International Journal of Database Theory and Application
Vol.9, No.7 (2016), pp.269-278
http://dx.doi.org/10.14257/ijdta.2016.9.7.23

Application of Improved BP Neural Network Algorithm in Data Mining Research

Yang liu

Chongqing Nanfang Translators college of SISU
23989060@qq.com

Abstract

With the development of network technology, the data capacity become more abundant. How to effectively manage the data, the retrieval more quickly, accurately, improve the data classification accuracy becomes crucial. The BP neural network algorithm with its learning speed, strong ability to adapt and is widely used in network in data mining. Exist but its convergence rate is not high and big error and other shortcomings, therefore, on the basis of traditional algorithm, an improved BP neural network algorithm is put forward. Low error, through experimental analysis, the improved algorithm convergence rate is better.

Keywords*: Data mining; Improved BP neural network; topology*

1. Introduction

The application of computer has deep into every aspect of life, and continues to affect our life. Since the birth of the computer, processing data and information has become one of the main job of it. Along with the wide application of computer technology and the advent of the era of information, we have in the information is everywhere, everywhere without time. In order to effectively use and found huge amounts of information that exist in a lot of knowledge, data mining technology arises at the historic moment, and with the help of many scholars and enterprises at home and abroad. Now Internet global link up, to realize the global information sharing. When we never leave home can understand world immediately, make friends, we are also from like a vast sea of information for how to grab a ladle knowledge of what we need. Network data mining technique combined with the traditional mining algorithm and the special characteristics of network information, the Internet information mining, let's get useful knowledge from the Internet. Nowadays, data mining technology on the network and web search technology rapid development, and then there are many famous company, on the other hand, the progress of the applied technology and promote the rapid development of the network data mining technology, the network data mining theory, whether in study or in applications as we need and we may need to keep pace with The Times. However, on the network data mining still have a lot of things to do, have a lot of problems need to study, there are many aspects to discuss [1-2].

In practice, how to fast and accurately obtain the information from the Internet has become an important standard of the user experience. In addition, the social informatization degree is higher and higher, the individuality service will become a trend, we hope that you can according to the characteristics and needs of the users system give personalized service, vary from person to person. This is especially need to extract knowledge from information in the targeted guidance to service in the future. Web data mining technology according to the different characteristics of network information from different directions at the information on the mining, make everyone can contain vast amounts of information from the Internet in the access to knowledge, they want as far as possible let users get accurate and the desired results. And with the use of multimedia

ISSN: 2005-4270 IJDTA

technology on the network, the development of network technology, multimedia information and the 3 d information mining is becoming more and more important, however, no matter in the mining process, the classification of information is necessary for mining and is of vital importance.

In the process of web content mining need to network data document classification effectively, it is found that the one of knowledge and its important part.When classifying network data documents, can a good classifier algorithm in the process of document data mining based on data acquisition, word processing, data processing and characteristics of library work effectively and finally complete the classification of the critical work.Data classification results directly affect the final result of web data mining, so the design of the data classifier algorithm is very important.A good classifier algorithm can work effectively on the basis of the icing on the cake, in front of the classification of network data document accurately and effectively, thus improve the performance of web content mining [3].

2. Related Works

2.1. Data Mining Technology

Dug up from huge amounts of data in data mining, it is implied in the reserves - the process of knowledge. Generally believe that data mining has broad sense and narrow sense, broad, also known as knowledge discovery in database, data mining from large, incomplete, noisy, not clear and random data, extract contains in it, people don't know in advance, but it is potentially useful information and knowledge of the process. Data mining technology in its narrow sense refers to the use of various analysis tools found in the huge amounts of data model and data of the relationship between process, is an important step in the process of knowledge discovery. Data mining is the term first appeared in the United States Detroit at the international joint conference on artificial intelligence symposium. A few years later, as the technology mature, a considerable amount of data mining products and application system also emerged, and won. In recent years, the personnel engaged in research and development of data mining in over 80 countries around the world, also from the research of data mining algorithm to the commercialization of the specific application. Data mining technology to cross the many other discipline knowledge, in different areas have different application, involves the database technology, statistics, pattern recognition, signal processing, artificial intelligence and machine learning disciplines. Data mining is a kind of decision support process, it is mainly based on artificial intelligence, machine learning, statistics, and so on technology, highly automated analysis of original data for inductive reasoning, dig out the potential patterns, predict the behavior of users and help enterprise decision-makers to adjust market strategy, reduce risk, make the right decisions. At present, the typical data mining research field with association rules, classification, clustering, prediction and web mining, *etc*. Data mining technology has become an important direction of applied research. Many large companies both at home and abroad are focused on the research and application of data mining technology, and developed many widely used data mining system.

Starting from the different applications, data mining can be divided into different categories. Depending on the type of database classification, according to the technology and method of classification, depending on the type of knowledge mining classification, according to the application field of data mining classification, *etc.*

1 according to the mining database type classification

Data mining must be based on a certain data sources, these data sources may be a variety of database. If was conducted on the basis of relational database, data mining is called a relational database; If was conducted on the basis of object-oriented database, data mining is called object-oriented databases. Similarly, for what type of database, is

based on this type of database data mining."Accordingly, and relationship - objects of data mining, transactional data mining, multimedia data mining, interpretation of data mining, data, spatial data mining, data mining/time sequence data mining, data mining, data warehouse, data mining, *etc.*

2 according to the technology and method of classification

According to the data mining technology and methods adopted by the data mining can be divided into the following categories: supervised and unsupervised data mining of data mining, found that drive them, interactive data mining, data mining machine data mining, statistical analysis of data mining, fuzzy data mining, biological technology, data mining and visualization of data mining.

3 depending on the type of knowledge mining classification

Depending on the type of knowledge mining can be divided into mining association rules, characteristics of rule mining, classification rule mining, clustering rule mining, mining sequence rules, deviation rule mining, outlier mining and so on.

According to the application domain classification mining. For different industries can be divided into: the telecoms industry's ethos of data mining, data mining of the financial industry, retail industry data mining in the field of data mining, medical and health, sports, such as data mining and data mining on the Internet [4-6].

Data mining is a repeated interaction process, require multiple steps associated with each other. And target according to the requirements of application, as well as the data sources and different meanings, step will have some changes, in general, the data mining process is divided into five stages: data preparation, data selection, data preprocessing, data mining and conversion mode and the evaluation of the model. The basic process of data mining is shown in Figure 1.

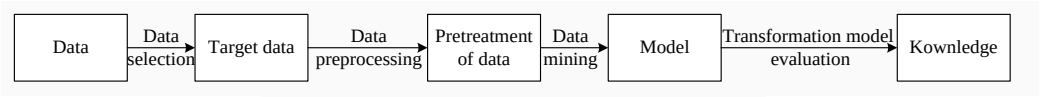

Figure 1. Basic Data Mining Process

2.2. The BP Neural Network Algorithm

Back neural network BP neural network is error, it is a kind of forward network, only to the prequel value process, did not return to the feedback network of neurons in layer, the number of neurons in each layer are not fixed, all neurons are arranged in the input layer and output layer and hidden layer. Connected to the output of the neuron is passed to it the next layer of neurons, the state of the connection weights will transfer control. In addition to the input layer, all other input layer neurons from a layer and its output weighted and related neurons. Neurons are both activation degree by the input, activation functions and threshold is decided by the combination results. The BP neural network can be divided into learning stage and working stage of two major parts. Learning phase of the input information from the input layer, along the hidden layer and transmitted to the output layer. On the transfer process, a layer of information will only be a layer of neurons state, under the influence of information in each layer will be fully processed. If the output and I hope the results don't match, has bigger error and output error are calculated, the reverse error transfer, and achieve the purpose of modifying weights of each layer neurons, as far as possible to the minimum error control. After repeated learning connection weights of each layer neurons are sure down, connection weights will not change during the work. Work, information from the input layer began to spread, and finally results in the output layer. Abnormalities is important in the BP neural network of error back propagation process, the work of the final output has a direct impact, but also indirectly has very important influence on the characteristics of the algorithm. Error calculation using the

objective function is defined as the actual output and usually want to output the error sum of squares of general gradient descent method is used to derive the calculation formula of sometimes also use other methods.

Online learning through such positive output and repeatedly back propagation error calculation, the error of the output will be more and more small, right to the final value converges to a fixed value, end of the learning process, the algorithm can work. During the learning network typically carried out in accordance with the normal working process of the strict training.

The training process of BP network learning algorithm is as follows:

(1) selected from a sample set sample, the sample information input to the network.

(2) network along the positive generated output.

(3) the calculation error of the actual output and the expected output.

(4) by the output layer along the original path reverse calculation, according to the network connection weights to adjust the right value, reduce the error.

(5) repeat the above work, making sample focusing all sample training error specified error precision is satisfied.

At this point, the network connection weights between nodes are fixed, the network is ready to, can work. At this point, you can use the Internet to deal with the unknown sample.

The BP neural network algorithm can be applied in many fields, it should be paid attention to when using the following features:

(1) BP neural network algorithm can be transformed into nonlinear problem, the problem of input, output, cumulative value using the gradient descent method, through calculating the error precision meet the requirements for many times, to join the hidden nodes to make results more accurate, normally according to the using performance usually set hidden layer to layer.

(2) The essay to realize nonlinear mapping of input and output. The BP neural network algorithm can complete the nonlinear mapping of input and output information, if the input number is m, n, number of output information is completed m d to n dimensional space transformation. BP network by superimposing the ordinary nonlinear processing unit, the stronger ability of nonlinear.

(3) global approximation network. f(x) function within the scope of considerable value in the variable x to a non-zero value. This makes the network within the global scope can keep normal work performance.

(4) the generalization ability. In the BP neural network, the general in order to obtain good generalization ability, in addition to the training sample set must exist, also need to test sample set for generalization ability test. In the process of online learning, the general system error will decrease with a lot of training, but sometimes the test sample system error, this shows that the network generalization ability weakened, work will produce unexpected results. In addition, the number of hidden layer and hidden layer node number also has a great influence to the network generalization ability, the general principle is: the network structure is simpler, better generalization ability [7-8].

BP network model is one of the important model of neural network, are widely used in many fields, but it also has many disadvantages: from the Angle of mathematics, a nonlinear optimization problem is sure to bring a local minimum point, and minimum point problems will affect the network performance. In addition, there is also the initial weights set has a significant effect on convergence speed; The algorithm convergence speed slower; Number of nodes in the hidden layer and hidden layer to determine the general selection according to the experience of the designer, there is no formula can rely on.

2.3. Improved BP Neural Network Algorithm

The structure of the BP neural network algorithm to choose the convergence rate of the network and rational utilization of network resources play an important role. In the determination of hidden layer in the determination of network structure is very complicated, if the number of hidden layer neurons is too little, the input vector is not accurate approximation; If the number of hidden layer neurons is too much, can make algorithm is complex, and waste of hardware resources. Usually the number of hidden layer determined by experience, but this method has limitations to solve the problem of complex network, thus the determination of hidden layer structure is very important. Can through the number of hidden layers and the number of hidden layer neurons to determine the hidden layer structure. We assume that the hidden layer is L, each layer of neurons with A_i, I from 1 to L, the weighting coefficient of each neuron is W_{iAi}, on the basis of fuzzy theory we can for each weight coefficient is given a U_{iAi} weighting method, we can according to the fuzzy concept definition of hidden layer structure, namely: H = $\{U_{iAi}/W_{iAi}\}$, according to the different network requirements, we can get to determine whether hidden layer neurons meet the threshold condition of the appropriate threshold.

$$\delta_{iAi} = \begin{Bmatrix} 1, \mu_{iAi} \geq \theta \\ 0, \mu_{iAi} < \theta \end{Bmatrix} \tag{1}$$

θ is the threshold. Algorithm after complete the calculation, the hidden layer is $M = \{\delta_{iAi}\mu_{iAi} / w_{iAi}\}$ accordingly. Usually additional momentum factor is a constant, constant introduction of usually plays a different role in network optimization. If the additional momentum factor is too small, the convergence rate is slow, the stability of the network is poor; If the additional momentum factor is too large, the algorithm excessive emphasis on the former weakened the derivative of the gradient descent of BP network. The change of the trend according to the error adjusting additional momentum, and the specific algorithm is [9-10]:

$$\begin{aligned} &\Delta E > \theta_1 \qquad \alpha(t+1) = \gamma_1 \alpha(t) \\ &\Delta E < \theta_2 \qquad \alpha(t+1) = \gamma_2 \alpha(t) \\ &\theta_2 < \Delta E < \theta_1 \qquad \alpha(t+1) = \alpha(t) \end{aligned} \tag{2}$$

3. The Application of Improved BP Network Algorithm in Data Mining

3.1. The Data Feature Extraction

Feature extraction in data classification system is very important, it directly with the result of classification and the classification of the follow-up work produce very big effect. A set of feature item must has the following two characteristics: (1) completeness, feature to be able to summarize all the data content;2 distinctiveness, feature sets to be able to make data between separate from each other. Tectonic characteristics itemsets in addition to strong generalization ability and profound knowledge, best with linguists according to the principle of extracting feature of common human manually.

3.2. The Structure Characteristics of Itemsets

Suppose you have classified data M group, general construction steps of the feature set according to the following process:

Step 1: to summarize the characteristics of the M set of data items, and each feature in the data to the statistics, the number of occurrences of the result set: the inductive available character D_1, D_2,...And D_M. And then according to the given threshold θ, remove the data, the characteristics of frequency is lower than the θ items available for each set of

data the new characteristics of collection: C_1 and C_2,..., C_M.

$$\begin{aligned} C_1 &=\geq \left\{T_1^1, T_2^1 ..., T_{N1}^1\right\} \\ C_2 &=\geq \left\{T_1^2, T_2^2 ..., T_{N2}^2\right\} \\ &... \\ C_M &=\geq \left\{T_1^M, T_2^M ..., T_{NM}^M\right\} \end{aligned} \tag{3}$$

Step 2: calculate M a collection and set, will agree to word, word escape and synonyms induction for the same feature, available and set: $C = C_1 U C_2 U ... U C_M = \{T_1, T_2, ... T_N\}$, then all data sets: the characteristics of a $\{T_1, T_2, ... T_N\}$.

3.3. The Structure of the Eigenvectors

Feature item set $\{T_1, T_2, ..., T_N\}$, each set of data items feature vector by each group of data in feature determined according to the number of episodes. Also should be specifically to high frequency characteristics of increasing data, specifically to low degree of the characteristics of the frequency decrease data. Specific steps are as follows:

Step 1: statistical characteristic result set $\{T_1, T_2, ..., T_N\}$ characteristics in the M set of data items in the number of occurrences of each group.

Step 2: according to the characteristics of the data generated by the following formula M group vector set $\{f(T_{m1}), f(T_{m2}),$ and $f(T_{mN})\}$, (m = 1, 2,..., M).

$$f(T_{mk}) = V_{mk} \lg(\frac{N}{N_k} + 0.5)(m = 1, 2, ..., M; k = 1, 2, ..., N) \tag{4}$$

Here, the V_{mk} characterization of feature T_k in data m, the number of occurrences of N represents the total number of data, N_k represents a feature of the total number of data Tk.

Step 3: to deal with the above feature vector, to get the final data characteristic vector $T_M = \{T_{m1}, T_{m2}, ..., T_{mN}\}$, (m = 1, 2,..., M).

The model will be the traditional BP neural network in structure, in order to in a variety of identification model can meet the requirements of precision. This model can improve the classification ability of data classification, but also corresponding network structure is complicated. In this article each subnets corresponding to a data classification, already so distracted the network load, and improved the precision of the network.

3.4. Data Partitioning

Now has divided the data set M group category, classification processing according to the following steps:

(1) to extract feature, generate feature vector;

(2)Diction of network parameters initialization: subnet number (h);Subnet number information input values (z);Error range £;Vector for α;Inertia coefficient of β.The cumulative number of repeated c; Most times repeatedly Max;

(3) the hidden layer and the threshold given initial value;

(4) to calculate the final output and classification error E;

(5) if E< £ or c > Max jump (7);

(6)to modify each layer of the weights and thresholds, c = c+1, jump to (4);

(7) get the final results, end of the training.

After training, you can use the parallel BP network for data classification [11-13].

4. The Experimental Results and Analysis

According to the Internet with tourism as the theme of the classification of the web site, tourism website is generally divided into the following eight kinds of conditions: (1) the

tourist attractions;(2) the travel guide;(3) the travel agency;(4) hotels;(5) car rental service,(6) tourist traffic,(7) overseas travel;(8) tourism comprehensive information.

Usually by the recall ratio and precision measuring algorithm, their corresponding calculation formula for:

(1) recall (H_i) = T_n/N, T_n class by the number of data to be correctly classified as H_i; The number of data in the N is originally belong to H_i.

(2) precision (H_i) = T_n/H_n, T_n is properly classified as the number of H_i class of data; H_n to be classified as the number of H_i class of data.

Use of baidu's web site to retrieve the above categories, select some of these pages, form 1500 sample data set. Select 1000 as the training sample set, the rest of 500 as test sample set. After analysis of all web pages and the organization, form feature, 81, 1500 web pages are reasonable coding, so that the algorithm can convenient for subsequent processing. As shown in Table 1 for part of the code [14-15].

Table 1. Web Coding Results

Num	Tourist attractions	The travel guide	Travel information	…	Document category	
					Category name	Category number
1	0.15	1.10	0.35	…	The travel guide	2
2	0.30	0.05	0.15	…	Car rental service	5
3	0.25	0.25	0.15	…	The travel agency	3
4	1.25	0.15	0.25	…	Tourist attractions	1
…	…	…	…	…	…	…
1500	0.20	0.15	0.20		Overseas travel	7

Subnet number is consistent with the classification number, each subnet has a hidden layer; Subnet number input information is summarized the characteristics of the item number; According to the new and improved BP algorithm and the experiment of concrete data can determine the subnet number of hidden layer nodes take 40;Per subnet has an output node; The final total results by binary number, take five nodes. Error precision £=0.05, the learning speed of α=0.15, inertia coefficient β=0.55, the highest repetitions Max = 5500, actual repeat 3860 times of convergence. From the training results can see the recall ratio and precision is high, such as in Table 2.

Table 2. Classification Result Table

category	Subclasses page number	Network identification number	Identify the correct number	Recall ratio (%)	Precision (%)
Tourist attractions	150	155	149	0.99	0.96
The travel guide	150	148	145	0.97	0.98
The travel agency	150	156	142	0.95	0.91
The hotel restaurant	150	152	143	0.95	0.94
Car rental service	150	142	138	0.92	0.97
Tourist traffic	150	145	140	0.93	0.97
Overseas	150	155	147	0.98	0.95

travel					
Tourism comprehensive information	150	147	136	0.91	0.93
Total	1200	1200	1140	0.95	0.95

Next use learn good network to test set of targeted 450 web pages are classified, the test results can be recall ratio and precision are more than 90%, close to comparing with the results of classification of training, indicate that network generalization ability is stronger, classification model has better performance. Based on the same training sample and test sample sets, using the three layers BP network topology, the input information every time 81, output information three at a time. Set the number of hidden layer nodes is 85, the repeat count for 12043 times, the accuracy of the testing sample set is 78%;Set the number of hidden layer nodes is 115, the repeat count for 9546 times, the accuracy of the testing sample set is only 66%.This shows that the new and improved BP algorithm and multiple subnets can effectively deal with the combination of high latitude parallel topology sample classification problem, while the normal BP network will be a slow convergence speed, easy to appear the phenomenon such as fitting, in this paper, the network up to avoid these problems very well.

5. Conclusion

Web data mining technology are introduced in detail in this paper, the semi structured information at the time of the network data mining, processing, and detailed introduction to the data in the process of data mining classification algorithm, the advantages and disadvantages of the main ideas of the classification algorithm and the algorithm, and on the basis of deep understanding of the BP neural network algorithm is combined with the new improved algorithm of BP neural network using multiple subnet topology structure of parallel classifying network data, raises the convergence rate and reduce the size of the classification of the classification error.

References

[1] M. Bramer, "Introduction to Data Mining", Principles of Data Mining. Springer London, **(2013)**, pp. 1-8.
[2] D. T. Larose and C. D. Larose, "An Introduction to Data Mining", Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition. John Wiley & Sons, Inc., **(2014)**, pp. 1-15.
[3] W. G. Touw, J. R. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels and S. A. F. T. van Hijum, "Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?", Briefings in Bioinformatics, vol. 14, no. 3, **(2013)**, pp. 315-326.
[4] X. Wu, X. Zhu and G. Q. Wu, "Data Mining with Big Data", IEEE Transactions on Knowledge & Data Engineering, vol. 26, no. 1, **(2014)**, pp. 97-107.
[5] M. Zorrilla and D. G. Saiz, "A service oriented architecture to provide data mining services for non-expert data miners", Decision Support Systems, vol. 55, no. 1, **(2013)**, pp. 399-411.
[6] B. A. Gionis, "Assessing data mining results via swap randomization", ACM TKDD, **(2013)**.
[7] D. Dutta, W. J. Hsu and G. Dommety, "Network data mining to determine user interest: US", US 8504488 B2, **(2013)**.
[8] D. Dutta, W. J. Hsu and G. Dommety, "Network data mining to determine user interest: US", US 8504488 B2, **(2013)**.
[9] F. Yu and X. Xu, "A short-term load forecasting model of natural gas based on optimized genetic algorithm and improved BP neural network", Applied Energy, vol. 134, no. 134, **(2014)**, pp. 102-113.
[10] S. Dong, D. D. Zhou and W. Zhou, "Research on Network Traffic Identification Based on Improved BP Neural Network", Applied Mathematics & Information Sciences, vol. 7, no. 1, **(2013)**, pp. 389-398.
[11] Y. Wang, C. Lu and C. Zuo, "Coal mine safety production forewarning based on improved BP neural network", International Journal of Mining Science & Technology, vol. 25, no. 2, **(2015)**, pp. 319-324.
[12] K. Li, K. Li and W. Zhang, "PCA Face Recognition Algorithm Based on Improved BP Neural Network", Computer Applications & Software, **(2014)**.
[13] W. Zhu, "Forecasting Railway Freight Volume Based on Improved BP Neural Network Model", Journal of Shijiazhuang Tiedao University, **(2014)**.
[14] M. You and Y. Li, "Automatic classification of the diabetes retina image based on improved BP neural

network", Control Conference (CCC), 2014 33rd Chinese. IEEE, **(2014)**.
[15] Z. Jian, J. Zuo and L. Jia, "Prediction of cement filling materials performance using improved BP neural network", Journal of Environmental Sciences, vol. 30, no. 4, **(2014)**, pp. 207-214.

Authors

Yang Liu, received the Master's degree of Engineering in Computer technology field from College of Computer Science of Chongqing University, China in 2009. She is currently researching on the bacterial foraging particle swarm algorithm, BP neural network algorithm, genetic algorithm, and the improved algorithm of ant colony algorithm in artificial intelligence.

