Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

# Coupled Interdependent Attribute Analysis on Mixed Data

**Can Wang, Chi-Hung Chi, Wei Zhou**
Digital Productivity Flagship, CSIRO, Australia
{can.wang, chihung.chi, wei.zhou}@csiro.au

**Raymond Wong**
University of New South Wales, Australia
wong@cse.unsw.edu.au

## Abstract

In the real-world applications, heterogeneous interdependent attributes that consist of both discrete and numerical variables can be observed ubiquitously. The usual representation of these data sets is an information table, assuming the independence of attributes. However, very often, they are actually interdependent on one another, either explicitly or implicitly. Limited research has been conducted in analyzing such attribute interactions, which causes the analysis results to be more local than global. This paper proposes the coupled heterogeneous attribute analysis to capture the interdependence among mixed data by addressing coupling context and coupling weights in unsupervised learning. Such global couplings integrate the interactions within discrete attributes, within numerical attributes and across them to form the coupled representation for mixed-type objects based on dimension conversion and feature selection. This work makes one step forward towards explicitly modeling the interdependence of heterogeneous attributes among mixed data, verified by the applications in data structure analysis, data clustering evaluation, and density comparison. Substantial experiments on 12 UCI data sets show that our approach can effectively capture the global couplings of heterogeneous attributes and outperforms the state-of-the-art methods, supported by statistical analysis.

## Introduction

In the era of big data, real-life mixed data sets are usually described by a mixture of heterogeneous interdependent attributes in diverse domains, including demography and finance. Here, *mixed data* is defined as a data set whose columns consist of both discrete and numerical attributes (i.e. with multiple heterogeneous attributes). The classical data representation is an information table, in which rows stand for objects and columns denote attributes. Each entry is designated a value of a particular attribute for a given object. This traditional way quantifies objects by associated multiple variables and assumes the independence of them.

The two key properties of mixed data focused in our study are its heterogeneity (Hunt and Jorgensen 2011) and interdependence (Cao, Ou, and Yu 2011). Tackling both properties simultaneously in data mining (Jia and Zhang 2008) processes is not an easy task because data scales/types are totally different, which challenges the calculation of similarity between mixed-type objects. There have been research efforts on mixed data, but most of them are partial and mainly address either heterogeneity or interdependence. For example, *k-prototype* (Huang 1998) tries to quantify similarity between mixed-typed objects, but without analyzing interdependence; *mADD* (Ahmad and Dey 2007) models only the interdependence of discrete attributes; and *mixture* (Hunt and Jorgensen 2011) only captures the interdependence of numerical attributes. Due to the complexity and variety of mixed data, addressing interdependence among heterogeneous mixed data in data mining is still a big open research question. This is the motivation of our paper.

Taking a fragment of synthetic data[1] as an example (i.e. Table 1), six objects are characterized by two categorical attributes and two continuous attributes; and they are divided into three classes. Each value only exhibits relevant information of its belonged attributes, but does not reflect any interaction with other variables. Based on such a table, many data mining techniques and machine learning tasks (Plant 2012) including clustering and classification have been performed. One of the critical parts in these applications is to study the pairwise distance between objects, which challenges the current work since the distance calculations for discrete and continuous variables are totally different. It is a significant research issue on how to integrate distinct mechanisms of distance computation in a proper manner.

A few metrics have been developed for mixed data, such as the weighted mixed metrics (*wmm* for short) proposed in (Huang 1998). Since objects $u_4$ and $u_6$ have identical values of $a_1^d$ and $a_2^d$, the *wmm* between them is only $0.535$, which is much smaller than that between $u_4, u_3$ (i.e. $1.311$) and nearly third of that between $u_6, u_5$ (i.e. $1.411$). It indicates that $u_4$ and $u_6$ stand a good chance to be clustered into the same group. However, in fact, $u_4$ and $u_3$ belong to $G_2$, $u_6$ and $u_5$ are labeled as $G_3$. The same phenomenon also applies for the partition of $u_3$ and $u_5$ with the *wmm* to be $0.861$.

This instance shows it is too limited to analyze mixed data by assuming all heterogeneous attributes to be independent.



[1] It is produced from a mixture of three Gaussian distributions with the first two variables categorized and predefined classes.

Table 1: A Synthetic Mixed Data Set ($FS$)

| data | $a_1^d$ | $a_2^d$ | $a_1^r$ | $a_2^r$ | class |
|---|---|---|---|---|---|
| $u_1$ | $X$ | $\beta$ | -1.133 | -4.743 | $G_1$ |
| $u_2$ | $Z$ | $\beta$ | -3.418 | -3.575 | $G_1$ |
| $u_3$ | $Y$ | $\alpha$ | -2.153 | -0.690 | $G_2$ |
| $u_4$ | $Z$ | $\alpha$ | -1.952 | 0.274 | $G_2$ |
| $u_5$ | $Y$ | $\alpha$ | 1.822 | 2.102 | $G_3$ |
| $u_6$ | $Z$ | $\alpha$ | 0.779 | 1.112 | $G_3$ |

In real-life applications, their data often consist of heterogeneous attributes interdependent on one another (Cao, Ou, and Yu 2011). From both practical and theoretical aspects, it is important to develop an effective representation scheme for mixed data sets by considering the interdependence relationships among their heterogeneous attributes, which include semantic couplings from domain knowledge and functional couplings from data dynamics itself. In this paper, we focus on discovering/learning the functional couplings that only rely on the given data without any domain knowledge.

There are indeed some simple solutions for mixed data when modeling the interdependence. This includes directly either converting discrete values into numerical values or discretizing numerical attributes into discrete attributes. One common practice is to assign numerical values to discrete attributes. For discrete variables such as color (e.g. red, blue, green), it is normal to assign 1, 2, 3 to them respectively. To quantify the distance between lengths (e.g. 1.60m, 1.75m), one way is to calculate the difference: 1.75m-1.60m=0.15m. However, it might be questionable to treat the distance between red and green as 3-1=2 and that of red and blue as 1. Alternatively, though it might not be a bad idea to discretize numerical attributes, such discretization usually leads to a mass of information loss. If we choose to discretize lengths, 1.75m and 1.60m might be put in the same bin or two adjacent bins, depending on the granularity of intervals used. It will inevitably end up in different clustering results. Moreover, how to select an appropriate discretization algorithm is equally complicated. Therefore, the similarity/distance calculation for mixed data is not as intuitive as it appears and different data types have to be appropriately integrated.

Several attempts have been made to model the interdependence within categorical attributes (Wang et al. 2011) and within numerical attributes (Wang, She, and Cao 2013) individually. For instance, in Table 1, the partially coupled distance (*pcd* for short) between $u_3, u_5$ is $0.772$, which is larger than that between $u_3, u_4$ (i.e. $0.454$) and that between $u_5, u_6$ (i.e. $0.644$) by using couplings via frequency, co-occurrence and correlation. It shows that the individual couplings have effectively captured certain hidden information out of data. However, the *pcd* of $u_4, u_6$ (i.e. $0.596$) lies between that of $u_3, u_4$ and that of $u_5, u_6$, which makes the allocation of $u_4$ and $u_6$ still unclear. The reason is that the *pcd* only caters for the relationships among discrete and continuous variables separately, leading to limited improvement as the interdependent relationships are only partially revealed. More often, heterogeneous variables are associated with one another via diverse interactions, which must not be limited within certain types of data. In addition, it is usually neither accurate nor efficient to ignore the coupling context and coupling weights in modeling interactive relationships in mixed data.

So based on the traditional information table, how to describe the global interdependence across heterogeneous attributes? How to quantify the coupling context and coupling weights? How to explicitly represent the original data by teasing out the implicit relationships? No work that systematically addresses these issues has been reported in the literature due to the complexity and variety of data sets in their heterogeneous types. Thus, this paper proposes the context-based coupled interdependent attributes analysis on mixed data. The key contributions are as follows:

– We model the interdependence within discrete attributes (intra-coupling and inter-coupling), within numerical attributes (intra-coupling and inter-coupling), and across them by addressing the coupling context and weights.

– A coupled representation scheme is introduced for mixed-type objects, which integrates the intra-coupled and inter-coupled interactions of heterogeneous attributes with the original information table based on feature conversion.

– The proposed coupled representation for mixed data is compared with the traditional representation by applying data structure analysis, data clustering and density comparison, revealing that the interdependence of heterogeneous attributes is essential to the learning process.

## Related Work

Many papers address the issue of mixed data clustering. Huang (Huang 1998) firstly introduced the *k-prototype* algorithm, which is an extension to the k-means algorithm for clustering data with categorical values. More recently, *SpectralCAT* (David and Averbuch 2012) was proposed for data in any type via automatic categorization and spectral clustering. However, it performs discretization on numerical data, resulting in a mass of information loss. In addition, neither of them considers any relationship hidden in the mixed data. More and more researchers now focus on analyzing the implicit interdependence relationships for the mixed data. *Mixture* model was summarized in (Hunt and Jorgensen 2011), in which discrete and numerical variables follow multinomial distribution and multivariate Gaussian distribution, respectively. The coupling of numerical attributes is reflected by covariance matrix. Ahmad and Dey present *mADD* to incorporate the interactions of categorical values in clustering (Ahmad and Dey 2007). The coupling of discrete attributes is quantified by co-occurrence matrix. Despite the current research progress, no work has been reported that systematically takes into account the global relationships among their discrete and numerical attributes.

## Interdependence of Heterogeneous Attributes

The usual way to represent data is to use an information table $S = (U, A)$. Universe $U = \{u_1, \cdots, u_m\}$ consists of finite data objects in rows. Dimension $A = A^d \cup A^r$ is a finite set of attributes in columns, where $A^d = \{a_1^d, \cdots, a_{n_1}^d\}$ and $A^r = \{a_1^r, \cdots, a_{n_2}^r\}$ are sets of discrete attributes and

numerical attributes, respectively. Table 1 is an information table composed of six objects $\{u_1, \cdots, u_6\}$, two discrete attributes $\{a_1^d, a_2^d\}$ and two numerical attributes $\{a_1^r, a_2^r\}$.

We adopt space conversion techniques for discrete and numerical attributes, in which the coupling relationships are teased out by also addressing the interactions across them.

## Attributes Coupling for Discrete Data

Discrete data is characterized by the first $n_1$ columns of $S$ (i.e. nominal attribute set $A^d$ with values). The domain of discrete attribute $a_j^d$ is $\mathcal{A}_j^d = \{a_j^d.v_1, a_j^d.v_2, \cdots, a_j^d.v_{t_j}\}$, indicating that $a_j^d$ has $t_j$ distinct attribute values. In Table 1, for example, $\mathcal{A}_1^d = \{X, Y, Z\}$, $\mathcal{A}_2^d = \{\alpha, \beta\}$, and $n_1 = 2$.

**Coupling in Discrete Attributes** Firstly, we represent the discrete data by converting the original space spanned by $n_1$ discrete attributes into a new space whose dimensionality is $T = t_1 + \cdots + t_{n_1}$. In other words, each original discrete attribute $a_j^d$ is expanded by its $t_j$ attribute values. The discrete data can be reconstructed as an $m \times T$ matrix $S^d$, where each new object $u^d$ is a $T$-dimension row vector and the $(j-1+p)$-th column corresponds to attribute value $a_j^d.v_p$.

A common way is to assign the matrix entry $S^d[u, v]$ to be 1 if object $u$ contains attribute value $v$, and otherwise 0. Each object has one value for each attribute, so every object in $S^d$ contains exactly $n_1$ 1s. However, such granularity is too coarse-grain since only 1 and 0 are considered. The relationships neither within nor between discrete attributes (i.e. $a_j^d$ and $a_k^d$) is explicated. Therefore, we propose the method to explicitly reveal the interdependence relationships, and refine the boolean matrix $S^d \in \{0, 1\}^{m \times T}$ by a soft matrix $FS^d \in [0, 1]^{m \times T}$.

Accordingly, we define the entry of soft matrix $FS^d$ based on the pairwise similarity between values of each discrete attribute. Specifically, for each row vector (i.e. object) $u^d$ in $FS^d$ and each attribute value $a_j^d.v \in \mathcal{A}_j^d$, we have:

$$u^d[a_j^d.v] = sim(f_j^d(u), a_j^d.v), \tag{1}$$

where $u^d[a_j^d.v]$ refers to the component related to $a_j^d.v$ of vector $u^d$; $f_j^d(u)$ returns the value of discrete attribute $a_j^d$ for object $u$; $sim(f_j^d(u), a_j^d.v)$ denotes the similarity between two values $f_j^d(u)$ and $a_j^d.v$ of $a_j$, which is expected to reveal the interactions of discrete attributes. In Table 1, $u_3^d[Z] = sim(Y, Z)$, where $Z$ is a value of $a_1^d$ and $Y = f_1^d(u_3)$.

As pointed out by (Wang et al. 2011), the coupled nominal similarity not only considers the intra-coupled similarity within an attribute by discrepancy on value frequency, but also concerns the inter-coupled similarity between attributes via value co-occurrence aggregation. Their proposed similarity has been verified to outperform the state-of-the-art similarity measures for discrete data in extensive experiments. Therefore, we adapt this coupled nominal similarity to fit in the function $sim(\cdot)$ in Equation (1).

**Coupling Context and Coupling Weights** Despite the effectiveness of coupled nominal similarity, it simply treats every pair of attributes to have an equal coupling weight, which is not always fair. The reason is that in most cases, some attributes are strongly related to each other, while some might only weak related. Thus, we develop a way below to measure to what extent and in what context two attributes are coupled, and how corresponding weights are assigned to such coupling relationships.

Inspired by (Ienco, Pensa, and Meo 2012), we define the coupling context for a pair of discrete attributes based on the relevance and redundancy. Our goal is that for each discrete attribute (e.g. $a_j^d$), selecting a subset of relevant but not redundant attributes. The relevance of attributes (i.e. $a_j^d$, $a_k^d$) is measured by symmetrical uncertainty ($SU$), defined as:

$$SU_{a_j^d}(a_k^d) = 2 * \frac{H(a_j^d) - H(a_j^d | a_k^d)}{H(a_j^d) + H(a_k^d)}, \tag{2}$$

where $H(a_j^d)$ and $H(a_j^d | a_k^d)$ are entropy and condition entropy of variables $a_j^d$ and $a_k^d$. It ranges from 0 to 1, in which 0 indicates the independence of variables and 1 represents that the value of $a_j^d$ can be completely predicted by $a_k^d$.

On the other hand, attribute $a_l^d$ is regarded as redundant with respect to $a_j^d$ if both conditions I and II are satisfied.

$$\text{I} : SU_{a_j^d}(a_l^d) \leq SU_{a_j^d}(a_k^d), \;\; \text{II} : SU_{a_k^d}(a_l^d) \geq SU_{a_j^d}(a_l^d).$$

Condition I specifies that $a_j^d$ is more relevant to $a_k^d$ than to $a_l^d$. Condition II quantifies that the relevancy of $a_l^d$ and $a_k^d$ is larger than that of $a_l^d$ and $a_j^d$. When representing discrete attribute $a_j^d$, $a_l^d$ is redundant since its close neighbor $a_k^d$ is far more than enough to perform this task well. Thus, $a_l^d$ must be kept away from the coupling context of $a_j^d$.

To calculate the coupling context of $a_j^d$, we firstly rank a set of candidate attributes $\{a_k^d\}$ according to a descending order of $\{SU_{a_j^d}(a_k^d)\}$. Each redundant attribute $a_l^d$ is determined by conditions I and II and then removed from the candidate set. While the coupling context $C_j^d$ is obtained for $a_j^d$, we also record the corresponding $SU$ value as the coupling degree of attributes. For a target discrete attribute $a_j^d$, the coupling weight $\gamma_{jk}$ for every attribute in the coupling context ($a_k^d \in C_j^d$) is defined as the normalized $SU$ value.

$$\gamma_{jk} = \frac{SU_{a_j^d}(a_k^d) - \min\limits_{a_k^d \in C_j^d} SU_{a_j^d}(a_k^d)}{\max\limits_{a_k^d \in C_j^d} SU_{a_j^d}(a_k^d) - \min\limits_{a_k^d \in C_j^d} SU_{a_j^d}(a_k^d)}. \tag{3}$$

For those redundant attributes $\{a_l^d\}$, we take $\gamma_{jl} = 0$. Thus, the coupling weight $\gamma_{jk} \in [0, 1]$ for all $1 \leq j, k \leq n_1$.

**Coupling across Discrete and Numerical Attributes** Till now, we have paid full attention to the interdependence relationships within discrete data. Those hidden between discrete and numerical attributes should also be addressed. Since the frequency and co-occurrence of values are used for discrete attributes, the coupling of mixed attributes is extracted based on an appropriate categorization of each numerical attribute.

Table 2: Coupled Representation ($FS^d$) for Discrete Data

| data | $X$ | $Y$ | $Z$ | $\alpha$ | $\beta$ |
|---|---|---|---|---|---|
| $u_1$ | 0.333 | 0 | 0.143 | 0.099 | 0.500 |
| $u_2$ | 0.143 | 0.364 | 0.600 | 0.099 | 0.500 |
| $u_3$ | 0 | 0.500 | 0.364 | 0.667 | 0.099 |
| $u_4$ | 0.143 | 0.364 | 0.600 | 0.667 | 0.099 |
| $u_5$ | 0 | 0.500 | 0.365 | 0.667 | 0.099 |
| $u_6$ | 0.143 | 0.364 | 0.600 | 0.667 | 0.099 |

We propose to perform the categorization in an automatic manner, in which k-means is applied to each numerical attribute while calculating the validity index (David and Averbuch 2012). This process is repeated using increasing numbers of categories and terminated when the first local maxima is found. The number of categories, which reaches the best validity index as a local maxima, is selected. As a result, the set of numerical attributes $A^r$ is transformed into a collection of categorical scales $A^{r\to d}$. We then merge this converted categorical data with the original discrete data to form a new discrete attribute set $A^D = \{A^d, A^{r\to d}\}$.

The soft object-value matrix $FS^d$ is updated, each entry is calculated as the coupled nominal similarity with the coupling weight $\gamma$ based on the new attribute set $A^D$. Regarding Equation (1), we then obtain the new entry of $FS^d$.

$$FS^d(u,v) = u^d[a_j^d.v] = sim(f_j^d(u), a_j^d.v) \quad (4)$$
$$= \delta_j^{Ia}(f_j^d(u), a_j^d.v) \cdot \sum_{a_k \in A^D, k\neq j} \gamma_{jk} \cdot \delta_{j|k}(f_j^d(u), a_j^d.v, \mathcal{A}_k),$$

where $\delta_j^{Ia}$ is the intra-coupled similarity between attribute values, and $\delta_{j|k}$ is the inter-coupled relative similarity between them, all the detailed formulae are specified in (Wang et al. 2011). We use $a_k \in A^D$ to integrate the couplings across discrete and numerical attributes in the process of selecting the coupling context from the newly expanded attribute value set $\mathcal{A}_k$ and the corresponding weights $\gamma_{jk}$.

Finally, we obtain an $m \times T$ matrix $FS^d$, which is new representation for discrete data involving interdependence relationships. Each object-value entry is quantified by the similarity between discrete values by taking into account the context from both selected discrete and numerical attributes. For Table 1, we have the coupled representation for discrete data, as displayed in Table 2. The entry of $(u_3, Z)$ is 0.364, which means the coupled similarity between attribute values $Y = f_1^d(u_3)$ and $Z$ is 0.364. In the coupled similarity calculation, the context of attribute $a_1^d$ is only $a_2^d$, while the context of attribute $a_2^d$ consists of $a_1^d$ and categorized $a_2^{r\to d}$ with the coupling weights of $0.655$ and 0.345, respectively.

## Attributes Coupling for Numerical Data

Numerical data $S^r$ is described by the last $n_2$ columns of $S$, i.e. continuous attribute set $A^r$. Most current research assumes the independence of attributes when performing data mining or machine learning tasks. In real-world data, attributes are more or less interacted and coupled via explicit or implicit relationships. Wang et al. (Wang, She, and Cao 2013) propose a framework of the coupled attribute analysis to capture the global dependency of continuous attributes, which integrates the intra-coupled interaction within an attribute (i.e. the correlations between attributes and their own powers) and inter-coupled interaction among different attributes (i.e. the correlations between attributes and the powers of others) to form coupled representation for numerical objects by the Taylor-like expansion.

For numerical data, we apply the coupling model presented in (Wang, She, and Cao 2013) by extracting the coupling context and weights as well as integrating the interdependence across numerical attributes and discrete attributes.

**Coupling Context and Weights** The coupling context of numerical data is specified as intra-coupling and inter-coupling context. For a given numerical attribute $a_j^r$, the intra-coupling context is selected from the powers $\langle a_j^r\rangle^x$, whose value is the $x$-th power of the corresponding value of attribute $a_j^r$. The powers of other attributes $\langle a_k^r\rangle^x (k \neq j)$ compose the candidates for the inter-coupling context.

An alternative criterion to choose context for $a_j^r$ depends on the p-value of Pearson's correlated coefficient $cor(\cdot)$:

$$cor(a_j^r, a_k^r) = \frac{\sum_{u\in U}(f_j^r(u)-\mu_j^r)(f_k^r(u)-\mu_k^r)}{\sqrt{\sum_{u\in U}(f_j^r(u)-\mu_j^r)^2}\sqrt{\sum_{u\in U}(f_k^r(u)-\mu_k^r)^2}},$$

where $\mu_j^r$, $\mu_k^r$ are the respective mean values of $a_j^r$, $a_k^r$. If the p-value of $cor(\langle a_j^r\rangle^x, \langle a_j^r\rangle^y)$ is smaller than $0.05$, $\langle a_k^r\rangle^y$ is a significant component of context for $\langle a_j^r\rangle^x$, vice versa. Otherwise, $\langle a_k^r\rangle^y$ must be excluded from the context of $\langle a_j^r\rangle^x$.

For each attribute power $\langle a_j^r\rangle^x$, the intra-coupling weight is $\mathbf{w_1} \cdot R^{Ia}$, and the inter-coupling weight is $\mathbf{w_2} \cdot R^{Ie}$. Here, $R^{Ia}$ is the intra-coupling correlation matrix, whose entry is the significant correlation $cor(\cdot)$ between attributes and their own powers. $R^{Ie}$ is the inter-coupling correlation matrix, whose entry is the significant correlation $cor(\cdot)$ between attributes and others' powers. $\mathbf{w_1}$ and $\mathbf{w_2}$ are the predefined parameters to make the coupling interactions resemble Taylor expansion (Jia and Zhang 2008).

**Coupling across Numerical and Discrete Attributes** Likewise, we address the interdependence relationships across numerical and discrete attributes. Based on the $m\times T$ matrix $FS^d$ obtained for discrete data, each column of $FS^d$ can be treated as a numerical attribute since each entry is a continuous value rather than the original category in $S^d$. Next, we merge $FS^d$ with the numerical data $S^r$ to get $FS$, which has $T + n_2$ columns (i.e., the first $T$ columns come from $FS^d$ and the following $n_2$ columns are from $S^r$). These columns correspond to $T + n_2$ new numerical attributes $\{a_j^{dr}\}$.

On the basis of this $m \times (T + n_2)$ matrix $FS$, the interdependent attribute analysis on numerical data is conducted to explore the coupling context and coupling weights for each new continuous attribute $a_j^{dr}$ by analyzing the powers of $a_j^{dr}$ and their correlations. The output is an updated $m \times L \cdot (T + n_2)$ matrix $\widehat{FS}$, where $L$ is the maximal powers. The entry for object $u$ and attribute power $\langle a_j^{dr}\rangle^x$ is:

$$\widehat{FS}(u, \langle a_j^{dr}\rangle^x) = \mathbf{u}^T \cdot (\mathbf{w_1} R^{Ia}(\langle a_j^{dr}\rangle^x) + \mathbf{w_2} R^{Ie}(\langle a_j^{dr}\rangle^x)), \quad (5)$$

Table 3: Normalized Coupled Representation ($\widehat{FS}$) for Categorical and Numerical Data

| data | $\langle X\rangle^1$ | $\langle X\rangle^2$ | $\langle Y\rangle^1$ | $\langle Y\rangle^2$ | $\langle Z\rangle^1$ | $\langle Z\rangle^2$ | $\langle\alpha\rangle^1$ | $\langle\alpha\rangle^1$ | $\langle\beta\rangle^1$ | $\langle\beta\rangle^2$ | $\langle a_1^r\rangle^1$ | $\langle a_1^r\rangle^2$ | $\langle a_2^r\rangle^1$ | $\langle a_2^r\rangle^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $u_1$ | 1.81 | 1.77 | -1.77 | -1.81 | -1.55 | -1.54 | -1.64 | -1.64 | 1.64 | 1.64 | -0.06 | -0.74 | -1.63 | 1.67 |
| $u_2$ | 0.02 | 0.61 | -0.61 | -0.02 | 0.84 | 0.84 | -0.85 | -0.85 | 0.85 | 0.85 | -1.23 | 1.88 | -0.86 | 0.82 |
| $u_3$ | -0.93 | -0.74 | 0.75 | 0.93 | -0.49 | -0.49 | 0.51 | 0.51 | -0.51 | -0.51 | -0.58 | 0.10 | 0.49 | -0.54 |
| $u_4$ | 0.02 | -0.71 | 0.70 | -0.02 | 0.84 | 0.84 | 0.66 | 0.66 | -0.66 | -0.66 | -0.48 | -0.10 | 0.65 | -0.66 |
| $u_5$ | -0.93 | -0.34 | 0.34 | 0.93 | -0.49 | -0.49 | 0.62 | 0.62 | -0.62 | -0.62 | 1.44 | -0.23 | 0.65 | -0.60 |
| $u_6$ | 0.02 | -0.59 | 0.58 | -0.02 | 0.84 | 0.84 | 0.70 | 0.70 | -0.70 | -0.70 | 0.91 | -0.91 | 0.70 | -0.68 |

Table 4: Coupled Representation ($\widehat{RFS}$) for Objects

| data | $\widehat{a_1}$ | $\widehat{a_2}$ | $\widehat{a_3}$ | $\widehat{a_4}$ | $\widehat{a_5}$ | class |
|---|---|---|---|---|---|---|
| $u_1$ | 5.685 | -1.383 | -0.231 | -0.216 | 0.010 | $G_1$ |
| $u_2$ | 1.878 | 2.773 | 0.311 | 0.485 | -0.021 | $G_1$ |
| $u_3$ | -1.689 | -0.111 | 1.309 | -0.813 | -0.105 | $G_2$ |
| $u_4$ | -1.921 | 0.637 | -0.884 | -0.618 | 0.173 | $G_2$ |
| $u_5$ | -1.867 | -1.378 | 0.882 | 0.850 | 0.097 | $G_3$ |
| $u_6$ | -2.086 | -0.537 | -1.386 | 0.312 | -0.154 | $G_3$ |

where $\mathbf{u}^T$ is a row vector for object $u$ with $L\cdot(T+n_2)$ attribute values in $FS$, $R^{Ia}(\langle a_j^{dr}\rangle^x)$ and $R^{Ie}(\langle a_j^{dr}\rangle^x)$ are columns vectors to specify the correlations of $\langle a_j^{dr}\rangle^x$ with its own powers and other attributes' powers, respectively.

For instance, we obtain the normalized coupled representation for categorical and numerical data based on Table 1 and Table 2 when maximal power $L=2$, as shown in Table 3. There are in total 14 columns, corresponding to newly produced 14 attributes by considering all the interactions.

## Coupled Representation for Mixed Objects

After exploring the value expansions for discrete data and attribute powers for numerical data, we obtain the coupled representation: $m\times L\cdot(T+n_2)$ matrix $\widehat{FS}$. It explicitly exhibits the coupling interactions within discrete data, within numerical data, and across them. However, the dimension of $\widehat{FS}$ is as large as $L\cdot(T+n_2)$. If discrete attributes $A^d$ have many categories, namely $T=t_1+\cdots+t_{n_1}$ is large, $\widehat{FS}$ is expected to have quite a lot of columns due to the multiplication of $T$ and $L$, even if $L=2$, shown in Table 3.

To improve the efficiency of subsequent data mining and machine learning tasks, it is preferable to reproduce/select a smaller number of attribute representatives to capture the characteristics and structure of the original data as much as possible. Many attribute reduction methods, including PCA (Grbovic, Dance, and Vucetic 2012), spectral and diffusion map (David and Averbuch 2012), and MDS (Lancewicki and Aladjem 2014), may serve this purpose. We choose PCA to do the attribute selection on $\widehat{FS}$ matrix in this paper since it is non-parametric when compared to other strategies. Finally, a reduced matrix $\widehat{RFS}$ with attributes projected along the directions of relatively greater variance is obtained.

For Table 1, we get the final coupled representation as shown in Table 4. The dimensionality has been reduced to 5 from 14, but most of the information is reserved. Accordingly, we obtain the normalized Euclidean distance between $u_4$ and $u_6$ is $1.189$ based on $\widehat{RFS}$, larger than both distances between $u_4,u_3$ (i.e. $1.147$) and between $u_6,u_5$ (i.e. $1.162$). Similarly, the normalized distance between $u_3$ and $u_5$ (i.e. $1.220$) is also greater than them. It means that $u_4,u_6$ and $u_3,u_5$ are unlikely to be clustered together, which is consistent with the real situation and verifies that our proposed coupled representation is effective in capturing the implicit interdependent relationships.

Table 5: Description of Data Sets

| Data Set | Object | Attribute | Discrete | Class | Short Form |
|---|---|---|---|---|---|
| zoo | 101 | 16 | 15 | 7 | zo |
| echo | 131 | 10 | 2 | 2 | ec |
| teach | 151 | 5 | 4 | 3 | te |
| hepatitis | 155 | 19 | 13 | 2 | hp |
| heart | 270 | 13 | 8 | 2 | ha |
| mpg | 392 | 7 | 2 | 3 | mp |
| credit | 663 | 15 | 9 | 2 | cr |
| australian | 690 | 14 | 8 | 2 | au |
| statlog | 1000 | 20 | 13 | 2 | st |
| contra | 1473 | 9 | 7 | 3 | co |
| thyroid | 3428 | 20 | 14 | 3 | th |
| abalone | 4168 | 8 | 1 | 21 | ab |

## Empirical Study

Experiments are performed on 12 UCI data sets, shown in Table 5[2]. Two data representation schemes are compared: the original representation $S$ and the coupled representation $\widehat{RFS}$. As suggested in (Wang, She, and Cao 2013), maximal power $L$ is assigned to be 3 or 4 , whichever performs better. The number of runs is set to be $100$ to obtain average results with their sample standard deviations. The number of clusters is fixed to be the number of real classes in each data.

### Cluster Structure Analysis

Experiments are designed to specify the internal data structure. The data representation is evaluated with the given labels and clustering internal descriptors: Relative Distance (RD), Davies-Bouldin Index (DBI) (Davies and Bouldin 1979), Dunn Index (DI) (Dunn 1974), and Sum-Distance (SD). RD is the ratio of average inter-cluster distance upon average intra-cluster distance; SD is the sum of object distances within all the clusters. As the internal criteria seek the clusters with a high intra-cluster similarity and a low inter-cluster similarity, larger RD, larger DI, smaller DBI,

[2]The "Discrete" column lists the number of discrete attributes.

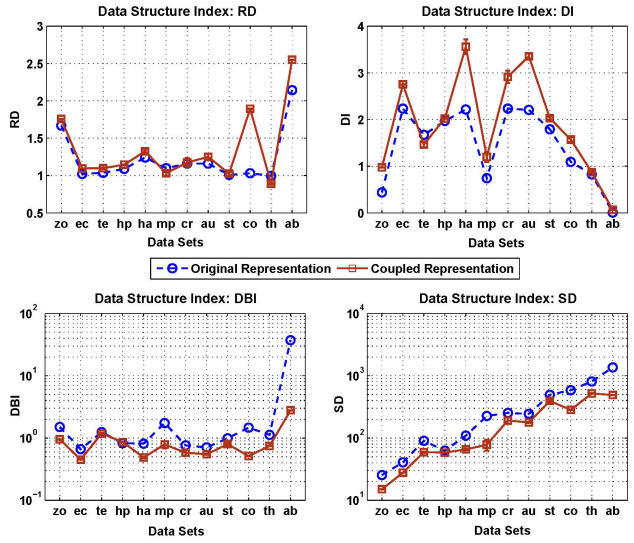


Figure 1: Data structure index comparisons on 12 data sets: average values with ± sample standard deviation error bars.

and smaller SD indicate the stronger cluster differentiation capability, which leads to a superior representation scheme.

The cluster structures produced by different representation schemes are analyzed on 12 data sets. The normalized results are shown in Figure 1. With the exception of only a few items (i.e. mp and th on RD, te on DI, hp on DBI), the corresponding RD and DI indexes for the coupled representation are larger than those for the original representation; and the associated DBI and SD indexes for the former are always smaller than those for the latter. It shows our proposed coupled representation, which effectively captures the global interactions within and between mixed attributes, is superior to the original in terms of differentiating objects in distinct clusters. All the results are supported by a statistical significant test at 95% significance level.

## Data Clustering Evaluation

Several clustering algorithms are in particular designed for mixed-type data, such as *k-prototype* (Huang 1998), *mADD* (Ahmad and Dey 2007), *mixture model* (Hunt and Jorgensen 2011), and *spectralCAT* (David and Averbuch 2012), which have been briefly introduced before. We also consider the *random* clustering as a baseline method. For our proposed strategy, we use k-means to perform the clustering task based on the coupled representation scheme, named *coupledMC*.

Table 6 reports the results in terms of an external measure: Accuracy. As described in (Cai, He, and Han 2005), the larger the accuracy, the better the clustering. The two highest measure scores of each experimental setting are highlighted in boldface. The row "Avg" displays the average values across all the 12 data sets. This table shows that *coupledMC* is always in the first two positions; in most cases, it outperforms all the other methods in all data sets. The maximal average improvement rate across all the data and all the methods is 69.76%, while the minimal is 12.80%. Statistical

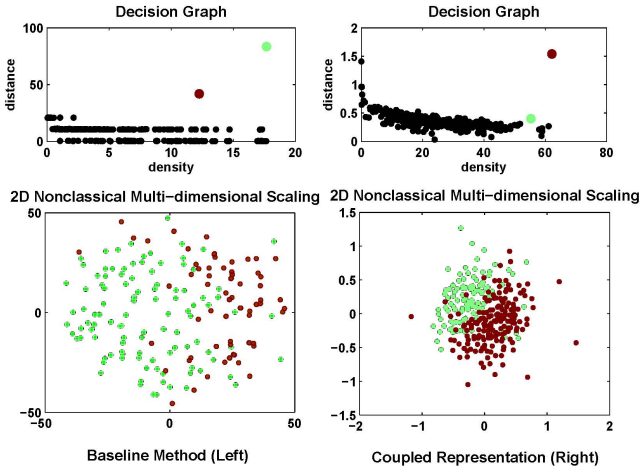


Figure 2: Clustering result on data set "heart".

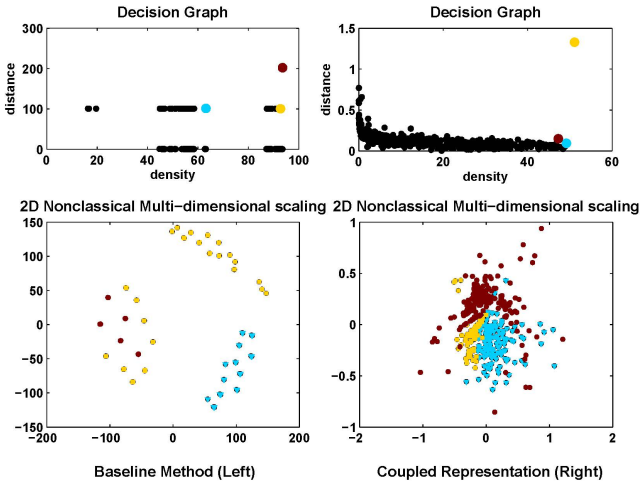


Figure 3: Clustering result on data set "mpg".

testing also supports that *coupledMC* performs better than others, at 95% significance level.

## Density Comparison

In this part, we use the newly published work (Rodriguez and Laio 2014) on clustering via distance and density to verify the superiority of our proposed coupled representation scheme. The original method is based on the idea that cluster centers are featured by a higher density than their neighbors and by a relatively large distance from points with higher densities. The authors have evidenced the power of algorithm in discovering clusters with any shape and dimensionality.

We adopt the mixed distance used by (Huang 1998) with this approach as the baseline method. The coupled representation scheme is also incorporated to make comparisons on the clustering quality. Figure 2 and Figure 3 show the decision graph to choose cluster centers and the 2D nonclassical multidimensional scaling of objects for data sets: heart and mpg, respectively. The solid circles in red, green, yel-

Table 6: Clustering Comparisons on Accuracy with ± Sample Standard Deviation

| Method | | *random* | *k-prototype* | *mADD* | *mixture* | *spectralCAT* | *coupledMC* |
|---|---|---|---|---|---|---|---|
| Accuracy | zoo | 0.260 ± 0.023 | 0.740 ± 0.095 | 0.764 ± 0.072 | **0.789 ± 0.074** | 0.765 ± 0.054 | **0.842 ± 0.008** |
| | echo | 0.537 ± 0.026 | **0.857 ± 0.016** | 0.763 ± 0.001 | 0.763 ± 0.002 | 0.682 ± 0.019 | **0.855 ± 0.051** |
| | teach | 0.388 ± 0.023 | 0.406 ± 0.028 | 0.423 ± 0.018 | 0.421 ± 0.027 | **0.439 ± 0.020** | **0.483 ± 0.030** |
| | hepatitis | 0.532 ± 0.021 | 0.698 ± 0.013 | 0.712 ± 0.034 | 0.735 ± 0.000 | **0.769 ± 0.024** | **0.806 ± 0.075** |
| | heart | 0.526 ± 0.017 | 0.781 ± 0.037 | **0.834 ± 0.004** | 0.804 ± 0.007 | 0.607 ± 0.068 | **0.870 ± 0.089** |
| | mpg | 0.366 ± 0.013 | 0.455 ± 0.006 | 0.479 ± 0.032 | 0.451 ± 0.001 | **0.622 ± 0.005** | **0.651 ± 0.044** |
| | credit | 0.515 ± 0.011 | **0.801 ± 0.001** | 0.543 ± 0.021 | 0.745 ± 0.006 | 0.556 ± 0.002 | **0.777 ± 0.106** |
| | australian | 0.513 ± 0.011 | **0.804 ± 0.016** | 0.641 ± 0.128 | 0.758 ± 0.032 | 0.556 ± 0.003 | **0.817 ± 0.035** |
| | statlog | 0.512 ± 0.008 | 0.522 ± 0.016 | **0.681 ± 0.000** | 0.603 ± 0.001 | 0.674 ± 0.007 | **0.707 ± 0.067** |
| | contra | 0.350 ± 0.006 | 0.406 ± 0.010 | 0.422 ± 0.010 | **0.435 ± 0.001** | 0.429 ± 0.005 | **0.443 ± 0.019** |
| | thyroid | 0.343 ± 0.005 | 0.364 ± 0.023 | 0.500 ± 0.059 | **0.739 ± 0.075** | 0.663 ± 0.070 | **0.904 ± 0.123** |
| | abalone | 0.072 ± 0.002 | 0.186 ± 0.003 | 0.177 ± 0.006 | 0.166 ± 0.005 | **0.187 ± 0.015** | **0.196 ± 0.007** |
| | Avg | 0.410 | 0.585 | 0.578 | **0.617** | 0.579 | **0.696** |

low and blue are the objects selected as cluster centers in the upper graphs (i.e. decision graph), and the points in different colors and shapes reflect the distinct clusters in the lower graphs. The density and distance scores for baseline method are not easily distinguishable, since many points have either the same density value or the same distance value. From the multi-dimensional scaling charts, we can see the number of projected points is rather limited in baseline method and they are mixed across different groups. In contrast, the distance-density-based clustering with coupling performs better in selecting cluster centers and thus partitioning objects better than the baseline method. So the coupled representation intrinsically exposes the data structure, in particular the hidden interdependence of heterogenous attributes.

## Conclusion

We have proposed the context-based coupled representation for mixed data via teasing out the relationships of interdependent attributes. The interdependence of heterogeneous attributes is exhibited as the couplings within discrete attributes, within continuous attributes, and across them. Several concepts, such as frequency, co-occurrence, categorization, correlation and powers, are used to build the coupling context and coupling weights for heterogeneous attributes. As a result, a coupled representation scheme is presented as a numerical matrix based on feature selection and conversion. Substantial experiments have verified that the coupled representation outperforms the original method on data structure, data clustering and density comparison, supported by statistical analysis. We are currently enriching the context-based coupled interdependent attribute analysis on mixed data by also addressing the couplings of objects and couplings of clusters. In the future, we will analyze the relationship between data characteristics and coupling interactions. In addition, semantic coupling based on domain knowledge or expert is expected to further improve the coupled representation of mixed data.

## References


Ahmad, A., and Dey, L. 2007. A k-mean clustering algorithm for mixed numeric and categorical data. *Data and Knowledge Engineering* 63:503–527.

Cai, D.; He, X.; and Han, J. 2005. Document clustering using locality preserving indexing. *IEEE TKDE* 17(12):1624–1637.

Cao, L.; Ou, Y.; and Yu, P. S. 2011. Coupled behavior analysis with applications. *IEEE TKDE* 24(8):1378–1392.

David, G., and Averbuch, A. 2012. SpectralCAT: categorical spectral clustering of numerical and nominal data. *Pattern Recognition* 45(1):416–433.

Davies, D., and Bouldin, D. 1979. A cluster separation measure. *IEEE TPAMI* 1(2):224–227.

Dunn, J. 1974. Well-separated clusters and optimal fuzzy partitions. *Cybernetics and Systems* 4(1):95–104.

Grbovic, M.; Dance, C. R.; and Vucetic, S. 2012. Sparse principal component analysis with constraints. In *AAAI 2012*, 935–941.

Huang, Z. 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery* 2(3):283–304.

Hunt, L., and Jorgensen, M. 2011. Clustering mixed data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1(4):352–361.

Ienco, D.; Pensa, R. G.; and Meo, R. 2012. From context to distance: learning dissimilarity for categorical data clustering. *ACM TKDD* 6(1):1.

Jia, Y., and Zhang, C. 2008. Instance-level semisupervised multiple instance learning. In *AAAI 2008*, 640–645.

Lancewicki, T., and Aladjem, M. 2014. Locally multidimensional scaling by creating neighborhoods in diffusion maps. *Neurocomputing* 139:382–396.

Plant, C. 2012. Dependency clustering across measurement scales. In *SIGKDD 2012*, 361–369.

Rodriguez, A., and Laio, A. 2014. Clustering by fast search and find of density peaks. *Science* 344(6191):1492–1496.

Wang, C.; Cao, L.; Wang, M.; Li, J.; Wei, W.; and Ou, Y. 2011. Coupled nominal similarity in unsupervised learning. In *CIKM 2011*, 973–978.

Wang, C.; She, Z.; and Cao, L. 2013. Coupled attribute analysis on numerical data. In *IJCAI 2013*, 1736–1742.