UDC 616.379-008.64-082 (519.767.6)

ShamiluuluS.[1*], Djakbarova U.[2]
Latuta K.N.[1], Baimuratov O.A.[1]
[1]Department of Computer Sciences Mathematics and Natural Sciences,
SuleymanDemirel University, Kaskelen, Kazakhstan
e-mail:shahriar.shamiluulu@sdu.edu.kz,konstatin.latuta@sdu.edu.kz,
olimzhon.baimuratov@sdu.edu.kz
[2]Medical Biology and Genetics Department, Faculty of Medicine,
International Ataturk Alatoo University, Bishkek, Kyrgyzstan
e-mail:umida.djakbarova@iaau.edu.kg,

COMPARATIVE ANALYSIS OF DIABETES PATIENTS MEDICAL DATA USING SUPERVISED MACHINE LEARNING ALGORITHMS AN EDUCATIONAL APPROACH

Abstract. The comparative analysis has been performed on tree supervised machine learning algorithms i.e., decision trees, logistic regression and multi-layer perceptron neural network over Waikato Environment for Knowledge Analysis tool on medical data for patients with two types of diabetes disorders. Diabetes-Mellitus refers to the metabolic disorder that happens from malfunction in insulin secretion and action. It is characterized by hyperglycemia.The diagnosis of diabetes is very important now days using various types of techniques.The dataset has been obtained from UCI machine learning repository for Pima Indian Diabetes patients. The purpose of the research study is to show the usage of different machine learning algorithms and performance metrics for educational purposes in teaching machine learning to students from information systems field in more efficient and non-trivial way. During the research the comparative analysis studies have been performed which revealed that less complex algorithms can be used for disease diagnosis and possess better performance when properly configured.

Keywords: Diabetes Mellitus, Supervised Learning, Performance Analysis, Clinical Decision Support Systems.

I. INTRODUCTION

Computer aided diagnosis plays an important role in medical field. Presently population increasing and medical institutions becoming larger, this opens door for useful decision support systems that can analyze large amount of information. It has been shown that the benefits of introducing machine learning into medical analysis are to increase the diagnostic accuracy, to reduce costs and to reduce human resources. In this study, the comparative performance analysis of three supervised machine learning algorithms are studiedi.e, decision trees, logistic regression and artificial neural network by using Waikato Environment for Knowledge Analysis machine learning toolbox [3] for educational purposes. The algorithms are tested over the Pima Indian diabetes dataset. Pima Indian Diabetes database had been examined with several different machine learning methods in the past [5-9]. Diabetes-Mellitus refers to the metabolic disorder that happens from malfunction in insulin secretion and action. It is characterized by hyperglycemia. There are two types of diabetes disorder but generally the symptomatic and lab results are same. The diagnosis of diabetes is very important now days using various types of techniques.

II. MATERIALS AND METHODS

A. Data Collection

The data set was obtained from the UCI Repository of Machine Learning Databases [3]. The data set was selected from a larger data set held by the National Institutes of Diabetes

and Digestive and Kidney Diseases. The patients in the Pima-Indian dataset are women at least 21 years old and living near Phoenix, Arizona, USA. The dichotomousoutcomeattribute takes the values ‘0’ or ‘1’, where ‘1’ means a positive test for diabetes and ‘0’ is a negative test for diabetes. There are 500 (65.1%) cases in class ‘0’ and 268 (34.9%) cases in class ‘1’. The dataset contains eight clinical findings which are:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index
7. Diabetes pedigree function
8. Age (years)

B. Decision Trees

A decision tree (DT) is a graph that uses a branching method to illustrate every possible outcome of a decision for particular. The goal in building a tree is to identify a best splitting attribute which is being found by Entropy and Information Gain. The detailed theoretical background regarding decision trees can be found here [2].

C. Logistic Regression

The logistic regression (LRM) has wide range of implications in medical research field. The LRM model is used for the classification of the attributes, which might help to classify the outcome. The distinctive feature of the model is that the outcome variable is dichotomous. The result is not bounded to a linear form. As a result, the created model can be used to classify a newly provided data via placing them in a model for the probability P, the detailed information is provided in [1].

D. Multi-Layer Perceptron Neural Networks

Multi-layer perceptron neural networks (MLP)are processing devices, whichclosely resemble a modelof the neuronal structure of the mammalian cerebral cortex. LargeMLPs might have hundreds or thousands of processor units, whereas a mammalian brain has billions of neurons with a corresponding increase in magnitude of their overall interaction and emergent behavior. Generally the neural network has three components or layers. The first layer is called input layer, through which it gets data inside network on our case disease related attributes. The second layer is called hidden where all operations performed. The last layer called out where network make final decision regarding patient’s condition. The detailed information regarding neural networks provided in [1, 2].

E. Simulated Program

The Waikato Environment for Knowledge Analysis (WEKA),is one of the best tools in teaching machine learning without going into details first. The tool is basedon Java platform that contains a large number of algorithms for data preprocessing, feature selection, classification, clustering, and finding the associative rule [4]. WEKA uses a common data representation format, making comparisons easy. It has three operation modes i.e., GUI, Command Line, and Java API.

F. Performance Measures

Evaluation of the classifier to measure the quality is commonly evaluated based on the data in the confusion matrix. Several standard measures have been defined for correct and incorrect classification results of the matrix. The most common practical measure to evaluate the performance is accuracy, which is defined as the proportion of the total number of instances that were classified correctly.

Recall is the mean proportion of actual positives which are correctly identified. Precision is the mean proportion of positives which are relevant. F-measure is a harmonic mean of recall and precision.

These performance metrics are calculated according to the data in the confusion matrix which are obtained by the WEKA tool.

III. SIMULATION RESULTS

In this study, the Pima dataset of patients with diabetes disorder, which containing 9 original features by using three machine learning methods i.e., DT (C4.5), LRM and MLP used for classification. The performance metrics like accuracy, recall, precision and f-measure along with error metrics forall features has been performed using 10-fold cross-validation.The simulations were performed by using WEKA3.8 machine learning tool.

	No of correct ins. with %	No of incorrect ins. with %	Build time
DT	567 - (74%)	201 - (26%)	0.12 sec
LRM	593 - (77%)	175 - (23%)	0.15 sec
MLP-NN	583 - (76%)	185 - (24%)	1.45 sec

Table 1: Accuracy metrics of ML algorithms

According to the results provided in Table 1, the LRM model outperforms remaining methods with the overall accuracy of 77% even though the MLP is considered one of top classification methods it came the second one in this race. During the analysis we have identified that this problem was due to overfitting issue.

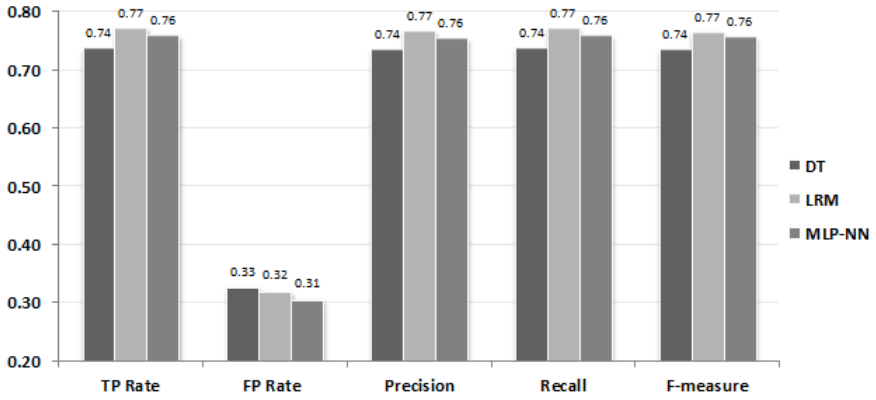

Figure 1: Performance measure metrics of ML algorithms

The Figure 1 contains the five different performance measure results for three machine learning algorithms. Based on the results the LRM method outperforms MLP by 1% and DT by 3% on overall. Even though the MLP considered the complex method for classification and prediction the complex nature of it, which contained one hidden layer with 5 nodes fall into problem of overfitting. In the literature survey we have find out that for three algorithms the accuracy metric was varying from partition to partition. For example, when we see the accuracy of C4.5 model, 77.08% in case of 75-25% training-testing partitions, 76.72% in case of 85-15% training-testing partitions and 75.32% in case of 90-10% training-testing partitions.

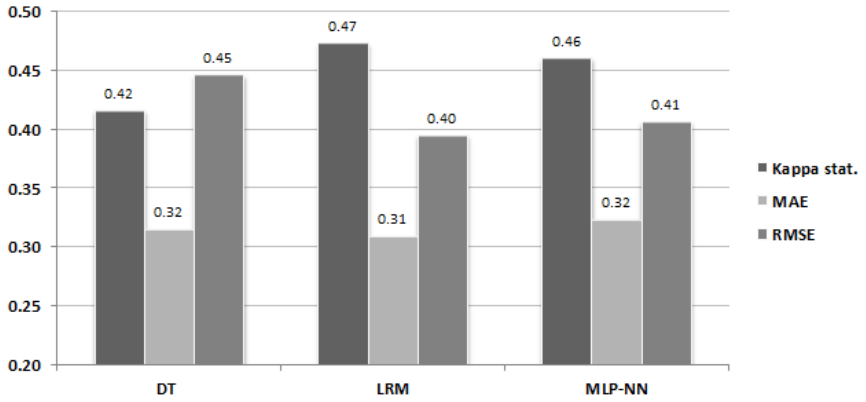

Figure 2: Error metrics of ML algorithms

The Figure 2 shows three error metric results i.e., Kappa statistics, Mean Absolute Error (MAE) and Root mean square error (RMSE) for simulated machine learning algorithms. The Kappa statistics measures the agreement of prediction with the true class. The value which is bigger than zero indicates that algorithm is not performed based on chance but rather taking more logical approach. The MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction and the value close to zero is better. It measures accuracy for continuous variables. The RMSE measures the average magnitude of the error and the value close to zero is better. Based on simulation results shown in Figure 2, the LRM outperforms all other methods.

IV. CONCLUSION

In this research study threesupervised machine learning algorithms were applied to the Pima Indians Diabetes (PID) medical dataset. The performance of LRM was the best for all performance and error metrics. The LRM method outperforms MLP by 1% and DT by 3% on overall. MLP is considered one of top classification methods it came the second one. During the analysis we have identified that this problem was due to overfitting issue. As a result shows that, LRM methods can be a good and practical choice to classify a medical data.

References:

1 Norvig P, Russell S. Artificial Intelligence: A Modern Approach, Prentice Hall. 2002.
2 Dunham MH. Data mining: Introductory and advanced topics, Pearson Education, 2006.
3 UCI Repository of Machine Learning Databases, University of California at Irvine, Department of Computer Science. Available: https://archive.ics.uci.edu/ml/datasets/Diabetes (Accessed: 7 Sept. 2016).
4 HallM., FrankE., and HolmesG., B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," ACM SIGKDD explorations newsletter, 2009. - Vol. 11, pp. 10-18.
5 Raj Anand, Vishnu Pratap Singh Kirar, Kavita Burse, " K-Fold Cross Validation and Classification Accuracy of PIMA Indian Diabetes Data Set Using Higher Order Neural Network and PCA ", IJSCE, 2013.-Vol. 2, Issue-6, pp. 436-438.
6 Y. Angeline Christobel, P.Sivaprakasam, " A New Classwise k Nearest Neighbor (CKNN) Method for the Classification of Diabetes Dataset", IJEAT, 2013. – Vol. 2, Issue-3, pp. 396-400.
7 Kumari V. Anuja, Chitra R. Classification of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications,2013. - Vol. 3, pp. 1797-1801.
8 Carpenter G.A., Markuzon N., "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases", Neural Networks, 1998. - 323-336 pp.

9 Deng, D., Kasabov, N., “On-line pattern analysis by evolving self-organizing maps”, Proc. of the 5th Biannual Conference on Aritificial Neural Networks and Expert Systems (ANNES), Dunedin, November, 2001. pp. 46-51.
10 Farahmandian M., Lotfi Y., Maleki I. Data Mining Algorithms Application in Diabetes Diseases Diagnosis: A Case Study. MAGNT Research Report, 2015. - Vol. 3, pp. 989-997.

UDC 004.8

Tolebi G.A.
M.Sc., Kazakh-British Technical University,
Almaty, Kazakhstan, e-mail: tolebi.glr@gmail.com

DESCRITION OF THE COMPUTATIONAL INTELLIGENCE TECHNIQUES FOR ADAPTIVE TRAFFIC SIGNAL CONTROLLER: REINFORCEMENT LEARNING

Аңдатпа. Аталмыш жұмыс бейімделетін бағдаршам жүйесін құру үшін қолданылатын есептеу интеллект әдістерін сипаттауға арналған. Атап айтқанда, бағалау арқылы үйрену технологиясының түрлерін қолдану жолдары берілген. Жолдағы жағдайға байланысты әрекет етуі өзгеретін басқарушыға қолданылатын алгоритмдерге салыстырмалы талдау жасалған.

Кілт сөздер. traffic signal controller, reinforcement learning, artificial neural network, actor-critic method, Q-learning.

I INTRODUCTION.

Traffic congestion one of the big problem for the big cities. It becomes serious issue, since growth of number of vehicles in magisterial. An extension of roads or building of new ones can be considered as one of the solution of current problem. However, it requires many expenses and human resources. Nowadays, Computational Intelligence (CI) is widely used to perform the applied problems. Extremely development of techniques of CI gives powerful tools for solving problems for nonlinear stochastic systems like traffic flows. Proper modeling of the traffic flows is complex task. Therefore, many researchers consider way of solution of problem without mathematical modeling of vehicles movement. In the traffic control problem traditional supervised learning techniques, such as Support Vector Machine, Random Decision Tree, feed-forward artificial neural network cannot be used. Because an environment is dynamic at the given problem. There are no labeled targets for learning. The traffic control model requires unsupervised learning technology. In the current work, evaluation of existing CI methods of traffic signal control is proposed. Methods for traffic signal management by applying Reinforcement Learning (RL) technique are described. This is more promised technique of CI, which is successfully applied to control problems.

In this paper Q-table, ANN and Actor-Critic implementation of RL is adopted to develop adaptive TSC for an intersection. The remaining part of this paper is organized as follows: Section II is literature review; Section III describes the RL and Q-table, ANN and Actor-Critic implementation and evaluation. And Section IV is conclusion.

II. BACKGROUND

Management of traffic flows was starting from the XIX century in London [1]. It was semaphores which are controlled manually. During the passing of time, the controller has been changed and has automotive control system. According to working principle, all TSC are divided into three groups: pre-timed, actuated and adaptive or intelligent. Pre-timed signal controllers have fixed time plan. Duration of phase time and whole cycle time, phase sequence are fixed. Optimal time for each phase, usually calculated based on the Webster formula and using historical data [2, 3]. It is a mostly used type of TSC. Advantage of fixed-time is simple algorithm, which can be easily implemented and action is predicted for the vehicles. Actuated