International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-4,April 2016
ISSN: 2395-3470
www.ijseas.com

ENHANCED SELF ORGANIZING MAP ALGORITHM FOR WEB MINING

C.SADHANA [1], Dr.L.Mary Immaculate Sheela[2]
[1] Research Scholar, Computer Science and Applications St.Peters University, chennai
[2] Professor, Department of Computer Application, R.M.D Engineering College, chennai

Abstract

My approach is automatic that it does not require users explicit input. Moreover, I take a systematic approach to collect and comprehend user activities. I provide a general framework for collecting, mining, and search/query personal usage data, which may be employed by various agents.

Neural based approach is used to analysis the performance of the clustering of the number of request. I propose an approach "ENHANCED SELF ORGANIZTING MAP" which is data visualization technique; it reduces the dimensions of data through the use of neural network. In previous study on SOM plot the similarities of data by grouping similar data items together, so they reduces dimension and display similarities SOM organize sample data, which are usually surrounded by similar samples ,similar samples are not always near each other .In ESOM I use users Clustering mining algorithm. ESOM can estimate the center and the number of clustering data set by" dissimilarity computing", it optimizes SOM neural network learning and improving clustering effect.

Keywords: *query personal, neural network, SOM, ESOM, dissimilarity computing.*

1. Introduction

The data set by dissimilarity computing optimize the Self organization Map (SOM)Neural Network. Enhanced self Organization Map(ESOM) Over comes the difficulty of establishing output nodes and affecting the Clustering effect about dataset.ESOM not need to link the Output node of the weights from input nodes as in SOM..ESOM is not only focus on Log files SOM algorithm extended to Cluster in ESOM on user registration information such as age, gender, income, region etc., ESOM can estimate the center and the number of clustering data set by" dissimilarity computing", it optimizes SOM neural network learning and improve clustering effect. In my proposal I going to design the ESOM which improve the lack of improvements can be well used in web log data mining. Improving the design of personalized business websites has broadened application prospects. ESOM can combine the user registration information such as age, gender, income, region to improve the access time.

2. Self Organizing Map

So far we have looked at networks with supervised techniques, in which there is target output for each input pattern, and the network learns to produce the required outputs. We now turn to unsupervised training, in which the networks learn to form their own classifications of the training data without external help. To do this we have to assume that class membership is broadly defined by the input patterns sharing common features, and that the network will be able to identify those features across the range of input patterns.

One particularly interesting class of unsupervised system is based on competitive learning,in which the output neurons compete amongst themselves to be activated, with the result that only one is activated at any one time. This activated neuron is called winner taken

2.1 Components of Self Organization

The self-organization process involves four major components:

Initialization: All the connection weights are initialized with small random values.

Competition: For each input pattern, the neurons compute their respective values of a *Discriminate function* which provides the basis for competition. The particular neuron with the smallest value of the discriminate function is declared the winner.

Cooperation: The winning neuron determines the spatial location of a topological neighbourhood of excited neurons, thereby providing the basis for cooperation among neighbouring neurons.

Adaptation: The excited neurons decrease their individual values of the discriminate function in relation to the input pattern through suitable adjustment of the associated connection weights, such that the response of the winning neuron to the subsequent application of a similar input pattern is enhanced. *Neuron* or simply the *winning neuron*. Such competition can be induced/implemented by having *lateral inhibition connections* (negative feedback paths) between the neurons.

The result is that the neurons are forced to organise themselves. For obvious reasons, such a network is called a *Self Organizing Map (SOM)*.

3. Overview of the SOM Algorithm

We have a spatially *continuous input space*, in which our input vectors live. The aim is to map from this to a low dimensional spatially *discrete output space*, the topology of which is formed by arranging a set of neurons in a grid. Our SOM provides such a nonlinear transformation called a *feature map*.

The stages of the SOM algorithm can be summarised as follows:

1. *Initialization* – Choose random values for the initial weight vectors w*j*.
2. *Sampling* – Draw a sample training input vector x from the input space.
3. *Matching* – Find the winning neuron *I*(x) with weight vector closest to input vector.
4. *Updating* – Apply the weight update equation D*wji t Tj I t xi wji* =h ( ) , (x)( ) (-).
5. *Continuation* – keep returning to step 2 until the feature map stops changing.

Next lecture we shall explore the properties of the resulting feature map and look at some simple examples of its application.

Dissimilarity/Similarity metric: Similarity is expressed in terms of a distance function, which is typically metric:*d*(*i*, *j*)

There is a separate "quality" function that measures the "goodness" of a cluster.The definitions of distance functions are usually very different for interval-scaled, boolean, categorical, ordinal and ratio variables.It is hard to define "similar enough" or "good enough" the answer is typically highly subjective.

3.1 Similarity and Dissimilarity Between Objects

Distances are normally used to measure the similarity or dissimilarity between two data objects

Some popular ones include: Minkowski distance:

where i = (xi1, xi2, …, xip) and j = (xj1, xj2, …, xjp) are two p-dimensional data objects, and q is a positive integer

If q = 1, d is Manhattan distance

$$d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+\ldots+|x_{ip}-x_{jp}|$$

If q = 2, d is Euclidean distance:

Properties

$d(i,j) \geq 0$

$d(i,i) = 0$

$d(i,j) = d(j,i)$

$d(i,j) \leq d(i,k) + d(k,j)$

Also one can use weighted distance, parametric Pearson product moment correlation, or other dissimilarity measures.

Dissimilarity between Binary Variables

3.2 Example

gender is a symmetric attribute the remaining attributes are asymmetric binary let the values Y and P be set to 1, and the value N be set to 0

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

4. Literature survey

4.1Clustering:

1.Different clustering algorithms including K-Means, Make Density Based, Farthest First, EM and Filtered on Face book 100 university dataset. The year is 2012 and the success rate or the results is According to the results of their experiment average accuracy rates for 10 university set are as follows:
K-means (64%), Make Density assed(55%). Farthest First(60%). EM(60%) and Filtered (57%).

2.An algorithm based on Business System Planning clustering algorithm along with Linked list data structure. In the year 2010 The Success rate is the edges between node is assumed to have the same weight, but in real world it might get changed.

3.K-Mean we can cover more Urls but SOM works better for larger number of cases. With increase in data, learning process of SOM becomes more accurate and we can consider larger number of clusters
.
4.The major problem with SOM is that they are vary computationally expensive which is a major drawback since as the dimensions of the data increases, dimension reduction visualization techniques become more improtant,but unfortunately then time to compute them also increase.

Advantages:

1.SOM are very easy to understand and simple.
2.Classification of data is depends upon the Strong similarities between objects.
3.It Reduces the dimensions of the data through use of SOM neural Networks.

Disadvantages:

1. SOM need a value for each dimension of each member of Samples in order to generate a map.
2. It includes the weight adjust, obtain network training and unsupervised organization learning.
3. SOM is very computationally expensive.
4. Time to calculate the dimension samples increases if dimension increases.
5. If scale of neighbors increases the number of distance the algorithm needs to compute increases exponentially.

5. Conclusion

Above result generated by the SOM network shows that our approach can effectively discover usage pattern. User's browsing behaviour can also predict based on the past history. With increase in data, learning process of SOM becomes more accurate and we can consider larger number of clusters. SOM is also efficient in time as compared to K-Means. For more accurate result we can provide output of K-Means as a input to SOM and use the results of SOM for generating recommendation System and analysis of the Web site.

References

[1] R. Kosala, H. Blockeel, "Web Mining Research: A Survey", SIGKKD Explorations, vol. 2(1), July 2000.

[2] Magdalini Eirinaki , Michalis Vazirgiannis, “Web Mining for Web Personalization”, ACM Transactions on Internet Technology, Vol. 3,No. 1, February 2003.

[3] J. Srivastava, R. Cooley, M. Deshpande, P.-N. Tan, “Web Usage Mining: Discovery And Applications Of Usage Patterns From Web Data”, SIGKKD Explorations, vol.1, Jan 2000.

[4] Vinita Shrivastava, “Web Usage Data Clustering Using Neural Network Learning”, IJRIM Vol. 1, No. 2 , June, 2011.

[5] Navin Kumar Tyagi, A.K. Solanki& Sanjay Tyagi, “An Algorithmic Approach To Data Preprocessing In WebUsage Mining” International Journal of Information Technology and Knowledge Management, Vol.2, No. 2, July-December 2010, pp.: 279-283.

[7] Masseglia, F., Poncelet, P., And Cicchetti, R. (1999). “WebTool: An integrated framework for data mining”, In Proceedings of the Ninth International Conference on Database and Expert Systems Applications (DEXA’99) (Florence, Italy, August1999, pp.: 892–901.