MÜHENDİSLİK FAKÜLTESİ

Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi
International Journal of Engineering Research and Development

UMAGD, (2020) 12(3), s65-s72.

10.29137/umagd.827683

Cilt/Volume:12 Sayı/Issue:3 Aralık/December 2020 Özel Sayı / Special Issue

Araştırma Makalesi / Research Article

Storage Requirement Estimation for Electronic Document Management System With Artificial Neural Networks

Zeynep Çetinkaya[1], Erdal Erdal[*1], Atilla Ergüzen [1]

[1]Engineering Faculty, Computer Engineering Department, Kırıkkale University, Kırıkkale, Turkey

***Başvuru/Received**: 18/10/2020* ***Kabul / Accepted**: 05/12/2020* ***Çevrimiçi Basım / Published Online:** 31/12/2020*

***Son Versiyon/Final Version**: 31/12/2020*

Abstract

Electronic document management systems are defined as the protection and management of the contents, formats and relational features of all kinds of documents created by an institution in the process of carrying out its activities. Storage areas are one of the important elements for electronic document management systems. With every transaction and activity transferred to electronic environment in institutions, the infrastructure and investments that should be allocated for Electronic document management systems storage areas increase and the forecast of this increase becomes more important over time. Artificial neural networks (ANN) approach has been used in many areas in recent years. Estimation studies in different fields have been made with ANN and it has been observed that successful results have been obtained. In this study, an ANN model is proposed to be used in estimating the storage area required for electronic document management systems. In this study using Kırıkkale University Electronic document management systems data, different ANN models were created, the most suitable models were determined, and the required storage area was estimated for the future periods.

Key Words

"Artificial Neural Networks, Electronic document management systems, Demand Forecast"

***Corresponding Author:** erdalerdal@kku.edu.tr*

1. Introduction

Records are the recorded information that is created to perform any individual or corporate business or constitutes evidence produced as a result of the work performed (Çiçek, 2009). Document is defined as records with a legally validated form that shows an action taken by the document manufacturer. Documents called corporate information are used to provide the necessary information for communication and decision making. Documentation is defined as a by-product carrying business activities and should be kept as long as it remains evidence of the business process (Dollar, 2002).

Within the scope of document management studies, the concept of document has a legal quality. The document includes all kinds of written and printed materials that are evident and are not doubtful before the law. Physical format of documents, storage environment, production date, or the conditions under which it was produced are not decisive (Çiçek, 2009; Dollar, 2002).

Documentation is to transform the information into text fixed on paper about the work done or planned to be made (Moses, 2005). This definition has undergone a natural change conceptually and structure in parallel with the developments in information technologies. Computer technology enables data sets that do not belong to any documentary format to be used in the business process and to be proof (Odabaş, 2008).

The fact that the new environments are much more dynamic, updatable, conducive to new data entries and offering more sharing opportunities have led to the change of the concept of document format. It is also foreseen that technological developments will continue to change this concept constantly (Odabaş, 2008; Shepherd and Geoffrey, 2003).

Document management is a system implemented to ensure corporate communication, to make the right decisions of the management, to examine the retrospective applications and to provide institutional control easily. The document management system has a life cycle that includes the document production, editing, distribution, storage, archiving of the document for certain situations, and destruction when the document loses its importance (Külcü, 2007).

Electronic document management system; is defined as the extraction of documents that may be evidence of corporate activities from all kinds of documentation created by institutions while performing their activities, to protect their content, format and relational features and to manage these documents in the process from production to destruction.

Electronic document management system, by ensuring that both internal and external correspondence is made electronic-based, correspondence in accordance with the standard file plan,shortening the delivery, initials and approval periods, reducing costs (labor, time and stationery),the establishment of the corporate electronic archive is targeted. With this system: delivering documents to the right person at the right time in the shortest way, tracking the status of documents during circulation, filing documents quickly and archiving electronically, instant access to documents with various search criteria according to their level of authority, with a single click, simultaneous access to the same information and document from multiple points is provided.

In a physical environment, performing the operations of a document during the life cycle causes serious costs. As a matter of fact, it has been found in researches that keeping a document, finding a wrongly filed document, re-preparing a lost document is equal to 20-45% of the total costs of document-related transactions (Delmar, Davidsson and Gartner, 2003). In such a situation, it has become inevitable for corporate documents to be produced, used, transmitted and stored electronically.

With the use of electronic document management system, savings are achieved in the following expenses:

- Paper savings will increase as the number of required outputs decreases in document preparation, approval, distribution and use.
- Courier and postage costs are reduced.
- Document filing, archiving and storage expenses are minimized.
- The cost of protection measures to be taken for increased security of documents and the cost of resources used to maintain the destroyed documents is reduced.
- With the decrease in current affairs, the number of personnel is reduced (Önaçan, Medeni and Özkanlı, 2012).

The electronic document management system actually triggered the formation of a paradox with facilitated the creation, processing, storage and sharing of a document. This paradox has made it necessary to determine how the documents created in the past years will be arranged, defined and put into use, as well as determining the methods to be followed in order to determine the potential amount of documents to be produced in the future (Çiçek, 2009).

ANN are defined as a technique that can be used successfully in the resolution of nonlinear problems and provide extremely dependable outcomes in line with the results obtained in the studies conducted (Öztemel, 2006). ANN are recognized as a technique established by demonstrating of the knowledge procedure, that is among the features of the human brain in electronic environment. This method is magnificently used in numerous topics such as creating estimation, new evidence, multiplexing, clustering, classification, association and generalization (Fausett, 1994; Zhang, Patuwo and Hu, 1998).

It will be meaningful to use ANN approach in estimating the number of potential documents to be produced in the future for the life cycle of documents, archiving processes and long-term retention periods (Çiçek, 2009). The structure of a simple cell created in artificial neural networks is as shown in Figure 1. This simple structure; consists of inputs, weight values, outputs, total function and activation function.

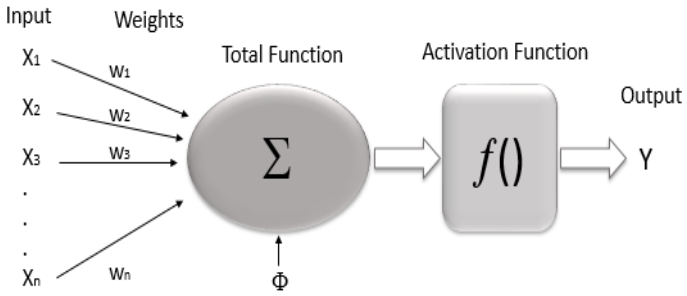

Figure 1. Artificial Neural Network Cell

Input is precise as data arrives the cell from other cells or external environments. These are strongminded by the instances that the system wants to learn. Weights are the defined values indicate the outcome of the cell in the input or previous layer on the next process. The inputs are combined with the weight value through the total function and connected to the next cell.

The total method presents the net input value to the cell. Dissimilar total method can be preferred at this stage. Weighted sum function, multiplication function, maximum or minimum functions and cumulative functions are examples of total functions. The most commonly used sum function is the weighted sum. Each incoming input value is multiplied by its own weight. Thus, net input to the cell is found. The weighted sum function is as given in Equation 1.

$$net = \sum_{i=1}^{n} WiXi + b \qquad (1)$$

Values gained because of the total method are passed through a nonlinear or linear differentiable method. This function is called the activation function. The net output of the cell is obtained as a result of the activation function. General equation representation is as in Equation 2. For activation functions, sigmoid function, step function, sinus function and hyperbolic tangent function can be given as examples.

$$y = f(net) = f(\sum_{i=1}^{n} WiXi + b) \qquad (2)$$

In ANN, various network representations have been created with the connections of layer elements, numerous aggregation and activation functions used, different learning strategies and different topologies resulting from learning rules. Perceptron, Adaline, Multilayer perceptron, Learning vector quantization, Hopfield, Self-organizing map, Cognitron, Adaptive resonance theory, Probability neural network are just a few of these ANN models created. Some of these artificial neural network models are used for classification operations (LVQ, ART, PNN, Counterpropagation), some others are customized for data association testing (Hopfield, Boltzmann). The MLP model is the most broadly selected model for estimation among these network constructions (Öztemel, 2006; Basheer and Hajmeer, 2000).

In multi-layer perceptron (MLP), cells - neurons are organized in stages. In this graded structure, there are intermediate layers in addition to the input and output layers. These layers between two layers are called hidden layers. MLP networks can be designed to contain one or more hidden layers. Figure 2 shows the general structure of the multi-layer perceptron.

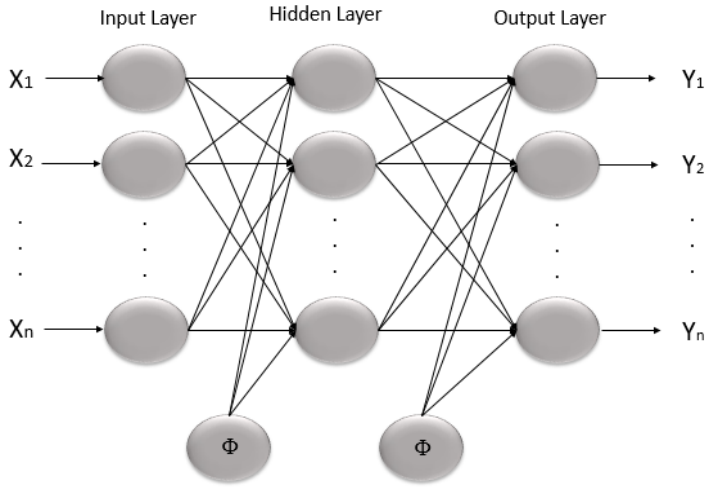

Figure 2. Multilayer Perceptron Model

Since it is understandable and precisely provable, back-propagation process is preferred in the training of MLP systems. This procedure is named back-propagation for it attempts to decrease errors by transmitting errors from output to input. The back-propagation procedure has a consulting learning (learning with teacher) structure and is the mostly preferred learning algorithm used in numerous developments (Fausett, 1994)

In teacher knowledge procedures, through the training of the net, a trial dataset consisting output values conforming to input and input values is assumed to the network. In consulting learning algorithms, weights in the learning phase are planned by minimizing the error function given in Equation 3.

$$E = \frac{1}{2} \sum_{k=1}^{m} (yk - tk)^2 \qquad (3)$$

In equivalence, yk shows the output manufactured by system, and tk shows the definite output value. In order to minimalize the error value of the net, weights between connections are rearranged and updated. Therefore, the system is envisioned to propagate the output nearby to the real output values (Zhang, Patuwo and Hu, 1998).

Technically, the elementary job of an ANN is to study the assembly in the trial dataset and oversimplify to perform the chosen job (Fausett, 1994). ANN has been extensively preferred in practically every field for the purpose of prediction in last years. Especially in non-linear time series, ANN's superior success compared to classical methods has been an significant issue in choosing this technique (Lewis, 1982).

EL-Saba and El-Amin (1999) have estimated long-term energy sales, total energy and peak load demand using artificial neural networks. This study shows that the information required for managers to make effective resource planning decisions can be provided by ANN and can be used for long-term forecasts with minimal error.

In Liu et al's (2003) study, an artificial neural network (ANN) technique was proposed as an effective way to model water demand in urban areas and formulate domestic water demand. Considering the data obtained from the study, it is seen that the correlation coefficients of the water demand prediction model using the artificial neural network are more than 90% for both training data and test data.

In a study by Çetinkaya and Erdal in 2019, an ANN-based model was proposed that predicts the daily food demand for students and staff at the Kırıkkale University cafeteria. According to the results of the study, it was seen that the model has low error rate and high performance.

In the research of Asılkan and Irmak (2009), it was tried to predict the future prices of used vehicles with ANN. To see the prediction accuracy of ANN, the results are compared with time series analysis methods. The fact that ANN gives more successful results in all data sets has shown that this method can be an alternative to classical statistical methods. Studies in different fields have shown that it is possible to use artificial neural networks for demand estimation.

2. Metarial and Metod

Kırıkkale University is an institution where information security is at the forefront, where speed of access to information and information reliability are important. For this reason, it required an advanced document management system and EBYS was launched as a solution in January 2015. With EBYS, it was ensured that the documents were collected in a central structure, their processes were accelerated, they were followed up, reported and managed. In addition, by creating data standards, all necessary data has been made compliant with these standards and archived in a secure way, enabling this information to be accessed quickly and effectively.

In-institution and out-of-institution documents issued at the university, documents coming to the university and all reports that have the characteristics of documents have started to be prepared through this system. This situation brings with it the need for a serious storage area as well as the benefits it provides.

The purpose of this study is to estimate the storage area that will be needed for the electronic document management system used in Kırıkkale University, depending on the number of documents to be produced in the future, using artificial neural networks.

In the study, the factors that will affect the need for storage space were determined first. The first of the effective factors is school periods. In the Fall and Spring school periods, it is seen that the documents produced are higher compared to the Summer period, thus the need for storage increases. Other factors determined are as listed below.

- Number of Academic Staff
- Number of Administrative Staff
- Number of students
- Documents issued for in institution
- Documents issued for out of institution
- Documents come from out of institution

The data gained as a result of determining the effective features are adapted to the assembly of the system so that they can be relocated as input to the ANN model. Data should be standardized to avoid needless data duplication and limit the data to a specific band.

Different methods are used in the literature for the normalization process. These can be list as: Min - Max rule, Median rule, Sigmoid normalization and Z-Score normalization (Jayalakshmi and Santhakumaran, 2011). One of these methods is the min-max method, which is the normalization method that finds the most common usage. The general formula of the min-max method used in the normalization process is given in Equation 4.

$$V' = \frac{V - V\_min}{V - \max - V\_min}\left(\max_{taget} - \min_{target}\right) + \min_{target} \qquad (4)$$

Here v: real value, v': normalized value, v_max: maximum input value, v_min: minimum input value, max_target: the upper limit of value, min_target: the lower limit of value.

In the min-max method, data entries are normalized within a predefined range. These range values are usually applied as [0, 1]. In this study, the input values to be normalized are preprocessed in the range value [0, 1]. The min-max method equation for the range values [0, 1] is as shown in equation 5 (Shalabi and Shaaban, 2006; Sola and Sevilla, 1997).

$$V' = \frac{V - V\_min}{V - \max - V\_min} \qquad (5)$$

After the normalization process of the data, ANN models were created, and the training-test process was started. MATLAB was preferred to carry out processes. Harmonized with the active issues, the number of input neurons of the ANN was strongminded to be 7 neurons. So as to regulate the most appropriate ANN prototypical, different artificial neural network models were created by changing the number of hidden layers and the number of neurons used in the hidden layer, and the performance values were compared.

In the study, monthly storage area amount data between January 2015 and April 2020 were used. Records that are far from normal or have missing data are separated from these data. 80% of the data gained as a result of the cleaning of noise data was used for education, and 20% was earmarked for testing.

In the study, many different ANN systems were created, and the training data prepared by pretreatment were presented to these networks. Test data was used to test the network models in which the learning process was carried out.

The designed artificial neural network models were evaluated using two different criteria. The first criteria is the mean square error (MSE) that reveals proximity difference between the predicted and the true value. It is desirable that this value be as small as possible.

The other evaluation criterion was chosen as the correlation coefficient (R). It reveals the closeness of the predicted values with the linear line formed by the real values. If the difference between the predicted and actual value is small, another word, ıf the R value closer to 1 the designed model is considered to have done well.

The models designed with ANN, is include 7 neurons in the input layer, 1 neuron in the output layer and the hidden layers prepared by using different number neurons for reduce the error rate. These models and the results obtained from these models are shown in Table1.

Table 1. Result obtained with using normalized data between [0,1]

Model	Number of Neurons	Training R	Validation R	Test R	All	MSE
1	7-3-5-1	0.88787	0.98975	0.96271	0.91949	0.00029254
2	7-3-7-1	0.96455	0.9857	0.96455	0.96037	0.00017394
3	7-3-8-1	0.86507	0.96346	0.97286	0.89123	0.00025499
4	7-3-10-1	0.96387	0.98994	0.97531	0.97812	0.00010108
5	7-5-5-1	0.92154	0.97738	0.89395	0.9201	0,0037697
6	7-5-7-1	0.96214	0.94995	0.95155	0.90736	0.0030171
7	7-5-8-1	0.96563	0.96894	0.93549	0.95104	0.00056888
8	7-5-10-1	0.84566	0.97257	0.83179	0.88206	0.00022034
9	7-7-7-1	0.73371	0.98858	0.84732	0.78773	0.00021507
10	7-7-10-1	0.94104	0.97766	0.96748	0.96119	0.00017227
11	7-8-8-1	0.95044	0.9226	0.96328	0.95596	0.00079471
12	7-8-10-1	0.91235	0.87792	0.96494	0.90418	0.0013207
13	7-10-5-1	0.80051	0.9376	0.87948	0.81794	0.0009469
14	7-10-7-1	0.73948	0.87411	0.8311	0.75942	0.0005859
15	7-10-10-1	0.819	0.98726	0.96496	0.89336	0,0087203

When the results of these models according to the determined performance criteria are examined; ıt is seen that the best result is the ANN model with numbered 4. This ANN model has 7-3-10-1 network structure and MSE value is the smalest with 0.00010108. Looking at the R values of the models, it is seen that the best results belong to the 4th model with Training R: 0.96387, Validation R: 0.98994, Test R: 0.97531 and All: 0.9781.

The MSE value, which is used as a performance criterion in artificial neural networks, was calculated ones again with using the test data allocated at the beginning of the study in order to quantity the analytical performance of the network constructions defined in the study. Mean Absolute Percent Error (MAPE) equation is used to determine performance and evaluate the best result. Performance criteria formulas are as presented in Equation 6 (Lewis, 1982). MSE and MAPE values obtained by using test data to determine performance criteria in the created ANN systems are given in Table 2.

$$MSE = \frac{\sum(y1 - y2)^2}{n}$$

$$MAPE = \frac{\sum\left|\frac{y1-y2}{y1}\right|}{n} * 100\ (\%) \qquad (6)$$

y1: actual value, y2: estimated values, n: number of samples

Table 2. MSE and MAPE values obtained using test data

Model	Number of Neurons	MSE	MAPE
1	7-3-5-1	0,00059	10,86376
2	7-3-7-1	0,000333	6,642618
3	7-3-8-1	0,000428	11,26686
4	7-3-10-1	0,000112	3,210506
5	7-5-5-1	0,003125	27,93325
6	7-5-7-1	0,003128	29,04445
7	7-5-8-1	0,001011	11,76617
8	7-5-10-1	0,000445	10,5565
9	7-7-7-1	0,000375	10,56968
10	7-7-10-1	0,000304	7,673828
11	7-8-8-1	0,00089	13,24089
12	7-8-10-1	0,00111	14,3777
13	7-10-5-1	0,000812	13,17937
14	7-10-7-1	0,000654	11,94267
15	7-10-10-1	0,002892	15,42667

There is an evaluation scale in the literature for the MAPE value. According to this scale, if the MAPE value is below 10%, the model is evaluated as "very good". If the MAPE value is between 10% and 20%, the model is called "good", and the models between 20% and 50% fall into the "acceptable" class. If this value exceeds 50%, the models are called "false" and are not evaluated.

3. Conclusion

In this study, it is suggested to use Artificial Neural Network based prediction models used in many fields in estimating the storage area, which is increasingly important for Electronic Document Management Systems and is necessary for the effective use of this system. Different artificial neural network models were created in the storage area estimation study using Kırıkkale University Electronic Document Management Systems data. The data separated as training and test data were processed with ANN models created using the MATLAB program. The results obtained are shown in tables. When these results are examined, it has been observed that the ANN model, which consists of an input layer with 7 neurons and an output layer with 1 neuron, with 3 - 10 neurons in its hidden layers, gives better results than other models. Mean square error value was found as 0,000112 and the MAPE value was calculated as 3,210506. According to the evaluation scale in the literature for the MAPE value, it was observed that the value obtained as a result of the study was less than 10% and entered the group defined as "very good". As a result of this study, it was seen that artificial neural networks can be used effectively with high accuracy in the problem of storage space estimation required in electronic document management systems.

References

Al-Saba, T. and El-Amin, I. (1999). Artificial neural networks as applied to long-term demand forecasting, Artificial Intelligence in Engineering 13 189–197.

Asilkan, Ö., Irmak S.(2009). Forecasting The Future Prices Of The Second-Hand Automobiles Using Artificial Neural Networks, Suleyman Demirel University The Journal of Faculty of Economics and Administrative Sciences, Vol.14, No.2 pp.375-391.

Basheer, I.A. and Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application, Journal of Microbiological Methods, 43 3–31

Çetinkaya Z. and Erdal E. (2019). Daily Demand Forecast With Artificial Neural Networks: Kırıkkale University Case, Institute of Electrical and Electronics Engineers, IEEE Xplore, 4th International Conference on Computer Science and Engineering, 19171793.

Çiçek, N. (2009). Modern Belgelerin Diplomatiği. İstanbul: Derlem Yayınları.

Delmar, F. and Davidsson, P. and Gartner, W. (2003). Arriving at the high growth firm. Journal of Business Venturing 18(2):pp. 189-216.

Dollar, C.M.(2002). Authentic electronic records: Strategies for long term access. Chicago: Cohasset Association.

Fausett, L. (1994), Fundamentals of Neural Networks: Architectures, Algorithms and Applications, Prentice Hall.

Jayalakshmi, T. and Santhakumaran, A. (2011). Statistical Nor-malization and Back Propagation for Classification, International Journal of Computer Theory and Engineering 3(1), pp. 1793-8201.

Külcü, Ö. (2007), Değişen Koşullarda Belge Yönetimi Çalışmaları ve Uluslararası Uygulamalar, XII. Türkiye'de İnternet Konferansı 8-10 Kasım, Ankara, s.57-81.

Kırıkkale Üniversitesi Elektronik Belge Yönetim Sistemi, http://bilgi.ebys.kku.edu.tr/

Lewis, C. D. (1982). Industrial and Business Forecasting Methods, Butterworths Publishing: London, s. 40

Liu, J., Savenije, H.H.G. and Xu, J. (2003). Forecast of water demand in Weinan City in China using WDF-ANN model, Physics and Chemistry of the Earth 28, 219–224

Moses, R. P. (2005). A Glossary of Archival and Records Terminology. Chicago: The Society of American Archivists.

Odabaş, H. (2008). Elektronik Belge Düzenleme Yaklaşımları ve Türkiye'de e-Devlet Uygulamalarında Elektronik Belge Yönetimi. Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 12(2), 121– 142.

Önaçan M.B.K., Medeni, T.D.,Özkanlı, Ö. (2012), Elektronik Belge Yönetim Sisteminin Faydaları ve Kurum Bünyesinde EBYS Yapılandırmaya Yönelik Bir Yol Haritası, Sayıştay Dergisi ,Sayı:85.

Öztemel, E. (2006). Yapay Sinir Ağları, Papatya Bilim Üniversite Yayıncılığı, İstanbul.

Shalabi, L.A. and Shaaban, Z. (2006). Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix, In: Proceedings of the International Conference on Dependability of Computer Systems, p. 207-214.

Shepherd, E. ve Geoffrey Y. (2003). Managing records a handbook of principles and practices. London: Facet Publishing.

Sola, J. and Sevilla, J. (1997). Importance of input data normalization for the application of neural networks to complex industrial problems, IEEE Transactions on Nuclear Science, v. 44, n. 3, p. 1464-1468.

Zhang, G., Patuwo, B. E. and Hu M. Y. (1998). Forecasting with Artificial Neural Networks: The State of The Art, International Journal of Forecasting, Vol.14, No-1, 35-62.