Mining Census Database With Generalized Self-Organizing Map Algorithm.

Sindhu Nair and Shalini Bhatia

Abstract— **The self-organizing map (SOM) is an unsupervised neural network which projects high-dimensional data onto a low-dimensional grid and visually reveals the topological order of the original data. Thus, SOM is an excellent tool in the exploratory phase of data mining. Self-organizing maps have been successfully applied to many fields, including engineering and business domains. The conventional SOM training algorithm based on Euclidean distance handles only numeric data. Consequently, the trained SOM is unable to reflect the correct topological order in case of categorical data. This paper applies the SOM algorithm by generalizing it for categorical data. This generalized self-organizing map model based on concept hierarchy specifies the similarity between categorical values via distance hierarchies. By measuring the distance in the hierarchies, the semantic relationships between the categorical data is revealed during the visualization. Experiments on real datasets conducted, demonstrate the effectiveness of the generalized SOM model.**

Index Terms— **Data Mining, Cluster Analysis, Self-Organizing Maps, Concept Hierarchy, Categorical Data, Distance hierarchy.**

I. Introduction

DATA mining processes can be divided into six sequential, iterative steps consisting of problem definition, data acquisition, data preprocessing and survey, data modeling, evaluation, and knowledge deployment. Various tools are used for the data survey step in data mining which have prominent visualization properties [7].

Data mining can be classified into two categories: descriptive data mining and predictive data mining . The former describes the data set in a concise and summarized manner and presents interesting general properties of the data. Cluster analysis is a part of descriptive data mining [1].

Cluster analysis has been widely used in numerous applications, including pattern recognition, data analysis, image processing and market research. By clustering, one can identify dense and sparse regions and, therefore, discover overall distribution patterns and interesting correlations among data attributes. In machine learning, clustering is an example of unsupervised learning. For this reason, clustering is a form of learning by observation rather than learning by examples. In conceptual clustering, a group of objects forms a class only if it is describable by a concept. This differs from conventional clustering, which measures similarity based on geometric distance. Conceptual clustering consists of two parts : (a) it discovers the appropriate classes (b) it forms descriptions for each class, as in classification [9].

Kohonen's self-organizing map (SOM) is an unsupervised neural network which projects high-dimensional data onto a low-dimensional grid. The projected data preserves the topological relationship of the original data. Hence, this ordered grid can be used as a convenient visualization surface for showing various features of the training data, for example, cluster structures [6]. The SOM is especially suitable for the data survey step in data mining as it has prominent visualization properties [7].

The conventional SOM training algorithm handles only numeric data since the distance computation to form clusters is based on the Euclidean distance. SOM is unable to process categorical data eg. For student data in a campus database, the department attribute is categorical. For sales transaction in a sales database, the product attribute is categorical while the sales-amount attribute is numeric. By generalizing the SOM model to Generalized Self-Organizing Map (GSOM) categorical data of various applications can be handled effectively for data mining.

Thus, the SOM can handle categorical data and mixed data such that it can process more diverse data and expand the applicability. The applications of SOM include image processing, process monitoring and control, speech recognition, flaw detection in machinery, business and management, information retrieval , medical diagnosis [4], time-series prediction, optimization as well as financial forecasting and management [3].

II. GENERALIZED SELF-ORGANIZING MAP

A. *Self-Organizing Map*

With self-organizing maps (SOMs), clustering is performed by having several units compete for the current object. The unit whose weight vector is closest to the current object becomes the winning or active unit. So as to move even closer to the input object, the weights of the winning unit are adjusted, as well as those of it's nearest neighbours. SOMs

Sindhu Nair is M.E. (Computer Engineering) student of Thadomal Shahani Engineering College, Mumbai, Maharashtra, India (e-mail:sindhu_knair@hotmail.com).

Shalini Bhatia is the Head, Computer Engineering Department, Thadomal Shahani Engineering College, Mumbai, Maharashtra, India (e-mail: shalini.tsec@gmail.com).

assume that there is some topology or ordering among the input objects, and that the units will eventually take on this structure in space. The organization of units is said to form a feature map, SOMs are believed to resemble processing that can occur in the brain and are useful for visualizing high-dimensional data in 2- or 3-D space. In SOM, there is one input layer and one special layer, which produces output values that compete. In effect, multiple outputs are created and the best one is chosen. This extra layer is not technically either a hidden layer or an output layer, but referred as the *competitive layer*. Nodes in this layer are viewed as a two-dimensional grid of nodes. Each input node is connected to each node in this grid. Propagation occurs by sending the input value for each input node to each node in the competitive layer[6].

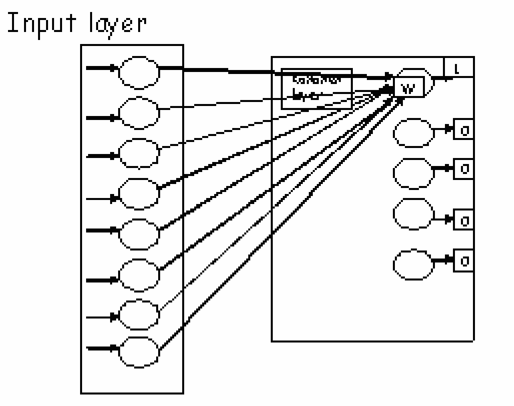

Fig 1. Architecture of SOM

The Kohonen network has two layers, an input layer and a Kohonen output layer. The input layer is a size determined by the user and must match the size of each row (pattern) in the input data file. Input Attributes determine number of input neurons while number of class attributes determine number of output neurons

B. Self-organizing Map training algorithm

Similar data patterns in the input space will be assigned to the same map unit or nearby units on the trained map.

BMU is the unit that is most similar to the input pattern.

Training an SOM using a dataset involves two key steps:

a)Determining the best matching unit (BMU)

b) Updating the BMU and its neighbours [6].

C. Categorical Data

Categorical Data are discrete data. Categorical attributes have a finite number of distinct values, with no ordering among the values. Examples include geographic location, job category and item type [9].

D .Concept Hierarchy

a) Has concept nodes and links where higher-level nodes represent more general concepts while lower-level nodes represent more specific concepts.

b) Extended with link weights: Each link has a weight representing a distance.

c) The distance between 2 concepts at leaf nodes is then defined as the total link weight between those 2 leaf nodes.

d) Links weights assigned by domain experts

E. Distance Hierarchy

a) Facilitates the representation and computation of the distance between categorical values.

b) General distance representation mechanism or both categorical and numerical values.

c) Models several distance computation schemes, including simple matching, binary transformation, and numeric subtraction.

d) Offers a unified platform for measuring the distance between mixed-type, numeric, and categorical data [10].

F. Example of Distance Hierarchy Computation

A point X in a distance hierarchy consists of anchor and offset denoted by X = (N_x, d_x) where anchor is the leaf node and the offset represents the distance from the root of the hierarchy to X. The distance between two points in a distance hierarchy is the total weight between them.

Let X = (N_x, d_x) and Y = (N_y, d_y) be two points, then, the distance between X and Y = $d_x + d_y - 2\, d_{LeastCommonPoint}(x,y)$ where $d_{LeastCommonPoint}(x,y)$ is the distance between the root and the least common point of X and Y [10].

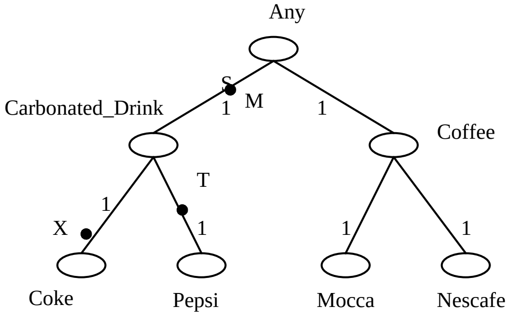

Fig 2 . Distance Hierarchy with link weight 1

In Fig 2,

M = (Pepsi, 0.3). Anchor (M) = Pepsi , Offset (M) = 0.3

Assume X = (Coke,2)
M = (Pepsi,0.3)
S = (Coke,0.3)
T = (Pepsi,1.3)

S = (Coke,0.3) equivalent to M = (Pepsi,0.3)

Distance between T,M is (1.3 + 0.3 – 2 * dist(LeastCommonPoint (T,M)) = (1.3 + 0.3 – 2 * 0.3) = 1

Distance between X,T is (2 + 1.3 – 2 * dist(LeastCommonPoint (X,T)) = (2 + 1.3 – 2 * 1) = 1.3

LeastCommonAncestor (T,M) is M

LeastCommonPoint (X,T) is a point at Carbonated Drink.

LeastCommonPoint (M,S) is M or S

LeastCommonPoint (M,X) is M

The distance between two data patterns is defined according to the mapping points in their associated distance hierarchies. All attribute values of two patterns are mapped to their hierarchies, and then, the distances between correspondent mapping points are aggregated.

A special type of distance hierarchy, called the numeric distance hierarchy for a numeric attribute is a degenerated one which consists of only two nodes and a link as shown in Fig 3.The two nodes are the root labeled by MIN and the leaf labeled by MAX. The only link has the weight w that is equal to the range of the numeric attribute.Used in binary encoding approach by associating each new binary attribute with a numeric distance hierarchy. Degenerated distance hierarchy with weight w = ($\max_A - \min_A$) for a numeric attribute A.

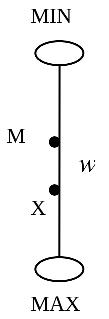

Fig 3. Numeric Distance Hierarchy

GSOM Training Algorithm

The GSOM training algorithm is as follows :

```
Input:  an n-dimensional training dataset D,
      a GSOM Model ( contains a map of n-dimensional units )
      a set of n distance hierarchies DH = { dh1,dh2,….,dhn}

Output : a trained GSOM

1.Initialize each unit m of the GSOM
2.For each pattern x in D,
        2.1 Identify its best matching unit from the GSOM.
              • Map the components of x and m to their distance hierarchies
              • Aggregate the distances of the mapping points in the hierarchies
              • Identify m that gives the minimum distance , d , to x as the best matching unit
        2.2 Adjust/update the BMU and it's neighbours
              • Calculate the adjustment by multiplying d with a learning rate and a neighbourhood function.
    Repeat till stop criteria met.
```

Once the distance between a training pattern and the BMU is determined, the adjustment amounts of the BMU and its neighbours are computed by multiplying the distance by a learning rate and a neighbourhood function. Then, the BMU and it's neighbours are adjusted by their respective adjustment amount such that the adjusted units become closer toward the training pattern. The adjusting of each component m_i of a GSOM unit m toward it's corresponding attribute x_i of a training pattern x is accomplished by moving the mapping point M of m_i in dh_i towards the mapping point X of x_i. Note that in the context of training a GSOM, it is always a point adjusted toward the other point that is located at a leaf node, because the value of a categorical attribute of a training pattern is always mapped to a leaf node.

Referring to Fig 2, during the adjusting phase, x_i is the adjusting aim of m_i. In terms of the points in their associated hierarchy, the mapping points, say X and M , of x_i and m_i form the adjustment path where M moves toward X along the path during adjusting. Let X be the mapping point which M and T move toward , the anchors of X, M *and* T be C(Coke), P(Pepsi) and P(Pepsi) respectively and the adjustment amount be δ and N_{LCA} be the least common ancestor of C and P.

N_{LCA} of C and P is the Carbonated_Drink.

Case 1 : If M is an ancestor of N_{LCA} and it does not cross N_{LCA} after adjustment , then the new M is ($P, d_M + \delta$) where d_M is the offset of M to the root.

Case 2 : If M is an ancestor of N_{LCA} and it crosses N_{LCA} after adjustment , then the new M is ($C, d_M + \delta$)

Note that whenever the adjusted point crosses it's least common ancestor N_{LCA} the point changes it's anchor to the anchor of the other point that it moves toward.

Case 3 : If N_{LCA} is an ancestor of T and T does not cross N_{LCA} after adjustment , then the new T is ($P, d_T - \delta$)

Case 4 : If N_{LCA} is an ancestor of T and T crosses N_{LCA} after adjustment , then the new T is ($C, 2d_{NLCA} - dT + \delta$) .

III. DESIGN FOR CENSUS DATASET

A. Distance hierarchies for the Real dataset.

For the clustering of the Real dataset using the GSOM Model, the categorical attributes considered are: Marital_status, Relationship, Education. So, the distance hierarchies have to be constructed for each of these categorical attributes. Hence, we begin with the design of the distance hierarchies of the categorical attributes of the real dataset.

The distance hierarchies for the categorical attribute Relationship is constructed. The root node is 'Any'. The different values for the attribute Relationship in the dataset are 'Husband', 'Wife', 'Not_in_family', 'Own_child', 'Unmarried', 'Other_relative'. The distance from the root node to the leaf node is assumed to be 1. (Ref fig 4)

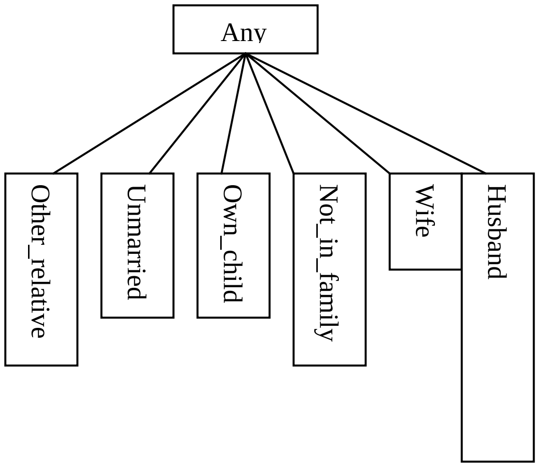

Fig 4 Distance hierarchies for *Relationship attribute* of the UCI Adult dataset

In the distance hierarchy for the categorical attribute Marital_status, the nodes ‘Single’ and ‘Couple’ are at a distance of 1 from the root node ‘Any’ while the actual values stored in the dataset are at a distance of 2 from the root node. The leaf nodes contain the actual values that are stored in the dataset. Thus, each of the distinct values for the categorical attributes becomes a leaf node in the distance hierarchies. (Ref fig 5)

Married_civ_spouse means that the person is married whose spouse is a civilian.Married_AF_spouse means that the person is married whose spouse is in the Armed Forces. The semantics of all other leaf nodes are as per their meaning in English.

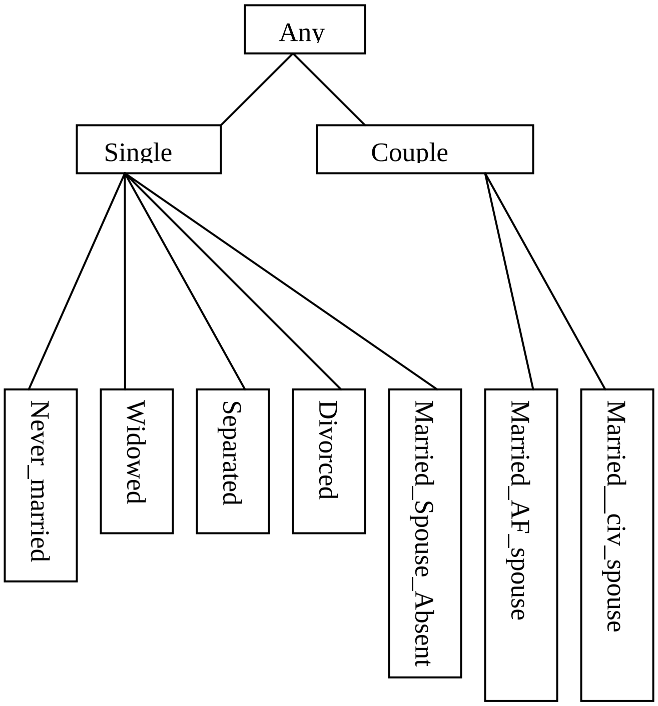

Fig 5.Distance hierarchies for *Marital_status attribute*_of the UCI Adult dataset

Similarly, in the distance hierarchy for the categorical attribute Education, the nodes created at a distance of 1 from the root node are ‘Little’, ‘Junior’, ‘HighSchool’, ‘College’,’Advanced’ while the actual values stored in the dataset are at a distance of 2 from the root node.

The leaf nodes contain the actual values that are stored in the dataset. Thus, each of the distinct values for the categorical attributes becomes a leaf node in the distance hierarchies. The leaf node contains the actual educational qualification.

The first step is the implementation of the distance hierarchies for each of the categorical attributes of the real dataset. The distance hierarchies for categorical attributes Education, Marital_status, and Relationship is implemented as text files which are used in the matlab training program. Input GSOM Model is the map model with initial weights at Epoch = 0 (that is, first training cycle). Offset which represents the distance from the root to the data points in the respective distance hierarchies are values ranging from 0 to 2.

Since the map-size is 400, and manually 7 groups have been identified for the Real dataset, that is, 7 clusters are to be formed. The function ‘rand’ in matlab gives a value between 0 and 1.

Using the rand function, a matrix of 400 rows (corresponding to map-size 400) and 7 columns (corresponding to seven groups or clusters of the real dataset) is created. The value at each position in the matrix is an offset value between 0 and 2. Such a matrix is assigned as the initial offset matrix for each of the categorical attribute in the real dataset. The weight values can be mapped to points in the distance hierarchies representing the distance from the root to the mapped data points

Here, three initial GSOM Models are created; one for each of the categorical attributes of the real dataset, that is , Education, Marital_status, and Relationship.

Thus, the tables below show that for Real dataset, we build the GSOM Model with Map-size = 400 unit. Offset varies from 0 to 2.

IV RESULTS AND DISCUSSIONS

A. .Output Display after the Real Census Dataset is trained using the GSOM Model

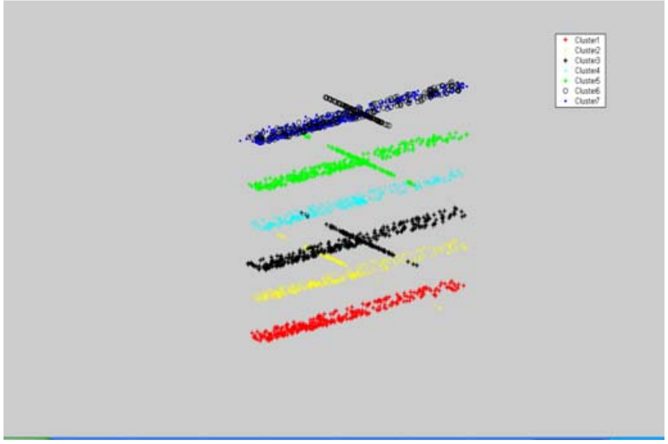

Fig 6 Output

TABLE 1
SALARY DISTRIBUTION IN INDIVIDUAL CLUSTERS GROUPED AS PER TRAINED GSOM

Cluster	>50K(%)	<=50K(%)
6	54.6539	45.3461
5	29.8551	70.1449
4	25	75
7	8.2569	91.7431
3	7.8471	92.1529
1	7.0175	92.9825
2	2.4775	97.5225
All	24.24	75.76

The seven clusters are formed manually such that in the first cluster the value of marital_status is ‘single’ or ‘couple’ and education is ‘little’ or ‘junior’.

In the second cluster the value of marital_status is 'single' and education is 'highSchool' In the third cluster the value of marital_status is 'single' and education is 'college'
In the fourth cluster the value of marital_status is 'single' and education is 'advanced'
In the fifth cluster the value of marital_status is 'couple' and education is 'highSchool'
In the sixth cluster the value of marital_status is 'couple' and education is 'college'
In the seventh cluster the value of marital_status is 'couple' and education is 'advanced'

As we have displayed the percentage of tuples > 50k and the percentage of tuples < 50k in each cluster from the real census dataset, in the above table, the clusters are arranged in the increasing order of percentage > 50k.
Hence, Cluster 6 with % (> 50 k) is 54.6539 and hence tops the table. It is followed by cluster 5 with % (> 50 k) is 29.8551. Cluster 2 has the least % (> 50 k) of 2.4775 and so is placed at the bottom of the table.
As we can see in the clustering of the GSOM Trained map, all the clusters with similar
% (> 50 k) are grouped together or lie in proximity to each other.

Therefore, cluster 6 with % (> 50 k) is 54.6539, cluster 5 with % (> 50 k) is 29.8551, cluster 4 with % (> 50 k) is 25 appear in proximity to each other in the trained GSOM Map. They are all positioned at the top of the trained GSOM Map.
Similarly, cluster 3, cluster 2 and cluster 1 are placed at the bottom of the GSOM Trained Map. The % (> 50 k) of cluster 3 is 7.8471, % (> 50 k) of cluster 1 is 7.0175, and % (> 50 k) of cluster 2 is 2.4775. All are these clusters have similar % (> 50 k) and hence are in proximity to each other.

We observe that cluster 1, cluster 2 and cluster 3 have education attribute as 'little', ' junior', ' highschool' which means that the education level is below average. Also the marital_status is single. Hence, as the education level is low, and they have no additional spouse income, the percentage of >50k in these clusters is low and thus justified.

Also cluster 6, cluster 5 and cluster 4 have education attribute as 'college', ' advanced', which means that the education level is above average. Also the marital_status is couple. Hence, as the education level is high, and they have additional spouse income, the percentage of >50k in these clusters is higher and thus justified.

V CONCLUSION

Experiments on real datasets have been conducted, and the results demonstrated the effectiveness of the generalized SOM model.

Categorical datasets from UCI Repository of ML Databases[Online] used in the experiments show that using the GSOM model the topological order of the input data is maintained for categorical dataset.
Thus, the applicability of SOM in data mining is expanded as it can process more diverse data.

VI REFERENCES

[1] J.Han and M.Kamber , Data Mining : Concepts and Techniques. San Mateo, CA: Morgan Kauffmann, 2001.

[2] C.J.Merz and P.Murphy (1996) UCI Repository of ML Databases.[Online].Available: http://www.cs.uci.edu/~mlearn/MLRepository.html

[3] D.R.Chen , R.F.Chang and Y.L.Huang," Breast Cancer diagnosis using self-organizing maps for sonography," Ultrasound Med. Boil., vol. 1 , no 26, pp. 4411, 2000

[4]J.Vesanto, E.Alhoniemi, J.Himberg, K.Kiviluoto and J.Parviainen, "Self-organizing map for data mining in Matlab: The SOM Toolbox," Simulation News Europe,vol 25, no. 54,1999

[5] T.Kohonen, "The self-organizing map," Proc IEEE, vol. 78, no 9, pp 1464 – 1480, Sep 1990.

[6] Juha Vesanto , Esa Alhoniemi, "Clustering of the Self-Organizing Map", vol. 11, no 3, May 2000

[7] S.Kaski, J.Sinkkonen, and J.Peltonen, "Bankruptcy analysis with self-organizing maps in learning metrics," IEEE Trans. Neural Netw., vol 12,no. 4, pp. 936-947, Jul 2001.

[8] Chung-Chian Hsu, "Generalizing Self-Organizing Map for Categorical Data", IEEE Trans. Neural Netw., vol. 17,no. 2,March 2006.

VII BIOGRAPHIES

Shalini Bhatia was born on August 08, 1971. She received the B.E. degree in Computer Engineering from Sri Sant Gajanan Maharaj College of Engineering, Amravati University, Shegaon, Maharashtra, India in 1993, M.E. degree in Computer Engineering from Thadomal Shahani Engineering College, Mumbai, Maharashtra, India in 2003.
She has been associated with Thadomal Shahani Engineering College since 1995, where she has worked as Lecturer in Computer Engineering Department from Jan 1995 to Dec 2004 and as Assistant Professor from Dec 2004 to Dec 2005. Since Jan 2006 she is looking after the department as the Head. Her research interests include neural networks, fuzzy systems, bioinformatics, intelligent systems, distributed computing, image processing, and advanced computer architecture. She has published a number of technical papers in National and International Conferences. She is an active member of CSI and also a member of Special Interest Group in Artificial Intelligenge (SIGAI) which is a part of CSI.

Sindhu S. Nair was born on September 5, 1975. She received the B E (Computers) from Padmabhushan Vasantdada Patil College of Engineering, Sion under the Mumbai University in 1997. Currently, she is pursuing an M E (Computers) from Thadomal Shahani College of Engineering, Bandra (W), Mumbai.
She worked as a Software Engineer at Seepz, Andheri from 1997 – 2001. Before joining the Information Technology Department at St. Francis Institute of Technology, Borivli, she worked as a Lecturer in the Computer Department at Thakur College of Engineering, Kandivli. Her areas of teaching and research interest are Advance Databases, Data Warehousing and Mining and Neural Networks.