Journal of Computer and Communications, 2018, 6, 76-83
http://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Classification of Hematological Data Using Data Mining Technique to Predict Diseases

Fahmida Akter[1], Md Altab Hossin[2], Golam Moktader Daiyan[3], Md. Motaher Hossain[4]

[1]Department of Computer Science and Engineering, Coxbazar International University, Chittagong, Bangladesh
[2]Department of Information Management and Ecommerce, University of Electronic Science and Technology of China, Chengdu, China
[3]Department of Computer Science and Engineering, East Delta University, Chittagong, Bangladesh
[4]Department of Computer Science and Engineering, University of Asia Pacific, Dhaka, Bangladesh
Email: fahmidaakter26@gmail.com, altabbd@163.com, daiyangolam@yahoo.com, motaher2426@yahoo.com

How to cite this paper: Akter, F., Hossin, M.A., Daiyan, G.M. and Hossain, M.M. (2018) Classification of Hematological Data Using Data Mining Technique to Predict Diseases. *Journal of Computer and Communications*, **6**, 76-83.
https://doi.org/10.4236/jcc.2018.64007

Received: March 30, 2018
Accepted: April 25, 2018
Published: April 28, 2018

Abstract

Over the years, the amount of information about patients and medical information has grown substantially. Moreover, due to an increase of blood diseases patients, conventional diagnostic tests have been using by the medical pathologists which are low in cost and result in an inaccurate diagnosis. To recognize optimal disease pattern from hematological data, a reliable prediction methodology is needed for medical professionals. Data mining approaches permit users to examine data from various dimensions, group it and sum up the relationships identified. Classification is a vital data mining technique with extensive applications. Classification algorithms are applied to categorize every item in a set of data into one of a known set of classes. The objective of this paper is to compare different classification algorithms using Waikato Environment for Knowledge Analysis and to find out a most effective algorithm for end-user functioning on hematological data. The most efficient algorithm found is Random Forest having accurateness at 96.47% and the overall time is taken to construct the model is 0.16 seconds which is more efficient than different existing works. On the contrary, Multilayer Perceptron classifier has the lowest accuracy of 75.29% with 1.92 seconds to construct the model.

Keywords

Data Mining, Random Forest, Multilayer Perception, Bayesian Network

1. Introduction

Data mining is the process of finding useful and relevant information from the various types of databases. Different approaches to data mining were suggested

to face the challenges of storing and processing all types of data [1]. Nowadays data mining has increasing applications in Medical Science, Railway and so on [2]. Data mining provides doctors to provide necessary treatments, and thus patients are treated better along with more cheap health services, becoming popular day by day [3] [4] [5] [6]. In pathology, it has become familiar with a strong technique in dealing with enormous pathological information to search knowledge that is given. Additionally, comparison of different classification techniques using WEKA (Waikato Environment for knowledge analysis) for blood-related data is a demanding task in medical science research. To find out better classification algorithms, it is hard to compare different classification algorithms in different collections of data [7]. The main concern is the classification of hematological data to predict diseases. With this purpose to perform better, hematological data analysis is divided into three phases: Hematological data collection, classification algorithms and evaluation of results and performance. Major data mining techniques are three which are known as regression, classification and clustering. The application of data mining now goes towards clinical research such as AML (Acute Myeloid Leukemia) where predictive model plays an important role [8] [9] [10] [11].

The remainder of this paper is organized as follows. Section 2 reviews the related works. Section 3 describes material and methods. Dataset and preprocessing are explained in Section 4. In Section 5, experimental results and discussion are illustrated. At Section 6, the conclusion is given.

2. Related Works

Several types of research have been made to evaluate the performance of data mining classification algorithms using WEKA. In the study [3] [12], the researchers evaluated the performance of data mining classification algorithm in WEKA. Another research in [1] compared different classification techniques using different datasets. The research in [2] compared the various clustering algorithms of WEKA tools. Moreover, performance analysis and evaluation of various data mining algorithms used for cancer cell classification had done [13]. This is also used in artificial intelligence and predicting abnormality in peripheral blood smear [14] [15]. Data mining classifiers were used in the study [16] to develop an automated diagnosis of thalassemia [17]. Also, analysis of various clustering algorithms of data mining on health informatics was performed [18]. The area of bioinformatics has also used data mining tools and various classification techniques which were compared [19]-[24]. Data mining techniques were also used to differentiate between the patients with a normal blood disease and patients with blood tumor [25]. Another study highlighted on contrasting of two classification techniques J48 and Random tree by means of WEKA to classify Sickle Cell Diseases (SCD). More recently, anemia has foreseen using different data mining classification algorithms [12] [26] [27] where J48 algorithm confirmed its best performance in classifying types of anemia [28]. Besides, WEKA

has been used in this experiment as hidden predictive information can be extracted using this algorithm from large database [29]. In addition, the experiment has been conducted for CBC (Compete Blood Count), which is quite rational to extract data using the intended algorithm as the WEKA is being employed for data mining widely.

3. Material and Methods

In this study, an open to all data mining tool WEKA (version 3.8.0) has been used. Two dissimilar data sets have been utilized and the performance of classification algorithms (classifiers) has been examined. The analysis has been carried out by SONY VIAO Windows version 8 with Intel® Core™ i3 Central Processing Unit, 1.70 Gigahertz Processor and 4 Gigabyte RAM. The data sets have been selected so that they vary in size, predominantly with the number of attributes. The hematological parameters consist of White blood cell o (WBC), Red blood cell count (RBC), Hemoglobin (Hb), Hematocrit (Hct), Mean corpuscular volume (MCV), Mean corpuscular hemoglobin (MCH), Mean corpuscular hemoglobin concentration (MCHC), Platelet count (PLT), Neutrophil count (NEU), Lymphocyte (LYMP), Monocyte (MONO), Eosinophil (EO), and Basophil (BASO) (SysMex 1000i Sysmexcorporation, Kobe, Japan). Hematological data were evaluated by the hand of a medical technologist. Data which are collected are allocated to multiple tags: indicative of anaemia of unceasing disorder, Eosinophilia, Microcytic hypochromic anaemia, Normocytic anaemia, Neutrophil leucocytosis, Neutrophilia, Non-specific findings, High ESR.

4. Dataset and Preprocessing

The dataset of experiment1 comprises of 425 samples and dataset of experiment 2 consists of 298 samples. The attributes characterize the Complete Blood Count (CBC) features as in **Table 1**.

In the preprocessing of the dataset, irrelevant attributes were eliminated, refilled the missing values and removed/refilled the outlier values on the outlier samples. **Table 2** represents the dataset attributes which are used in this investigation.

5. Result and Discussion

In this study, the experiment that employs the data mining classifiers will be separated into two branches: the experimentation with full and reduced features. The outcomes from these two branches and in-depth classification accuracy analysis highlighting on the classification errors will be displayed in following sections. Three experiments were conducted in each type: the first one is to measure the performance of the Random Forest Tree classifier; the second one is to measure the performance of the Bayesian Network classifier, the third one to measure the performance of the Neural network (Multilayer Perceptron). The

Table 1. CBC test features.

Shortcut	Term	Male normal value	Female normal value
WBC (cmm)	White Blood Cell	4000 – 11,000	
RBC (million/cmm)	Red Blood Cell	5.0 ± 0.5	4.3 ± 0.5
HB (g/dl)	Hemoglobin	15.0 ± 2.0	13.5 ± 1.5
HCT (I/I)	Hemoglobin	0.45 ± 0.05	0.41 ± 0.005
MCV (ft)	Mean Cellular Volume	92 ± 9	
MCH (pg)	Mean Cellular Hemoglobin	29.5 ± 2.5	
MCHC (g/dl)	Mean Cellular Hemoglobin Concentration	33.0 ± 1.5	
PLT (/Cmm)	Platelet Count	150,000 – 400,000	
NEU	Neutrophils (%)	40 - 75	
LYMP	Lymphocytes (%)	20 - 40	
MONO	Monocytes (%)	2 - 10	
EO	Eosinophils (%)	2 - 6	
BO	Basophils (%)	<1.0	

Table 2. Dataset attributes.

Attribute	Data type	Attribute role
SEX	Binomial	Regular
WBC	Integer	Regular
RBC	Integer	Regular
HB	Integer	Regular
HCT	Integer	Regular
MCB	Integer	Regular
MCH	Integer	Regular
MCHC	Integer	Regular
PLT	Integer	Regular
NEU	Integer	Regular
LYMP	Integer	Regular
MONO	Integer	Regular
EO	Integer	Regular
BO	Integer	Regular
Hematological Comments	Nominal	Label

feed-forward back-propagation neural network classifier was regulated with 500 training cycles, learning rate 0.3, and momentum 0.2.

5.1. Experiment with Full Features

In these experiments, whole traces aspects of each sample were used. The Random Forest tree classifier gives an accuracy of 96.47%, the Neural Network (Multilayer Perceptron) presents accuracy of 75.29%, and finally, the Bayesian network classifier provides accuracy of 84.70% as shown in Figure 1 and in Table 3.

5.2. Experiment with Reduced Feature

The results from these experiments are given in Table 4. The Random Forest Tree classifier puts the accuracy of 86.44%, while the Neural Network classifier provides accuracy of 52.54% and the Bayesian Network classifier gives an accuracy of 74.57% as shown in Figure 2 and in Table 4.

After considering Figure 1, Figure 2 and Table 5, it is seen that the maximum accuracy is 96.47% and the minimum accuracy is 52.54%. It can be concluded that Random Forest tree classifier is better than other classifiers considered.

6. Conclusion

This paper evaluated and investigated three preferred classification algorithms based on WEKA. By utilizing the hematological data, the superlative algorithm found is Random Forest Classifier with an accuracy of 96.47% and the total time taken to build the model is at 0.16 s. Neural Network has the accuracy of 52.54% which is the lowest accuracy in comparison with others, which is an affirmative side of this study. These results will aid the researchers to get competent results for a particular dataset. The finding will help users to analyze disease in minimal time which is a good contribution of this study.

Table 3. Dataset attributes.

Algorithm (Total Instances 425)	Correctly Classified Instances % (Value)	Incorrectly Classified Instances % (Value)	Time Taken (Second)	Kappa Statistic
Random Forest Tree	96.47% (82)	3.53% (3)	0.16	0.9535
Neural Network	75.29% (64)	24.71% (21)	2.06	0.6772
Bayesian Network	84.70% (72)	15.30% (13)	0.1	0.7985

Table 4. Simulation result of experiment 2.

Algorithm (Total Instances 425)	Correctly Classified Instances % (Value)	Incorrectly Classified Instances % (Value)	Time Taken (Second)	Kappa Statistic
Random Forest Tree	86.447% (51)	13.56% (8)	0.16	0.9535
Neural Network	52.54% (31)	47.46% (28)	1.53	0.3363
Bayesian Network	74.57% (44)	25.43% (15)	0.2	0.6456

Table 5. Comparison of various classifiers.

Name of the Classifier	Experiment 1	Experiment 2
Random Forest Tree	96.47	86.44
Neural Network	75.29	52.54
Bayesian Network	84.70	74.57

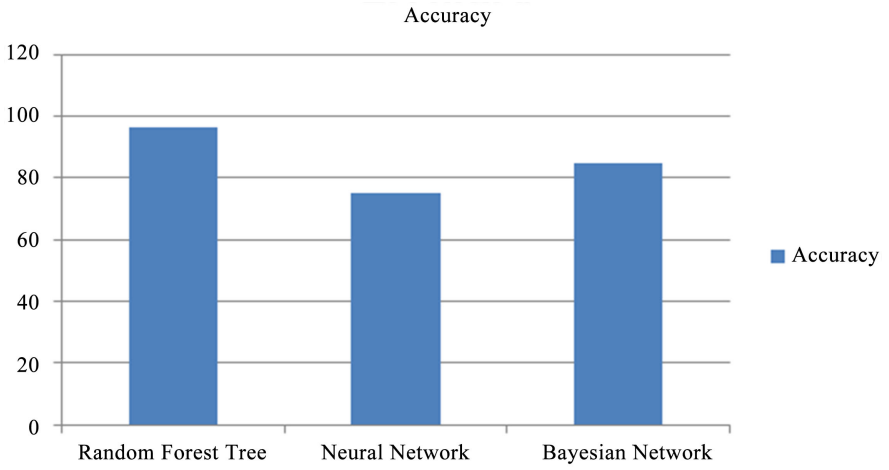

Figure 1. Classifiers accuracy value for Experiment 1.

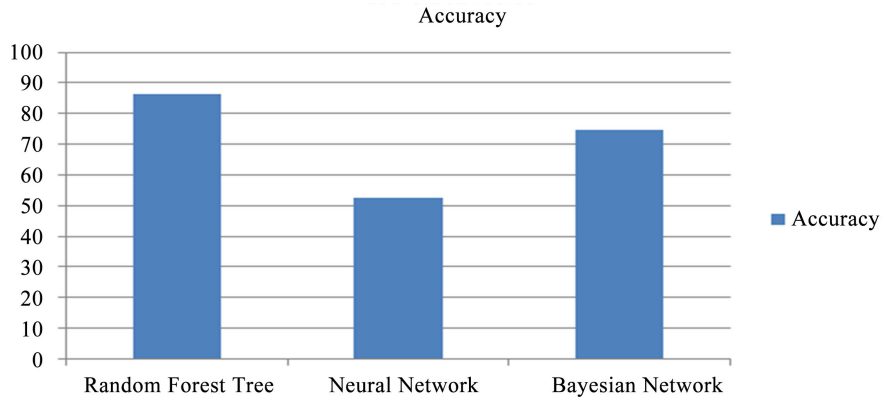

Figure 2. Classifiers accuracy value for Experiment 2.

References

[1] Kaur, P., Singh, M. and Josan, G.S. (2015) Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector. *Procedia Computer Science*, **57**, 500-508. https://doi.org/10.1016/j.procs.2015.07.372

[2] Zierk, J., Hirschmann, J., Toddenroth, D., Prokosch, H.U., Rauh, M. and Metzler, M. (2016) A Bioinformatics Approach to Pediatric Hematology Reference Intervals. *Klinische Pädiatrie*, **228**, A45. https://doi.org/10.1055/s-0036-1582522

[3] Salvithal, N.N. and Kulkarni, R.B. (2013) Evaluating Performance of Data Mining Classification Algorithm in Weka.

[4] Vaithiyanathan, V., *et al.* (2013) Comparison of Different Classification Techniques Using Different Datasets. *International Journal of Advances in Engineering & Technology*, **6**, 2.

[5] Narendra, S., Bajpai, A. and Litoriya, R. (2012) Comparison the Various Clustering Algorithm of Weka Tools. *International Journal of Emerging Technology and Advanced Engineering*, **2**, 73-80.

[6] Singhal, S. and Jena, M. (2013) A Study on WEKA Tool for Data Preprocessing. *International Journal of Innovative Technology and Exploring Engineering* (*IJITEE*), **2**.

[7] Khan, S.A., Epstein, J.H., Olival, K.J., Hassan, M.M., Hossain, M.B., Rahman, K.B.M.A., Elahi, M.F., *et al.* (2011) Hematology and Serum Chemistry Reference Values of Stray Dogs in Bangladesh.

[8] Shouval, R., Bondi, O., Mishan, H., Shimoni, A., Unger, R. and Nagler, A. (2014) Application of Machine Learning Algorithms for Clinical Predictive Modeling: A Data-Mining Approach in SCT. *Bone Marrow Transplantation*, **49**, 332-337; https://doi.org/10.1038/bmt.2013.146

[9] Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V., Paschka, P., Roberts, N., Potter, N.E., Heuser, M., Thol, F., Bolli, N., Gundem, G., Van Loo, P., Martincorena, I., Ganly, P., Mudie, L., McLaren, S., O'Meara, S., Raine, K., Jones, D., Teague, J., Butler, A.P., Greaves, M.E., Ganser, A., Döhner, K., Schlenk, R., Döhner, H. and Campbell, P.J. (2016) Genomic Classification and Prognosis in Acute Myeloid Leukemia. *The New England Journal of Medicine*, **374**, 2209-2221. https://doi.org/10.1056/NEJMoa1516192

[10] Dankowski, T. and Ziegler, A. (2016) Calibrating Random Forests for Probability Estimation. John Wiley & Sons Ltd., Hoboken.

[11] Zhang, W., Ma, D. and Wei, Y. (2014) Medical Diagnosis Data Mining Based on Improved Apriori Algorithm. *Journal of Networks*, **9**, 1339-1345. https://doi.org/10.4304/jnw.9.5.1339-1345

[12] Chung, H.J., Park, C.H., Han, M.R., Lee, S., Ohn, J.H., Kim, J. and Kim, J.H. (2005) ArrayXPath II: Mapping and Visualizing Micro-Array Gene-Expression Data with High Dimension. *Nucleic Acids Research*, **33**, W621-W626. https://doi.org/10.1093/nar/gki450

[13] Nookala, G.K.M., Orsu, N., Pottumuthu, B.K. and Mudunuri, S.B. (2013) Performance Analysis and Evaluation of Different Data Mining Algorithms Used for Cancer Classification. *International Journal of Advanced Research in Artificial Intelligence*, **2**, 49-55.

[14] Raviya, K.H. and Gajjar, B. (2013) Performance Evaluation of Different Data Mining Classification Algorithm using WEKA. *Indian Journal of Research*, **2**, 19-21.

[15] Saichanma, S., Chulsomlee, S., Thangrua, N., Pongsuchart, P. and Sanmun, D. (2014) The Observation Report of Red Blood Cell Morphology in Thailand Teenager by using Data Mining Technique. *Advances in Hematology*, **2014**, Article ID: 493706.

[16] Othman, B., Fauzi, M. and Shan Yau, T.M. (2007) Comparison of Different Classification Techniques using WEKA for Breast Cancer. 3*rd Kuala Lumpur International Conference on Biomedical Engineering*, Kuala Lumpur, 11-14 December 2006, 520-523.

[17] Elshami, E.H. and Alhalees, A.M. (2012) Automated Diagnosis of Thalassemia Based on Data Mining Classifiers. In: *The International Conference on Informatics and Applications*, The Society of Digital Information and Wireless Communication, 440-445.

[18] Saxena, P. and Lehri, S. (2013) Analysis of Various Clustering Algorithms of Data Mining on Health Informatics. *International Journal of Computer & Communica-*

tion Technology, **6**, 108-112.

[19] Vijayarani, S. and Muthulakshmi, M. (2013) Comparative Analysis of Bayes and Lazy Classification Algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, **2**, 3118-3124.

[20] Satish Kumar, D., Saeb, A.T.M. and Al Rubeaan, K. (2013) Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics. *Computer Engineering and Intelligent Systems*, **4**, 28-38.

[21] Pandey, R., Guru, R.K. and Mount, D.W. (2004) Pathway Miner: Extracting Gene Association Networks from Molecular Pathways for Classifying and Predicting the Biological Significance of Gene Expression Microarray Data. *Bioinformatics*, **20**, 2156-2158. https://doi.org/10.1093/bioinformatics/bth215

[22] Wahbeh, A.H., *et al.* (2011) A Comparison Study between Data Mining Tools over Some Classification Methods. *International Journal of Advanced Computer Science and Applications*, 18-26.

[23] Solanki, A.V. (2014) Data Mining Techniques using WEKA Classification for Sickle Cell Disease. *International Journal of Computer Science and Information Technologies*, **5**, 5857-5860.

[24] Sharma, T., Sharma, A. and Mansotra, V. (2016) Performance Analysis of Data Mining Classification Techniques on Public Health Care Data. *International Journal of Innovative Research in Computer and Communication Engineering*, **4**, 11381-11386.

[25] Alaa, M. and Shurrab, A.H. (2017) Blood Tumor Prediction using Data Mining Techniques. *Health Informatics—An International Journal*, **6**, 23-30.

[26] Alkrimi, J.A., Jalab, H.A., George, L.E., Ahmad, A.R., Suliman, A. and Al-Jashamy, K. (2015) Comparative Study using Weka for Red Blood Cells Classification. *International Journal of Medical*, *Health*, *Pharmaceutical and Biomedical Engineering*, **9**, 19-22.

[27] Rajesh, K. and Sangeetha, V. (2012) Application of Data Mining Methods and Techniques for Diabetes Diagnosis. *International Journal of Engineering and Innovative Technology*, **2**, 224-229.

[28] Hasani, M. and Hanani, A. (2017) Automated Diagnosis of Iron Deficiency Anemia and Thalassemia by Data Mining Techniques. *International Journal of Computer Science and Network Security*, **17**, 326.

[29] Jagtap, S.B. and Kodge, B.G. (2013) Census Data Mining and Data Analysis using WEKA. *International Conference in Emerging Trends in Science*, *Technology and Management*, Singapore.