Clustering Hierarchical Data Using Self-Organizing Map: A Graph-Theoretical Approach

Argyris Argyrou

HANKEN School of Economics
Arkadiankatu 22, 00101 - Helsinki, Finland
argyris.argyrou@hanken.fi

Abstract. The application of Self-Organizing Map (SOM) to hierarchical data remains an open issue, because such data lack inherent quantitative information. Past studies have suggested binary encoding and Generalizing SOM as techniques that transform hierarchical data into numerical attributes. Based on graph theory, this paper puts forward a novel approach that processes hierarchical data into a numerical representation for SOM-based clustering. The paper validates the proposed graph-theoretical approach via complexity theory and experiments on real-life data. The results suggest that the graph-theoretical approach has lower algorithmic complexity than Generalizing SOM, and can yield SOM having significantly higher cluster validity than binary encoding does. Thus, the graph-theoretical approach can form a data-preprocessing step that extends SOM to the domain of hierarchical data.

Keywords: Clustering, hierarchical data, SOM, graph theory.

1 Introduction

The Self-Organizing Map (SOM) [1] represents a type of artificial neural network that is based on unsupervised learning; it has been applied extensively in the areas of dimensionality reduction, data visualization, and clustering [2]. The original formulation of SOM uses the Euclidean distance as a similarity metric [3, p.4], and hence its domain of application is restricted to metric spaces [4]. SOM has been extended to non-metric spaces by using generalized means and medians as the distance measures and the batch variant of SOM [4]; for example, speech recognition [5], and clustering of protein sequences [6]. An online algorithm for SOM of symbol strings was provided by [7]. However, neither a metric distance nor a string metric (e.g. Levenshtein distance) can yield meaningful results in the domain of hierarchical data, and thus the application of SOM in this domain remains an open issue. For example, consider clustering the data: {cat, rat, mouse}. A string metric would find that {cat} and {rat} are more closely related to each other than {rat} and {mouse} are, while a metric distance would produce meaningless results.

J.C. Príncipe and R. Miikkulainen (Eds.): WSOM 2009, LNCS 5629, pp. 19–27, 2009.

To address this issue, prior studies have suggested two main techniques that transform hierarchical attributes into numerical attributes. First, the most prevalent technique encodes a categorical attribute in binary terms $\{1,0\}$, where 1 and 0 denote the presence and absence of an attribute respectively. The binary encoding is then treated as a numerical attribute in the range $\{1,0\}$. Second, Hsu [8] introduced Generalizing SOM (GSOM), whereby a domain expert describes a set of categorical data by means of a concept hierarchy, and then extends it to a distance hierarchy in order to represent and calculate distances between the categorical data. However, both techniques suffer from theoretical and practical limitations.

Motivated by this open issue, the paper puts forward a graph-theoretical approach that processes hierarchical data into a numerical representation, and thus renders them amenable for clustering using SOM. To elaborate, based on graph theory, the paper encodes a set of hierarchical data in the form of a rooted and ordered tree. The root vertex represents the complete set of the hierarchical data, and each vertex represents a sub-set of its "parent" vertex. An edge between a pair of vertices is assigned a weight, which can be any positive real number, representing the distance between the two vertices. Thus, the distance between a pair of vertices, v_i and v_j, is the sum of the weighted-edges that exist in the path from v_i to v_j. The paper uses a level-order traversal algorithm to calculate the distances between each vertex and all other vertices. This process yields a symmetric distance matrix $D = (d_{ij})_{nn}$, where n is the number of vertices, and d_{ij} the distance between v_i and v_j.

In the present case, the paper encodes the animals that are contained in the zoo-dataset [9] in the form of a rooted and ordered tree, and calculates the distances between all pairs of animals by using a level-order traversal of the tree, as shown in Fig. 1. The symmetric distance matrix $D = (d_{ij})_{nn}$ thus derived forms the numerical representation of the zoo-dataset, where $n = 98$ reflecting the number of animals, and d_{ij} denotes the distance between a pair of animals. The distance metric d_{ij} satisfies the conditions of a metric space, as follows [10, p.65]: (i) $d_{ij} \geq 0$, (ii) $d_{ij} = 0$ if and only if $i = j$, (iii) $d_{ij} = d_{ji}$, and (iv) $d_{iz} \leq d_{ij} + d_{jz}$. Each row in D represents an animal, and becomes an input vector – $\boldsymbol{x}_j \in \mathbb{R}^{98}$, $j = 1, 2, \ldots 98$ – to SOM.[1]

The paper trains two SOMs, batch and sequence, for each of the two representations of the zoo-dataset, original binary encoding and paper's graph-theoretical approach. For each of the four combinations, the paper selects one hundred samples by using bootstrap; and for each of the 400 bootstrapped samples, it trains a SOM with a Gaussian neighborhood and an 8 x 5 hexagonal lattice. The paper evaluates the quality of each SOM in terms of: (i) the entropy of clustering, (ii) quantization error, (iii) topographic error, and (iv) the Davies-Bouldin index. Based on these quality measures, the paper uses the Wilcoxon rank-sum test at the one-tailed 5% significance level to assess whether the graph-theoretical

[1] The distance matrix D is symmetric, and hence the number of observations (i.e. animals) is equal to the number of dimensions (i.e. 98), and selecting either rows or columns as input vectors to SOM would yield the same result.

approach can yield significantly better SOM than binary encoding does. Further, the paper compares the algorithmic complexity of the graph-theoretical approach with that of GSOM.

The results suggest that the graph-theoretical approach enjoys a lower algorithmic complexity than Generalizing SOM does, and can yield SOM having significantly higher cluster validity than binary encoding does.

The paper's novelty and contribution lie in the formulation of the graph-theoretical approach, and its application as a data-preprocessing step that can extend SOM to the domain of hierarchical data.

The paper proceeds as follows. Section 2 describes briefly the SOM algorithm, binary encoding, and Generalizing SOM. Section 3 formulates the graph-theoretical approach. Section 4 outlines the design of experiments, and section 5 presents and discusses the results. Section 6 presents the conclusions.

2 Background and Related Work

2.1 The SOM Algorithm

In the context of this study, the SOM algorithm performs a non-linear projection of the probability density function of the 98-dimensional input space to an 8 x 5 2-dimensional hexagonal lattice. A neuron i, $i = 1, 2, \ldots 40$, is represented by XY coordinates on the lattice, and by a codevector, $\boldsymbol{m}_i \in \mathbb{R}^{98}$, in the input space. The formation of a SOM involves three processes [11, p.447]: (i) competition, (ii) co-operation, and (iii) adaptation. First, each input vector, $\boldsymbol{x} \in \mathbb{R}^{98}$, is compared with all codevectors, $\boldsymbol{m}_i \in \mathbb{R}^{98}$, and the best match in terms of the smallest Euclidean distance, $\| \boldsymbol{x} - \boldsymbol{m}_i \|$, is mapped onto neuron i, which is termed the best-matching unit (BMU):

$$BMU = \underset{i}{\operatorname{argmin}} \{ \| \boldsymbol{x} - \boldsymbol{m}_i \| \} \quad . \tag{1}$$

In the co-operation process, the BMU locates the center of the neighborhood kernel h_{ci}:

$$h_{ci} = a(t) \cdot \exp \left[- \frac{\| r_c - r_i \|^2}{2\sigma^2(t)} \right] \quad . \tag{2}$$

where r_c, $r_i \in \mathbb{R}^2$ are the radius of BMU and node i respectively, t denotes discrete time, $a(t)$ is a learning rate, and $\sigma(t)$ defines the width of the kernel; $a(t)$ and $\sigma(t)$ are monotonically decreasing functions of time [3, p.5].

In the adaptive process, the sequence-training SOM updates the BMU codevector as follows:

$$\boldsymbol{m}_i(t+1) = \boldsymbol{m}_i(t) + h_{ci}(t) \left[\boldsymbol{x}(t) - \boldsymbol{m}_i(t) \right] \quad . \tag{3}$$

The batch-training SOM estimates the BMU according to (1), but updates the BMU codevector as [12, p.9]:

$$\boldsymbol{m}_i(t+1) = \frac{\sum_{j=1}^{n} h_{ci}(t) \boldsymbol{x}_j}{\sum_{j=1}^{n} h_{ci}(t)} \quad . \tag{4}$$

To carry out the experiments, the paper uses both sequence-training (3) and batch-training (4) SOM.

2.2 Binary Encoding and Generalizing SOM

Binary encoding converts a categorical variable into a numerical representation consisting of values in the range $\{1, 0\}$, where 1 and 0 denote the presence and absence of an attribute respectively. The binary encoding of each categorical datum is then treated as a numerical attribute for SOM-based clustering.

To overcome the limitations associated with binary encoding, Hsu [8] introduced Generalizing SOM (GSOM). Briefly, a domain expert extends a concept hierarchy, which describes a data domain, to a distance hierarchy by associating a weight for each link on the former. The weight represents the distance between the root and a node of a distance hierarchy. For example, a point X in distance hierarchy $dh(X)$ is described by $X = (N_X, d_X)$, where N_X is a leaf node and d_X is the distance from the root to point X. The distance between points X and Y is defined as follows:

$$\mid X - Y \mid = d_X + d_Y - 2d_{LCP(X,Y)} \ . \tag{5}$$

where $d_{LCP(X,Y)}$ is the distance between the root and the least common point of X and Y.

3 The Graph-Theoretical Approach

3.1 Preliminaries

A comprehensive review of graph theory lies beyond the scope of this paper; a textbook account on this subject can be found in [10]. For the purposes of this paper, it suffices to define a tree as a special type of graph, $G = (V, E, w)$, that satisfies at least two of the following three necessary and sufficient properties: (i) G is acyclic, (ii) G is connected, and (iii) $\mid E \mid = \mid V \mid -1$; any two of these properties imply the third [10, p.8]. Let $T = (V, E, w)$ be a tree that is: (i) rooted, with v_0 the root vertex, and (ii) ordered, which means that there is a

Table 1. Notations and definitions

$G = (V, E, w)$	A graph
$V = \{v_1, v_2, \ldots v_n\}$	Set of vertices
$E = \{e_1, e_2, \ldots e_m\}$	Set of edges
$w : E \rightarrow \mathbb{R}^+$	Function assigning a positive real number to an edge
$\mid V \mid$	Degree of graph, cardinality of V
$\mid E \mid$	Order of graph, cardinality of E
$e = \{v_i, v_j\}$	Edge connecting vertices v_i and v_j
$d_{ij} = w(e)$	Distance between v_i and v_j
$D = (d_{ij})_{nn}$	Distance matrix

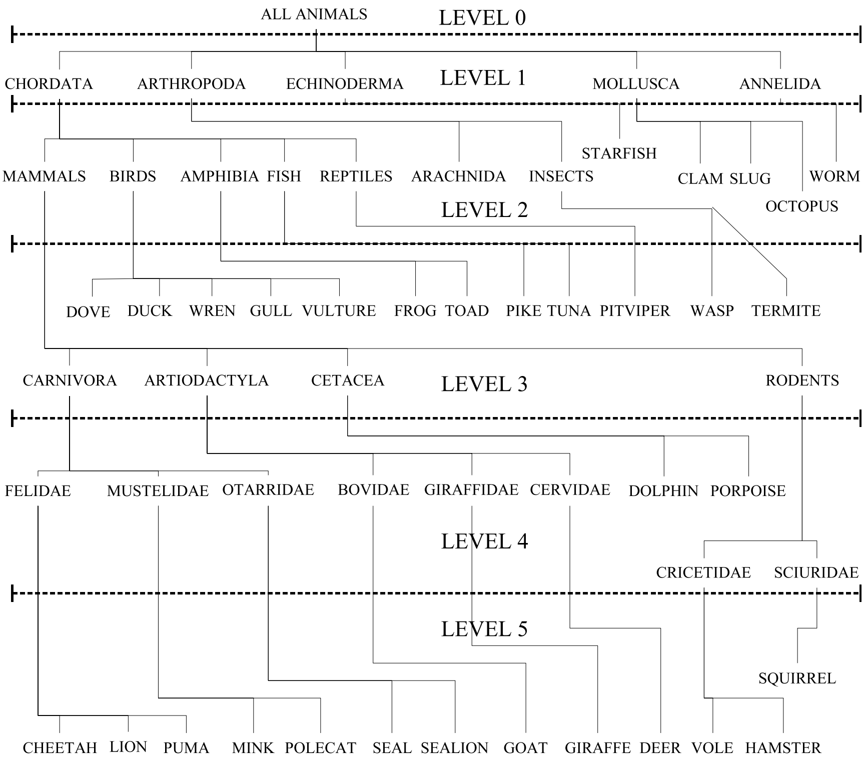

Fig. 1. Extract from the graph-theoretical representation of the zoo-dataset

linear ordering of its vertices such that for each edge $e = \{v_i, v_j\}$ then $v_i < v_j$. It can be easily deduced that in tree T: (i) all vertices excluding v_0 have at most one "parent" vertex, (ii) at least one vertex has no "child" vertices, and (iii) there is a unique path between any two vertices. A tree can be traversed in a level-order way; such a traversal starts from the root vertex, v_0, and proceeds from left-to-right to visit each vertex at distance d from v_0 before it visits any vertex at distance $d + 1$, as shown in Fig. 1.

3.2 Description

The graph-theoretical approach is motivated by the observation that hierarchical variables have a set of states that can be ranked in a meaningful order. For example, consider the variable "size" having five states: {very big, big, medium, small, very small}. It is obvious that {very big} matches {big} more closely than it matches {very small}. However, this piece of information is lost if binary encoding is used, because such an encoding produces a dichotomous output: a state either matches another state or does not.

The graph-theoretical approach operates in three phases. First, it encodes a set of hierarchical data in the form of a rooted and ordered tree. The root vertex represents the complete set of hierarchical data, and all other vertices are ordered in such a way that each vertex represents a sub-set of its "parent" vertex. The edges indicate the covering relation between the vertices. For example, consider a finite order set P; $x, y \in P$; $T = (V, E, w)$; and $v_x, v_y \in V$ correspond to x and y respectively. If x is covered by y (i.e. $x \prec y$), then v_x is a "child" vertex of v_y. Each edge is assigned a weight, which can be any positive real number (i.e. $w : E \rightarrow \mathbb{R}^+$).

Second, the graph-theoretical approach traverses the tree in a level-order manner in order to calculate the distances between the root vertex and all other vertices. The distance between the root vertex v_o and a vertex v_i is the sum of the weighted-edges that exist in the unique path between v_o and v_i. This calculation has an algorithmic complexity of $O\left(\mid V \mid\right)$. To calculate the distances for all pairs of vertices, the graph-theoretical approach designates each vertex as the root vertex and repeats the level-order traversal. Thus, the all-pairs distances can be obtained in $O\left(\mid V \mid^2\right)$. This process yields a symmetric distance matrix $D = \left(d_{ij}\right)_{nn}$, where d_{ij} denotes the distance between vertex v_i and vertex v_j, $d_{ij} > 0$ for all $i \neq j$, $d_{ij} = 0$ if and only if $i = j$, $d_{ij} = d_{ji}$, and $d_{iz} \leq d_{ij} + d_{jz}$.

Finally, the distance matrix D constitutes the numerical representation of the set of hierarchical data and each of its rows becomes an input vector to SOM.

4 Data and Experiments

The design of experiments consists of six steps. First, the zoo-dataset [9] contains 101 animals that are described by one numerical attribute and 15 binary attributes, and classified into seven groups. The paper eliminates the instances "girl" and "vampire" for obvious but unrelated reasons, and one instance of "frog", because it appears twice.

Second, to apply the graph-theoretical approach to the zoo-dataset, the paper uses none of the original attributes. Instead, it uses a "natural" taxonomy that classifies animals based on their "phylum", "class", and "family". This taxonomy can be expressed as a tree (Fig. 1), where the root vertex stands for the complete set of animals. For the experiments, the weight for each edge is set to 1 (i.e. $w : E \rightarrow 1$), though it can be any positive real number and different for each edge. The paper calculates the distances of all pairs of vertices by using a level-order traversal of the tree, and thus derives a distance matrix that makes up the numerical representation of the zoo-dataset.

Third, for each representation of the zoo-dataset, original binary encoding and the paper's graph-theoretical approach, the paper uses bootstrap to draw one hundred random samples with replacement. Fourth, for each bootstrapped sample, the paper trains two SOMs, batch and sequence, with a Guassian neighborhood and an 8 x 5 hexagonal lattice. Fifth, the paper evaluates each SOM in terms of four quality measures: (i) the entropy of clustering, (ii) quantization error, (iii) topographic error, and (iv) the Davies-Bouldin index. Sixth, based on the quality measures, the paper uses the Wilcoxon rank-sum test at the one-tailed 5% significance level to assess whether the graph-theoretical approach can yield significantly better SOMs than binary encoding does.

Table 2. Wilcoxon rank-sum test

SOM-Training	H(Z)	QE	TE	DBI
Batch	A<B	A<B	N.S	A<B
Sequence	A<B	A<B	N.S	A<B

Further, the paper compares the algorithmic complexity of the proposed graph-theoretical approach with that of Generalizing SOM [8]. An experimental comparison was not possible, because GSOM was not available.[2]

4.1 Quality Measures

The quantization error, QE, and topographic error, TE, have been extensively reviewed in the literature pertinent to SOM. Thus, this section concentrates on two cluster validity indices: (i) the Davies-Bouldin index, and (ii) the entropy of clustering.

The Davies-Bouldin index [13], DBI, is defined as:

$$DBI = \frac{1}{C}\sum_{i=1}^{C} \max_{i \neq j} \left\{ \frac{\Delta(C_i) + \Delta(C_j)}{\delta(C_i, C_j)} \right\} \quad . \tag{6}$$

where C is the number of clusters produced by SOM, $\delta(C_i, C_j)$, and $\Delta(C_i)$ and $\Delta(C_j)$ the intercluster and intracluster distances respectively.

Following [14], the entropy of clustering Z, H(Z), can be defined as:

$$H(Z) = -\sum_{j=1}^{C} \frac{m_j}{m} \sum_{i=1}^{K} \frac{m_{ij}}{m_j} log_2 \frac{m_{ij}}{m_j} \quad . \tag{7}$$

where C is the number of clusters produced by SOM, $K = 7$, the number of groups of animals in the zoo-dataset, m_{ij} is the number of animals in group i that are clustered by SOM in cluster j, m_j is the size of cluster j, and m is the size of all clusters.

5 Results and Discussion

The results (Table 2) suggest that the graph-theoretical approach yields SOMs having statistically significant lower entropy of clustering, quantization error, and Davies-Bouldin index than binary encoding does. In contrast, the difference in topographic error is not significant. Further, the results are invariant to the two SOM-training algorithms, batch and sequence.

Referring to Table 2, A and B stand for the graph-theoretical approach and binary encoding respectively, $A < B$ denotes that the difference between the

[2] Personal correspondence with the author of Generalizing SOM.

two approaches is statistically significant at the one-tailed 5% significance level, whereas N.S implies that a significant difference does not exist.

To compare the algorithmic complexity of the graph-theoretical approach with that of GSOM [8], the paper assumes that GSOM is applied to the zoo-dataset, and that GSOM uses this paper's tree (Fig. 1) as its distance hierarchy. As discussed in Sect. 2.2, GSOM entails the following three tasks: (i) calculate distances from the root to all nodes, a level-order traversal of the tree has $O(\mid V \mid)$ complexity; (ii) find the all-pairs least common point (LCP), the current fastest algorithm has $O\left(\mid V \mid^{2.575}\right)$ complexity [15]; and (iii) calculate distances from the root to all LCPs, this takes $O\left(l\right)$, where l is the number of LCPs.

Therefore, the algorithmic complexity of GSOM is $O\left(\mid V \mid^{2.575}\right)$, and hence higher than the quadratic complexity, $O\left(\mid V \mid^{2}\right)$, of the graph-theoretical approach.

5.1 Critique

The proposed graph-theoretical approach is not impervious to criticism. Like binary encoding, it increases the dimensionality of the input space in direct proportion to the number of states a hierarchical variable has. In turn, the dimensionality of the search space increases exponentially with the dimensionality of the input space, a phenomenon aptly named "the curse of dimensionality" [16, p.160]. Further, it assumes that the hierarchical data are static, and hence a deterministic approach is sufficient. To deal with this limitation, future research may explore a probabilistic variant of the graph-theoretical approach.

6 Conclusions

The paper's novelty and contribution lie in the development and application of a data-preprocessing step that is based on graph theory and can extend SOM to the domain of hierarchical data. The results suggest that the proposed graph-theoretical approach has lower algorithmic complexity than Generalizing SOM, and can yield SOM having significantly higher cluster validity than binary encoding does. Further, the graph-theoretical approach is not confined only to SOM, but instead it can be used by any algorithm (e.g. k-means) to process hierarchical data into a numerical representation. Future research may consider a probabilistic variant of the graph-theoretical approach as well as its application in the area of hierarchical clustering. Notwithstanding its limitations, the paper presents the first attempt that uses graph theory to process hierarchical data into a numerical representation for SOM-based clustering.

Acknowledgments and Dedication. I am much indebted to Mr. Tom Linström, Prof. Anders Tallberg, Dr. Andriy Andreev, and Dr. Sofronis K. Clerides for their insightful comments and suggestions. This paper is dedicated to my first mentor, the late Mr. Marios Christou.

References

1. Kohonen, T.: Self-Organizing Maps, 2nd edn. Springer Series in Information Sciences, vol. 30. Springer, Heidelberg (1997)
2. Vesanto, J.: Data Exploration Process Based on the Self-Organizing Map. Doctoral dissertation, Helsinki University of Technology, Espoo, Finland (May 2002)
3. Kohonen, T., Hynninen, J., Kangas, J., Laaksonen, J.: Som-pak: The self-organizing map program package. Technical Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland (1996)
4. Kohonen, T., Somervuo, P.: Self-organizing maps of symbol strings. Neurocomputing 21(1-3), 19–30 (1998)
5. Kohonen, T., Somervuo, P.: Self-organizing maps of symbol strings with application to speech recognition. In: Proceedings of the First International Workshop on Self-Organizing Maps (WSOM 1997), pp. 2–7 (1997)
6. Kohonen, T., Somervuo, P.: How to make large self-organizing maps for nonvectorial data. Neural Networks 15(8-9), 945–952 (2002)
7. Somervuo, P.J.: Online algorithm for the self-organizing map of symbol strings. Neural Networks 17(8-9), 1231–1239 (2004)
8. Hsu, C.C.: Generalizing self-organizing map for categorical data. IEEE Transactions on Neural Networks 17(2), 294–304 (2006)
9. Asuncion, A., Newman, D.: UCI Machine Learning Repository. School of Information and Computer Sciences, University of California, Irvine (2007), `http://archive.ics.uci.edu/ml/datasets/Zoo`
10. Jungnickel, D.: Graphs, Networks and Algorithms. Algorithms and Computation in Mathematics, vol. 5. Springer, Berlin (English edition, 2002)
11. Haykin, S.: Neural Networks. A Comprehensive Foundation, 2nd edn. Prentice Hall International, Upper Saddle River (1999)
12. Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J.: Som toolbox for matlab 5. Technical Report A57, SOM Toolbox Team, Helsinki University of Technology, Espoo, Finland (2000)
13. Davies, D., Bouldin, D.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 1(2), 224–227 (1979)
14. Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal 27, 379–423, 623–656 (1948)
15. Czumaj, A., Kowaluk, M., Lingas, A.: Faster algorithms for finding lowest common ancestors in directed acyclic graphs. Theoretical Computer Science 380, 37–46 (2007)
16. Maimon, O., Rokash, L. (eds.): The Data Mining and Knowledge Discovery Handbook, 1st edn. Springer, New York (2005)