Spatio-Temporal Data Evolutionary Clustering Based on MOEA/D

Jingjing Ma, Yanhui Wang
Xidian University
Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xi'an, 710071, China
yhwang86@yeah.net

Maoguo Gong, Licheng Jiao
Xidian University
Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xi'an, 710071, China
{gong, lchjiao}@ieee.org

Qingfu Zhang
University of Essex
School of Computer Science & Electronic Engineering, Colchester, CO4 3SQ, UK
qzhang@essex.ac.uk

ABSTRACT
Clustering the data evolve with time, which is termed evolutionary clustering, is an emerging and important research area in recent literature of data mining, and it is very effective to cluster the dynamic data. It needs to consider two conflicting criteria. One is the snapshot quality function; the other is the history cost function. Most state-of-the-art methods combine these two objectives into one and apply a single objective optimization method for optimizing it. In this paper, we propose a new evolutionary clustering approach by using a multi-objective evolutionary algorithm based on decomposition (MOEA/D) to optimize these two conflicting functions in evolutionary k-means algorithm (EKM). The experimental results demonstrate that our algorithm significantly outperforms EKM.

Categories and Subject Descriptors
I.5.3 [**Pattern Recognition**]: Clustering — *Algorithms*; I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search— *Heuristic methods*

General Terms
Algorithms.

Keywords
Evolutionary Clustering, Multiobjective Optimization, Evolutionary Algorithm, Spatio-temporal Data, Data Mining.

1. INTRODUCTION
In many clustering applications of data mining, some changes maybe occur in some features of the data to be clustered, which contain both long-term trend due to concept drift and short-term variation due to noise [2]. One wants to maintain long-term trend, but smooth short-term variation. This issue has raised a challenge to traditional clustering algorithms. Evolutionary clustering is developed for addressing this issue. The term, which is investigated in the KDD/Data-mining community in 2006 [1], does not mean clustering by using evolutionary algorithms but means clustering the data evolve with time, also known as spatio-temporal data.

Chakrabarti et al. [1] proposed the first evolutionary clustering framework which considers two conflicting functions. One is the snapshot quality function (*sq*), which reflects the quality of the clustering at the current timestamp; the other is the history cost function (*hc*), which ensures that the clustering evolves smoothly over time. Some methods (e.g., [2, 7, 8, 6]) based on this framework have been proposed for temporal evolutionary data clustering.

Most state-of-the-art methods combine two objectives into one and use single objective optimization methods for optimizing it. This paper treats it as a bi-objective problem and directly applies multi-objective techniques for finding trade-off solutions.

In this paper, we propose EKM-MOEA/D, a novel evolutionary k-means algorithm for handling the two conflicting criteria. Our method uses the multi-objective evolutionary algorithm based on decomposition (MOEA/D) [4]. Two criterion functions in evolutionary k-means (EKM) are optimized at the same time. Our approach is able to provide multiple options for the users. In addition, the combination of evolutionary optimization with EKM improves the global search ability of the algorithm.

2. DESCRIPTION OF EKM-MOEA/D
In this section, we describe our proposed evolutionary clustering algorithm, termed as Evolutionary K-means Algorithm based on MOEA/D (EKM-MOEA/D). EKM-MOEA/D uses MOEA/D to optimize the two criterion functions: the snapshot quality (*sq*) and the history cost (*hc*). The lower value of *sq* means that the clustering quality should be good. The lower value of *hc* means that the clustering should not deviate much from the previous clustering.

2.1 Objective Function in EKM-MOEA/D
The *sq* in EKM [1] is defined as $sq = \sum_{x \in U} \min_{c \in C} \|c - x\|$, where $C = \{c_{1,\ldots,}c_k\}$ is the set of the current clustering centers, $x \in U$ is the data point vector. The *hc* in EKM [1] is defined as $hc = \sum_{f:[k]\to[k]} \left\| c_i - c^{t-1}_{f(i)} \right\|$, where $C = \{c_{1,\ldots,}c_k\}$ is the set of the current clustering centers, $C^{t-1} = \{c_{1,\ldots,}c_k\}$ is the set of the history clustering centers, f is a function that maps centers of C to centers of C^{t-1}, which means finding a closest center in C^{t-1} to each center in C in the best possible way.

In this paper, we use MOEA/D [4] to minimize the *sq* and the *hc* simultaneously to get the approximation of the PF. We will optimize a bi-objectives problem, namely, $f_1 = sq$, $f_2 = hc$.

GECCO'11, July 12–16, 2011, Dublin, Ireland.
ACM 978-1-4503-0690-4/11/07.

3. EXPERIMENTS

In order to analyze the performance of EKM-MOEA/D, the comparison of EKM-MOEA/D with EKM [1] is made on UCI datasets. In our experiments, we use EKM-MOEA/D to generate 101 approximate optimal solutions. In EKM, we let α increases from 0 to 1 with a step of 0.01 to get a PF with the same number of solutions. We apply EKM 100 times, each time with different α values, to generate the same number of approximate solutions. In addition, we use the C-metric [3] to illustrate the comparison. In experiments, we performed 30 independent runs on each test dataset. Box plots [5] are used to illustrate the distribution of these statistical values of the coverage of the two sets.

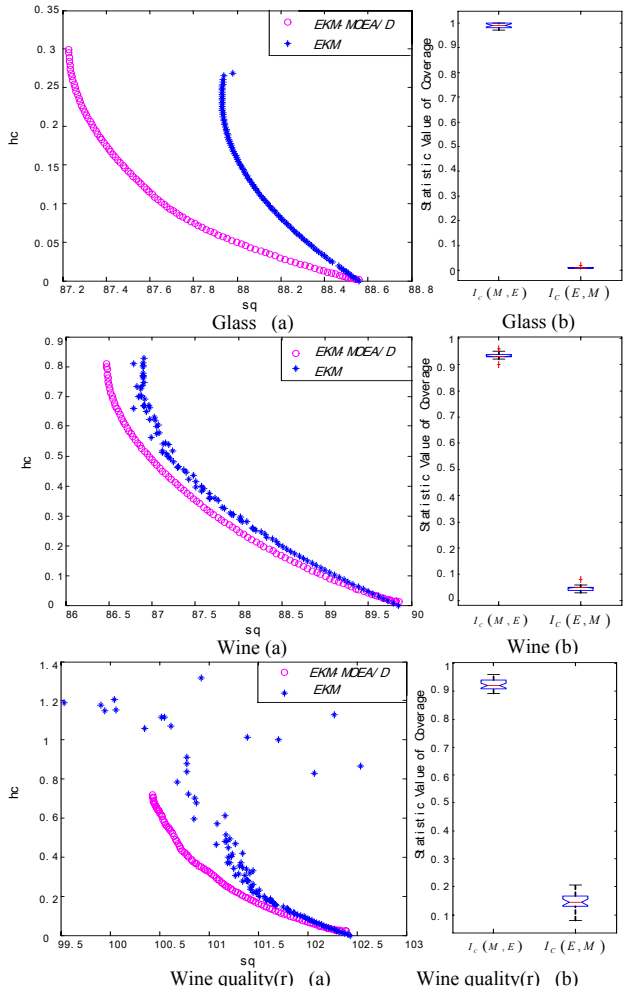

Figure 1: (a)The PF obtained by EKM-MOEA/D and EKM in testing the three UCI datasets, illustrated by a set of 101 uniform points on them; (b) Box plots of the Coverage of the two sets. M denotes the solution set obtained by EKM-MOEA/D, E denotes the solution set obtained by EKM.

As preprocessing, we need divide each dataset into several parts to represent different timestamps data. In our study, we only consider the previous two timestamps data. We cluster the first timestamp data using the k-means for ten times and choose the best cluster centers, which make SQ minimum, as the history cluster centers in clustering the second timestamp data. The specific information of test datasets is shown in Table 1. Parameter setting in MOEA/D can be seen in [4].

Table 1. Description of test datasets

Datasets	Dim	Cluster	Num	Timestamps
glass	10	2	214	2
Wine	13	3	178	2
Wine(qr)	11	4	1599	10

It can be seen intuitively from Figure 1 that EKM-MOEA/D has a better performance than EKM. The better performance can be expressed as follows: Firstly, EKM-MOEA/D conquers the unstable solution caused by improper choice of random initial centers in EKM and can find a better global optimization than EKM. More importantly, EKM-MOEA/D provides a set of PF for the users to choose from instead of the solutions in different weight parameters in EKM, both the SQ and the HC will not be worse at least.

4. CONCLUDING REMARKS

In this paper, a novel evolutionary clustering algorithm EKM-MOEA/D is proposed. We have showed that EKM-MOEA/D has a better improvement in SQ and HC and more stable solutions than EMK under different weight parameters. EKM-MOEA/D might help us to obtain a set of Pareto-optimal solutions for the users to select, but not need to choose the weight parameters in advance. We only apply our approach to test two timestamps data in UCI datasets in this paper. Next, we plan to apply our approach to test more temporal variable data and extend our algorithms to real applications.

5. REFERENCES

[1] D. Chakrabarti, R. Kumar, and A. Tomkins. 2006. Evolutionary Clustering. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, Philadelphia, PA, USA, pp. 554-560.

[2] Y. Chi, X. Song, D. Zhou, K. Hino, and B. L. Tseng. 2007. Evolutionary spectral clustering by incorporating temporal smoothness. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, San Jose, California, USA, pp. 153-162.

[3] J. D. Knowles and D. W. Corne. 1999. The Pareto archived evolution strategy: A new baseline algorithm for multiobjective optimization. In Proc. Congr. Evol. Comput.,Washington, D.C., Jul. pp. 98–105.

[4] H. Li and Q. Zhang. 2009. Multiobjective Optimization Problems with Complicated Pareto Sets, MOEA/D and NSGA-II, IEEE Trans on Evolutionary Computation, vol. 12, no 2, pp 284-302.

[5] McGill, R., Tukey, J., and Larsen, W. 1978. Variations of boxplots. The American Statistician,32: 12–16.

[6] R. Shankar, G. Kiran, V. Pudi. 2010. Evolutionary Clustering using Frequent Itemsets. ACM StreamKDD'10, July 25.

[7] T. Xu, Z. M. Zhang, P. S. Yu, and B. Long. 2008. Dirichlet Process Based Evolutionary Clustering. In Proceedings of IEEE International Conference on Data Mining, pp. 648-657, December 2008.

[8] T. Xu, Z. M. Zhang, P. S. Yu, and B. Long. 2008. Evolutionary Clustering by Hierarchical Dirichlet Process with Hidden Markov State," In Proceedings of the Eighth IEEE International Conference on Data Mining, pp. 658-667.