Active Citizenship by Knowledge Management & Innovation
19–21 June 2013 · Zadar, Croatia
make learn
Management, Knowledge and Learning International Conference 2013

KNOWLEDGE DISCOVERY FROM MIXED DATA BY ARTIFICIAL NEURAL NETWORK WITH UNSUPERVISED LEARNING

Chung-Chian Hsu Chien-Hao Kung
Department of Information Management
National Yunlin University of Science and Technology
Yunlin, Taiwan
hsucc@yuntech.edu.tw

Abstract:

Knowledge discovery or data mining from massive data is a hot issue in business and academia in recent years. Real-world data are usually of mixed-type, consisting of categorical and numeric attributes. Mining knowledge from massive, mixed data is challenge. To explore unknown data, visualized analysis allows users to gain some initial understanding regarding the data and to prepare for further analysis. Self-organizing map (SOM) has been commonly used as a visualized analysis tool due to its capability of reflecting topological order of the high-dimensional data in a low-dimensional space. Interesting patterns can thus be discovered by visual clues, possibly leading to discovery of valuable knowledge. In previous studies, an extended SOM has been proposed to visualize mixed-type data. However, the model works under the setting of supervised learning in order to measure the similarity between categorical values. In this article, we propose a model which can work under the setting of unsupervised learning so that neither class attribute nor domain expert is required. Experimental results are reported to demonstrate effectiveness of the proposed approach.

Keywords: Information technology; Knowledge discovery; Self-organizing map; Visualization; Unsupervised learning.

1. INTRODUCTION

Big data commonly existing in modern corporations may contain valuable knowledge or hidden pattern. Analyzing big data to unveil the patterns is a hot topic nowadays. However, analyzing massive amount of data is not an easy task. Not only is the volume of the data huge but also real-world data usually consist of different types of attributes such as categorical and numeric attributes. Most of algorithms handle only one type of values. When mixed data are encountered, a preprocess transforming one type of the data to the other is performed prior to using the algorithms. A typical method to transform a categorical attribute is 1-of-k coding which converts a categorical attribute to a set of binary attributes. The binary attribute corresponding to the categorical value is set to one and the others zero. The 1-of-k coding scheme has one disadvantage: Semantics embedded in categorical values is lost after transformation. For instance, the set of binary values does not reflect the semantics that a drink Coke is more similar to Pepsi than to Latte.

Self-organizing map (SOM) is a popular neural network which has been applied to visualized clustering analysis (Vesanto & Alhoniemi, 2000). The model can project high-dimensional data to a low-dimensional space, typically, a two dimensional one. Consequently, the high-dimensional data become visible and can be analyzed on the two dimensional map. Moreover, when projecting data to a low-dimensional space, SOM can preserve the topological order in the data. That is, data close to one another in the data space are also near to one another on the map. However, when mixed-type data are encountered and 1-of-k is resorted to convert the data, topological order shown on the projection map will be distorted due to the loss of semantics after the transformation.

To address the issue, we proposed a generalized SOM model or GSOM (Hsu, 2006) which not only allows processing mixed-type data directly but also can preserve the semantics embedded in categorical values. In the extended model, we exploited a data structure distance hierarchy, consisting of nodes, links and weights, to represent the relationship between values. A distance hierarchy is a tree. Each link is associated with a weight representing the distance between a parent and a child node. The distance between two values is measured by the total weight of the path. A categorical value is represented as a leaf in the tree. It is easy to see that distance hierarchy can be used to reflect semantic similarity between categorical values. A categorical value is more similar to another value in the same subtree than to a value not in the subtree. Consequently, by incorporating distance hierarchy into the SOM, the extended GSOM is able to process categorical values and reveal proper topological order on the map as well.

Nevertheless, distance hierarchies used in GSOM and the other mixed-type SOMs (Hsu & Lin, 2012; Tai & Hsu, 2012) were constructed in a supervised manner, i.e., by domain experts or requiring existence of a class attribute in the data. In the case of presence of a class attribute, the similarity between two categorical values can be measured based on co-occurrence extent between categorical values and class labels. If the two values are associated with the class labels in a similar way, the values are deemed similar. According to the idea, pairwise distance of categorical values in a categorical attribute can be measured and then distance hierarchy for that attribute can be constructed by an agglomerative clustering algorithm. However, in a real-world dataset, there may be no class attribute.

The study intends to address the issue so that even domain experts are not available and a class attribute does not exist in the dataset. We can construct distance hierarchies for representing similarity between two categorical values. We propose an approach to integrating an unsupervised learning scheme of distance hierarchy with the extended SOM.

2. PRELIMERARY

2.1. Transformation by 1-of-*k* Coding

1-of-*k* coding is a method which transforms a categorical value into a set of k binary, numeric values. The attribute which corresponds to the categorical value is set to one and otherwise zero. Picture 1 illustrates a simple example. As can be seen, the semantics that Green Tea is more similar to Oolong Tea than to Latte is lost after the transformation if Euclidean distance is used to measure the similarity between transactions. Among those four binary attributes, any two of the four transactions are different to each other in two attributes.

Picture 1: (a) A simplified mixed-type dataset, and (b) the transformed dataset by using 1-of-k coding.

ID	Drink type	Price	Qnt	Class
1	Green Tea	30	10	T
2	Black Tea	25	10	T
3	Latte	60	10	C
4	Cappuccino	50	5	C

(a)

ID	Green Tea	Black Tea	Latte	Cappuccino	Price	Qnt	C
1	1	0	0	0	30	10	T
2	0	1	0	0	25	10	T
3	0	0	1	0	60	10	C
4	0	0	0	1	50	5	C

(b)

2.2. Co-occurrence between Feature and Class Label

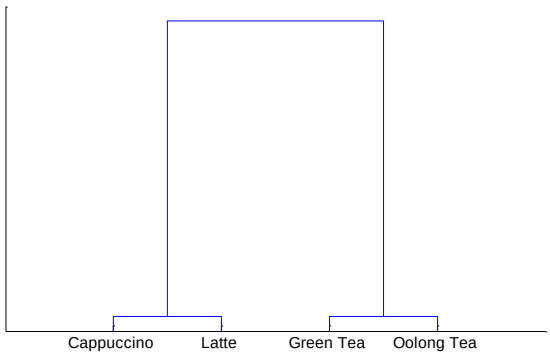
In (Hsu & Lin, 2012), we used a supervised approach, referred to as COFC in this study, which exploits co-occurrence information between feature values and class label. COFC defines the distance between yi and yj in a categorical attribute Y as follows:

$$d(y_i, y_j) = \frac{1}{|C|}\sum_{c \in C}\left|conf(y_i \Rightarrow c) - conf(y_j \Rightarrow c)\right| \quad (1)$$

where c is a class label in the class attribute C and $\mathrm{conf}(y_i \Rightarrow c)$ denotes the ratio of co-occurrence of y_i and c with respect to y_i.

After acquiring the pairwise distance matrix of values of a categorical attribute, distance hierarchy can be constructed by using the popular agglomerative hierarchical clustering algorithm. The tree structure distance hierarchy can be used to represent the distance relationship of a categorical value with the others by the path length. As shown in Picture 2, Green Tea is closer (more similar) to Oolong Tea than to Latte.

Picture 2: An example of distance hierarchy constructed from a pairwise distance matrix by using hierarchical clustering.



3. METHODOLOGY

The method which we are going to propose is unsupervised, i.e., does not require existence of class attribute, unlike COFC which requires the presence of class attribute in the data. The idea of the approach is depicted as follows.

Assume A and B are two distinct values in a categorical attribute. A and B are deemed to be similar or have a small distance if the situation of A co-occurs with the values in the other feature attributes is very similar to the situation of B co-occurs with those values,

3.1. Distance between Categorical Values

We were inspired by the method DILCA or DIstance Learning for Categorical Attributes (Ienco, Pensa, Meo, 2012), a robust algorithm for computing the distance between any pair of values of a specific categorical attribute with respect to other attributes, referred to as context attributes. However, in DILCA, the authors considered only categorical attributes as context attributes. When a numeric attribute presents, the attribute is discretized before the algorithm is applied. In this study, we devise a new formula which takes into account both categorical and numeric attributes.

Our unsupervised approach, referred to as DCUL, to calculating the distance between two categorical values in a target attribute Y uses the information of context attributes X of Y. A context attribute can be categorical or numeric. For a categorical context attribute $X_c \in X$, the distance between y_i and y_j in Y is dependent on the difference of conditional probabilities of y_i and y_j given x_k in X_c. For a numeric context attribute X_n, the distance is dependent on the difference between the averages of the values in X_n which co-occurs with y_i and y_j, respectively. With consideration of all context attributes, the distance between y_i and y_j is defined as follows.

$$d(y_i, y_j) = \frac{1}{|X|}\left(\sum_{X_c \in X}\sqrt{\frac{\sum_{x_k \in X_c}\left(P(y_i|x_k) - \mathrm{P}(y_j|x_k)\right)^2}{|\mathrm{X_c}|}} + \sum_{X_n \in \mathrm{X}}\frac{\left|Avg(X_{n,j}) - Avg(X_{n,i})\right|}{Max(X_n) - Min(X_n)}\right) \quad (2)$$

where $Avg(X_{n,*})$ is the average which comes from the numeric values of co-occurrence of y_* in context attribute X_n and we use the deviation between the maximum and the minimum value in X_n to normalize the value to [0,1].

3.2. An illustrative Example

Three approaches, 1-of-*k* coding, COFC and DCUL for measuring the distance between categorical values have been presented. To illustrate the calculation of the distance, we use an example on a simple synthetic dataset. Table 1 is a toy dataset for illustration. By 1-of-*k*, A1 will be transformed to a list of two binary attributes, namely, <a, b>. The value of A1 of the first transaction is therefore transformed to <1, 0> while that of the fourth <0, 1> accordingly.

Table 1: A SIMPLE ILLUSTRATIVE DATASET.

A1	A2	A3	Class
a	50	L	Yes
a	20	M	No
a	80	H	No
b	60	N	Yes
b	65	M	Yes
b	90	H	No

The algorithm COFC measures dissimilarity between two categorical values of a categorical attribute with respect to the class attribute by using Eq. (1). Accordingly, the distance between **a** and **b** in **A1** is $d(\mathrm{a},\mathrm{b}) = \left(\frac{1}{2}\right)(|\mathrm{conf(a} \Rightarrow 1) - \mathrm{conf(b} \Rightarrow 1)| + |\mathrm{conf(a} \Rightarrow 2) - \mathrm{conf(b} \Rightarrow 2)|) = \frac{1}{2}\left(\left|\frac{2}{3} - \frac{1}{3}\right| + \left|\frac{1}{3} - \frac{2}{3}\right|\right) = 0.33.$

The proposed DCUL considers context attributes instead as shown in Eq. (2). Accordingly, the distance between **a** and **b** is therefore $d(\mathrm{a},\mathrm{b}) = \frac{1}{2}\left(\sqrt{\frac{D_3}{4}} + \frac{|D_2|}{90-20}\right)$. Note that D_2 and D_3 are the context attributes, **A2** and **A3**, in the example. $D_2 = \left|Avg(D_{2,a}) - Avg(D_{2,b})\right| = |50 - 71.66| = 21.66$, and $D_3 = \left(\frac{a}{H} - \frac{b}{H}\right) + \left(\frac{a}{M} - \frac{b}{M}\right) + \left(\frac{a}{L} - \frac{b}{L}\right) + \left(\frac{a}{N} - \frac{b}{N}\right) = 0 + 0 + 1 + 1 = 2$. Therefore, the distance is $d(\mathrm{a},\mathrm{b}) = \frac{1}{2}\left(\sqrt{\frac{2}{4}} + \frac{|21.66|}{90-20}\right) = 0.51.$

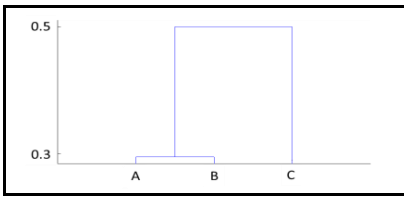
3.3. Transfer the Hierarchies to CPM Matrice

In addition to representing the distance between two categorical values, distance hierarchy facilitates update of categorical component of a neuron with a mixed-type prototype. During the update, the position of the parent node of two leaf values in the hierarchy is required (Hsu & Lin, 2012). To ease the update, a matrix CPM recording the distance between the root and the parent-node position is generated. The entries in the diagonal are set to 0.5. Picture 3(a) and (b) illustrate the process. Picture 3(c) shows the CPM matrix where the value of (A, C) is zero indicating the parent of A and C is the root. The hierarchy is normalized so that the distance between the root and each leaf is 0.5.

Picture 3: (a) a pairwise distance matrix, (b) distance hierarchy constructed from (a) by agglomerative hierarchical clustering, and (c) the CPM matrix from the hierarchy in (b).

	A	B	C
A	0	0.3	0.4
B	0.3	0	0.5
C	0.4	0.5	0

(a)



(b)

	A	B	C
A	0.5	0.2	0
B	0.2	0.5	0
C	0	0	0.5

(c)

3.4. Evaluation Measures

We use two measures, *mean square error* (MSE) and *entropy*, to evaluate the projection results of GMixSOM with different approaches to handling categorical values. MSE measures the average distance between the input instance and the weight of its best match unit and is defined as follows.

$$\text{MSE} = \frac{\sum_{i=1}^{|X|} ndist(x_i, BMU_i)^2}{|X|} \tag{3}$$

where $|X|$ is the number of input data and x_i is an instance of X. BMU_i is the prototype of the neuron into which x_i is projected. The *ndist*(.,.) is the normalized distance between the two arguments. The distance is normalized by the number of attributes to allow performance comparison among different approaches. Note that 1-of-*k* coding increases the number of attributes.

Entropy measures consistence of class labels of instances projected in neurons and is defined by the weighted average of entropies of individual neurons.

$$\text{Entropy} = \frac{|X_n|}{|X|} \sum_{n=1}^{N} Entropy(n) \tag{4}$$

where $|X_n|$ is the number of data projected into neuron *n*.

4. EXPERIMENTAL RESULTS

We evaluate validity of the proposed approach by using DCUL to construct distance hierarchies for use of GMixSOM. Using the MSE, QE, and entropy measures, we compare projection results by GMixSOM with the 1-of-*k* coding scheme and with using distance hierarchies constructed by COFC.

4.1. Synthetic Dataset

The two numeric attributes of the four-attribute synthetic dataset which has 1600 instances were generated by Gaussian functions with designated means and standard deviations.

Picture 4 shows the two distance hierarchies of attributes Dept and Drink generated by the hierarchical clustering algorithm taking as input of the distance matrix produced by DCUL. When running the unsupervised DCUL, the class attribute was excluded and only the other three were considered context attributes. The result reflects the characteristics of categorical values in the dataset. For example, the drinks of the same type, say fruit juices, were grouped together in the hierarchy.

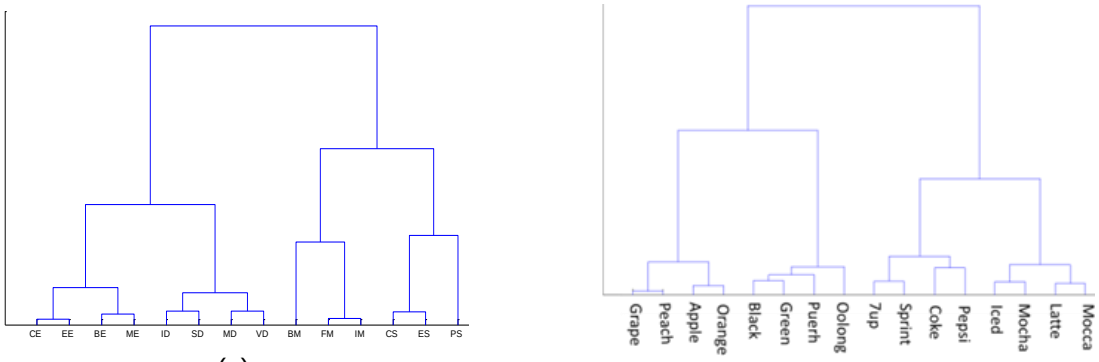
The supervised COFC computes the distance with respect to the class attribute. COFC also clustered the categorical values in each of Dept and Drink to four groups. The distance between any two values in each group, e.g., EE and CE, is zero since the confidence of the class value with respect to given EE and CE is the same. In contrast, the distance between values in different groups, e.g., EE and VD, is 1.

Table 2: A synthetic mixed-type dataset Drink.

Dept	*Drink*	*N1(μ, σ)*	*N2(μ, σ)*	*Class*	*Count*
EE CE ME BE	Mocca Latte Iced Mocha	(90, 10)	(600, 50)	E	400
VD MD SD ID	Coke Pepsi 7Up Sprint	(70, 8)	(800, 40)	D	400
BM IM FM	AppleJuice OrangeJuice PeachJuice GrapeJuice	(40, 5)	(300, 10)	M	400
PS CS ES	OolongTea GreenTea Black Tea PuerhTea	(10, 2)	(100, 5)	S	400

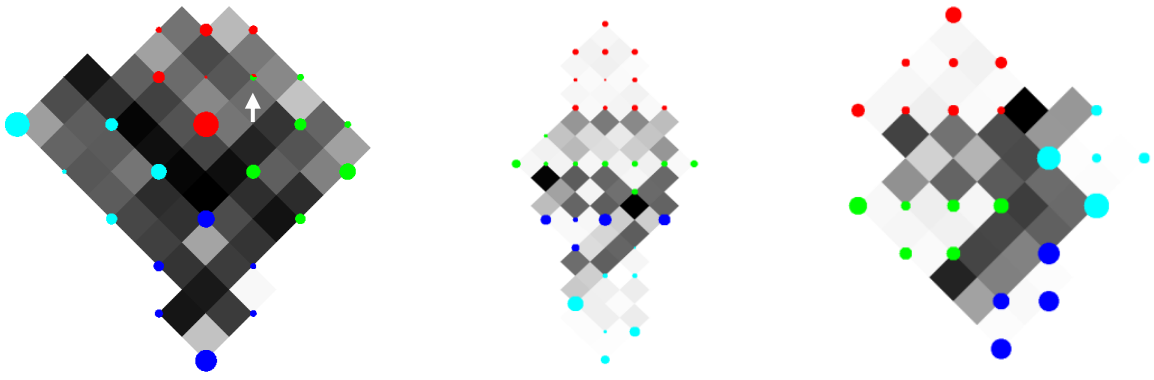
Picture 5 shows the projection result by GMixSOM. Each color corresponds to one class. The size of spots is proportional to the number of instances projected to the neuron. The background grey level indicates the distance between the prototypes of two neurons: the darker, the more distant. As can be seen, the projections in Picture 5(b) and (c) have clearer boundary between different classes compared to that of Picture 5(a). Moreover, one neuron in the upper region of Picture 5(a), indicated by a white arrow, contains instances from different classes, i.e., the red and the green. This is not seen in Pictures 5(b) and (c) which were generated with using distance hierarchies for categorical attributes.

Picture 4: Distance hierarchies of (a) the Dept. attribute and (b) the Drink attribute of dataset Drink constructed by DCUL



(a) (b)

Picture 5: The projection result of the synthetic dataset by GMixSOM with (a) 1-of-*k* coding, (b) DCUL-DH and (c) COFC-DH.



(a) (b) (c)

4.2. Real Datasets

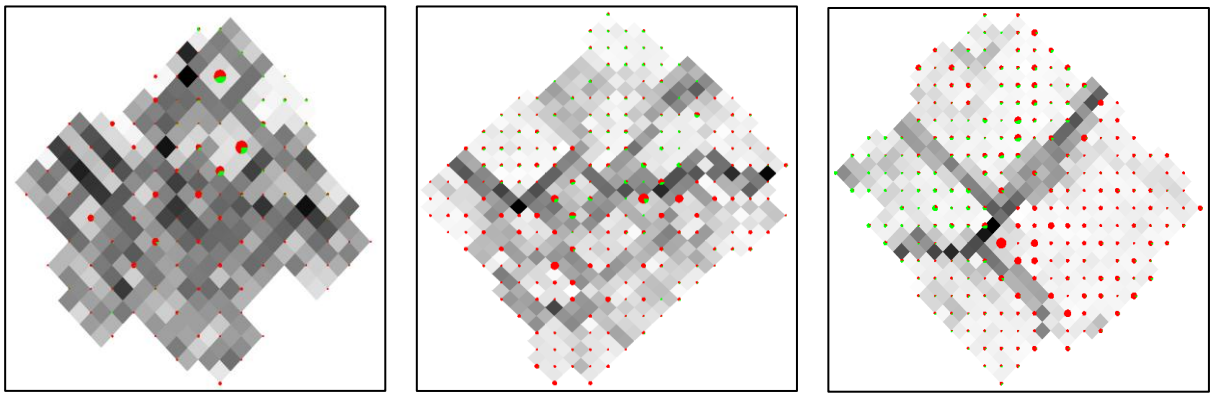
Three datasets, Adult, Australian Credit Approval (ACP), and Contraceptive Method Choice (CMC), from the UCI repository (Blake & Merz, 1998) are used. In GMixSOM, parameters are set according to the suggestion in the SOM toolbox of Matlab. Note that the learning rate decreased linearly, $\alpha(t) = \alpha(0) \times (1.0 - t/T)$ where t and T denote the current iteration and the total training iterations, respectively, and $\alpha(0)$ denoting the initial learning rate which was set to 0.95. The summary of the three datasets is shown in Table 3. Data Adult originally has 14 feature attributes. We retained the seven attributes used in (Hsu, 2006; Hsu & Lin, 2012).

The projection results are shown in Picture 6. By inspecting background colors, Picture 6(a) by GMixSOM with 1-of-*k* shows most obscure boundary. It is not easy to visually analyze the map. Picture 6(b) and (c) by GMixSOM with DCUL-DH and COFC-DH, respectively, show clearer boundary.

Table3: The three real-world mixed-type datasets for the UCI dataset repository.

Dataset	Data Points	Categorical Attributes	Numerical Attributes	Total Attributes	Class
Adult	48,842	3	4	7	2
ACP	690	8	6	14	2
CMC	1,473	4	5	9	3

Picture 6: The projection result of dataset Adult by GMixSOM with (a) 1-of-*k* coding, (b) DCUL-DH and (c) COFC-DH.



(a) (b) (c)

Table 4: Comparison between different similarity algorithms on real-world datasets.

Dataset	Measure	1 of k	DCUL	COFC
Adult	MSE	0.008	0.006	**0.002**
	Entropy	0.560	0.529	**0.509**
ACP	MSE	0.053	**0.035**	0.045
	Entropy	0.554	0.489	**0.480**
CMC	MSE	0.060	**0.044**	0.046
	Entropy	1.416	1.405	**1.392**

In Picture 6(b), there is a boundary in the middle and most of the instances with class label salary > 50K, indicated by the green color, are located in the upper half of the map. In Picture 6(c), the map can be roughly divided to four regions. The instances with salary > 50K are mainly located in the upper-left, especially, the left-most region.

Table 4 shows the result from COFC has the lowest entropy since the distance hierarchies were constructed with respect to class attribute in a supervised manner. The results from the unsupervised

DCUL outperformed those from 1-of-*k* and are inferior to those from COFC by small differences. As to MSE, DCUL is superior to 1-of-*k* in all three datasets and to COFC in two datasets except for Adult.

5. CONCLUSIONS

We proposed to integrate an unsupervised approach to constructing distance hierarchies for representing the distance between two categorical values. Our approach alleviates the shortcoming of the previous models GSOM and the like which require the existence of class attribute in the dataset or domain experts. The experimental results demonstrate that the performance of the new model is superior to that by using 1-of-*k* coding and comparable to that by using the supervised approach COFC.

ACKNOWLEDGEMENT

This research is supported by National Science Council, Taiwan under grant NSC 100-2410-H-224-003-MY2.

REFERENCE LIST

Blake, C.L. & Merz, C.J. (1998). UCI Repository of Machine Learning Datasets. Dept. Inform. Comput. Sci., University of California, Irvien. Available: http://www.ics.uci.edu/~mlearn/ML-Repository.html.

Hsu, C.-C. (2006). Generalizing self-organizing map for categorical data. *IEEE Transactions on Neural Networks,* 17(2), 294-304.

Hsu, C.-C. & Lin, S.-H. (2012). Visualized Analysis of Mixed Numeric and Categorical Data via Extended Self-Organizing Map. *IEEE Transactions on Neural Networks and Learning Systems,* 23(1), 72-86,.

Tai, W.-S. & Hsu, C.-C. (2012). Growing Self-Organizing Map with cross insert for mixed-type data clustering. *Applied Soft Computing,*12, 2856-2866.

Ienco, D., Pensa, R. G., & Meo, R. (2012). From Context to Distance: Learning Dissimilarity for Categorical Data Clustering. *ACM Transactions on Knowledge Discovery from Data,*6(1), 25 pages,.

Vesanto, J. and Alhoniemi, E. (2000). Clustering of the Self-Organizing Map. *IEEE Transactions on Neural Networks,* 11(3), 586-600.