INFLUENCE OF PRIORS OVER MULTI-TYPED OBJECT IN EVOLUTIONARY CLUSTERING

L. Visalatchi[1] and Dr. T. Meyyappan[2]

[1]Associate Professor, Dept. of Information Technology,
Dr. Umayal Ramanathan College for women, Karaikudi.
{visaramki@gmail.com, visaramki@yahoo.com}
[2]Professor, Dept. of Computer Science and Engineering,
Alagappa University, Karaikudi.
{meyslotus@yahoo.com}

ABSTRACT

In recent years, Evolutionary clustering is an evolving research area in data mining. The evolution diagnosis of any homogeneous as well as heterogeneous network will provide an overall view about the network. Applications of evolutionary clustering includes, analyzing, the social networks, information networks, about their structure, properties and behaviors. In this paper, the authors study the problem of influence of priors over multi-typed object in evolutionary clustering. Priors are defined for each type of object in a heterogeneous information network and experimental results were produced to show how consistency and quality of cluster changes according to the priors.

KEYWORDS

Evolutionary Clustering, priors, consistent clusters.

1. INTRODUCTION

In the past few years, attraction towards dynamic networks such as information networks and social networks were enormous due to their ubiquity. Earlier, traditional clustering methods [3]were used to summarize the structure of these networks. But it was found that the traditional clustering method does not consider the changes that occur quite frequently in the dynamic networks and it provides only static information.

To cope up with the continuously evolving properties of dynamic networks, Deepayan chakrabarti[1] defined the evolutionary clustering. He defines the evolutionary clustering as the problem of processing timestamped data to produce sequence of clustering, that is, a clustering for each timestep of the system. The emphasis of timestep is stated, with the sense that the clustering sequences should be similar to the one at the previous stamp and also should accurately reflect the data at the current timestamp. Thus, consistency of the cluster is much more important over time. In dynamic networks evolutionary clustering [9]10] is also useful in maintaining consistency, noise removal, cluster correspondence and smoothing. Evolutionary clustering methods use a technique called temporal smoothness to maintain the natural correspondence among the clusters at every timestamp.

Jan Zizka (Eds) : CCSIT, SIPP, AISC, PDCTA - 2013
pp. 193–200, 2013. © CS & IT-CSCP 2013 DOI : 10.5121/csit.2013.3621

With evolutionary clustering, the evolution diagnosis provides the details about changes, in the objects overtime. These details, helps us to specifically determine properties [11] of the object in the clusters. In order to diagnose, various metrics for measuring the quantification and quality of the clusters [7] were proposed.

In this work, probabilistic generative model is used for initialization during the start of the clustering process. The priors are set based on the representativeness of the object in the current cluster and the cluster of the previous time stamp. These priors have certain effect in producing consistent and quality clusters. As well the priors could be set for either target or attribute objects at different time granularities which also make its impact in the creation of clusters.This varying effect of priors is studied when set to different nodes at different time granularities in a heterogeneous network.

2. RELATED WORK

Many studies on clustering problems were now turned towards evolutionary clustering. Deepayan chakrabarti[1] proposed a frame work of evolutionary clustering, in which the emphasis on producing the sequence of clusters reflecting the changes on the current data at any timestamp. The temporal similarity is maintained with the correlation of two time series, T and T – 1. Kumar et.al studied the [5][6] evolution and structure of online social networks by segmenting the networks into regions. With maximumlikelihood principle they present a microscopic analysis of the evolution of social networks. The concept of consistency is incorporated by[2] YunChi. et.al in finding evolutionary patterns of themes [8]in the textstream. They achieve temporal smoothness by preserving cluster quality and preserving cluster membership.

Particle and density [4] based evolutionary clustering method was proposed by Min Soo Him. He proposes cost-embedding technique for achieving temporal smoothness. Aggarwal et. al [7] presented analysis of information Networks, in which various metrics were proposed to diagnose the network apart from producing clusters aside. The initialization of cluster is set using probabilistic generative model, maximum likelihood techniques and expectation maximization approach to produce consistent and quality clusters. However they do not deal with defining priors to different type of nodes. We defined the priors to multi type objects and studied its influence over the consistency and other quality of clusters.

3. EFFECT OF PRIORS OVER MULTI TYPED OBJECT

Our work is greatly inspired by the work done by C.Aggarwal et.al. , in which they propose an evolutionary algorithm that generate net-cluster tree with temporal smoothness. We exploit the variation of E-Netclus by defining priors over target type objects as well as attribute type objects and comparing the results to evaluate the cluster quality. The following are the steps in the method.

Step 1: Initialization
Step 2: Ranking
Step 3: Computing posterior probability for each target object
Step 4: Computing posterior probability for attribute type object
Step 5: Cluster adjustment

We explain each of the steps in further subsections.

3.1. Initialization:

In general priors are the values defined intuitively (at the time of starting) before initializing a cluster. Then these prior values were propagated throughout each cluster which plays important role in the correct alignment of the nodes in the corresponding cluster.

In the case of, previous knowledge about number of clusters that may arise, the prior probability of an object can be defined with a high value. Prior probability = $\{P(O \mid C_k)\}_{k=1}^{K}$ where, O is the object, k is number of clusters. Based on this the successive probabilities are calculated using its representativeness in the current cluster within the net cluster trees with previous time stamp. The smoothness of the cluster sequence also depends on these prior probabilities. In our work as we know the number of clusters, we fix the initial value for the prior to be high, ranges between 0 to 1. With this prior probability initial clusters were generated for target objects.

3.2. Ranking:

To simplify the complex nature of evaluating the similarity between the objects in multi typed heterogeneous network, ranking plays a crucial role. As ranking provides weights to the object, it helps to align the object in a cluster in a better way. A probabilistic generative model is used to model the probability of generation of cluster sequence and ranking is done based on the representativeness of the object in current snapshot at time step t and also includes the representativeness of the object in the previous time step t-1. This is known as authority ranking. In our work frequency based approach authority ranking is used for ranking the term, authors, conferences etc. For example the authors who have published more papers will be having high rank. And thus, the overall probability P (O $|T_o$, C_k) is influenced by the priors and the ranking function.

3.3. Computation of posterior probability for target object:

Maximum likelihood technique is applied to compute the posterior probability $(p(c_k^t \mid o))$ for each target object in a cluster sequence. This is achieved by maximizing the probability function iteratively which is

$$\sum_{i=1}^{|O|} P^t(C_k \mid o_i) / |O|$$

3.4. Computation of posterior probability for attribute object:

The posterior probability of each attribute object $p(c_k^* \mid o)$ in each cluster tree is computed using the posterior probability of target objects of neighboring cluster trees. The priors defined over the objects influences the opt matching of clusters at all levels.

3.5. Cluster adjustment:

The average distance with respect to centroids in each cluster is computed and the objects are assigned to the nearest cluster. Steps 2, 3 and 4 were repeated iteratively, until there is only a smaller ratio of change in the clusters.

4. EVALUATION OF CLUSTER QUALITY

In Multi typed information networks, the awareness of prior with ranking provides more efficient way of producing clustering than with traditional clustering. The quality of clusters could be evaluated by measuring the consistency and compactness of the cluster generated. Entropy of cluster also determines the quality of the cluster. We define the functions to measure the above stated properties. Consistency of a cluster C is the average of the consistencies of all the levels.

$$\text{Consistency}(C,t_1,t_2) = \frac{1}{|O|}\sum_{o\in O} \frac{\sum_{k=1}^{k} b_k(O)_{t1}\, x b_k(O)_{y2}}{\sqrt{\sum_{k-1}^{k} b_k(O)_{t1^2}}\sqrt{\sum_{k=1}^{k} b_k(O)_{t_2^2}}}$$

Average ratio of intra cluster similarity to inter cluster similarity can be defined as compactness. Higher value of compactness implies better quality of clustering. Quality of cluster is also implied with entropy. Lower entropy implies high quality of clustering.

$$\text{Compactness } C = \frac{1}{|O|}\sum_{o\in O} \frac{\sum_{k=1}^{k} b_k(O)_{t1}\, x b_k(O)_{y2}}{\sqrt{\sum_{k-1}^{k} b_k(O)_{t1^2}}\sqrt{\sum_{k=1}^{k} b_k(O)_{t_2^2}}}$$

Where O is the set of target objects, $S(O_{ki},C_k)$ measures the similarity and C_k is the centroid of the cluster.

$$\text{Entropy } E = -\frac{1}{|O|}\sum_{k=1}^{k} \sum_{0=1}^{|0k|} b_K(0)\log(b_k(O))$$

5. EVOLUTIONARY ANALYSIS

The cluster tree views could be used in diagnosing the evolution with various metrics that provide more informative views such as cluster merge rate and split rate, continue rate, appearance rate and disappearance rate. An Object may continue to be the member of a cluster if shows increasing rate of probability. This probability should be compared with the timestamp t and t -1.

$$\text{Continue rate of cluster } C_i = \frac{1}{|O|}\sum_{o\in O} \min\left(\frac{b_i(O)_t}{b_i(O)_{t-1}},1\right)$$

As well, if the membership probability of an object to a particulars cluster increases over time then it is said to be merging with the particulars cluster, if the membership probability, decreases, then it is said to be splitting out from the corresponding cluster.

$$\text{Merge rate of cluster } C_i = \frac{1}{|O|}\sum_{o\in O} \min\left(\frac{b_i(O)_t - b_i(O)_{t-1}}{b_i(O)_t},0\right)$$

When, most of the object were missing in the in the previous timestamp, then they are said to be a new cluster.

$$\text{Appearance rate} = \frac{\sum_{o\in O'} b_c(O)_t}{\sum_{o\in O''} b_c(O)_t}$$

If most of the object of the cluster were not present in time t + 1 and present time t, then that cluster is said to be disappearing.

5. EXPERIMENTS AND RESULTS

Experiments were performed with synthetic dataset resembling a subset of DBLP network. It consists of data objects such as author, paper, conference and terms over the years 2008 to 2011. Terms are the keywords taken from the paper title. This dataset contains approximately 2000 papers(with 1000 terms) written by 1400 authors in 16 conferences in the four areas of data mining, information retrieval, machine learning and databases. The time stamp is the year which is considered as period interval. In this work, time stamp is taken as one year. The paper (node) is treated as target object and others as attribute types. The number of clusters was set to 4. The prior weight is fixed at 0.8.

The purpose of this experiment is to evaluate the effectiveness of prior in constructing cluster sequences and there by diagnosing the evolution. Initially clustering is performed with the synthetic dataset and later quantification consistency, compactness and entropy were studied by setting prior to the attribute type term, author and conference.

It was found that the rate of consistency and compactness were increased when the priors are set to author node rather than to term node and contrastingly the rate of consistency is high and compactness is found lower when the priors are set to conference node. Table 2 shows the varied rate of consistency, compactness, entropy of the clustering when the priors were changed.

Table 1. Sample data consisting of different node types

S.No	Authors	Paper	Conference
1	Philipp Stuermer Anthony Bucci Branke Pablo Funes Elena Popovici	Analysis of coevolution for worst case optimization	GECO
2	Alexandros Agapitos,Michael ONeill ,Anthony Brabazon	Promoting the generalisation of genetically induced trading rules	CFE
3	Edwin D de Jong, Anthony Bucci	Objective Set Compression Test Based Problems and Multiobjective Optimization	IADIS
4	Martin Anthony	Probability theory in machine learning	CFE
5	Walterio Mayol Cuevas,Ezra Winston	Improving Image Sets Through Sense Disambiguation and Context Ranking	IEEE

Table : 2 Variation in consistency and compactness, entropy according to the node types

Node type	Term	Author	Conference
Consistency	0.18616	5.5454	2.6704
Compactness	77.51477	87.94260	66.11830
Entropy	0.15928	0.27244	1.27943

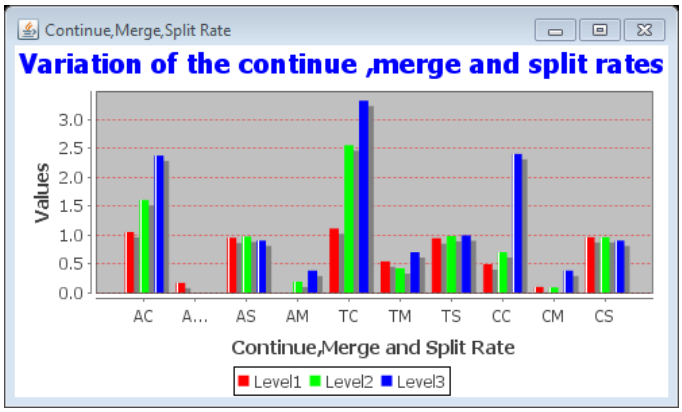

Figure.1 Varition of continue, merge and split rate while priors set to term node

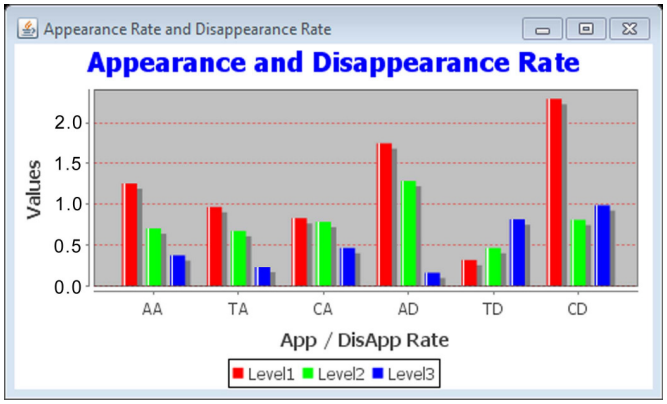

Figure.2 Varition of appearance and disappearance rate while priors set to term Node

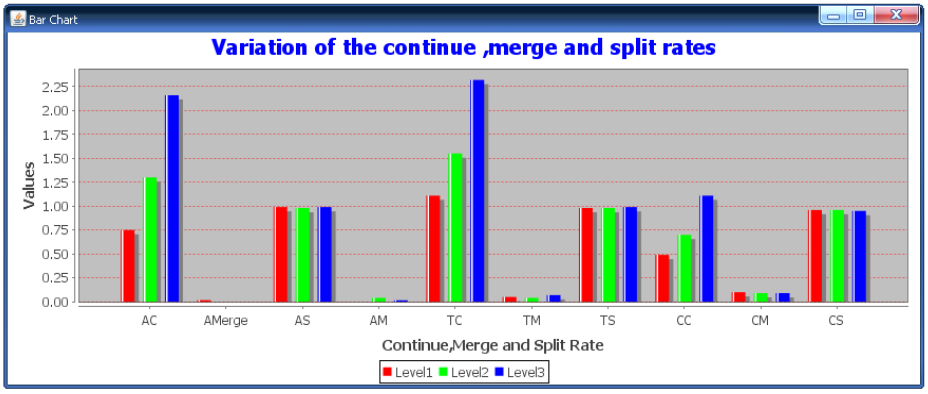

Figure.3 Varition of continue, merge and split rate while priors set to author node

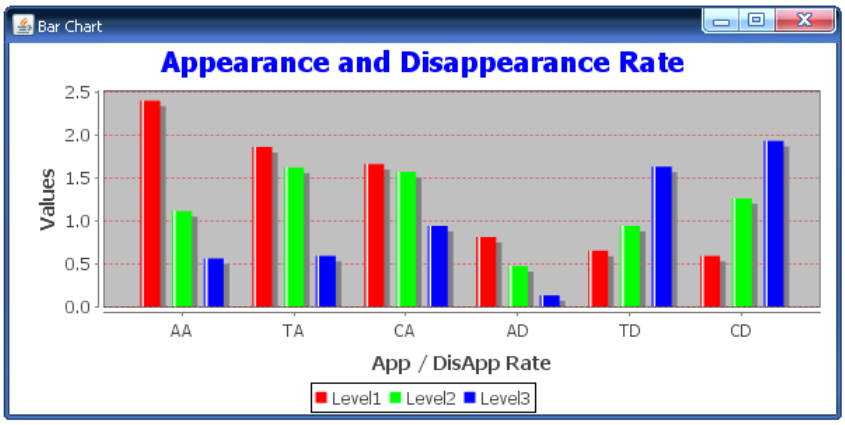

Figure.4 Varition of appearance and disappearance rate while priors set to author node

The observation clearly shows that the setting of priors to different nodes at different time granularities influences the characteristic change in the evolution. The above result shows that if priors were used for term, conference nodes, these object, often overlaps and gets correlated since more authors would use titles with common terms which would result in lower rate of consistencies and compactness. Similarly when priors were set to conference nodes, we find quite lower consistency and compactness. We believe, that many authors would publish papers in more than one conference, so that the clustering is confused between the current snapshot and the historical clusters which might result in lower consistency. Meanwhile the author node when set as prior nodes result in higher rate of consistency. As lower entropy implies better quality cluster we average the consistency, compactness and entropy rates. With this it is observed that quality clusters were produced when priors are set over author node. This is because it is a unique node type, and so that there will be no confusions with the current and prior information and it tends to converge to a better local maxima of log likelihood.

Apart from these quality measures, the evolution study of other metrics such as appearance, disappearance, continue, merge and split rates shows similar variations. The merge rate is lower than split rate, these rates do not vary. Likewise the appearance and disappearance rate, shows same variations irrespective of prior set to different nodes. When the prior node is set as term, the disappearance rate is higher than appearance rate, since author select specialized areas specifically rather than sub areas. But when prior node is set as author node, we find the disappearance rate is lower than appearance rate.

6. CONCLUSION:

In this paper, the authors studied the problem of prior setting over different nodes in heterogeneous network and proposed a new method. The proposed method was implemented with java code and the results were interpreted. Our study analyses the influence of priors and showed that qualitative and quantification properties of clusters differ according to prior setting. Our observation depicts clearly that the quality of the clusters can be affected by prior setting and plays a major role in evolution diagnosis. In future work we will further examine this influence of priors over the social and information networks with extended evolutionary version of K-means clustering algorithm, Dirichlet HTM model and the issues of evolutionary metrics will also be tested. To reduce the burden of iterations, the algorithm will be stuffed with additional features such as automatic inference of prior weight with prior setting. We believe the result would be an interesting turning point in evolutionary clustering.

REFERENCES

[1] D. Chakrabarti. R. Kumar, and A. Tomkins. Evolutionary clustering. KDD(2006)
[2] Y. Chi, X. Song, D. Zhou, K. Hino, and B. Tseng. Evolutionary spectral clustering by incorporating temporal smoothness. KDD (2007)
[3] V. Illango, R. Subramanian, V. Vasudevan. Cluster Analysis Research Design model, problems, issues, challenges, trends and tools. IJCSE Vol.3No. 8 August 2011
[4] M. Kim and J. Han. A Particle-and-Density Based Evolutionary tenng Method for Dynamic Networks. VLDB (2009)
[5] R.Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. KDD (2006)
[6] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. KDD (2008)
[7] Manish Gupta, Charu C. Aggarwal, Jiawei Han, Yizhou Sun, Evolutionary Clustering and Analysis of Bibliographic Networks. Advances in Social Networks analysis and Mining (2011)
[8] Q, Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. KDD(2005)
[9] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. Rankclus: inte¬grating clustering with ranking for heterogeneous information network analysis. EDBT (2009)
[10] Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. KDD (2009)
[11] L. Tang, H. Liu, J. Zhang, and Z. Nazer. Community evolution in dynamic multi-mode networks. KDD (2008)