Report for seminar: Topics in Data Mining

[on study of the paper: Oliveira, MÃąrcia, and JoÃčo Gama. "A framework to monitor clusters evolution applied to economy and finance problems."]

Zijian Zhang
Subject area: Informatik
Matriculation number: 3184680
zhangzijian0523@gmail.com

ABSTRACT

In this report a short summary about studying of paper *A framework to monitor clusters evolution applied to economy and finance problems.*(MCEEF) is presented. At first are a couple of real world's problems evolved being mentioned by the MCEEF paper introduced. Then we go into the framework, which is described by the paper, to see how in the world is the framework been defined, what functionalities does it have or which types of data phenomenons could it describe. Then the result and performance of appliance of this framework are to be illustrated. At last shortages that still exist within the framework are discussed and according to the paper we could also mention some probable improvement on the framework.

1. RESEARCHES REVIEW

Cluster structure and clustering problems of a bundle of data are fascinating topics in studying properties of data analysis. Being discussed in [3] and so on. static data cluster was deeply researched. However, considering the variation of data alone with the time goes by, it dosen't make seance that clusters already gotten always keep as-is. A cluster could birth, die, slit and attributes of (clusters of) data set could illustrate interesting properties. That's why the temporal attributes, a.k.a. transitions, were discussed at first in different aspects, such as generic patterns[13][7][8], clusters[21][12][23] or association rules[5][6].

At the aspect of static data, frameworks such as FOCUS[13] and MH-DIFF[8] provided a formal way to describe the structural component and measure component by researching the clustering of data set. Those frameworks also inspired the research on the differences or deviations between two cluster sets, or those between clustering of two snapshots of data set. With the help of tree structures built in t_i and t_{i+1}, distance of two models are described using a set of editing operations.

Further researches like PAM(Automated Pattern Monitor)[6] expand the method of detection of derivation by combining GRM(Generic Rule Model) and monitoring using threshold as two different phases. That figured out a clearer way to go deep into the monitoring of clusters' evolution.

Based on the result of studying of social networks, an event-based framework[12][4] for characterizing the evolution of interaction networks was produced. The social networks were considered as time fluctuating graphs. The temporal behavior was categorized in to several critical events, such as *formation*, *dissolution*, *split* and *merge* of clusters (communities), or *join*, *leave*, *appearance* or *disappearance* of individuals(entities).

Confronted data streams, data analyzer usually focus on unsupervised learning of the patterns attend in data. Naturally clustering should to be main direction of research proceeded[2][1][9][18][11][15][17]. Also because of the efficiency and scalability of algorithms. A corresponded generic tool to understand, visualize and diagnose differences between evolving data streams were also developed. Within these approaches, the concept of velocity density estimation are suggested been split into both temporal and spatial velocity profiles. While temporal velocity profiles provide the variation of the density rate over a given section of time; the spatial velocity profile describes the movement of data in a feature space within a fixed time point. Furthermore the events of data points are expanded with *coagulation*, *dissolution* and *shift*. Also this framework could be extended to high-dimensional data streams.

According to the writer of this paper, there were still two research works are closely related to theirs, thus MONIC framework[21] and it's improved version MONIC+ framework[20]. While MONIC uses a data aging function to detect the evolution of clustering, MONIC+ is less generic and introduces proper heuristic to capture transitions between different clusterings, which is independent of the clusters discovery process. Although MONIC and it's improvement were published by a same author, they seems to use differentmetric to detect evolution and do not provide a visualize approach. While this paper the author provide an approach which improves his former work[19] and uses the same way, i.e. weighted bipartite graphs, to monitor the variation of clustering. Thus according to the author, those four papers are mutually substantive.

2. SHORT DESCRIPTION OF MEC FRAMEWORK

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2016 ACM. ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

To proceed the research on evolution of clusters. The essay at first provides two critical point about it: definition of clustering and the concept of temporal evolution. Our main interested part is definition and recognition of specific patterns of behaviour of clusters within time interval $[t_i, t_{i+\Delta t}](\Delta t = 1, ..., T - i)$.

2.1 Definition of clusters and clustering

The clusters, or a clustering of a set of data, is denoted as υ is a specific partitioning of a data set D into K partitions $\upsilon = C_1, ..., C_i, ...C_k$ and have three main properties:

1.clusters should be *disjoint*(or mutually exclusive)

$$C_i \cap C_j = \emptyset, \forall i \neq j$$

2.clusters should be *collectively exhausted*

$$\cup_{i=1}^{K} C_i = D$$

3.Data points, or according to the essay, observations, 'seem' to be closer to it's partners within a same data cluster than those in other clusters.

The MEC framework defines its clusterings in a enumerative way, i.e. defining a cluster by its members. Thus alone the variation of time, if there are two clusters that contain totally same data points, they are treated as a same cluster. This methodology is also known as extensional definition of clusters. Let $\vec{x_i}, (i = 1, ..., N)$ be a vector defined within N-dimensional real space $\vec{x_i} = (x_{i,1}, x_{i,2}, ...x_{i,n})$. The snapshot at time point t of a cluster C_j can be defined as such:

$$C_j(t) = \vec{x_1}, ..., \vec{x_m}$$

where m is the number of data points with in the cluster $C_j(t), j = (1, ...k)$, k is the number of clusters and $t = (1, ..., T)$, where T corresponds to the last analyzed time stamp

2.2 Taxonomy of transitions

According to the essay there are in total eight types of transitions ever mentioned[21][12][23][2][9][2][17][16][4]. However, there only five of them are considered here, they are:

1. Birth - generation of a new cluster.
2. Death - disappearance of an existed cluster.
3. Split - separation of one cluster into two or more clusters
4. Merge - fusion of two or more clusters as one cluster
5. Survival - a cluster that does not suffer any of the above transitions

Defined externally, these five basic type of variation stand for the related transition of the whole *Clustering* between two adjacent time points. The terminology of those transaction are at first introduced by Spiliopoulou[21] as in Table 1.

Those terminology could be intuitively illustrated by the Figure 1.

Table 1: Terminology for clusters' transitions

Mathematical Notation	Description
$\emptyset \rightarrow C_u(t_{i+\Delta t})$	Cluster's Birth
$C_m(t_i) \rightarrow \emptyset$	Cluster's Death
$C_m(t_i) \rightarrow C_1(t_{i+\Delta t}), ...C_r(t_{i+\Delta t}) \wedge$ $C_m(t_i) = \cup_{k=1}^{r} C_k(t_{i+\Delta t})$	Split of a cluster into r clusters
$C_1(t_i), ..., C_p(t_i) \rightarrow C_u(t_{i+\Delta t}) \wedge$ $\cup_{k=1}^{p} C_k(t_i) = C_u t_{i+\Delta t}$	Merge of p clusters into one cluster
$C_m(t_i) \rightarrow C_u(t_{i+\Delta t})$	Cluster's Survival

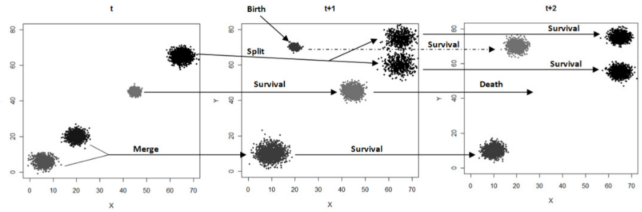

Figure 1: Demonstration of five types of evolution

2.3 Tracking and Mapping

The central point of determinate a transaction is finding the mapping relationship between all of the points of two clusters over a specific time interval $[t_{i-\Delta t}, t_i]$. To do so, its critical to discover overlapping region with in the feature space or, in other words, the matches of clusters. In this essay the mapping relationship between two clusters are defined with a conditional probability. The clustering of two time stamps are considered as a bipartite graph. Each cluster within either time point are treated as a vertex of corresponding party of bipartite graph. The Edges of the graph connect every pair of possible connections between clusters obtained at consecutive time points. And the weight of edges are assigned by the conditional probabilities between two of them. Under this deterministic context it could describe the similarity between them thus translate the transitions into a graphical way to visualize them. If the conditional probability of two clusters outstrips a threshold, it could be convinced that those two clusters are basically the same clusters, or the later is a cluster after several translations of former one. The structure of bipartite graph is showed in Figure 2.

And the formal definition of weighted bipartite graph could is such:

Given the clusterings $\upsilon_i, \upsilon_{i+\Delta t}$, a graph G=(U, V, E) can

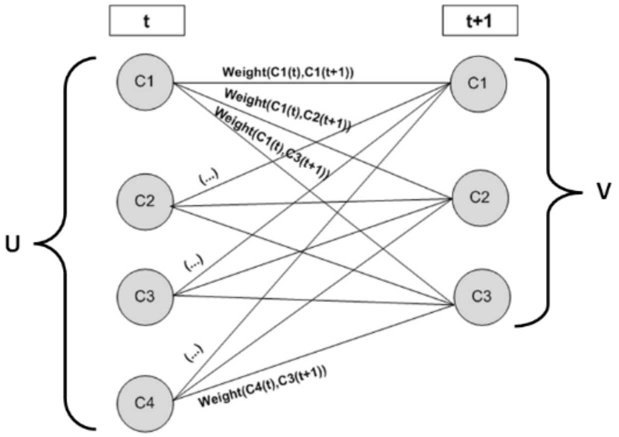

Figure 2: Demonstration of five types of evolution

be constructed, where U represents the first subset of vertices (clusters of t_i), V represents the second subset of vertices (clusters of $t_{i+\Delta t}$), and E denotes a set of weighted edges between any pair of clusters belonging to υ_i and $\upsilon_{i+\Delta t}$. Formally, the weight assigned to the edge connecting clusters $C_m(t_i)$ and $C_u(t_{i+\Delta t})(m = (1, ...k_{t_i}))$and $u = (1, ..., k_{t_{t_{\Delta} t}})$, where k_{t_i} and $k_{t_{i+\Delta t}}$ are the numbers of clusters returned by a given clustering algorithm in time points t_i and $t_{i+\Delta t}$, respectively) are estimated in accordance with the conditional probability:

$$\begin{aligned} &weight(C_m(t_i), C_u(t_{i+\Delta t})) \\ &= P(X \in C_u(t_{i+\Delta t})|X \in C_m(t_i)) \\ &= \frac{\sum P(x \in C_m(t_i) \cap C_u(t_{i+\Delta t}))}{\sum P(x \in C_m(t_i))} \end{aligned}$$

where X is the set of data sets contained by cluster $C_m(t_i)$ and $P(X \in C_u(t_{t+\Delta t})|X \in C_m(t_i))$ means that under the condition of given X belongs to cluster $C_m(t_i)$ the probability that X is contained by C_u at $t_{i+\Delta t}$.

The formula above only works under situations of crisp clustering. If we want to make it adequate to fussy clusters, some adjustment on the definition of metrics have to be done. For instance, it's mentioned in the essay that for fuzzy clustering, the *union* of two clusters A and B could be defined as:

$$A \cup B = max[A(x), B(x)]$$

and analogous, the *intersection* of them could be modified as:

$$A \cap B = min[A(x), B(x)]$$

In order to find the intersection of two clusters on different time points, each observations within the data set should be structurally identical. That is to say, we have to identify each observation on every time point and track it from it's occurrence to disappearance. Intuitively it may sound like a drawback of performance of the framework, but according to the essay this tracking of identical observations can be of major interest.

To detect different transactions between adjacent time points, the paper also introduces a threshold called **split threshold** λ and a **survival threshold** τ, which is based on the design of MONIAC's framework's external transitions[21]. The mathematical description of transitions are in Table 2

Now are some explanations to the huge table above:

A cluster is **born**, when the weight of edges which connect this cluster to each clusters of previous time point is under the survive threshold τ, or, in other word, it's not possible to find a predecessor from which this cluster transforms not 'so much'.

A cluster **dies**, when the weight of all edges link from it self to a successor are under the survival threshold.

A cluster **splits** into a set of successive clusters, when at least two clusters $C_u(t_{i+_{\Delta}t})$ and $C_v(t_{t+\Delta t})$ exist, whose edges' weights are equal or exceed the split threshold λ given that they are split from $C_m(t_i)$ (corresponding the conditional probability). And at the same time their sum is equal, or exceed the survival threshold.

A set of clusters **merge**, when there are at least two different clusters form t_i, whose connections to a given cluster of $t_{i+\Delta t}$ are equal, or exceed the survival threshold.

A cluster **survives**, when there is one and only one cluster at time point $t_{\Delta t}$, the link between whom and it has weight equal, or exceed the survival threshold.

3. THRESHOLDS' SENSITIVITY ANALYSIS

After defining the toxicology of transactions, we still have to face to a problem of determination of two threshold values. Thus the writer begins with a threshold's sensitivity analysis. The experiments are conducted under that all values of survival threshold within the range [0.5, 1], which is to say, if a cluster survives, there must be at least a half of observations within time point t_i that maintain till the time point $t_{\Delta t}$. And for the values of the split threshold, they all fall into [0, 0.4]. The analysis is performed using *ceteris paribus*.

3.1 Survival Threshold

Before the conduction of experiment, the writer gives a logical hypothesis that one would expect that the higher the value of the survival threshold τ, the smaller the number of survival and merges detected. It wold also be expected that the increase of τ leads to a greater number of births, death and splits. Also according to the result of experiment with some value of survival threshold the transaction *merge* could be replaced by *survive*, since the weight assigned to one of the edges became lower than the threshold. And the number of *death*s increases for higher values of the threshold.

3.2 Split Threshold

Logical hypothesises also can be made for split threshold λ. It's expected that the higher its value, the greater the number of deaths and *birth*s, and lower the number of *split*s. It could also be inferred that number of *merge*s and *survival*s are not affected by the variation of λ, based on the formal definition of these transitions.

3.3 Discussion

The thresholds' sensitivity analysis concludes that it is survival threshold that lead to completely different results, while the split threshold dosen't has a significant impact on the behavior of framework.

4. REAL WORLD OBSERVATION TESTS

After analysing the parameters, the writer at last decides to use $\lambda = 0.2$ and $\tau = 0.6$ describing the method of identification of transitions, this paper also conducts several experiment with the data from real world, especially from finance territory. Data extracted from Banco de Portugal's Central Balance-Sheet Database and *The data Page* of New York University - Leonard N. Stern School of Business in order to capture and understand the general trends of cluster evolution. The result does work well for the large amount of data set and perform a intuitive accuracy.

5. PROS AND CONTRAS OF THE PRESENT WORK

The advantagement of present work is that in the paper a framework of monitoring evolutions of clustering structures overtime be introduced. Also the various types of transactions are categorized and the frame work is tested with real-world case using economic and financial data setd.

Table 2: Formal definition of the external transitions of a cluster represented by enumeration

Transitions' Taxonomy	Notation	Formal Definition
Cluster's Birth	$\emptyset \rightarrow C_u(t_{i+\Delta t})$	$0 < weight(C_m(t_i), C_u(t_{i+\Delta t})) < \tau\forall_m$
Cluster's Death	$C_m(t_i) \rightarrow \emptyset$	$weight(C_m(t_i), C_u(t_{i+\Delta t})) < \lambda\forall_u$
Cluster's Split	$C_m(t_i) \rightarrow C_1(t_{i+\Delta t}), ...C_r(t_{i+\Delta t})\wedge$ $C_m(t_i) \subset C_1(t_{i+\Delta t}), ...C_r(t_{i+\Delta t})$	$(\exists_u\exists_v : weight(C_m(t_i), C_u(t_{i+\Delta t})) \geq \lambda\wedge$ $weight(C_m(t_i), C_v(t_{i+\Delta t})) \geq \lambda)\wedge$ $\sum_{u=1}^{r} weight(C_m(t_i), C_u(t_{i+\Delta t})) \geq \tau$
Cluster's Merge	$C_1(t_i), ..., C_p(t_i) \rightarrow C_u(t_{i+\Delta t})\wedge$ $C_1(t_i), ..., C_p(t_i) \subset C_u(t_{i+\Delta t})$	$(weight(C_m(t_i), C_u(t_{i+\Delta t}) \geq \tau)\wedge$ $\exists C_p \in \upsilon_i \setminus C_m : weight(C_p(t_i), C_u(t_{i+\Delta t})) \geq \tau$
Cluster's Survival	$C_m(t_i) \rightarrow C_u(t_{i+\Delta t})$	$(weight(C_m(t_i), C_u(t_{i+\Delta t})) \geq \tau)\wedge$ $\nexists C_p \in \upsilon_i \setminus C_m : weight(C_p(t_i), C_u(t_{i+\delta t})) \geq \tau$

However, it still contain some of the shortages. One of them is as mentioned in the paper, this framework dosen't anapt to the fuzzy clustering situation, moreover the complexity of identification of a cluster or a clustering is too high. From which another disadvantage of this work could be seen, that the dimentions and quantity of experimental data are rather low. The other shortages could be identified if we extend our vision to other papers, which are to be mentioned

6. RELATED WORKS AND EXTENSIONS

According to the work of Michael Hahsler et.al.[14], the orders that which transition happens implicates some properties. The largest contribution of works of Michael Hahsler et.al. are:

1. It builds a Markov model transitions along the time flow, which discribes the relationship of adjacent cluster transitions of the data stream

2. In the paper it was mentioned that the identificational data of clusterings should be saved syntheticly, i.e. only the statistical data of a clustering should be saved and thus the data complexity of clustering representation are substantially decreased.

3. In order to prevent the current result from being significantly effected by old data, A lazy fading strategy was introduced. Where a expotential factor $2^{-\lambda}$ was used to multiplicate with nearst predecessor. Since this factor is accumulative it could be transformed to $2^{-\lambda\Delta t}$ and multiplicated only when the historical data is needed.

However, the work of Michael Hasler et.al. also has its disadvantage. For example it doesn't consider the situation that a cluster *dies*, instead, using fading factor to represent one cluster is not effective anymore. Another example is that they only considered there are only one kind of transition occurs between situations. But the data set in real world more than one transitions come out between two observation time point. So there are no exact correspondance between every time point and situation.

Another splendid work from F.E. Correa et.al.[10] trade observations with more than one properties over time as a 3-dimensional tensors and applyied Tucker decomposition[22] on them. Since Tucker decomposition is an already existant algorithm, it is easy to find the already existant optimized best practice of it. Also the conduction of Tucker decomposition significantly compresses the data and thus decrease the process complexity. Since the observation with properties over times are treated as a three dimensional tensor, an analysis over three dimensions, especially that over time dimention the transition of observations could be identified.

The main disadvantagement of F.E. Correa et.al. is their framwork doesn't presente an intuitive result. Being illustrated on several 2-D plane over the dimension to analyze, the result usually requires a further implication. And it provided only a general vision on the overall scope of data sets, instead of a micro-scope on behaviour of every clusters. Another frustration is in the region of compute complexity. In the work of F.E. Correa et.al. a tradeoff between computational complexity and explained variation are mentioned during choosing the parameters for Tucker decomposition. Solving this tradeoff requires some priori knowledge, which may fade the applicability of this algorithm.

7. REFERENCES

[1] C. C. Aggarwal. A framework for diagnosing changes in evolving data streams. *ACM SIGMOD '03*, pages 575–586, 2003.

[2] C. C. Aggarwal. On change diagnosis in evolving data streams. *IEEE Transaction on Knowledge and Data Engineering*, 17:587–600, 2005.

[3] A.K.Jain, M.N.Murty, and P.J.Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3), September 1999.

[4] S. Asur, S. Parthasarathy, and D. Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. *ACM Transactions on Knowledge Discovery from Data*, 3:1–36, 2009.

[5] S. Baron and M. Spiliopoulou. *Monitoring change in mining results.* Springer Berlin/Heidelberg, 2001.

[6] S. Baron and M. Spiliopoulou. *Monitoring the evolution of web usage patterns.* Springer Berlin/Heidelberg, 2004.

[7] I. Bartolini, P. Ciaccia, I. Ntoutsi, M. Patella, and Y. Theodoridis. The panda framework for comparing patterns. *Data and Knowledge Engineering*, 68(2):244–260, 2009.

[8] S. S. Chawathe and H. Garcia-Molina. Meaningful change detection in structured data. *SIGMOD Record*, 26:26–37, 1997.

[9] K. Chen and L. Liu. Detecting the change of clustering structure in categorical data streams. *SIAM International Conference on Data Mining*, 2006.

[10] F. Correa, M. Oliveira, J. Gama, P. Corrêa, and

J. Rady. Analyzing the behavior dynamics of grain price indexes using tucker tensor decomposition and spatio-temporal trajectories. *ELSEVIER Computers and Electronics in Agriculture*, 120:72–78, 2016.

[11] S. Elnekave, M. Last, and O. Maimon. *Proceeding of the 23th International Conference on Data Engineering*, 2007.

[12] T. Falkowski, J. Bartelheimer, and M. Spiliopoulou. Mining and visualizing the evolution of subgroups in social networks. *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 52–58, 2006.

[13] V. Ganti, J. Gehrke, and R. Ramakrishnan. A framework for measuring changes in data characteristics. *PODS'99*, pages 126–137, 1999.

[14] M. Hahsler and M. H. Dunham. Temporal structure learning for clustering massive data streams in real-time. *SLAM Conference on Data Mining 2011*, 2011.

[15] P. Kalnis, N. Mamoulis, and S. Bakiras. *On discovering moving clusters in spatio-temporal data.*, volume 3633. Springer Berlin/Heidelberg, 2005.

[16] S. Kaur, V. Bhatnagar, S. Mehta, and S. Kapoor. Concept drift in unlabeled data stream. *Technical report*, 2009.

[17] T. Li, S. Ma, and M. Ogihara. Entropy-based criterion in categorical clustering. *ICML'04*, pages 68–, 2004.

[18] L. O'Callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani. Streaming-data algorithms for high-quality clustering. *ICDE'02*, pages 685–694, 2002.

[19] M. Oliveira and J. Gama. *Bipartite graphs for monitoring clusters transitions.*, volume 6065. Springer Berlin/Heidelberg, 2010.

[20] M. Spiliopoulou, I. Ntoutsi, and Y. Theodoridis. Tracing cluster transitions for different cluster types. *Control and Cybernetics*, 38:239–259, 2009.

[21] M. Spiliopoulou, I. Ntoutsi, Y. Theodoridis, and R. Schult. Monic: Modeling and monitoring cluster transitions. *12th ACM SIGKDD'06*, pages 706–711, 2006.

[22] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31, September 1966.

[23] H. Yang, S. Parthasarathy, and S. Mehta. A generalized framework for mining spatio-temporal patterns in scientific data. *11th ACM SIGKDD'05*, pages 716–721, 2005.