How to organize data with measurement errors?

Yuhua Qian
School of Computer and Information Technology
Shanxi University
Taiyuan, China
jinchengqyh@126.com

Jiye Liang
School of Computer and Information Technology
Shanxi University
Taiyuan, China
ljy@sxu.edu.cn

***Abstract*—Clustering analysis is an outstanding contribution to data mining and knowledge discovery from large-scale data, which has become an important research direction. Many excellent researches have been developed, whereas there is a blind point which is not addressed. How to organize data with measurement errors? What is needed for this purpose is the concept of error number– an error number which is defined as $a(x)\pm\Delta_{a(x)}$, where $a(x)$ is the measurement value of an object x under an attribute a, and $\Delta_{a(x)}$ is the value of its measurement error. In this paper, we will explore tentatively data clustering with measurement errors through employing the framework of k-means clustering algorithm, which can be regarded as an introductory work.**

***Index Terms*—Clustering Analysis; Measurement Error; Distance; k-means Algorithm**

I. Introduction

As an unsupervised learning process, clustering is a challenging problem, especially in cases of high-dimensional data sets [3, 4, 5, 8]. The data sets are usually represented as a data table in which each dimension is regarded as an attribute [10, 11]. Under a given attribute, in real-life applications, the value of each object is often derived from practical measure by a measure instrument. That is to say, measurement error denotes the variation between measurements of the same quantity on the same individual. It is inevitable that there is a measurement error between measurement value and real value, which is a very common phenomenon in many data analysis fields. For example, in medical information systems there may exist some criteria for which is is impossible to accurately quantify, such as weight, temperature and blood pressure. In image processing issue, we are often faced with the image data contains a lot of noise, in which the gray value of each pixel are effected by errors. In the weather forecast, due to instrument error and air's uncertainty, we also can not accurately measure all indicators. Many other measurement tasks encountered the same situation. For characterizing this case, scientists usually denote these data by a error number $a(x)\pm\Delta_{a(x)}$, where $a(x)$ is the measurement value of an object x under an attribute a and $\Delta_{a(x)}$ is the value of its measurement error. Where the errors are coming from for a measurement issue? In fact, errors are caused by two subcomponents, random error and systematic error.

Generally, the generation of errors will effect the accuracy of data analysis. We need to quantify these measurement errors as accurately as possible. One approach is to perform repeated measurements on the same objects under a given feature and calculate the mean and deviation of them. If we consider the cost of repeated measurements, the other approach often adopted is to estimate the distribution of errors, which have uniform distribution and normal distribution, and so on.

Generally speaking, for this case, clustering analysis is always based on measurement values of objects. However, many data sets include the features with errors that is associated with the nature of data. Clustering result derived measurement values of objects may not reveal the real partition status of these data sets. This presents a very important problem, that is, how to develop a clustering method for data with measurement errors. The objective of this paper is to explore tentatively data clustering with measurement errors.

II. Error-number distance

General speaking, we perform clustering analysis via measurement values of objects, but not their real values. It is inevitable that there is a measurement error between measurement value and real value. In this section, we attempt to give a distance for measuring the dissimilarity between two numeric objects with measurement errors in k-means objective function. For convenience, a numeric value with a measurement error is called an error-number in this paper.

We assume the set of objects to be clustered is stored in a database table T defined by a set of attributes, $a_1, a_2, \cdots, a_m$. Each attribute a_j describes a domain of values, denoted by $V(a_j)$, associated with a defined semantic and a data type. In this study, we only consider a general data type, numeric, and assume other types used in database systems can be mapped to this type [1]. A given numeric domain consists of real numbers [6, 7, 8].

Let $X=\{X_1, X_2, \cdots, X_n\}$ be a set of n objects, in which X_i is represented as a vector $(x_{i,1}, x_{i,2}, \cdots, x_{i,m})$, $x_{i,j}\in V_{a_j}$, $1\le j\le m$. An object X in T can be logically represented as a conjunction of attribute-value pairs $a_1 = x_{i,1}\wedge a_2 = x_{i,2}\wedge\cdots\wedge a_m = x_{i,m}$, where $x_{i,j}\in V_{a_j}$ for $1\le j\le m$. One writes $X_i = X_k$ if $x_{i,j} = x_{k,j}$ for each $j\le m$. The relation $X_i = X_k$ does not mean that X_i and X_k are the same object in the real-world database, but rather that the two objects have equal values in attributes $a_1, a_2, \cdots, a_m$.

Given two objects with measurement errors $X_i = (x_{i,j}\pm\Delta x_{i,j},\ j\le m)$ and $X_k = (x_{k,j}\pm\Delta x_{k,j},\ j\le m)$, we define two operators "$+$" and "$-$" as follows:

978-1-4577-0653-0/11/$26.00 ©2011 IEEE

(1) $X_i + X_k = (|x_{i,j} + x_{k,j}| \pm |\Delta x_{i,j} + \Delta x_{k,j}|,\ j \le m)$,
(2) $X_i - X_k = (|x_{i,j} - x_{k,j}| \pm |\Delta x_{i,j} - \Delta x_{k,j}|,\ j \le m)$.

For convenience, we also can denote an object X_i with measurement errors by combing the corresponding measurement-value vector $M(X_i)$ and measurement-error vector $\Delta(X_i)$, that is $X_i = M(X_i) \pm \Delta(X_i)$, where $M(X_i) = (x_{i,1}, x_{i,2}, \cdots, x_{i,m})$ and $\Delta(X_i) = (\Delta x_{i,1}, \Delta x_{i,2}, \cdots, \Delta x_{i,m})$.

Given two error-numbers X_i and X_k, in this paper, the dissimilarity measure $\mathcal{D}(X_i, X_k)$ is defined by the following two forms:

$$\mathcal{D}_1(X_i, X_k) = \sum_{j=1}^{m} \delta(x_{i,j} \pm \Delta x_{i,j},\ x_{k,j} \pm \Delta x_{k,j}), \quad (1)$$

where $\delta(x_{i,j} \pm \Delta x_{i,j},\ x_{k,j} \pm \Delta x_{k,j}) = |x_{i,j} - x_{k,j}| + |\Delta x_{i,j} - \Delta x_{k,j}|$;
and

$$\mathcal{D}_2(X_i, X_k) = \sqrt{\sum_{j=1}^{m} \theta(x_{i,j} \pm \Delta x_{i,j},\ x_{k,j} \pm \Delta x_{k,j})}, \quad (2)$$

where $\theta(x_{i,j} \pm \Delta x_{i,j},\ x_{k,j} \pm \Delta x_{k,j}) = (x_{i,j} - x_{k,j})^2 + (\Delta x_{i,j} - \Delta x_{k,j})^2$.

We call each of these two dissimilarity measures error-number distance.

Sometimes, we can look forward a given error number as an interval number. For example, given an error number $a(x) \pm \Delta_{a(x)}$, where $a(x)$ is the measurement value of an object x under an attribute a and $\Delta_{a(x)}$ is the value of its measurement error, we can write it as an interval number $[a(x) - \Delta_{a(x)},\ a(x) + \Delta_{a(x)}] = [a^L(x),\ a^U(x)]$. Based on this consideration, one also can employ several distances in interval data analysis for measuring the difference between two error numbers.

Souza and De Carvalho [12] introduced adaptive city-block distance to dynamic clustering algorithm for symbolic interval data, which is defined as

$$\mathcal{D}_3(X_i, X_k) = \sum_{j=1}^{m} (|x_{i,j}^L - x_{k,j}^L| + |x_{i,j}^U - x_{k,j}^U|), \quad (3)$$

where $x_{i,j}^L = x_{i,j} - \Delta_{x_{i,j}}$ and $x_{i,j}^U = x_{i,j} + \Delta_{x_{i,j}}$.

De Carvalho et al. [2] proposed dynamic clustering algorithm for symbolic interval data by considering adaptive Hausdorff distance with the following form

$$\mathcal{D}_4(X_i, X_k) = \sum_{j=1}^{m} \max\{|x_{i,j}^L - x_{k,j}^L|,\ |x_{i,j}^U - x_{k,j}^U|\}, \quad (4)$$

where $x_{i,j}^L = x_{i,j} - \Delta_{x_{i,j}}$ and $x_{i,j}^U = x_{i,j} + \Delta_{x_{i,j}}$.

According to the above consideration, we also can define the following distance

$$\mathcal{D}_5(X_i, X_k) = \sum_{j=1}^{m} \sqrt{|x_{i,j}^L - x_{k,j}^L|^2 + |x_{i,j}^U - x_{k,j}^U|^2}. \quad (5)$$

The above five distances are all used to calculate the difference between two error numbers. In the framework of k-means algorithm, each of these distances can be used to determine initial k cluster centers and judge the class labels of objects at each iteration, and the differences in clustering results will be produced by different error-number distances. In this study, we will employ the five kinds of distances for verify the reasonableness and effectiveness of the proposed point of view that data clustering considering measurement errors maybe a much better selection.

In this study, we will explore tentatively clustering for data with measurement errors in the framework of k-means algorithm.

The objective of clustering a set of n error numbers into k clusters is to find Z that minimizes

$$f(Z) = \sum_{j=1}^{k} \sum_{X \in \omega_j} \mathcal{D}^2(X, Z_j), \quad (6)$$

where $k(\le n)$ is a known number of clusters, $Z = \{Z_1, Z_2, \cdots, Z_k\}$, and Z_j is the jth cluster center with the numeric attributes $a_1, a_2, \cdots, a_m$.

Minimization of f in (6) is formalized in the following EKM algorithm:

Algorithm 1. The k-means algorithm based on error-number distance (EKM)

Input: A set of objects with measurement errors $X = \{X_1, X_2, \cdots, X_n\}$ and attributes $a_1, a_2, \cdots, a_m$;
Output: k clusters.

(1) Using classical max-min distance means, choose k points as initial cluster centers: $Z_1^0, Z_2^0, \cdots, Z_k^0$. Let $l = 0$.

(2) If $\mathcal{D}(X_i, Z_c^{(l)}) = \min_j\{\mathcal{D}(X_i, Z_j^{(l)})\}$, $i = 1, 2, \cdots, |U|$, then put X_i into the cluster $\omega_c^{(l+1)}$, which generates new clusters $\omega_j^{(l+1)}$ $(j = 1, 2, \cdots, k)$.

(3) Compute the center of each new cluster by

$$Z_j^{(l+1)} = \frac{1}{n_j^{(l+1)}} \sum_{X_i \in \omega_j^{(l+1)}} X_i,\ j = 1, 2, \cdots, k,$$

where $n_j^{(l+1)}$ is the number of objects in the cluster $\omega_j^{(l+1)}$.

(4) If $Z_j^{(l+1)} = Z_j^{(l)}$ $(j = 1, 2, \cdots, k)$, then the algorithm stop; otherwise, $l = l + 1$, goto (2).

In the above algorithm, the time complexity of determining class labels of n objects at each iteration is $O(kn)$. Let the iterative times be p when the algorithm stop, one easily knows that the time complexity of EKM algorithm is $O(pkn)$.

To evaluate the performance of clustering algorithms, one often considers three measures: 1) accuracy (AC), 2) precision (PE), and 3) recall (RE) [12]. Objects in an lth cluster are assumed to be classified either correctly or incorrectly with respect to a given class of objects. Let the number of correctly

classified objects be a_l, let the number of incorrectly classified objects be b_l, and let the number of objects in a given class but not in a cluster be c_l. The clustering accuracy, precision and recall are defined as follows:

$$AC = \frac{\sum_{l=1}^{k} a_l}{n},\ PR = \frac{\sum_{l=1}^{k}(\frac{a_l}{a_l+b_l})}{k},\ RE = \frac{\sum_{l=1}^{k}(\frac{a_l}{a_l+c_l})}{k},$$

respectively.

III. Experimental analysis

In order to evaluate the performance of EKM clustering algorithm, two synthetic data sets with measurement errors and six public data sets are employed. Our aim is to explain a viewpoint that the clustering result considering measurement errors may be much closer to real classification status of data sets than that only derived from data with measurement values.

A. Experimental methodology

In the experimental analysis, we reconstruct the value of each object under a given attribute through generating an acceptable measurement error.

Let $X = \{X_1, X_2, \cdots, X_n\}$ be a set of n objects, $\{a_1, a_2, \cdots, a_m\}$ be a set of m attributes, in which X_i is represented as a vector $(x_{i,1}, x_{i,2}, \cdots, x_{i,m})$, $x_{i,j} \in V_{a_j}$, $1 \leq j \leq m$. For $\forall a_j$, one can obtain an error-number by the following approach:

$$x_{i,j} \rightarrow x_{i,j} \pm \Delta x_{i,j},$$

$\Delta x_{i,j}$ is a random number from the interval $[0, s(j)]$, where

$$s(j) = (\max\{V_{a_j}\} - \min\{V_{a_j}\})/\varepsilon.$$

ε is a given error range, called error ratio.

To facilitate further discussion, we set the error ratio into a reasonable interval $[10, 100]$ in the following experiments. If the error ratio is too bigger (> 100), an error number will be much closer to the corresponding measurement value and the clustering results obtained using EKM algorithm may be also much closer to those using k-means algorithm. If the error ratio is too smaller (< 10), it implies that the error range can not be tolerable in practical applications. To verify the performance of EKM algorithm, in the experimental study, the parameter ε are assigned as 10, 20, 30, 40, 50, 60. 70, 80, 90 and 100, respectively.

In the experimental analysis, we think of error modeled as two probability density functions, uniform distribution and normal distribution. For uniform distribution, through generating a random number y in the interval $[0,\ 1]$, one can obtain such a random number $y \times s(j)$ that falls in the interval $[0, s(j)]$. Hence, we can construct a set of error numbers in which errors satisfy the uniform distribution in the interval $[0, s(j)]$. For normal distribution, we first generate a random number y with the standard normal distribution in which the mean is zero and the variance equals one, then enlarge this number y to the interval $[-s(j), s(j)]$. Through the above preprocessing, we can transform a given data set into the corresponding data set with measurement errors with different distributions.

In the proposed EKM algorithm, the abstract distance can be calculated by each of formulas (1)-(5), denoted by $1th$-distance, $2th$-distance, $3th$-distance, $4th$-distance and $5th$-distance, respectively. The different distances will produce the different clustering results. It is gratifying that the clustering results induced by each of these distances are statistically much better than those by k-means without measurement errors.

B. Synthetic data sets

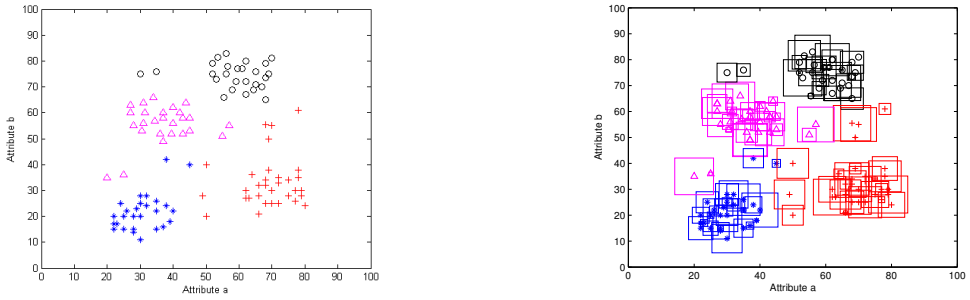
In this paper, we consider one numeric data sets with measurement errors. This data set is characterized by attribute a and attribute b.

The synthetic data set has 115 points scattered among four classes, which is shown in Fig. 1. In Fig. 1, a numeric data set is displayed on left side, in which the value of each object is a measurement value. Whereas the figure on right side displays a data set with measurement errors in which each point is an error number satisfying the uniform distribution with $\varepsilon = 10$. From (a) and (b) in Fig. 1, it can be seen that the clusters induced by a clustering algorithm should be four classes. The configuration of this data set aims to illustrate clustering result obtained from data with measurement errors may be much better than that only derived from data with measurement values.

Through using classical k-means algorithm and the proposed algorithm EKM with 1th distance, we organize the data set in Fig. 1. Two different clustering results are acquired, which are shown in (a) and (b) of Fig. 2. It is clear that the clustering result by EKM algorithm with measurement errors is much closer to real classes than that by k-means algorithm.

In order to evaluate the advantage of EKM algorithm in the context of statistics, we employ 10-fold cross validation method with various error ratios from 10 to 100 for this purpose. Table 1 shows the summary results for both algorithms, in which the errors are estimated by the uniform distribution and the normal distribution and five error-number distances are calculated respectively.

According to Table 1, the performance of EKM with measurement errors is consistently better than the original k-means algorithm without measurement errors for AC, PR, and RE. It is deserved to point out that no matter what kind of error-number distance, the value of each of AC, PR and RE almost decrease as the value of the error ratio varies from 10 to 100. It is because that an error number will be much closer to the corresponding measurement value and the clustering results obtained using EKM algorithm may be also much closer to those using k-means algorithm. In order to more clearly display, Fig. 3 displays variation of AC, PR and RE with various error ratios on 1th synthetic data set reconstructed by $\varepsilon = 10$ and the uniform distribution. In the figure, the x-axis is the error ratio of error data used to clustering analysis and y-axis is the statistical values of AC, PE and RE. Note that the last column in each sub-figure is the corresponding statistical value of each of AC, PE and RE in the clustering results induced by k-means algorithm, which provides a base line to evaluate the performance of clustering results. It can



(a) Real classes with measurement values

(b) Real classes with measurement errors ($\varepsilon = 10$)

Fig. 1: Real classes on 1th synthetic data set

be easily seen that each of these three indices are much better than that index obtained by k-means algorithm. In a word, it is easy to see that the clusters from 1st synthetic data set by EKM is statistically much closer to its real partition than those by k-means without measurement errors.

IV. Conclusion

Clustering analysis is an outstanding contribution to data mining and knowledge discovery from large-scale data. Many excellent researches have been developed, whereas there is a blind point which is not addressed, that is how to organize data with measurement errors? As an introductory work, this paper explores tentatively data clustering with measurement errors through employing the framework of k-means clustering algorithm. This work has four central contributions: 1) It reveals several existing problems in data clustering without measurement errors, 2) it offers several dissimilarity measures, called error-number distance, which can be used to calculate the difference between two error numbers, and 3) through using the framework of k-means algorithm, it develops a clustering algorithm (EKM) for deriving clustering results from a data set with measurement errors. The overall experimental results demonstrating on two synthetic data sets and five real-life data sets show that the proposed EKM algorithm can improve the performance of k-means algorithm and obtaining much better clustering results. The results show that data clustering with measurement errors is a very noteworthy research issue.

Acknowledgment

This work was supported by the national natural science foundation of China (No. 60903110, 60773133, 70971080), the national high technology research and development program of China (No. 2007AA01Z165), the natural science foundation of Shanxi province, China (No. 2008011038, 2009021017-1).

References

[1] M. R. Chmielewski and J. W. Grzymala Busse, "Global discretization of continuous attributes as preprocessing for machine learning," *Int. J. Approx. Reasoning*, vol. 15, no. 4, pp. 319-331, 1996.

[2] F. A. T. De Carvalho, R. M. C. R. Souza, M. Chavent, and Y. Lechevallier, "Adaptive Hausdorff distances and dynamic clustering of symbolic data," *Pattern Recog. Letters*, vol. 27, no. 3, pp. 167-179, 2006.

[3] I. Guyon and A. Elisseeff, "An introduction to variable feature selection," *J. Mach. Learn. Res*, vol. 3, pp. 1157-1182, 2003.

[4] Q. H. Hu, Z. X. Xie, and D. R. Yu, "Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation," *Pattern Recognition*, vol. 40, pp. 3509-3521, 2007.

[5] Q. H. Hu, D. R. Yu, Z. X. Xie, and J. F. Liu, "Fuzzy probabilistic approximation spaces and their information measures," *IEEE Trans. Fuzzy Syst.*, vol. 14, no. 2, pp. 191-201, 2006.

[6] Z. X. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining Knowledge Discovery*, vol. 2, no. 3, pp. 283-304, 1998.

[7] Z. X. Huang, "A Fuzzy k-modes algorithm for clustering categorical data," *IEEE Transactions on Fuzzy Systems*, vol. 7, no. 4, pp. 446-452, 1999.

[8] H. Liu and R. Setiono, "Feature selection via discretization," *IEEE Trans. Knowl. Data Eng.*, vol. 9, no. 4, pp. 642-645, 1997.

[9] K. N. Michael, M. J. Li, Z. X. Huang and Z. Y. He, "On the impact of dissimilarity measure in k-modes clustering algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 503-507, 2007.

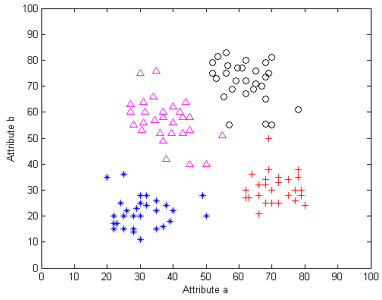
[10] Y. H. Qian, J. Y. Liang, and C. Y. Dang, "Consistency measure, inclusion degree and fuzzy measure in decision tables," *Fuzzy Sets and Systems*, vol. 159, pp. 2353-2377, 2008.

[11] Y. H. Qian, J. Y. Liang, Witold, and C.Y. Dang, "Positive approximation: an accelerator for attribute reduction in rough set theory," *Artif. Intell.*, vol. 174, pp. 597-618, 2010.

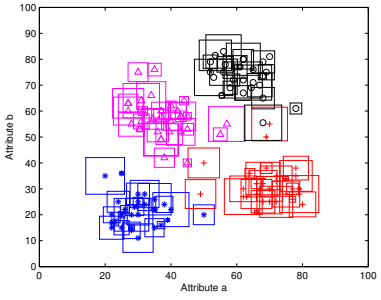
[12] R. M. C. R. Souza, F. A. T. De Carvalho, "Clustering of interval data based on city-block distances," *Pattern Recog. Letters*, vol. 25, no. 3, pp. 353-365, 2004.

TABLE I: Variation of AC, PR and RE with various error ratios on 1th synthetic data set

Distance	EKM Error ratio	Uniform distribution AC	PR	RE	Normal distribution AC	PR	RE
1th distance	10	0.9223 ± 0.0000	0.9233 ± 0.0022	0.9217 ± 0.0004	0.9078 ± 0.0078	0.9096 ± 0.0069	0.9086 ± 0.0069
	20	0.9194 ± 0.0044	0.9203 ± 0.0055	0.9191 ± 0.0042	0.9078 ± 0.0049	0.9095 ± 0.0036	0.9086 ± 0.0044
	30	0.9184 ± 0.0048	0.9193 ± 0.0059	0.9182 ± 0.0044	0.9078 ± 0.0049	0.9095 ± 0.0036	0.9086 ± 0.0044
	40	0.9146 ± 0.0073	0.9160 ± 0.0067	0.9147 ± 0.0066	0.9078 ± 0.0049	0.9095 ± 0.0036	0.9086 ± 0.0044
	50	0.9126 ± 0.0061	0.9140 ± 0.0057	0.9130 ± 0.0056	0.9078 ± 0.0049	0.9095 ± 0.0036	0.9086 ± 0.0044
	60	0.9126 ± 0.0061	0.9140 ± 0.0057	0.9130 ± 0.0056	0.9078 ± 0.0049	0.9095 ± 0.0036	0.9086 ± 0.0044
	70	0.9126 ± 0.0061	0.9140 ± 0.0057	0.9130 ± 0.0056	0.9078 ± 0.0049	0.9095 ± 0.0036	0.9086 ± 0.0044
	80	0.9126 ± 0.0061	0.9140 ± 0.0057	0.9130 ± 0.0056	0.9078 ± 0.0049	0.9095 ± 0.0036	0.9086 ± 0.0044
	90	0.9126 ± 0.0061	0.9140 ± 0.0057	0.9130 ± 0.0056	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056
	100	0.9117 ± 0.0052	0.9130 ± 0.0048	0.9121 ± 0.0048	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056
2th distance	10	0.8777 ± 0.0089	0.8815 ± 0.0082	0.8806 ± 0.0080	0.8757 ± 0.0058	0.8798 ± 0.0052	0.8788 ± 0.0055
	20	0.8757 ± 0.0058	0.8798 ± 0.0052	0.8788 ± 0.0055	0.8757 ± 0.0058	0.8798 ± 0.0052	0.8788 ± 0.0055
	30	0.8757 ± 0.0058	0.8798 ± 0.0052	0.8788 ± 0.0055	0.8757 ± 0.0058	0.8798 ± 0.0052	0.8788 ± 0.0055
	40	0.8757 ± 0.0058	0.8798 ± 0.0052	0.8788 ± 0.0055	0.8757 ± 0.0058	0.8798 ± 0.0052	0.8788 ± 0.0055
	50	0.8757 ± 0.0058	0.8798 ± 0.0052	0.8788 ± 0.0055	0.8748 ± 0.0029	0.8790 ± 0.0029	0.8780 ± 0.0029
	60	0.8748 ± 0.0029	0.8790 ± 0.0029	0.8780 ± 0.0029	0.8748 ± 0.0029	0.8790 ± 0.0029	0.8780 ± 0.0029
	70	0.8748 ± 0.0029	0.8790 ± 0.0029	0.8780 ± 0.0029	0.8748 ± 0.0029	0.8790 ± 0.0029	0.8780 ± 0.0029
	80	0.8748 ± 0.0029	0.8790 ± 0.0029	0.8780 ± 0.0029	0.8748 ± 0.0029	0.8790 ± 0.0029	0.8780 ± 0.0029
	90	0.8748 ± 0.0029	0.8790 ± 0.0029	0.8780 ± 0.0029	0.8748 ± 0.0029	0.8790 ± 0.0029	0.8780 ± 0.0029
	100	0.8748 ± 0.0029	0.8790 ± 0.0029	0.8780 ± 0.0029	0.8748 ± 0.0029	0.8790 ± 0.0029	0.8780 ± 0.0029
3th distance	10	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056	0.9087 ± 0.0064	0.9101 ± 0.0054	0.9095 ± 0.0057
	20	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056
	30	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056
	40	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056
	50	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056
	60	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056
	70	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056
	80	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056
	90	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056
	100	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056
4th distance	10	0.9204 ± 0.0039	0.9214 ± 0.0055	0.9200 ± 0.0035	0.9097 ± 0.0076	0.9115 ± 0.0068	0.9103 ± 0.0068
	20	0.9165 ± 0.0048	0.9178 ± 0.0044	0.9164 ± 0.0042	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056
	30	0.9155 ± 0.0062	0.9166 ± 0.0065	0.9156 ± 0.0058	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056
	40	0.9146 ± 0.0058	0.9158 ± 0.0062	0.9148 ± 0.0052	0.9068 ± 0.0064	0.9086 ± 0.0054	0.9078 ± 0.0056
	50	0.9126 ± 0.0061	0.9140 ± 0.0057	0.9130 ± 0.0056	0.9078 ± 0.0065	0.9094 ± 0.0057	0.9087 ± 0.0056
	60	0.9136 ± 0.0052	0.9147 ± 0.0049	0.9138 ± 0.0049	0.9078 ± 0.0065	0.9094 ± 0.0057	0.9087 ± 0.0056
	70	0.9126 ± 0.0043	0.9137 ± 0.0040	0.9129 ± 0.0042	0.9078 ± 0.0065	0.9096 ± 0.0059	0.9087 ± 0.0056
	80	0.9126 ± 0.0043	0.9137 ± 0.0040	0.9129 ± 0.0042	0.9078 ± 0.0065	0.9096 ± 0.0059	0.9087 ± 0.0056
	90	0.9117 ± 0.0052	0.9130 ± 0.0048	0.9121 ± 0.0048	0.9078 ± 0.0065	0.9096 ± 0.0059	0.9087 ± 0.0056
	100	0.9107 ± 0.0073	0.9121 ± 0.0068	0.9113 ± 0.0064	0.9078 ± 0.0065	0.9096 ± 0.0059	0.9087 ± 0.0056
5th distance	10	0.9379 ± 0.0048	0.9374 ± 0.0061	0.9386 ± 0.0058	0.9272 ± 0.0078	0.9282 ± 0.0073	0.9290 ± 0.0071
	20	0.9379 ± 0.0048	0.9374 ± 0.0061	0.9386 ± 0.0058	0.9282 ± 0.0089	0.9290 ± 0.0077	0.9298 ± 0.0078
	30	0.9359 ± 0.0064	0.9360 ± 0.0064	0.9369 ± 0.0064	0.9291 ± 0.0098	0.9299 ± 0.0087	0.9309 ± 0.0089
	40	0.9379 ± 0.0048	0.9376 ± 0.0062	0.9386 ± 0.0058	0.9291 ± 0.0098	0.9299 ± 0.0087	0.9309 ± 0.0089
	50	0.9379 ± 0.0048	0.9376 ± 0.0062	0.9386 ± 0.0058	0.9301 ± 0.0073	0.9305 ± 0.0062	0.9315 ± 0.0061
	60	0.9340 ± 0.0058	0.9339 ± 0.0057	0.9347 ± 0.0056	0.9282 ± 0.0064	0.9290 ± 0.0057	0.9297 ± 0.0052
	70	0.9320 ± 0.0075	0.9323 ± 0.0069	0.9333 ± 0.0068	0.9282 ± 0.0064	0.9290 ± 0.0057	0.9297 ± 0.0052
	80	0.9301 ± 0.0073	0.9307 ± 0.0066	0.9316 ± 0.0065	0.9282 ± 0.0064	0.9290 ± 0.0057	0.9297 ± 0.0052
	90	0.9291 ± 0.0076	0.9299 ± 0.0070	0.9309 ± 0.0068	0.9282 ± 0.0064	0.9290 ± 0.0057	0.9297 ± 0.0052
	100	0.9223 ± 0.0174	0.9235 ± 0.0157	0.9241 ± 0.0167	0.9282 ± 0.0064	0.9290 ± 0.0057	0.9297 ± 0.0052
	c-means						
		AC	PR	RE	AC	PR	RE
		0.8748 ± 0.0029	0.8790 ± 0.0029	0.8780 ± 0.0029	0.8748 ± 0.0029	0.8790 ± 0.0029	0.8780 ± 0.0029

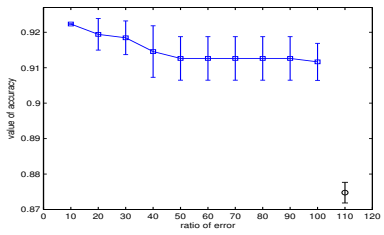


(a) Clusters with k-means algorithm

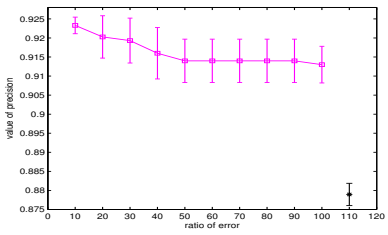


(b) Clusters with EKM algorithm ($\varepsilon = 10$)

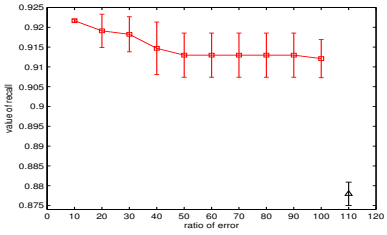
Fig. 2: Clusters with k-means algorithm and those with EKM algorithm on 1th synthetic data set



(a) Real classes



(b) Clusters with k-means algorithm



(c) Classes with measurement errors ($\varepsilon = 10$)

Fig. 3: Variation of AC, PR and RE with various error ratios on 1th synthetic data set