Mixture Kernel Radial Basis Functions Neural Networks for Web Log Classification

Dash Ch. Sanjeev Kumar[1], Pandia Manoj Kumar[1], Dehuri Satchidananda[2], and Cho Sung-Bae[3]

[1] Department of Computer Science, Silicon Institute of Technology, Patia, Bhubaneswar-24, Odisha, India
[2] Department of Information and Communication Technology, Fakir Mohan University, Vyasa Vihar, Balasore-756019, Odisha, India
[3] Soft Computing Laboratory, Department of Computer Science, Yonsei University, 50 Yonsei-ro, Sudaemoon-gu, Seoul 120-749, South Korea.
{sanjeev_dash,manoj_pandia}@yahoo.com, satchi.lapa@gmail.com, sbcho@yonsei.ac.kr

Abstract. With the immense horizontal and vertical growth of the World Wide Web (WWW), it is becoming more popular for website owners to showcase their innovations, business, and concepts. Along side they are also interested in tracking and understanding the need of the users. Analyzing web access logs, one can understand the browsing behavior of users. However, web access logs are voluminous as well as complex. Therefore, a semi-automatic intelligent analyzer can be used to find out the browsing patterns of a user. Moreover, the pattern which is revealed from this deluge of web access logs must be interesting, useful, and understandable. A radial basis function neural networks (RBFNs) with mixture of kernels are used in this work for classification of web access logs. In this connection two RBFNs with different mixture of kernels are investigated on web access logs for classification. The collected data are used for training, validation, and testing of the models. The performances of these models are compared with RBFNs. It is concluded that mixture of appropriate kernels are an attractive alternative to RBFNs.

Keywords: Neural networks, Radial basis function neural networks, Classification, Web log, Mixture kernel.

1 Introduction

The web has influenced a lot to both users (visitors) as well as the web site owners. The web site owners are able to reach to all the targeted audience domestically and internationally. On the other side visitors are also availing those facilities. Over the last decades, the proliferation of information on the web has resulted in a large repository of web documents stored in multiple websites. These plethora and diversity of resources have promoted the need for developing a semi-automatic mining technique on the WWW, thereby giving rise to the term of web mining [14, 15].

S.C. Satapathy et al. (Eds.): Proc. of Int. Conf. on Front. of Intell. Comput., AISC 199, pp. 1–9.
DOI: 10.1007/978-3-642-35314-7_1

Every website contains multiple web pages, each of which has: 1) contents: which can be in any form, e.g., text, graphics, multimedia etc; 2) structure: containing links from one page to another; and 3) usage: users accessing the web pages. According to this the area of web mining can be categorized like Table 1.

Table 1. Web Mining Categorization

Type	Structure	Form	Object	Collection
Usage	Accessing	Click	Behavior	Logs
Content	Pages	Text	Index	Pages
Structure	Map	Hyperlinks	Map	Hyperlinks

Mining the contents of web pages is called as “Content Mining”. Similarly, mining the links between web pages and the web access logs are respectively called “Structure” and “Web Usage” mining.

Web servers record and mount up data about user interactions whenever request for resources are received. Analyzing the web access logs of different web sites can help understand the user behavior and the web structure, thereby improving the design of this colossal collection of resources. To proceed towards a semi-automatic intelligent web analyzer, obviating the need for human intervention, we need to incorporate and embed computational or artificial intelligence into web analyzing tools. However, the problem of developing semi-automated tools in order to find, extract, filter, and evaluate the users desired information from unlabeled, distributed, and heterogeneous web access logs data is far from being solved. To handle these characteristics and overcome some of the limitations of existing methodologies, RBFN (a member of computational intelligence family) seems to be a good candidate. In this paper RBFN with mixture of kernels is used to design a classifier for web access logs classification. Li et al. [17] and Junjie et al. [18] works of RBFN for classification of web usage data are the source of motivation to carry out this work.

Over the last several decades multilayer perceptron (MLP) network was the popular network architectures and used in most of the applications. In MLP, the weighted sum of the inputs and bias term are passed to activation level through a transfer function to produce the output, and the units are arranged in a layered feed-forward topology called Feed Forward Neural Network (FFNN). In particular, MLP using back-propagation learning algorithm has been successfully applied to many applications. However, the training speed of MLP is typically much slower than those of feed-forward neural network comprising of single layer. Moreover, the problems such as local minima trapping, saturation, weight interference, initial weight dependence, and over-fitting make MLP training difficult [19]. Additionally, it is also very difficult to determine the parameters like number of neurons in a layer, and number of hidden layers in a network, thereby deciding a proper architecture is not trivial. These issues create the source of motivation to choose RBFNs over other alternatives to solve the problem focused in this paper.

Radial basis function neural network is based on supervised learning. RBFN networks are also good at modeling nonlinear data and can be trained in one stage rather than using an iterative process as in MLP. Radial basis function networks [1, 2, 3] have been studied in many disciplines like pattern recognition [4], medicine [5], multi-media applications [6], computational finance [7], software engineering [9], etc. It is emerged as a variant in late 1980's, but its root entrenched in much older pattern recognition, numerical analysis and other related fields [8]. The basic architectural framework of RBFN and its improvement by considering mixture of kernels are discussed in Section 3.4.

The rest of the paper is organized as follows. In Section 2, we cover briefly about the basics of RBFNs, as because this research is just improvement over RBFNs. Section 3 presents the framework of mixture kernels based RBFNs for classification. Sections 4 discuss the experimental study followed by conclusions and future work in Section 5.

2 Radial Basis Function Neural Networks

A RBFN is a multilayer and feed forward network which is often used for interpolation in multidimensional space. The basic architecture of a three-layered network is shown in Figure 2. A RBFN has three layers consisting of input, hidden, and output layers. The input layer receives input data. The hidden layer transforms the data from the input space to the hidden space using a non-liner function. The output layer, which is linear, yields the response to the network. Its training procedure is usually divided in to two stages. First, the centers and widths of the hidden layer are determined by various ways like clustering algorithms such as k-means algorithm [10], vector quantization [11], decision trees [12], or self-organizing future maps [13]. In this paper random initialization of centers and width is considered and then fine tuned over iterations. Second the weights connecting the hidden layer with the output are determined by singular value decomposition (SVD) or least mean square (LMS) algorithm. Here SVD is used for optimizing the weights. The number of basis functions controls the complexity and the generalization ability of RBF network.

The argument of the activation function of each hidden unit in an RBFN computes the Euclidean distance between an input vector and the center of the unit.

$$\Phi_i(x) = \exp\left(-\frac{\| x-\mu_i \|^2}{2\sigma_i^2}\right) \tag{1}$$

where ||...|| represents Euclidean norm, μ_i, σ_i and ϕ_i are center, width and the output of the ith hidden unit. The output layer, a set of summation units, supplies the response of the network.

The commonly used radial basis functions are enumerated in Table 2 as follows:

Table 2. Kernel functions used in RBFNs

Name of the Kernel	Mathematical Formula
Gaussian Functions	$\Phi(r)=\exp\left(-\frac{x^2}{2\sigma^2}\right)$ width parameter $\sigma>0$
Generalized Multi-Quadric Functions	$\Phi(r)=\left(x^2+\sigma^2\right)^{\beta}$ Parameter $\sigma>0$, $1>\beta>0$
Generalized Inverse Multi Quadric Functions	$\Phi(x)=(x^2+\sigma^2)^{-\alpha}$
Thin Plate Spline Function	$\Phi(x)=x^2\ln(x)$
Cubic Function	$\Phi(x)=x^3$
Linear Function	$\Phi(x)=x$

3 The Framework of Mixture Kernel Based RBFNs

Figure 1 describes overall framework of our research. Subsection 3.1 - 3.3 describes the web usage mining and steps required for preprocessing the dataset. Subsection 3.4 describes RBFN with mixture kernels for classification of web access logs.

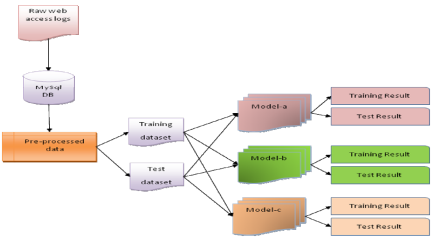

Fig. 1. Block diagram of web log classification

3.1 Web Usage Mining and Preprocessing

While visiting a website user navigates through multiple web pages. Each access to a web page is stored as a record in log files. These log files are managed by web servers. The format of log files varies from server to server. Apache web server maintains two different log file formats namely 1) Common Log Format, and 2) Combined Log Format.

The raw data from access log file are not in a state to be used for mining. These records must be pre-processed at first. Pre-processing [16] involves 1) Data cleaning, 2) User identification, and 3) Session identification.

For pre-processing we have followed the database approach where all the raw records are inserted into a table and the pre-processing tasks such as data cleaning,

user identification and session identification are done using SQL queries. This approach gives better flexibility and faster execution.

First the raw web access log records are scanned from log files and inserted into a table which is designed in MySQL database. As we know query execution is easy and also the time taken for query execution is much less in comparison to file processing. Once all records are inserted successfully, the steps of preprocessing can be executed.

In this paper we have used the web access log data from www.silicon.ac.in. Silicon institute of technology is one of the premiere technical institutes in Odisha, India. We have collected the records between 22-Oct-2010 04:15:00 to 24-Oct-2010 04:05:48. The total number of records in the file is 12842.

3.2 Data Transformation

We have written a java program which reads each line from the log file and insert in to the table in MySQL database.

3.3 Data Cleaning

For data cleaning we need to remove all the records which contain the request for files like jpg, gif, css, js, etc. For this we execute the delete query.

Total number of records reduced to 1749 from 12842 after the data cleaning. Every unique page is assigned a PageID. The total unique web pages are found to be 97.

3.4 Mixture Kernels Based RBFNs

Suppose we take the region R to be a hypercube with sides of length h centered on the point x. Its volume is then given by

$$V=h^d . \tag{2}$$

We can find an expression for K, the number of points which fall within this region, by defining a kernel function $H(u)$, also known as Parzen window given by

$$H(u)=\begin{cases} 1 & |u_j|<1/2 \\ 0 & \text{otherwise} \end{cases} \qquad j=1,\ldots,d \tag{3}$$

so that $H(u)$ corresponds to a unit hypercube centered at the origin. Thus for all data points x^n, the quantity $H((x-x^n)/h)$ is equal to unit y if the point x^n falls inside a hypercube of side h centered on x, and is zero otherwise. The total number of points falling inside the hypercube is then simply

$$K=\sum_{n=1}^{N} H\left(\frac{x-x^n}{h}\right) \tag{4}$$

Mixture models are a divide-and-conquer learning method derived from the mixture estimation paradigm [21] that is heavily studied in artificial neural network research

[22]. They reduce the complexity by decomposing learning tasks and variance by combining multiple kernels.

A probability density function $p(x)$ can be modeled as a superposition of component densities as

$$p(x) = \sum_{j=1}^{M} P(j)\Psi_j(x) \tag{5}$$

Such a representation is called mixture distribution and the coefficients $P(j)$ are called the mixing parameters, where p is the prior probability of the component density ψ_j. The parameters of such a mixture model can be obtained using maximum likelihood [22]. These priors are chosen to satisfy the constraints.

$$\sum_{j=1}^{M} \psi(j) = 1 \tag{6}$$

$$0 \le \psi(j) \le 1 \tag{7}$$

Similarly, the component density functions $p(x \mid j)$ are normalized so that

$$\int p(x \mid j)dx = 1 \tag{8}$$

and hence can be regarded as class-condition densities. To generate a data point from the probability distribution [13], one of the component j is first selected at random with probability $\psi(j)$, and then a data point is generated from the corresponding component density p(x|j). Fatai [21] was done a comparative study of the application of Gaussian Mixture Model (GMM) and Radial Basis Function (RBF) in biometric recognition of voice.

Recall that RBF network contains three layers. In the hidden layer, the number of neurons is decided by leaning process. The two proposed models including RBFNs (model-a) have used nine neurons in the hidden layer. In model-b we have used combination of Gaussian, multi-quadric, and inverse multi-quadric functions.

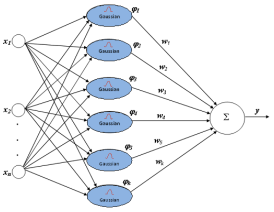

Fig. 2. (Model-a)

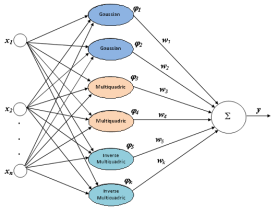

Fig. 3. (Model-b)

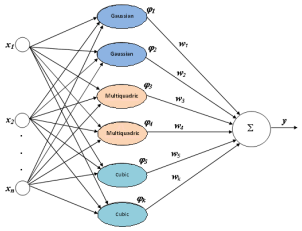

Fig. 4. (Model-c)

4 Experimental Study

4.1 Description of Dataset and Parameters

For classification purpose we have categorized the access time into two categories, i.e. from 00:00 hours to 12:00 hours as category 1 and 12:00 to 24:00 as category 2. Similarly we have grouped all the region of access into two, i.e. IP addresses accessing from within India as belongs to category 1 and IP addresses accessing from outside India as in category 2. To find the region from IP address we have used the website www.ip2location.com [18].

Our class attribute is frequent. For this we have considered two values i.e., 1 for not frequent and 2 for frequent.

This dataset includes 237 instances and each of which has 4 attributes, e.g., ID, request count, timeid, and locationid. The instances are classified into 2 classes (i.e., every instance either belongs to class 1 or 2). Class 1 has 129 instances, and class 2 contains 108. None of the attributes contain any missing values.

4.2 Results and Analysis

The results obtained from our experimental study are described in Table 3 according to different kernels like Gaussian functions, multi-quadric functions, inverse multi-quadric function, and cubic function. It is observed that model-b (Fig.3) [mixture of Gaussian, multi-quadric, and inverse multi-quadric RBF network] gives better result as compared to model-a (Fig.2) [RBFN with Gaussian] and model-c (Fig.4) [mixture of Gaussian, multi-quadric, and cubic] for web log dataset.

Table 3. Classification accuracy obtained from simulation of model-a, model-b, and model-c

Name of the Kernel	Hidden Units	Training	Testing	Average Accuracy
RBFN (with Gaussian)	9	83.7209	51.9380	67.82945
Mixture RBFN (Gaussian, multi-quadric, and inverse multi-quadric)	9	75.1938	76.7442	75.969
Mixture RBFN (Gaussian, multi-quadric, and cubic)	9	74.4186	76.7442	75.5814

5 Conclusion and Future Work

In this paper, we have compared RBFNs with two mixture models of RBFNs for classification on web log data. The average accuracy of second model (i.e., model-b)

gives better result than the first and third model. From the analysis we conclude that the web pages of www.silicon.ac.in are accessed most frequently at daytime inside India. As we have applied the classification on the attributes region and time, similarly this can be extended to consider some more attributes like user agent and referrer. Here we have considered only Gaussian, multi-quadric, inverse multi-quadric, and cubic kernel. By evolutionary technique this can be further extended to use other combination of kernels for higher accuracy.

References

1. Powell, M.J.D.: Radial Basis Functions for Multi-variable Interpolation: A Review. In: IMA Conference on Algorithms for the Approximations of Functions and Data, RMOS Shrivenham, UK (1985)
2. Broomhead, D.S., Lowe, D.: Multivariable Functional Interpolation and Adaptive Networks. Complex Systems 2, 321–355 (1988)
3. Buhmann, M.D.: Radial Basis Function Networks. In: Encyclopedia of Machine Learning, pp. 823–827 (2010)
4. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Academic Press, San Diego (2008)
5. Subashini, T.S., Ramalingam, V., Palanivel, S.: Breast Mass Classification Based on Cytological Patterns Using RBFNN and SVM. Expert Systems with Applications 36(3), 5284–5290 (2009)
6. Dhanalakshmi, P., Palanivel, S., Ramalingam, V.: Classification of Audio Signals Using SVM and RBFNN. Expert Systems with Applications 36(3), part 2, 6069–6075 (2009)
7. Sheta, A.F., De Jong, K.: Time Series Forecasting Using GA Tuned Radial Basis Functions. Information Sciences 133, 221–228 (2001)
8. Park, J., Sandberg, J.W.: Universal Approximation Using Radial Basis Function Networks. Neural Computation 3, 246–257 (1991)
9. Idri, A., Zakrani, A., Zahi, A.: Design of Radial Basis function Neural Networks for Software Effort Estimation. International Journal of Computer Science 7(4), 11–17 (2010)
10. Moody, J., Darken, C.J.: Fast Learning Networks of Locally-Tuned Processing Units. Neural Computation 6(4), 281–294 (1989)
11. Falcao, A., Langlois, O.T., Wichert, A.: Flexible Kernels for RBF Networks. Neurocomputing 69, 2356–2359 (2006)
12. Ghodsi, A., Schuurmans, D.: Automatic Basis Selection Techniques for RBF Networks. Neural Networks 16, 809–816 (2003)
13. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, New York (1995)
14. Hu, C., Zong, X., Lee, C.W., Yeh, J.H.: World Wide Web Usage Mining Systems and Technologies. Journal of Systemics, Cybernetics and Informatics 1(4), 53–59 (2003)
15. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N.: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations 1(2), 1–12 (2000)
16. Li, Z., He, P., Lei, M.: Applying RBF Network to Web Classification Mining. Journal of Communication and Computer 2(9) (2005) ISSN 1548-7709
17. Junjie, C., Rongbing, H.: Research of Web Classification Mining based on RBF Neural Network. In: Proceedings of Control, Automation, Robotics and Vision Conference, vol. 2, pp. 1365–1367 (2004)

18. IP to location mapping, http://www.ip2location.com
19. Dehuri, S., Cho, S.B.: A Comprehensive Survey on Functional Link Neural Networks and an Adaptive PSO-BP Learning for CFLNN. Neural Computing and Applications 19(2), 187–205 (2010)
20. Anifowose, F.A.: A Comparative Study of Gaussian Mixture Model and Radial Basis Function for Voice Recognition. International Journal of Advanced Computer Science and Applications 1(3), 1–9 (2010)
21. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley, New York (1973)
22. Jordan, M.I., Jacobs, R.A.: Hierarchical Mixtures of Experts and the EM Algorithm. Neural Computation 6, 181–214 (1994)