TELKOMNIKA Telecommunication, Computing, Electronics and Control
Vol. 18, No. 4, August 2020, pp. 1777~1783
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v18i4.13849

Enhancement of student performance prediction using modified K-nearest neighbor

Saja Taha Ahmed[1], Rafah Al-Hamdani[2], Muayad Sadik Croock[3]
[1,2]The Informatics Institute for Postgraduate Studies, Iraqi Commission for Computers & Informatics (IIPS-ICCI), Iraq
[3]Computer Engineering Department, University of Technology, Iraq

Article Info

Article history:

Received Aug 9, 2019
Revised Feb 25, 2020
Accepted Mar 18, 2020

Keywords:

Consuming time
Educational data mining
Moments
KNN
Prediction

ABSTRACT

The traditional K-nearest neighbor (KNN) algorithm uses an exhaustive search for a complete training set to predict a single test sample. This procedure can slow down the system to consume more time for huge datasets. The selection of classes for a new sample depends on a simple majority voting system that does not reflect the various significance of different samples (i.e. ignoring the similarities among samples). It also leads to a misclassification problem due to the occurrence of a double majority class. In reference to the above-mentioned issues, this work adopts a combination of moment descriptor and KNN to optimize the sample selection. This is done based on the fact that classifying the training samples before the searching actually takes place can speed up and improve the predictive performance of the nearest neighbor. The proposed method can be called as fast KNN (FKNN). The experimental results show that the proposed FKNN method decreases original KNN consuming time within a range of (75.4%) to (90.25%), and improve the classification accuracy percentage in the range from (20%) to (36.3%) utilizing three types of student datasets to predict whether the student can pass or fail the exam automatically.

Corresponding Author:

Saja Taha Ahmed,
The Informatics Institute for Postgraduate Studies,
Iraqi Commission for Computers & Informatics (IIPS-ICCI), Iraq.
Email: sajataha@ymail.com

1. INTRODUCTION

The growth of the internet and communication technologies have contributed to the dissemination of e-learning to support certain countries confronting a rising scarcity of instructors [1]. It is evident that understudies, or individuals in general, who are looking for information can achieve this effectively and with minimal effort at any time and anywhere. This encourages various colleges and instructive organizations to adopt an online learning framework with an extension of the student data volume. However, e-learning has a lot of impediments and challenges and drop-out rates for students are more common than conventional learning [2, 3].

The educational data mining (EDM) is utilized to develop a model that can influence e-learning system because data gathered from e-learning system often exceeds large numbers of students [4]. The involvement of different variables that EDM can exploit for model building helps the educational system to perform better. The developed model can support the decision making of educational institutions and universities about future of their students, for example, distinguishing effective students from a given course and perceiving students who will drop out or fail to pay more consideration during course progress [5].

The K- the nearest neighbor is one of the simplest EDM algorithms [6]. It is computationally simple based on similarity measures such as a Euclidean distance metric with majority voting of the K closest training sample class assigned to the test sample [7]. KNN is an instance-based learner, sometimes called a lazy learner, as it defers the training until a new student (test sample) should be classified (i.e. there is no training phase) with most the power relies on matching scheme [8]. KNN has some cons that can be listed as [9, 10]:

- Computational overhead is extremely high as each new student needs to calculate the distance to all training samples.
- The capacity requirement is huge in proportion to the training size set.
- KNN with multidimensional data sets has a minimum accuracy rate.

Researchers offered a variety of techniques for dealing with the issues of traditional KNN algorithm and improving its performance. The authors of [11] proposed that the genetic algorithm (GA) and KNN were combined to improve the classification performance. GA was used to instantly pick up k-neighbors and calculate the distance to classify the test samples. The proposed method was compared with the traditional KNN, CART and SVM classifiers. The results showed that the proposed method reduced complexity and improve accuracy.

The authors in [12] solved the large sample computation problem using a cure clustering algorithm with KNN to obtain representative samples of the original dataset for text categorization. The proposed method classified 6500 news essays from 8 categories of Sina websites with improved computation speed compared to traditional KNN but did not enhance the accuracy of KNN, which is considered as a major limitation of the proposed method.

The author in [13] focused on improving the performance of KNN by combining local mean based KNN with distance weight KNN. The proposed method was applied to four datasets from UCI, kaggle, and keel, in addition to a real dataset from public senior high school. The obtained results appeared that the classification accuracy of the proposed method compared to KNN was increased, but this research ignored the complexity of execution time resulted from the mixing of proposed methods.

The KNN computational complexity for classifying a single new instance is O(n), where n is a number of training samples [14]. Therefore, in this study, the prototype storage, computation time and accuracy have a great deal of analysis. This paper proposed a solution by introducing an acceleration scheme to overcome KNN drawbacks via a combination of moment descriptors with traditional KNN. The moment descriptors have been utilized well in multimedia research for various applications, such as musical similarity and song year prediction [15], speed up color image fractal compression [16] and enhance fractal audio compression [17]. The training set will be arranged into subsets; samples belong to the same subset have similar descriptor number. The proposed FKNN does not have to test each new sample (i.e., compute its distance) with all training samples. but, each test sample (new student) when the proposed FKNN computes its descriptor value is matched only with a predetermined subset of training samples which has similar descriptor value. This significantly reduces the execution time (comparison distance time) and memory requirements. In addition, each training subset is formed on the basis of a weighted moment descriptor that captures the importance of selected attributes for different samples, this enables each training subset to contain the most similar samples. It, in turn, increases the accuracy of a classification and avoids double majority classification (i.e. misclassification).

2. RESEARCH METHOD

This study will include two phases as a part of the methodology, as follow:

2.1. Dataset collection and preparation

The collection, preprocessing and feature selection of data sets are done based on our research work of [18]. This study used three datasets, the first being the Iraqi student performance prediction dataset, which is collected through applying (or submitting) questionnaire in three Iraqi secondary schools for both applicable and biology branches of the final stage during the second semester of the 2018 year and uploaded to [19] with full description. While the second and third datasets (student alcohol consumption dataset), are obtained from UCI Portugal [20], which incorporates two datasets: student-mat.csv and student-por.csv. Dataset preprocessing includes the following steps:

- Dataset encoding: the dataset contains attributes of various data types, for instance: binary, interval, numeric and categorical (nominal, ordinal). The KNN requires data to be in the numerical formulation. This is due to that there are many feature encoding methods for transforming categorical data to numeric ones, such as label encoding or integer encoding, one-hot encoding, binarized and hashing. In this research, the datasets are encoded using Label Encoder, which is the most common method to transform categorical features into numerical labels. Numerical labels are always being between 0 and (#attribute_value-1).

- Dataset normalization: in the machine learning algorithm where the distance plays a vital role like KNN, the datasets must be normalized for a better predictor (i.e. avoid misclassification) and to efficiently train the algorithm. The normalization is the process of scaling attribute values within a specific range (such as 0 to 1), in a manner that all attributes have approximately similar magnitudes. This research normalizes the attribute values using Min-Max normalization at the range [-1, 1].
- Feature selection: the results of the proposed feature selection in [18] obviously show that the highest performance accuracy is achieved by social factors in combination with marks. This research selects top eight features subset based on Pearson correlation ranking criteria. Feature subset of Iraqi dataset includes the following questions: "Q37 Worry Effect", "Q20 Family Economic Level", "Q25Reason of study", "Q27 Failure Year", "Q8 Father Alive", "Q17 Secondary Job", "Q33 Study Hour", "Q23 Specialization", while UCI.student-por.csv feature subset includes: "Q10 reason" , "Q8 Mjob", "Q21 internet", "Q3 address", "Q7 Fedu", "Q6 Medu", "Q13 study time", "Q20 higher", and UCI.student-mat.csv feature subset has "Q17 paid", "Q8 Mjob", "Q1 sex", "Q3 address", "Q10 reason", "Q7 Fedu", "Q20 higher", "Q6 Medu features".

2.2. The proposed method

In this study, the proposed FKNN utilizes the concept of moment descriptor which is a set of parameters that describe the distribution of material [21]. The main idea is the similarity between attributes value of new student (test sample) and previously registered students (trained examples), since if two samples have same descriptors they are going to have approximately similar performance. From this point of view, this research comes out with the contribution of enhancing the performance of KNN by employing moment descriptor to pre-classify students. This strategy uses the descriptors as a reference indicator to pre-classify the training samples into groups with a specific descriptor value based on social and academic factors. The reason for adopting this classification concept is that the descriptor of each student represents a signature to differentiate the student behavior. Therefore, instead of making the full search during distance computation with the whole training set, only a subset of these samples is computed. The moments are determined by exploiting two first-order moments D1 and D2, as shown mathematically in (1):

$$D_{1,2} = \sum_{j=0}^{v-1} W[j] * S[i,j] \quad (1)$$

where v is the length of the feature subset. S represents attribute j of sample i in a dataset. W is a mathematical representation chosen for a better separation control. The adopted weights in this research are:

$$w1[j] = Cos[\frac{2j}{v-1}\pi] \quad (2)$$

$$w2[j] = \begin{cases} \frac{2j}{v-1} & if\ j \leq \frac{v-1}{2} \\ \frac{-[2(v-1)-j]}{v-1} & otherwise \end{cases} \quad (3)$$

where j=0 … v-1, The descriptor of both training and testing sample is determined using the following mathematical (4):

$$Des = \frac{D_1^2 - D_2^2}{D_1^2 + D_2^2} \quad (4)$$

The proposed FKNN needs to determine the index value for each sample (i.e. student), the determined descriptor (Des) is converted into integer value within range [0, No_sub], where No_sub is the number of training subsets, the descriptor index value for each sample (Des_Index) is computed using the (5):

Des_Index=Round (Absolute(Des)* No_sub) (5)

In addition, the proposed FKNN needs to construct a data structure (DS) to warranty faster access to the samples. This data structure contains the identification number and descriptor index for all training samples. The samples of a data structure are arranged in ascending order according to samples' descriptor index. Therefore, all samples that have the same descriptor will form a class (i.e. training sample subset) in contiguous locations. The pre-classification of a training set into subsets is clarified by the following pseudo-code written as Algorithm 1.

The next step is to calculate the frequency for each descriptor index in the sorted data structure (DS) and set an array of pointer to indicate the start and end for each training subset. In such a way, the limitations

of each subset are indicated by pointers that act as leading signs to reach the intentional class immediately. The steps for building an array of pointers are illustrated in the following pseudo-code, presented as Algorithm 2.

Algorithm 1. Training Set Pre-Classification

```
Input: Training sample as a matrix [# students, # attributes]
Output: Sorted Data Structure contains samples classifying according to their
descriptor index field value.

Define DS as a data structure which is an array of records contains two elements
student descriptor index and his positioning (or, identifier) in training set.
Set the number of descriptor classes to No_sub.
For each index j of feature vector length // Calculate weights in the range of
feature vector length.
 Begin
      Compute w1[j] using equation 2
      Compute w2[j] using equation 3
End
For each student i in the Training set
Begin
      For each attribute j in the feature vector
          Compute D1 and D2 based on equation 1.
Compute Descriptor of student i (Des) by using equation 4
Compute Sample Descriptor Index (Des_Index) using equation 5
Set DS[i]. Index=Des_Index
Set DS[i]. Identifier=i
End
Sort elements of data structure (DS) according to descriptor field.
Return DS
```

Algorithm 2. Pointers

```
Input: Sorted data structure (DS) of samples' descriptors and identifiers
Output: an array of pointers Pointer[#No_sub]
Define Freq [#No_sub] as an array of integer hold the occurrences of descriptor
index (Des_Index) in DS.
For each student i in training dataset
Begin
      Set X=DS[i]. Index
      Increment Freq[X] by one
End
Set Pointer [0] =0
For each value n in No_sub
Set Pointer[n]=Pointer[n-1] + Freq[n-1]
Return Pointer
```

After completing the task of sorting training samples, the matching process takes place by applying KNN. When samples of training subset are arranged at contiguous locations since they shared a similar descriptor index, as a result, each test sample is only compared with the specific training subset based on its descriptor index. Absolutely, this training subset has fewer samples than those found within the full training dataset. In addition, the best similar samples (in terms of their attributes) are most probably available in this training subset that has similar descriptor index. This led to a substantial reduction in running time of KNN and improves the accuracy of classification. The similarity measurement is based on the Euclidean distance between the test sample and samples of the training subset. The calculated distances are stored in a sorted ascending order array. If the distance has zero value, the label of the corresponding sample is considered as target class directly, otherwise, the k training sample is picked out and the target class of the new sample is determined by the use of the majority voting concept. The following pseudo-code (algorithm 3) explains the steps involved in applying the proposed (FKNN) for test samples:

Algorithm 3. The Proposed FKNN

```
Input: test set as matrix [#students, # attributes]
Output: target class for a test samples
For each student t in the test set
Begin
    //Define DS2 to contain distance and training sample identifier.
    Calculate descriptor of student t using equation 4
    Calculate descriptor index of student t to get Des-test using equation 5
   //Determine the start and end index for training subset which has the same
descriptor index as student t
    using an array of the pointer.
```

```
    Set Start=Pointer[Des-test]
    Set End=Freq[Des-test] +Pointer[Des-test]
    Set x=0
    For each training sample i in the range from Start to End
     Begin
      //Calculate Distance between new student and train students which have the
same
       descriptor in the sorted data structure DS
       Set ID=DS[i]. Identifier
       For each attribute j in feature vector         //calculate Euclidean
distance
```

$$\textbf{Distance}[x] = \sqrt{\sum_{j=0}^{v-1} \textbf{newstudent}[t, j] - \textbf{trainstudent}[ID, j]}$$

```
       Set DS2[x]. Distance=Distance[x].
       Set DS2[x]. ID=ID.
       Increment x by one
     End
Sort DS2 according to the distance in ascending order
If DS2[0]. Distance is equal to zero,
                 then the target class is the label of the sample that has DS2[0].
ID.
Else Begin
         //Matching via Majority Vote
         Pick the first K entries from Distance
         Get the labels of the selected K entries
         End
End
Return the target class of the majority K labels
```

3. RESULTS AND ANALYSIS

The experiments and the application system are performed based on visual studio.net C# 2015. The evaluation of the proposed method is performed using holdout validation, which splits datasets into two sets: 70% training and 30% test. Accuracy (ACC) is considered to measures the degree to which the instances correctly classified by the machine learning algorithm in proportion to the entire tested instances [22]. As mentioned earlier, the main aim of this work is the prediction of student performance. For this purpose, the target class label is formulated for each dataset, which can be either "Pass" or "Fail". There are three averages of G1, G2, and G3 in the UCI dataset with values from 0 to 20. Therefore, if the student has a grade equal to or greater than 10, it should be classified under the "Pass" label, otherwise, it should be classified as a "Fail" label. In Iraqi dataset, grade values are within range of (0-100). If the student has a grade equal or higher than 50, it should be defined within the "Pass" label, otherwise is classified as "Fail" student.

The students' performance of the UCI datasets is predicted based on final semester grades (G3) as the objective class. The Iraqi dataset prediction of the target class is done using the second-semester average (Avg2). In this work; for the purpose of comparing results among datasets, certain parameters must be established such as the number of descriptor classes (i.e. a number of bins) which set to a value of five and the value of K considered to be three nearest neighbors.

In the perspective of traditional KNN issues, the proposed FKNN has proved that it runs faster for all test samples than traditional KNN since FKNN requires a smaller number of comparisons based on the distance calculation of each new sample information from a subset of training data containing the same descriptor index as the new sample. This can also reduce memory requirements significantly. In contrast to traditional KNN, it is being slow because of the dependency on the exhaustive search of each new sample with all training data and requires more memory capacity to store distances of whole training samples. Figure 1 indicates that the running time of the proposed method is improved compared to the traditional KNN time.

Comparison a common way to measure the processing effect is to compare the outcome of interest before processing with that after processing. The percentage change measures an item's change in value relative to its original value. Suppose x is the baseline value, y is the post-processing value. The Percentage change can be calculated using (6) [23]:

$$PC=((X-Y)/X) *100 \quad (6)$$

Table 1 summarizes the percentage change of running time based on the results shown in Figure 1. It can be seen that the proposed FKNN reduces the time complexity of the traditional KNN by (90.25 %), (87.53 %), and (75.4 %) for Por, Math, and Iraq, respectively.

The performance of the proposed FKNN achieves better classification accuracy than traditional KNN for all datasets. This is due to that the proposed FKNN relies on the weighted moment descriptor samples to construct training subsets that have higher class discriminatory information. This can lead to getting the best

matching distance among new and training samples. Via this selective scheme, the misclassification problem of traditional KNN can be significantly overcome. Referring to Table 2 the proposed method obtains the highest accuracy of 100% for Iraqi student performance dataset. In addition, it can be seen that the proposed method is able to enhance classification accuracy for final semester grade prediction (G3) by obtaining the percentage change in accuracy of (36.3%), (23.7%), and (20%), for Por, Math and Iraq datasets, respectively.

Table 3 shows a comparison of the proposed FKNN with the research work of [24]. This research uses Por dataset from UCI to predict student performance based on eight features G2, G1, failures, higher, Medu, school, studytime, Fedu. In addition, a comparison of the proposed FKNN with the research work of [25]. This research uses Math dataset from UCI to predict student performance based on 19 features including the class attribute: sex, famsize, address, pstatus, medu, fedu, mjob, fjob, traveltime, studytime, schoolsup, higher, internet, romantic, freetime, Dalc, Walc, health, success. It is clear that the proposed DDT surpass all methods utilized in these researches for two UCI (Por and Math) datasets.

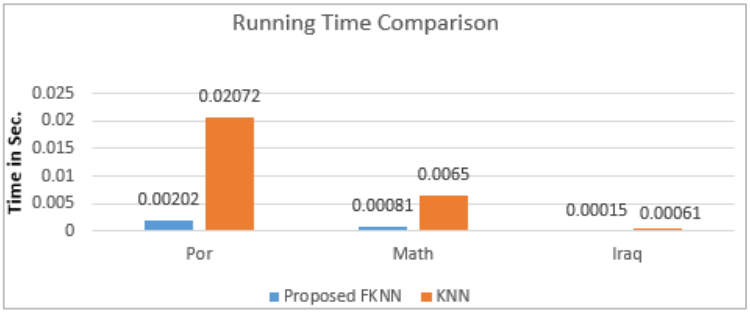

Figure 1. The comparison of the proposed FKNN and traditional KNN running time for final grade prediction

Table 1. Percentage change in running time of proposed FKNN

Dataset	PC Time
Por	90.25%
Math	87.53%
Iraq	75.4%

Table 2. The accuracy improvement of the proposed method vs. traditional KNN for final grade prediction

Datasets	Traditional KNN	Proposed FKNN	Accuracy Differences	PC in Accuracy
Por	69.2	94.3	25.1	36.3%
Math	78.1	96.6	18.5	23.7%
Iraq	83.3	100	16.7	20%

Table 3. Accuracy comparison of our proposed DDT and other methods for UCI datasets

Dataset	Research Work	Method	Accuracy
Por	[24] (2019)	Naïve Bayes	73.18 %
		Decision Tree	76.27 %
		RandomTree	67.95 %
		REPTree	76.73%
		JRip	74.11 %
		OneR	76.73 %
		SimpleLogistic	73.65%
		ZeroR	30.97%
	Our Proposed Model	The Proposed FKNN	94.3%
Math	[25] (2016)	PCF with k-medoids algorithm	65.82 %
		PCF with k-means algorithm	63.50%
	Our Proposed Model	The Proposed FKNN	96.6

4. CONCLUSIONS

This study presented the FKNN algorithm which combines the sample indexing mechanism with KNN to deal with the major problems of the traditional KNN. The computational overhead, memory requirement, multidimensionality (the number of samples) and misclassification problems were substantially reduced due to the pre-classification of training data based on the descriptors used by the selective search strategy. The classification accuracy was enhanced using the proposed FKNN method since the training sample grouped according to their similarity. The results showed a significant enhancement in accuracy with the highest increase reached to (36.3%) and improved the computation time of KNN with the highest time reduction reach to 90.25% for UCI.student-pro.csv dataset. The adopted experiments confirmed that the proposed FKNN outperformed the traditional KNN for all educational data sets. Therefore, the proposed algorithm was very useful for the real-time system such as e learning environment and could be used for larger datasets.

REFERENCES

[1] Andersson A., "Seven major challenges for e-learning in developing countries: Case study eBIT, Sri Lanka," *International Journal of Education and Development using ICT*, vol. 4, no. 3, pp. 45-62, 2008.

[2] Hasibuan Z. A., "Step-Function Approach for E-Learning Personalization," *TELKOMNIKA Telecommunication, Computing, Electronics and Control,* vol. 15, no. 3, pp. 1362-1367, 2017.

[3] Dogruer N., Eyyam R., Menevis I., "The use of the internet for educational purposes," *Procedia-Social and Behavioral Sciences,* vol. 28, pp. 606-611, 2011.

[4] Akibu M. A., Mokhairi M., Suhailan S., "The patterns of accessing learning management system among students," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, no. 1, pp. 15-21, 2019.

[5] Mining T. E., "Enhancing teaching and learning through educational data mining and learning analytics: An issue brief," *Proceedings of conference on advanced technology for education*, 2012.

[6] Agrawal R., "K-nearest neighbor for uncertain data," *International Journal of Computer Applications*, vol. 105, no. 11, pp. 133-16, 2014.

[7] Wisit L., Sakol U., "Image classification of malaria using hybrid algorithms: convolutional neural network and method to find appropriate K for K-Nearest neighbor," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 1, pp. 382-388, 2019.

[8] Garcia E. K., Feldman S., Gupta M. R., Srivastava S., "Completely lazy learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 9, pp. 1274-85, 2009.

[9] Hall M. A., "Correlation-based feature selection for machine learning," PhD Thesis, 1999.

[10] Alizadeh H., Minaei-Bidgoli B., Amirgholipour S. K., "A new method for improving the performance of k nearest neighbor using clustering technique," *Journal of Convergence Information Technology*, vol. 4, no. 2, pp. 84-92, 2009.

[11] Suguna N., Thanushkodi K., "An improved k-nearest neighbor classification using genetic algorithm," *International Journal of Computer Science Issues*, vol. 7, no. 2, pp. 18-21, 2009.

[12] Chen S., "K-nearest neighbor algorithm optimization in text categorization," *IOP Conference Series: Earth and Environmental Science*, 2018.

[13] Syaliman K. U., Nababan E. B., Sitompul O. S., "Improving the accuracy of k-nearest neighbor using local mean based and distance weight," *Journal of Physics: Conference Series*, vol. 978, no. 1, pp. 28-30, 2018.

[14] Hassanat A. B., Abbadi M. A., Altarawneh G. A., Alhasanat A. A., "Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach," *International Journal of Computer Science and Information Security*, vol. 12, no. 8, 2014.

[15] Foster P., Mauch M., Dixon S., "Sequential complexity as a descriptor for musical similarity," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1965-77, 2014.

[16] George L. E., Al-Hilo E. A., "Speeding-up Fractal Color Image Compression Using Moments Features Based on Symmetry Predictor," *2011 Eighth International Conference on Information Technology: New Generations*, 2011.

[17] Bedan A. K., George L. E., "Speeding-up fractal audio compression using moment descriptors," *Lambert Academic Publishing (LAP)*, 2013.

[18] Saja T. A., Rafah S. H., Muayad S. C., "EDM Preprocessing and Hybrid Feature Selection for Improving Classification Accuracy," *Journal of Theoretical and Applied Information Technology*, vol. 96, no 1, no. 1992-8645, 2019.

[19] Saja Taha, "Iraqi Student Performance Prediction," *Mendeley Data*, 2018. [Online]. Available: http://dx.doi.org/10.17632/smgx6s5pwr.1, DOI: 10.17632/smgx6s5pwr.1.2018

[20] Cortez P., Silva A. M., "Using data mining to predict secondary school student performance," *EUROSIS*,2008.

[21] Ghiringhelli L. M., Vybiral J., Ahmetcik E., Ouyang R., Levchenko S. V., Draxl C., Scheffler M., "Learning physical descriptors for materials science by compressed sensing," *New Journal of Physics*, vol. 19, no. 2, 2017.

[22] M. Z. H. J., Hossen A., Hossen J., Raja J. E., Thangavel B., Sayeed S., "AUTO-CDD: automatic cleaning dirty data using machine learning techniques," *TELKOMNIKA Telecommunication, Computing, Electronics and Control,* vol. 17, no. 4, pp. 2076-2086, 2019.

[23] Tu Y. K., "Testing the relation between percentage change and baseline value," *Scientific Reports*, vol. 6, pp. 1-8, 2016.

[24] Salal Y. K., Abdullaev S. M., Kumar M., "Educational Data Mining: Student Performance Prediction in Academic," vol. 8, no. 4C, pp. 54-59, 2019.

[25] Sati N. U., "Prediction of Students'success in Mathematics by a Classification Technique Via Polyhedral Conic Functions," *The Eurasia Proceedings of Educational & Social Sciences*, 2016.