the body. These chemicals often makes blood cells called platelets become stickier, making it easy to form clots in the blood vessels. Pipe and cigar smokers the risk is not as high as segaret, but still at risk for heart disease than non-smokers. The number of cigarettes in the suction also affected, the risk increases according to the level of consumption that is lightweight (less than 10 cigarettes a day), moderate (10-20 cigarettes a day) and heavy smokers (more than 20 cigarettes per day). The reason doctors strongly advised to stop smoking because it is a risk factor that can control themselves. Moreover, the patient will begin to feel the benefit when stopped. Although the risk of heart disease is not as low as nonsmokers, the result will be approached about a year later.

f)Stress

Often we found stress is the cause of heart attacks, but scientifically it is actually hard to prove. There are several factors trigger heart attacks, such as sport suddenly ¬tiba overwhelming emotions, can lead to heart attacks even though this is rare, believe it or not, during World War II that a lot of stress on civilians and the military, the number of civilians who had a heart attack have decreased.

CONCLUSION AND SUGGESTION

To predict heart disease, one of the data mining algorithm that can be used is a C4.5 algorithm. In addition to generating a simple model that is easy to understand, this algorithm also has a high degree of accuracy in the classification process attribute. C4.5 algorithms are also quite reliable in processing numerical data and a nominal increase in the prediction of heart disease, yet C4.5 algorithms still have weaknesses in handling high-dimensional data, it's him backward elimination algorithm can be applied to the process of selecting attributes in order to improve the prediction of attack heart disease in patients.

Based on the results that have been concluded in stages over about a prediction of heart disease, it can put forward some suggestions as follows:

1. To generate higher predictive of the study, needed cleaning (data cleansing) of the input data is inconsistent and the data is corrupted or the so-called garbage data, the initial processing stage.
2. Adding a greater amount of data and attributes more, so that the measurement results will be obtained better, or use other methods are better than backward elimination.

BIBLIOGRAPHY

1. Anbarasi, E. Anupriya dan N.CH.S.N.Inyegar pada tahun 2010 menggunakan UCI *Heart Disease* Data Set dengan algoritma *Naives Bayes.*
2. Abdul Rohman, Tahun 2013 menggunakan *Heart Disease* UCI Data Set dengan algoritma C4.5 berbasis Adaboost akurasi 92,24 %.
3. Edy tahun 2012 menggunakan *Heart Disease* UCI Data Set dengan algoritma C4.5 Barbasis Genetika dengan akurasi 77.40%.
4. Freund dan Scaphire tahun 1995. Adaboost atau adaptive boosting.
5. Srinivas Dr. Raghavendra Rao dan Dr. A. Govardhan pada tahun 2010 menggunakan *Heart Disease* UCI Data Set.
6. (Han & Kamber, 2006, p. 301). Atribut dengan data yang lebih seragam, dan pohon keputusan yang sesederhana.