The Korean Communications in Statistics
Vol. 13 No. 3, 2006, pp.719-732

Cluster Analysis with Balancing Weight on Mixed-type Data[1)]

Seong San Chae[2)], Jong-Min Kim[3)] and Wan Youn Yang[4)]

Abstract

A set of clustering algorithms with proper weight on the formulation of distance which extend to mixed numeric and multiple binary values is presented. A simple matching and Jaccard coefficients are used to measure similarity between objects for multiple binary attributes. Similarities are converted to dissimilarities between i th and jth objects. The performance of clustering algorithms with balancing weight on different similarity measures is demonstrated. Our experiments show that clustering algorithms with application of proper weight give competitive recovery level when a set of data with mixed numeric and multiple binary attributes is clustered.

Keywords : Agglomerative clustering algorithm; mixed-type attribute; association coefficient.

1. Introduction

Clustering algorithms partition a data set into several disjoint groups such that objects in the same group are similar to each other according to some dissimilarity metric. Most clustering algorithms work with numeric data, but there has been work on clustering categorical data (Huang, 1998; Ordonez, 2003; Chae and Kim, 2005). Cluster analysis on categorical data is not as clear as on numeric data. Moreover, clustering on large and high dimensional numeric and categorical data is not easy to work.

The standard hierarchical clustering methods can handle data with numeric and categorical values (Everitt, 1993; Jain and Dubes, 1988) using dissimilarity suggested by Gower (1971), and other dissimilarity measures (Gowda and Diday, 1991; Gower and Legendre, 1986). However, the formulation of distance between i

1) This work was partially supported by RIC(R) grants from Traditional and Bio-Medical Research Center, Daejeon University (RRC04713, 2005) by ITEP.
2) Professor, Department of Applied Statistics, Daejeon University, Daejeon, 300-716, Korea,
3) Associate Professor, Division of Science and Mathematics, University of Minnesota, Morris, MN 56267, USA.
4) Associate Professor, Department of Applied Statistics, Kyungwon University, Seongnam, 461-701, Korea.
Correspondence : ywy@kyungwon.ac.kr

th and jth objects on the mixed-type data has not been studied extensively in clustering with mixed numeric and multiple binary values. Huang (1998) studied K -means algorithms for clustering large data sets with categorical values, suggesting the dissimilarity between ith and jth mixed-type objects. In his formulation, the weight was used to avoid favoring either type of attribute. However the weight was considered only on the categorical attributes and the range of the weight values was varied from 0.0 to infinite depending on the data.

This work focuses on clustering a set of data with mixed numeric and multiple binary values. New formulation of distance based on proper weight that competitive or superior to Gower (1971) is suggested. Rand's (1971) C statistic serves as the measure of the retrieval abilities(or, reproducibility) and the agreements(or, correspondence) of clustering algorithms based on how they partition the object space. When C is 1.0, the partition produced by clustering algorithm is identical to the structure within data treated, that is $0.0 \leq C \leq 1.0$. The extensive studies on using the concept of retrieval and agreement of on Rand's C statistics could be found in Chae, DuBien and Warde (2006). For the purpose of study, (0.0, -0.5) is known as single linkage, (0.0, 0.0) as average linkage, (0.0, 0.5) as complete linkage, (-0.25, 0.0) and (-0.5, 0.0) as representations of the flexible strategy, and (-0.5, 0.75) as recommendation by DuBien and Warde (1987), are used.

2. Gower and Suggested Distances

Gower (1966, 1967) has shown that distances satisfying triangle inequality from similarities can be done only if the matrix of similarities is nonnegative definite. With the nonnegative definite condition and with the similarity, S_{ij}, between ith and jth objects, $d_{ij} = \sqrt{1 - S_{ij}}$ has the properties of distance. Then function as Euclidean distance as dissimilarity measure between the ith and jth objects was defined by Gower (1971) as shown below.

$$d_{ij} = \sqrt{1 - S_{ij}} = \sqrt{1 - \frac{\sum_{l=1}^{r} w_{ijl} s_{ijl}}{\sum_{l=1}^{r} w_{ijl}}}$$

$$= \sqrt{1 - \frac{\sum_{l=1}^{c} w_{ijl} s_{ijl} + \sum_{l=c+1}^{r} w_{ijl} s_{ijl}}{\sum_{l=1}^{c} w_{ijl} + \sum_{l=c+1}^{r} w_{ijl}}}$$

$$= \sqrt{1 - \frac{\sum_{l=1}^{c} w_{ijl}\left(1 - \frac{|x_{il} - x_{jl}|}{R_l}\right) + \sum_{l=c+1}^{r} s_{ijl}}{\sum_{l=1}^{c} w_{ijl} + \sum_{l=c+1}^{r} w_{ijl}}}$$

$$= \sqrt{1 - \frac{c - \sum_{l=1}^{c} \frac{|x_{il} - x_{jl}|}{R_l} + \sum_{l=c+1}^{r} s_{ijl}}{c + \sum_{l=c+1}^{r} w_{ijl}}}$$

$$= \sqrt{\frac{\sum_{l=1}^{c} \frac{|x_{il} - x_{jl}|}{R_l} + \sum_{l=c+1}^{r} (w_{ijl} - s_{ijl})}{c + \sum_{l=c+1}^{r} w_{ijl}}}$$

where R_l is a range of lth variable and $w_{ijl} = 1.0$ for continuous variables. For binary variables, $s_{ijl} = 1.0$ if $x_i = x_j$ and $s_{ijl} = 0.0$ otherwise, and w_{ijl} is typically 1.0 or 0.0 depending on whether or not the comparison is considered valid for the lth variables, that is a formulation of Jaccard coefficient.

In Gower's formulation, it was not considered that either one of variables (continuous or binary) had significant effect on calculating distance obtained from continuous or discrete variables, assigning equal weight. Thus, clusters from applying clustering algorithms are easily occupied by one kind of variable types. To protect or enlarge this phenomena, it is necessary to consider assigning reasonable weights depending on the characteristic of data treated.

At this point, we define reasonable and comparable dissimilarity measure between ith and jth objects as

$$d_{ij}^{*} = \tau_{ij} \sum_{l=1}^{c} \frac{1}{c}\left(\frac{|x_{il} - x_{jl}|}{R_l}\right) + (1 - \tau_{ij})\sqrt{1 - A_{ij}}$$

$$= \tau_{ij} \sum_{l=1}^{c} \frac{1}{c}\left(\frac{|x_{il} - x_{jl}|}{R_l}\right) + (1 - \tau_{ij}) \sqrt{1 - \frac{\sum_{l=c+1}^{r} s_{ijl}}{\sum_{l=c+1}^{r} w_{ijl}}}$$

where τ_{ij} is a balancing weight to avoid favoring either type of attributes or a dominating weight to enlarge favoring one type of attributes, satisfying $0.0 \leq \tau_{ij} \leq 1.0$, R_l is a range of lth variable in quantitative values. For binary values, $\sum_{l=c+1}^{r} s_{ijl}$ and $\sum_{l=c+1}^{r} w_{ijl}$ are varied depending on the similarity measures

preferred by researchers. The weight value of τ_{ij} for each pair of ith and jth objects will differently effect on the result of clustering process. With this formulation, A_{ij} might be varied as simple matching, Jaccard, and Yule coefficients. Weights of 0.0 are assigned when variable l is unknown for one or both variables.

In finding the weight value of τ_{ij} for each pair of ith and jth objects, let ρ_{ij}^c and ρ_{ij}^d be any reasonable similarity measures for the quantitative and the multiple binary variables, respectively. From the equation above, let $d_{ij}^* = \tau_{ij} d_{ij}^c + (1-\tau_{ij}) d_{ij}^d$.

Then the τ_{ij} for each pair of ith and jth objects is assigned as

$$\tau_{ij} = \begin{cases} 1.0 - \dfrac{|\rho_{ij}^c|}{|\rho_{ij}^c| + |\rho_{ij}^d|}, & \text{if} \quad 1.0 < \dfrac{|\rho_{ij}^c|}{|\rho_{ij}^d|} \\ 1.0 - \dfrac{|\rho_{ij}^d|}{|\rho_{ij}^c| + |\rho_{ij}^d|}, & \text{if} \quad 1.0 > \dfrac{|\rho_{ij}^c|}{|\rho_{ij}^d|} \\ 0.5, & \text{if} \quad |\rho_{ij}^c| = |\rho_{ij}^d|, \end{cases}$$

where $-1.0 \le \rho_{ij}^c,\ \rho_{ij}^d \le 1.0,\ i = 2,3,\ldots,n,\ j = 1,2,\ldots,n-1,\ i > j$.

The rationale behind this formulation is as follows: Euclidean distance is a measure of dissimilarity and, in order to have equivalence with similarity measures, it is necessary to divide it by the range. Because the significance of distance from either one of variables (quantitative or binary) is different, the τ_{ij} is designed to balance these cases by including an option as a weight. Depending on the pairwise comparison between ith and jth objects, the weights are changed and are used to avoid favoring either type of variables.

In this study, Pearson correlation coefficient, ρ_{ij}^c, for the quantitative variable and product moment correlation coefficient, ρ_{ij}^d, for the multiple binary variables, are used. However, any reasonable measures might be used instead of ρ_{ij}^c and ρ_{ij}^d, if they correspond each other in measuring similarity between the ith and jth objects within different types of variables.

3. Design of Simulation Study

Suppose a sample of size N is observed with m variables on each data point. The $N \times m$ matrix of measurements, say X, might be $X_{(N\times m)} = X^N = [X_1 X_2 \cdots X_{N-1} X_N]$ where X_i represents a $m \times 1$ vector of measurement on the

ith objects. Then a cluster, y_h, is simply a nonempty subset of the object space, and a clustering, $Y = (y_1, y_2, \ldots, y_K)$, is any partition of the object space, if the following three conditions hold:

(1) For every $y_h \in Y, y_h \not\subset \varnothing$;
(2) If $y_h, y_l \in Y$ and $y_h \neq y_l$, then $y_h \cap y_l = \varnothing$;
(3) $\cup_{h=1}^{K} y_h = X$.

Some notations useful for understanding a cluster, a clustering, an hierarchy and an agglomerative clustering method can be found in DuBien and Warde (1987).

Let Y represent the "true" structure of the N data points with number of clusters, K, and $Y^{[N,K]}$ be a certain type of rearrangement of Y with K clusters. Let Y^a denote a clustering that result from applying an agglomerative clustering algorithm to the N data points with number of clusters, K. Then Rand's (1971) $C(Y, Y^a)$ is a measure of the "retrieval" ability of the agglomerative clustering algorithm to the true structure for K. And $C(Y^a, Y^b)$ is a measure of the "agreement" between 'a' and 'b' clustering algorithms.

Investigating the "retrieval" ability and "agreement" of clustering algorithms using Rand's (1971) C statistic, some of the structural parameters considered in this study are defined as follows:

1) N, the number of data points in $X_{N\times r} = (Z_c : Z_d)$, where Z_c is a $N\times c$ matrix and Z_d is a $N\times d$ matrix;
2) c and d are the numbers of continuous and binary variables, respectively, with $r = c + d$;
3) n_k, the size of the kth cluster generated from each population;
4) δ, the distance between mean vectors;
5) R, the correlation matrix of the form,

$$R=\begin{pmatrix} A & B & B \\ B & A & B \\ B & B & A \end{pmatrix}, \quad A=\begin{pmatrix} 1.0 & \rho & \rho \\ \rho & 1.0 & \rho \\ \rho & \rho & 1.0 \end{pmatrix}, \quad B=\begin{pmatrix} \eta & \eta & \eta \\ \eta & \eta & \eta \\ \eta & \eta & \eta \end{pmatrix},$$

where $\rho = 0.5, 0.8$ and $\eta = -0.2, 0.2$.

For convenience, the number of data points in Z_c is $N = 60$, the number of variables is $c = 9$, and the number of clusters is $k = 3$ in this study. Then a brief summary of data structure may be outlined as follows:

$$Z_c \sim MVN(\mu_g, \Sigma)$$

where $g=1,..,k$, $i=1,2,...,N$. The number of data points are split into $k=3$ populations of size $(n_1-n_2-n_3)=[(20-20-20),\ (30-20-10)]$, and the mean vectors μ_g, $g=1,2,3$ are constrained by an equilateral triangle spatial configuration,

$$\mu_1' = (0.0 \quad \delta_c \quad \delta_c \quad \delta_c \quad 0.0 \quad \delta_c \quad \delta_c \quad \delta_c \quad 0.0),$$
$$\mu_2' = (\delta_c \quad 0.0 \quad \delta_c \quad \delta_c \quad \delta_c \quad 0.0 \quad 0.0 \quad \delta_c \quad \delta_c),$$
$$\mu_3' = (\delta_c \quad \delta_c \quad 0.0 \quad 0.0 \quad \delta_c \quad \delta_c \quad \delta_c \quad 0.0 \quad \delta_c)$$

so that the Euclidean distances between mean vectors are $\delta=\delta_c\times\sqrt{6.0}$. For this study we set $\delta=4.0, 6.0$.

Currently, computer programs which generate "multiple binary" data treat Z_{di} as a multiple binary random variable are not available. One is not able to randomly generate a multiple observation in which each variable is an outcome of a Bernoulli trial. There is no correlation structure associated with the generation. However, mixed type attributes are considered to have strong relationship in real set of data. For convenience, a set of multiple binary samples was generated from a multivariate normal random variable with the reduced correlation matrix of $R_{6\times 6}$ with mean vector μ_g, $g=1,2,3$ for only six variables among nine. Each variable for the multiple binary attributes($d=6$) was transformed to a Bernoulli random variable by translating the normal z_c value for each variate to "1" if $z_c \le \delta-1.0$, "0" otherwise, for $g=1$; "1" if $z_c \le \delta$, "0" otherwise, for $g=2$; "1" if $z_c \le \delta+1.0$, "0" otherwise, for $g=3$. A multiple binary data, Z_d, was generated from a multivariate normal random variable with the reduced correlation matrix $R_{6\times 6}$. Finally, a set of mixed-type data, $X_{N\times r}=(Z_c : Z_d)$, with three clusters was generated.

With this design, the results from clustering algorithms applied to generated data were observed by investigating the "retrieval" ability and "agreement" of clustering algorithms using Rand's (1971) C statistic. The values of C representing the recovery of true structure for the six $(\beta,\ \pi)$ clustering algorithms were generated by the following steps:

1) An object space $X_{N\times r}$ of data points was generated;
2) The distance converted using the formula $d_{ij}=\sqrt{1-S_{ij}}$, where S_{ij} is the similarity between each pair of data points in X, was computed and stored in lower triangular matrix order by rows as the vector D_1 for Gower's method;
3) The distance d_{ij}^c, between each pair of data points in Z_c, was computed and stored as the vector D_2^c;
4) The distance converted from association coefficient using the formula

$d_{ij}^{d} = \sqrt{1 - A_{ij}}$, where A_{ij} is the similarity between each pair of data points in Z_d, was computed and stored as the vector D_2^d;

5) ρ_{ij}^{c} and ρ_{ij}^{d} between each pair of data points in Z_c and Z_d, were computed and stored as the vectors $\Lambda_c$and Λ_d, respectively;

6) In favor of balancing, the τ_{ij} for each pair of ith and jth objects is assigned as,

$$\tau_{ij} = \begin{cases} 1.0 - \dfrac{|\rho_{ij}^{c}|}{|\rho_{ij}^{c}| + |\rho_{ij}^{d}|}, & \text{if} \quad 1.0 < \dfrac{|\rho_{ij}^{c}|}{|\rho_{ij}^{d}|} \\ 1.0 - \dfrac{|\rho_{ij}^{d}|}{|\rho_{ij}^{c}| + |\rho_{ij}^{d}|}, & \text{if} \quad 1.0 > \dfrac{|\rho_{ij}^{c}|}{|\rho_{ij}^{d}|} \\ 0.5, & \text{if} \quad |\rho_{ij}^{c}| = |\rho_{ij}^{d}|, \end{cases}$$

where ρ_{ij}^{c} and ρ_{ij}^{d} in the vectors $\Lambda_c$and Λ_d, respectively;

7) Then the distance d_{ij}^{*}, between each pair of data points in X, was computed and stored in lower triangular matrix order by rows as the vector D_2 for our suggested method;

8) Each of the six clustering algorithms was applied to D_1 and D_2 to produce a clustering, Y^a;

9) Each of the six clustering algorithms was applied to D_1 and D_2 to produce clusterings, Y^a for $K = 3$;

10) For each of the clustering, $Y^a$from above steps, $C(Y, Y^a)$ for retrieval of 'a' algorithm, $C(Y^a, Y^b)$ for agreement between 'a' algorithm and 'b' algorithm were calculated for the six clustering algorithms.

For each setting of the structural parameters, the above sequence of steps was replicated 100 times. Then the sample mean and variance of the C statistic, $\overline{C}$, were computed for each of the six agglomerative clustering algorithms. Consequently, $\overline{C}$ result is examined and compared to quantify the "retrieval" ability for each of the clustering algorithms, and the "agreement" between clustering algorithms based on Gower's d_{ij} and our d_{ij}^{*} calculated from mixed-type attributes for each setting of the structural parameters.

4. Simulation Results and Discussions

Based on the data from each setting of the various structural parameters, all results from the comparative study will be discussed with agglomerative clustering algorithms defined with (β, π) and association coefficients. However, discussion is

made only on the results using simple matching and Jaccard coefficients since the retrieval is not good when Yule coefficient is used. The results from the simulation study are not independent of the fixed structural parameters which were specified previously. The results based on the settings $(n_1 - n_2 - n_3) = (20 - 20 - 20)$ will be discussed since the recovery levels of clustering algorithms were not significantly different in our simulation study. Since the results from the single linkage are the worst among the six clustering algorithms in simulation study, it is excluded from further discussion. For our convenience, the results using the vectors, D_1 and D_2, are only presented for comparison on two dissimilarity measures.

In <Table 1>, the results are given as $\overline{C}$ computed over 100 replications for each setting of the various structural parameters and for each of the five agglomerative clustering algorithms using continuous variable only, binary variable only, and Gower's distance, d_{ij}, with both types of variables, that based on $\sum_{l=1}^{c} \frac{|x_{il} - x_{jl}|}{R_l}$, Jaccard and simple matching coefficients, respectively. As shown in <Table 1>, the retrieval abilities of clustering algorithms are varied depending on the choice of variable, (dis)similarity types on the set of data generated with settings of structural parameters. The recovery levels of clustering algorithms using Gower's d_{ij} decrease or increase depending on the choice of clustering algorithms compare to using only one type of variables.

For the comparison with using Gower's d_{ij}, the recovery levels of clustering algorithms with our suggested distance, d_{ij}^* are given in the form of $\overline{C}(Y, Y')$ for the cases of Jaccard, and simple matching coefficients in <Table2>. As presented in <Table 2>, the clustering algorithms, (-.25, .0) and (-.5, .0) that are flexible strategies, with Jaccard and simple matching coefficients give high recovery levels compare to the results from other combinations of clustering algorithms and methods of weight.

<Table 1> The $\overline{C}(Y, Y^a)$ representing retrieval from clustering algorithms using Gower's d_{ij} and d_{ij}^*

	δ	4.0				6.0			
	ρ	0.5		0.8		0.5		0.8	
Distance	Algorithm/ η	-.2	.2	-.2	.2	-.2	.2	-.2	.2
Conti.-only	(.0,.0)	.6340	.6847	.5717	.6434	.9433	.9328	.9209	.9035
	(.0,.5)	.6394	.6721	.5822	.5978	.9214	.8683	.7972	.7706
	(-.25,.0)	.9144	.8944	.8922	.8916	.9984	.9988	.9986	.9980

	(-.5,.0)	.9339	.9305	.9341	.9309	.9985	.9981	.9977	.9979
	(-.5,.75)	.8426	.8439	.7990	.7963	.9908	.9880	.9778	.9798
Binary only (Simple)	(.0,.0)	.7183	.7136	.7098	.7105	.7819	.7480	.7597	.7371
	(.0,.5)	.7182	.7165	.7246	.7086	.7529	.7395	.7458	.7438
	(-.25,.0)	.7630	.7521	.7387	.7272	.8133	.7886	.7716	.7579
	(-.5,.0)	.7675	.7523	.7257	.7299	.7977	.7843	.7564	.7441
	(-.5,.75)	.5992	.6063	.6135	.6161	.6146	.6308	.6366	.6478
Binary only (Jaccard)	(.0,.0)	.4422	.4413	.4534	.4672	.4726	.4433	.5043	.5182
	(.0,.5)	.5988	.5677	.5785	.5828	.6633	.6081	.6460	.6137
	(-.25,.0)	.6120	.5771	.6005	.5739	.7316	.6940	.7086	.6801
	(-.5,.0)	.7080	.6651	.6638	.6274	.7372	.6980	.7018	.6712
	(-.5,.75)	.5964	.6030	.6126	.6153	.6146	.6308	.6366	.6478
Mixed Gower (Jaccard)	(.0,.0)	.7111	.6922	.7143	.6949	.7427	.7330	.7408	.7324
	(.0,.5)	.7076	.7036	.6913	.6831	.7289	.7223	.7148	.7054
	(-.25,.0)	.7275	.7279	.7277	.7266	.8210	.8670	.7831	.8157
	(-.5,.0)	.8839	.8612	.8254	.7927	.9933	.9916	.9931	.9928
	(-.5,.75)	.8182	.8109	.7790	.7823	.9891	.9774	.9759	.9590
Mixed (Simple)	(.0,.0)	.6612	.6985	.5982	.6960	.8964	.9661	.8754	.9678
	(.0,.5)	.5777	.6285	.5661	.6062	.6394	.8079	.5940	.7361
	(-.25,.0)	.9431	.9606	.9296	.9607	.9993	.9995	.9973	.9992
	(-.5,.0)	.9595	.9726	.9615	.9692	.9992	.9998	.9974	.9996
	(-.5,.75)	.7501	.8129	.7013	.7801	.9230	.9852	.8782	.9703
Mixed (Jaccard)	(.0,.0)	.6120	.6870	.5943	.6711	.8781	.7557	.8780	.9536
	(.0,.5)	.5861	.6150	.5597	.6022	.6388	.7836	.6042	.7311
	(-.25,.0)	.9407	.9528	.9183	.9579	.9996	.9991	.9968	.9995
	(-.5,.0)	.9526	.9645	.9467	.9629	.9984	.9993	.9977	.9991
	(-.5,.75)	.7205	.8003	.6986	.7551	.9230	.9822	.8566	.9640

<Table 2> The $\overline{C}(Y^a, Y^b)$ representing agreement from clustering algorithms using Gower's d_{ij} and d_{ij}^*

		δ	4.0				6.0			
		ρ	0.5		0.8		0.5		0.8	
Distance	Algorithm	Algorithm/ η	-.2	.2	-.2	.2	-.2	.2	-.2	.2
Mixed Gower (Jaccard)	(.0,.0)	(.0,.5)	.9126	.9025	.8873	.8877	.9606	.9519	.9642	.9460
		(-.25,.0)	.9187	.9029	.9127	.8958	.8882	.8402	.9147	.8789
		(-.5,.0)	.7745	.7677	.8198	.8121	.7377	.7313	.7368	.7277
		(-.5,.75)	.7956	.7740	.8179	.7870	.7419	.7397	.7459	.7473
	(.0,.5)	(-.25,.0)	.9203	.9195	.8954	.8957	.8831	.8356	.9048	.8656
		(-.5,.0)	.7716	.7822	.8070	.8168	.7244	.7210	.7126	.7013
		(-.5,.75)	.7994	.7898	.8048	.7904	.7280	.7294	.7220	.7238
	(-.25,.0)	(-.5,.0)	.8047	.8153	.8515	.8702	.8180	.8622	.7805	.8127

		(-.5,.75)	.8168	.8131	.8466	.8257	.8202	.8731	.7905	.8279
	(-.5,.0)	(-.5,.75)	.8301	.8502	.8359	.8375	.9882	.9753	.9777	.9596
Mixed (Simple)	(.0,.0)	(.0,.5)	.5740	.5870	.5474	.5714	.6325	.8044	.5890	.7267
		(-.25,.0)	.6581	.6901	.6035	.6919	.8963	.9657	.8734	.9673
		(-.5,.0)	.6599	.6973	.5907	.6884	.8960	.9660	.8737	.9674
		(-.5,.75)	.5864	.6313	.5490	.6229	.8462	.9517	.7852	.9388
	(.0,.5)	(-.25,.0)	.5763	.6297	.5663	.6061	.6395	.8079	.5937	.7358
		(-.5,.0)	.5766	.6292	.5655	.6048	.6395	.8078	.5938	.7361
		(-.5,.75)	.5693	.6127	.5658	.5913	.6331	.8034	.5900	.7252
	(-.25,.0)	(-.5,.0)	.9344	.9566	.9147	.9500	.9988	.9995	.9967	.9989
		(-.5,.75)	.7374	.8055	.6900	.7732	.9228	.9848	.8767	.9696
	(-.5,.0)	(-.5,.75)	.7427	.8140	.6965	.7775	.9226	.9850	.8768	.9701
Mixed (Jaccard)	(.0,.0)	(.0,.5)	.5515	.5701	.5546	.5556	.6193	.7710	.5979	.7189
		(-.25,.0)	.6081	.6795	.5898	.6674	.8780	.9548	.8759	.9532
		(-.5,.0)	.6085	.6820	.5904	.6611	.8777	.9550	.8772	.9528
		(-.5,.75)	.5419	.6260	.5333	.5913	.8248	.9393	.7636	.9199
	(.0,.5)	(-.25,.0)	.5878	.6168	.5574	.6075	.6388	.7837	.6044	.7309
		(-.5,.0)	.5868	.6174	.5597	.6007	.6389	.7835	.6041	.7309
		(-.5,.75)	.5700	.5976	.5529	.5894	.6363	.7809	.5965	.7246
	(-.25,.0)	(-.5,.0)	.9344	.9458	.9135	.9391	.9984	.9989	.9957	.9988
		(-.5,.75)	.7078	.7866	.6898	.7443	.9229	.9820	.8549	.9640
	(-.5,.0)	(-.5,.75)	.7126	.7984	.6952	.7457	.9221	.9819	.8548	.9631

With the design described, more similar clusterings are retrieved by applying clustering algorithms when d_{ij}^{*} with methods of weight is used. The values of $\overline{C}$ from the clustering algorithms show essential differences depending on the choice of distance measures and methods of weight between the ith and jth objects. This implies that the use of d_{ij}^{*} has an effect on the recovery of the true clusters in the data from mixed-type attributes.

5. Application to Real Data and Discussion

The use of different distances prior to applying the agglomerative clustering algorithm are investigated on the financial performance data(Affi and Clark, 1990). For convenience, the 25 companies with 7 variables was used as the data set with three clusters that identified by different kinds. Details on 7 variables might be found in Affi and Clark (1990). To obtain a set of mixed data, the correlation matrix was calculated and principal component analysis was applied. Then it was found that 3 variables ROR5(percent rate of return on total capital), NPM1(percent net profit margin) and PAYOUTR1(annual dividened divided by the 12-months earnings per share) were different from the other four variables. At this point,

ROR5 and PAYOUTR1 were transformed to binary variables '1' if the values of each variable are less than the medians of each variables; '0', otherwise. And NPM1 was transformed to binary variables '1' if the values of each variable are larger than the medians of each variables; '0', otherwise.

For each of companies, the sizes of clusters to which it belongs are (14-5-6) in the Chemical, Health, and Supermarket with 25 companies. Hence the clusters are identified by the six agglomerative clustering algorithms using Gower's distance, d_{ij}, and our suggested distance, d_{ij}^{*}, based on Jaccard coefficients between ith and jth companies. For comparison, the recovery levels of the six clustering algorithms on the data with continuous variables originally given by Affi and Clark (1990) are presented in the following <Tables 3-4>.

As shown in <Table 3>, the recovery level of clustering algorithms on the "company defined clusters" is increased or decreased by using our suggested distance depending on association coefficients. When the results of using simple matching coefficient are considered, the recovery levels of clustering algorithms, complete linkage, (.0, .5), and (-.5, .75) are high with d_{ij}^{*}, while the recovery levels of average linkage-(.0, .0), and one of flexible strategies-(-.25, .0) are high with d_{ij}. If simple matching coefficient with d_{ij}^{*} is used, the results are better than Jaccard coefficient with Gower's d_{ij} except for average linkage-(.0, .0). Among the five clustering algorithms, a clustering algorithm suggested by DuBien and Warde (1979) gives great performance on using simple matching and Jaccard coefficients with our suggested distance, d_{ij}^{*}, if the clusters identified by Affi and Clark (1990) are well defined.

<Table 3> The $\overline{C}(Y, Y^{a})$ values representing retrieval from applying clustering algorithms on the data from Affi and Clark (1990)

	Gower's d_{ij}	Suggested d_{ij}^{*}	
Algorithm/ac	Jaccard	Simple	Jaccard
(.0,.0)	.6800	.5900	.5900
(.0,.5)	.5533	.9367	.7300
(-.25,.0)	.6800	.7200	.5900
(-.5,.0)	.6267	.7200	.6967
(-.5,.75)	.5233	.9367	.9367

In <Table 4>, the agreements of the clusterings from the six clustering algorithms are different for the cases using two distances, d_{ij} and d_{ij}^{*}. By using those agreements among clusterings, a natural basis for organizing companies

depending on their financial performance might be obtained depending on the characteristic of data. In recovering the clusters that identified by Affi and Clark (1990), the reproducibility of clusters using d_{ij}^{*} are better than using d_{ij} if simple matching coefficient is used.

The highest recovery levels in <Table 3> were given by using complete linkage, (.0, .5), and (-.5, .75) with simple matching coefficient. This implies that two algorithms reproduce the clusters defined on the data more closely than the other algorithms. Further, the results from two clustering algorithms agree perfectly since the agreement is 1.0, as presented in <Table 4>.

<Table 4> The $\overline{C}(Y^{a}, Y^{b})$ values representing agreement between clustering algorithms on the data from Affi and Clark (1990)

		Gower's d_{ij}	Suggested d_{ij}^{*}	
Algorithm	Algorithm/ac	Jaccard	Simple	Jaccard
(.0,.0)	(.0,.5)	.7400	.6267	.7767
	(-.25,.0)	1.000	.7767	1.000
	(-.5,.0)	.7000	.7767	.7867
	(-.5,.75)	.5633	.6267	.6267
(.0,.5)	(-.25,.0)	.7400	.7500	.7767
	(-.5,.0)	.7400	.7500	.8300
	(-.5,.75)	.7633	1.000	.7500
(-.25,.0)	(-.5,.0)	.7000	1.000	.7867
	(-.5,.75)	.5633	.7500	.6267
(-.5,.0)	(-.5,.75)	.7633	.7500	.6800

6. Conclusion

This work focuses on clustering mixed numeric and multiple binary values. The dissimilarity measures between ith and jth objects as d_{ij}^{*} is suggested instead of d_{ij}. In calculating the dissimilarity, τ_{ij} is a weight to balance the two parts to avoid favoring either type of attributes, satisfying $0.0 \leq \tau_{ij} \leq 1.0$. For binary values, similarity measures, A_{ij}, are varied depending on the association coefficients defined by researchers. The weight value of τ_{ij} for ith and jth objects will differently effect on the result of clustering process.

The results of applying clustering algorithms on d_{ij}^{*} and d_{ij} were compared. As shown in the results from simulations with mixed-type data sets, the retrieval ability of the clustering algorithms was significantly improved using d_{ij}^{*} using Jaccard coefficient. Then simple matching coefficient instead of Jaccard coefficient

was used to calculating the distance d_{ij}^{*} in comparison with Gower's distance, d_{ij}. by using the mixed data generated from Affi and Clark (1990). A clustering algorithm, (-.5, .75), suggested by DuBien and Warde (1979) gives great per -formance on using Jaccard and simple matching coefficients with our suggested distance, d_{ij}^{*}.

In the concept of agreements among several different clusterings, we might have more confidence in identifying the clusters using measures of distance d_{ij}^{*} instead of d_{ij}. Further, the highest recovery levels were given by using average linkage and (-.5, .75), implying that two algorithms reproduce the clusters defined on the data more closely than the other algorithms.

References

[1] Affi, A.A. and Clark, V. (1990). *Computer-Aided Multivariate Analysis.* Van Nostrand Reinhold Company, New York.

[2] Asparoukhov, O.K. and Krzanowski, W.J. (2001). A comparison of discriminant procedures for binary variables. *Computational Statistics & Data Analysis,* Vol. 38, 139-160.

[3] Chae, S.S., DuBien J.L. and Warde, W.D. (2006). A method of predicting the number of clusters using Rand's statistic. *Computational Statistics & Data Analysis,* Vol. 50, 3531-3546.

[4] Chae, S.S. and Kim, J.I. (2005). Cluster analysis using principal coordinates for binary data. *The Korean Communications in Statistics,* Vol. 12, 683-696.

[5] DuBien, J.L. and Warde, W.D. (1987). A comparison of agglomerative cluster -ing methods with respect to noise. *Communications in Statistics, Theory and Method,* Vol. 16, 1433-1460.

[6] Everitt, B. (1993). *Cluster Analysis.* 3rd edition, John Wiley & Sons.

[7] Gowda, K.C. and Diday, E. (1991). Symbolic clustering using a new dissimi -larity measures. *Pattern Recognition,* Vol. 24, 567-578.

[8] Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika,* Vol. 53, 325-338.

[9] Gower, J.C. (1967). A comparison of some methods of cluster analysis. *Bio -metrics,* Vol. 23, 623-637.

[10] Gower, J.C. (1971). A general coefficient of similarity and some of its proper -ties. *Biometrics,* Vol. 27, 857-871.

[11] Gower, J.C. and Legendre, P. (1986). Metric and Euclidean properties of dis -similarity coefficients. *Journal of Classification,* Vol. 3, 5-48.

[12] Huang, Z. (1998). Extensions to the k-means algorithms for clustering large

data sets with categorical values. *Data Mining and Knowledge Discovery,* Vol. 2, 283-304.

[13] Jain, A.K. and Dubes, R.C. (1988). *Algorithms for Clustering Data.* Prentice Hall.

[14] Lee, J.J. (2005). Discriminant analysis of binary data with multinomial distri -bution by using the iterative cross entropy minimization estimation. *The Korean Communications in Statistics*, Vol. 12, 125-137.

[15] Ordonez, C. (2003). Clustering binary data streams with K-means. *In 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery.*

[16] Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, Vol. 66, 846-850.

[Received July 2006, Accepted July 2006]