International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

# Data Mining for Heart Disease Dataset Using Genetic Algorithm with J48 Classifier

**Arunpreet Veen[1], Deepak Aggarwal[2]**

[1]Scholar at Department of CSE, BBSBEC, Punjab, India

[2]Professor, Department of CSE, BBSBEC, Fatehgarh Sahib, Punjab, India

**Abstract:** ***Classification has been used for extraction of hidden patterns available in dataset. Classifier has been used for prediction of class labels to the testing dataset using classification methodology. In this paper heart disease prediction classification has been done using J48 classifier with genetic approach for attribute selection. Genetic approach selects optimizes attributes subset for classification that can provide better classification accuracy. In this paper result of purposed classifier has been illustrated.***

**Keywords:** Genetic Algorithm, J48, crossover, mutation and information gain

## 1. Introduction

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods. Consequently, data mining consists of more than collection and managing data, it also includes analysis and prediction. Classification technique is capable of processing a wider variety of data than regression and is growing in popularity. There are several applications for Machine Learning (ML), the most significant of which is data mining. People are often prone to making mistakes during analyses or, possibly, when trying to establish relationships between multiple features. This makes it difficult for them to find solutions to certain problems. Machine learning can often be successfully applied to these problems, improving the efficiency of systems and the designs of machines. Numerous ML applications involve tasks that can be set up as supervised. In the present paper, we have concentrated on the techniques necessary to do this. In particular, this work is concerned with classification problems in which the output of instances admits only discrete, unordered values. Our next section presented Decision Tree Induction. Section 3 described Bayesian Network whereas k-nearest neighbor classifier described in section 4. Finally, the last section concludes this work.

## 2. Review of Literature

**Li Liu "Robust dataset classification approach based on neighbor searching and kernel fuzzy c-means"** Dataset classification is an essential fundament of computational intelligence in cyber-physical systems (CPS). Due to the complexity of CPS dataset classification and the uncertainty of clustering number, this paper focuses on clarifying the dynamic behavior of acceleration dataset which is achieved from micro electro mechanical systems (MEMS) and complex image segmentation. To reduce the impact of parameters uncertainties with dataset classification, a novel robust dataset classification approach is proposed based on neighbor searching and kernel fuzzy c-means (NSKFCM) methods. Some optimized strategies, including neighbor searching, controlling clustering shape and adaptive distance kernel function, are employed to solve the issues of number of clusters, the stability and consistency of classification, respectively. Numerical experiments finally demonstrate the feasibility and robustness of the proposed method.

**Muhammad Shakil Pervez "Feature selection and intrusion classification in NSL-KDD cup 99 dataset employing SVMs",** Intrusion is the violation of information security policy by malicious activities. Intrusion detection (ID) is a series of actions for detecting and recognizing suspicious actions that make the expedient acceptance of standards of confidentiality, quality, consistency, and availability of a computer based network system. In this paper, we present a new approach consists with merging of feature selection and classification for multiple classes NSL-KDD cup 99 intrusions detection dataset employing support vector machine (SVM). The objective is to improve the competence of intrusion classification with a significantly reduced set of input features from the training data. In supervised learning, feature selection is the process of selecting the important input training features and removing the irrelevant input training features, with the objective of obtaining a feature subset that produces higher classification accuracy. In the experiment, we have applied SVM classifier on several input feature subsets of training dataset of NSL-KDD cup 99 dataset. The experimental results obtained showed the proposed method successfully bring 91% classification accuracy using only three features and 99%classification accuracy using 36 features, while all 41 training features achieved 99% classification accuracy.

**Omprakash Chandrakar "Empirical Study to Suggest Optimal Classification Techniques for Given Dataset"** Classification techniques play an important role in Data Mining. Large numbers of classification techniques have been proposed in the literature. No single algorithm can be considered optimal for all type of data set. Accuracy of classification result highly depends on the selection of classification algorithms. Different classification techniques produce different results for the same data set. Thus finding the optimal algorithm for the given data set is a challenge. The outcome of this research work can be useful in selecting most suitable classifier for the given dataset. Research

Methodology: To determine the effectiveness of various classification algorithms, authors run some well-known classification algorithms against some standard datasets. Effectiveness of various algorithms is measured on the basis of average accuracy, time taken to build classification model, mean absolute error etc. Results: Based on the comparative study of the experiment results, authors suggest the optimal algorithm for different categories of datasets.

**Datta H. Deshmukh "Improving classification using preprocessing and machine learning algorithms on NSL-KDD dataset"** Classification is the category that consists of identification of class labels of records that are typically described by set of features in dataset. The paper describes a system that uses a set of data pre-processing activities which includes Feature Selection and Discretization. Feature selection and dimension reduction are common data mining approaches in large datasets. Here the high data dimensionality of the dataset due to its large feature set poses a significant challenge. In Pre-processing with the help of Feature selection algorithm the various required features are selected, these activities helps to improve the accuracy of the classifier. After this step various classifiers are used such as Naive Bayes, Hidden Naive Bayes and NB-Tree. The advantage of Hidden Naive Bayes is a data mining model that relaxes the Naive Bayes Method's conditional Independence assumption. Also the next Classifier used is NB-Tree which induces a hybrid of decision tree classifiers and Naive Bayes classifiers which significantly improves the accuracy of classifier and decreases the Error rate of the classifier. The output of the proposed method are checked for True positive, True negative, False positive, False negative. Based on these values the Accuracy and error rate of each classifier is computed.

**Pooja Sharma "Classification algorithms on a large continuous random dataset using rapid miner tool"** Classification is widely used technique in the data mining domain, where scalability and efficiency are the immediate problems in classification algorithms for large databases. Now a day's large amount of data is generated, that need to be analyses, and pattern have to be extracted from that to get some knowledge. Classification is a supervised machine learning task which builds a model from labeled training data. The model is used for determining the class; there are many types of classification algorithms such as tree-based algorithms (C4.5 decision tree, j48 decision tree etc.), naive Bayes and many more. These classification algorithms have their own pros and cons, depending on many factors such as the characteristics of the data. We can measure the classification performance by using several metrics, such as accuracy, precision, classification error and kappa on the testing data. We have used a random dataset in a rapid miner tool for the classification. Stratified sampling is used in different classifier such as J48, C4.5 and naive Bayes. We analyzed the result of the classifier using the randomly generated dataset and without random dataset.

**RajibSarkar"Singer based classification of song dataset using vocal signature inherent in signal"** Singer based classification of song data is important in the applications like, organized archival and indexing of music data, music retrieval. In a song, singing voice is mixed with accompanying instrument signal. To extract the vocal characteristics of the singer, the effect of non-voiced part is to be minimized. In this work a simple methodology is proposed to remove the non-voiced segments and to reduce the impression of instruments from the voice-dominating signal. To extract the vocal signature, proposed features extract the variation pattern of zero crossing rate and short term energy. In broad sense, the features try to capture the range of pitch and energy over which a singer mostly operates. This is motivated by the way a human being tries to identify a singer. Finally, singer based classification is done using multi-layer perceptron network. Experiment is carried out with artist20 dataset and 63%classification accuracy is achieved. Comparison with reported works on the same dataset shows that the performance of the proposed simple methodology is better than the majority and very close to others.

## 2. Methodology

Data mining is the processing of raw information for extraction of valuable information. Information has been extracted using various data mining approaches that divide whole datasets into different segments. These different segments have been used for extraction of value able information on the basis of various rules of weight functions.

In the purposed work data classification has been done for heart disease dataset. Heart disease data set contains different attributes for dataset. Heart disease dataset contain 13 different attributes and 14$^{th}$ class label attribute. In the process of classification these attributes have been used for generation of different rules for classification process. In the processing of heart disease prediction system these attributes have been used by different tree based classifiers. Tree based classifier utilizes different attributes for division of dataset. Dataset attributes have been explained below.

**Table 3.1:** Heart Disease Dataset Description

| Dataset Attributes | Description |
|---|---|
| AGE | In years |
| Sex | 1-Male/0- Female |
| Chest Pain Type | 1: typical type 1 angina, 2 typical type angina, 3:non-angina pain; 4: asymptomatic |
| Resting Blood Pressure | BP in mg (numeric) |
| Serum Cholesterol | In md/dl |
| Fasting Blood Sugar | If >120 1 else 0 |
| Resting electrocardiographic | 0-normal,1-having ST-T wave abnormality,2-showing probable or definite left ventricular hypertrophy |
| Maximum heart rate achieved | Numeric value |
| Exercise induced angina | 0-no, 1-yes |
| Old-Peak | ST depression induced by exercise relative to rest |
| Slope of the peak exercise ST segment | 1-unsloping, 2- Flat, 3-downsloping |
| number of major vessels | 0-3 numeric |
| Thal | 3 = normal; 6 = fixed defect; 7 = reversible defect, (nominal) |

Table 3.1 represents dataset attributes description used for data classification. Dataset attributes are important for detection of disease to a patient. These 13 attributes are interrelated on each other that provide better classification.

In the purposed work data mining has been done using J48 classifier. J48 classifier is a tree based classifier that has been used for classification by generating pruning tree for classification.

J48 classification tree has been used for analysis of predicted variable that is known as independent variable based on attributes of the dataset. In j48 tree has been developed using highest information gain attribute.

Classification tree has been developed by analyzing attribute that provides similarity of class labels on the basis of highest gain information attribute. Which attribute provides highest gain information that is selected as a root tree and other attributes gain information has been computed for root node and divide tree to different leave nodes.

In the purposed work genetic algorithm has been implemented for selection of best attributes using cross over mutation process. Genetic algorithm has been used crossover, selection mutation and mutation process.

In this process genetic approach use dataset from attribute selection based on merit and scaled values. Merit value has been generated from initial population that has been solved by the classification approach and generated new generation based on cross over and mutation process.

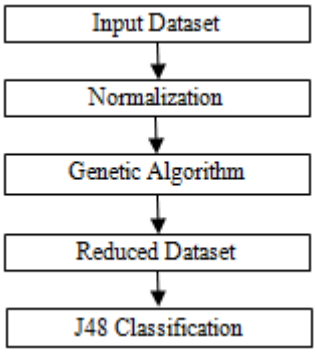


**Figure 3.1:** Flow of purposed model

This figure represents flow of the purposed classifier that has been used for heart disease dataset classification. In the purposed work reduction of the dataset attributes have been using genetic optimization process. Genetic algorithm selects best attributes on the basis of scaled functions.

## 3. Results

In the purposed work heart disease data classification has been done using different optimization approach for dataset attribute selection with tree based classifier. In the tree based classifier different approaches have been used for selection of best attributes from the dataset. Generic approach has been used for attribute selection. Genetic approach selects different merits on the basis of scaled function using various subsets. For each subset new generation has been evolved using crossover and mutation process. Crossover and mutation use parent class variable for generation of new generation and best feature subset has been used from these generation.

```
        Population size: 20
        Number of generations: 20
        Probability of crossover:  0.6
        Probability of mutation:  0.033
        Report frequency: 20
        Random number seed: 1

Initial population
merit           scaled          subset
 0.30391         0.36787        1 3 4 5 6
 0.26126         0.25873        1 3 7
 0.32092         0.41139        2 4 5
 0.1714          0.02882        6
 0.3413          0.46352        4 5 6 7
 0.31763         0.40296        3 4 5 7
 0.19645         0.09292        4
 0.31548         0.39747        2 7
 0.19645         0.09292        4
 0.29331         0.34075        1 4 5 6
 0.18964         0.0755         1 5
 0.29352         0.34129        1 2 4
 0.24277         0.21143        2
 0.27498         0.29384        1 6 7
 0.22358         0.16233        3 4
 0.19645         0.09292        4
 0.25367         0.23932        4 6
 0.21293         0.13509        3 6
 0.36562         0.52575        1 2 5 6 7
 0.28626         0.32271        1 2 5
```

**Figure 4.1:** Initial population for genetic algorithm

This figure represents initial population that has been used for selection of best attributes from the feature subset. On the basis of these initial feature subsets new subsets have been generated and best sub set has been evolved.

```
Generation: 20
merit          scaled         subset
 0.38982        0.42236       1 2 4 5 6 7
 0.38982        0.42236       1 2 4 5 6 7
 0.38982        0.42236       1 2 4 5 6 7
 0.38982        0.42236       1 2 4 5 6 7
 0.38982        0.42236       1 2 4 5 6 7
 0.3897         0.42123       1 2 3 4 5 6 7
 0.38744        0.40033       2 3 4 5 6 7
 0.38744        0.40033       2 3 4 5 6 7
 0.38982        0.42236       1 2 4 5 6 7
 0.3897         0.42123       1 2 3 4 5 6 7
 0.38964        0.42074       2 4 5 6 7
 0.38982        0.42236       1 2 4 5 6 7
 0.3897         0.42123       1 2 3 4 5 6 7
 0.38964        0.42074       2 4 5 6 7
 0.3442         0                1 4 5 6 7
 0.38982        0.42236       1 2 4 5 6 7
 0.38744        0.40033       2 3 4 5 6 7
 0.38744        0.40033       2 3 4 5 6 7
 0.36687        0.20988       2 3 5 6 7
 0.38982        0.42236       1 2 4 5 6 7

Attribute Subset Evaluator (supervised, Class (nominal): 8 class):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 1,2,4,5,6,7 : 6
```

**Figure 4.2:** Generated population using genetic algorithm

This figure represents new generated feature subset after implementation of genetic algorithm. This approach use best 6 attributes for classification process for heart disease dataset.

```
chest <= 3
|   thal <= 6
|   |   oldpeak <= 1.6: absent (58.0/2.0)
|   |   oldpeak > 1.6
|   |   |   number_of_major_vessels <= 0
|   |   |   |   maximum_heart_rate_achieved <= 163: present (4.0/1.0)
|   |   |   |   maximum_heart_rate_achieved > 163: absent (2.0)
|   |   |   number_of_major_vessels > 0: absent (2.0)
|   thal > 6
|   |   oldpeak <= 1.9
|   |   |   number_of_major_vessels <= 0
|   |   |   |   age <= 58: absent (10.0/1.0)
|   |   |   |   age > 58: present (3.0)
|   |   |   number_of_major_vessels > 0
|   |   |   |   maximum_heart_rate_achieved <= 158: absent (4.0/1.0)
|   |   |   |   maximum_heart_rate_achieved > 158: present (2.0)
|   |   oldpeak > 1.9: present (5.0)
chest > 3
|   number_of_major_vessels <= 0
|   |   thal <= 6
|   |   |   exercise_induced_angina <= 0
|   |   |   |   age <= 59: absent (13.0)
|   |   |   |   age > 59: present (3.0/1.0)
|   |   |   exercise_induced_angina > 0
|   |   |   |   oldpeak <= 0.8: absent (2.0)
|   |   |   |   oldpeak > 0.8: present (2.0)
|   |   thal > 6
|   |   |   oldpeak <= 0.2
|   |   |   |   age <= 42: present (2.0)
|   |   |   |   age > 42: absent (2.0)
|   |   oldpeak > 0.2: present (11.0)
|   number_of_major_vessels > 0: present (45.0/2.0)
```

**Figure 4.3:** J48 pruning tree representation

This figure represents pruning tree that has been generated using J48 classifier. On the basis of this tree chest is the best attribute that has been selected as root node. Chest is highest gain information attribute that can be used for classification. After chest thal and maximum number of vessels have been used for division of other attributes to generate classifier model.

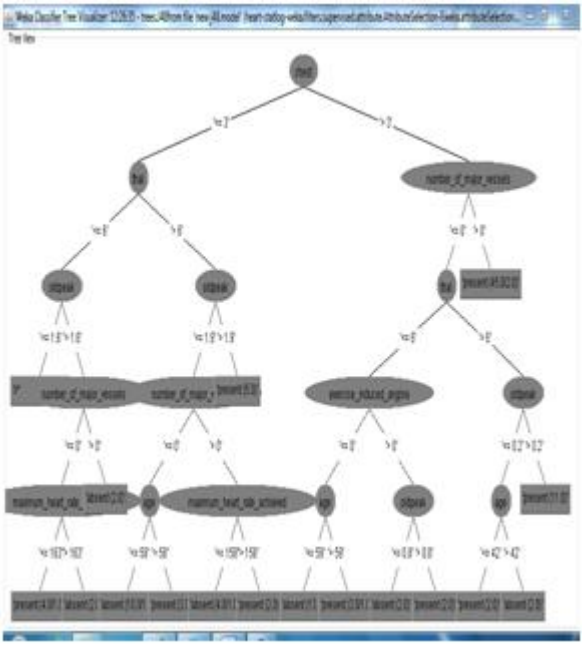

**Figure 4.4:** Tree structure of J48 classifier

This figure represents classification tree developed using J48 classification model. This tree contains 17 leave nodes and total size of tree is 33. On the basis of tree dataset has been used for classification of heart disease dataset.

Dataset has been uploaded to the system and dataset has been normalized using different filters after this genetic approach has been implemented on the dataset for reduction of dataset attributes. Genetic approach computes best subset from all the attributes using different generations. Best feature subset is age, chest, maximum heart rate achieved, exercise induced angina, oldpeak, number of major vessels and thal has been best selected attributes. Using these attributes classifier has been developed and dataset has been classified. After dataset classification various parameters have been analyzed for performance evaluation of classifier.

After classification model that has been used for classification of heart disease dataset. Classification model classify different instances using predicted class. After process of predicted class labeling various parameters has been analyzed for performance evaluation of purposed system.

**Table 5.1:** Compassion table for different classifiers using various parameters

| Parameter | GA+J48 | J48 |
|---|---|---|
| Correctly Classified | 140 | 125 |
| Incorrectly Classified | 30 | 45 |
| Precision | 0.823 | 0.736 |
| Recall | 0.824 | 0.735 |
| F-measure | 0.823 | 0.736 |
| TP Rate | 0.824 | 0.735 |
| FP Rate | 0.21 | 0.266 |
| ROC Area | 0.851 | 0.764 |

This table represents various parameters for performance evaluation of purposed system. These parameters are important for performance evaluation of a classifier.

## 4. Conclusion & Future Scope

Data mining is the process of extraction of different relations between different instances available in the dataset. Data mining classification classify different data attributes into different classes based on properties of dataset. In this paper heart disease prediction has been done using tree based classifier that develops pruning tree for dataset classification. Various classification parameters have been analyzed in purposed work. By analyzing these performances evaluation parameter we can state that purposed approach provides better classification than simple J48 classifier.

## References

[1] DeepaliChandna "Diagnosis of Heart Disease Using Data Mining Algorithm", IEEE Conf. on International Journal of Computer Science and Information Technologies, 2015, pp. 1678-1680.
[2] Thuraisingham "Data Mining for Malicious Code Detection and Security Applications", 978-0-7695-4406-9, 4 – 5, IEEE, 2011.
[3] Asghar, S. "Automated Data Mining Techniques: A Critical Literature Review" 978-0-7695-3595-1, 75 – 79, IEEE, 2009.
[4] M.Akhiljabbar a, "Heart Disease Prediction System using Associative Classification and Genetic Algorithm", IEEE, 2012.
[5] Ashish Kumar Sen1 "A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level"ISSN 2319-7242Volume 2, 2663-2671, IEEE, 2013.
[6] Li Liu "Robust dataset classification approach based on neighbor searching and kernel fuzzy c-means" IEEE/CAA Journal of AutomaticSinica, pp. 235 – 247, 2015.
[7] Muhammad Shakil Pervez "Feature selection and intrusion classification in NSL-KDD cup 99 dataset employing SVMs" IEEE International Conference on Software, Knowledge, Information Management and Applications, pp. 1–6, 2014.
[8] OmprakashChandrakar "Empirical Study to Suggest Optimal Classification Techniques for Given Dataset" IEEE International Conference on Computational Intelligence & Communication Technology, pp. 30 – 35, 2013.
[9] Datta H. Deshmukh "Improving classification using preprocessing and machine learning algorithms on NSL-KDD dataset" IEEE International Conference on Communication, Information & Computing Technology,pp. 1–6,2015
[10] Pooja Sharma "Classification algorithms on a large continuous random dataset using rapid miner tool" IEEE International Conference on Electronics and Communication Systems, pp. 704–709, 2015.
[11] Rajib Sarkar, "Singer based classification of song dataset using vocal signature inherent in signal" NationalConference on Computer Vision, Pattern Recognition, Image Processing and Graphics, pp. 1 – 4, 2015.
[12] Pramod Kumar Yadav, K.L.Jaiswal2, "Intelligent Heart Disease PredictionModel Using Classification Algorithms", *IJCSMC, Vol. 2, Issue, pp.102*–107, IEEE, 2013.
[13] ShamsherBahadur Patel, Pramod Kumar Yadav, Dr. D. P.Shukla, "Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques" ISSN: 2319-2380, 2319-2372. Volume 4, Issue 2, PP 61-64, IEEE, 2013.
[14] Dr.Motilal C., "Role of Data Mining Techniques in Healthcare sector in India", 2320-6691, 158-160, IEEE, 2013.
[15] Nassar, O.A. "The integrating between web usage mining and data mining techniques" 243–247, IEEE, 2013.