http://ansinet.com/itj

ITJ

ISSN 1812-5638

# INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 10 (12): 2420-2426, 2011
ISSN 1812-5638 / DOI: 10.3923/itj.2011.2420.2426


# Information-theoretic Agglomerative K-means

Yanfeng Zhang, Xutao Li, Yunming Ye, Xiaofei Xu and Shengchun Deng
Department of Computer Science, Harbin Institute of Technology, China

**Abstract:** Agglomerative K-means is a clustering algorithm of K-means type. The algorithm has good properties because of its insensitiveness to the locations of initial centers and its effectiveness in determining the number of clusters. In present study, we extend the agglomerative K-means from information theoretic view and develop a new clustering algorithm, Information-Theoretic Agglomerative K-means. Different from the agglomerative K-means, we propose a new objective function employing the Kullback-Leibler divergence to measure the dispersion of clusters. Based on this objective function, we derive the updating formulas of centers and membership for objects associated to different centers and then develop an efficient algorithm. Experimental results on both well-separated and overlapped data suggested that the proposed clustering algorithm is not only promising in obtaining good clustering performance but also effective in identifying the number of clusters.

**Key words:** Agglomerative K-means, information theory, kullback-leibler divergence, number of clusters

## INTRODUCTION

Clustering is an important unsupervised task in data mining field. Based on the differences of methodologies, clustering algorithms can mainly be classified into four categories, including partitional clustering, hierarchical clustering, density-based clustering and spectral clustering. K-means is one of the most widely studied and used clustering algorithms.

K-means preserves two important and well-known problems. One is that the clustering results of K-means are sensitive to the locations of initial centers and the other is that the number of cluster is hard to determine in real applications. In the literature, both problems have been extensively studied and many different methods have been proposed. Pelleg and Moore (2000) proposed to determine the number of clusters based on Bayesian Information Criterion. Hamerly and Elkan (2003) studied the number of clusters using hypothesis test. Feng and Hamerly (2007) further developed this method to determine the number of clusters for high dimensional data. Likas *et al*. (2003) proposed to determine the number of clusters by using Minimum Description Length. Likas *et al*. (2003) proposed a global k-means clustering algorithm which attacks the local optimization problem by obtaining k clusters incrementally. Krishna and Murty (1999) proposed to use genetic algorithms to optimize the clustering results. Lu *et al*. (2004) further developed this method and developed a fast genetic algorithm. Arthur and Vassilvitskii (2007) studied and proposed a strategy to select better locations for initial centers to improve clustering results of K-means. Recently, Li *et al*. (2008) proposed an agglomerative K-means algorithm which is not only able to alleviate the local optimizing problem but also able to determine the number of clusters. Based on this algorithm, Li *et al*. (2010) proposed a hierarchical clustering scheme for clustering nested and multi-density data sets.

In present study, we extend the agglomerative K-means from the information theoretic view and develop a new clustering algorithm, Information-Theoretic Agglomerative K-means. Different from the agglomerative K-means, we propose a new objective function employing the Kullback-Leibler divergence to measure the dispersions of clusters. Based on this new objective function, we derive the updating formulas of centers and membership for objects associated to different centers and then develop an efficient algorithm. Experimental results on both well-separated and overlapped suggested that the proposed clustering algorithm is not only promising in obtaining good clustering performances but also effective in identifying the number of clusters.

## INFORMATION-THEORETIC AGGLOMERATIVE K-MEANS

Assume there are n objects and there are m attributes for each object. Let $D = \{D_1, D_2,..., D_n\}$ represent the set of n objects and let the i-th object be represented as $D_i = \{d_{i,1}, d_{i,2},..., d_{i,m}\}$, where $d_{i,1}$ is the value of j-th attribute of this object. Since we will consider the

**Corresponding Author:** Yunming Ye, Department of Computer Science, Harbin Institute of Technology, China

clustering problem form information theory which is unable to consider negative data, we preprocess the data by transforming them into non-negative ones as follows:

$$q_{i,l} = \frac{d_{i,l} - \min(l)}{\max(l) - \min(l)}$$

where, min ($l$) and max ($l$) represent the smallest value and the largest value of $l$-th dimension in the data set D, respectively.

In order to partition the n objects into k clusters, we modify the objective function of agglomerative K-means as follows:

$$\min_{U,P} Q(U,P) = \sum_{i=1}^{n}\sum_{j=1}^{k} u_{i,j}\left(\sum_{l=1}^{m} q_{i,l}\log\frac{q_{i,l}}{p_{j,l}} - q_{i,l} + p_{j,l}\right) + \lambda\sum_{i=1}^{n}\sum_{j=1}^{k} u_{i,j}\log u_{i,j} \quad (1)$$

subject to:

$$\sum_{j=1}^{k} u_{i,j} = 1,\ \ u_{i,j} \in (0,1],\ \ i = 1,2,\ldots,n$$

where, $u_{i,j}$ indicates the membership that i-th object is associated to j-th cluster and $p_{j,l}$ indicates the value of the l-th dimension in the j-th cluster. Remarkably, the basic difference between our objective function and the objective function of agglomerative K-means (Li *et al*., 2008) is that we employ the Kullback-Leibler divergence to measure the cluster dispersion. Since, the Kullback-Leibler divergence is asymmetric, we can also have another objective function as follows:

$$\min_{U,P} Q(U,P) = \sum_{i=1}^{n}\sum_{j=1}^{k} u_{i,j}\left(\sum_{l=1}^{m} p_{j,l}\log\frac{p_{j,l}}{q_{i,l}} - p_{j,l} + q_{i,l}\right) + \lambda\sum_{i=1}^{n}\sum_{j=1}^{k} u_{i,j}\log u_{i,j} \quad (2)$$

subject to:

$$\sum_{j=1}^{k} u_{i,j} = 1,\ \ u_{i,j} \in (0,1],\ \ i = 1,2,\ldots,n$$

In this study, we consider both the objective functions (Eq. 1 and 2). Based on them, we derive different updating formulas for centers and membership of objects associated to them. We refer to the Eq. 1 as method (1) and the Eq. 2 as method (2). Interestingly, we can rewrite the Eq. 1 (or the Eq. 2) from the theoretic view:

$$\min_{u,p} u\left(H(q,p) - H(q)\right) + \lambda H(u)$$

where, H (q, p) represents the cross entropy of random variables q and p and H (q) represents the entropy of random variable q. Because of this fact, we name our algorithm as Information Theoretic Agglomerative K-means.

Next we derive the updating formulas for centers and membership of objects associated to them. Following the deriving procedure by Li *et al*. (2008), one still has similar updating formula for the membership matrix U as follows:

$$u_{i,j} = \frac{\exp\left(-\frac{KL_{i,j}}{\lambda}\right)}{\sum_{t=1}^{k}\exp\left(-\frac{KL_{i,t}}{\lambda}\right)} \quad (3)$$

where, for method (1):

$$KL_{i,j} = \sum_{l=1}^{m} q_{i,l}\log\frac{q_{i,l}}{p_{j,l}} - q_{i,l} + p_{j,l}$$

and for method (2):

$$KL_{i,j} = \sum_{l=1}^{m} p_{j,l}\log\frac{p_{j,l}}{q_{i,l}} - p_{j,l} + q_{i,l}.$$

To obtain the updating formula for membership of objects associated to centers, we derive for methods (1) and method (2), respectively. Taking partial derivative on the right hand of Eq. 1 with respect to $p_{j,l}$ and setting it zero, we have:

$$\frac{\partial Q}{\partial p_{j,l}} = -\sum_{i=1}^{n} u_{i,j}\frac{q_{i,l}}{p_{j,l}} + \sum_{i=1}^{n} u_{i,j} = 0$$

Solving the equation, we have $p_{j,l}$ for method (1):

$$p_{j,l} = \frac{\sum_{i=1}^{n} u_{i,j}\, q_{i,l}}{\sum_{i=1}^{n} u_{i,j}} \quad (4)$$

Taking partial derivative on the right hand of Eq. 2 with respect to $p_{j,l}$ and setting it zero, we have:

$$\frac{\partial Q}{\partial p_{j,l}} = \sum_{i=1}^{n} u_{i,j}\left(1 + \log p_{j,l} - \log_{i,l} - 1\right) = 0$$

Solving the equation, we have $p_{j,l}$ for method (2):

$$p_{j,l} = \exp\left(\frac{\sum_{i=1}^{n} u_{i,j}\log q_{i,l}}{\sum_{i=1}^{n} u_{i,j}}\right) \quad (5)$$

Given a fixed value for λ, a clustering result can be obtained by updating centers and membership of objects in terms of formulas derived above until the process converges. However, the interesting property of this algorithm is that the centers tend to agglomerate as the value of lambda increases. Because of this property, the algorithm can obtain a serial of clustering results with different number of clusters by increasing the value of lambda gradually and then it determines the final result by selecting the most stable one from the serial of results.

**Algorithm:** In this subsection, we summarize the algorithm of information-theoretic agglomerative K-means (both method (1) and method (2)) as Algorithm 1.

The algorithm resembles the agglomerative K-means and it needs a input $K_{max}$ for specifying the largest number of clusters. Note that $K_{max}$ must be larger than the genuine number of clusters so that the genuine number of clusters can be identified. In the algorithm, the number of clusters and final clustering results are determined by increasing λ to have the second term (the negative entropy term) in our objective function make more contribution. This term can control the number of clusters to be formed. As λ increases, the centers tend to agglomerate and we have clustering results with different number of clusters. After that, we select the most stable result with respect to the change of λ.

Algorithm 1: Information-Theoretic Agglomerative K-means method (1) (or method (2))

**Input:** $K_{max}$ represents the largest number of clusters
**Output:** The clustering result with suitable number of clusters
**Procedure:**
1: Randomly initialize $K_{max}$ centers;
2: Set $\lambda = 0.001$, $t = \lambda$;
3: Update centers P as Eq. 4 (or Eq. 5 for method (2)) and update membership matrix U as Eq. 3 until it converges;
4: Check the distances between current centers, merge the centers that are close enough and re-update centers P and membership matrix U;
5: If the number of clusters decreases, record the clustering result.
6: If current number of cluster is not equal to one, set $\lambda = \lambda + t$ and goto Step 3; otherwise goto Step 7;
7: Output the clustering result with the largest λ changing interval during which the number of clusters does not decrease.

## EXPERIMENTS

Here, we present three experimental results to show the effectiveness of the proposed algorithm. In the first experiment, we show the result of the proposed algorithm on synthetic well-separated data set. In the second experiment, we show the result of the proposed algorithm on synthetic overlapped data set. In the third experiment, we show the result of the proposed algorithm on UCI data set. In all the three experiments, we set $K_{max}$ to be 10.

**Evaluation metric:** To evaluate the clustering results, we use the rand index which measures the qualities of resulting clusters with the differences between them and true partitions. The rand index is computed as follows:

$$\text{rand index} = \frac{a+d}{a+b+c+d}$$

where, a is the number of object pairs that are in the same clusters in resulting partitions and in the same clusters in true partitions, b is the number of object pairs that are in different clusters in resulting partitions and in the same clusters in true partitions, c is the number of object pairs that are in the same clusters in resulting partitions and in different clusters in true partitions, d is the number of object pairs that are in different clusters in resulting partitions and in different clusters in true partitions. Note that the higher the rand index is, the better the clustering result is.

**Experiment 1:** Here, we generate a 2D well-separated data set with 3000 objects and 3 inherent clusters, shown as in Fig. 1. We apply the proposed method (1) and method (2) to this data set. For comparison, we apply the agglomerative K-means to this data set as well.

In Fig. 2 (a-c), we show how the number of clusters changes against the values of λ. We can see that the lambda is the most stable when the number of clusters is three. The facts demonstrate the proposed method (1) and method (2) can determine the number of clusters as the agglomerative K-means does. The resulting clustering accuracies are as follows: method (i) 1.0000, method (ii) 0.9987 and the agglomerative K-means 1.0000. According to the curves in Fig. 2 (a-c) and the resulting clustering accuracies, we can see that the proposed Information-Theoretic Agglomerative K-means is comparable to the

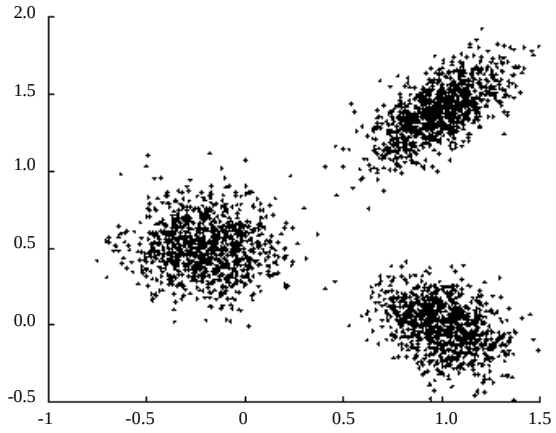


Fig. 1: The well-separated data set

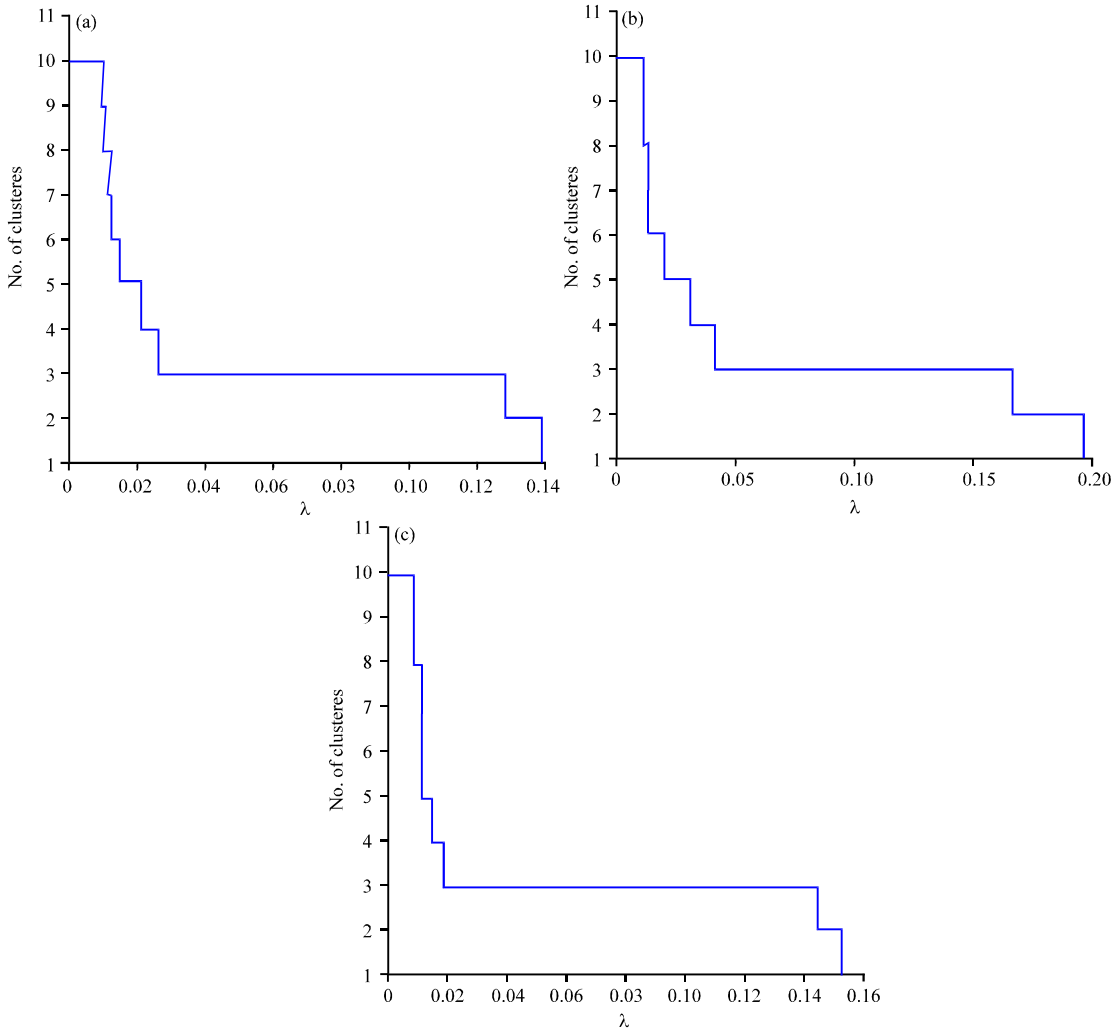


Fig. 2 (a-c): The number of clusters changes against the values of λ in experiment 1. (a) method (1), (b) method (2) and (c) the agglomerative K-means

agglomerative K-means no matter on the identification of genuine number of clusters or on resulting well-qualified clusters for well-separated data sets.

**Experiment 2:** Here, we generate a 2D overlapped data set with 5000 objects and 5 inherent clusters, shown as in Fig. 3. We apply the proposed method (1) and method (2) to this data set. For comparison, we apply the agglomerative K-means to this data set as well.

In Fig. 4 (a-c), we show how the number of clusters changes against the values of λ. We can see that the lambda is the most stable when the number of clusters is five. The facts demonstrate the proposed method (1) and method (2) can determine the number of clusters as the agglomerative K-means does. Besides, we can see that lambda is also very stable when the number of clusters is four because the cluster in the center are seriously overlapped with the other four. The resulting clustering accuracies are as follows: method (1) 0.9792, method (2) 0.9756 and the agglomerative K-means 0.9783. The accuracy of method (1) is slightly better than those of method (2) and agglomerative K-means. Both results demonstrate that the proposed Information-Theoretic Agglomerative K-means is comparable to the agglomerative K-means no matter on the identification of genuine number of clusters or on resulting well-qualified clusters for overlapped data sets.

**Experiment 3:** Here, we apply the proposed Information-Theoretic Agglomerative K-means to the UCI wine data set which is the result of a chemical analysis of wines grown in the same region in Italy but derived from three

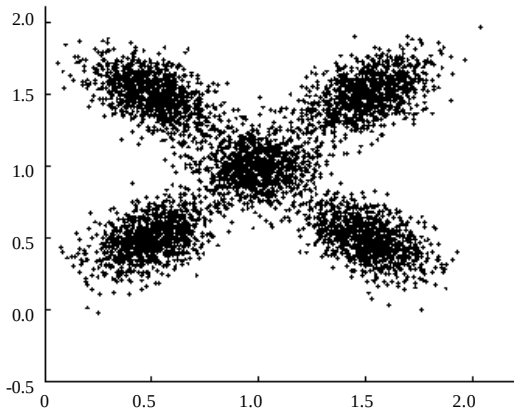


Fig. 3: The overlapped data set

different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. Therefore, this data is composed of 13 attributes with 3 different classes. For comparison, we apply the agglomerative k-means to this data set as well.

In Fig. 5 (a-c), we show how the number of clusters changes against the values of λ. The proposed method (1) and agglomerative K-means are able to determine the number of clusters accurately. However, we see that the method (2) cannot identify the number of clusters and λ, is the most stable when the number of clusters is two. The resulting clustering accuracies are as follows: the proposed method (1) achieves the clustering accuracy 0.9331; the proposed method (2) achieves the clustering accuracy 0.9040 (when the number of clusters is three);

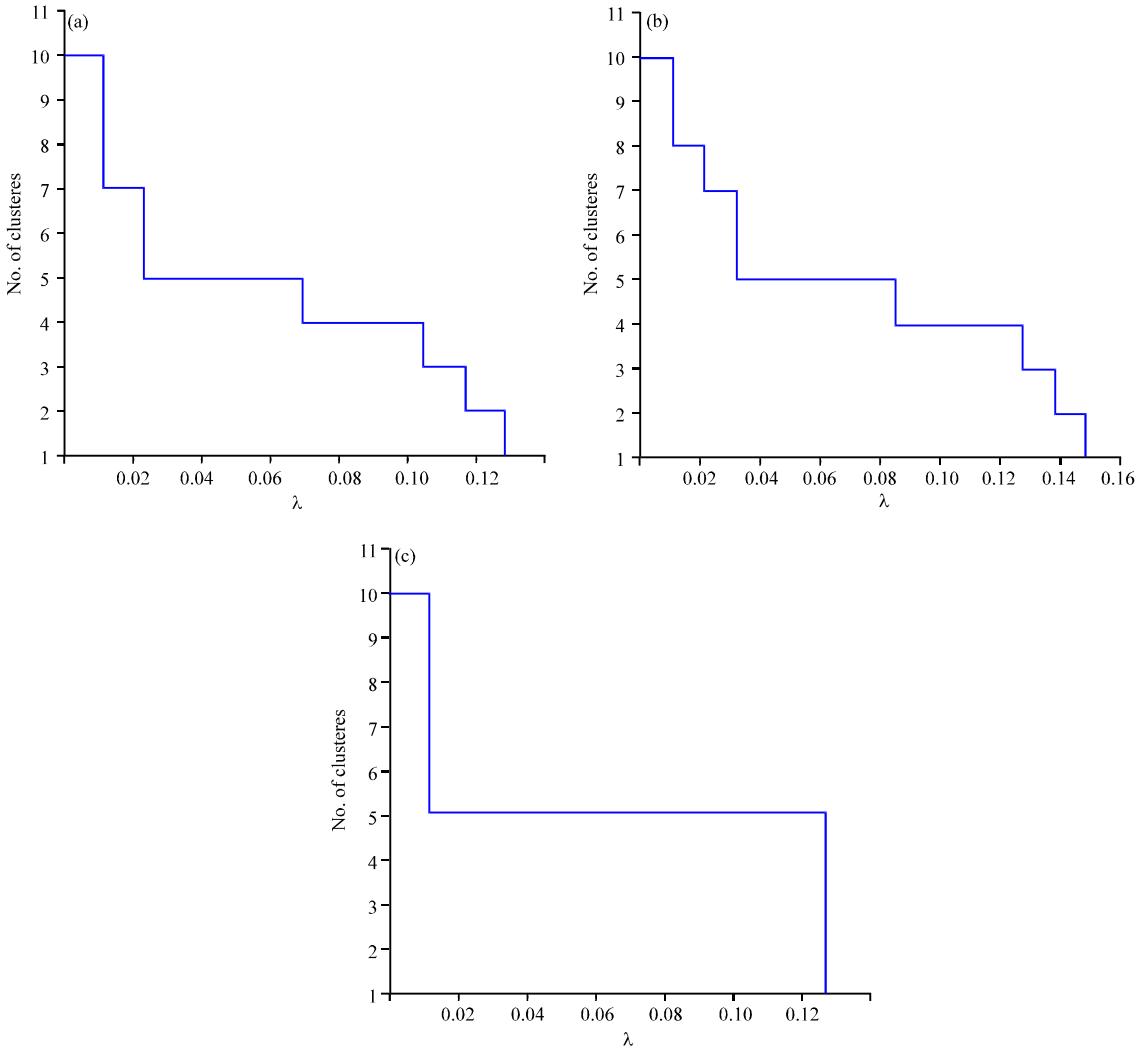


Fig. 4 (a-c): The number of clusters changes against the values of λ in experiment 2. (a) method (1), (b) method (2) and (c) the agglomerative K-means

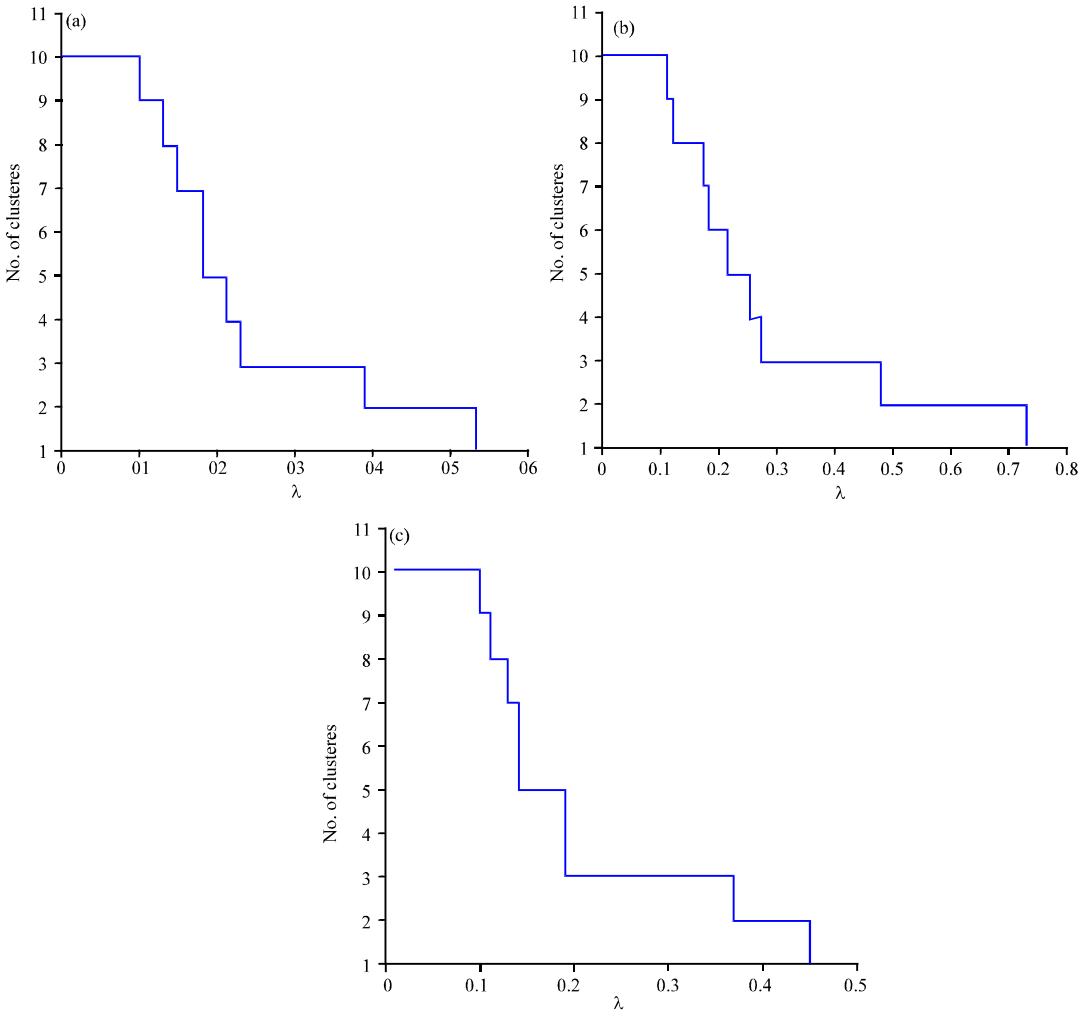


Fig. 5 (a-c): The number of clusters changes against the values of λ in experiment 3. (a) method (1), (b) method (2) and (c) the agglomerative K-means

the agglomerative K-means achieves the clustering accuracy 0.9251. We see that the clustering result of method (1) is better than those of method (2) and agglomerative K-means.

In summary, all the experimental results have shown that the proposed algorithm is comparable to the agglomerative K-means, no matter on determining the number of clusters or on yielding well-qualified clustering results. Moreover, for the proposed two methods, we find that method (1) seems better than method (2) according to the experimental results.

## CONCLUSIONS

In present study, we propose the Information-Theoretic Agglomerative K-means which is an extension of agglomerative K-means from the information-theoretic view. The algorithm is not only able to determine the number of clusters but also able to obtain well-qualified clustering accuracies. Experimental results on synthetic data sets and UCI data set demonstrate that the proposed Information-Theoretic Agglomerative K-means is comparable to the agglomerative K-means.

## ACKNOWLEDGMENTS

Y. Ye's research is supported in part by NSFC under Grant No. 61073195 and Shenzhen Science and Technology Program under Grant No. CXB201005250024A. S. Deng's research is supported in part by NSFC under Grant No.61073051.

## REFERENCES

Arthur, D. and S. Vassilvitskii, 2007. k-means++: The advantages of careful seeding. Proceedings of the 18th Annual ACM-SIAM Symposium of Discrete Analysis, Jan. 7-9, New Orleans, Louisiana, pp: 1027-1035.

Feng, Y. and G. Hamerly, 2007. PG-means: Learning the number of clusters in data. Adv. Neural Inform. Proc. Syst., 19: 393-400.

Hamerly, G. and C. Elkan, 2003. Learning the k in k-means. http://u.math.biu.ac.il/~louzouy/courses/seminar/kmeans1.pdf

Krishna, K. and M. Murty, 1999. Genetic K-means algorithm. IEEE Trans. Syst. Man Cybernet. Part B. Cybernet., 29: 433-439.

Li, M.J., M.K. Ng, Y.M. Cheung and J.Z. Huang, 2008. Agglomerative fuzzy K-means clustering algorithm with selection of number of clusters. Knowl. Data Eng., 20: 1519-1534.

Li, X., Y. Ye, M.J. Li and M.K. Ng, 2010. On cluster tree for nested and multi-density data clustering. Pattern Recognit., 43: 3130-3143.

Likas, A., N. Vlassis and J.J. Verbeek, 2003. The global k-means clustering algorithm. Pattern Recognit., 36: 451-461.

Lu, Y., S. Lu, F. Fotouhi, Y. Deng and S.J. Brown, 2004. FGKA: A fast genetic k-means clustering algorithm. Proceedings of the 19th ACM Symposium on Applied Computing, Nicosia, Cyprus, March 14-17, ACM, pp: 622-623.

Pelleg, D. and A. Moore, 2000. X-means: Extending K-means with efficient estimation of the number of clusters. Proceedings of the 17th International Conference on Machine Learning, (ICML'00), Morgan Kaufmann, pp: 727-734.