Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 50 (2015) 551 – 553

Procedia
Computer Science

2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

PROGONOSIS AND MODELLING OF BREAST CANCER AND ITS GROWTH NOVEL NAIVE BAYES

B.Durgalakshmi, V.Vijayakumar

** B.Durgalakshmi, VIT UNIVERISITY-CHENNAI,INDIA*
** V. Vijayakumar, VIT UNIVERISITY-CHENNAI,INDIA*

Abstract

Cancer is a crucial disease which kills the people worldwide. To reduce cancer death rate is to detect it earlier. It depends upon the age, blood group, food habits, genetic combination, and heredity. Even though predicting cancer is generally clinical and biological in nature, in general we used some of the computation methods and artificial intelligence to predict breast cancer through images and rough set values.In this paper we are using revised edition of classifier (i.e) Mahout Naive Bayes and the performance and accuracy can be calculated

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

Keywords: Naive Bayes, Hadoop.Feature Selection

1. Introduction

The primary cause of death among the population is cancer.Breast cancer is painstaking persistent cancer in women, with more than one million cases and virtually 600,000 deaths occurring worldwide annually. It represents 20.5 % of the total general of cancer cases in worldwide and a percentage of 36.2. % of breast cancer is measured as a treatable one. Early finding helps to save thousands of victims.we all know cancer is a disease in which makes the human cells turn into abnormal and shape more cells in an uncontrolled way or dead cells.

In case of breast cancer, the cancer causing germs starts in the tissues that make up the breasts. So finally these cells may form a accumulation called a tumor. It may also close to tissues and invade other cells and extend to lymph nodes and rest of the body. so many diseases comes out of this cancer one. The scenario is principal factor in diagnosis of disease that will immediately follow treatment on particular stage of cancer. In past few years more number of research is going on breast cancer diagnosis. Many of the authors the described the breast cancer prognosis problem by employing many techniques includes artificial neural Networks, machine learning and data mining.

2. Section

In this section 2 describes about the Literature survey. Section 3 Points out feature selection methods and features taken for this work. Section 4 denotes the proposed methods, and finally section 5 describes the conclusion and future work.

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)
doi:10.1016/j.procs.2015.04.102

3. Literature Survey

For prognosis of breast cancer, many methods are deploying to predict the cancer. Decision tree, classifiers plays a major role in depicting materialize the proper methods for these types of problems. In decision tree because if some parameters taken as feature selection can shown not more significant in the decision process, then their rejection can be optional, which would shorten the whole system. Different algorithms such as ID3, ID4 and C5 (new updated version of C4.5), were used in breast cancer research, but many attributes were obtained as considerable predictive factors, which would extremely complicate the architecture of the final neural network system. An appropriate predictive feature selection method is necessary. A new method called control of induction by sample division method has been formed to perform adaptive pruning with predictive control.

A mixture of machine learning algorithms can be for classification. So most commonly applied classifiers are the Naive Bayes (NB), SVM classifier and the decision tree (J48) algorithm along with a combination of Meta learning algorithms. A meta algorithm is an algorithm that combines results from other so-called base classifiers. The Naive Bayes algorithm is based on Bayes Theorem and is applied on simple probabilistic classifier broadly used for more real data sets WEKA (Waikato Environment for Knowledge Analysis) software to permit access to different classifiers.

WEKA includes numerous standard machine learning techniques which enables the user to apply machine learning to derive useful knowledge and helps to make decision from databases that are too large and hard be analyzed in person. Machine learning algorithms are different in applying the data set features. By using feature reduction the classifier performance can be improved. The problems in trees are all features are not able to process in the decision tree. The simple Naive Bayes (NB) classifier is used in our work as a reference algorithm ant is said to be benchmark for other algorithms.

In the previous work for actual data sets the naive bayes performs well. Generally the bayes classifier assumes that all features which we are selecting are independent of one another. In the identification of gene identification network Bayesian networks can be used as we earlier said Naive Bayes has been used by several researchers for gene selection and classification for predicting cancer.

The advantage of this classifier which enhances classification accuracy..Neural network architectures reaches the accuracy of the neural classifiers reaches 93%. Then the predictions were further evaluated by means of analysis through the Kaplan-Meier approximation method. That paper also evaluates the performance of feature reduction and classification algorithms on the training and test dataset. Gin general many methods that apply for the prediction and classification of a particular cancer with high accuracy & precision still remains a challenge.

4. Feature Section

This proposed scheme has been evaluated using three real datasets of human sequences of cancer or non-cancer and particularly breast/non-breast cancer . The quality of data is important for data mining projects and many of quality indicators can be used. Accuracy and consistency are the two foremost indicators of data quality. The definition of each feature in the dataset should be analyzed and clearly defined in a way. If the feature or property which we used as a characteristic feature is not sufficiently well defined, the definition has to be corrected or improved.

In case of statistical methods, the known rule of thumb that the number of instances should be the exponential number of used features. These kinds of particular rule is not clear in data mining, but the input features should fit into a classifier which in turn helps the model as a prognostic one. When we are having numerous choice of features, an huge effort is usually put into allowing for how the features can be reduced. In the past selection method, the aim of feature selection methods is to improve model accuracy in which automatically reduce the cost and complexity of classifiers.

Generally based upon the estimated importance, significant features from a dataset is chosen. A rating scheme is designed to estimate the value and important to the necessary application. Extensive ranges of automatic feature selection methods are available. In other way, the size of the data can be minimized considerably and irrelevant features can be deleted from the dataset.

5. Proposed Method

In this paper we are applying the Mahout naive bayes classifier for prediction. The origin of naive bayes from Apache. It runs from hadoop environment. the problem in the existing naive bayes is skewed data problem and weight measurement error is there. In a overview it runs on the hadoop environment which uses map-reduce programming paradigm.

It process the unstructured data also which is the extra ordinary feature of Hadoop.In general three procedures for mining data is said to be (i) Data pre-processing, ,(ii) data training , (iii) Data testing. The data pre-processing step forms all the necessary operations to create the dataset in the appropriate format that is required for prognosis, In our work cancer format is a cluster of files containing text which is converts them into a sequence file format. Because map reduce paradigm can handle sequence file format. If the data are from different sources we can use SQOOP to extract data from different sources.

Data training is the critical part; from the original dataset, when we extract the information that should be relevant to our predicting tasks and we hav to model the frequency in it. Data evaluation is important, as the data should fit into both the training and test phases. In this mahout Naive bayes works well. Other classifiers like Decision forest and Logistic Regression also works but the performance measure may vary. In the workflow, the first process of converting the raw data into sequence file format, After converting it, these file formats arfe transformed to sparse vectors.

Then applying these vectors to naive bayes trainer and model has been calculated and test. The main advantage of mahout classifier which gives the best performance and accuracy measure of 94%.The result is depend upon the features in which we chosen. With addition of classifier, some Meta learning methods can also be used.

6. Conclusions

As we said earlier, so many methods are in process of predicting and classification of cancer especially in breast with high accuracy is still remains a challenge. In this paper we used the mahout which is emerging and gives maximum throughput of accuracy. Depends upon the dataset and feature selection the values may change, In future we will work on selecting the different feature to test and estimate performance and accuracy.

REFERENCES

[1] Lawrence,james,David,Jaime “Tackling the Poor Assumptions of Naive Bayes Text Classifiers” in 2003.

[2] Thair Nu Phyu “Survey of Classification Techniques in Data Mining” *Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009, Hong Kong.*

[3] M.C.P. de Souto, R.B.C. Prudencio, R.G.F. Soares, D.S.A. de Araujo, I.G. Costa, T.B. Ludermir, A. Schliep, Ranking and selecting clustering algorithms using a meta-learning approach, *IEEE International Joint Conference on Neural Networks, 2008. IJCNN 2008.*

[4] Okun,H.Priisalu,”Dataset complexity gene expression based classification using ensembles of k-nearest neighbors, ArtificialIntelligence in Medicine45(2–3)(2009)”.

[5] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, S. Levy, “A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis”.

[6] http.// apache.mahout.org