Journal of AI and Data Mining
Vol 6, No 2, 2018, 409-415
DOI: 10.22044/JADM.2017.989

A New Hybrid model of Multi-layer Perceptron Artificial Neural Network and Genetic Algorithms in Web Design Management Based on CMS

M. Aghazadeh and F. Soleimanian Gharehchopogh*

Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran.

Received 15 September 2016; Revised 31 March 2017; Accepted 20 June 2017
Corresponding author: bonab.farhad@gmail.com (F. S. Gharehchopgh).

Abstract
The size and complexity of websites have grown significantly during the recent years. In line with this growth, the need to maintain most of the resources has been intensified. Content Management System (CMS) is software that has been presented in accordance with the increased demands of the users. With the advent of CMSs, factors such as domains, pre-designed module development, graphics, optimization, and alternative support have become the factors that influenced the cost of the software and web-based projects. Consecutively, these factors have challenged the previously introduced cost estimation models. This paper provides a hybrid method in order to estimate the cost of the websites designed by CMSs. The proposed method uses a combination of Genetic Algorithm (GA) and Multi-layer Perceptron (MLP). The results obtained are evaluated by comparing the number of correctly classified and incorrectly classified data, and Kappa coefficient, which represents the correlation coefficient between the sets. According to these results, the Kappa coefficient on testing dataset equals 0.82% for the proposed method, 0.06% for GA, and 0.54% for MLP Artificial Neural Network (ANN). Based upon these results, it can be said that the proposed method can be used as a considered method in order to estimate the cost of websites designed by CMSs.

Keywords: *Genetic Algorithm, Multi-Layer Perceptron Artificial Neural Network, Website Cost Estimation, Content Management System.*

1. Introduction

Some researchers have stated that the number of web pages on big websites double annually [1, 2]. However, in order to develop websites, it is required to have websites of different areas that have business-to-business relations, multi-language websites, and intranets, which integrate suppliers and business partners [3]. The number of people who increases different contents is increasing daily. These content helpers might have different writing and reception styles that demand the personal content and enhanced performance of the websites. Different media also need to manage both text and images [4, 5]. These factors make a website content management a compulsive priority.

The phrase content management of a constant source is ambiguous [6]. There are various definitions among different individuals. Some consider it as a platform that can be purchased, and the others see it as a set of approved methods or a new way to create business [7]. Since there are different definitions for content management in different areas, it has no precise and solid definition. Generally, a content management system (CMS) software is installed on a web server, and the users with different access rights can enter their own domain. According to their access rights, every user can generate, edit, and publish the content as well as information. This function is known as a management role, and it is common for almost all CMSs [8].

This paper has been organized as follows: Section 2: Related works; Section 3: Basic concepts; Section 4: Proposed method; Section 5: Evaluation criteria and data analysis; Section 6: Conclusions and future works.

2. Related works

In the recent years, many methods have been proposed for website management, their design,

and production cost estimation. For example, by proposing the WEBMO model, which is a developed COCOMO II and one of the algorithmic methods, the researchers [9] have tried to estimate the effort rate for web projects. Likewise, this model calculates the cost of web projects: first by using estimators like web objects and then complexity coefficient table for object criteria, system operators, and operands calculation process along with selecting criteria complexity in the areas of e-commerce, financial and commercial applications, portals, and information services. WEBMO differs from COCOMO II by having nine instead of 7 cost drivers and variables instead of fixed capacity. This model eventually uses Pred (N) standard to evaluate the proposed model. E. Mendes et al. [10] affiliated their measurements and requirements to the number of Use Cases, total number of entities in entity-relationships, total number of pages in entity-relationship, total number of nodes in a navigation graph, number of anchors in navigation graph, and effort in person hours for designing web pages. Then by counting the number of HTML pages, the number of media files, the total number of phrases used in JavaScript code, and the cascading style sheets, the total number of internal and external links on each page, the number of media differences on each page using Case-based technique and using data extracted from 25 databases, they have tried to estimate the cost and effort of new projects. They have used the MMRE, MdMRE, and Pred (25) evaluation criteria to assess the proposed model. In another research [11], the COBRA model has been presented to estimate the cost and effort of web projects using the data from small companies. The COBRA model is a method that aims to develop understandable cost estimation of a particular company. It uses expert views and the data used in previous projects to estimate the cost and effort of new projects. Using this model, they have tested and evaluated the web objects presented by Reifer in 2000, which were completed in 12 projects, and then they proposed a model using expert views and linear regression estimation, which has been evaluated using the MMRE and Pred (25) standards. CWADEE is also another quick estimation method developed by a group of leading experts of software engineering in the University of Chile [12].

Other researchers [13] have also developed a model to estimate the cost of CMSs. They have claimed that their proposed model can be used to estimate the cost and effort of development and design based on these systems. Furthermore, they have pointed out that the data from other content management projects has been studied. Finally, bagging predictor has been used in the linear regression model. The size of this project has been evaluated with point method by means of object modification. In order to find out different objects, their classification, and their complexity in the project, a questionnaire has been used. In order to help the project managers, a final effort has been estimated using the project size and other factors involved. For a better rate of performance of the system and the model, production characteristics, overall system characteristics, experience, and ability of the developer have been used. This model was the result of assessment and evaluation of 12 web-based projects.

In [14], a model has been proposed to estimate the cost and effort of web-based object-oriented applications. This method is a combination of the leading scholars' and researchers' theories in web projects cost estimation. This research work, which is based upon case-based reasoning, selects three similar projects from finished web-based projects, and estimates the cost and effort of the current object-oriented web-based project in accordance with the cost and effort of the finished projects. The results from the previous works have been used to assess the method.

The Naïve Bayesian algorithm has also been proposed to estimate the cost of designing content management system websites [15]. Assessments have been carried out on 99 web projects. The proposed method had a classification accuracy of 55%, and the incorrect classification accuracy was 45%. The results obtained show that the proposed method is an efficient model to estimate the cost of websites.

The hybrid of K-Nearest Neighbor (KNN) models and MLP (ANN) to estimate the cost of web project that follow CMS has been proposed [16]. Evaluation was conducted on 99 web project. The results obtained show that the accuracy precision in the hybrid model equals 0.95, and the kappa coefficient is 0.93; if compared with the MLP and KNN models, it looks much better.

The FRBFN model [17] which is a hybrid of fuzzy logic and ANN is proposed to estimate web. Evaluation was conducted on 53 web projects with 9 features. Fuzzy C-Means is used for project clustering. The results obtained show that the FRBFN model, according to the number of various clusters, has different MMRE and PRED. With increase in the number of clusters, the MMRE value decreases, and the PRED value increases.

3. Basic Concept

3.1. Effective Factors

After identifying the main requirements, a website cost is estimated based on the effective factors. The effective factors used in this paper that estimate the cost of websites are shown in figure 1.

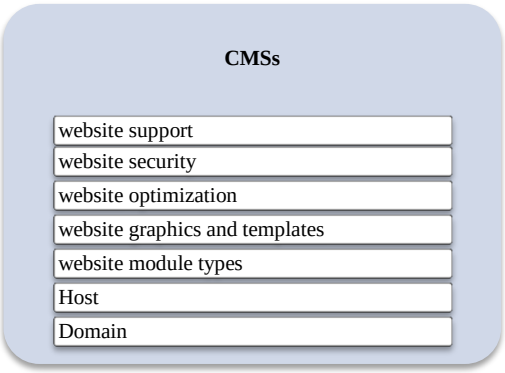

Figure 1. Factors affecting final cost of CMSs.

3.2. Genetic algorithm

Genetic algorithm (GA) is inspired by genetics and Darwin's evolutionary theory, and is based upon natural selection and survival of the fittest. GA was proposed by Holland in 1970 [18]. It is a stochastic optimization algorithm, which is suitable for complex problems with unknown search space. Like other evolutionary algorithms, GA is population-based, and has its specific parameters. The way of determining these parameters affects the performance of this algorithm. These parameters include: 1. Population 2. Fitness function 3. Chromosome representation 4. Algorithm Operators such as cross-over, mutation, and selection.

3.3. MLP (ANN)

MLP is based upon a computational unit named perceptron. A perceptron gets a vector of inputs with real values, and calculates a linear combination of the inputs. If the result is bigger than the threshold, the perceptron output is 1; otherwise, it is -1. MLP ANNs are among the most practical ANNs. These networks are capable of do a non-linear mapping with arbitrary precision by appropriately selecting the number of layers and neurons, which are mostly not large [19].

4. Proposed method

In this paper, a hybrid of GA and MLP (ANN) was used to estimate the cost of web-based software projects. In this model, the design cost of websites based on CMSs is estimated. Figure 2 shows the process of the proposed method.

In the proposed method, initially, the dataset was pre-processed, and noisy data was removed. Subsequently, the data was randomly divided into 80% training and 20% testing data. After dividing the data, the training data was used for the training phase of the proposed method. GA is used to find the most optimal parameters for MLP ANN.

After determining the testing and training datasets, MLP ANN was run. The MLP ANN used here consists of three layers:

The first layer includes the inputs that consist of 6 neurons.

The second layer acts as a hidden layer, which consists of 10 neurons, and uses the sigmoid transfer function as the activation function. Considering that the number of neurons in the hidden layer has a significant impact on the performance of the algorithm, therefore, in the proposed method based on trial and error and the iteration of the algorithm with different numbers of neurons, it can be concluded that using 10 neurons results in a better performance.

The third layer is the output layer, and has 7 neurons. In the MLP ANN, GA is used in the second layer to update the information. In each iteration, after applying the activation function, it updates the weights based on defined chromosomes using cross-over and mutation operators during consecutive iterations. In the next step, the updated weights are re-injected into the MLP ANN, and the iterations continue as long as the desired error rate has not been reached.

In the proposed method, the data whose cost has not been properly defined is used as fitness function which, in each stage, tries to reduce it.

In this paper, for evaluation of the fitness function, the accuracy criterion is used according to (1). In (1), TN represents the number of samples whose true category are negative, and the classification algorithm also correctly specifies their category as negative. TP represents the number of samples whose true category is positive, and the classification algorithm also correctly specifies their category as positive. FP represents the number of samples whose true category are positive, and the classification algorithm also correctly specifies their category as positive whose true category are positive, and the classification algorithm handles them mistakenly as negative.

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

After the hybrid algorithm work is completed, the results obtained from the training phase are injected to the data in the testing dataset, and the

testing phase is completed as well. After the training and testing stages, the results obtained are evaluated based on the Kappa method and counting the number of right and wrong answers, and the results obtained are displayed as graphs and tables.

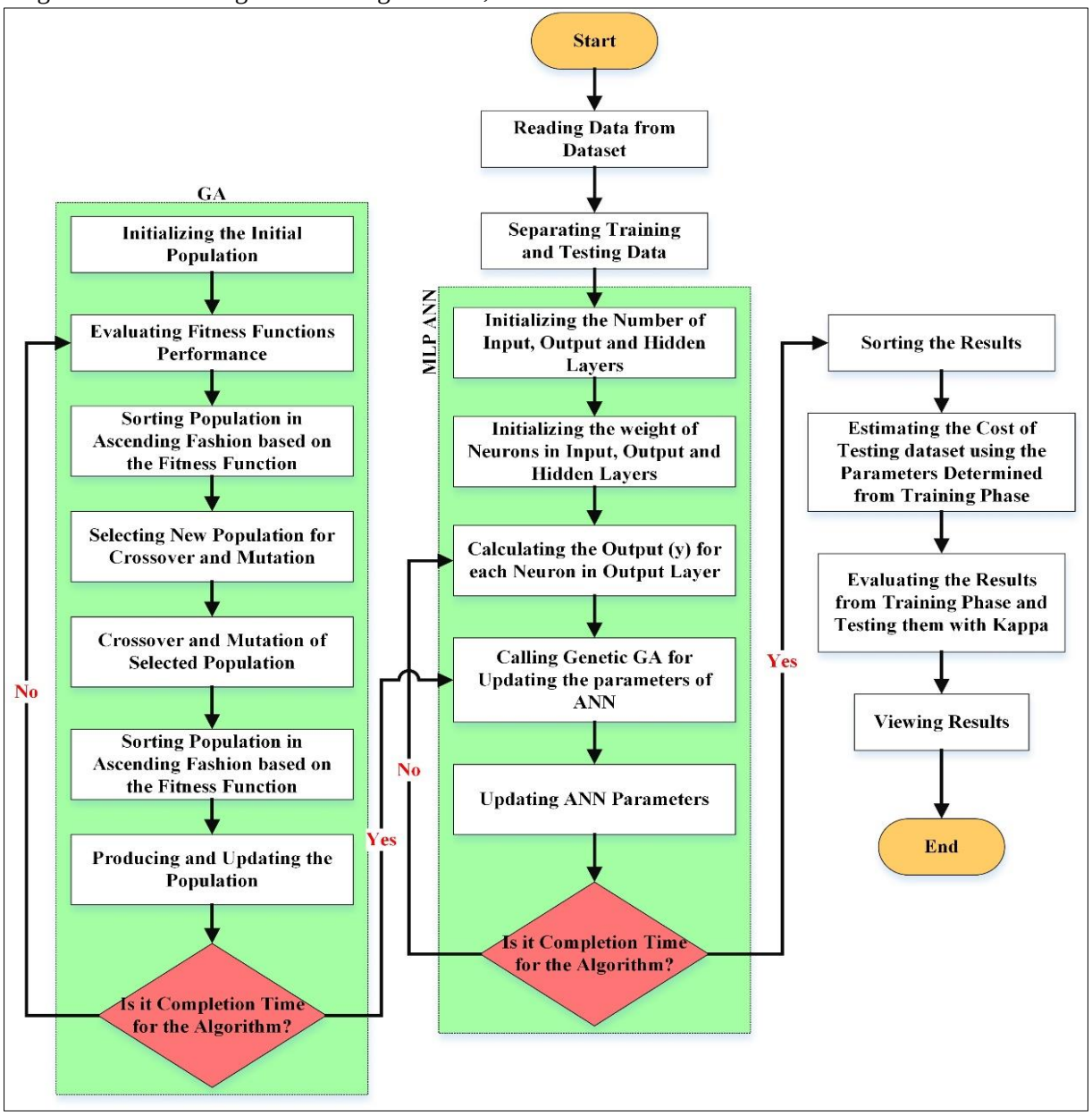

Figure 2. Process of proposed method.

5. Evaluation and results

In this paper, the Kappa coefficient and the number of correct and incorrect classifications were used to determine the accuracy of the proposed method. The dataset provided by Khaze [15] was used as the dataset. The dataset includes 99 web-based projects with 7 essential features created to specify the cost of designing the websites.

The Kappa coefficient is used to evaluate the reliability of the nominal and classifiable data. The Kappa method was proposed, for the first time, by Cohen [20]. Equation (2) is used to calculate the error coefficient.

$$kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (2)$$

In (2), Pr(a) is the relative observed agreement among the sets, and Pr(e) is the hypothetical probability of the agreement. The results of the evaluation of the proposed algorithm are shown in tables 1 and 2 and figures 2 and 3. Table 1 shows the values obtained to determine the Kappa coefficient for testing the dataset. Since the costs of the websites in the dataset are divided into six

types, the number of sets is considered as six as well.

Table 1. Values obtained for Kappa coefficient by proposed method on testing dataset.

	Type1	Type2	Type3	Type4	Type5	Type6
Type1	2	0	0	0	0	0
Type2	0	0	0	0	0	0
Type3	0	0	2	1	0	0
Type4	0	0	2	3	0	0
Type5	0	0	0	0	2	2
Type6	0	0	0	0	0	6

Table 2 shows the values obtained for Kappa coefficient, the number of correctly and incorrectly classified data for training and testing datasets. In this table, the Kappa coefficient obtained reflects the appropriate performance of the proposed method.

Table 2. Results of evaluation of proposed method.

Materials	Training dataset	Testing dataset
Number of correctly classified data	73	15
Number of incorrectly classified data	7	4
Kappa coefficient	0.95	0.82

Figure 3 reflects the results of comparing the proposed method with GA and MLP (ANN).

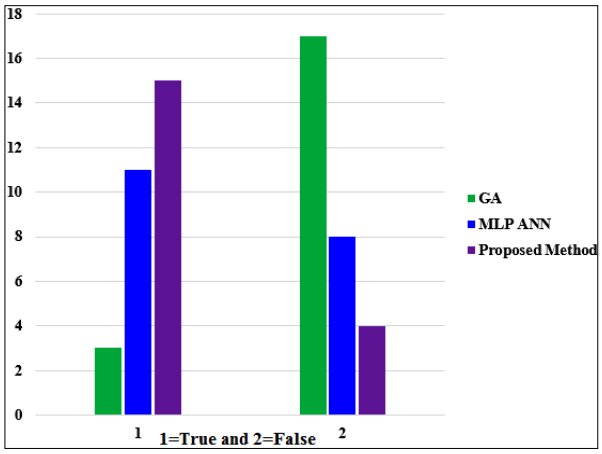

Figure 3. Comparison of proposed method with other two algorithms over testing dataset.

The comparison is based upon the number of correctly and incorrectly classified data over the testing dataset. Based on this figure, the proposed algorithm has a better performance than the other two algorithms.

Figure 4 shows the results of comparing the proposed method with GA and MLP ANN. The comparison is based upon the number of correctly and incorrectly classified data over the training dataset. Based on this figure, the proposed algorithm has a better performance than the other two algorithms.

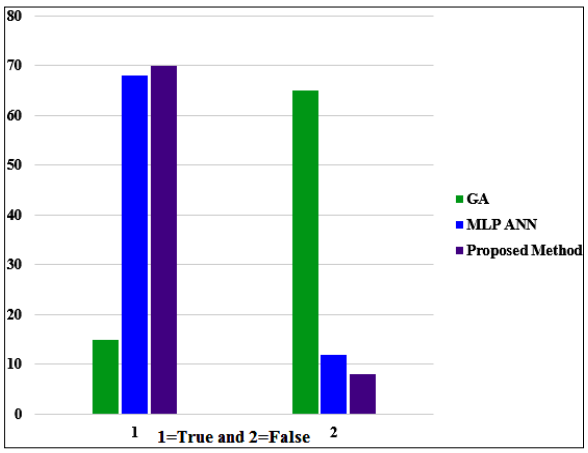

Figure 4. Comparison of proposed method with other two algorithms over training dataset.

The results of comparing the proposed method with GA and MLP ANN are shown in table 3. This table shows the number of correctly classified data, incorrectly classified data, and Kappa coefficient for the testing dataset. The Kappa coefficient for the proposed method equals 0.82%, which is an indication of an appropriate performance of proposed method.

Table 3. Results of proposed method for testing dataset.

Method	Number of correctly classified data	Number of incorrectly classified data	Kappa coefficient
GA	4	15	0.06
MLP ANN	11	8	0.54
Proposed method	15	4	0.82

Table 4 shows the results of the proposed model with 200 times of iteration of GA. Since with more iteration, diversity in GA increases, as a result, achieving an optimal solution is more probable, cross-over and mutation operators have a better chance of finding the right samples. Therefore, GA with 200 cycles of iterations in creating the optimal solution and escaping from local minimum serves best.

Table 4. Results of proposed model with 200 times GA iteration.

Materials	Training Dataset	Testing Dataset
Number of correctly classified data	76	17
Number of incorrectly classified data	4	2
Kappa coefficient	0.97	0.85

Table 5 shows the results of the proposed model with 5 layers in MLP. As shown in this table, diagnose accuracy is increased. Since MLP with increase in the number of layers can better train

and test the data, and the iteration of layers reduces the error and detection of the best weight for features. If the amount of weight is to be given to a precise characteristic, then the output error reduces, and classification accuracy increases.

Table 5. Results of proposed model with 5 layers in MLP.

Materials	training dataset	Testing dataset
Number of correctly classified data	77	16
Number of incorrectly classified data	3	3
Kappa coefficient	0.97	0.86

The present paper proposes a hybrid model of GA and MLP (ANN) to achieve a higher accuracy and lower rate of error in cost estimation.

7. Conclusion and future works

With increase expansion of virtual communications, having a website is a necessity. Given the importance of cost estimation of the websites and the lack of a one hundred percent accurate method, it was tried in this work to propose a new method based on machine-learning algorithms to estimate the cost of the websites provided by content management systems. In this method, a combination of genetic algorithm and MLP ANN is used. The results obtained from the proposed method were evaluated according to the Kappa coefficient and the number of correctly and incorrectly classified data. According to the results obtained, the Kappa coefficient for the proposed method was equal to 0.82%, while it was 0.06% for genetic algorithm and 0.54 percent for MLP ANN. Based upon these results, it can be said that the proposed method has an appropriate performance in estimating the cost of websites provided by content management systems.

References

[1] Martino, S. D., Ferrucci, F., Gravino, C., & Sarro, F. (2016). Web Effort Estimation: Function Point Analysis vs. COSMIC, Information and Software Technology, vol. 72, pp. 90-109.

[2] Gharehchopogh, F. S. (2011). Neural Networks Application in Software Cost Estimation: A Case Study, 2011 International Symposium on Innovations in Intelligent Systems and Applications, pp. 69-73, IEEE, Istanbul, Turkey, 15-18 June 2011.

[3] Ghatasheh, N., Faris, H., Aljarah, I. & Iand Al-Sayyed, R. (2015). Optimizing Software Effort Estimation Models Using Firefly Algorithm. Journal of Software Engineering and Applications, vol. 8, pp.133-142.

[4] Friedlein, A. (2003). Maintaining and Evolving Successful Commercial Web Sites: Managing Change, Content, Customer Relationships, and Site Measurement, Morgan Kaufmann.

[5] Ceke, D. & Milasinovic, B. (2015). Early effort estimation in web application development, Journal of Systems and Software, vol. 103, pp. 219-237.

[6] Clark, D. (2007). Content Management and the Separation of Presentation and Content, Technical Communication Quarterly, vol. 17, no.1, pp. 35-60.

[7] Marvi, H., Esmaileyan, Z. & Harimi, A. (2013). Estimation of LPC coefficients using Evolutionary Algorithms, Journal of Artificial Intelligence & Data Mining, vol. 1, no. 2, pp.111-118.

[8]Han, Y. (2004). Digital Content Management: The Search for a Content Management System, Library Hi Tech, vol. 22, no. 4, pp. 355-365.

[9] Reifer, D. J. (2004). Web Development: Estimating Quick-To-Market Software, Software IEEE, vol. 17, no. 6, pp. 57-64.

[10] Mendes, E., Mosley, & Watson, N. I. (2002). A Comparison of Case-Based Reasoning Approaches, International Conference on World Wide Web, pp. 272-280, 2004.

[11] Ruhe, M., Jeffery, R. & Wieczorek, I. (2003). Cost Estimation for Web Applications; 25th International Conference in Software Engineering, pp. 285-294, 2003.

[12] Sergio, F., Ochoa, M., Cecilia, B. & German, P. (2003). Estimating the Development Effort of Web Projects in Chile, LA-WEB '03 Proceedings of the First Conference on Latin American Web Congress, Page 114, IEEE Computer Society Washington, DC, USA, 2003.

[13] Aggarwal, N., Prakash, N. & Sofat, S. (2010). Content Management System Effort Estimation Using Bagging Predictors, Technological Developments in Education and Automation, pp 19-24, 2010.

[14] Suharjito, R. (2012). FHS Web EE: An Effort Estimation Model for Web Application, International Conference on Advances Science and Contemporary Engineering, Procedia Engineering 50, pp. 613-622, 2012.

[15] Khaze, S. R., Ghaffari, A. & Masdari, M. (2013). Using the Naïve Bayes Algorithm for Web Design Cost Estimation with Content Management System, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 11, pp. 999-1007.

[16] Jahatloo, L. E. & Jafarian, A. (2015). A New Approach with Hybrid of Artificial Neural Network and K-Nearest Neighbor Algorithms in Cost Estimation of CMS based Web Sites Designing, Journal of Scientific Research and Development, vol. 2, no. 7, pp. 134-140.

[17] Idri, A., Zakrani, A., Elkoutbi, M. & Abran, A. (2007). Fuzzy Radial Basis Function Neural Networks

for Web Applications Cost Estimation, Innovations in Information Technologies, pp. 576-580.

[18] Holland, J. H. (1975). Adaptation in Natural and Artificial Systems, ANN Arbor: The University of Michigan Press.

[19] Samani, B. H., Jafari, H. H. & Zareiforoush, H., (2017). Artificial Neural Networks, Genetic Algorithm and Response Surface Methods: the Energy Consumption of Food and Beverage Industries in Iran, Journal of Artificial Intelligence & Data Mining, vol. 5, no.1, pp. 79-88.

[20] Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, vol. 20, no. 1, pp. 37-46.

نشریه هوش مصنوعی و داده کاوی

ارائه مدل ترکیبی جدید با استفاده از الگوریتمهای شبکه عصبی مصنوعی پرسپترون چند لایه و ژنتیک در مدیریت طراحی وب سایت‌های CMS

مریم آقازاده و فرهاد سلیمانیان قره چپق*

گروه مهندسی کامپیوتر، واحد ارومیه ،دانشگاه آزاد اسلامی ، ارومیه، ایران.

ارسال ۲۰۱۶/۰۹/۱۵؛ بازنگری ۲۰۱۷/۰۳/۳۱؛ پذیرش ۲۰۱۷/۰۶/۲۰

چکیده:

اندازه و پیچیدگی وب سایت‌ها در طول سالیان اخیر رشد قابل توجهی پیدا نموده است و در راستای آن نیاز به حفظ اکثر منابع تشدید شده است. سیستم‌های مدیریت محتوا، نرم‌افزارهایی هستند که با افزایش تقاضاهای کاربران ارائه شدند. با ظهور سیستم‌های مدیریت محتوا عواملی همچون دامنه‌ها، توسعه ماژول‌های از پیش طراحی شده،گرافیک، بهینه‌سازی و پشتیبانی جایگزین عوامل موثر بر هزینه پروژه‌های نرم‌افزاری و پروژه‌های تحت‌وب سنتی شده است و مدل‌های پیش‌بینی هزینه‌هایی که قبلا معرفی شده بود را به چالش کشانیده است. در این مقاله به ارائه روش ترکیبی برای تخمین هزینه وب سایت‌های طراحی شده توسط سیستم‌های مدیریت محتوا پرداخته شده است. در روش پیشنهادی از ترکیب دو الگوریتم ژنتیک و شبکه عصبی مصنوعی پرسپترون چند لایه استفاده شده است. و نتایج حاصله از روش پیشنهادی براساس ضریب کاپا و تعداد داده‌های درست و اشتباه دسته‌بندی شده مورد مقایسه و ارزیابی قرار گرفته است. براساس نتایج بدست آمده مقدار ضریب کاپا در مجموعه‌داده آزمون برای روش پیشنهادی برابر با ۰,۸۲ درصد، برای الگوریتم ژنتیک برابر با ۰,۰۶ درصد و برای شبکه عصبی مصنوعی پرسپترون چند لایه برابر با ۰,۵۴ درصد می‌باشد. براساس این نتایج می‌توان گفت که روش پیشنهادی می‌تواند به‌عنوان یکی از روش‌های مطرح در تخمین هزینه وب سایت‌های طراحی شده توسط سیستم‌های مدیریت محتوا مطرح گردد.

کلمات کلیدی: الگوریتم ژنتیک، شبکه عصبی پرسپترون چند لایه، تخمین هزینه وب‌سایت‌ها، سیستم‌های مدیریت محتوا.