Special Issue Published in International Journal of Trend in Research and Development (IJTRD),
ISSN: 2394-9333, www.ijtrd.com

Rule Based Mining for Semi-Cluster Network

[1]D.GopalaKrishnan, [2]Dr.D.Santhi Jeslet
[1]Research scholar, [2]Professor & Head,
[1,2]Department Of Computer Science, [1,2]M.G.R Arts and Science College, Hosur, India

Abstract: The research we study the performance of semi-cluster based network algorithms in networking , a new algorithm to address the shortcomings of both routing and clustering based as well as the traditional algorithms. Due to the unstable nature of MANETs memory based search algorithms such as clustering based methods become efficient and practical as the network mobility increases. Our simulations show that ant based routing algorithms suffer from lack of accuracy while incurring extensive memory usage as well as valuable processing power and bandwidth costs in highly dynamic environments. The proposed semi-cluster based routing algorithm exploits a dual-mode approach. Each node can independently operate in a local mode as well as in global mode using the information provided by rule based mining. The network topology changes are constantly monitored. When the rate of topology change gets too high for converge efficiently, a node switches to local mode to rely less on the information learnt through ants. Our simulations show how this switching approach significantly improves the performance of the network by quickly adapting to the dynamics of the environment.

I. BASIC CONCEPTS

Semi-clustering techniques vary tremendously in how they use distances to determine groupings and there is no universal definition of what a semi-cluster is or what properties it must have. Before a formal definition of semi-clustering is presented, some of the notations are introduced.

Let R= {A1,A2,.., Am} be a relation schema and r be a relation over R where |R|=m and |r|=n. The convention used is that the symbols from the end of the alphabet X, Y refers to the sets of attributes and symbols from the beginning A, B refer to single attributes.

A semi-cluster is a set of tuples. For a specific set of attributes X, certain restrictions are placed on the properties of these tuples when projected on X. For this reason, it can be said that a semi-cluster is “defined on” X and the semi-cluster is denoted as Cx.

A possible quality measure on a one dimensional semi-cluster is the range or smallest interval consisting all points or on two dimensions, the area of the smallest bounding box. However, the area does not reflect the density or coverage of points within the semi-cluster. Hence, it is chosen to use a common measure from statistics, the average pair-wise (intra-semi-cluster distance) or diameter of a semi-cluster . δx is used to denote a distance metric on values in the attribute set X, such as the Euclidean or Manhattan distance.

- *Definition 1*

The diameter d on X of a set of tuples S = {ti: 1 ≤ i ≤ N} is the average pair wise distance between tuples projected on X.

In order to find semi-clusters in finding factions, the quality of semi-clustering is restricted using thresholds on the semi-cluster size and the diameter.

- *Definition 2*

A semi-cluster Cx defined on a set of attributes X is any subset of r that satisfies the following for some density threshold d0x and the frequency threshold s0.

The first criterion ensures that the semi-cluster is sufficiently dense. The second criteria ensures that the semi-cluster is frequent, i.e., that it is supported by a sufficient number of tuples.

II. SEMI-CLUSTER AND PREDICTION

Prediction is used to build models that help predict future data values. The difference is that classification predicts the categorical label of a tuple, while prediction models a continuous-valued function. The process of classification begins by identifying one of the attributes of the tuples as the class label. The data set that is used to build the model is called the training data set. Because the tuples in the training data has a provided class label, this is a supervised learning method. In the second step, the model is evaluated. Usually this involves a test data set that is independent of the training data. The model is used to classify the test data and the result is compared to the class labels of the test data. If a high ratio of correctly classified tuples is obtained, the model can be used to classify new tuples with unknown class label. It is important to keep test and training data separate. Most classification methods are susceptible to over fitting, that is given enough training they learn the structure of the training data. Classification rules, decision trees or mathematical formulae can for example, represent a learned model.

III. SEMI-CLUSTERING

The job of assigning tuples to pre-defined class labels is referred to as classification, whereas the task of discovering classes to which the tuples belongs to is referred to as semi-clustering. In general, semi-clustering is categorized as an unsupervised learning method, since there are no labeled data to train the algorithm. Hence, in semi-clustering, the data is grouped into semi-clusters. A general description of a semi-cluster is that the tuples that lie in the same semi-cluster should be similar, and they should be dissimilar to tuples that are not in the same semi-cluster. Semi-clustering can be used as a preprocessing step for classification. It can also be used as a tool by itself, to identify different segments in the data. The most often used measure of evaluating a semi-cluster is the attribute distance. It is always preferred when using semi-clustering that, the distance between tuples belonging to the same semi-cluster to be less than the distance to tuples in different semi-clusters. When distance is used as measure, a metric is needed. A general form is the Minkowski metric:

When q = 2, it is more known as the Euclidean metric. Setting q = 1 gives the Manhattan or "city block" distance. So to assess the result of a semi-clustering algorithm, the simplest and most widely used criterion function is the sum-of-squared-error function where c is the number of semi-clusters, Di is the tuples in semi-cluster i and mi is the mean value of the tuples in that semi-cluster.

For interval data, there is a whole field devoted to the discovery and analysis of data groupings that reflect the relative distances between data points (Srikant et al 1995).

IV. THE PROPOSED SEMI-CLUSTERING BASED MODEL

The conventional hierarchical semi-clustering algorithms such as single-link and complete-link suffer higher time complexity. As a result, a recent trend is to develop hybrid-semi-clustering algorithms that exploit the advantages of both hierarchical and partitioned algorithms. Hence, a semi-clustering algorithm Birch (Tian Zhang et al 1996) has been utilized for finding factions which are treated as a group of semi-clusters. The idea is to use a standard semi-clustering algorithm to identify the intervals of interest followed by the construction of Coalescent Dataset in order to check the applicability of the rules outside of the dataset. The semi-clustering algorithm uses a single partitioning of the attributes into disjoint sets (Xi) over which there is a meaningful metric. Most often, each Xi an individual attribute or a small set of closely related attributes over similar domains. The semi-clusters are created incrementally and represented by a compact summary. The summaries produced in the first phase are then used for the construction of the Coalescent Dataset approach.

V. SEMI-CLUSTER$ METHODOLOGY

BIRCH (Balanced Iterative Reducing and Semi-clustering using Hierarchies) is an incremental and hierarchical semi-clustering algorithm for large databases. The strongest point of the Birch algorithm is its support for very large databases (main memory is lower than the size of the DB).

There are two main building components in the Birch algorithm:

The hierarchical semi-clustering component and the main memory structure component. The idea of a hierarchical semi-clustering is illustrated the algorithm starts with single point semi-clusters (every point in a database is a semi-cluster, Semi-clustering Feature CF shown in figure 4.2(a)). Then it groups the closest points into separate semi-clusters (Figure 4.2(b), Figure 4.2(c)), and continues, until only one semi-cluster remains (Figure 4.2(d)). The computation of the semi-clusters is done with a help of distance matrix (O(n2) large) and O (n2) time.

VI. SYNTHETIC DATA SET

The popular method for generating transactions containing associations, originally proposed by (Agrawal and Srikant 1994) was utilized. The same probability distributions and the notations were followed as mentioned in (Agrawal and Srikant 1994). Transaction sizes are typically semi-clustered around a mean, with a small portion of transactions having many items. Typical sizes of inter-transaction itemsets are also semi-clustered around a mean with a small portion of frequent transaction itemsets having a large number of items across different transactions.

A set L of the potentially frequent transaction itemsets which may span across different transactions were first generated and a frequent inter transaction itemset from L was assigned to corresponding transactions. Therefore, the number of potentially frequent itemsets for a faction was set to |L|. A potentially frequent inter transaction itemset was generated by first picking the size of the itemset from a Poisson distribution with mean equal to |I|. The maximum size of the potentially frequent inter - transaction itemsets is |max (|I|)|. Items and their intervals in the first frequent inter-transaction itemset are chosen randomly, where item is picked up from one to | ∑ | , and its interval is picked up from Zero to R-1. To model the phenomenon that frequent inter-transaction itemsets often have common items and intervals, some fraction of items and their intervals in subsequent itemsets are chosen from the previous itemsets generated. An exponentially distributed random variable with mean equal to the correlation level was used to decide this fraction for each itemset. The remaining items and their intervals are picked up at random. In the datasets used in the experiments, the correlation level is set to 0.5. After generating all the items and the intervals for a frequent inter-transaction itemset, each of its intervals was revised by subtracting the minimum interval value of this frequent itemset. In this way, the minimum interval of each potentially frequent inter-transaction itemset is always zero.

VII. EXTENSIONS

Clustering Rule-Based generated by the basic association Semi-Cluster Rule-Based model are referred to as Boolean association Clustering Rule-Based in the mining literature, since the only relevant information in each database transaction is the presence or absence of an item. Many kinds of Clustering Rule-Based have been proposed in the research literature as extensions to Boolean Association Semi-Cluster Rule-Based Mining. These include hierarchical, quantitative, categorical, cyclic, constrained and sequential Clustering Rule-Based. Each of the above can be described as below:

Hierarchical Semi-Cluster Rule-Based: It is possible to extract a semantically richer set of Clustering Rule-Based, called hierarchical Semi-Cluster (Srikant et al 1995), from a transaction database, if an 'is-a' hierarchy over the set of items in the database is provided. For example, given that sweaters and ski jackets are both instances of winter wear, the Clustering Rule-Based output could contain a "pseudo-item" called a winter wear to denote "either sweater or ski jacket or both". An example of hierarchical Clustering Rule-Based could be winter wear hiking boots.

Quantitative and Categorical Clustering Rule-Based: Relational tables in most businesses and scientific domains have richer attribute types than the Boolean attributes considered in the basic problem for transactional databases. Attributes can be quantitative (e.g., age, income) or categorical (e.g., zip code). The problem of mining association Clustering Rule-Based over such attributes in relational databases has been addressed in (Srikant et al 1996). An example of such a Clustering Rule-Based would be, if Age is between 30..39 and Married = yes, then the Number of Cars = 2.

Cyclic Semi-Cluster Rule-Based: These Semi-Cluster Rule-Based, proposed in (Ozden et al 1998), are association Clustering Rule-Based that display regular cyclic variation over time. For example, the user may wish to compute association Clustering over networks data to observe seasonal variation where certain Clustering Rule-Based are true at approximately the same month each year. Discovering such Semi-Cluster and their periodicities may reveal interesting information that can be used for prediction and decision-making.

Constrained Semi-Cluster Rule-Based: In (Ng et al 1998), the authors propose constrained Clustering Rule-Based as a means of specifying constraints (including domain, class and SQL-style aggregate constraints) which are to be satisfied by the antecedent and consequent of a mined association Clustering Rule-Based.

Sequential Clustering Rule-Based: While standard Boolean association Clustering Rule-Based find associations between items within a single transaction, sequential Clustering Rule-Based proposed in (Agrawal et al 1995), discover associations between items purchased at different times.

Association Clustering Rule-Based mining is an important component in mining all of the above types of patterns. Previous works on generating hierarchical, quantitative and categorical Clustering Rule-Based e.g. (Srikant et al 1995, Srikant et al 1996) have shown that, even if they require some preprocessing, these problems are finally reducible to BAR mining. For cyclic and constrained Clustering Rule-Based, the authors in (Ozden et al 1998, Ng et al 1998), have integrated their techniques with existing BAR-mining algorithms. In (Agrawal et al 1995), the strategy recommended for mining sequential Clustering Rule-Based include a preprocessing stage that consists of standard BAR mining. These examples, combined with the fact that BAR-mining can be successfully applied for classification and clustering tasks indicate that BAR mining is an important high-impact problem.

IX. IMPLEMENTATION OF CLUSTRTING

A key issue that needs more concern when using Association Semi-Cluster Mining is the soundness of the Clustering external to the data set from which they are generated. Clustering are usually the derivative of the patterns in a specific data set. When a different phenomenon occurs, the transformation in the set of Clustering obtained from the new dataset could be significant. This work provides a Group based Mining of Association Clustering (G-MAR) model by paying special attention as how the variation between two different settings affects the changes of the Clustering, based on the notion of fine partitioned groups termed as factions. Using this G-MAR model, a simple technique called Coalescent Dataset, is proposed to get a fine approximation of the set of Clustering for a new situation. The approach proposed, works independently of the core mining process and can be easily implemented with all variations of the Clustering mining techniques.

Also, a fuzzy clustering based Association Clustering Mining system is proposed in targeting customers to improve networks which improvises the G-MAR model by predicting networks based on customer needs and functional features. For a potential customer arriving the store, which customer group one should belong to according to customer needs, what are the preferred functional features or products that the customer focuses on and what kind of offers will satisfy the customer etc., finds to be the key factor in targeting customers to improve networks. Generally, a transactional database is created to record all the products purchased by the customer.

To focus on the market segment that each customer falls into, the transaction database can be grouped into different semi-clusters based on the customer needs.

Research on studying the aspects of association Semi-Cluster Rule-Based mining includes improving the performance of Semi-Cluster Rule-Based generation (Agrawal et al 1994, Agrawal et al 1995, Brin et al 1997, Mannila et al 1994, Park et al 1995, Rajamani et al 1999, Savasere et al 1998, Sarawagi et al 1998, Toivonen 1996), extending the scope of association Semi-Cluster Rule-Based mining to cover diverse data types and data sources (Brin et al 1997, Miller et al 1997, Ozden et al 1998, Srikant et al 1995, Srikant et al 1996), constraint-based and user-guided approaches to association Semi-Cluster Rule-Based mining (Klementtien 1994, Ng et al 1998) and the usage of the discovered association Semi-Cluster Rule-Based for further data mining processes. The main contribution of this thesis to the association Semi-Cluster Rule-Based mining in the area of data mining is to define a new model with a technique that can be used to broaden the applicability of association Semi-Cluster Rule-Based. The model can be depicted as: Given the data available from the earlier cases, a set of factions for the new situation, and the proportions of the factions expected in the new situation, sample the original dataset according to the new proportions (i.e., select random transactions from different factions, but select these factions with probabilities according to new proportions), and, finally, learn association Semi-Cluster Rule-Based from the new sample. The sample that is constructed from the original dataset is termed as the Coalescent Dataset.

From the above outline, it can be seen that the main point of the proposed model is to formulate and sample the Coalescent Dataset, which is independent of any core mining algorithm. Existing algorithms as how to discover association Semi-Cluster Rule-Based proposed by various researchers can be used for association Semi-Cluster Rule-Based generation. For different cases, different algorithms can be used. For example, if the taxonomies (is-a hierarchies) over items is available in the data set, the algorithm proposed in (Srikant et al 1995) can be used, since it has studied the problem of mining association Semi-Cluster Rule-Based, where taxonomies over items are available. If the classification hierarchy over items is available and the parallel algorithm is preferred, then the various algorithms proposed to discover association Semi-Cluster Rule-Based can be used.

CONCLUSION

The use of data mining in enrollment management is a fairly new development. Current data mining is done primarily on simple numeric and categorical data. In the future, data mining will include more complex data types. In addition, for any model that has been designed, further refinement is possible by examining other variables and their relationships. Research in data mining will result in new methods to determine the most interesting characteristics in the data. As models are developed and implemented, they can be used as a tool in enrollment management.

References

[1] Data mining concept and technique-book of jeiwei Han-2015

[2] Data mining concept textbook-charu agarwal-2013

[3] Data mining concept introductory and advanced topic-margrate H. dunham-2014

[4] Statistical Analysis and Data Mining: The ASA Data Science Journal First published: 16 August

[5] 2016- Nickolay T. Trendafilov Sara Fontanella

[6] Statistical Analysis and Data Mining: The ASA Data Science Journal First published: 16 July 2017- Giovanni Montana