Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Comparative Analysis of Classification Algorithms for the Prediction of Leukemia Cancer

[1]Durairaj M, [2]Deepika R
[1]Assistant Professor, [2]Research Scholar
[1, 2] School of Computer Science Engineering & Applications,
Bharathidasan University, Trichy, Tamilnadu, India

Abstract— Classification algorithms of data mining often used in the prediction of medical data analysis. Many researchers have been working on improving the performance of existing algorithms in terms of minimizing the time taken to build the model and maximizing the prediction accuracy of the proposed model. But still certain classification algorithms suffer from manipulating large datasets with more number of attributes. The main objective of this paper is to compare the behavior of traditional classification algorithms with respect to leukemia cancer dataset which contains 7130 attributes with 72 records. The results are analyzed using two factors such as prediction accuracy and time.

Keywords— leukemia, classification, data mining, accuracy, dataset

I. INTRODUCTION

The process of uncovering the existing knowledge in large data sources is called data mining. Data mining tools predict weather, future trends and behaviors, students' behavior, allowing companies to make proactive, knowledge-driven decisions and medical diagnosis. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by reflective tools typical as decision support systems. Data mining tools can answer not only to business questions but also for medical data analysis that is critical process to resolve.

Clinical databases contain information about patients and their medical conditions. Relationships and patterns within this data could provide new medical knowledge. Data mining methods assist physicians in numerous ways right from the interpretation of complex diagnostic tests, merging information from multiple sources and providing support for differential diagnosis and patient-specific prognosis. Leukemia is a cancer that starts in blood stem cells. Stem cells are basic cells that develop into different types of cells that have different jobs. Blood stem cells develop into either lymphoid stem cells or myeloid stem cells. Leukemia develops when blood stem cells in the bone marrow change and no longer grow or behave normally. These abnormal cells are called leukemia cells. Over the time, leukemia cells crowd out normal blood cells and prevent from doing their jobs.

Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). Here, a generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemia's as a test case. A class discovery procedure automatically discovered the distinction between Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL) without previous knowledge of these classes. An automatically derived class predictor can determine the class of new leukemia cases. The results demonstrate that the feasibility of cancer classification is based solely on gene expression monitoring and suggest a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge.

In section 1, the leukemia cancer and classification algorithms of data mining are briefly introduced. In section 2, related works are briefly discussed. In section 3 briefs the three types of classifiers. Section 4 discusses the experimentation and results. Section 5 concludes with the future direction of this work.

II. RELATED WORKS

Kaishi Li, et,al., [1] Illustrated that the feature extraction of microarray genes has a greater impact on its classification and clustering as it is taken as input to any network. The use of gene expression data in discriminating two types of very similar cancers acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) presented in Classification results are reported in using methods other than neural networks. This paper explored the role of the feature vector in classification. In order to achieve best results in learning algorithm, feature subset selection method should be applied on to the dataset.

A. Zibakhsh, et.al., [2] presented a new memetic algorithm that has the ability of filtering interpretable and précised fuzzy if–then rules over cancer data. The concept of memetic algorithms with the Multi-View fitness

function were introduced as a first approach. Multi-View fitness function that was presented considered two kinds of evaluating procedures. Procedure one, was placed within the main progressive structure of the algorithmic rule, evaluates every single fuzzy if–then rule per the required rule quality (the evaluating procedure doesn't think about alternative rules). However, the second procedure determines the standard of every fuzzy rule per the full fuzzy rule set performance. Compared to classic memetic algorithms, these forms of memetic algorithms enhance the rule discovery method considerably.

Gopala Krishna Murthy Nookala, et.al., [3] presented a comprehensive comparative analysis of fourteen totally different classification algorithms and their performance has been evaluated by victimization three different cancer information sets. The results indicated that none of the classifiers outperformed all others in terms of the accuracy on all the three information sets. Most of the algorithms performed higher because the size of the info set is augmented. They counseled the users to not follow a selected classification methodology and will judge totally different classification formulas and choose the classic algorithm.

Shweta Kharya, et.al., [4] discussed various data processing approaches used for carcinoma identification and prognosis. Carcinoma identification is distinctive of benign from malignant breast lumps and carcinoma Prognosis predicts once carcinoma is to recur in patients that have had their cancers excised. This study paper summarized varied review and technical articles on carcinoma identification and prognosis conjointly they targeted on current analysis being dole out victimization the info mining techniques to reinforce the carcinoma identification and prognosis.

Cheng-Mei Chen, et.al., [5] established a survival prediction model for cancer of the liver victimization data processing technology. The information were collected from the cancer registration data source of a medical center in Northern Taiwan between 2004 and 2008 comprised of 227 patients were recently diagnosed with cancer of the liver throughout this point. Nine variables relating cancer of the liver survival were analyzed victimization t-test and chi-square take a look at. Six variables showed vital. Artificial Neural Network (ANN) and Classification and Regression Tree (CART) were adopted as prediction models. The models were tested in 3 conditions such as one variable from clinical stage, six vital variables, and nine variables (significant and non-significant). The result were guaranteed five year endurance with the output prediction.

Xiangchun Xiong et.al., [6] discussed on three methods to diagnose breast cancer, namely Mammography, FNA (Fine Needle Aspirate) and Surgical biopsy. They used FNA with a Data mining and Statistics method to achieve a best result. They combined some statistical methods with data mining methods to find the unsuspected relationships. They explored that statistics and data mining techniques can offer great promise in helping us to uncover patterns in the data.

Chun-Hui Wu,et.al., [7] employed few data mining techniques to explore hidden knowledge among meridian energy of prostate cancer from 213 patients' health examination data including patient demographics and evaluations for the Prostate-Specific Antigen (PSA) blood test as well as the meridian energy. The findings were considered as helpful reference in diagnosis and treatment of prostate cancer for TCM(Traditional Clinic Medicine) physicians. This study provided new scientific and quantitative information for TCM Traditional Chinese Medicine physician in clinical practice of prostate cancer. TCM physicians would benefit from a better understanding of the relationship between meridian energy system and prostate cancer.

Soltani Sarvestani, et.al., [8] collected datasets for breast cancer knowledge discovery and invoked various data mining techniques to find out the percentage of disease development. Thus, the result helped in selecting a reasonable treatment of the patient. This work also indicated that statistical neural networks can be effectively used for breast cancer diagnosis to help oncologists.

Yao Liu, et.al., [9] implemented a classifier for the prediction of lung and breast cancer, which are the most common cancers for both men and women. This work examined the usefulness of the new rule pruning procedure, and showed the proposed procedure resulted a positive influence on the accuracy for both DPSO and PSO. The proposed approach was compared with popular data mining algorithms on classifying breast cancer and lung cancer, and demonstrated that DPSO combined with the rule pruning is effective in predicting common types of cancer. Experiment showed that the new pruning method increased the classification accuracy, and the new approach is found to be effective in cancer prediction.

Nashat, et.al.,[10] presented a method to find a clustering pattern of the genes involved in breast cancer. They designed a Growing Hierarchical Self-Organizing Map (GHSOM) to mine gene microarray data. They applied their technique on 24,481 genes of DNA microarray of breast tumor samples. Result revealed 17 genes that are likely to be correlated with four breast cancer marker genes.

Sivagowry S, et. al., [11] presented a method for Medical Data Mining, which is a domain of challenge and involves a lot of imprecision and uncertainty. Provision of quality services at affordable cost is the major challenge faced in the health care organization. This work proposed a methodology, which applied Data mining technique for the prediction of heart disease. This proposed methodology successfully diagnosed heart disease in early response time.

Durairaj M, et.al., [12] discussed three important things that are considered to be unique from other works, firstly it presents the importance of data mining approaches on medical data mining. Furthermore it, thoroughly surveys the related work that has been carried out so far on cancer prediction. Finally, this paper compares the correct accuracy level of the reviewed papers and discusses the key facts that are obtained from the results. Through the results, it was noted, that the correct accuracy of the classification algorithms are not stable and they differ from one another.

III. METHODOLOGY

A. Naïve Bayes Classifier

A Naie Bayes classifier is a simple probabilistic classifier based on Bayes' theorem and is particularly suited when the dimensionality of the inputs are high. NaïveBayes classifier assumes that the classes for classification are independent. Though this is rarely true Bayesian classification, there are some theoretical reasons for the apparent unreasonable efficiency achieved.

$$P(c|d) = \frac{P\left(\frac{c}{d}\right)P(c)}{P(d)} \quad \text{——————————} \quad (1)$$

and the classifier is $c^* = \arg maxc\, P(c|d)$

Naïve Bayes is used as a classifier in various real world problems like email grouping, document classification, spam detection, mail priority sorting and content detection [11].

B. Decision Tree Classifier

Decision trees are popular methods for inductive inference. They are robust to noisy data and learn disjunctive expressions. A decision tree is a k-array tree in which each internal node specifies a test on some attributes from input feature set representing data. Each branch from a node corresponds to possible feature values specified at that node and every test results in branches, representing varied test outcomes. The decision tree induction basic algorithm is a greedy algorithm constructing decision trees in a top-down recursive divide-and-conquer manner

$$info\ o(D) = -\sum_{i=1}^{n} p_i\ log_2(p) \quad \text{——————————} \quad (2)$$

Where p_i is the probability that arbitrary vector in D belongs to class c_i [12].

C. Lazy Classifier (LC)

Lazy learners store the training instances and do no real work until classification time. Lazy learning is a learning method in which generalization beyond the training data is delayed until a query is made to the system where the system tries to generalize the training data before receiving queries. The main advantage gained in employing a lazy learning method is that the target function will be approximated locally such as in the k-nearest neighbour algorithm. Because the objective function is approximated locally for each query to the system, lazy learning systems can concurrently solve multiple problems and deal successfully with changes in the problem arena.

IV. EXPERIMENTATION AND RESULTS

For our experiment, we have taken the leukemia cancer dataset from UCI (University of California Irvine) machine learning data repository for the purpose of assessing the performance of typical classification algorithms on a dataset contains larger number of attributes. The dataset contains 7130 attributes including the class attributes which explores the gene expression information of leukemia cancer. Out of 72 instances, 38 instances have chosen as training samples and the remaining 34 have chosen as test samples. The evaluation is conducted using WEKA, the most popular and widely used data mining tool for the analysis of data. A classification knowledge flow is built for three sets of above discussed classifiers to get the time and accuracy results of the predictive models. Fig. 1 depicts the sample knowledge flow model of the naïve Bayes classifier.

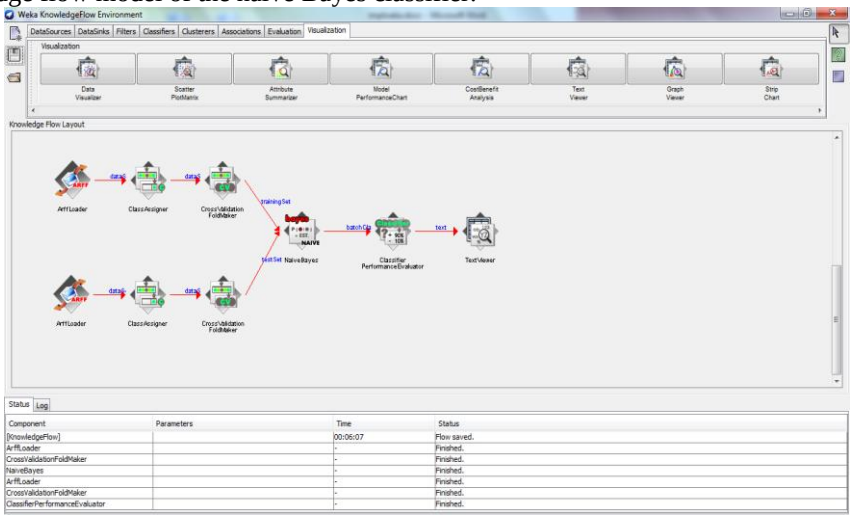

Fig 1. Knowledge Flow of Naïve Bayes Classifier for Leukemia

The knowledge flow is constructed using the leukemia training and test datasets. The step by step process of the knowledge flow is explained in the following algorithm.

- Select Knowledge Flow Button on the Weka GUI Chooser
- Drag two arff loaders from data Sources for configuring both training and test data
- Select Two Class Assigners from Evaluation for assigning the class attributes of the training and test data
- Set the cross Validations Folder Makers as 10 cross fold as default

- Drag the NaïveBayes Classifier from Evaluation and input both training and test data
- Input batch classifier of the Naïve Bayes Classifier to the Classifier Performance Evaluator of Evaluations
- Input text of Classifier Performance Evaluator to Text Viewer to view results

The knowledge flow of Naïve Bayes classifier has got 91.17 % of prediction accuracy with both the training and test dataset of leukemia cancer. The accuracy of the classifier is represented in Fig. 2

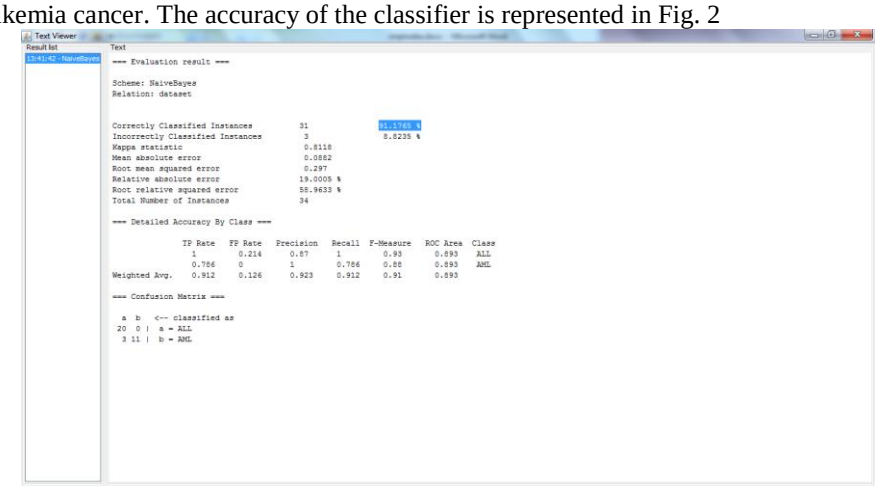

Fig 2. Prediction Accuracy of Naïve Bayes Classifier for Leukemia

The same dataset is evaluated with other classification algorithms that are available in weka 3.6 and the accuracy of those algorithms is computed and presented in Table 1. in terms correctly and incorrectly classified instances

Table 1: Prediction Accuracy of the Traditional Algorithms

S. No	Classifier Type	Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	Prediction Accuracy
1	Baysian Classifier	Naïve Bayes	31	3	91.1765 %
2	Lazy Classifier	IBk	24	10	82.3582 %
3	Decision Tree Classifier	J48	31	3	91.1765 %

Table 1 explicates that both baysian classifier and decision tree classifier have demonstrated the highest and same prediction accuracy of 91.17 %, though the number of attributes is high. Hence, those algorithms can be compared with another factor called time to consider the best among them. On the other hand, IBk algorithm of Lazy Classifier has only demonstrated 70.58 % which seems that those types of algorithm are not well-suited for handling large attributed data sets. The Naïve Bayes algorithm gives the accuracy result for predict by classification of such algorithms. It is shown as a graphical representation in fig. 1. In this graph it comparing and give the 91.17% of the accuracy result. So it is the best accuracy value of the prediction.

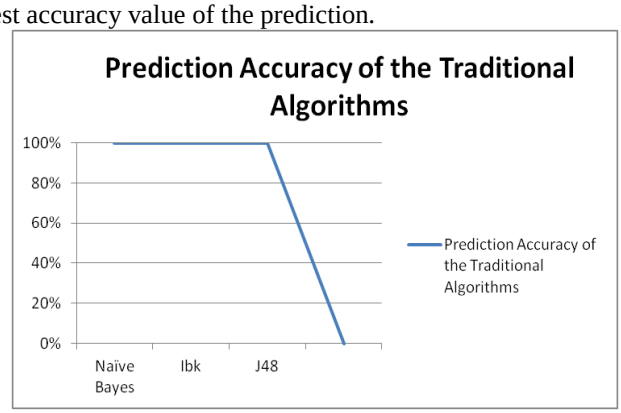

Figure1: Represents Prediction of Accuracy

In addition to the comparison of accuracy, the time taken to build the model is also compared. Table 2. The t able describes that the naïve bayes algorithm has taken less time of 0.16 seconds to produce a good prediction model than the other two algorithms. J48 algorithms have only varied with the minor difference in time.

Table 2: Time taken to build model

S. No	Algorithm	Accuracy
1	Naïve Bayes	0.16 seconds
2	IBk	0.02 seconds
3	J48	0.41 seconds

The naïve Bayes algorithm is calculate the accuracy value in the short time of comparing those algorisms. It is shown as a graph in Fig.2. In this graph the naïve bayes algorithm is take the time 0.16 seconds. So it is the fastest elapsed time from the other algorithms.

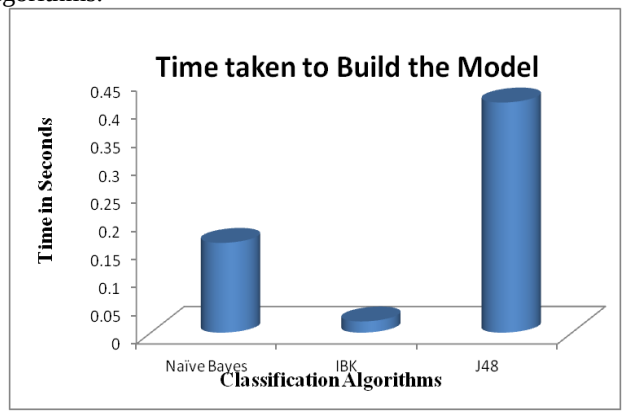

Fig 2: Represents Time taken to build Model

V. CONCLUSION

This paper presents a comparative study on the performance of various classifiers of data mining over high dimensional data. For the comparative analysis and experiments, a dataset with 7130 attributes is taken. Three types of classifiers have chosen for the study, they are baysian classifier, decision tree classifier and lazy classifier. From the results it is identified that naïvebayes classifier is able to build good prediction model with 91.17% with less time of 0.16 seconds. As a future dimension of this work, the accuracy of the data mining classification algorithms will be compared with the statistical model in order to propose suitable model for effective cancer prediction.

REFERENCES

[1] Kaishi Li, Meixue Yang, Gaurav Sablok, Jianping Fan, Fengfeng Zhou, *"Screening features to improve the class prediction of acute myeloid leukemia and myelodysplastic syndrome",* ELSEVIER, 348–354, 2013

[2] A. Zibakhsh, M.SanieeAbadeh, *"Gene selection for cancer tumor detection using a novel memetic algorithm witha multi-view fitness function"*, ELSEVIER, Engineering Applications of Artificial Intelligence, 1274–1281,2013

[3] Gopala Krishna Murthy Nookala, Bharath Kumar Pottumuthu, Nagaraju Orsu, Suresh B. Mudunuri, *"Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification",* (IJARAI) International Journal of Advanced Research in Artificial Intelligence,Vol. 2, No.5, 2013

[4] Shweta Kharya, "Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), April 2012

[5] Cheng-Mei Chen, Chien-Yeh Hsu, Cheng-Mei Chen and Chien-Yeh Hsu," Prediction of Survival in Patients with Liver Cancer using Artificial Neural Networks and Classification and Regression Trees" Seventh International Conference on Natural Computation 2011.

[6] Xiangchun Xiong, Yangon Kim, Yuncheol Baek, Dae Wong Rhee, Soo-Hong Kim, "*Analysis of Breast Cancer Using Data Mining & Statistical Techniques*", IEEE, 0-7695-2294-7/05, 2005

[7] Chun-Hui Wu, Kwoting Fang, Ta-Cheng Chen," Applying data mining for prostate cancer", IEEE, 978-0-7695-3687-3/09,2009.

[8] A. Soltani Sarvestani, A. A. Safavi, N.M. Parandeh, M.Salehi, "*Predicting Breast Cancer Survivability Using Data Mining Techniques*", IEEE, 978-1-4244-8666-3, 2010.

[9] Yao Liu and Yuk Ying Chung, "*Mining Cancer data with Discrete Particle Swarm Optimization and Rule Pruning*", IEEE, -1- 61284-704-7,2011.

[10] Nashat Mansour, Rouba Zantout, Mirvat El-Sibai," *Mining Breast Cancer Genetic Data",* IEEE, 978-1-4673-4714-3, 2013

[11] Sivagowry .S, Durairaj. M and Persia.A, 2013" An Empirical Study on applying Data Mining Techniques for the Analysis and Prediction of Heart Disease'' international conference on information communication and embedded systems icices.

[12] Durairaj.M, Deepika.R," Prediction Of Acute Myeloid Leukemia Cancer Using "Datamining-A Survey",2015

