MEDWELL PUBLISHING
International Journal of Soft Computing 4 (1): 10-15, 2009
ISSN: 1816-9503

A New CMAC Neural Network Model for Content-based Web Page Classification

[1]Somaiyeh Dehghan and [2]Amir Masoud Rahmani
[1]Department of Computer Engineering, Islamic Azad University, Ilkhchi Branch, Iran
[2]Department of Computer Engineering, Islamic Azad University,
Science and Research Branch, Tehran, Iran

Abstract: The rapid growth of World Wide Web in recent years, makes it necessary for search engines to classify this data into categories. The automatic classification of web pages deals with text information, structure information and hyperlink information of web pages, which focus research on automatic classification of web text information, that is classification based on content. A major difficulty of content based web page classification is how to deal with high dimensional feature space. Based on the analysis done, CMAC neural network showed faster learning in high dimensional problems. In the present study, the new CMAC neural network model is proposed for use in content based web page classification. The results show that the proposed model is more useful than any other algorithms.

Key words: Web mining, Web Content Mining (WCM), web page classification, Cerebellar Model Arithmetic Computer (CMAC) neural network, feature selection, web document representation

INTRODUCTION

With the rapid growth of information on the World Wide Web, automatic classification of web pages has become important for effective indexing and retrieval of web documents. The automatic classification of web pages deals with text information, structure information and hyperlink information of web pages, which focus research on automatic classification of web text information, that is classification based on content (Liang, 2003). A major difficulty of this application is the high dimensionality of the feature space. A common approach to representing a text document is to use a bag of words that appear in the document. Since, a web page can contain thousands of words, the feature space for representing web pages is potentially huge. Few machine learning systems can handle such a large number of features.

Due to the diversity of web pages' content, complexity of structure and other characteristics, it is very difficult to improve the accuracy of automatic classification. Of course the researchers have designed different kinds of classifiers in order to solve this problem such as: Naive Bayes classifier (Fern´andez *et al*., 2006; Tomar *et al*., 2006), K-neighbor Clustering (Shi, 2002), SOM neural network (Zhang, 2002), Support Vector Machine (SVM) (Dumais, 2000; Xue, 2006) each of which has its own unique characteristics and applications. For example, Naive Bayes classifier is based on the assumption that features to be classified are orthogonal and subject to polynomial independent uniform distribution, K-neighbor Clustering assumes that no sample of another class appears in the known neighboring area, while SOM neural network is an order-holding mapping but requires more training time and SVM in spite of its strong classifying ability, demands a way of solving the problem through quadratic programming (Liang, 2003).

Based on the past analysis's (Albus, 1975; Palacios *et al*., 2006) because of its rapid learning qualification, capability of a better local generalization, guaranteed convergence and easy implementation by the hardware, the Cerebellar Model Arithmetic Computer (CMAC) has high potentiality in the production of effective data mining techniques.

In the present study, we have introduced the new CMAC neural network model for content based web page classification.

WEB PAGE REPRESENTATION AND FEATURE EXTRACTION

The first step in the process of content based web page classification is the choice of features. The purpose of feature selection is to find a subset of attributes, which can describe the documents the best with regard to classification process or by their choice the learning

Corresponding Author: Somaiyeh Dehghan, Department of Computer Engineering, Islamic Azad University, Ilkhchi Branch, Iran

algorithm has the most accuracy. Since, web data is unclear and disorganized, before the choice of features, there is need for some preprocessing steps on web documents. First the web pages are converted to HTML tag free text structures, then by eliminating the punctuation and stopwords (Frakes and Baeza-Yates, 1992) is purified and all the words in the document are changed to either small or capital letters. After this the number of terms remained in the text corpus is still a lot, so by the use feature extraction algorithm a number of key words are chosen for the description of the documents.

In order to select the features, first of all the web documents are converted to feature vectors. For displaying the documents in the form of vectors, there are different models such as: Vector Space Model (VSM) or Term Frequency-Inverse Document (TFID) (Salton *et al*., 1975), Boolean representation (Markov and Larose, 2007) and bags of words or Term Frequency (TF) (Markov and Larose, 2007). After converting the documents to feature vectors by one of the algorithms of feature extraction such as, Correlation-based Feature Selection (CFS) (Hall, 1998), Information Gain attribute evaluation (Quinlan, 1986) and Similarity-based Feature Selection (Kononenko, 1994) a number of terms are selected as features.

In this study, we have applied Boolean representation (Markov and Larose, 2007) method for representation of web documents. In Boolean representation method first the term-document matrix is made. In the term-document matrix columns represent terms and lines represent the documents. In Boolean representation method, which is the simplest model for displaying the documents, each term is considered as a Boolean attribute. As a result each cell of matrix consists 1 (if the term exists in the document) and/or 0 (if the term does not exists in the document).

After displaying the documents in the form of feature vectors, in order to determine, which term to select as the feature, the we used Information-based or Information Gain attribute evaluation (Quinlan, 1986). Information Gain attribute evaluation is based on entropy (Markov and Larose, 2007).

Let, S be a set of document vectors from k classes, C_1, C_2, ..., C_k. Then the number of vectors in S is $|S| = |S_1| + |S_2| + ... + |S_k|$, (where S_i is the set of vectors belonging to class) . The entropy is the average information needed to predict the class of an arbitrary vector in S. It is defined to be according to Eq. 1:

$$H(S) = -\sum_{i=1}^{k} P(C_i)\log P(C_i) \quad (1)$$

Where, the probability of class C_i is calculated as the proportion of instances in it (i.e., $P(C_i) = |S_i|/|S|$).

Assume now that attribute A has m values v_1, v_2, ..., v_m. Then A splits the set S into m subsets, A_1, A_2, ..., A_m, each including the documents that has value v_i for A. The entropy in the split based on attribute A is defined to be according to Eq. 2:

$$H(A_1, A_2, ..., A_m) = \sum_{i=1}^{m} \frac{|A_i|}{|S|} H(A_i) \quad (2)$$

where, $H(A_i)$ is the entropy of the class distribution in set A_i.

After splitting a set of documents the entropy in the split decreases. The best split (produced by the best attribute) will put in each A_i documents from a single class and thus, its entropy will be 0. The information gain (Quinlan, 1986) measures the quality of a split, respectively an attribute) by the decrease of entropy: that is according to Eq. 3:

$$Gain(A) = H(S) - H(A_1, A_2, ..., A_m) \quad (3)$$

CONVENTIONAL CMAC

A conventional CMAC model was introduced Albus (1975). The Conventional CMAC is a neural network, which models the structure and the operation of the cerebrum, part of the brain whose main function is to coordinate and control eyes, hands, fingers, arms and legs. The model is based on associative memory, which uses a lookup-table technique. The block diagram of CMAC model is shown in Fig. 1.

Figure 2 illustrates an example of CMAC structure with 2 s_1 and s_2 input variables. Each input variable is divided into a number of discrete sections called blocks. Figure 2, these sections are shown as A, B and C for s_1 and a, b and c for s_2. By shifting each block a small distance (called an element), different blocks arc created in different layers. For example: For A, B and C of's$_1$ input variable, we have D, E and F in the next layer. The blocks in 2 s_1 and s_2 input state form a space called hypercube. These spaces are named Bb, Ee, Ih. Each hypercube is a memory cell, which stores and retrieves data (Rahmani, 2003).

Output mapping: The output corresponding to the inputs is saved as information or weights in N_e association memory cells or in N_e hyper cubes. In order to obtain the outputs the Eq. 4 is used:

$$y(s) = \alpha^T(s)W = \sum_{j=1}^{M} \alpha_j(s)W_j \quad (4)$$

where, s is α distinguished state input, M is the volume of the memory and is equal to the number of the hyper cubes. If the location of memory j is covered by one of hyper cubes of input state s, a_j (s) = 1otherwise will be 0.

Learning algorithm: In CMAC model, supervised learning is used to adjust the stored weights in memory cells. Learning is iterative and based on global error generalization. From this view it resemble to standard back propagation algorithm. Learning algorithm has two steps: Applying the inputs and obtaining the outputs and Adjusting weights by comparing the obtained outputs with the desired ones.

The first step is obtained by the Eq. 4 or output mapping and the second one is obtained through the Eq. 5:

$$W_{new} = W_{old} + \frac{\alpha}{N_e}\alpha(s)(\hat{y}(s) - y(s)) \quad (5)$$

Where, α is the learning rate such that $0<\alpha<1$, N_e is the number of block elements or the number of layers, $\hat{y}(s)$ is the amount of desired output and $(\hat{y}(s) - y(s))$ is the error amount for the training sample.

The error is multiplied by a(s) in order to adjusting the weights of the locations of memory, which are covered by hyper cubes. CMAC neural network has generalization potentiality such that if applied input to the network, which the network has not been taught to those inputs, the suitable output is generate.

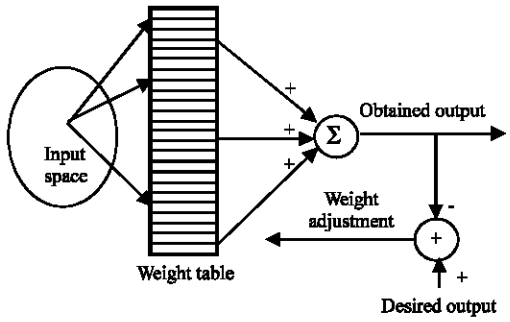

Fig. 1: The block diagram of CMAC

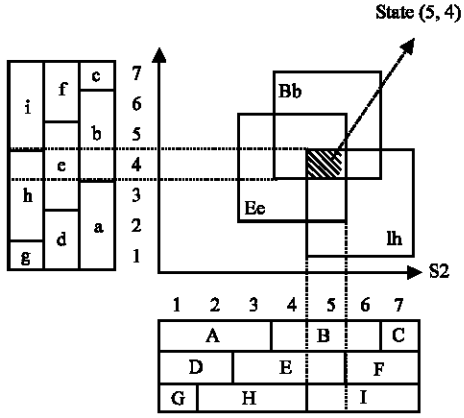

Fig. 2: The CMAC structure with 2 inputs

THE PROPOSED NEW CMAC NEURAL NETWORK MODEL

The problem in web page classification is with high dimensional feature space, which increases the required memory for CMAC. For solving this problem and adaptation CMAC neural network for classifying web pages based on content, we proposed the new CMAC model, which is shown in Fig. 3.

In content based web page classification, the presence or absence of features or the number of feature occurrences together with keyword of each class, determine the class of the web pages.

Thus, each class has at least one keyword and if the number of the classes is C = $\{c_1,..., c_m\}$, the least number of keywords is also K = $\{k_1, ..., k_m\}$. Also, by using Information Gain attribute evaluation (Quinlan, 1986), N features extracted from the web pages, such that Features $(c_i) = \{x_1, ..., x_N\}$.

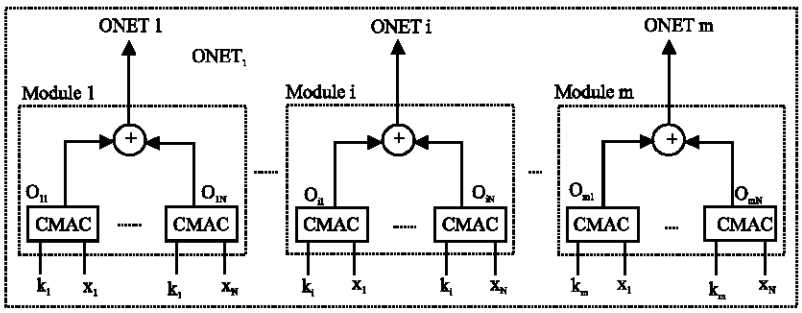

Fig. 3: The new CMAC model

In the new CMAC model the number of modules is equal to the number of classes, that is M = {M_1, ..., M_m}. Within each module equal to the selected features from documents there are 2 inputs CMAC that is N 2 inputs CMAC. Then inputs of each CMAC within each module M_i is the keyword of class C_i and one of the selected feature from the same class. In the new CMAC model the output of each module ($ONET_i$) is the sum of two-inputs CMAC outputs.

Output mapping: In this model output of each CMAC is the stored weights in hyper cubes, which covered the input space (k_i, x_j), which is calculated through Eq. 6:

$$O_{ij}(s) = \alpha^T(s)W = \sum_{j=1}^{M} \alpha_j(s)\ W_j \quad (6)$$

where, s is a distinguished state input, M is the volume of the memory and is equal to the number of the hyper cubes. If the location of memory j is covered by one of hyper cubes of input state s, $\alpha_j(s)$ = 1otherwise will be 0.

Also, the output of each module ($ONET_i$) is the sum of CMAC outputs when the number of occurrence of feature x_j in training sample is not zero, that is: frequency(x_j)≥1.

For this purpose, we define the vector P = [p_1 p_2 p_j ... p_N]. If feature x_j appears in the sample web document, p_j = 1 otherwise p_j = 0. Thus, the output of module i is obtained through the Eq. 7:

$$ONET_i = \sum_{j=1}^{N} O_{ij} \times p_j \quad (7)$$

where, O_{ij} is the output of jth CMAC in module i, N is the number of features, which has been extracted form web pages and p_j shows the presence or absence of feature x_j in the training sample.

Learning algorithm: The learning algorithm in the new CMAC model is similar to the conventional CMAC. This model makes use of supervised learning based on global error generalization in order to adjust the stored weights in hyper cubes. Learning Rule obtained through the Eq. 8:

$$W_{new} = W_{old} + \frac{\alpha}{N_P \times N_e} a(s)(\hat{ONET}_i - ONET_i) \quad (8)$$

where, α is the learning rate such that $0<\alpha<1$, N_p is the number of non zero elements of vector P, N_e is the number of layers, $\hat{ONET}_i$ is the amount of desired output and $(\hat{ONET}_i - ONET_i)$ is the error amount for the training sample. The error is multiplied by $\alpha(s)$ in order to adjusting the weights of the locations of memory, which are covered by hyper cubes.

EXPERIMENTAL RESULTS

Dataset used: In order to show, the learning efficiency of the proposed model, the dataset of Syskill and Webert web pages ratings (Pazzani *et al*., 1996) were used for testing the new CMAC model. Pazzani *et al*. (1996) suggested an intelligent agent by the name of Syskill and Webert, which was designed to help users identify their favorite web pages on a special topic. This system provides an user interface so that users can rate the web pages to 3 different interesting classes: hot, medium, or cold. The results of the rating are recorded as user's profile with positive and negative examples. In this system each user has a set of profiles for every topic. The dataset of Syskill and Webert web pages ratings consists of HTML web pages together with the single user's rating on four distinct topics: Bands-recording artists, Goats, Sheep and Biomedical. This rating consists of the following 5 records: the name of source HTML web pages, the level of rating (cold, medium, hot), URL of the web site, the visited data and the title of web page.

In most researches, the medium level of this dataset because of its few number is merged with cold pages (Pazzani *et al*., 1996). Table 1 shows the rating subjects of single user, different interesting levels and the total number of web pages.

Data preparation: First of all, we converted web pages of Syskill and Webert dataset into HTML tag free text, then by using Weka software (Markov and Larose, 2007) the

Table 1: The used topic in the experiments

	Number of pages			
Topic	Hot pages	Medium pages	Cold pages	Total number of pages
Bands	15	7	39	61
Goat	32	1	37	70
Sheep	14	0	51	65
BioMedical	32	3	101	136

Table 2: Four labels assigns to each document in classification tasks

True Positive (TP):	Actual positive and predicted as positive
False Positive (FP):	Actual negative but predicted as positive
True Negative (TN):	Actual negative and predicted as negative
False Negative (FN):	Actual positive but predicted as negative

Table 3: The accuracy of various classification algorithms and the new CMAC model

	Algorithms							
Topic	N Neigh	ID3	Percept	BackP	PEBLS	Bayes	Rocchio	New CMAC Model
Bio medical	74.50	70.2	73.20	76.00	74.60	77.30	77.50	72.50
Bands	74.40	70.7	71.40	73.10	74.50	73.40	73.70	78.00
Goats	62.00	64.7	66.30	67.00	62.70	62.90	69.40	74.00
Sheep	79.30	78.4	78.90	80.50	79.30	81.50	78.80	84.40
Average accuracy on 4 topics	72.55	71.0	72.45	74.15	72.77	73.77	74.85	77.22

documents in the dataset were purified with the omission of punctuations and stopwords (Frakes and Baeza-yates, 1992) and all the words were changed into small letters. Then by using Weka, Boolean representation of documents was prepared. By the help of Weka, we used Information Gain attribute evaluation algorithm (Quinlan, 1986) for selecting features from web pages and a number of informative terms as features were selected. Then whole documents were converted feature vectors, which is learnable by the new CMAC model.

Performance measure: The major measurement criterion for classification process is *Accuracy*. As the most common clustering and classification problems involve 2 classes, they are usually called positive and negative. Also, the original class labels are referred to as actual and those determined by the classification algorithm are called predicted. According to this terminology, the Classes-to-Class evaluation assigns to each document one of the following 4 labels in Table 2 (Markov and Larose, 2007).

With regard to Table 2, *Accuracy* is the ratio of the number of predicted corrects (TP + TN) to total predicted ones, which is represented in Eq. 9:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (9)$$

Experiment: In this experiment the reported results by Pazzani *et al*. (1996), which used the seven classification algorithms for learning user profiles on Syskill and Webert rating web pages, compared with the results of the proposed new CMAC neural network results. For comparison the identical conditions were used. In this experiment a training set by the size of 20 samples of web pages was used.

Also for selecting 128 features, Information Gain attribute evaluation (Quinlan, 1986) and for constructing feature vectors, Boolean representation model (Markov and Larose, 2007) were used. Table 3 shows the *Accuracy* of seven algorithms Pazzani *et al*. (1996) and new CMAC model in the identical conditions.

The results of Pazzani *et al*. (1996) show that Bayesian Classifier has the best efficiency. The results obtained from experiment show that the new CMAC model in average has better efficiency with respect to algorithms.

CONCLUSION

In this study, the new CMAC model was introduced for content based web page classification, which is capable of learning the profiles of users. The experimental results show that the new CMAC model has fast learning and suitable generalization and when compared with other Classification algorithms has more Accuracy.

REFERENCES

Albus, J.S., 1975. Data Storage in the Cerebellar Model Articulation Controller (CMAC). J. Dynamic Syst. Measurement and Controller. Trans. ASME, pp: 228-233.

Dumais, S. and H. Chen, 2000. Hierarchical Classification of Web Content. Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieva.

Fern´andez, V.F., R.M. Unanue, S.M. Herranz and A.C. Rubio, 2006. Naive Bayes Web Page Classification with HTML Mark-Up Enrichment. Proc. Int. Multi-Conference on Comput. Global Inform. Technol., pp: 48.

Frakes, W. and R. Baeza-Yates, 1992. Information Retrieval: Data Structures and Algorithms. Prentice Hall, Englewood Cliffs, NJ.

Hall, M.A., 1998. Correlation-based feature selection for machine learning. PhD Thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand.

Kononenko, I., 1994. Estimating attibutes: Analisys and extensions of relief. Proc. Seventh Eur. Conf. Machine Learning, Springer-Verlag, pp: 171-182.

Liang, J.Z., 2003. Chinese Web page classification based on self-organizing mapping neural networks. Proc. Fifth Int. Conf. Computational Intelligence and Multimedia Applications, ICCIMA, pp: 96-101.

Markov, Z. and D.T. Larose, 2007. Data mining the web: Uncovering patterns in web content, structure and usage. Wiley.

Palacios, F., X. Li and L.E. Rocha, 2006. Data Mining based on CMAC Neural Networks. 3rd International Conference on Electrical and Electronics Engineering, pp: 1-4.

Pazzani, M., J. Muramatsu and D. Billsus, 1996. Syskill and Webert: Identifying Interesting Web Sites. Proceeding of the 13th National Conference Artificial Intelligence, pp: 54-59.

Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn., 1 (1): 81-106.

Rahmani, A.M., 2003. TD-CMAC Intelligent Network for Short-Term Maximum Temperature Prediction. 4th International Arab Conference on information technology (ACIT 2003). Neural Network, Dec. Alexandria Egypt, 1: 273- 280.

Salton, G., A. Wong and C.S. Yang, 1975. A Vector Space Model for Automatic Indexing. Communications of the ACM, 18 (11): 613-620.

Shi, Z.Z., 2002. Knowledge Discovery (in Chinese). Beijing, Tsinghua University Press.

Tomar, G.S., S. Verma and A. Jha, 2006. Web Page Classification using Modified Naive Bayesian Approach. IEEE Region 10 Conf. TENCON, pp: 1-4.

Xue, W., W. Huang and Y. Lu, 2006. Web Page Classification Based on SVM. Proceedings of the 6th World Congress on Intelligent Control and Automation, Dalian, China, pp: 6111- 6114.

Zhang, Y.Z., 2002. Web page's Text Information Mining Based on Content (in Chinese). Postdoctoral Report, Tsinghua University.