INTERNATIONALJOURNALOFENGINEERING SCIENCES&MANAGEMENT

ANALYSIS AND CLASSIFICATION OF ALGORITHMS FOR BREAST-CANCER & DISEASES PREDICTION USING THE DATA MINING WEKA TOOL

[1]Poorva Pathak and [2]Chinmay Bhatt

[1]M.Tech Scholar, Department of Computer Science Engineering,
SRK University, Bhopal (M.P.)

[2]Assistant Professor, Department of Computer Science Engineering,
SRK University, Bhopal (M.P.)

ABSTRACT

In the 21st century medical & healthcare facilities are at their disposal vast amounts of breast cancer patients' data. The identical analysis of available data can be used to facilitate more patients and will lead the decision-making efficiency. During the study it was found that extraction of relevant knowledge from this data and act upon it in a timely manner is challengeable. To troubleshoot this problem and turn into knowledge, use of efficient computing and data mining tools has been suggested. The classification and analysis of this data can aid in developing expert systems for decision support in breast cancer and other diseases. Also this can reduce the cost, the waiting time, and be uses as troubleshooting tools to liberate medical practitioners for more research and reduce errors and mistakes. A research work has been conducted in SRBC Bhopal and analytical data has been used for classification of algorithms for breast Cancer \disease prediction using the data mining Weka Tool. Effective data mining tools can assist in early detection of diseases such as breast cancer.

Keywords- Weka tool; Diseases Prediction;Data mining;

INTRODUCTION

This research includes analysis and classification results of the diagnosis of breast cancer using data mining techniques. The world is facing the problem of many diseases and Breast cancer is the most common cancer disease among Women from many decades. The two types of breast cancer observed more, i.e. malignant and benign, the malignant tumour develops when cells in the breast tissue divide and grow abnormally without the normal controls on cell death and cell division. Many developed and industrialized nations such as Canada, the United States, Australia, and countries in Western Europe witnessed the highest incidence rates of breast cancer. Although breast cancer is the second leading cause of cancer death in women, still the survival rate is high once it is detected early [1].

Medical database today have taken the shape of a large number of terabytes. Within these masses of information, data of vital importance is hidden. Because of these vast measures of information, it makes one wonder at that point, "How would you make important determinations about this information?" Data mining addresses this investigation.

Breast Cancer classification

Cancer in human body developed because of abnormal development of cells which are the building blocks that make up all tissues and organs of the body. When normal cells grow, old get damaged, they die, and new cells take their place. Sometimes, this process goes wrong. New cells form when the body doesn't need them, and old or damaged cells don't die as they should. The build-up of extra cells often forms a mass of tissue called a lump, growth, or tumour [2].

Types of tumours benign (not cancer) or malignant (cancer): *Benign tumours* Benign tumours are usually not harmful. They rarely invade the tissues around them. These tissues don't spread to other parts of the body and can be removed and usually don't grow back. *Malignant tumours* Malignant tumours are opposite to that of benign tumours. Malignant tumours are very harmful and may be threat to life and can invade nearby organs and tissues .These tissues can spread to other parts of the body. They can often be removed but sometimes grow back[3].

REVIEW OF LITERATURE

This section contains the description of the literature that has been done on Breast Cancer Analysis and classification techniques. Comparative study of different classification techniques is summarized with advantages and disadvantages. Most of the key features, methods are mentioned below with respective limitations and benefits that make our work unique. Clinical diagnosis of breast cancer helps in predicting the malignant cases. Various common methods used for breast cancer diagnosis are Mammography, Positron Emission Tomography, Biopsy and Magnetic Resonance Imaging. This section includes the review of various technical and review articles on data mining techniques applied in breast cancer diagnosis. Ireaneus. Y et al., discussed about the tumors in early detection of Breast Cancer Using SVM Classifier Technique. The mammogram is divided into three main stages [4].

The first step involves an enhancement procedure and image enhancement techniques. To makes certain features it is easy to modify the colors or intensities in image by increasing the signal and noise ration. Then the features are extracted from the segmented mammogram. The next stage involves the classification using SVM classifier. K. Rajesh, et.al [5] classified about SEER breast cancer data into the groups of "Carcinoma in situ" and "Malignant potential" using C4.5 algorithm. They obtained an accuracy of 94% in the training phase and an accuracy of 93% in the testing phase. They have compared the performance of C4.5 algorithm with other classification techniques.

Vanaja S., K. Ramesh kumar, [6] discussed about theC4.5 classification algorithm is the extraction of ID3 algorithm. It support continuous attributes and shows the best accuracy on attribute with missing values. The information gain by attribute measurement, which indicates the percentage by given attribute and separate dataset according to their final classification. Both C4.5 and C5.0 can produce classifiers expressed either decision trees or rule sets. In many applications, rule sets are preferred because they are simpler and easier to understand than decision trees, butC4.5's rule set methods are slow and need more memory.C5.0 embodies new algorithms for generating rule sets and the improvement is substantial.

During the 1980s and 1990s, PCs had steadily gained momentum on different occasions since the inception of the investigation, educational ventures appeared, new courses were introduced, and sponsorships were opened. All referenced components encouraged experts to focus on new philosophies for neural network application, correction and prediction, foreshadowing, and investigation

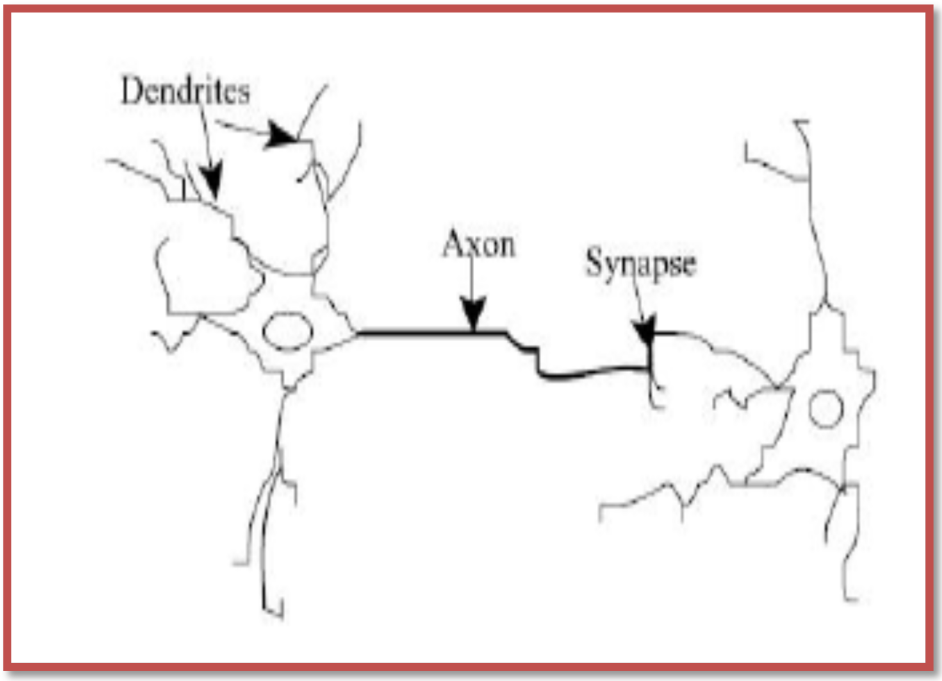

Fig:. Structure of Human neuron

RESEARCH DESIGN

This research was a positivist research (scientific) and dependent on using the principles of prediction upon preceding history and data manipulations (manipulating in such manner, doesn't include change in information structure or qualities. however, manipulating data is the way toward filling missing element value, data normalization, treating noisy data, so on)

This research design was based on survey, observation and reasoning as a tool for understanding a certain problem or behavior, i.e. this was an experimental research design. The design involved manipulations towards variables and prediction s on the root of preceding examination or history. The study was concerned with what could be the cause of a particular relationship and what the effects of that relationship could be.

Methods

This section contains two sub-sections. In the first subsection, data mining is described, and in the second subsection, classification algorithms are explained.

Data Mining

Knowledge discovery for DM techniques indication to take out useful patterns and relationships from embrace Databases. Due to the gigantic total of data, and to obtain valuable output, a systemic technique that was applied in the research was represented in the figure

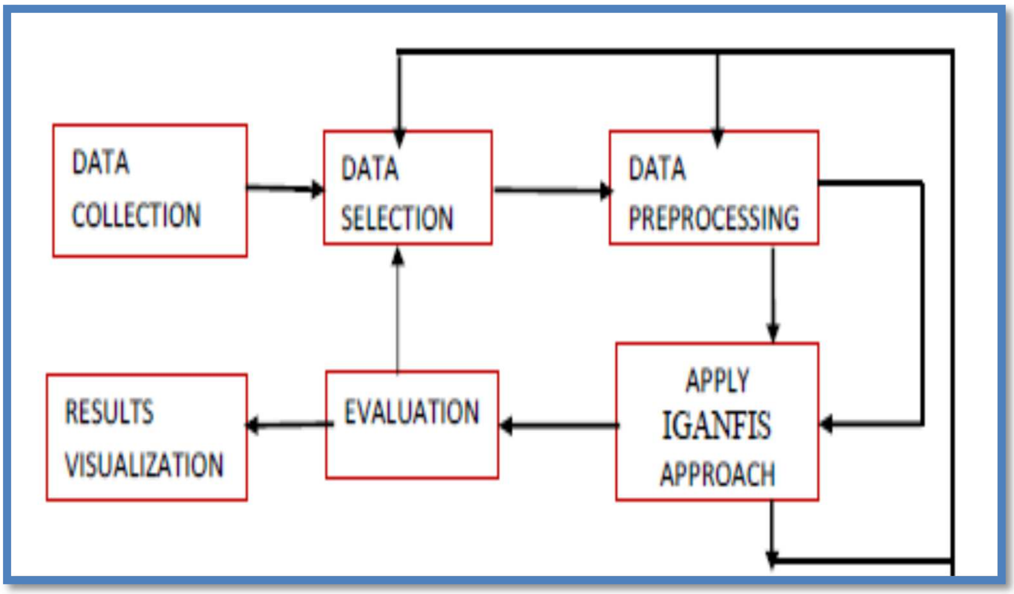

Figure: Data mining Conceptual framework

Data mining is a statistical process whereby data is taken from a data warehouse and compiled, organized and interpreted. Examination of the relevant literature [7-8] has identified that Weka software, a free data mining tool, is frequently used and a productive tool. Weka, which was developed by Waikato University in New Zealand, implements data mining algorithms using Java. This open-source software is consists of machine learning algorithms for data mining tasks. Weka applies algorithms for data pre-processing, data visualization, classification, clustering, regression, and association rules. Classification algorithms are the most commonly applied data mining method [9].

Recent studies have shown the data mining algorithms application in medicine. Various machine learning and data mining algorithms can be implemented in the field of medicine as a second opinion diagnostic tool and as a tool for the information extraction stage of the knowledge discovery process in databases system.

The main reason to choose Weka was is that it is frequently used as per the requirement of literature. In the begining stage, the dataset from other sources was simulated under a single data set and transformed to *arff* format - Weka's file format. Following these procedures, classification algorithms were run in Weka software to determine whether patients had breast cancer/tumors.or other deceases.

Classification Algorithms

Classification algorithms predict one or more discrete variables as required, based on the other attributes in the dataset. Data mining software is required to run the classification algorithms. In Weka, there are 8 classes, namely bayes, functions, lazy, meta, mi, misc, rules and trees. In our study, the class, mi, did not run and produce any results because of the data format[10].

This section includes the definitions of the classification algorithms which were successful and efficient as well as an explanation of how they work, are presented. In other words, the Weka classes and algorithms that produced the most successful results which are Bagging, IBk, Random Committee, Random Forest, and Simple Classification and Regression Tree algorithms were identified, and are described below.

- J48 decision tree
- LMT Tree
- REP Tree
- Naive Bayes
- Fuzzy

The study used UCI machine learning repository. The SRBC Databases attributes were collected from RKDF medical college digital fine needle aspirate (FNA) of breast mass. A summary of the SRBC datasets from UCI that were used in this study are shown in table 1.

Table 1: Sarvepalli Radhakrishnan Breast cancer (SRBC) Parameter

Attribute	**Domain**
Clump Thickness	1-10
Uniformity of Cell Size	1-10
Uniformity of Cell Shape	1-10
Marginal Adhesion	1-10
Bare Nucleoli	1-10
Single Epithelial Cell Size	1-10
Bland Chromatin	1-10
Normal Nucleoli	1-10
Mitoses	1-10
Class	(2 for benign, 4 for malignant)

This work used SRBC (Original), Sarvepalli Radhakrishnan Prognosis breast cancer (SRPBC), and Sarvepalli Radhakrishnan Diagnosis breast cancer (SRDBC), from UCI repository to find the best classifiers. Our intention here was to come up with the best combination of classifiers that best classifies breast cancer patient's data; these data sets were represented as in the table 2

Table 2: SRBC (Original), SRDBC, and SRPBC dataset

Dataset	Number of features	Number of Instances	Number of Classes
Sarvepalli Radhakrishnan Breast cancer (SRBC) Original	11	699	2
Sarvepalli Radhakrishnan Diagnosis breast cancer (SRDBC)	32	569	2
Sarvepalli Radhakrishnan Prognosis breast cancer (SRPBC)	34	198	2

CONCLUSION

In this paper, the different data mining and machine learning algorithms are studied to predict the breast cancer using different data sets and different data mining algorithms. The accuracy depends upon the data mining algorithms. Future work can be done to analyze the data set of breast cancer using data mining classification algorithms because classification algorithms show better result than the clustering algorithms. To focus on the accuracy and precision, classification algorithms can show the better performance in predicting Breast Cancer.

REFERENCE

[1].G. Holmes; A. Donkin and I.H. Witten (1994). "Weka: A machine learning workbench". Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia. Retrieved 2007-06-25.

[2].Vikas Chaurasia, Saurabh Pal "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability", International Journal of Computer Science and Mobile Computing, (2014) Vol. 3, Issue 1, pp. 10 – 22.

[3].Rajni Bedi and Ajay Shiv Sharma "Classification Algorithms for Prediction of Lumbar Spine Pathologies", IEEE International Conference on advanced informatics for computing research, (2017) pp. 42–50

[4].Y.Ireaneus Anna Rejani, Dr.S.ThamaraiSelvi, "Early Detection Of Breast Cancer Using Svm Classifier Technique", Int. Journal on Computer Science and Engineering, Vol. 1, Issue 3, 2009, pp. 127-130.

[5].Rajesh,k., Dr.SheilaAnand,"Analysis of SEER Dataset for Breast Cancer Diagnosis usingC4.5 Classification Algorithm", IJARCCE,Vol.1, Issue 2, 2012, pp. 2278-1021.

[6].Vanaja, S., K. Rameshkumar, "Performance Analysis of Classification Algorithms on Medical Diagnoses-a Survey", Journal of Computer Science, Vol. 11, 2015, pp. 32-33.

[7].Danacı, M., Çelik, M., & Akkaya, A. E. (2010). Prediction and diagnosis of breast cancer cells using data mining methods. *ASYU'2010*, Kayseri, Turkey, 9-12.

[8].Aydın, E. A. & Keleş, M. K. (2017). Breast cancer detection using K-nearest neighbors data mining method obtained from the bow-tie antenna dataset. *International Journal of RF and Microwave Computer-Aided Engineering, 27*(6). https://doi.org/10.1002/mmce.21098

[9].Kaya Keleş, M. (2016). A Comparison of Statistical Methods and Data Mining Methods. *Papers on Social Science, ICOMEP Special Issue*, 20-24.

[10]. Kaya, M., Keleş, A. E., & Oral, E. L. (2014). Construction Crew Productivity Prediction by Using Data Mining Methods. *Procedia - Social and Behavioral Sciences*, 141, 1249-1253. https://doi.org/10.1016/j.sbspro.2014.05.215