PREDICTION OF CONCRETE STRENGTH WITH DATA MINING METHODS USING ARTIFICIAL BEE COLONY AS FEATURE SELECTOR

YAPAY ARI KOLONİSİNİ ÖZNİTELİK SEÇİCİ OLARAK KULLANARAK VERİ MADENCİLİĞİ YÖNTEMLERİ İLE BETON DAYANIMINI TAHMİN ETME

Mümine KAYA KELEŞ
Department of Computer Engineering
Adana Science and Technology University
Adana, Turkey
mkaya@adanabtu.edu.tr

Abdullah Emre KELEŞ
Department of Civil Engineering
Adana Science and Technology University
Adana, Turkey
aekeles@adanabtu.edu.tr

Ümit KILIÇ
Department of Computer Engineering
Adana Science and Technology University
Adana, Turkey
ukilic@adanabtu.edu.tr

***Abstract*— Concrete which is a highly complex material is the most basic input of the construction industry. Because of its strength, concrete is one of the most preferred structural building materials. In the ready-mixed concrete sector, there is an increasing need for earthquake resistant structures due to the fact that some producers produce out of control and poor quality. Ready-mixed concrete is a product whose quality can only be understood at the end of the 28th day if it is only controlled by taking the sample by the user. In this study, a data mining study was conducted on the factors affecting the 28-day compressive strength of concrete using the Concrete Slump Test Data Set from UCI Machine Learning Repository. The Artificial Bee Colony Algorithm is used as a feature selection method in order to determine the important ones of the concrete components, which are cement, slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate, affecting concrete strength and tried to predict the strength with data mining algorithms. As a result of the study, it was observed that Random Forest Algorithm gave the highest success rate with 91.2621% accuracy using only 3 features, which are cement, fly ash, and water. This means that it is possible to predict the compressive strength of concrete with a ratio above 90% by using a smaller number of concrete components.**

***Index Terms*— Artificial Bee Colony, Data Mining, Feature Selection, Prediction of concrete strength, 28-Day compressive strength.**

I. INTRODUCTION

Since concrete is a building material frequently used in construction, it is important to determine its strength. The accurate determination of the strength value has always been a critical issue, because concrete consists of a combination of aggregation, cement, water and additive materials. The reason for this is that concrete, which is a structural building material in different colors, surfaces and shapes, can easily be crushed, transported, placed, compacted and surface smoothed.

The concrete strength test, which is an empirical test, measures the consistency and workability of fresh concrete before it is adjusted. Consistency refers to the ease and homogeneity of concrete that can be mixed, placed, compressed and finished.

Basically, the most important feature of the concrete is the high strength. At the same time, it is expected that concrete will be equipped with features such as being economical, resistant to chemical deterioration and showing resistance to fire.

In this study, several data mining classification algorithms using the Artificial Bee Colony (ABC) algorithm as feature selection method was established predict the strength of concrete.

In this paper, the performance of ABC-based feature selection method proposed by Schiezaro and Pedrini [1] and updated by us is evaluated on Concrete Slump Test data set [2] from UCI Machine Learning Repository. The aim of this paper is to achieve high F-measure results by reducing the number of features used in the classification process using our updated ABC-based feature selection method. The classification accuracy rate and F-measure value are used and the F-Measure value is assessed for fitness value in this paper. The Accuracy Rate [3] and F-Measure [4] formulas are shown as in Equation 1 and 2, respectively. TP, TN, FP, FN mean True Positive, True Negative, False Positive, and False Negative, respectively.

978-1-5386-6878-8/18/$31.00 ©2018 IEEE

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$F - Measure = 2 * \frac{(\frac{TP}{TP + FP}) * (\frac{TP}{TP + TN})}{(\frac{TP}{TP + FP}) + (\frac{TP}{TP + TN})} \quad (2)$$

This paper is organized as follows: in the second section the material and method is given. The Material and Method section is divided into 4 categories including Data Set, data Pre-processing, Classification algorithms, Feature Selection, respectively. A brief explanation of Artificial Bee Colony algorithm and our updated version of ABC algorithm are given in this section. The third section includes some experimental evaluation results showing the effectiveness of our updated feature selector using ABC algorithm and these results are discussed. Finally, the last section concludes the study and gives some future work.

II. MATERIAL AND METHOD

In this section, the data used in this study, the data pre-processing phases, the used classification algorithms and the proposed feature selection method for this study are explained.

A. Data Set

The Concrete Slump Test Data Set from UCI Machine Learning Repository contains 103 observations and 10 variables including 7 concrete ingredient variables like Cement, Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, and Fine Aggregate, and 3 concrete property variables like Slump, Flow, and Compressive Strength [5,6]. Also each observation symbolizes a concrete mixture constitutes of different amounts of ingredients. The compressive strength variable planned to be predicted in this study indicates the 28-day compressive strength of each mixture of concrete, measured in mega pascals (MPa).

B. Data Pre-processing

Since the aim of the study is to predict the strength of the concrete, the other 2 output values slump and flows variables are removed from the data set. The data set is divided into 10 folds by 10-fold cross validation method. This method partitions the original sample into a training set a test set to train and evaluate the model.

C. Classification Algorithms

Four different classification algorithms, such as Random Forest, Sequential Minimal Optimization (SMO), K-Nearest Neighborhood (KNN), and Decision Table are used in the method to be able to evaluate performance and to get F-measure values as their fitness values on WEKA.

D. Feature Selection

Artificial Bee Colony algorithm is an optimization algorithm that inspired by intelligent foraging behavior of honey bees. In ABC model, there are three types of bee groups such as employed bees, onlooker bees and scout bees [6]. The probable solutions are represented by food sources and the food sources have a employed bee that assigned to them. Half of the colony is derived from employed bees and the other half includes onlooker bees. The number of each type of bee groups are equal. Also, number of food sources are equal to number of bees in each group. Employed bees explore food sources and their fitness quality represented by F-measure values. Onlooker bees get information about sources and exploit them. Scout bees produce new food sources to be replaced with exhausted ones.

The steps of ABC algorithm can be summarized as follows [7]:

1. Initially, food sources are explored using Equation 3 in order to find possible solutions.

$$X_{ij} = X_j^{min} + \text{rand}(0,1)(X_j^{max} - X_j^{min}) \quad (3)$$

where i=1,2,3,…,SN, j=1,2,3,…,D such that SN is the number of food sources and D is the number of optimization parameters.

2. After initialization, employed bees and start to explore the neighborhood of the food sources using Equation 4.

$$V_{ij} = X_{ij} + \Phi_{ij}(X_{ij} - X_{kj}) \quad (4)$$

where indices j is a random integer in the range [1,D] and k ϵ {1,2,3,…,SN} and k has to be different from i. For each food source X_i, new food source V_i is determined by changing one parameter of X_i. Φ_{ij} is a uniformly distributed real random number in range [-1,1].

3. After getting V_i, the fitness values re calculated for food sources using Equation 5.

$$fitness_i = \begin{cases} \frac{1}{1 + f_i} & \text{if } f_i \geq 0 \\ 1 + abs(f_i) & \text{if } f_i < 0 \end{cases} \quad (5)$$

where f_i is the cost function for the solution V_i.

4. Quality information of food sources are shared with onlooker bees. Onlookers evaluate the information obtained from employed bees and choose a food source has best selection probability. The probability is calculated by using Equation 6.

$$P_i = \frac{fitness_i}{\sum_{i=1}^{SN} fitness_i} \quad (6)$$

5. The onlookers become the employed bees. Neighborhood determination process for the selected food sources are carried out as explained in the steps between 2 and 4.

6. The counter which has been updated during search process is checked. If it is greater than maximum limit, then the food source is considered exhausted and abandoned. Abandoned source is replaced with randomly created food source that produced by the scout bee. [1]

An ABC-based feature selection method is created by updating and reiterating some steps of the pre-proposed method in "[1]". In the ABC-based feature selection method, each food source is represented as N-sized bit vector where *N* is the total number of feature. In food sources, if a bit's value is 1 that means the corresponding feature is selected in the feature subset created as a probable solution. If the value is 0, that means the corresponding feature is not a part of feature subset. Food sources store their quality and in the method, the quality is produced as F-measure value by classifiers. The classifiers produce fitness value by using created vector bits.

As it can be seen Fig. 1, the method starts by producing bit vectors. Only one bit of each vector is set as 1 at the beginning. Employed bees determine neighborhoods of the vectors and the neighbors are sent to the classifier to obtain F-measure as fitness values. Comparisons of fitness values are carried out and bit vectors with the best fitness are assigned to onlooker bees. Onlookers turn into employed bees to execute employed bee section. New N-sized bit vectors different than memorized ones are produced by scout bees. Steps are iterated until termination criteria are met.

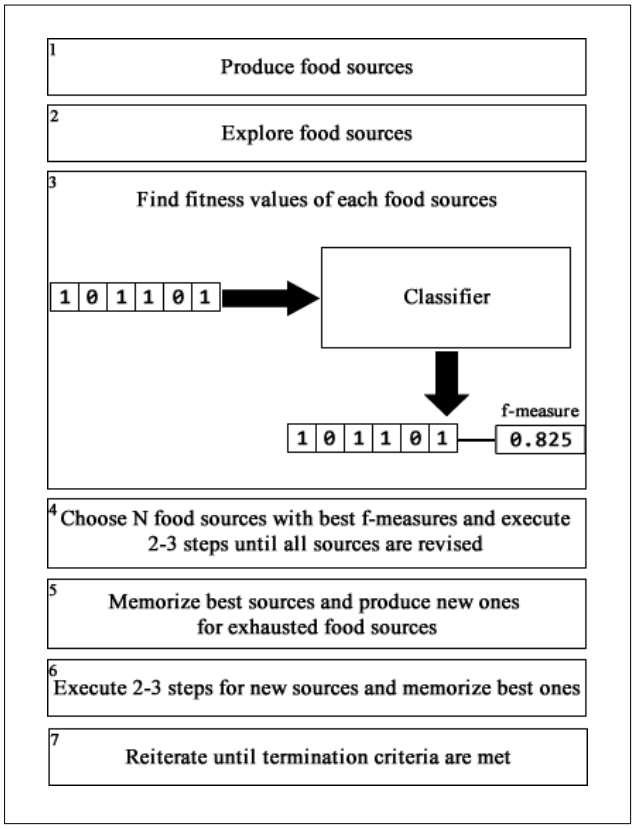

Fig. 1. The steps of ABC-based feature selection method

III. EXPERIMENTAL RESULTS AND DISCUSSION

Experiments have been carried out on a computer with Intel Core I5-6400 2.70 Ghz and 16 GB RAM. The updated ABC based feature selection method was implemented Java programming language. The classification process is carried out on Weka. The classification results by using all features and by using selected features is shown in Table I and Table II.

TABLE I. CLASSIFICATION ACCURACY AND F-MEASURE RESULTS WITH ALL FEATURES

Classifier	Accuracy	F-Measure
KNN	84.466 %	0.847
Random Forest	91.2621 %	0.913
SMO	87.3786 %	0.868
Decision Table	88.3495 %	0.888

From WEKA [8], KNN (IBk) from Lazy section, Random Forest from Trees section, SMO from Functions section and Decision Table from Rules section have been selected to evaluate the performance of the method. Classification accuracy and F-measure values of the raw dataset have been received. Random Forest achieved the best accuracy rate and F-measure values, 91.2621 % and 0.913 respectively, by using all of 10 attributes.

After applying feature selection, all results have been obtained from WEKA 3.8 software with the selected features. The feature selection method selected features with the four classification algorithms and these selected features were sent to the classifiers as input in WEKA software. In the updated pre-proposed method and results acquisition process, a 10-fold cross-validation is used and the modification rate (MR) is used as 0.3. The iteration number which is the termination criterion is set to 50 and k is selected as 3 for KNN classifier.

All features selected by the proposed method achieved better or equal results. As it can be seen in Table II, using Random Forest as classifiers while obtaining fitness values in the proposed method is produced 91.2621% and 0.913 which is best accuracy and F-measure performance, respectively. The result is same with the best result raw dataset has, but this result has been achieved using fewer features.

IV. RESULTS

This paper presents an updated pre-proposed feature selection method based on the Artificial Bee Colony (ABC) algorithm. For the raw dataset, the best classification accuracy rate and F-measure results were 91.2621% and 0.913, respectively. Fewer features have given the same result with the method.

For future work, hybridizing ABC based feature selection method with some other algorithms and focusing on working to find alternative ways to evaluate fitness value of food sources are planned.

TABLE II. CLASSIFICATION ACCURACY AND F-MEASURE RESULTS WITH SELECTED FEATURES USING DIFFERENT CLASSIFIERS

Classifier used in proposed method to obtain fitnesses	**Number of selected features**	**KNN**		**Random Forest**		**SMO**		**Decision Table**	
		Accuracy	**F-Measure**	**Accuracy**	**F-Measure**	**Accuracy**	**F-Measure**	**Accuracy**	**F-Measure**
KNN	4	89.3204 %	0.893	91.2621 %	0.913	88.3495 %	0.876	88.3495 %	0.888
Random Forest	3	89.3204 %	0.890	91.2621 %	0.913	87.3786 %	0.861	88.3495 %	0.888
SMO	4	89.3204 %	0.893	91.2621 %	0.913	88.3495 %	0.876	88.3495 %	0.888
Decision Table	2	85.4369 %	0.855	86.4078 %	0.865	87.3786 %	0.864	88.3495 %	0.888

ACKNOWLEDGMENT

This study was supported by Scientific Research Projects Commission Unit of Adana Science and Technology University under Grant Number: 18332001 and Grant Number: 18103004. Please send all your correspondence to aekeles@adanabtu.edu.tr which is the e-mail address of our Corresponding Author.

REFERENCES

[1] M. Schiezaro, H. Pedrini, “Data feature selection based on Artificial Bee Colony algorithm”, EURASIP Journal on Image and Video Processing, Vol. 1, pp. 47, 2013.

[2] I. C. Yeh, "Modeling slump flow of concrete using second-order regressions and artificial neural networks," Cement and Concrete Composites, Vol.29, No. 6, pp. 474-480, 2007.

[3] M. Shouman, T. Turner, “Using decision tree for diagnosing heart disease patients”, *Proceedings of the 9-th Australasian Data Mining Conference (AusDM’11)*, 01-02 December 2011, Ballarat, Australia, 2011.

[4] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation", Journal of Machine Learning Technologies, Vol. 2, No. 1, pp. 37–63, 2011.

[5] UCI Machine Learning Repository, 2018, Concrete Slump Test Data Set, https://archive.ics.uci.edu/ml/dataset/Concrete+Slump+Test, accessed: 10.05.2018

[6] D. Karaboga, “An idea based on honey bee swarm for numerical optimization”, Vol. 200. Technical report-tr06, Erciyes University, Engineering Faculty, Cmputer Engineering Department, 2005.

[7] D. Karaboga, B. Akay, “A modified artificial bee colony algorithm for real-parameter optimization”, Inf. Sci., Vol. 192, pp. 120-142, 2012.

[8] WEKA, https://cs.waikato.ac.nz/ml/weka/. Accessed: 14 June 2018.