Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Analysis of Heart Disease using in Data Mining Tools Orange and Weka

Sarangam Kodati[1]

[1] Sri Satya Sai University of Technology and Medical Science

Received: 9 December 2017 Accepted: 2 January 2018 Published: 15 January 2018

Abstract

Health care is an inevitable task to be done in human life. Health concern business has become a notable field in the wide spread area of medical science. Health care industry contains large amount of data and hidden information. Effective decisions are made with this hidden information by applying patient; however, with data mining these tests could be reduced. But there is a lack of analyzing tool according to provide effective test outcomes together with the hidden information, so and such system is developed using data mining algorithms for classifying the data and to detect the heart diseases. Data mining acts so a solution by many healthcare problems. Naïve Bayes, SVM, Random Forest, KNN algorithm is one such data mining method which serves with the diagnosis regarding heart diseases patient. This paper analyzes few parameters and predicts heart diseases, thereby suggests a heart diseases prediction system (HDPS) based total on the data mining approaches

Index terms— data mining, weka, orange, heart disease, data mining classification techniques.

1 Analysis of Heart Disease using in Data Mining

Tools Orange and Weka Sarangam Kodati ? & Dr. R. Vivekanandam ? Abstract-Health care is an inevitable task to be done in human life. Health concern business has become a notable field in the wide spread area of medical science. Health care industry contains large amount of data and hidden information.

Effective decisions are made with this hidden information by applying patient; however, with data mining these tests could be reduced. But there is a lack of analyzing tool according to provide effective test outcomes together with the hidden information, so and such system is developed using data mining algorithms for classifying the data and to detect the heart diseases. Data mining acts so a solution by many healthcare problems. Naïve Bayes, SVM, Random Forest, KNN algorithm is one such data mining method which serves with the diagnosis regarding heart diseases patient. This paper analyzes few parameters and predicts heart diseases, thereby suggests a heart diseases prediction system (HDPS) based total on the data mining approaches.

Keywords: data mining, weka, orange, heart disease, data mining classification techniques.

I.

2 Data Mining

ata mining is concerned together with the method of computationally extracting unknown knowledge from vast sets of data. Extraction of useful knowledge from the enormous data sets and providing decision-making results for the diagnosis or remedy of diseases is very important. Data mining can stand used to extract knowledge by analyzing and predicting some diseases. Health care data mining has a large potential according to discover the hidden patterns among the data sets about the medical domain. Various data mining methods are available with their suitability dependent on the healthcare data. Data mining applications in health care can have a wonderful potential and effectiveness. It automates the process of finding predictive information in large databases. Disease prediction plays an important role in data mining. Finding of heart disease requires the performance of some

tests on the patient. However, use of data mining techniques can reduce the number of tests. This reduced test set plays a significant role in performance and time. Health care data mining is an important task because it allows doctors to see which attributes are more important for diagnosis such as age, weight, symptoms, etc. This will help the doctors diagnose the disease more efficiently. Knowledge discovery in databases is the method of finding useful information and patterns into data. Knowledge discovery within databases can be do using data mining. It makes use of algorithms after extract the information and patterns derived by the knowledge discovery in databases process. Various stages of knowledge discovery in databases process are highlighted in Fig. **??**.

3 Fig. 1: KDD Process

Various stages concerning knowledge discovery of databases method are described as follows. In Selection stage, that obtains the different data resources. In preprocessing stage, it removed the unwanted missing and noisy data and furnished the clean data which execute format in accordance including a common format of transform stage. Then data mining techniques are applied according to get desired output. Finally into the between the signification stage, that will present the result after end user in a meaningful manner. instead of predicting the discrete class the outcome is a numeric value. c) Association rule mining: The association and patterns between the some attributes are extracted or from its attributes, rules are created. The rules and patterns are used predicting the categories or classification of the test data. d) Clustering: The grouping of similar instances into clusters takes place. The challenges or drawbacks considering this type of machine learning is that we have according to first identify clusters and assign a new instance according to these clusters [8].

4 II.

5 Data Mining Techniques

Out of this four types of learning methods, we need to identify the algorithm as performs better. The application of data mining methods depends on the types of data which is fitted to be used in the techniques, or solving data mining troubles depend on the types of data to stand used and the selection about data mining technique which is most suitable for the data used.

6 III.

7 Machine Learning

Machine learning (ML), employed as like a method in data science, is the process of programming computers after learning from past experiences (Mitchell, 1997Machine Learning seeks to develop algorithms to that amount learn out of data directly with little or no human intervention. Machine Learning algorithms perform a range of tasks such so like prediction, classification, or decision making. Machine Learning stems from artificial intelligence research and has become an essential aspect of data science. Machine learning begins with input so a training data set. In this phase, the Machine Learning algorithm employs the training dataset after learning from the data and structure patterns. The learning phase outputs a model so much is used by way of the testing phase. The testing phase employs any other dataset, applies the model from the training phase, and results are presented for analysis. The overall performance regarding the test dataset demonstrates the model's ability in conformity with performing its task against data. Machine learning extends beyond a statically coded set regarding statements into statements, so a lot are dynamically generated based as regards the input data.

IV.

8 Open Source Softwares

Open source has, in the minds regarding many, come to be synonymous with free software **??**Walters, 2007). Open source software is software where the development then the source code are made publically available and designed after denying everyone the right according to exploit the software **??**Laurent, 2004). Open source general refers in conformity with the source code concerning the application being freely and openly available because of modifications. Two such examples of open source licenses are the GPL, or general people consent (GNU.org, 2015a), then GNU(GNU.org, 2015b). Anyone be able to develop extensions then customizations about open source software; though, charging a fee for certain things to do is typically prohibited by using a public license agreement whereby any modifications to the source code automatically become public domain. Communities emerge around software with developers worldwide extending open source software.

V.

9 Heart Diseases

The highest mortality in both India and abroad is due to heart disease. So it is vital time to check this death toll by correctly identifying the disease between initial stage. The matter becomes a headache for all medical doctors both in India and abroad. Nowadays doctors are adopting many scientific technologies and methodology for both identifications or diagnosing not only the common disease but also many fatal diseases. The successful treatment is continually attributed to right and accurate diagnosis. Doctors may also sometimes fail to take

accurate decisions while diagnosing the heart disease about a patient, therefore heart disease prediction systems which use machine learning algorithms assist in such cases to get accurate results [1].

10 VI.

11 Heart Disease Dataset

The dataset used for this work is from UCI Machine Learning repository from which the Cleveland heart disease dataset is used. The dataset has 303 instance and 76 attributes. However, only 14 attributes are used of this paper. These 14 attributes are the consider factors for the heart disease prediction [8]. Even though it has 303 instances as only 297 are completed and the remaining rows contained missing values and removed out of the experiment.

12 VII. Overview of Data Mining Tools

Data mining has a wide number of applications ranging from marketing and advertising about goods, functions and products, artificial intelligence research, biological sciences, crime investigations to high-level government intelligence. Due to its widespread usage and complexity involved in building information mining applications, a vast number of Data mining tools hold been developed over decades. Every tool has its advantages and disadvantages. [6] Within data mining, there is a group of tools that have been developed by a research community and data analysis enthusiasts; he are provided free of the price using one on the existing open-source licenses. An open-source development model means that the tool is a result of a community effort, not necessarily supported by a single

13 Global Journal of Computer Science and Technology

Volume XVIII Issue I Version I 18 Year 2 018 ( ) C organization but alternatively the result regarding contributions from an international and informal development team. This development style affords a means on incorporating the various experiences Data boring gives many excavation techniques according to extract data from databases. Data mining tools predict future trends, behaviors, allowing business according to make proactive, knowledge-driven decisions. The development and application concerning data mining algorithms require the use of very powerful software tools. As the number of accessible tools continues by grow the choice of the most suitable tool becomes increasingly difficult. [6] The top 6 open source tools available because data mining is briefed as below.

Data mining tools like Weka and Orange are used to perform various data mining techniques. The first step of the methodology consists of selecting a number of available open source data mining tools in accordance with being tested. Many open data mining tools are available for free on the Web. After surfing the Internet, some tools were chosen; including the Waikato Environment for Knowledge Analysis (WEKA) durability and Orange Canvas.

14 VIII.

15 Weka

The Waikato Environment for Knowledge Analysis (WEKA) [7] is an open source software and machine learning toolkit introduced by Waikato University, New Zealand. WEKA helps several standard data mining tasks as data preprocessing, clustering, classification, regression, visualization and feature selection New algorithms can also be implemented the usage concerning WEKA with existing data mining and machine learning techniques. WEKA gives a number sources because loading data, which include files, URLs then databases. It helps file formats include WEKA"s own ARFF format, CSV, Lib SVMs format, and C4.5's format. Many evaluation criteria are also provided of WEKA certain as confusion matrix, precision, recall, true positive and false negative, etc. Some of the advantages of WEKA tool includes Open source, platform independent and portable, graphical user interface and contains a very vast collection of different data mining algorithms.

16 IX.

17 Orange

Orange is an open source machine learning technology or data mining software. Orange can be used for explorative data analysis and visualization **??**3]. It gives a platform for experiment selection, predictive modeling, and recommendation systems and can be used of genomic research, biomedicine, bioinformatics, and teaching. Orange is always preferred when the factor of innovation, quality, or reliability is involved[10], [4].

X.

18 The Comparative Study

The methodology of the study constitutes regarding collecting a set of free data mining and knowledge discovery tools according to be tested, specifying the data sets to be used, and selecting a set of classification algorithm

according to test the tools' performance. Demonstrates the overall methodology followed for fulfilling the goal of its research.

19 b) Recall

It is defined as the average probability of complete retrieval. Recall= True positives/True positives + False negative

20 c) Navie Bayes

When the dimensionality of the inputs is high, the Naïve Bayes Classifier method is particularly suited. The problem including the Naïve Bayes Classifier is so that assumes all attributes are independent on each other which in general cannot be applied. Naive Bayes is harder to debug and understandable [2]. Naive Bayes used into robotics and computer vision. In naive Bayes, decision tree perform poorly. Comparative analysis of precession and recall analyzing for heart disease data sets precession in Orange 82.4% and Recall 80.6%. In WEKA precession 83.7% and Recall 83.7 %.Compare to Orange tool and WEKA, weka is best precession and Recall.

21 d) Support Vector Machine

Support Vector Machines proved themselves to be very fine into a variety of pattern classification tasks and accordingly received a great deal of attention recently. Support vector machine is a supervised machine learning technique. The SVM algorithm predicts the occurrence about heart disease by ability on plotting the disease predicting attributes regarding the multidimensional hyperplane or classifies the classes optimally by creating the approach between two data clusters [5]. This algorithm attains high accuracy by the use regarding nonlinear features called kernels. Comparative analysis of precession and recall analyzing for heart disease data sets precession in Orange 81.7% and Recall 70.5%. In WEKA precession 81.8% and Recall 81.9 %.Compare to Orange tool and WEKA, weka is best precession and Recall.

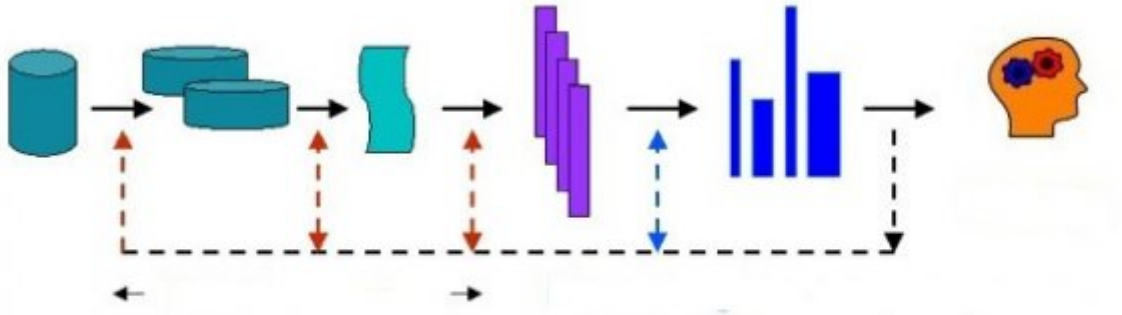
22 e) Random Forest

Random Forest is essentially an ensemble of unpruned classification trees. It gives excellent performance concerning a number about practical problems, largely because such is not sensitive to noise in the dataset, and it is not subject to overfitting. It works fast and generally exhibits a substantial performance improvement over many other tree-based algorithms. Random forests are built by combining the predictions on a number of trees, each of which is trained within isolation. There are three main choices to stand performed when constructing a random tree. Comparative analysis of precession and recall analyzing for heart disease data sets precession in Orange 77.9% and Recall 73.4%. In WEKA precession 81.8% and Recall 81.9 %.Compare to Orange tool and WEKA, weka is best precession and Recall.

23 f) KNN Classifier

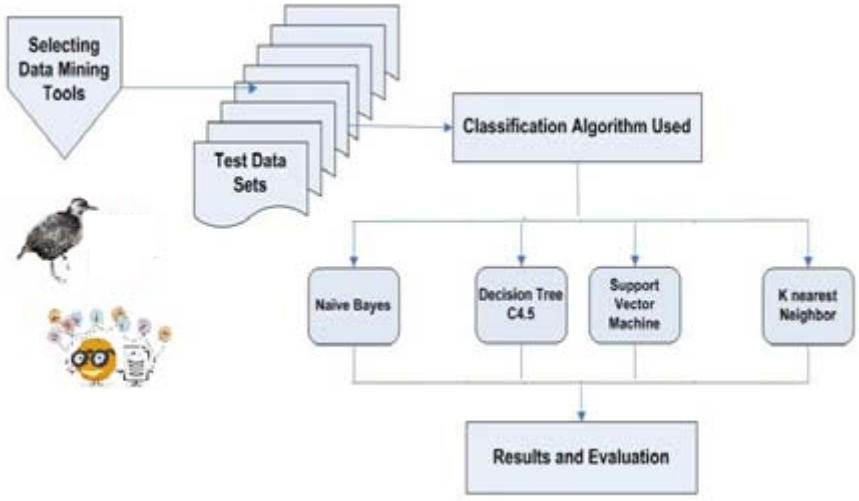
K-nearest neighbor is a sophisticated approach for classification that finds a group of K objects in the training documents that are close to the test value. To classify an unlabeled object, the distance between it object and labeled object is computed and it's K nearest neighbors are identified. Classification accuracy commonly depends of the choice value of K and will be better than that of using the nearest neighbor classifier [9]. For vast data sets, K can be larger to reduce the error. Choosing K can be done experimentally, where a number concerning patterns taken out from the training set can be categorised using the remaining training patterns for different values over k. The value of K which gives the least error in classification will be chosen. If same class is shared in various of K-nearest neighbors, then per-neighbor weights of as class are added together, and the resulting weighted sum is used as the likelihood score of that class with respect to the test document [8].Comparative analysis of precession and recall analyzing KNN for heart disease data sets precession in Orange 58% and Recall 54.7%. In WEKA precession 75.3% and Recall 75.2 %. Compare to Orange tool and WEKA weka is best precession and Recall.
[1]

[1]



2

Figure 1: Fig. 2 :



3

Figure 2: Fig. 3 :

Data F usion

De Dimensionalization Classification Visualization - - noising

D

The most frequently used Data Mining techniques are specified below: a) Classifica

a set of classified examples (training set) and uses it for training the algorithms. With the trained algorithms, classification of the test data takes place based over the patterns and r

Author ?: Research Scholar, Department of Computer Science and Engineering, Sri Satya Sai University of Technology and Medical Science, Sehore, Bhopal, Madhya Pradesh, India.
e-mail: k.sarangam@gmail.com
Author ?: Professor, Director in Muthayammal Engineering College, Namakkal, India.

b) Numeric predication: This is a variant of classification learning with the excepti

Figure 3: Raw Data Target Data Preprocessed Data Transformed Data Patterns Knowledge Data preprocessing Pattern Recognition Interpreting Results Knowledge

1

Figure 4: Table 1 :

showing best classification

Algorithm classification Average	Precessio n in Orange	Recall in Orange	Precession in WEKA	Recall in WEKA
Naïve base classifier	0.824	0.806	0.837	0.837
SMO or Support Vector Machine	0.817	0.705	0.84	0.8365
Random Forest	0.779	0.734	0.818	0.819
1BK or K-Nearest Neighbor	0.58	0.547	0.753	0.752

Figure 5:

[Mikut and Wiley (2011)] , Ralf Mikut , Markus Reischl Wiley . *Interdisciplinary Reviews: Data Mining and Knowledge Discovery* September/ October 2011. 1 (5) p. .

[Gosain and Kumar (2009)] 'Analysis of health care data using different data mining techniques'. A Gosain , A Kumar . *IAMA 2009. International Conference on*, 2009. July 2009. 6 p. . (Intelligent Agent & Multi-Agent Systems)

[Majali et al. ()] 'Data mining techniques for diagnosis and prognosis of cancer'. J Majali , R Niranjan , V Phatak , O Tadakhe . *International Journal of Advanced Research in Computer and Communication Engineering* 2015. 4 (3) p. .

[Iyer et al. ()] 'Diagnosis of diabetes using classification mining techniques'. A Iyer , S Jeyalatha , R Sumblay . *IJDKP* 2015. 5 (1) p. .

[Tan ()] 'Neighbor-weighted K-nearest neighbor for unbalanced text corpus'. S Tan . `http://orange.biolab.si/` *Expert Systems with Applications* 2005. 28 (4) p. .

[Orange Data Mining Library Documentation Release] *Orange Data Mining Library Documentation Release*, 3. (Orange Data Mining)

[Prerana et al. ()] 'Prediction of Heart Disease Using Machine Learning Algorithms-Naïve Bayes, Introduction to PAC Algorithm, Comparison of Algorithms and HDPS'. T H M1 Prerana , N Shivaprakash , C2 , N3 Swetha . *International Journal of Science and Engineering* 2347-2200. 2015. 3 (2) p. .

[Kirkby ()] *WEKA Explorer User Guide for version 3-3-4*, R Kirkby . 2002. University of Weikato