Hierarchical Clustering and Association Rule Discovery Process for Efficient Decision Support System

Bobby D. Gerardo[1], Yung-Cheol Byun[2,*], and Bartolome Tanguilig III[3]

[1] Institute of ICT, West Visayas State University
Luna St., Lapaz, Iloilo City, Philippines
bgerardo@wvsu.edu.ph
[2] Dept. of Computer Engineering, Jeju National University
Jeju City, Korea
ycb@jejunu.ac.kr
[3] Technological Institute of the Philippines, Cubao, Quezon City
bttanguilig_3@yahoo.com

Abstract. This paper proposed a model based on hierarchical Clustering and Association Rule, which is intended for decision support system. The proposed system is intended to address the shortcomings of other data mining tools on the processing time and efficiency when generating association rules. This study will determine the data structures by implementing the cluster analysis which is integrated in the proposed architecture for data mining process and calculate for associations based on clustered data. The results were obtained using the proposed system as integrated approach and were rendered on the synthetic data. Although, our implementation uses heuristic approach, the experiment shows that the proposed system generated good and understandable association rules, which could be practically explained and use for the decision support purposes.

Keywords: Data mining, decision support system, clustering, association rules.

1 Introduction

Often, there are many attributes or dimensions that are contained in the database, and it is possible that subsets of such dimensions are highly associated with each other. The dimensionality of a model is determined according to the number of input variables used. Clustering can be used to group data into clusters so that the degree of association is strong between members of the same cluster and weak between members of different clusters [1], [9]. Thus, each cluster describes the class to which its members belong. For that reason, cluster analysis can reveal similarities in data which may have been otherwise impossible to find.

Data cubes allow information to be modeled and viewed in multiple dimensions and such cubes are then defined by the dimensions and facts [1]. They defined

* Corresponding author.

T.-h. Kim et al. (Eds.): FGCN 2011, Part II, CCIS 266, pp. 239–247, 2011.

dimensions as entities with respect to which an organization wants to keep records of. Data cubes may be used in theory to answer query quickly, however, in practice they have proven exceedingly difficult to compute and store because of their inherently exponential nature [7].

Moreover, issues that other researchers observed in the data mining tasks were computing speed, reliability of the approach for computation, heterogeneity of database, and vast amount of data to compute [1], [2], [7].

This paper explore the formulation of the cluster analysis technique as integrated component of the proposed model to partition the original data prior to implementation of other data mining tools. The model that we proposed uses the hierarchical nearest neighbor clustering method and apriori algorithm for association mining implemented on transactional databases.

2 Related Studies

Association rule mining tasks includes finding frequent patterns, associations, or causal structures among sets of items or objects in transactional databases and relational databases. Data mining uses various data analysis tools such as from simple to complex and advanced mathematical algorithms in order to discover patterns and relationships in dataset that can be used to establish association rules and make effective predictions.

2.1 Cluster Analysis

The goal of cluster analysis is categorization of attributes like consumer products, objects or events into clusters or groups, so that the degree of correlation is strong between members of the same cluster and weak between members of different clusters. Each group describes the class in terms of the data collected to which its members belong. It may show structure and associations in data, although not previously evident, but are sensible and useful once discovered. The results of cluster analysis [9] may contribute to the definition of a formal classification scheme, such as in taxonomy for related animals, insects or plants; suggest statistical models with which to describe populations; indicate rules for assigning new cases to classes for identification and diagnostic purposes; provide measures of definition, size and change in what previously were only broad concepts.

2.2 Apriori Algorithm

There are varieties of data mining algorithms that have been recently developed to facilitate the processing and interpretation of large databases. One example is the association rule algorithm, which discovers correlations between items in transactional databases. The Apriori algorithm is used to find candidate patterns and those candidates that receive sufficient support from the database are considered for transformation into a rule. This type of algorithm works well for complete data with

discrete values. Some limitations of association rule algorithms, such as the Apriori is that only database entries that exactly match the candidate patterns may contribute to the support of that candidate pattern. In the past years, there were lots of studies on faster, scalable, efficient and cost-effective way of mining a huge database in a heterogeneous environment. Most studies have shown modified approaches in data mining tasks which eventually made significant contributions in this field. However, there are limitations on generated rules, like producing enormous, unclear and sometimes irrelevant rules.

3 System Architecture

The proposed architecture for the data mining system is shown in Figure 1. Its refinement is presented in the subsequent sections.

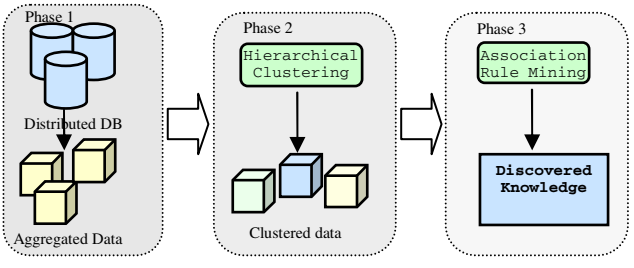

Fig. 1. The general view of the proposed model

Figure 1 shows the proposed three phase architecture, where the first phase is the data preprocessing stage that performs data extraction, transformation, loading and refreshing. This will result to an aggregated data cubes as shown in the same figure. Phase 2 shows the implementation of the hierarchical nearest neighbor clustering, while Phase 3 is the implementation of Apriori algorithm to generate rules.

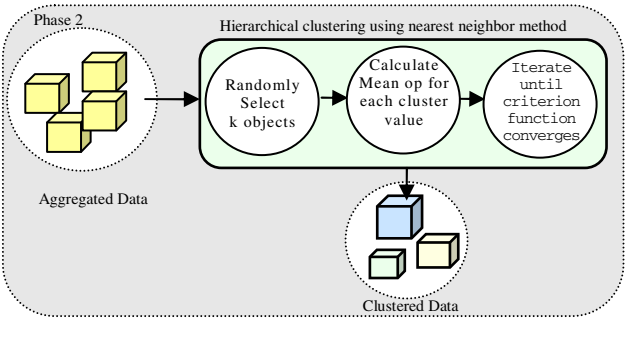

Fig. 2. Refinements of the model at Phase 2

Figure 2 shows the implementation of the cluster analysis using the hierarchical nearest neighbor clustering algorithm while Figure 3 is the implementation of association rule discovery method. Figure 3 shows the refined view of Phase 3. In this illustration, association rule algorithm is used as part of the data mining process. The successions of transforms for association rule algorithm which are represented by bubbles are shown in the shaded rectangle.

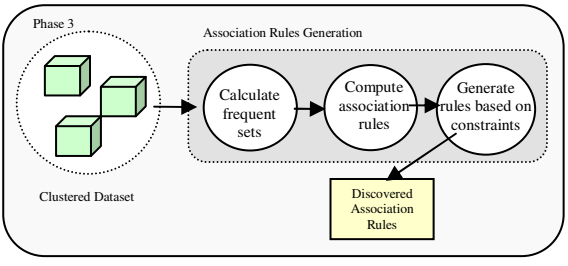

Fig. 3. Refinements of the model at Phase 3

Phase 3 is the final stage in which the association rule algorithm will be implemented to generate the association rules. This calculates for the frequent itemsets and then compute for the association rules using the threshold for support and confidence. The output is given by the last rectangle showing the discovered rules. In this study, the discovered rules are provided in the table showing the support count and the strength of its confidence which are presented in section 5.

The process allows the data to be modeled and viewed in multiple dimensions. Cluster analysis will generate partitions of the dataset, and then the association rule discovery process will be employed. The data cubes will reveal the frequent dimensions, thus, could generate rules from it. The final stage is utilization of the result for decision support. The proposed architecture will implement the association rule generation on a clustered database and would expect better data mining results.

4 Cluster Analyses and the Proposed Model

Among the most popular hierarchical clustering methods are Nearest-Neighbor, Farthest-Neighbor, and Minimal Spanning Tree while for non- hierarchical methods are K-Means, Fuzzy K-Means, and Sequential K-Means. This study put more emphasis on the use of hierarchical method as shown in the experimental results.

4.1 Types of Cluster Analysis

Cluster analysis is a method used for partitioning a sample into homogeneous classes to create an operational classification. Such classification may help formulate hypotheses concerning the origin of the sample, describe a sample in terms of a typology, predict the future behavior of population types, optimize functional processes for business site locations or product design, assist in identification as used

in the medical sciences, and measure the different effects of treatments on classes within the population [9].

Nearest-Neighbor clustering is one of the simplest agglomerative hierarchical clustering methods, which is also known as the nearest neighbor technique. The defining feature of the method is that distance between groups is defined as the distance between the closest pair of objects, where only pairs consisting of one object from each group are considered [10]. An agglomerative hierarchical clustering procedure produces a series of partitions of the data, P_n, P_{n-1} until P_1. The first P_n consists of n single object clusters, while the last P_1 consists of single group containing all n cases. At each particular stage the method joins together the two clusters which are closest together or are most similar. Figure 4 shows the algorithm for the nearest neighbor clustering.

Given:
A set X of objects $\{x_1,...,x_n\}$, A distance function $dis(c_1,c_2)$

1. **for** $i = 1$ to n
   $c_i = \{x_i\}$
   end for
2. $C = \{c_1,...,c_b\}$
3. $l = n+1$
4. **while** C.size > 1 **do**
   a) (c_{min1},c_{min2}) = minimum $dis(c_i,c_j)$ for all c_i,c_j in C
   b) remove c_{min1} and c_{min2} from C
   c) add $\{c_{min1},c_{min2}\}$ to C
   d) $l = l + 1$
   end while

Fig. 4. The nearest neighbor clustering algorithm

Although there are several other hierarchical clustering methods, in this study, the nearest neighbor had been utilized as part of our proposed model for clustering the data.

4.2 Implementation of the Proposed Model

The models in section 3 as reflected in Figures 1, 2 and 3, respectively, will be implemented in a heuristic process. In our experiment, we will calculate for the clustered data based on the proposed model. And then the outputs will be processed by implementing the Apriori for association mining. Tables showing the comparison of the results on original dataset, the proposed model and the discovered rules are presented in section 5.

5 Simulation and Results

The simulation was done on the database containing 30 attributes comprising of six (6) major dimensions and a total of 1,000 tuples of e-commerce and transactional

types of data. The evaluation platforms utilized in the study were IBM compatible computer, Windows OS, C++, and Python.

For the purposes of illustrating the database used in the experiment, we present the Dataset showing partially the data as revealed in Table 1. The abbreviated notations for the attributes stand as follows: A_n= books and its corresponding subcategories, B_n = Electronics, C_n = Entertainment, D_n= Gifts, E_n = Foods, and F_n = Health. Furthermore, A_n Book attribute is consist of subcategories like A_1= Science, A_2=social, A_3=math, A_4=computer, A_5=technology, A_6=religion, and A_7=children books. Other dimensions are written with notations similar to that of A_n. The discrete values indicated by each record are corresponding to the presence or absence of the attribute in the given tuples. Supposed that we consider the problem of determining how often consumers buy products and the probability of purchasing some items online. The results which will be presented in the subsequent sections will answer this problem.

Most literatures assumed that the hierarchical clustering procedure is suitable for binary or counts data types [1], [7], [8], [10]. The method that we considered for cluster analysis, which is integrated in proposed model is just suitable for the dataset that we assumed. Consumers respond to questions by giving their agreement or disagreement on buying some products online.

5.1 Hierarchical Clustering Results

The simulation will identify relatively homogeneous groups of cases based on selected characteristics. It is observed that a total of 4 clusters had been created and the group membership of each case is shown in Table 1. In the clustering result, the minimum distance of each case indicates its membership to the cluster. In summary, cluster 1 has a total of 433 cases (43.3%), cluster 2 has 235 cases (23.5%), cluster 3 has 165 (16.5%) and cluster 4 has 167 cases (16.7%).

5.2 Comparison of Data Mining Result after the Implementation of the Model

The data mining results using the two approaches are shown on Table 1 which also shows their corresponding values. The same table presents the number of cases that belong to the respective clusters.

After implementing the clustering, we then employed the association rule algorithm (Apriori property). The results is shown in Table 2. The use of such algorithm is for discovering association rules that can be divided into two steps: (1) find all itemsets (sets of items appearing together in a transaction) whose support is greater than the specified threshold. Itemsets that meet the minimum support threshold are called frequent itemsets, and (2) generate association rules from the frequent itemsets. All rules that meet the confidence threshold are reported as discoveries of the algorithm.

Table 1. Clusters and the Discovered Rules, Support ≥0.90

		Original Dataset		Clustered (Cases, 433, 235, 165, 176)	
Clusters →	All (1,000)	1 (433)	2 (235)	3 (165)	4 (176)
Number of Rules→	1,758	1,154	708	650	548

Table 2. Comparison of the Discovered Rules

Models	Discovered Rules (showing first 5 rules generated)	Support	Confidence
Original (1,758 rules)	A6=Buy -> A2=Buy F4=Buy	0.935	0.942
	A6=Buy -> A2=Buy A3=Buy	0.927	0.934
	A6=Buy -> A2=Buy C4=Buy F4=Buy	0.916	0.922
	A6=Buy -> A2=Buy D2=Buy F4=Buy	0.915	0.921
	A6=Buy -> A2=Buy A3=Buy C4=Buy	0.910	0.916
Cluster Analysis			
1(1154rules)	A6=Buy -> A2=Buy F4=Buy	0.924	0.939
	A6=Buy -> A2=Buy A3=Buy	0.919	0.934
	A6=Buy -> A2=Buy D2=Buy F4=Buy	0.905	0.920
	A6=Buy -> A2=Buy C4=Buy F4=Buy	0.903	0.918
	A6=Buy -> A3=Buy F4=Buy	0.901	0.915
2 (708 rules)	A6=Buy -> A2=Buy F4=Buy	0.912	0.923
	A6=Buy -> A2=Buy A3=Buy	0.910	0.921
	A6=Buy -> A2=Buy	0.963	0.974
	A6=Buy -> A2=Buy C4=Buy	0.942	0.953
	A6=Buy -> A2=Buy D2=Buy	0.940	0.951
3 (650 rules)	D2=Buy -> F3=Buy	0.921	0.938
	D2=Buy -> F2=Buy	0.909	0.926
	D2=Buy -> C4=Buy	0.958	0.975
	D2=Buy -> A2=Buy	0.958	0.975
	D2=Buy -> A3=Buy	0.945	0.963
4 (548 rules)	D2=Buy -> F3=Buy	0.922	0.939
	D2=Buy -> F2=Buy	0.910	0.927
	D2=Buy -> A2=Buy	0.958	0.976
	D2=Buy -> C4=Buy	0.952	0.970
	D2=Buy -> A6=Buy	0.946	0.963

The result only shows the first five rules generated for each of the cluster. The support threshold that we set prior to the experiment was 0.90. In the original dataset, those who buy A6 (books on religion) will most likely buy A2 (books on social science) and F4 (Health supplement) with support of 0.935 and confidence of 0.942 (94.2% probability of buying). The same fashion of explanation and analysis could be done to other rules.

In cluster 1, those who buy A6 (books on religion) will most likely buy A2 (books on social science) and F4 (Health supplement) with support of 0.924 and confidence of 0.939 (93.9%). Similar approach of analysis could be made for other rules in this cluster. And a similar fashion of explanation could also be done for other rules discovered such as in clusters 2, 3 and 4, respectively.

In principle, there would be an improvement in processing time since the computation of rules is based on chunks of data, i.e. clustered data. Shorter processing time had been observed to compute for smaller clusters attributes implying faster and ideal processing period than processing the entire dataset.

5.3 Further Analysis and Implications

The blending of cluster analysis and association rule generation in the proposed model specifically isolate groups of correlated cases using the hierarchical nearest neighbor clustering and then using of the extended data mining steps like the algorithm for association rule generation. The model identify relatively homogeneous groups of cases based on selected characteristics and then employed the Apriori algorithm to calculate for association rules. This resulted to some partitions where we could conveniently analyze specific associations among clusters of attributes. This further explains that the generated rules were discovered on clusters indicating highly correlated cases which will eventually implies simplification of analysis of the result, thus beneficial to be used for decision support purposes.

6 Conclusions and Recommendations

The model reveals clusters that have high correlation according to predetermined characteristics and generated isolated but imperative association rules based on clustered data which in return could be practically explained for decision support purposes. The rules generated based on clustered attributes indicates simple rules, thus it could be efficiently used for decision support system such as in policy making or top level decision making. For future works, upgrade of the model based on extended clustering methods like divisive and non-hierarchical clustering may be needed to check if it performs well with other mechanisms.

Acknowledgments. This research was financially supported by the Ministry of Education, Science Technology (MEST) and Korea Industrial Technology Foundation (KOTEF) through the Human Resource Training Project for Regional Innovation.

References

[1] Han, J., Kamber, M.: Data Mining Concepts & Techniques. Morgan Kaufmann, USA (2001)

[2] Pressman, R.: Software Engineering: a practitioner's approach, 6th edn. McGraw-Hill, USA (2005)

[3] Hellerstein, J.L., Ma, S., Perng, C.S.: Discovering actionable patterns in event data. IBM Systems Journal 41(3) (2002)

[4] Multi-Dimensional Constrained Gradient Mining, ftp://fas.sfu.ca/pub/cs/theses/2001/JoyceManWingLamMSc.pdf

[5] Chen, B., Haas, P., Scheuermann, P.: A new two-phase sampling based algorithm for discovering association rules. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2002)

[6] Margaritis, D., Faloutsos, C., Thrun, S.: NetCube: A Scalable Tool for Fast Data Mining and Compression. In: 27th Conference on Very Large Databases (VLDB), Roma, Italy (September 2001)

[7] Han, E.H., Karypis, G., Kumar, V., Mobasher, B.: Clustering in a high-dimensional space using hypergraph models (1998), http://www.informatik.uni-siegen.de/~galeas/papers/general/Clustering_in_a_High-Dimensional_Space_Using_Hypergraphs_Models_%28Han1997b%29.pdf

[8] Cluster Analysis defined, http://www.clustan.com/what_is_cluster_analysis.html

[9] Determining the Number of Clusters, http://cgm.cs.mcgill.ca/soss/cs644/projects/siourbas/cluster.html#kmeans

[10] Using Hierarchical Clustering in XLMiner, http://www.resample.com/xlminer/help/HClst/HClst_intro.htm

[11] Ertz, L., Steinbach, M., Kumar, V.: Finding Topics in Collections of Documents: A Shared Nearest Neighbor Approach. In: Text Mine 2001, Workshop on Text Mining, First SIAM International Conference on Data Mining, Chicago, IL (2001)

[12] Hruschka, E.R., Hruschka Jr., E.R., Ebecken, N.F.F.: A Nearest-Neighbor Method as a Data Preparation Tool for a Clustering Genetic Algorithm. In: SBBD, pp. 319–327 (2003)