DOI: http://dx.doi.org/10.26483/ijarcs.v13i3.6832

ISSN No. 0976-5697

Volume 13, No. 3, May-June 2022

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

DATA MINING IN DISEASE PREDICTION

Archana Thakur
School of Computer Science and IT,
Devi Ahilya Vishwavidyalaya, Indore – 452001,
Madhya Pradesh, India.

Abstract: Enhancement in information technology has led to design of many applications in the field of crop disease recognition. Disease recognition applications generate voluminous data. The disease related data can be processed using data mining techniques to predict various diseases. Data mining is a field of analysing, extracting data for furnishing new knowledge which represents the relationship between different patterns of data. Some of the data mining methods include classification, clustering, prediction and association rule mining. In the present work data mining is used for disease prediction.

Keywords: Data mining; clustering; classification; prediction.

I. INTRODUCTION

All Major advances in information technology yields extreme expansion of data in disease recognition application [1]. Disease recognition data consists of data related to various crop diseases and data of treatment cost. This voluminous data is created from various sources and formats. Disease related data may contain irrelevant attributes and missing data. Application of different data mining methods is a vital approach to mine knowledge from huge disease related data. Data mining offers a variety of methods to mine knowledge from large disease data sets. Data mining methods of classification, clustering and association rule mining can be employed to examine data and extract significant information. Various applications of data mining in disease recognition include forecasting the future occurrences of diseases based upon the prior data collected from various diseases, identification of disease based upon crop's data, analyzing the treatment costs and demand of various resources, preprocessing of noisy and missing data, reducing the time to wait for the recognition of various diseases. Data mining tools like Orange, Weka and Rapid miner [2, 3, 4] are used in the present work to monitor and forecast better results for disease recognition task. The novel and existing data mining tools and method are employed for recognition of disease(s) to enhance the disease recognition services in a cost-efficient manner and thereby reducing the time for disease prediction. The organization of the work is as under. Section II explains the basic concepts of data mining. The different data mining methods useful in disease prediction are discussed in Section III. The data mining tools helpful for disease recognition are discussed in section IV. Results and discussions are presented in section V. Conclusions and scope of future work are discussed in section VI.

II. BASIC CONCEPTS OF DATA MINING

Data mining is a stream of extracting unknown knowledge from huge data sets. Extraction of helpful knowledge from the massive data sets and offering decision-making results for the identification and treatment of diseases is the most vital step. Data mining can be used to mine knowledge by examining and forecasting different diseases. Data mining for disease recognition has enormous potential to determine the concealed patterns in the datasets of various diseases. Data mining methods can be used with their appropriateness depending upon the disease data. Data mining applications in disease prediction field are efficient and have enormous potential. It automates the process of searching predictive information in huge databases. Disease prediction plays a vital role in data mining. Searching of a disease needs the performance of a number of tests upon the crop. However, use of data mining methods, can decrease the number of tests. This reduced test set plays an important role in terms of time and performance. Data mining in disease prediction is a vital field as it permits experts to observe which attributes are more significant for prediction, for example, leaf symptoms, seed symptoms, temperature etc. This will help the disease experts to identify the disease and suggest treatment more efficiently. Knowledge discovery in databases can be achieved using data mining. Knowledge discovery in databases is the process of determining helpful information and patterns in data. It employs the algorithms to extract the information and patterns derived by the new knowledge discovery in data mining process. During the selection stage, it acquires the data from various resources. In preprocessing stage, it eliminates the undesired noisy, missing data and furnishes the clean data which can be transformed to a common format in transformation stage. Then data mining methods are applied to acquire the desired output. Finally, during the interpretation stage, it presents the result to the end user in a significant manner.

III. DATA MINING METHODS

Data mining methods like clustering, classification and association rule mining are highly useful in disease data analysis.

A. *Classification*

One of the famous machine-learning based data mining methods is classification. It is employed to classify each instance in a set of data into one of predefined set of groups or classes. It exercises mathematical methods like linear

programming, Decision Trees, Artificial Neural Networks and Statistics to categorize the data into various classes or groups. Recent classification methods offer more intelligent methods for effective forecasting of diseases [5]. Different types of classification methods include Support Vector Machine, Naive Bayes, Discriminant Analysis, Random Forest and Decision Trees.

B. Clustering

Clustering is a data mining method that creates clusters of objects that have comparable traits using automatic method. It defines the classes and assigns objects in them where the class is unknown. Different types of cluster methods include Fuzzy CMeans (FCM), Rough-Fuzzy CMeans (RFCM), K-means, Rough C-means (RCM), Robust RFCM (rRFCM), hierarchical and Gaussian mixture.

C. Association rule mining

Association rule mining is a very famous method for determining interesting relations amongst various data in big databases. It is meant to recognize well designed rules discovered in databases employing various methods of importance based upon the input dataset. It is the data mining process of determining the rules, searching numerous patterns, correlations, associations, or some causal structures among groups of items that may monitor associations and causal objects amongst sets of items. Understand customer buying habits by searching associations & correlations amongst the various items that customers put up in their "shopping basket". The major applications of association rule mining include basket data analysis, cross-marketing and catalog design. The above data mining methods can be employed for the prediction of various diseases [6].

IV. DATA MINING TOOLS

Data mining tools like RapidMiner, Weka and Orange are used to analyze the performance of data mining method.

A. RapidMiner

RapidMiner [8] is an open-source data mining tool which offers a flexible environment for data mining processes. It has the ability of drag-and-drop which is exercised to create the dataflow. It offers different file formats. Classification, regression and clustering jobs can be performed easily with various learning algorithms. This tool holds up a large number of the Classification and Regression Trees (CART), Decision Trees, Association rule mining method, Clustering algorithms and several features are provided for data preprocessing, filtering, normalization and data analysis. It can import data from various standard and traditional databases.

B. Orange

Orange [9] is an open-source data mining tool designed at the laboratory of Bioinformatics at the University of Ljubljana. The applications can be executed exercising the scripting and visual programming. Python library is provided for manipulation of data and widget alteration. Here programming is done by placing widgets on the canvas and joining their inputs and outputs. This tool is appropriate for various machine learning and data mining algorithms. It can be easily exercised by both researchers of data mining and naive users who want to design and test their own algorithms. It gives us benefit of reusing as much of the code as possible.

C. Weka

The Waikato Environment for Knowledge Analysis (WEKA) [7] is a popular open-source software and data mining toolkit set up by Waikato University, New Zealand. It supports various standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. Novel algorithms can also be executed using WEKA with current data mining and machine learning methods. It provides different sources for loading the data, including files, URLs and databases. It supports file formats including ARFF format, CSV, Lib SVMs format, and C4.5's format. Various performance evaluation measures are also offered in WEKA such as confusion matrix, precision, recall, true positive and false negative, etc. Some of the major benefits of Weka include Open source, platform independent and portable software, and easy to use graphical user interface with a very large collection of different data mining algorithms.

V. RESULTS AND DISCUSSION

Random Forest is a well-known classification algorithm. It is an ensemble of various Decision Trees. Random Forest algorithm is executed on RapidMiner, Orange and Weka for predicting various vegetable crop diseases [10]. The performance observations are presented in Table 1 and Fig.2.

Classification method	Tools	Prediction Accuracy (in %)
Random Forest	Weka	96
	Orange	85
	RapidMiner	95

Table 1- Prediction Accuracy observations

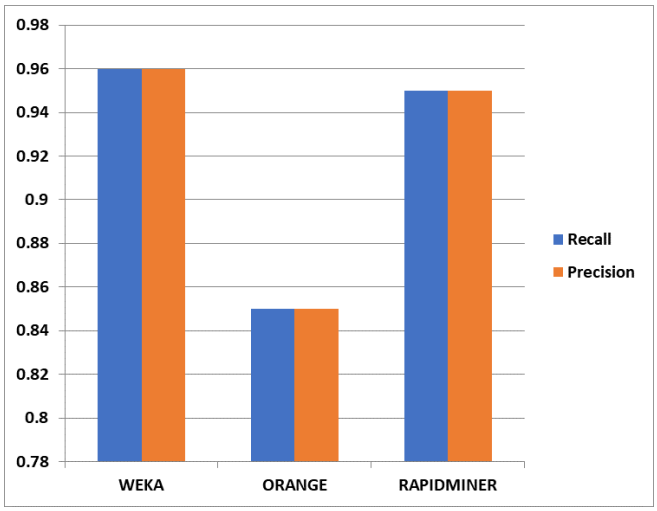

Fig. 2 – Precision and Recall observations.

VI. CONCLUSIONS AND FUTURE SCOPE OF WORK

Based upon comparative performance investigation the following results have been observed from the above-mentioned data mining tools. Weka is an easy to learn and the best tool for a beginner. It has many in-built and experimental features and no previous knowledge of coding is needed. RapidMiner is the only tool which is independent of limitations of any language and has good statistical and predictive analytical ability. Orange and RapidMiner are the data mining tools that are useful for advanced users. Both require prior

knowledge of coding. The performance differences observed on different tools are due to treatment with null values, missing values present in data. Performance measures like True Negatives and False Positives may also differ while executing an algorithm on different data mining tools. As a part of future work, more diseases can be considered with different classification methods. Some different data mining methods such as clustering, association rule mining can be employed to examine the performance of different data mining tools.

VII. REFERENCES

[1] Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, I. Chouvarda, "Machine learning and data mining methods in diabetes research", Computational and structural biotechnology journal, pp.104-116, 2017.

[2] P.H. Patil, S. Thube, B. Ratnaparkhi, K. Rajeswari, "Analysis of Different Data Mining Tools using Classification, Clustering and Association Rule Mining", International Journal of Computer Applications, Vol. 93, Issue No. 8, pp. 35-39, 2014.

[3] D. Usha Rani, "Survey on data mining tools and techniques in medical field", International Journal of Advanced Networking & Applications, Vol. 8, Issue No. 5, pp. 51-54, 2017.

[4] S.K. Devi, S. Krishnapriya, D. Kalita, "Prediction of Heart Disease using Data Mining Techniques", Indian Journal of Science and Technology, Vol. 9, Issue No. 39, pp. 1-5, 2016.

[5] N. Bhatla, K. Jyoti, "An analysis of heart disease prediction using different data mining techniques", International Journal of Engineering, Vol. 1, Issue No. 8, pp.1-4, 2012.

[6] M. Fatima, M. Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic", Journal of Intelligent Learning Systems and Applications, Vol. 9, Issue No. 1, pp.1-16, 2017.

[7] http://www.cs.waikato.ac.nz/ml/weka/ [Last accessed: 22/08/2017].

[8] https://rapidminer.com/ [Last accessed: 22/08/2021].

[9] https://orange.biolab.si/ [Last accessed: 22/08/2021].

[10] A. Chaudhary, R. Thakur, S. Kolhe, R. Kamal, "A particle swarm optimization based ensemble for vegetable crop disease recognition", Computers and Electronics in Agriculture, Vol. 178, pp. 1-7, 2020.