Apply Distance Hierarchy and Dimensionality Reduction to Classification of Mixed-Type Data

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

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

Abstract

An integrated dimensionality reduction technique with distance hierarchy which can handle mixed-typed data, reduce dimensionality of the data, and visualize data on a 2D map is proposed. There are two aspects of the integration. First, distance hierarchy (DH) is applied to handle categorical values which are mapped to the DH. In contrast to 1-of-*k* coding, DH considers semantics inherent in categorical values and therefore topological order in the data can be better preserved. Second, t-SNE is employed to reduce data dimensionality which transforms the data in a high-dimensional space to a low-dimensional space. t-SNE is better than other counterparts in separating classes in the lower dimensional space. We use weighted K-NN to evaluate classification performance of using DH and using 1-of-*k* coding in the original data space and in the projection space. We demonstrate the superiority of using DH against 1-of-*k* coding by analyzing four real-world datasets.

Keyword: classification of mixed-type data, distance hierarchy, 1-of-*k* coding, dimensionality reduction, t-SNE

Apply Distance Hierarchy and Dimensionality Reduction to Classification of Mixed-Type Data

I. INTRODUCTION

1.1 Motivation

High dimensional data such as commercial transactions, medical history and so on are incessantly produced in varied domains. A large amount of data usually contains much hidden knowledge. Most data consists of categorical and numeric attributes at the same time. However, many data-mining algorithms cannot process mixed-type data. Most of the algorithms can analyze either only numeric or categorical data.

Complex data including numerical and categorical values are referred to as mixed-type data. For example, credit card data of customers include numerical type such as revenue, age, and the number of parents and categorical type such as education and job. Various algorithms of data-mining techniques cannot directly deal with categorical attributes. Therefore, conversion methods have been proposed that transform categorical to numerical attributes. The 1-of-*k* coding is a well-known conversion method. There are some drawbacks of 1-of-*k* coding. First, 1-of-*k* transforms a categorical value to a vector of binary values in which semantics inherent in the values is lost. Second, the data can't retain its original topological structure. Due to the above two points, the conversion could affect accuracy of algorithms such as K-NN classifier, K-mean clustering, SOM, etc.

On the other hand, high dimensional data suffers the issue of curse of dimensionality. Numerous problems occur in high dimensional data. For instance, computational cost increases with the increased data dimensionality. In order to analyze high-dimensional data, we need to reduce data dimensionality. Dimensionality reduction methods have been proposed such as t-distributed stochastic neighbor embedding (t-SNE) (Hinton 2008b), principal component analysis (PCA) (Pearson 1901), and classical multidimensional scaling (MDS) (Borg 2005), etc.

1.2 Objective

In this study, a method of dimensionality reduction with distance hierarchy (DRDH) is proposed, which integrates distance hierarchy (Hsu 2006) to keep semantics for categorical values and to apply t-SNE (Hinton 2008b) for dimensionality reduction to visualize the data on a two dimensional map. We will investigate whether classification performance by processing categorical values with DH is better than that by 1-of-*k*. 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 in dimensionality reduction.

We are going to compare classification performance in the original data space and the map (or reduced) space with respect to different treatment to categorical values. In particular, we will use K-NN to evaluate the performance in the data space and the map space by using the DH method and the 1-of-*k* coding for handling categorical values. In addition, we will study the impact of weighting to the neighbors of the input on classification accuracy.

1.3 Organization

This study consists of five sections. In Section 2, we review conversion methods of categorical values, dimensionality reduction methods, and weighted K-nearest neighbor classifier. In Section 3, we present DRDH which has three important components include data preprocessing, dimensionality reduction as well as K-NN classifier. In Section 4, we describe the experimental setup and the results of four real-world mixed-type datasets. In Section 5, the experimental results are described.

II. LITERATURE REVIEW

2.1 1-of-*k* coding

The 1-of-*k* coding is a method for converting categorical type to numerical type which transforms a categorical value to a vector of binary values. 1-of-*k* coding transforms a nominal attribute to k unique values that is a set of k binary attributes and one of the binary attributes relates to one of the nominal values. For instance, the drink type is a nominal attribute in a drink dataset and its nominal values include black tea, oolong tea, black coffee and latte. Converting the drink type attribute of nominal values to four attributes each of which data type is binary (as shown in Table 1).

Some disadvantages of 1-of-*k* coding that lead to reduced accuracy of the algorithms (Lin 2009). First, 1-of-*k* transforms a categorical value to a vector of binary values in which semantics inherent in the values is lost. Second, data can’t reflect its original data structure on the projection map. Therefore, we adopt distance hierarchy to solve the drawbacks of 1-of-*k* coding.

Table 1. 1-of-*k* coding coverts nominal attribute to k binary attributes.

Id	Drink type	Price	Amount
1	Black Tea	20	5
2	Oolong tea	25	5
3	Black coffee	45	5
4	latte	55	8

Id	Black Tea	Oolong Tea	Black Coffee	Latte	Price	Amount
1	1	0	0	0	20	5
2	0	1	0	0	25	5
3	0	0	1	0	45	5
4	0	0	0	1	55	8

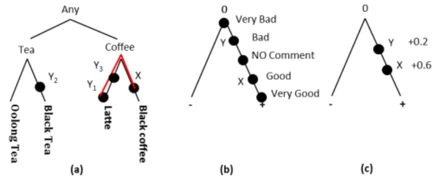


Figure 1. The distance hierarchies for category, ordinal and numeric type. (a) categorical hierarchy (b) ordinal hierarchy (c) numeric hierarchy

2.2 DH

Distance hierarchy is a data structure for representing mixed-type data and it can keep semantics inherent mixed-type data point (Hsu 2006). It is a hierarchical tree structure that includes nodes, links, and weights. In the distance hierarchy, each node on behalf of a concept which higher-level nodes represent general concept and lower-level nodes represent specialized concept likes coffee in the higher-level and latte in the lower-level. Each link represents the semantic relationship and is associated with a weight representing the distance between the two nodes. The distance between two mixed-type data points in distance hierarchy is defined by the total link weight between two mixed-type data points. There are two ways of assigning the link weight. First, link weight is assigned by objective calculation (Das et al. 1997; Palmer et al. 2003). Second, link weight is manually assigned based on expert's domain knowledge.

A point in 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. X and Y are at the same position where both points are equivalent, denoted as $X \equiv Y$. The lowest common ancestor $LCA(X,Y)$ of two points represents the most specialized common ancestor node of X and Y. For example, in Figure 1a, *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

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

In Figure 1a. , $LCP(X, Y_1) = \text{Coffee}$, $LCP(Y_1, Y_2) = \text{Any}$ and $LCP(Y_1, Y_3) = Y_3$. On the other hand, the distance between two points in the hierarchy is calculated by the total weight between the two points define as

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

where $d_P, d_Q\ and\ LCP(P,Q)$ represent distance of point P , Q and LCP(P,Q) from location to the root.

For instance in Figure 1a, Assume X=(Black coffee, 1.8), Y_1=(Latte, 1.7), Y_2=(Black Tea, 1.7), and Y_3=(Latte, 1.2). LCA of X and Y_1 is Coffee and LCP equals LCA. LCA of X and Y_2 is Any and LCP equals Any. LCA of Y_1 and Y_3 is Coffee and LCP equals Y_3. The distance between X and Y_1 is $|(\text{Black coffee}, 1.8) - (\text{Latte}, 1.7)| = 1.8 + 1.7 - 2 \times 1.0$=1.5, the distance between X and Y_2 is $|(\text{Black coffee}, 1.8) - (\text{Black Tea}, 1.7)| = 1.8 + 1.7 - 2 \times 0 = 3.5$, Y_1 and the distance between Y_3 is $|(\text{Black Tea}, 1.7) - (\text{Latte}, 1.2)| = 1.7 + 1.2 - 2 \times 1.2 = 0.5$.

Every attribute of the dataset can be mapped to a distance hierarchy which can be categorical type, ordinal type or numeric type. A categorical value is mapped to a point at the leaf node labeled by the same value, and numerical value is mapped to a point on a link of its numerical hierarchy. Moreover, ordinal value is converted to a numeric value and processed as a numeric value. The ordinal attribute is associated with numerical type distance hierarchy and distributed in the same interval to the right edge.

2.3 Dimensionality reduction method

Dimensionality reduction methods transform the data in the high-dimensional space to low-dimensional space. Many dimensionality reduction methods have been proposed such as t-Distributed stochastic neighbor embedding (t-SNE) (Hinton 2008b), principal component analysis (PCA) (Pearson 1901), classical multidimensional scaling (MDS)(Borg 2005), etc. The data set in the high-dimensional space is defined as $X = \{x_1, x_2, x_3, \dots, x_n\}$ and low-dimensional space is defined as $Y = \{y_1, y_2, y_3, \dots, y_n\}$. The high-dimensional space transforms to low-dimensional space (Bunte et al. 2011) is defined by

$$f: X \to Y. \quad (3)$$

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

$$d_{x_{ij}} = f_{d_x}(x_i, x_j). \quad (4)$$

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

$$d_{y_{ij}} = f_{d_y}(y_i, y_j). \tag{5}$$

The situation presents where function f_{d_y} can be Euclidean distance for MDS, or joint probability for t-SNE. Finally, the error of projection in low- dimensional space is called cost function which is defined as

$$\mathcal{E} = f_{\mathcal{E}}(d_x, d_y). \tag{6}$$

The function $f_{\mathcal{E}}$ can be minimized by weighted least squared error for MDS, and Kullback-Leibler divergences for t-SNE, etc.

2.4 t-distributed stochastic neighbor embedding

The performance of various dimensionality reduction methods on artificial datasets is better than real datasets (Hinton 2008b). Moreover, several methods can't retain the local and the global structure of the data in low-dimensional space. t-distributed stochastic neighbor embedding has been proposed. The performance of t-SNE is better than that of other dimensionality reduction methods.

t-SNE is based on transforming the original data space distance between data points into conditional probabilities which represent similarities. Reference (Hinton 2008b) indicates t-SNE including four main components. The conditional probability of data points in high-dimensionality space is defined by

$$p_{ij} = \frac{exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right)}{\sum k \neq i \; exp\left(-\frac{\left\|x_k - x_i\right\|^2}{2\sigma^2}\right)}. \tag{7}$$

The conditional probability of data points in low-dimensionality space is defined by

$$q_{ij} = \frac{(1 + \left\|y_i - y_j\right\|^2)^{-1}}{\sum k \neq i \; (1 + \left\|y_k - y_i\right\|^2)^{-1}}. \tag{8}$$

The goal is to minimize the Kullback-beibler divergence between joint probability distribution of high-dimensionality space P and low-dimensionality space Q. The cost function is

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}} \;, \tag{9}$$

where

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \;.$$

It uses gradient descent to minimize the cost function. The gradient of the Kullback-beibler divergence is defined as

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} + q_{ij})(y_i - y_j)\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1}. \tag{10}$$

2.5 K-nearest neighbor

K-nearest neighbor algorithm (Cover et al. 1967) is one of the most popular and simplest methods for classification. It's 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 record data point, the algorithm uses class labels of nearest neighbors to determine the class label of unknown record (i.e., in Figure 2a.). However, it's sensitive to noise points if k is too small, and the neighborhood may include points from other classes if k is too large (i.e., in Figure 2b.). On the other hand, the mechanism of weighting can be applied to K-nearest neighbor (Dudani 1976). It can be useful to weight the neighbors by their distance with the testing instance. The nearer neighbors are weighted more to increase their importance than the more distant neighbors.

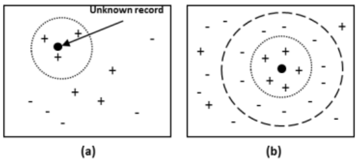


Figure 2. K-NN classify class for unknown data points.

Ⅲ. METHODOLOGY

To improve topological order of mixed-type data including categorical, ordinal and numeric type, we propose to exploit distance hierarchy to measure the distance in the various types of value pairs. To reduce dimensionality, t-distribution stochastic neighbor embedding is employed. We integrate distance hierarchy scheme with t-SNE. The model improves the existing models by retaining semantics inherent in categorical values, solving complex calculation and curse of dimensionality.

3.1 Construct Distance Hierarchies

In data preprocessing phase, the mixed-type data can be preprocessed by using distance hierarchy. We use distance hierarchy algorithm to construct relationship with mixed-type data points. The different attribute types are presented with different types of distance hierarchy. There are two ways for constructing distance hierarchies: manual and automated approaches. 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.

However, some applications don’t have existing hierarchies or the values of categorical attributes are encrypted due to privacy. In this situation, we can manually construct the hierarchies by using domain knowledge or apply the idea in (Das et al. 1997; Palmer et al. 2003) to automatically construct the hierarchies from the dataset.

The automated method of constructing hierarchies includes two components. In the first component, 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 et al. 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 is defined by (Hsu et al. 2012).

$$d(\mathrm{A},\mathrm{B}) = \sum_{D \in P} |conf(\mathrm{A} \Rightarrow D) - conf(\mathrm{B} \Rightarrow D)|, \tag{11}$$

where $conf(\mathrm{A} \Rightarrow D)$ represents the frequency of co-occurrence of A and D.

In the second component, 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. The dengrogram is constructed by applying a hierarchical clustering algorithm on the matrix. The clustering can be performed with single link, average link or complete link.

Algorithm 1 constructs distance hierarchy h_i from different attribute A_i. If construction type of hierarchy is manual, distance hierarchies are constructed by using manual manner based on domain knowledge. Otherwise, distance hierarchies are constructed by using automated manner. There are two phases in the automated construction hierarchy. First, the similarity $d(a_{ij}, a_{ik})$ between two categorical values in the feature attribute is calculated and distance matrix $D_{C_i\times C_i}$ is created. Second, the hierarchical clustering algorithms apply to construct distance hierarchies h_i based on distance matrix $D_{C_i\times C_i}$ of each attribute.

```
Algorithm 1: Construct Distance Hierarchies

Input: A $m_{high}$ dimension of mixed-type dataset $X = \{x_1, x_2, \dots, x_n\}$,
        a type of hierarchical clustering algorithm HC
Output: A set of distance hierarchy $H = \{h_1, h_2, \dots, h_{m_{high}}\}$
Step1. Chose distance hierarchies construction type
if construction type is manual then
 Step2.1. Construct manually
  Construct distance hierarchies manually based on domain knowledge
  else
 Step2.2. Construct automatically
    for i=1 to $m_{high}$
     Step2.2.1 Calculate similarity between two categorical values in the feature attribute
      if $A_i$ is category
        for each different attribute values $a_{ij} \in A_i \in X$
           $d(a_{ij}, a_{ik}) = \sum_{D \in P, j \neq k} |conf(a_{ij} \Rightarrow D) - conf(a_{ik} \Rightarrow D)|$
        end for
       Step2.2.1.1 Apply hierarchical clustering algorithm to construct distance hierarchies
        switch HC
          case single link:
           $h_i$ = single_link($D_{C_i\times C_i}$)
          case average link:
           $h_i$ = average_link($D_{C_i\times C_i}$)
          case complete link:
           $h_i$ = complete_link($D_{C_i\times C_i}$)
        end switch
     Step2.2.2 Construe numeric distance hierarchy
      else if $A_i$ is numeric then
        $h_i$ = numeric_hierarchy($A_i$)
     Step2.2.3 Construe ordinal distance hierarchy
      else if $A_i$ is ordinal then
        $h_i$ = ordinal_hierarchy($A_i$)
      end if
    end for
end if
```

3.2 Dimensionality reduction with distance hierarchy

Several problems exist with analyzing high dimensional data. For instance, computational cost increases with the increased dimensionality. Despite of the proposed dimensionality reduction methods, the performance of these methods on artificial datasets is better than real-world datasets. In section 2, we discussed t-SNE presented in (Hinton 2008b). The performance of t-SNE is better than that of 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 in Algorithm 2 which includes mapping original data points to distance hierarchies, computing similarity between data points on distance hierarchies, and projecting data to the lower data space.

In the mapping phase, we use different types of distance hierarchy to map different types of attributes. Distance hierarchies can be divided into three types including categorical, ordinal, and numeric one. First, a categorical value is mapped to a point at the leaf node labeled by the same value. Second, a numerical value is mapped to a point on the link of its numerical hierarchy. Finally, an ordinal value needs to be converted to a numeric value and then handled by the same way as the numeric one.

In the phase of similarity computation, a distance matrix of pair-wised data points is constructed by summing the weights between two points in individual DHs. The total weight between two points is defined by Eq. (1) and (2).

In the projection phase, the data is projected onto 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.

Algorithm 2: dimensionality reduction with Distance hierarchy

```
Input: A $m_{high}$ dimension of mixed-type dataset $X = \{x_1, x_2, \dots, x_n\}$,
       a set of distance hierarchy $H = \{h_1, h_2, \dots, h_{m_{high}}\}$,
       number of low-dimensionality $m_{low}$,
       t-SNE cost function parameters $perplexity$
Output: A $m_{low}$ dimensional of low-dimensionality space $Y = \{y_1, y_2, \dots, y_n\}$,
       a distance matrix of original data space $D_{n \times n} = \{d_{11}, d_{12}, \dots, d_{nn}\}$
Step 1. Map attribute values $a_i$ on distance hierarchies $h_i$.
  for each $x \in X$
    map attribute value $a_i \in x$ on distance hierarchy $h_i \in H$
  end for
Step2. Calculate distance $d_{ij}$ between data points on original data by distance hierarchy
  for each $x \in X$
    $d_{ij}$=Calculate distance by DH($x_i$, $x_j$) (using Equation 1 and 2)
  end for
```

Step3. *Reduce* m_{high} *dimension of high-dimensionality space* X *to* m_{low} *dimension of low-dimensionality space* Y

$Y = \text{t-SNE}(D_{n\times n}, m_{low}, perplexity)$ (using t-SNE dimensionality reduction method)

3.3 Classification by Using K-NN for Distance Hierarchy Dimensionality Reduction

To evaluate classification performance of DRDH, we use K-nearest neighbor algorithm. K-NN is one of the most popular and simplest methods for classification. It is said that weighting the neighbors by their distance to the testing instance is beneficial. The nearer neighbors are weighted more to increase their importance than the more distant neighbors. Therefore, we investigate the performance of K-NN with different weighting mechanisms to classify unknown records which are created by DRDH.

There are three phases for weighted K-NN using on data space and map space which include decomposition data into training data and testing data, weighting for distance between two data points, and prediction class label for unknown record. The process of classification is presented in Algorithm 3.

In the first phase, we decompose the dataset into training data and testing data. The training data of data space is denoted as $D_{testing}$ and map space is denoted as $Y_{training}$. The testing data of data space is denoted as $D_{testing}$ and map space is denoted as $Y_{testing}$. In the weighting phase, the distance of pairwise data point $d_{x_{ij}}$ and $d_{y_{ij}}$ can be compressed or enlarged through transformation function f. There are two weighting processes with different transformation functions which are defined in Table 2. In the prediction phase, determining class labels for testing data in map space by k nearer neighbors is defined by

$$y_{label_{testing}} = \underset{v}{argmax} \sum\nolimits_{(y_{training,i} y_{testing,i}) \in D_y} w_{y_{i\square}} \times I(v = y_{testing,i}). \tag{12}$$

Determining class labels for testing data in data space by k nearer neighbors is defined by

$$x_{label_{testing}} = \underset{v}{argmax} \sum\nolimits_{(x_{training,i} x_{testing,i}) \in D_x} w_{x_{ij}} \times I(v = x_{testing,i}). \tag{13}$$

The $(y_{training,i}, y_{testing,i})$ is a training example from k-nearest neighbors D_y, $(x_{training,i}, x_{testing,i})$ is a training example from k-nearest neighbors D_x, w_{ij} is weighting to distance between testing data point and training data point, and I is an indicator function returning 1 if the condition is evaluated true and 0 otherwise.

Algorithm 3: K-NN Classifier for Generalized Distance Hierarchy Dimensionality Reduction

```
Input:  A $m_{low}$ dimensional of low-dimensionality space $Y = \{y_1, y_2, \cdots, y_n\}$,
        a distance matrix of original data space $D_{n \times n}$,
        neighbors parameter k.
Output: the predicated class labels for testing data in data space $D_{label_{testing}}$ and map space$Y_{label_{testing}}$
Step 1. Divide into training and testing data for data and map space
  Divide training data $Y_{training}, D_{training}$ and testing data $Y_{testing}, D_{testing}$
Step 2. Weight distance for each nearer neighbor
  if space is data space then
    $w_{ij} = 1/f(d_{x_{ij}})$
  else
    $w_{ij} = 1/f(d_{Y_{ij}})$
  end if
Step 3. Determine class labels for testing data by k nearer neighbors
  if data is map space then
    $y_{label_{testing}} = \arg\max_{v} \sum_{(y_{training,i}, y_{testing,i}) \in D_y} w_{ij} \times I(v = y_{testing,i})$
  else
    $x_{label_{testing}} = \arg\max_{v} \sum_{(x_{training,i}, x_{testing,i}) \in D_x} w_{ij} \times I(v = x_{testing,i})$
  end if
```

Table 2. Weighting of the distance used in K-NN.

K-NN type	Method	Weighting Function	Transformation function f
Weighted-KNN	Tf	$w_{ij} = \begin{cases} \frac{1}{f(d_{ij})}, if\ weight = Exp \\ \frac{1}{f(d_{ij})+1}, otherwise \end{cases}$	$\mathrm{Log}(d_{ij}+1) = \log(d_{ij}+1)$ $\mathrm{Pow}(d_{ij}, 2) = (d_{ij})$^2 $\mathrm{Exp}(d_{ij}) = $e^$(d_{ij})$ $\mathrm{Sqrt}(d_{ij}) = (d_{ij})$^0.5 $\mathrm{Org}(d_{ij}) = d_{ij}$
Original-KNN	W1	$w_{ij} = 1$	$f(d_{ij}) = d_{ij}$

Ⅳ. EXPERIMENTS

To evaluate DRDH, we present experiments in which DRDH is compared to dimensionality reduction with 1-of-*k* coding, referred to as DR1K hereafter, on map space. Classification accuracy by K-NN with distance hierarchy and 1-of-*k* coding on data space is also compared. Furthermore, the effect of distance weighting on K-NN is investigated as well.

4.1. Experimental data

We used four real-world datasets from the UCI machine learning repository (Merz et al. 1996). The Adult dataset 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. We selected 7 attributes according to (Hsu 2006) 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.

The Car evaluation dataset has 1728 data points of 6 categorical and one class attribute indicating unacc, acc, good, and v-good. The distribution is about 77.023% of unacc, 22.222 % of acc, 3.993 % of good, or 3.762 % of v-good.

The Australian credit approval dataset has 690 data points of 14 attributes including 8 categorical, 6 numerical and one class attribute indicating 1(+) and 0(-). The distribution is about 44.5% of 1 and 55.5% of 0. We selected 11 attributes according to a chi-square test which includes five categorical (A4, A5, A6, A8 and A9) and six numeric attributes (A2, A3, A7, A10, A13 and A14).

4.2. Experimental setup

We start by using distance hierarchy or 1-of-*k* to convert categorical values to numeric values in the data preprocessing stage. The distance hierarchies were automatically constructed by using a hierarchical clustering algorithm with average link, complete link and single link.

Due to high computation complexity of t-SNE, random sampling was used to choose 6000 data points for the two large datasets Adult and Nursery. For the other two small datasets Car evaluation and Australian credit approval, all of the instances were used.

Parameters of t-SNE (Hinton 2008a) were set according to the suggestion in the reference (Hinton 2008a). However, we need to modify the parameter *Perplexity* in some situation, for instance, pairwise data points of dataset are high degree of similarity which will cause the cost function closer to zero in the initial stage for t-SNE.

For evaluation, the holdout method 80%-20% was applied to the large datasets and 10-fold cross-validation was applied to the small datasets. The results of dimensionality reduction were further used to classify the testing by the K-NN method. Classification by K-NN on the original data space with distance hierarchy and 1-of-*k* is also conducted. The detail of experimental parameter setting is shown in Table 3.

Table 3. The parameter setting of experiments for four real world datasets.

Parameter Setting	**Adult dataset**	**Nursery dataset**	**Car evaluation dataset**	**Australian credit approval dataset**
Data preprocessing	1-of-*k* method DH with average link, complete link and single link			
Perplexity of t-SNE	1ofk sets *Perp*=230 Single sets *Perp*=50 Average sets *Perp*=50 Complete sets *Perp*=50	1ofk sets *Perp*=100 Single sets *Perp*=30 Average sets *Perp*=30 Complete sets *Perp*=30	1ofk sets *Perp*=30 Single sets *Perp*=30 Average sets *Perp*=30 Complete sets *Perp*=30	1ofk sets *Perp*=30 Single sets *Perp*=30 Average sets *Perp*=30 Complete sets *Perp*=30
K-NN	The weighted K-NN and the original K-NN are applied to evaluate performance on the data space and map space. Moreover, weighted K-NN transform distance with different weight functions to weight distance between pair of data point. The transformation functions of weighting distance are shown in Table 2.			
K value	K value sets 1 to 31.			
Evaluation	Holdout	Holdout	10-fold Cross-validation	10-fold Cross-validation

4.3. Analysis of experimental results

To analyze the four experimental results by different point of views, we collect the accuracy of experiments in which the experimental results is analyzed by the best accuracy and the average accuracy. The best accuracy is defined by equation (14). The average accuracy is defined by equation (15). The k value of best accuracy is analyzed in which is defined by equation (16).

$$Accuracy_{bset} = \max_{k}(accuracy_{k=1}, \dots, accuracy_{k=31}). \tag{14}$$

$$Accuracy_{avg} = \frac{1}{31}\sum_{k=1}^{31} accuracy_{k} . \tag{15}$$

$$K_{best} = arg\max_{k}(accuracy_{k=1}, \dots, accuracy_{k=31}). \tag{16}$$

4.3.1 The accuracy of 1-nearest neighbor

Via accuracy of 1-nearest neighbor classification, we like to prove that structure of the data is better reflected by measuring distance between data with distance hierarchy rather than with 1-of-*k* coding.

According to Table 4, in the projection space DRDH yields better results than DR1K and in the data space DH outperforms 1-of-*k* as well. In Car Evaluation, the accuracy of DH is significantly better than that of 1-of-*k* coding either on the map space or on the data space. In addition, all results on the data space are better than those on the projection space. The dataset lost some data structure by applying dimensionality reduction. The extent of loss by DR1K is worse than by DRDH.

4.3.2 The k value of best accuracy

Table 5 shows the *k* value which results in the best accuracy. Except for Australian Credit Approval, most datasets have a small *k* when distance hierarchy was used. Moreover, with 1-of-*k* more nearest neighbors are required to obtain the best result.

4.3.3 Comparison accuracy of different space

The best accuracy in different spaces is compared in the Figure 3. The accuracy with differently weighted K-NNs was indicated by Exp, Log, Org, Pow, or Sqrt. The accuracy with a K-NN without weighting was indicated by Non. The results show most of accuracy on the data space is better than on the map space. The high dimensional space converted to a low dimension space will lose information. We found that DRDH yields better results than DR1k on the map space.

On the other hand, the average accuracy is compared in Figure 4. The average accuracy is overall evaluation accuracy for weighted K-NN and original K-NN on the different space. For the original K-NN, the results of the average accuracy and the best accuracy are consistency which most results show data space yields better results than map space. The comparison average accuracy of weighted K-NN is different form comparison best accuracy. The weighted K-NN is beneficial to improve accuracy in the projection space more than in the data space. We found that the results of average accuracy were improved with weighted K-NN in the projection space. For instance, the Figure 4 obviously shows that the average accuracy is improved with weighted K-NN for the Adult dataset and the Nursery dataset in the projection.

4.3.4 Comparison accuracy of 1ofk and DH

In the Figure 3 and the Figure 4 prove consistently result which DH better than 1-of-*k* coding on the data space and on the map space. Most experimental results indicate DHRD with complete or average link outperforms DHRD with single link.

4.3.5 Comparison accuracy of transformation function with K-NN

Figure 3 did not show significant differences in performance by the different transformation functions. Because all of the transformation functions got the same best accuracy in the both space. However, in Figure 4 regarding average performance, most experimental results indicate the expansive weighting schemes which expands the original distance are better than the compressive methods. The accuracy ranking of the schemes are Exp > Pow > Org > Sqrt > Log.

Table 4. The accuracy of 1-nearest neighbor for the four datasets.

Dataset	Space	1ofk	Average	Complete	Single
Car Evaluation	**Data**	0.7920 (0.1241)	**0.9033 (0.0764)**	**0.9033 (0.0764)**	**0.9033 (0.0764)**
	Map	0.7673 (0.0475)	0.8912 (0.0767)	0.8958 (0.0780	0.8755 (0.0851)
Australian	**Data**	0.7913(0.0472)	0.8130(0.0320)	0.8014(0.0459)	**0.8145(0.0371)**
	Map	0.7652(0.0429)	0.8043(0.0490)	0.7942(0.0604)	0.7942(0.0409)
Adult	**Data**	0.8592	0.8667	**0.8675**	0.8617
	Map	0.8400	0.8583	0.8533	0.8467
Nursery	**Data**	0.9458	**0.9675**	**0.9675**	0.9558
	Map	0.8642	0.9625	0.9608	0.9558

Table 5. The k value of best accuracy for (a) Car Evaluation dataset (b) Australian dataset (c) Adult dataset (d) Nursery dataset.

Coding	Space	Exp	Log	Org	Pow	Sqrt	Non
1ofk	**Data**	5	5	5	5	5	5
	Map	31	31	31	29	31	11
Average	**Data**	3	3	3	3	3	1
	Map	1	3	1	1	3	1
Complete	**Data**	1	3	3	3	3	1
	Map	1	1	1	1	1	1
Single	**Data**	3	3	3	3	3	1
	Map	3	3	3	3	3	1

(a)

Coding	Space	Exp	Log	Org	Pow	Sqrt	Non
1ofk	**Data**	25	25	25	25	25	23
	Map	5	5	5	5	5	9
Average	**Data**	19	19	19	19	19	19
	Map	21	11	15	17	13	13
Complete	**Data**	17	17	17	17	17	17
	Map	17	13	17	15	13	17
Single	**Data**	21	17	21	21	25	25
	Map	15	19	19	19	19	21

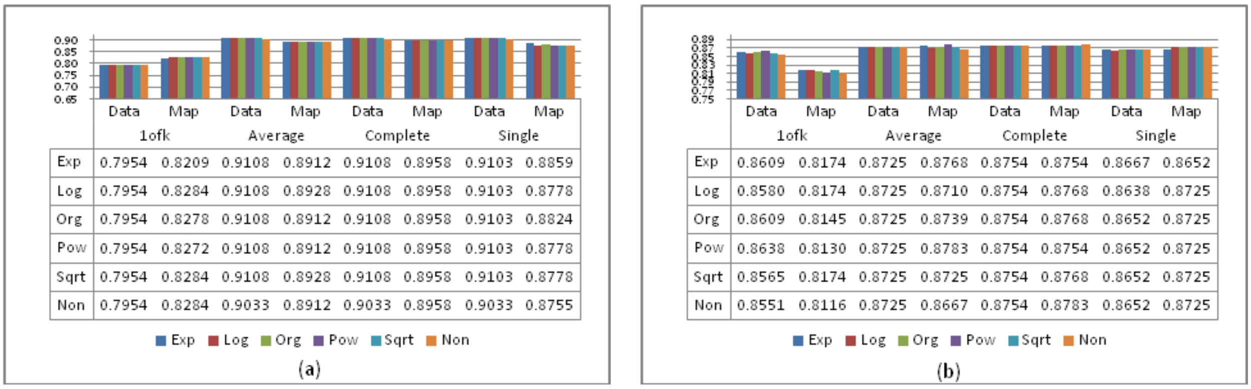
(b)

Coding	Space	Exp	Log	Org	Pow	Sqrt	Non
1ofk	**Data**	1	1	1	1	1	1
	Map	1	1	1	1	1	1
Average	**Data**	1	1	1	1	1	1
	Map	1	1	1	1	1	1
Complete	**Data**	1	1	1	1	1	1
	Map	1	1	1	1	1	1
Single	**Data**	1	1	1	1	1	1
	Map	1	1	1	1	1	1

(c)

Coding	Space	Exp	Log	Org	Pow	Sqrt	Non
1ofk	**Data**	15	11	11	15	11	11
	Map	13	1	3	11	3	1
Average	**Data**	1	1	1	1	1	1
	Map	1	1	1	1	1	1
Complete	**Data**	1	1	1	1	1	1
	Map	1	1	1	1	1	1
Single	**Data**	1	1	1	1	1	1
	Map	1	1	1	1	1	1

(d)

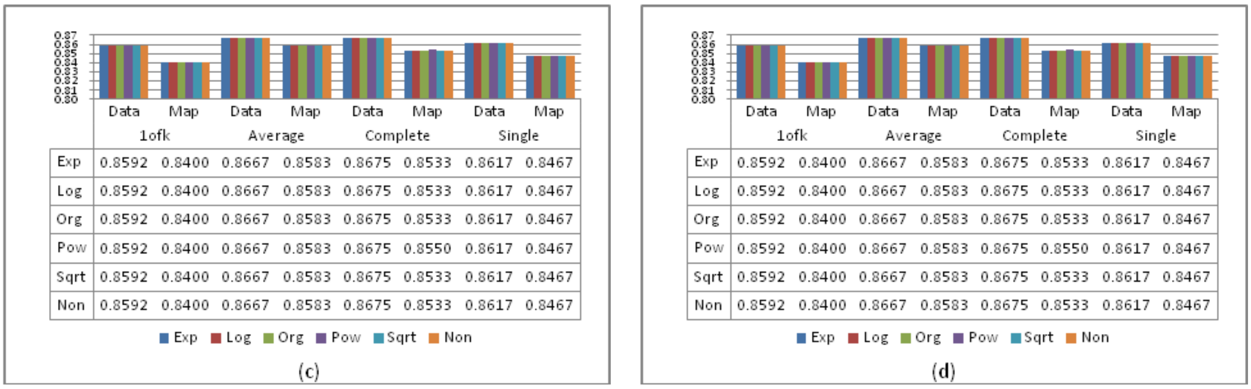


	Data	Map	Data	Map	Data	Map	Data	Map
	1ofk		Average		Complete		Single	
Exp	0.7954	0.8209	0.9108	0.8912	0.9108	0.8958	0.9103	0.8859
Log	0.7954	0.8284	0.9108	0.8928	0.9108	0.8958	0.9103	0.8778
Org	0.7954	0.8278	0.9108	0.8912	0.9108	0.8958	0.9103	0.8824
Pow	0.7954	0.8272	0.9108	0.8912	0.9108	0.8958	0.9103	0.8778
Sqrt	0.7954	0.8284	0.9108	0.8928	0.9108	0.8958	0.9103	0.8778
Non	0.7954	0.8284	0.9033	0.8912	0.9033	0.8958	0.9033	0.8755

(a)

	Data	Map	Data	Map	Data	Map	Data	Map
	1ofk		Average		Complete		Single	
Exp	0.8609	0.8174	0.8725	0.8768	0.8754	0.8754	0.8667	0.8652
Log	0.8580	0.8174	0.8725	0.8710	0.8754	0.8768	0.8638	0.8725
Org	0.8609	0.8145	0.8725	0.8739	0.8754	0.8768	0.8652	0.8725
Pow	0.8638	0.8130	0.8725	0.8783	0.8754	0.8754	0.8652	0.8725
Sqrt	0.8565	0.8174	0.8725	0.8725	0.8754	0.8768	0.8652	0.8725
Non	0.8551	0.8116	0.8725	0.8667	0.8754	0.8783	0.8652	0.8725

(b)



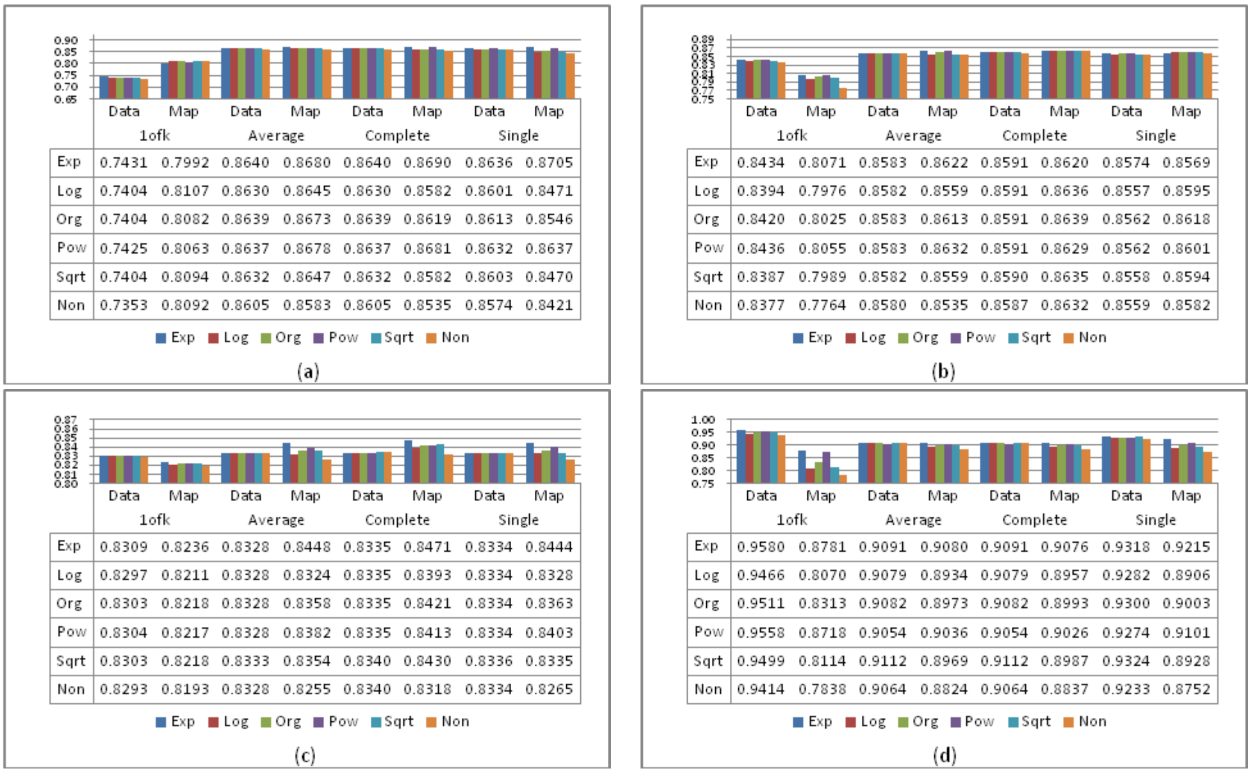
(c)

	Data	Map	Data	Map	Data	Map	Data	Map
	1ofk		Average		Complete		Single	
Exp	0.8592	0.8400	0.8667	0.8583	0.8675	0.8533	0.8617	0.8467
Log	0.8592	0.8400	0.8667	0.8583	0.8675	0.8533	0.8617	0.8467
Org	0.8592	0.8400	0.8667	0.8583	0.8675	0.8533	0.8617	0.8467
Pow	0.8592	0.8400	0.8667	0.8583	0.8675	0.8550	0.8617	0.8467
Sqrt	0.8592	0.8400	0.8667	0.8583	0.8675	0.8533	0.8617	0.8467
Non	0.8592	0.8400	0.8667	0.8583	0.8675	0.8533	0.8617	0.8467

(d)

	Data	Map	Data	Map	Data	Map	Data	Map
	1ofk		Average		Complete		Single	
Exp	0.8592	0.8400	0.8667	0.8583	0.8675	0.8533	0.8617	0.8467
Log	0.8592	0.8400	0.8667	0.8583	0.8675	0.8533	0.8617	0.8467
Org	0.8592	0.8400	0.8667	0.8583	0.8675	0.8533	0.8617	0.8467
Pow	0.8592	0.8400	0.8667	0.8583	0.8675	0.8550	0.8617	0.8467
Sqrt	0.8592	0.8400	0.8667	0.8583	0.8675	0.8533	0.8617	0.8467
Non	0.8592	0.8400	0.8667	0.8583	0.8675	0.8533	0.8617	0.8467

Figure 3. The best accuracy of different space with different type K-NN. (a) Car Evaluation dataset (b) Australian dataset (c) Adult dataset (d) Nursery dataset



(a)

	Data	Map	Data	Map	Data	Map	Data	Map
	1ofk		Average		Complete		Single	
Exp	0.7431	0.7992	0.8640	0.8680	0.8640	0.8690	0.8636	0.8705
Log	0.7404	0.8107	0.8630	0.8645	0.8630	0.8582	0.8601	0.8471
Org	0.7404	0.8082	0.8639	0.8673	0.8639	0.8619	0.8613	0.8546
Pow	0.7425	0.8063	0.8637	0.8678	0.8637	0.8681	0.8632	0.8637
Sqrt	0.7404	0.8094	0.8632	0.8647	0.8632	0.8582	0.8603	0.8470
Non	0.7353	0.8092	0.8605	0.8583	0.8605	0.8535	0.8574	0.8421

(b)

	Data	Map	Data	Map	Data	Map	Data	Map
	1ofk		Average		Complete		Single	
Exp	0.8434	0.8071	0.8583	0.8622	0.8591	0.8620	0.8574	0.8569
Log	0.8394	0.7976	0.8582	0.8559	0.8591	0.8636	0.8557	0.8595
Org	0.8420	0.8025	0.8583	0.8613	0.8591	0.8639	0.8562	0.8618
Pow	0.8436	0.8055	0.8583	0.8632	0.8591	0.8629	0.8562	0.8601
Sqrt	0.8387	0.7989	0.8582	0.8559	0.8590	0.8635	0.8558	0.8594
Non	0.8377	0.7764	0.8580	0.8535	0.8587	0.8632	0.8559	0.8582

(c)

	Data	Map	Data	Map	Data	Map	Data	Map
	1ofk		Average		Complete		Single	
Exp	0.8309	0.8236	0.8328	0.8448	0.8335	0.8471	0.8334	0.8444
Log	0.8297	0.8211	0.8328	0.8324	0.8335	0.8393	0.8334	0.8328
Org	0.8303	0.8218	0.8328	0.8358	0.8335	0.8421	0.8334	0.8363
Pow	0.8304	0.8217	0.8328	0.8382	0.8335	0.8413	0.8334	0.8403
Sqrt	0.8303	0.8218	0.8333	0.8354	0.8340	0.8430	0.8336	0.8335
Non	0.8293	0.8193	0.8328	0.8255	0.8340	0.8318	0.8334	0.8265

(d)

	Data	Map	Data	Map	Data	Map	Data	Map
	1ofk		Average		Complete		Single	
Exp	0.9580	0.8781	0.9091	0.9080	0.9091	0.9076	0.9318	0.9215
Log	0.9466	0.8070	0.9079	0.8934	0.9079	0.8957	0.9282	0.8906
Org	0.9511	0.8313	0.9082	0.8973	0.9082	0.8993	0.9300	0.9003
Pow	0.9558	0.8718	0.9054	0.9036	0.9054	0.9026	0.9274	0.9101
Sqrt	0.9499	0.8114	0.9112	0.8969	0.9112	0.8987	0.9324	0.8928
Non	0.9414	0.7838	0.9064	0.8824	0.9064	0.8837	0.9233	0.8752

Figure 4. The average accuracy of different space with different type K-NN. (a) Car Evaluation dataset (b) Australian dataset (c) Adult dataset (d) Nursery dataset

V . CONCLUSION

We proposed a method of dimensionality reduction with distance hierarchy (DRDH) which can handle mixed-typed data and reduce data dimensionality. DRDH benefits from two important properties: DH considers semantics inherent in categorical values and therefore topological order in the data can be preserved better. The classes in the lower dimensional space can be better separated.

The experimental results prove 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 on the map space. The data space also gives consistent outcome that DH outperforms 1-of-*k* coding. Moreover, the weighted K-NN is beneficial in improving accuracy in the projection space more than in the data space. The average accuracy is significantly improved with expansive distance weighting schemes on K-NN in the projection space.

ACKNOWLEDGEMENTS

The work is supported by the National Science Council, Taiwan, under grant NSC 100-2410-H-224-003-MY2.

REFERENCES

Borg, I., Groenen, P. *Modern Multidimensional Scaling: theory and applications, (2nd ed.)*, New York: Springer-Verlag, 2005.

Bunte, K., Biehl, M., and Hammer, B. "Dimensionality reduction mappings," Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on, 2011, pp. 349-356.

Cover, T., and Hart, P. "Nearest neighbor pattern classification," *Information Theory, IEEE Transactions on* (13:1) 1967, pp 21-27.

Das, G., Mannila, H., and Ronkainen, P. "Similarity of Attributes by External Probes," *In Knowledge Discovery and Data Mining*) 1997, pp 23--29.

Dudani, S. A. "The Distance-Weighted k-Nearest-Neighbor Rule," *Systems, Man and Cybernetics, IEEE Transactions on* (SMC-6:4) 1976, pp 325-327.

Hinton, L. v. d. M. a. G. "http://homepage.tudelft.nl/19j49/t-SNE.html,") 2008a.

Hinton, L. v. d. M. a. G. "Visualizing data using t-SNE," *Journal of Machine Learning Resarch* (9) 2008b, pp 2579-2605.

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

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

Lin, S.-H. "Apply Extended Self-Organizing Map to Analyze Mixed-Type Data," National Yunlin University of Science & Technology, 2009.

Merz, C. J., and Murphy, P. "http://www.ics.uci.edu/~mlearn/MLRepository.html," 1996.

Palmer, C. R., and Faloutsos, C. "Electricity based external similarity of categorical attributes," *Lecture notes in computer science*) 2003, pp 486-500.

Pearson, K. "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*) 1901, pp 2:559-557.

應用階層距離結合降維度於分類混合型資料

黃韋皓
國立雲林科技大學資訊管理學系
g9923752@yuntech.edu.tw

許中川
國立雲林科技大學資訊管理學系
hsucc@yuntech.edu.tw

摘要

資料探勘中很多技術只能處理種類或數值型資料，無法對同時含有種類及數值的混合型資料進行處理與分析。此外，高維度資料使得運算成本過高及不易分析。以往分析混合型資料，需將種類型資料先進行二元編碼處理，但經由二元編碼後，會失去其原有的拓撲結構及語意關係，t 分配隨機鄰居嵌入法是降低維度技術之一，其投射績效比其他技術較傑出。本研究藉由整合距離階層及 t 分配隨機鄰居嵌入法，進而提出階層距離降維度技術來處理混合型資料並降低其資料維度。我們利用加權型最近鄰居法評估真實資料集在不同維度空間之分類績效，實驗結果證明本研究所提方法能改善投射結果。

關鍵詞：分類混合型資料、距離階層 、1-of-*k* 編碼、降維度、t 分配隨機鄰居嵌入法