INTEGRATED DIMENSIONALITY REDUCTION TECHNIQUE FOR MIXED DATA INVOLVING CATEGORICAL VALUES

Chung-Chian Hsu
National Yunlin University of Science and Technology Yunlin, Taiwan
hsucc@yuntech.edu.tw

Wei-Hao Huang
National Yunlin University of Science and Technology Yunlin, Taiwan
g9923752@yuntech.edu.tw

Abstract:
An extension to the recent dimensionality-reduction technique t-SNE is proposed. The extension facilitates t-SNE to handle mixed-type datasets. Each attribute of the data is associated with a distance hierarchy which allows the distance between numeric values and between categorical values be measured in a unified manner. More importantly, domain knowledge regarding semantic distance between categorical values can be specified in the hierarchy. Consequently, the extended t-SNE can reflect topological order of the high-dimensional, mixed data in the low-dimensional space.

Keywords: information technology; dimensionality reduction; categorical data; t-SNE

1. INTRODUCTION

High-dimensional data such as sales transactions, medical records and so on are incessantly produced in various domains. Large amount of data usually contain hidden patterns which may be useful to decision makers. Nevertheless, high dimensionality of real-world data suffers several issues, including increased computational cost and curse of dimensionality which causes the definition of density and the distance between points become less meaningful (Tan, Steinbach, & Kumar, 2006). In order to analyze high-dimensional data in a low-dimensional space, one can resort to dimensionality reduction techniques. Many such methods in the literature have been proposed like principal component analysis (PCA) (Hotelling, 1933), classical multidimensional scaling (MDS) (Torgerson, 1952), and t-distributed stochastic neighbor embedding (t-SNE) (L. van der Maaten & G. Hinton, 2008), etc. However, the studies were conducted in the context of numeric data.

Most of real-world datasets consist of categorical and numeric attributes at the same time. For example, the data of credit card applications include numeric attributes such as annual salary, age, and the amount of various savings, and categorical attributes such as education, job, position, and marital status. Nevertheless, many knowledge discovery algorithms do not process mixed-type data. The algorithms analyze either only numeric or categorical data and tackle the deficiency by transforming one type of the data to the other.

For the algorithms which handle only numeric data, 1-of-*k* coding is a popularly adopted method which converts each categorical value to a vector of binary values. However, the method suffers several drawbacks. First, the transformed data increase its dimensionality and so do its computational cost. Second, 1-of-*k* transforms a categorical value to a vector of binary values in which semantics embedded in the categorical values is lost. As a result, the new data do not retain its original topological structure. Moreover, the conversion could affect accuracy or performance of algorithms such as *k*-NN classifier, *k*-mean clustering, SOM, etc.

In this study, a method of dimensionality reduction integrated with a data structure, distance hierarchy, for dissimilarity calculation between categorical values is proposed. The aim of the integration is to facilitate the recent dimensionality reduction technique t-SNE (L. van der Maaten & G. Hinton, 2008) to handle mixed-type datasets in the way of preserving semantics in categorical values. We verified the proposed approach by a synthetic dataset to see whether topological order can be reflected in a two-dimensional plane. In addition, we investigated whether classification performance by processing categorical values with distance hierarchy (DH) coding scheme is better than that with 1-of-*k* coding. Furthermore, we want to explore whether data processing that is based on DH and 1-of-*k* will affect data analysis in a lower data space resulted by dimensionality reduction.

2. PRELIMINARY

2.1. 1-of-k coding

The 1-of-*k* coding is a method for converting categorical attribute to numeric attribute by transformation of a categorical value to a vector of binary values. Specifically, a categorical attribute with a domain of *k* distinct values is transformed to a set of *k* binary attributes such that every binary attribute relates to one of the categorical values. A categorical value in a data point is thus converted to a vector in the new data point in which the value of the corresponding binary attribute is set to one and the others zero. For instance, **Drink** in Picture 1 is a categorical attribute with a domain, say, {**Green Tea**, **Oolong Tea**, **Mocha**, **Latte**}. **Drink** is converted to four binary attributes accordingly. In the transformed data, the value **Green Tea** is represented by the binary vector <1 0 0 0>.

As seen in Picture 1, the dimension is increased from three to six in this toy dataset. Furthermore, for the binary vector part, any two of the vectors from the transformed dataset yield the same Euclidean distance, namely, $\sqrt{2}$. In other words, the semantic similarity that Green Tea is more similar to Oolong Tea than to coffee Latte is lost in the transformed vectors in terms of the Euclidean distance.

Picture 1: Transform categorical attributes with 1-of-*k* coding

Id	Drink	Price	Amount
1	Green Tea	20	5
2	Oolong Tea	30	5
3	Mocha	20	5
4	Latte	30	5

⇩

Id	Green Tea	Oolong Tea	Mocha	Latte	Price	Amount
1	1	0	0	0	20	5
2	0	1	0	0	30	5
3	0	0	1	0	20	5
4	0	0	0	1	30	5

2.2. Dimensionality reduction method

Dimensionality reduction methods transform the data in the high-dimensional space to a low-dimensional space. Many dimensionality reduction methods have been proposed such as principal component analysis (PCA) (Hotelling, 1933), Sammon mapping (Sammon, 1969), classical multidimensional scaling (MDS) (Torgerson, 1952), Locally Linear Embedding (Roweis & Saul, 2000), t-Distributed stochastic neighbor embedding (t-SNE) (L. van der Maaten & G. Hinton, 2008), etc. The data set in the high-dimensional space is defined as $X = \{x_1, x_2, x_3, \dots, x_n\}$ and is defined as $Y = \{y_1, y_2, y_3, \dots, y_n\}$ in the low-dimensional space. The high-dimensional space transforms to the low-dimensional space is defined (Bunte, Biehl, & Hammer, 2011) by

$$f: X \rightarrow \tilde{Y}$$

A general principle of dimensionality reduction includes three components which are characteristics of the data, characteristics of projection and error measure (Bunte et al., 2011). First, the distance or similarity between data points in the original data space is represented as

$$d_{x_{ij}} = f_{d_x}(x_i, x_{j.}\tilde{)}$$

The function f_{d_x} can be Euclidean distance for MDS, or joint probability for t-SNE. Second, the distance or similarity in the low-dimensional space is defined as

$$d_{y_{ij}} = f_{d_y}(y_i, y_{j.}\tilde{)}$$

Similarly, function f_{d_y} can be Euclidean distance for MDS, or joint probability for t-SNE. Finally, the error of projection in the low- dimensional space is referred to as cost function and is defined as

$$\mathcal{E} = f_{\mathcal{E}}(d_x, d_{y.}\tilde{)}$$

The function f_ε is minimized by weighted least squared error for MDS, and Kullback-Leibler divergences for t-SNE.

t-distributed stochastic neighbor embedding was proposed recently by Maaten and Hinton (L. van der Maaten & G. Hinton, 2008). The performance of t-SNE is better than that of other dimensionality reduction methods including Sammon mapping (Sammon, 1969), Isomap (Tenenbaum, Silva, & Langford, 2000) and Locally Linear Embedding (Roweis & Saul, 2000).

2.3. t-distributed stochastic neighbor embedding

The idea behind t-SNE is to construct a probability distribution in the low-dimensional space which is similar to the probability distribution constructed in the high-dimensional space. In the probability distribution, similar data points have a high probability of being picked, while dissimilar points have a low probability of being picked. t-SNE includes four main components (L. van der Maaten & G. Hinton, 2008). The conditional probability of point *j* with respect to point *i* in the high-dimensional space is defined by

$$p_{j|i} = \frac{exp\,(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2})}{\sum k \neq i\; exp(-\frac{\left\|x_k - x_i\right\|^2}{2\sigma^2})}.$$

The probability distribution of data points in the low-dimensional space is defined by

$$q_{ij} = \frac{(1 + \left\|y_i - y_j\right\|^2)^{-1}}{\sum k \neq l\,(\,1 + \|y_k - y_l\|^2)^{-1}}.$$

The goal is to minimize the Kullback-Leibler divergence between the probability distribution *P* in the high-dimensional space and Q in the low-dimensional space. The cost function is defined by

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \text{where} \quad p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}.$$

The cost function can be minimized by using gradient descent. The gradient of the Kullback-Leibler divergence is defined as

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} + q_{ij})(y_i - y_j)\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1}.$$

2.4. *k*-nearest neighbors Classifier

k-nearest neighbors algorithm (Cover & Hart, 1967) is one of the most popular methods for classification. It is a type of supervised learning that has been employed in various domains such as data mining, image recognition, patterns recognition, etc. To classify an unknown data point, the algorithm uses class labels of nearest neighbors to determine the class label of the unknown point. The unknown point is assigned to the class label which occurs most frequently among the set of nearest training neighbors. The classifier is sensitive to noise points if *k* is too small while the neighborhood may include points from other classes if *k* is too large. On the other hand, the weighting with respect to the distances between the unknown point and the neighbors may affect classification result as well

(Dudani, 1976). In principle, the nearer the training point is closer to the unknown point, the larger weight the training point shall have.

3. METHOD

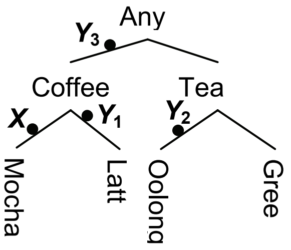
To improve topological order of mixed-type data including categorical and numeric type in the low-dimensional map by dimensionality reduction techniques, we propose to exploit distance hierarchy to measure distance between categorical values. Specifically, we integrate distance hierarchy scheme with t-SNE. The integrated model improves the existing models by retaining semantics embedded in categorical values and avoiding increased dimensionality unlike 1-of-*k* coding.

3.1. Distance Hierarchy

Distance hierarchy (C.-C. Hsu, 2006) is a data structure for representing similarity relationship among values. The hierarchical structure consists of nodes, links, and weights. In distance hierarchy, each node represents a concept. The nodes in the higher levels represent general concepts and the nodes in the lower levels represent specific concepts. Each link is associated with a weight. Link weights can be assigned manually by domain experts or computed according to the relationship of categorical values with the values in the other attributes.

The distance (or similarity) between two values can be measured by the total weight between the two corresponding points of the two values in the hierarchy. As a result, the hierarchy can naturally be used to model similarity relationship among categorical values. In particular, the values (or points) under the same parent node are more similar to one another than to those under another parent node.

Picture 2: Portion of a distance hierarchy for categorical attribute Drink with four projected data points



Specifically, a point in a distance hierarchy is presented by an anchor and a positive offset, denoted as $X = (N_X, d_X)$, representing a leaf node and the distance from the root to *X*, respectively. *X* is an ancestor of *Y* if *X* is in the path from *Y* to the root in the hierarchy. If *X* and *Y* are at the same position, the two are called equivalent, denoted as *X* ≡ *Y*. The lowest common ancestor $LCA(X,Y)$ of two points represents the most specific common tree node of *X* and *Y*. For example, in Picture 2, *LCA* of *X* and Y_1 is **Coffee**. *LCA* of Y_1 and Y_2 is **Any**. The common point *LCP(P,Q)* of two points is defined by

$$LCP(P,Q) = \begin{cases} P\ or\ Q, if\ P \equiv Q \\ P, if\ P\ is\ an\ ancestor\ of\ Q \\ Q, if\ Q\ is\ an\ ancestor\ of\ P \\ LCA(P,Q), otherwise \end{cases} . \qquad (1)$$

In Picture 2, $LCP(X, Y_1) = \mathbf{Coffee}$, $LCP(Y_1, Y_2) = \mathbf{Any}$ and $LCP(Y_1, Y_3) = Y_3$. The distance between two points in the hierarchy is calculated by the total weight between the two points define as

$$|P - Q| = d_P + d_Q - 2d_{LCP(P,Q)} , \qquad (2)$$

where $d_P, d_Q and\ LCP(P,Q)$ represent the distances of point *P* , *Q* and *LCP*(*P*,*Q*) to the root, respectively.

For instance in Picture 2, assume the weight of each link is one, *X* = (Mocha, 1.8), Y_1 = (Latte, 1.2), Y_2 = (Oolong, 1.7), and Y_3 = (Mocha, 0.4). The distance between *X* and Y_1 is | (Mocha, 1.8) – (Latte, 1.2) | = 1.8 + 1.2 – 2 ×1.0 = 1.0, the distance between *X* and Y_2 is | (Mocha, 1.8) – (Oolong, 1.7) | = 1.8 + 1.7 – 2 × 0 = 3.5, the distance between Y_1 and Y_3 is | (Latte, 1.2) – (Mocha, 0.4) | = 1.2 + 0.4 – 2 × 0.4 = 0.8.

Every attribute of the dataset, which can be categorical, ordinal, or numeric, is associated with one distance hierarchy and every attribute value can then be mapped to the distance hierarchy associating with the attribute. A categorical value is mapped to a point at the leaf node labeled by the same value, and numeric value is mapped to a point on a link of its numeric hierarchy. Moreover, ordinal value is converted to a numeric value and processed as a numeric one. Consequently, the distance between categorical values and that between numeric values can be calculated in the same manner by mapping the values to their associated distance hierarchy and then aggregating the weight between the points.

3.2. Construct Distance Hierarchies

There are two ways for constructing distance hierarchies: manual and automated approach. In some domain, there are existing hierarchies ready for use such as the hierarchy of the International Classification for Diseases (ICD) in medicine, the hierarchy of ACM's Computing Classification System (CCS) in computer science, and product classification systems in retail sales.

Some domains do not have existing hierarchies or the values of categorical attributes are encrypted due to privacy consideration. In this situation, we can manually construct the hierarchies by using domain knowledge or apply the idea in (Das, Mannila, & Ronkainen, 1997; Palmer & Faloutsos, 2003) to automatically construct the hierarchies from the dataset.

The automated method for constructing hierarchies includes two parts. In the first part, an approach to quantifying the dissimilarity between two categorical values is to measure the relationship between the values and an external probe. If two categorical values have about the same extent of co-occurrence with the external probe, the values are regarded as similar (Palmer & Faloutsos, 2003). Therefore, the dissimilarity between two categorical values, *A* and *B*, in a feature attribute with regard to the set of labels in class attribute *P*, as the external probe, is defined by (C. C. Hsu & Lin, 2012)

$$d(A,B) = \sum_{D\in P} |conf(A \Longrightarrow D) - conf(B \Longrightarrow D)|, \qquad (3)$$

where $conf(A \Longrightarrow D)$ represents the probability of co-occurrence of *A* and *D* with respect to *A*.

In the second part, the distance matrix of all pair of categorical values in a categorical attribute is calculated, which is denoted as $D_{C\times C}$ where C is the number of categorical values. We can then apply an agglomerative hierarchical clustering algorithm on the matrix to construct a dengrogram, which can be used as the distance hierarchy of the categorical attribute. The distance between two clusters can be measured by single link, average link or complete link.

3.3. Dimensionality reduction with distance hierarchy

In section 2, we discussed t-SNE the performance of which is better than that of many other dimensionality reduction methods. However, the original t-SNE was studied under numeric, instead of mixed-type, data. In this work, we integrate t-distribution stochastic neighbor with distance hierarchy for mixed-type data.

The dimensionality reduction with distance hierarchy (DRDH) consists of three components including mapping the original data points to distance hierarchies, computing similarity between data points on the distance hierarchies, and projecting the data to the low-dimensional space. In the mapping phase, we use different types of distance hierarchy to map different types of attributes, numeric, ordinal and categorical ones.

In the phase of similarity computation, distance hierarchy associated with categorical attribute is constructed first. Then, a pair-wise distance matrix for the values of categorical attribute is constructed. Each pair-wise distance is calculated by summing the link weights between the two points which correspond to the two categorical values. The weight between two points is computed by Eq. (2). The distance matrix is inputted to the dimensionality reduction algorithm.

In the projection phase, the data is projected into a lower dimensional space by using t-SNE. The t-SNE algorithm consists of four steps which include similarity calculation between data points in the original data space, similarity calculation between data points in the map space, cost calculation between the data and the map space, and minimization of the cost function by using gradient descent.

3.4. Evaluation

For the real-world datasets, we cannot evaluate the performance by inspecting the projection maps since we do not know the structure of the data in the high-dimensional space. Therefore, instead we resorted to the *k*-nearest neighbors (or *k*-NN) classification, which is one of the most popular classification algorithms. The idea behind *k*-NN is that the instances with the same class label usually gather together. We evaluate the distance hierarchy coding and the 1-of-*k* coding scheme by comparison of classification accuracy on the testing points. The comparison was conducted in the data space as well as in the map space. The real-world dataset is divided to the training subset with the size 2/3 of the dataset and the testing subset with the size 1/3 of the dataset. The class label of an instance in the testing set is determined by the class labels of the training instances located in the neighborhood of the testing instance.

In particular, a training instance in the data space is denoted by (x_i, y_i) while one in the map space is denoted by (x_i^m, y_i). In classification, the class label y_t for a testing instance (x_t, y_t) by *k* nearest neighbors in the data space and in the map space is defined by Eq. (4) and (5), respectively,

$$y_t = \underset{v}{argmax} \sum_{x_i \in D_{x_t}} w_{it} \times I(v = y_i), \qquad (4)$$

$$y_t = \underset{v}{argmax} \sum_{x_i^m \in D_{x_t^m}} w_{it}^m \times I(v = y_i). \qquad (5)$$

$D_{x_t}$is the set of *k*-nearest neighbors of x_t in the data space. $D_{x_t}^m$is the set of *k*-nearest neighbors of x_t in the map space. *v* is a class label. I is an indicator function returning 1 if the condition is evaluated true and 0 otherwise. w_{it} and w_{it}^m are the weighting in the data and in the map space, respectively, according to the distance between the testing sample and the training sample. The closer the training sample to the testing sample, the larger the weighting value is. In this study, we do not consider the distance weighting; That is, the values of *w* were set to one. Therefore, the class label of the testing instance is assigned to the label of the majority in the set of its *k* nearest neighbors.

4. EXPERIMENTS

To evaluate the integrated model, we conduct experiments in which DRDH is compared to dimensionality reduction with 1-of-*k* coding, referred to as DR1K hereafter, in the data space as well as in the map space. One synthetic dataset was designed to facilitate the inspection of projection result in the two-dimensional map space. For real-world datasets, classification accuracy by *k*-NN with distance hierarchy and 1-of-*k* coding was compared.

4.1. Experimental Data

We used one synthetic dataset and two real-world datasets from the UCI machine learning repository (Merz & Murphy, 1996).

The synthetic dataset as shown in Table 1 includes 240 data points of which each consists of two categorical and one numeric attribute. The dataset mainly includes nine groups of the data. The numeric values in each group were generated and followed a Gaussian distribution with designated mean and standard deviation. The distance hierarchies were manually designed, the one in Fig. 3 for attribute Dept. and the one similar to Fig. 2 for attribute Drink.

The real-world dataset Adult has 48,842 data points of 15 attributes including 8 categorical and 6 numeric and one class attribute indicating salary >50K or ≤50K. The distribution is about 76% of >50K and 24% of ≤50K. Following the reference (C.-C. Hsu, 2006), we used the seven attributes which includes three categorical (Marital-status, Relationship and Education) and four numeric attributes (Capital-gain, Capital-loss, Age and Hours-per-week).

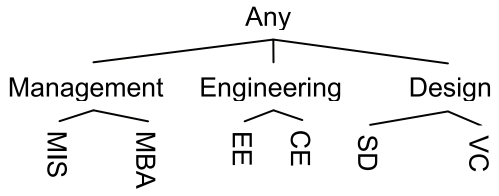
The Nursery dataset has 12,960 data points of 8 categorical and one class attribute indicating not_recom, recommend, very_recom, priority, and spec_prior. The distribution is about 33% of not_recom, 0.015% of recommend, 2.531% of very_recom, 32.917% of priority, and 31.204% of spec_prior.

Due to high computation complexity of t-SNE, random sampling was used to draw 6000 data points for each real-world dataset.

Table 1: The synthetic mixed-type dataset SynMix1

Group	Dept.	Drink	Amount(μ, σ)	Count
1	MIS	Coke	(500, 25)	60
2	MBA	Pepsi	(400, 20)	30
3	MBA	Pepsi	(300, 15)	30
4	EE	Latte	(500, 30)	60
5	CE	Mocha	(400, 20)	30
6	CE	Mocha	(300, 15)	30
7	SD	GreenTea	(500, 25)	60
8	VC	Oolong	(400, 20)	30
9	VC	Ooloing	(300, 15)	30

Picture 3: The distance hierarchy for attribute Dept. of dataset SynMix1



4.2. Experimental Setup

We start respectively by using 1-of-*k* to convert categorical values to numeric values and by mapping categorical values to corresponding distance hierarchies. The distance hierarchies were automatically constructed by using an agglomerative hierarchical clustering algorithm with average link. Parameters of t-SNE were set according to the suggestion in the reference (Laurens van der Maaten & Geoffrey Hinton, 2008).

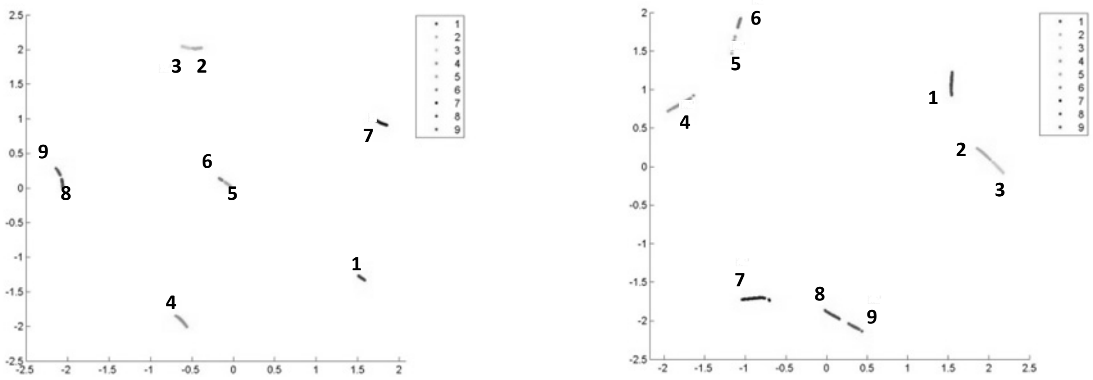
For evaluation, the holdout method 80%-20% was applied in classification. The results of dimensionality reduction were further used to classify the testing instances by the *k*-NN method. Classification by *k*-NN in the original data space with distance hierarchy and 1-of-*k* coding was also conducted.

4.3. Sythetic Dataset

According to the dataset shown in Table I, groups 1, 2, and 3 are intuitively similar to one another since they are of management in the Department attribute and of carbonated drink in the Drink attribute. Analogously, groups 4, 5, and 6 are similar, and so are groups 7, 8, and 9.

Picture 4 shows the projection results of dataset SynMix1. As shown in Picture 4(b) which is the projection by using distance hierarchy coding for categorical attributes, the projection appropriately reflects the structure of the dataset. In particular, groups 1, 2, and 3 are projected close to one another in the same region on the map. So are groups 4, 5, and 6, and groups 7, 8, and 9 as well. In contrast, the projection in Picture 4(a) which is the result by using 1-of-*k* coding does not reflect correct topology. For instance, group 1 is far apart from groups 2 and 3 on the map. So are group 4 from groups 5 and 6, and group 7 from groups 8 and 9.

Picture 4: The projection result of SynMix1 dataset by using coding scheme (a) 1-of-*k* with *Perp* = 220 and (b) distance hierarchy with *Perp* = 220



(a)

(b)

4.4. Real Datasets

As shown in Table 2, classification accuracy by using 1-nearest neighbor classifier with distance hierarchy coding scheme is better than that with 1-of-*k* coding on the two testing datasets. This result holds in the data space as well as the map space.

What is the best accuracy if more neighbors of training instances around the testing instance are taken into account? We set *k* from 1 to 31 with a step of 2. That is, 31 nearest neighbors are considered at most. As shown in Table 3, the best performance usually occurred when *k* was set to one except for the case of Nursery with 1-of-*k* coding in the data space. In that case, the *k* value was 11 and the best performance was 0.9533 which is still lower than that (i.e., 0.9675) of its counterpart, namely, with distance hierarchy in the data space.

Table 2: The accuracy of 1-nearest neighbor for the two real-world datasets

Dataset	Space	1-of-*k*	Dis. Hier.
Adult	Data	0.8592	0.8667
	Map	0.8400	0.8583
Nursery	Data	0.9458	0.9675
	Map	0.8642	0.9625

Table 3: The k value from the range of 1 to 31 with a step of 2 which yields the best performance

Dataset	Space	1-of-*k*	Dis. Hier.
Adult	Data	1	1
	Map	1	1
Nursery	Data	11	1
	Map	1	1

5. CONCLUSION

We proposed a method of dimensionality reduction integrated with data structure distance hierarchy (DRDH) which can handle mixed-typed data and reduce data dimensionality. DRDH has one advantage over the traditional 1-of-*k* coding for converting categorical attributes in that DH considers semantics embedded in categorical values and therefore topological order in the data can be preserved better. The classes projected in the lower dimensional space can be better separated.

The experimental results demonstrated that the structure of the data is more properly reflected by measuring distance between data with distance hierarchy rather than with 1-of-*k* coding. The DRDH yields superior classification results than DR1K in the map space. In the data space, experiments also gave consistent outcome that DH outperforms 1-of-*k* coding.

ACKNOWLEDGMENT

This work is supported by the National Science Council, Taiwan, under grant NSC 102-2410-H-224-019-MY2

REFERENCE LIST

1. Bunte, K., Biehl, M., & Hammer, B. (2011). *Dimensionality reduction mappings*. Paper presented at the IEEE Symposium on Computational Intelligence and Data Mining (CIDM).
2. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on, 13*(1), 21-27.
3. Das, Gautam, Mannila, Heikki, & Ronkainen, Pirjo. (1997). Similarity of Attributes by External Probes. *In Knowledge Discovery and Data Mining*, 23--29.
4. Dudani, Sahibsingh A. (1976). The Distance-Weighted k-Nearest-Neighbor Rule. *Systems, Man and Cybernetics, IEEE Transactions on, SMC-6*(4), 325-327.
5. Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 23*, 417-441.
6. 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. doi: Doi 10.1109/Tnnls.2011.2178323
7. Hsu, Chung-Chian. (2006). Generalizing self-organizing map for categorical data. *Neural Networks, IEEE Transactions on, 17*(2), 294-304. doi: 10.1109/TNN.2005.863415
8. Maaten, L. van der, & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research, 9*, 2579-2605.

9. Maaten, Laurens van der, & Hinton, Geoffrey. (2008). http://homepage.tudelft.nl/19j49/t-SNE.html.
10. Merz, C. J., & Murphy, P. (1996). http://www.ics.uci.edu/~mlearn/MLRepository.html.
11. Palmer, C. R., & Faloutsos, C. (2003). Electricity based external similarity of categorical attributes. *Advances in Knowledge Discovery and Data Mining, 2637*, 486-500.
12. Roweis, S.T., & Saul, L.K. (2000). Nonlinear dimensionality reduction by Locally Linear Embedding. *Science, 290*(5500), 2323-2326.
13. Sammon, J.W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers, 18*(5), 401–409.
14. Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*: Addison Wesley.
15. Tenenbaum, J.B., Silva, V. de, & Langford, J.C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science, 290*(5500), 2319–2323.
16. Torgerson, W.S. (1952). Multidimensional scaling I: Theory and method. *Psychometrika, 17*, 401-419.