Predictive Accuracy For Two Diabetes Datasets Using Ant-Miner Algorithm

Nur Hadirah Khairul Anwar, Rizauddin Saian

Abstract: Data mining is helpful in turning a large amount of data into useful knowledge, therefore, it is beneficial in the medical field. The knowledge that being extracted can assist medical experts in enhancing the diagnosis and treatment of disease. Diabetes mellitus (DM) is a serious health challenge in most developed countries. It is a condition in which the body produces an insufficient amount of insulin to regulate the amount of sugar in the blood. Two diabetes datasets used in this study is Pima Indian diabetes dataset and Frankfurt Germany diabetes dataset. The aim of this paper is to improve predictive accuracy for diabetes by implementing Ant-Miner algorithm to the diabetes datasets and the results obtained will be compared with the result derived from using the different machine learning model such as Naïve Bayes, AdaBoostM1, K-nearest neighbors and RIPPER.

Index Terms: AdaBoostM1, Ant-Miner, Data mining, Diabetes, K-nearest neighbors, Naïve Bayes, RIPPER

1 INTRODUCTION

IN a real world application, there exist a huge amount of data that cannot be managed by normal human being [1]. A human being is incapable of handling many data because we have limited form of memory where we cannot memorize every detail of the data. Thus, this is where data mining plays its role. Data mining is also viewed as knowledge from data or KDD [2]. There are several steps involve in KDD which are 1) Data cleaning for the purpose of eliminating noise and unreliable data, 2) Data integration in which multiple sources of data might be combined, 3) Data selection where we select which data we want to analyze, 4) Data transformation so that the data now is in the form that is suitable for mining, 5) Data mining in which we choose which algorithm will be suitable to fit in the model to extract the data pattern, 6) Pattern evaluation where we analyze the data pattern that being extract and the last one is knowledge presentation where we derived knowledge from the pattern we get [2]. The healthcare sector is one of the areas that apply data mining in its daily activity. Related areas that usually used data mining are the medical device industry, pharmaceutical industry and hospital management [3]. Data mining can help healthcare filed in doing a prediction and diagnosis of the treatment. It used data mining because of large amounts of data involved and the pattern that being extracted from using data mining technique can reveal the useful knowledge. The knowledge derived from using data mining can help in better decision making and increase efficiency in working [4].

2 LITERATURE REVIEW

2.1 Diabetes Mellitus

Diabetes mellitus has become a serious health challenge in the 21st century. It is the seventh leading cause of fatality in the United States. Diabetes happens when the body loses the capability to produce sufficient amount of insulin to normalize the sugar intake in our blood. Diabetes is classified into four clinical cases which are type 1, type 2, other specific types of diabetes because of the other cases and the last one is gestational diabetes which happens to pregnant women [5]. Modern lifestyle and practicing unhealthy eating habits can lead to obesity and then to diabetes. Diabetes does not have a cure and one way for the patients to survive this disease is by making sure that the level of blood sugar is normal without serious high or low blood sugars. This can be achieved by doing an exercise, practicing healthy eating habits, taking oral diabetes medication or using insulin to normalize sugar amount in blood [6].

2.2 Ant Colony Optimization for Classification

Ant Colony Optimization algorithm was firstly introduced by Marco Dorigo in year 1992 [7]. This algorithm takes the inspiration from the behavior of real ants during their scavenging for food. During food hunting, ants will deposit a chemical substance namely pheromone along the trail. The purpose of pheromone is to keep them on the right track and avoid them from losing their trail. Ants will converge to the trail that has higher amount of pheromone evaporation. The first Ant Colony Optimization algorithm for classification is Ant-Miner algorithm. Ant-Mine algorithm was introduced by Parpinelli, Lopes and Freitas [8] to generate classification rules. The classification rules are in the form of IF-THEN rules. Classification is one of the tasks in data mining and one of the examples that doing classification is medical diagnosis. Example of research that classify medical data using Ant-Miner algorithm is research by Wahid and Al-Mazini [9] where they used it to classify cervical cancer dataset.

2.3 Other Classification Algorithm

Common data mining used in the medical field such as classification, association, Naïve Bayes, clustering and decision tree with the purpose of making a prediction about the disease in the early stage to help the medical practitioners perform the future diagnosis [10]. K-nearest neighbor is one of the important algorithms in Artificial Intelligence used for the diagnose of disease. Genetic Programming was developed by Aslam, Zhu and Nandi [11] to classify diabetes patients. Two classifier algorithms involved in this study are the K-nearest neighbor and the Support Vector machine to test the selected feature. Another study that implements K-nearest neighbor algorithm to predict diabetes is a study conducted by Krati, Saxena, Khan and Singh [12]. The obtained result show that

- *Nur Hadirah Khairul Anwar is currently a postgraduate student in Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perlis Branch, Arau Campus, 02600 Arau, Perlis, Malaysia. E-mail: hadirah1995@gmail.com*
- *Rizauddin Saian (corresponding author) is currently a senior lecturer in Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perlis Branch, Arau Campus, 02600 Arau, Perlis, Malaysia. E-mail: rizauddin@uitm.edu.my*

when the k value increases, accuracy and error rate will also increase. A fuzzy system is used for the diagnose of diabetes. Ganji and Abadeh [13] used FCS-ANTMINER that is based on the Ant Colony classification system to extract a set of fuzzy rules with the purpose of performing diagnosis on diabetes disease. An expert system, for example, is best known to diagnose type 1 diabetes such as research done by Lalka and Jain [14]. Besides that, fuzzy Mamdani model is also well known in fuzzy system for the identification of diabetes like research done by Bhandari and Kumar [15] where they used Mamdani-type and Sugeno-type-fuzzy expert systems with the use of multiple parameters for diabetes diagnosis and compared the performance. Type 2 diabetes is known as non-insulin dependent diabetes mellitus (NIDDM) or adult onset diabetes. Mostly, type 2 diabetes is because of older age, obesity, non-active person and because of family history [5]. There are several machine learning methods that are used to predict type 2 diabetes for example like research done by Yue, Xin, Kewen and Chang [16] that use quantum particle swamp optimization (QPSO) algorithm and weighted least square support vector machine.

3 METHODOLOGY

For this study, we will used two different diabetes datasets, the Frankfurt Germany diabetes dataset and Pima Indian diabetes dataset. Both datasets were obtained from Kaggle dataset website. In the origin, Frankfurt Germany diabetes dataset contains 2000 number of dataset and Pima Indian diabetes dataset has 768 of dataset. In these datasets, there are 8 attributes which are the number of times the patient conceived, plasma glucose concentration after 2 hours in a oral glucose tolerance test, diastolic blood pressure, triceps skinfold thickness, 2-hour serum insulin, body mass index, diabetes pedigree function and age. Normalization is the process of eliminating noise and irrelevant data. After the normalization process, Frankfurt Germany diabetes dataset only left 1035 and Pima Indian diabetes dataset has 392 number of datasets. For these datasets, normalization is applied to eliminate missing value. In Pima Indian diabetes dataset five had glucose at 0 value, 11 had 0 BMI, 28 had a blood pressure at 0, 192 had skinfold thickness reading at 0 and 140 had serum insulin levels of 0. In Frankfurt Germany diabetes dataset 13 patients had glucose at 0, 28 had BMI at 0, 71 had a blood pressure at 0, 488 patients that had 0 value of skinfold thickness reading and 365 had insulin level at 0 value. None of these makes any sense in term of the medical data hence, we deleted all of that data thus only left data without missing value. By eliminating irrelevant data, a better prediction of accuracy can be made. It is a time consumption if we want to find the 0 value and eliminate it manually, therefore, we highlight one by one column and press ctrl F to find the value we want as shown in Fig. 1. After we click find all then we press Ctrl and A to select all and then go to home tab, in cells group click delete and next click delete sheet rows as shown in Fig. 2.

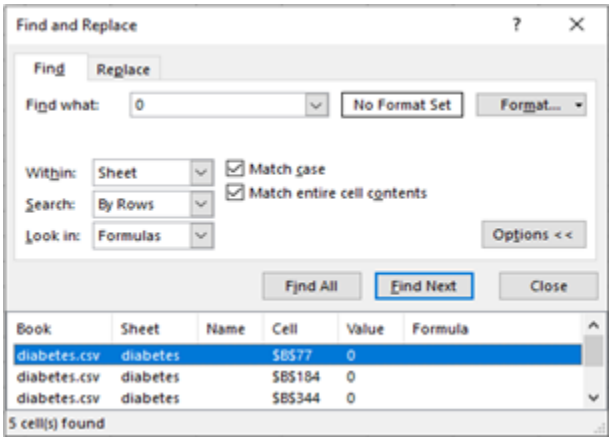

***Fig. 1**. Find command in Microsoft Excel.*

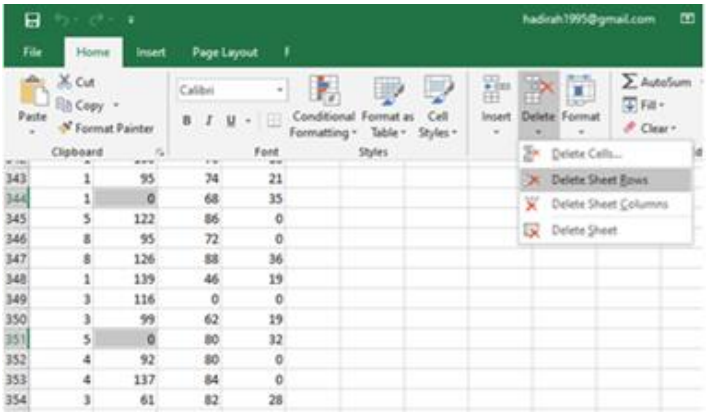

***Fig. 2**. Delete sheet rows in Microsoft Excel.*

After we eliminate all the missing values in both data, then the data need to be discretized first before we can classify the data. All the attributes in the datasets is in the form of numeric therefore discretization process is a must. The purpose of discretization is to change from numeric form to nominal form. Both datasets will be discretized using two methods which are equal-width and equal-frequency. The type of discretization used for both datasets is an unsupervised discretizes attribute. Frankfurt Germany diabetes dataset and Pima Indian diabetes dataset will be run in Gui Ant Miner to measure the accuracy. Gui Ant Miner can only analyze the nominal attribute and that is why we need to discretize the datasets first. In Gui Ant Miner, we only change the number of ants but everything else like cross validation fold = 10, min. cases per rule = 5, max. uncovered cases = 10, rules for convergence = 10 and a number of iterations = 100 remains the same. Another classification algorithm like Naïve Bayes, AdaBoostM1, K-nearest neighbor and RIPPER will be run using WEKA application to predict the accuracy.

4 RESULTS AND DISCUSSIONS

For Frankfurt Germany diabetes dataset, we will compare the accuracy obtained using Ant-Miner algorithm with Naïve Bayes and AdaBoostM1 while Pima Indian diabetes dataset however will be compared with K-nearest neighbor and RIPPER algorithm. Table 1 shows the result accuracy for Frankfurt Germany diabetes dataset using Ant-Miner algorithm with the number of ants is set to 10 compared with the accuracy value

obtained for Naïve Bayes and AdaBoostM1 by using WEKA application. Fig. 3 illustrates the accuracy value obtained using the different machine learning model for Frankfurt Germany diabetes dataset in the line chart.

TABLE 1
ACCURACY VALUE FOR FRANKFURT GERMANY DIABETES DATASET

Algorithm	Accuracy
Ant-Miner	79.79%
Naïve Bayes	79.61%
AdaBoostM1	77.10%

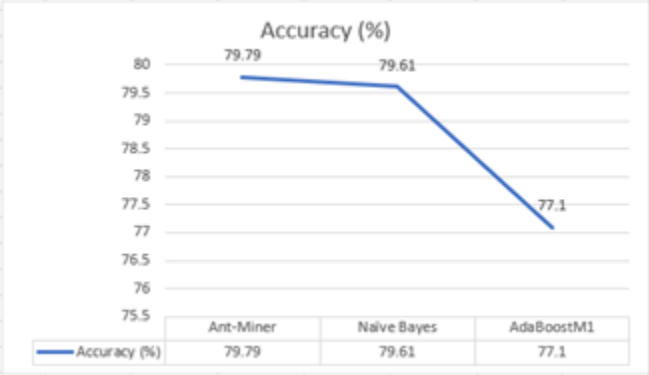

Fig. 3*. Line chart of accuracy value for Frankfurt Germany diabetes dataset.*

From the accuracy result above we can see that the accuracy value for the Frankfurt Germany diabetes dataset when using the Ant-Miner algorithm is higher compared to another classification algorithm. Table 2 shows the accuracy result for Pima Indian diabetes dataset obtained by using the different machine learning models. The highest accuracy obtained for the dataset is when we implement Ant-Miner algorithm which is 73.64% compared to other algorithms. Accuracy value for Pima Indian diabetes is illustrated using line chart as shown in Fig. 4.

TABLE 2
ACCURACY VALUE FOR PIMA INDIAN DIABETES DATASET

Algorithm	Accuracy
Ant-Miner	73.64%
RIPPER	72.95%
K-nearest neigbor	71.93%

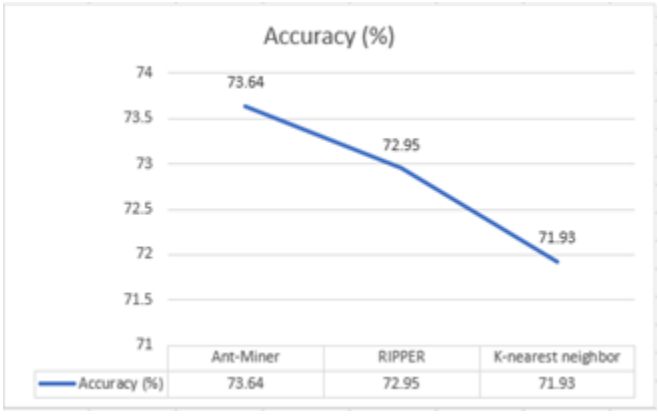

Fig. 4. Line chart of accuracy value for Pima Indian diabetes dataset.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we use two different diabetes datasets, the Frankfurt Germany diabetes dataset and the Pima Indian diabetes dataset. Various machine learning model involved in this study like Naïve Bayes, AdaBoostM1, K-nearest neighbor and RIPPER. The main algorithm that we used in this research is Ant-Miner to make a comparison in term of accuracy value. For Frankfurt Germany diabetes dataset, it will be compared between accuracy value obtained by using Ant-Miner and when using Naïve Bayes and AdaBoostM1. Accuracy value for Pima Indian diabetes however will be compared when using Ant-Miner and accuracy obtained when implementing K-nearest neighbor and RIPPER algorithm. The result shows that the method of using Ant-Miner algorithm for both diabetes datasets obtained good classification accuracy compared to other algorithms. For future work, this paper suggests to make an improvement to the existing algorithm Ant-Miner and doing an analysis to the attribute in the dataset, for example, like choose the most influential attribute that leads to diabetes disease and ignore attribute that less influential.

ACKNOWLEDGMENT

The authors wish to thank the Ministry of Higher Education Malaysia for funding this study under the Fundamental Research Grant Scheme, FRGS/1/2018/STG06/UITM/02/3 and RMC, Universiti Teknologi MARA, Malaysia for the administration of this study.

REFERENCES

[1] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann. 2016.
[2] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier. 2011.
[3] M. Durairaj and V. Ranjani, Data mining applications in healthcare sector: a study. International Journal of Scientific & Technology Research, vol. 2, no. 10, pp. 29-35, 2013.
[4] I. Taranu and others, Data mining in healthcare: decision making and precision. Database Systems Journal, vol. 6, no. 4, pp. 33-40, 2016.
[5] D. Atlas, International diabetes federation. IDF Diabetes Atlas, 7th Edn. Brussels, Belgium: International Diabetes Federation, 2015.
[6] M. Nilashi, O. Ibrahim, M. Dalvi, H. Ahmadi, and L. Shahmoradi, Accuracy improvement for diabetes disease classification: a case on a public medical dataset. Fuzzy Information and Engineering, vol. 9, no. 3, pp. 345-357, 2017.
[7] M. Dorigo, Optimization, learning and natural algorithms. PhD Thesis, Politecnico Di Milano, 1992.
[8] R. S. Parpinelli, H. S. Lopes, and A. A. Freitas, An ant colony algorithm for classification rule discovery. In Data mining: A heuristic approach. IGI Global, pp. 191-208, 2002.
[9] J. Wahid and H. F. A. Al-Mazini, Classification of Cervical Cancer using Ant-Miner for Medical Expertise Knowledge Management, 2018.
[10] A. Azrar, M. Awais, Y. Ali, and K. Zaheer, Data Mining Models Comparison for Diabetes Prediction. International Journal of Advanced Computer Science and Applications, vol. 9, no. 8, pp. 320-323, 2018.

[11] M. W. Aslam, Z. Zhu, and A. K. Nandi, Feature generation using genetic programming with comparative partner selection for diabetes classification. Expert Systems with Applications, vol. 40, no. 13, pp. 5402-5412, 2013.
[12] D. Krati Saxena, Z. Khan, and S. Singh, Diagnosis of diabetes mellitus using k nearest neighbor algorithm. International Journal of Computer Science Trends and Technology (IJCST), vol. 2, no. 4, pp. 36-43, 2014.
[13] M. F. Ganji and M. S. Abadeh, A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis. Expert Systems with Aplications, vol. 38, no. 12, pp. 14650-14659, 2011.
[14] N. Lalka and S. Jain, Fuzzy based expert system for diabetes diagnosis and insulin dosage control. International Conference on Computing, Communication & Automation, IEEE, pp. 262-267, 2015.
[15] V. Bhandari and R. Kumar, Comparative analysis of fuzzy expert systems for diabetic diagnosis. International Journal of Computer Applications, vol. 132, no. 6, pp. 8-14, 2015.
[16] C. Yue, L. Xin, X. Kewen, and S. Chang, An intelligent diagnosis to type 2 diabetes based on QPSO algorithm and WLS-SVM. 2008 International Symposium on Intelligent Information Technology Applization Workshops, IEEE, pp. 117-121, 2008.