CLUSTERING CST TELEPHONE BILLS USING SOM ALGORITHM

Karma Wangchuk[1], Dorji Phuntsho[2] Basant Subba[3] Tshering[4]

Department of Information Technology, College of Science and Technology
Royal University of Bhutan

[1]2010039@cst.edu.bt, [2]2010020@cst.edu.bt, [3]2010003@cst.edu.bt,
[4]tshering.cst@rub.edu.bt

ABSTRACT

Whenever the system is used, human supervision is required. Self OrganizingMap(SOM) is an un-supervised visualization and analysis tool used for high dimensional data. This study is proposed to study and analyze the voucher expenses incurred by the College of Science and Technology's staffs using SOM Algorithm. The outcome of this study revealed coherent evidence of the range of expenditure incurred by office phones used by individual staff. It will help management to monitor the telephone call expenditure.

Keywords: *SOM, Clustering, Algorithm*

1. Introduction

There are technologies available for reduction of human effort and give the optimum throughputs that gradually makework easier. The Self Organizing Maps (SOM) is one of the system technologies where it's been used asvisualization and analysis tool for high dimensional data. Basically, SOM is used for clustering, dimensionality reduction, classification, sampling, vector quantization and data-mining.

SOM has the same principle of the workings of the brain. The certain part of the brain is responsible for specific skills and tasks which are activated by identification of a certain types of pattern. Similarly, SOM organize information spatially where similar concepts are mapped to adjacent areas and that constitute a trademark of the SOM.

The basic idea of the SOM is to make classification of the available resources based on the certain types of available patterns. In SOM, patterns which are similar will activate similar areas.In this study, multidimensional telephone bill is analyzed using SOM Algorithm.

The sample initialization, Random sampling and dot product are used from Initialization, Sampling and Distance respectively in our experiment of using SOM to cluster real data.

Section two contains brief theoretical review of

Self Organizing Map. In section three under research methodology, system description and dataset are presented. Theexperiment and result are discussed in section four. Finally conclusion is presented in section five.

2. THEORETICAL BACKGROUND

2.1 SELF ORGANIZING MAP

The Professor TeuvoKohonen designed a SOM belonging to an artificial neural network data analysis technique (T.Kohonen, 2001). This technique creates a sheet-like artificial neural network, the cells of which becomes specially tuned to various input signal patterns or classes of patterns through an unsupervised learning process storing information in such a way, that any topological relationships within the training set are maintained(Tsheringet al,2008). SOM is used as visualization and analysis tool for high dimensional data. It is based on unsupervised learning. Some of the SOM applications are:clustering, dimensionality reduction, classification, sampling, vector quantization and data-mining.

The Self Organizing Map is a single layer feed forward network where the output neurons are arranged in low dimensional grid. Each input is connected to all output neurons. Attached to every neuron there is a weight vector with the same dimensionality as the input vectors. The numbers of input dimensions are usually higher than the output grid dimensions as it mainly used for dimensional reduction. Each node has a specific topological position (x, y coordinate in the lattice) and contains vector of weight of the same dimensions as the input vectors. The neurons are usually located in the nodes of the two dimension grid with rectangular or hexagonal cells. The neurons also interact with each other and distance between the neurons on the map lattice governs the degree of the interaction. Let us consider the input vector x of n dimensional space of real numbers, Rn forming SOM in a regular two dimensional grid (array) consisting of p and r columns (array of p x r element).

Each neuron is equipped with weight connection w(i,j) that forms an n-dimensional weight vector connection as shown in (a).

$$\mathbf{w(i,j)=[w_1(i,j), w_2(i,j),\ldots w_n(i,j)]}\quad(a)$$

It computes distance function d(w,x) between its connection and the corresponding input x as shown in(b).

$$\mathbf{y(i,j)=d(w(i,j),x)}\quad(b)$$

where (i,j) denotes certain (i,j) position of the neuron in the array and x is an input to all neurons.Each of the all input x affects all neurons. The neuron with the shortest distance between the input and the connections becomes activated to the highest extent and is called the winning neuron. It coordinates are represented by (i_0,j_0), more precisely as shown in (c)

$$\mathbf{(i_0,j_0)=arg\ min(i,j)d(w(i,j),x)}\quad(c)$$

The winner neuron is rewarded to all allow it to modify the connections so to bring even closer to the output data. The update mechanism is governed by the expression below:

$$\mathbf{w\_new(i_0,j_0)=w(i_0,j_0) + \alpha(x- w(i_0,j_0))}\quad(d)$$

where α denotes the learning rate, $\alpha > 0$.

2.2 SOM ALGORITHM

The Self Organizing Map algorithm is a step by step procedure for which given initial state and end state are defined.

(a) Initialization:

The first and the foremost thing, done before feeding the raw data is initialization. Initialization means to give each weight of the output node a random vector value. The dimensionality of the vector values put into must match the dimensionality of the raw data. When the raw data consist of

five arrays, then the vectors must have five elements. The initialization can be done using one of the methods given below:

(i) Random Initialization
(ii) Sample Initialization
(iii) Linear initialization

The different initialization may well produce quite different results. This is where SOM become a little tricky. A good initialization can be determined experimentally.

(b) Sampling:

The Sampling in SOM involves picking of sample x from input distribution with some probability. It can be also being done through one of the available technique.

(i) Systematic Sampling
(ii) Random Sampling

(c) Distance:

The concept of nonnegative dissimilarity is an essential component of any form of clustering that help navigate through the data space and form cluster. The closeness of the patterns is based on the concept of distance functions with respect to their geometry (RuiXu, 2005). The distance function is chosen from one of the following:

(i) The Dot Product
(ii) The Euclidean Distance
(iii) Hamming Distance
(iv) Tchebyshev
(v) Minkkowski

(d) Similarity matching(BMN).

When n dimensional input vector is introduced to the work, the reference vector in the network, that is closest to the input vectors gets activated to the highest extend and is called the Best Matching Node(BMN) winner neuron.

If the coordinates are denoted by (i_0,j_0), more precisely (e) is used for Eucledian distance and (f) is used for dot product.

$$(i_0,j_0)=\arg \min(i,j)d(w(i,j),x) \quad (e)$$

$$(i_0,j_0)=\arg \max(i,j)d(w(i,j),x) \quad (f)$$

(e) Updating

The winning neuron or the best matching node matches x, because it is the winner of this competition, winner neuron is rewarded and allowed to modify the connection so to become more closer to the input data.

The update mechanism is governed by (g).

$$w\_new(i_0,j_0)=w(i_0,j_0) + \alpha(x- w(i_0,j_0)) \quad (g)$$

where αdenotes a learning rate, $\alpha>0$, the higher the learning rate, the more intensive the updates of the connection becomes.

(f) 6. Iteration

Continue by picking another sample until the map become stable with numbers of iterations. The iterations are specified in advanced or the learning terminates once there are no significant changes in the connection of the neurons (Tshering et al, 2008).

2.3 ARCHITECTURE OF SOM ALGORITHM

The fig 1 represents architecture of SOM algorithm.

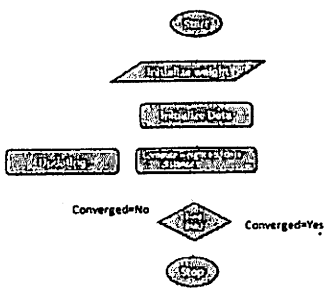

Fig.1 Architecture of SOM

The data are feed into the SOM. The initialization of weight for each data takes place. In the third stage there is a computation of weight and data distance. Then decision is made by matching the best match unit. If it isnot converged, data is sent for updationand if it's converged it terminates.

3. RESEARCH METHODOLOGY

3.1 SYSTEM DESCRIPTION

The proposed system is shown in Fig 2. The data is selected from the dataset randomly. These data are process using SOM toolbox in which the functionalitiesare defined. Then execution or the process starts as given in the SOM algorithm: initialization, sampling, distance, similarity matching, updating and iteration. Finally, the output of the system is displayed.

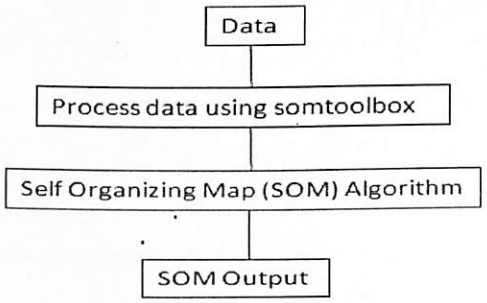

Fig.2 Proposed system

3.2 DATASET

The six months telephone bills from July 2012 to March 2013 are used as dataset. The telephone bills incurred by staffs were compiled on monthly basis asin table 1.

Table 1. Telephone bills

Name	July	Aug	Sept	Oct	Nov	Dec	Jan	Feb	Mar	Total
Store Assistant	528	639	468	283	355	405	526	632	500	4336
Head, IT	133	414	509	1328	1283	1008	1115	1324	1547	8641
Head IR	1268	1476	1534	1416	1265	1200	1669	1267	997	12092
Finance Officer	932	1521	750	1401	860	564	973	801	960	8762
Civil Engineering Dept.	148	135	148	96	234	129	334	386	856	2466
Chief Provost	688	725	1623	511	688	200	815	1121	757	7128
Procurement Manager	409	458	550.4	513	552	852	845	922	946	6047.4
Estate Manager	741	736	493	465	540	277	836	940	874	5902
Head, EED	401	371	404	325	211	231	218	227	506	2894
ICT	146	268	100	154	113	150	681	152	144	1948
Program Leader ECE	100	100	119	156	126	100	100	100	651	1552
Training Manager	509	815	2189	1913	704	1972	705	711	846	10364
Construction Manager	270	287	368	108	337	244	682	587	618	3501
Library	148	165	121	337	261	117	120	163	277	1709
Dean AA	100	100	100	186	145	100	139	104	100	1074
Civil(LRC)	517	878	523	904	695	681	704	348	527	5777
Cheku Dorji Machine Lab	197	100	184	165	115	100	262	101	213	1437

4. EXPERIMENTATION

The staffs' telephone bills from July 2012 to March 2013 were collected from the office of administration. The output from the SOMmapped staffs into high, medium and minimal cluster base on telephone bill amount.

The output from this study helps management to study nature of expenditure incurring from office telephone.

4.1 SOFTWARE PREPARATION

The data is preprocessed arranging in a required format(Fig. 3) and saved as *staff_voucher_data.txt*. A*som.m* file is edited to call data file as:

sDiris =
som_read_data(*'staff_voucher_data.txt');*

The *som.m* file is run. The SOM's Algorithm, execution begins via. Initialization of the data, sampling, distance calculation, similarity matching, updating and iteration and output is generated.

4.2 DATA PREPARATION AND EXECUTION OF ALGORITHM

Staffs were considered as a sample and months as a variable. Therefore, there are 17 samples and 9 variables. The telephone data is given as an input and clustered.

Table. 2 Raw data

Name	July	Aug	Sept	Oct	Nov	Dec	Jan	Feb	Mar	Total	
Store Assistant	528	639	468	283	355	405	526	632	500	4336	A
Head, IT	113	414	509	1328	1283	1008	1115	1324	1547	8641	B
Head IIR	1268	1476	1534	1416	1265	1200	1669	1267	997	12092	C
Finance Officer	932	1521	750	1401	860	564	973	801	960	8762	D
Civil Engineering Dept.	148	135	148	96	234	129	334	386	856	2466	E
Chief Provost	688	725	1623	511	688	200	815	1121	757	7128	F
Procurement Manager	409	458	550.4	513	552	852	845	922	946	6047.4	G
Estate Manager	741	736	493	465	540	277	836	940	874	5902	H
Head, EED	401	371	404	325	211	231	218	227	506	2894	I
ICT	146	268	100	154	113	190	681	152	144	1948	J
Program Leader ECE	100	100	119	156	126	100	100	100	651	1552	K
Training Manager	509	815	2189	1913	704	1972	705	711	846	10364	L
Construction Manager	270	287	368	108	337	244	682	587	618	3501	M
Library	148	165	121	337	261	117	120	163	277	1709	N
Dean AA	100	100	100	186	145	100	139	104	100	1074	O
CMI(LRC)	517	878	523	904	695	681	704	348	527	5777	P
Chelku Dorji Machine Lab	197	100	184	165	115	100	262	101	213	1437	Q

The total amount of the expenditure was calculated and added as new column for sorting in excel. To display output, using lengthy was impossible. For thisreason, capital letters were attached to plot in U matrix Map. The multidimensional data after passing through SOM algorithm produced easy to visualize output in hexagonal 6x18 sizes U Matrix Map.

```
528 639 468 283 355 405 526 632 500 A
113 414 509 1328 1283 1008 1115 1324 1547 B
1268 1476 1534 1416 1265 1200 1669 1267 997 C
932 1521 750 1401 860 564 973 801 960 D
148 135 148 96 234 129 334 386 856 E
688 725 1623 511 688 200 815 1121 757 F
409 458 550.4 513 552 852 845 922 946 G
741 736 493 465 540 277 836 940 874 H
401 371 404 325 211 231 218 227 506 I
146 268 100 154 113 190 681 152 144 J
100 100 119 156 126 100 100 100 651 K
509 815 2189 1913 704 1972 705 711 846 L
270 287 368 108 337 244 682 587 618 M
148 165 121 337 261 117 120 163 277 N
100 100 100 186 145 100 139 104 100 O
517 878 523 904 695 681 704 348 527 P
197 100 184 165 115 100 262 101 213 Q
```

Fig 4preprocessed data

4.3 RESULT INTERPRETATION

The SOM algorithm clustered raw multidimensional data andproducedsignificant output as in Fig. 5.

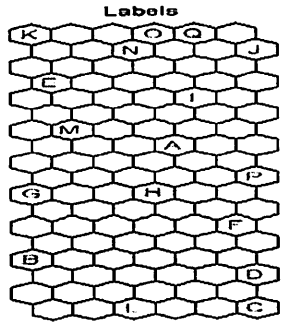

Fig. 5. Output from SOM tool box

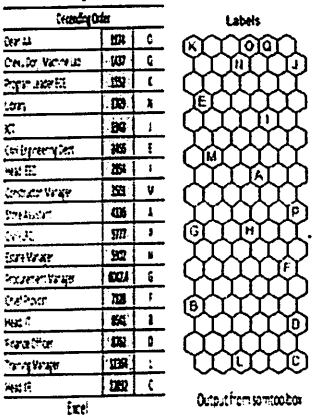

Fig. 6 Validation of SOM output with Excel

Using excel total column is arranged into descending order. When compared with SOM output, the highest expenditure incurred was C with Nu.12092, L with Nu.10364, D with Nu.8762, B with Nu.8641,O with Nu.1074.Same pattern was observed in the SOM output as L,C,D,B, O. thus from this study it is proven that SOM can be used to cluster multidimensional data.

It is seenin the figure that C, D , L and B are in one cluster. Here, Head IIR, Training Manager, Finance Officer and Head IT are in one cluster and they belong to the cluster of staffs who are incurring highest expenditure.

Thus, from the output we are clear that high expenditure were incurred by C, D, L

and B and medium by F,G,H P,A ,I and M and minimal by E,J,N,K,Q and O.
Therefore from data Fig 7 consisting three clusters can be constructed for revaluation of facts.

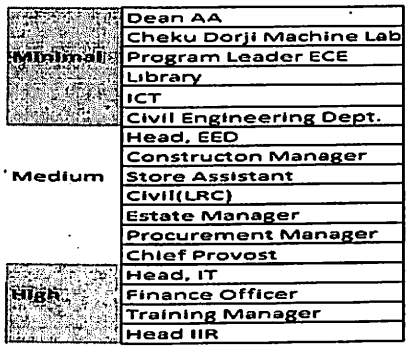

Fig 7 Output Meaning

5. CONCLUSION

In this study, we studied SOM Algorithm and successfully deployed in investigation of telephone bills. From the result, it is concluded that SOM Algorithm can be used in mining and analysis of huge data. In our study, it revealed the statistic correctly.
DechenPelki et al (2012) innovated Hybrid PCAK Algorithm and validated our findings. Their finding drew exactly the same conclusion.

REFERENCE

Abidi, S., &Ong, J.(2000). Automated data clustering based on a synergy between self-organizing neural networks and k-means clustering techniques. Proceedings of IEEE TENCON, 568-573.

Bhavana, S.Kuril, D.C and Tshering 2001;:"a neural network approach to the detection of high impedance fault on distribution feeder"; National Institute of Technology, Waranal, Andra Pradesh, India.

D. Karma, G.Tashi, Y.Lobzang, D. Tshering, and R.Phurpa May 2010; "hybrid ksom approach to high impedance fault" detection in the distribution feeder", College of Science and Technology, Rinchending, Phuentsholing, Chhukha, Bhutan.pp.21-3

Hiroshi,D., &Takeshi,T.(2007).Mapping of the Genome Sequences Using Two stage Self Organizing Maps. 6th International conference on Self Organizing Map(WSOM).

Hollmen, Jaakko 1996; "Self Organizing Map(SOM)"; URL http://www.cis.hut.fi/jhollmen/dippa/node9.html

Hollmen,J.(1996). Process Modelling Using the Self-Organizing Map.Master's thesis, Helsinki University of Technology.

Teknomo, Kardi;2006; "What is K-Mean Clustering?"; URL,http//:people.revoledu.com/kardi/tutorial/kMean/WhatIs.htm

Tshering, Somjit Arc in.(2008); "clustering ict indicators of Bhutan using hybrid self-organizing map:ksom and hsom"; Master thesis, KnonKaenUniversity,Thailand.

DechenPelki, DechenWangmo and Tshering, (2012), "Clustering CST Telephone Bills using Hybrid PCAK Algorithm", ZorigMelong, Vol (1), No (1), 2013.