International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in; editorijoer@gmail.com

Vol.5., Issue.6, 2017
Nov-Dec

RESEARCH ARTICLE

ISSN: 2321-7758

COMPARATIVE ANALYSIS OF DATA AND TEXT MINING TOOLS IN BIOINFORMATICS

SARANGAM KODATI[1], Dr. R VIVEKANANDAM[2]

[1]Research Scholar, Department of Computer Science and Engineering, Sri Satya Sai University of Technology and Medical Science, Sehore,Bhopal,Madhya Pradesh , (India)

[2]Professor, Department of Computer Science and Engineering,Sri Satya Sai University of Technology and Medical Science,Sehore,Bhopal, Madhya Pradesh , (India)

ABSTRACT

Heart disease has been recognized as like the leading cause of death throughout the world.. The researchers accelerating their research works to boost a software with the help machine learning algorithm which perform help doctors to take a decision regarding both prediction and diagnosing of heart disease. Unraveling the molecular complexities of human heart disease, particularly end-stage heart disease, can be achieved by combining multiple investigative approaches in bioinformatics. There are several parts to the problem. Each patient is the product of a complex set of genetic variations, different degrees of influence of diets and lifestyles, and usually heart transplantation patients are treated with multiple drugs. we recognize the need for bioinformatics to make sense of the large quantities of data that will flow from our laboratories. Thus, we plan to provide meaningful molecular descriptions of a number of different conditions that result in terminal heart failure. The main objective over this research paper is predicting the heart disease regarding a patient the use of data mining ,text mining tools and machine learning algorithms. Comparative study concerning Naïve Base Classifier, K-nearest neighbor, Support vector machines and Random Forest the precision and recall about machine learning algorithms is performed through a graphical representation concerning the results.

Keywords--Data mining, data mining classification algorithms, Data mining and Text mining tools, Heart disease, Bioinformatics .

.

1. INTRODUCTION

In the past decade, heart disease has been the leading cause of death in different continents and countries in the world, regardless of the income level of countries . According to WHO report, heart disease is the leading cause of death across the world, accounting for 7.2 million deaths, i.e., 12.8% of all fatalities in the world [1], illustrates deaths from heart disease across the world (scale: 1:100000). According to recent research predictions, cardiovascular diseases will become the leading cause of death up to 2030. Although cardiovascular diseases have been identified as the leading cause of death in the world in the past decade, they have been introduced as the most preventable and controllable diseases. The complete and correct treatment of a disease depends on the timely diagnosis of that disease . An accurate and systematic tool for identifying high-risk patients and extracting data for timely diagnosis of heart disease seems a critical need.

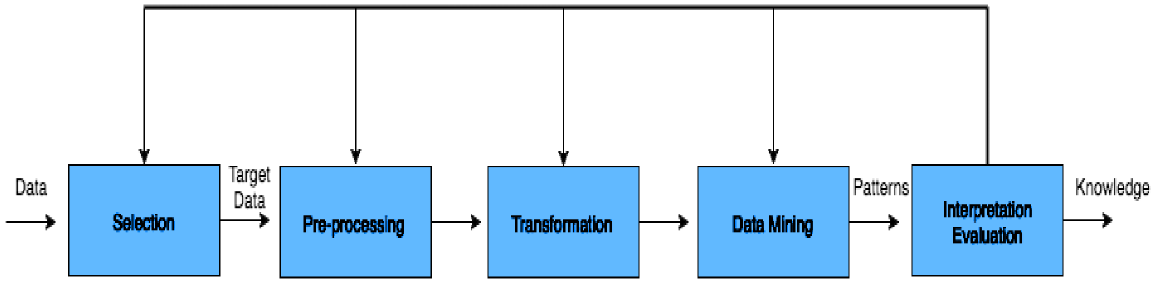

Fig 1: Knowledge Discovery in Databases Process(KDD)

Knowledge Discovery (KDD) Process – Data mining—core of knowledge discovery process Pattern Evaluation Data Mining Task-relevant Data and Data Warehouse Data Cleaning Data Integration Databases ,Data cleaning for remove Noise and Inconsistent Data ,Data integration for where multiple data sources may be combined ,Data selection for mining by performing summary or aggregation operation ,Data Mining • An essential Process where intelligent methods are applied to extract data patterns, Pattern Evolution to identify the truly interesting patterns representing knowledge based on interestingness measures, Knowledge representation where visualization and knowledge representation techniques are used to present mined knowledge to users

Every day, modern computer-based systems collect large amounts of data using automatic data record systems in different fields[2]. Data mining technology is the product of the evolution of database technology, IT and storage devices [3]. The current challenges is according to make data mining and knowledge discovery systems applicable in accordance with a wider range regarding domains [4]. Researchers are adopting data mining strategies to diagnose one of a kind diseases who consists of diabetes , stroke , cancer and heart disease . Considering the high rate about cardiovascular induced fatalities[5], researchers have tried in imitation of adopt data mining structures to diagnose heart disease [7].

2. DATA MINING AND TEXT MINING IN BIOINFORMATICS

Data mining according to bioinformatics include heart disease , protein feature domain detection, characteristic motif detection, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein or gene interplay network reconstruction, data cleansing[6], and protein sub-cellular location prediction. For example, microarray technologies are used according to predict a patient’s outcome. On the basis concerning patients’ genotypic microarray data, their survival time and risk about tumor metastasis or recurrence can remain estimated. Machine learning be able stay used for peptide identification through mass spectroscopy. Correlation amongst fragment ions among a tandem mass spectrum is crucial amongst reducing stochastic mismatches for peptide identification by using database searching. Bioinformatics is the area that combines computer science, information technology, and biology. Tools provided through bioinformatics help scientists analyze and explain various types regarding data, including sequences over amino acids, numerical or textual data[18]. Research areas in the area of bioinformatics[6] include heart disease, genome annotation, literature mining, and analysis of many other biological subjects[8]. Beside others, literature mining is the key area to that amount deals with the analysis and interpretation about textual data and it is done by using the help of the text mining methods.

3. DATA MINING AND TEXT MINING TOOLS

3.1 WEKA Tool

Waikato Environment because Knowledge Analysis or WEKA is an open source software, developed into Java, issued under the GNU General Public License. Weka is basically a collection of machine learning algorithms because data mining tasks, such so data pre-processing, visualization, classification, regression and clustering[16].

3.2 Orange Tool

Orange is an open source machine learning technology and data mining software. Orange can remain used for exploration data analysis and visualization[14]. It gives a platform because experiment selection, predictive modeling, and recommendation systems and to be used among gnomic research, bio medicine, bioinformatics, and teaching. It is intended because both experienced users and researchers about machine learning any want to prototype new algorithms while reusing as like much of the code as much possible, and because of those just entering the field who perform either write short Python scripts because of data analysis and enjoy into the robust while easy-to-use visible programming environment[15]. Orange is always preferred so the component of innovation, quality, or reliability is involved. Orange includes a range about techniques, such and data management and preprocessing, supervised and unsupervised learning, performance analysis, and a range of data and model visualization techniques.

3.3 Text Analyzer Tool

Text analyzer is Free software utility which allows you to find the most frequent phrases and frequencies of words. The text analyzer is support with heart disease and bioinformatics. Non-English language texts are supported. It also counts number of words, characters, sentences and syllables. Also calculates lexical density.

4. PRECISION AND RECALL

Recall (**R**) and Precision (**P**) are measures that are based on confusion matrix data. Recall (R) is the portion of instances that have true positive class and are predicted as positive. On the other hand, Precision (P) is the probability of that a positive prediction is correct as shown in

$$\mathbf{R} = \frac{\mathbf{TP}}{\mathbf{CN}} \text{ and } \mathbf{P} = \frac{\mathbf{TP}}{\mathbf{RN}}$$

Classification Accuracy (**Acc**) is the most used measure that evaluates the effectiveness of a classifier by its percentage of correctly predicted instances as in

$$\mathbf{ACC} = \frac{\mathbf{TP + TN}}{\mathbf{N}}$$

5. HEART DISEASE DATASET

The data used in this study is the benchmark Cleveland Clinic Foundation Heart disease data set available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease. The data set has 76 raw attributes. However, all of the published experiments only refer to 13 of them. The data set contains 303 rows of which 297 are complete. Six rows contain missing values and they are removed from the experiment[17].

5. DATA MINING CLASSIFICATION ALGORITHMS

5.1. Naïve Base Classifier

Naïve Bayes is a simple probabilistic classifier based about applying Bayes theorem with strong independence assumptions. A more descriptive term because underlying probability model would be "independent feature model". In simple terms, a Naïve Bayes classifier assumes as the presence of a specific feature of a class is unrelated in imitation of the presence of someone other feature.[10] Depending on the specific makeup of the chance model, Naïve Bayes classifiers can be trained very sufficiently of a supervised learning setting. In many practical applications, parameter discernment for Naïve Bayes models use the approach of maximum likelihood. the predictive accuracy is reduced.

5.2 . Support Vector Machine

Support Vector Machines proved themselves to be very effective in a variety of pattern classification tasks and thus received a great deal of attention recently. Support vector machine is a supervised machine learning technique. The SVM algorithm predicts the occurrence of heart disease by plotting the disease predicting attributes in multidimensional hyper plane and classifies the classes optimally by creating the margin between two data clusters. This algorithm attains high accuracy through the use of nonlinear functions called kernels[9]. Each support vector is characterized with an equation describing the boundary line of each class. Support Vector Machine[8].

5.3. Random Forest

Random Forest consists of decision trees. Every decision tree is formed by subset of training data which randomly selected. The decision tree is a approach because displaying a series regarding laws that are leading to a category or value. The difference between the methods of decision tree is that how the distance to be measured. Decision trees that are used in accordance with predict the cluster variables called classification trees because they are located the samples within clusters or classes. Every decision tree in Random Forest provides results for classification and final results of

Random Forest, is that most of the trees have announced. To build Random Forest, such can lie preserved a number about decision trees so much need to exist of the forests. One over the advantages concerning Random Forest is that it requires insignificant preprocessing. Also there is no need to choose the required variables at the beginning and Random Forest model itself chooses the useful variables [9].

5.4 . K-Nearest Neighbor

Nearest Neighbor algorithms are among the simplest on all machine learning algorithms. The thought is to memorize the training employ and afterwards after predict the label regarding some new instance about the basis of the labels regarding its closest neighbors into the training set. The rationale behind such a technique is based on the assumption that the features so much are used to construct the area points are relevant to theirs labeling of a course to that amount makes close by factors in all likelihood to have the same label[11]. Furthermore, within some situations, even when the training set is immense, finding the nearest neighbor can be done extremely fast.

6. CLASSIFICATION ALGORITHMS TABLES

After selecting the data sets, a number of classification algorithm are chosen for conducting the test. Many classification algorithms such as Naïve Bayes (NB) algorithm, K Nearest Neighbor (KNN) algorithm, Support Vector Machine (SVM) algorithm, and RandomForest algorithm.

Table 1: Classification Algorithm Precession and Recall in Orange Tool

Algorithm classification Average	Precession in Orange	Recall in Orange
Naïve base classifier	0.824	0.806
Support Vector Machine	0.817	0.705
RandomForest	0.779	0.734
K-Nearest Neighbor	0.58	0.547

Table 2: Classification Algorithm Precession and Recall in WEKA Tool

Algorithm classification Average	Precession in WEKA	Recall in WEKA
Naïve base classifier	0.837	0.837
SMO or Support Vector Machine	0.84	0.8365
Random Forest	0.818	0.819
1BK or K-Nearest Neighbor	0.753	0.752

7. RESULTS

In this section, we study the diagnosis Bioinformatics of heart disease, according to results of data mining techniques such classification algorithms such as Naïve Bayes (NB) algorithm, K Nearest Neighbor (KNN) algorithm, Support Vector Machine (SVM) algorithm, and RandomForest algorithm. That each one explained before then we compare them to decide which is more accurate in the diagnosis of heart disease. The data set has 76 raw attributes. However, all the published experiments only refer to 13 of them. The data set contains 303 rows of which 297 are complete According to the results Figures 7, respectively, shows the comparison of the accuracy of these criteria: Precision and Recall using Orange tool and Weka Tool..

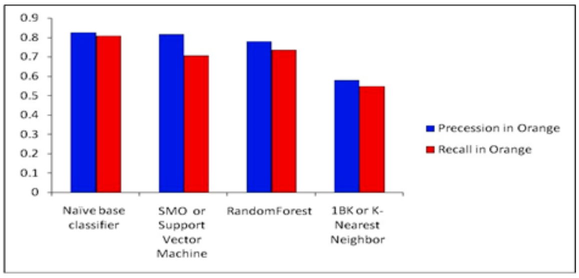

Fig 2: Classification Algorithm Graph for Precession and Recall in Orange Tool

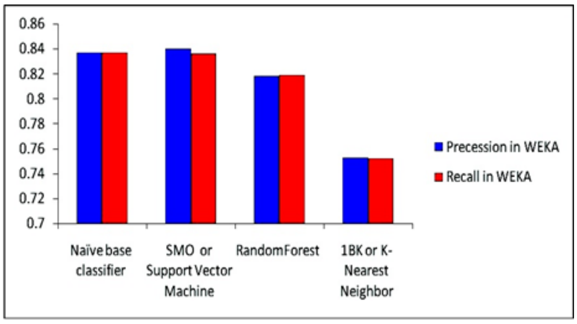

Fig 3: Classification Algorithm Graph for Precession and Recall in WEKA Tool

8. CONCLUSION

Heart disease is the lead cause concerning death in the world. It accounts for 7.2 million deaths, i.e., 12.8% regarding fatalities among the world. Although cardiovascular diseases bear been recognized so the leading cause of death in the past decade, they are the most preventable and controllable diseases at the same time. Deaths from bioinformatics of heart diseases show an ever-increasing trend. We also choose one dataset from heart available at UCI machine learning repository. Comparative analysis concerning precession and recall weka is the best overall performance compared to an orange. main objective concerning this paper is to compare the data mining tools on the basis of theirs classification precession and recall. According to the result of three data mining tools used in this paper, such has been observed so different data mining tools are furnishing different results concerning same data set with different classification algorithm. WEKA and ORANGE are showing best classification Precession and Recall. In future, more disease dataset can be used for classification methods, and other data mining techniques such as clustering can be used according to compare the performance of various data mining tools.

REFERENCES

[1]. World Health Organization (2013) Deaths from coronary heart disease.

[2]. Dunham, M. H., Sridhar S, Data Mining: Introductory and Advanced Topics, Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006 .

[3]. Panzarasa S, Quaglini S, Sacchi L, Cavallini A, Micieli G, et al. (2010) Data mining techniques for analyzing stroke care processes. In the Proc. of the 13th World Congress on Medical Informatics.

[4]. Ralf Mikut and Markus Reischl Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 1, Issue 5, pages 431–443, September/October 2011.

[5]. Lakshmi K, Krishna MV, Kumar SP (2013) Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability. International Journal of Scientific and Research Publications 3: 1-10.

[6]. M H Dunham, Data Mining: Introductory and Advanced Topics, Prentice Hall, 2002.

[7]. Obenshain MK (2004) Application of data mining techniques to healthcare data. Infection Control and Hospital Epidemiology 25: 690-695.

[8]. Soman K P, Loganathan R and Ajay V, “Machine Learning with SVM and Other Kernel Methods”, PHI, India, 20

[9]. Ho, Tin Kam (1995). Random Decision Tree (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp.278–282.

[10]. Prerana T H M1, Shivaprakash N C2 , Swetha N3 ”Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes,Introduction to PAC Algorithm, Comparison of Algorithms and HDPS” International Journal of Science and Engineering Volume 3, Number 2 – 2015 PP: 90-99 ©IJSE Available at www.ijse.org ISSN: 2347-2200

[11]. S. TAN,"Neighbor-weighted K-nearest neighbor forunbalanced text corpus", Expert Systems with Applications, Vol. 28, No. 4, pp. 667-671, 2005.

[12]. A. S. Abdullah, R. R. Rajalaxmi, “A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier”, International Conference on Recent Trends in Computational Methods, Communication and Controls (ICON3C 2012), ICON3C(3), pp.22-25, April 2012.

[13]. Das R, Turkoglu I, Sengur A (2009) Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications 36: 7675-7680.

[14]. http://orange.biolab.si/

[15]. Orange Data Mining, ‘Orange Data Mining Library Documentation Release 3’.

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

[17]. V.A. Medical Center, Long Beach Clinic Foundation, “Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease, [Last Accessed 5 November 2015].

[18]. A. M. Cohen and W. R. Hersh, ―A survey of current work in biomedical text mining,‖ (Briefings in bioinformatics, vol. 6, no. 1, pp. 57–71, 2005.)