International Journal of Advances in Computer and Electronics Engineering
Volume: 1 Issue: 1, May 2016, pp. 23–27

A Modified Bcbimax Biclustering Algorithm for Market Segmentation

[1]S. Madhu Rupa, [2]R.Balamurugan
[1]PG Scholar, Department of Computer Science and Engineering,
Bannari Amman Institute of Technology, Tamil Nadu, India
[2]Assistant Professor, Department of Computer Science and Engineering,
Bannari Amman Institute of Technology, Tamil Nadu, India
[1]madhu478@ymail.com, [2]balacse05@gmail.com

Abstract: *Market segmentation plays a crucial role in design and development of the product. It separates a large number of customers into meaningful groups who share similar characteristics, requirements and behaviors. This is mainly used to match diverse customer needs or to deploy resources effectively. Hence, it enables companies to increase the opportunities of market success. Market segmentation can be implemented based on the customer pain point. It contains customer's inconvenience, annoying or frustration towards a product. Biclustering based market segmentation by using customer pain points. Different from one way clustering, biclustering will cluster both row and column which is associated with customer and customer pain points. This is mainly used to identify the homogenous subgroup of customers with common characteristics towards a subset of a segmentation variable. But, there is the loss of data during discretization and also there occurs overlapping problems in the existing system. Hence that can be solved by means of using similarity score.*

Keyword: *Biclustering; BCBimax, Similarity score;*

1. INTRODUCTION

Data mining the analysis step of the Knowledge Discovery in Databases (KDD) process. The data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Market segmentation plays a crucial role in product development and has become an essential part of product innovation [6]. According to the information, companies can develop new types of products to match diverse customer needs or deploy resources effectively to manufacture a product for the most potential segment [2].

1.1 Clustering

Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. The one way clustering algorithms aim to divide a set of objects into groups (clusters) by finding a one-way division of data to produce clusters where customers behave similarly over all the segmentation variables [5]. Clustering is also used in outlier detection applications such as detection of credit card fraud. These clustering methods obtain a global model rather than a local model, failing to discover subgroups of customers who have similar characteristics on partial variables, especially in high-dimensional data is the major drawback [9].

1.2 Biclustering

Biclustering is a popular approach to analyze patterns in a dataset, especially those of biological origin such as expression data. Biclustering performs better than classical clustering techniques under certain data sets, since it can simultaneously cluster both rows and columns of matrix. Given a set of m rows in n columns (i.e., an m x n matrix), the biclustering algorithm generates biclusters – a subset of rows which exhibit similar behavior across a subset of columns, or vice versa. The detection of biclusters is an NP-hard problem and the computational complexity is very high [12]. Test clustering can solve the high-dimensional sparse p0.roblem, which means clustering text and words at the same time.

When clustering text, it needs to think about not only the words information, but also the information of words clusters that was composed by words. Then according to similarity of feature words in the text, will eventually cluster the feature words. This is called co-clustering which is represented in

International Journal of Advances in Computer and Electronics Engineering
Volume: 1 Issue: 1, May 2016, pp. 23–27

Figure 1. The idea of biclustering method is an NP-hard problem and the computational complexity is very high. In the practical scenarios, customers share similarly only on a small fraction of variables, such as knowledge, need, attitude, interest and loyalty status [11]. The customer pain points reflect customers core concerns, main interests and emergent needs for products, thus identifying groups of customers who have similar pain points is more beneficial for companies to achieve accurate market segmentation and positioning. Because the pains that make customers uncomfortable, annoying or frustrating towards a product, normally result from deficiencies, shortcomings, problems, or defects of the product [3]. Thus, biclustering was proposed to discover subgroups of customer that share similar transcriptional behaviors over a subset of conditions in a microarray experiment.

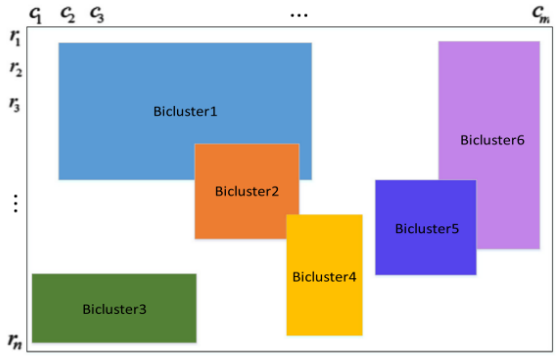

Figure 1 Representation of the biclustering model

1.3 Problem Statement

In an existing work the data is collected from the customer then that can be converted into binary data. In order to get the binary data, data transformation process is executed. So because of discretization there occurs data loss. So this leads to incorrect outcome. The second drawback, this method avoids overlapping, so this made one customer to assign one segment, so in some practical situation it is not possible to find the proper results.

2. METHODOLOGY

2.1 Bcbimax Algorithm

The BCBimax algorithm of biclustering technique is introduced to conduct market segmentation using customer pain points. The original algorithm generates overlapping biclusters, which means that an object can be classified into multiple subgroups. The BCBimax algorithm of biclustering technique is introduced to conduct market segmentation using customer pain points [10]. The BCBimax method that derives from the BiMax method was first put forward to serve as a reference method or a baseline for comparison of main biclustering algorithms employed in expression data as represented in Figure 2.

The data structure is expressed by a binary matrix, where the rows represent customers and the columns represent customer pain points. A set of n customers (rows) form customer pain points (columns) is recorded as a binary matrix E, where an element e_{ij} being 1 represents that customer i chooses pain points j and otherwise e_{ij} equals 0. A bicluster (R , C) corresponds to a subset of customers R that collectively experience a subset of pain points , which means that a pair (R,C) defines a submatrix of E where all elements are 1. But there comes a question that if the value of an element e_{ij} is 1, by definition, it is a bicluster itself. It makes no sense to find such a pattern. On the contrary, the real goal is to discover all biclusters that are inclusion-maximal.

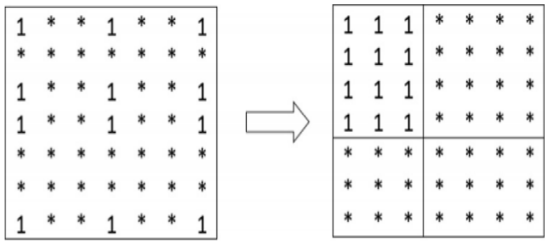

Figure 2 BcBiMax algorithm

The idea of the BCBimax algorithm lies in partitioning a binary matrix E into three sub-matrices, one of which contains only 0s and thus can be removed. Then the remaining two sub-matrices U and V are processed recursively with the algorithm; the recursion stops if the current matrix represents a bicluster that contains only 1s. For the purpose of prohibiting overlaps, the bicluster with the maximum number of 1 is stored and the next bicluster is searched from the data which has excluded the rows of the already found bicluster. In conclusion, the BCBimax algorithm adopts a recursive divide and a conquer strategy to enumerate all biclusters in a binary matrix E. The detail process is illustrated as follows;

Step1. Choose a random row containing a mixture of 1s and 0s to divide the original matrix E into two column sets: C_U and C_V, as shown in Fig. 3. If there are no such rows that fit the criterion, all elements of the matrix either equal 1, in which case the entire matrix is a single bicluster, or all elements equal 0, in which case, it has no biclusters.

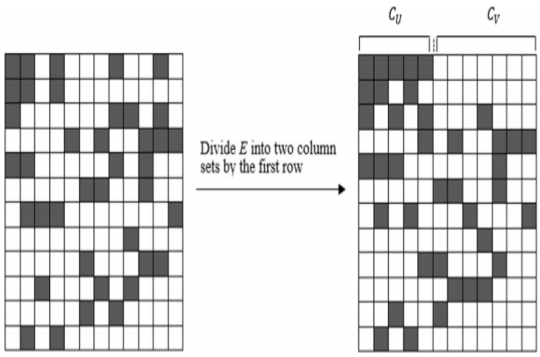

Figure 3 Step 1 of BCBimax algorithm.

International Journal of Advances in Computer and Electronics Engineering
Volume: 1 Issue: 1, May 2016, pp. 23–27

Step 2. Divide the m rows into three sets, R_U are rows with 1s only in column set C_U , R_W are rows with 1s in both C_U and C_V, R_V are rows with 1s only in C_V .
Step 3. Construct two submatrices U = (R_U U R_W, C_U) and V = (R_W U R_V, C_U U C_V) delete the empty submatrix formed by (R_U, C_V). After being rearranged the matrix looks like the matrix drawn in Figure 4.
Step 4. Process recursively U and V through repeating step 1 to 3 until the pre-set minimum size of matrix is found and report those matrices with only 1s. If U and V do not share any rows or columns, i.e., R_W is empty, these two matrices can be processed independently from each other. Nevertheless, if U and V over-lap, having a set rows of R_W in common, there is a possibility that a bicluster in U has some rows in V. This situation may cause that part of the bicluster to be a maximal bicluster in V but not in E. To avoid this error, it requires to generate biclusters in V that must contain at least one column from each column set in Z, where Z is the set of all C_V column sets in the current call stack.

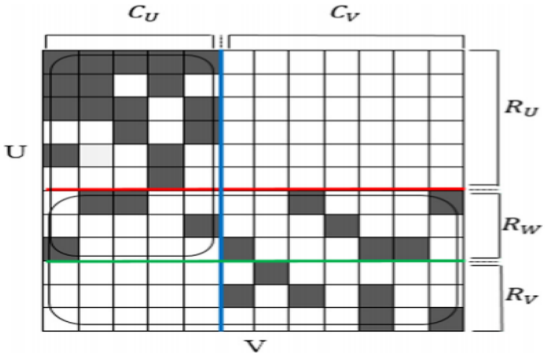

Figure 4 Step 2 and Step 3 of BCBimax algorithm.

Step 5. Save the biggest matrix with only 1s as a bicluster, and delete this bicluster's rows from the data to restart.
Step 6. Continue step 1 to 5 until there is no new bicluster can be found.

3. EXECUTION FLOW

The flowchart of using biclustering method to conduct market segmentation based on customer pain points includes five interrelated parts as depicted in Fig. 5. This denotes the execution of market segmentation using Bcbimax with similarity score. The development of Internet and bigdata drive companies to collaborate with customers in product innovation. For one thing, the spread of Internet allows companies to contact with end-consumers directly and more efficiently at a lower cost [8].

3.1 Data Collection

This is the first step in market segmentation, which is used to collect data from the customer. The companies can capture the customer experience more effectively and less costly through online surveys, online feedback forums, online reputation monitoring, digital experience replay and online focus group. Companies can collect customer pain points directly and easily from virtual brand communities [1, 7].

3.2 Similarity Score Calculation

Let A(I, J) be an n · m matrix of real numbers, where I = {1, 2, . . . , n} is the set of customers and J = {1, 2, . . . ,m} is the set of customer pain points. The element a_{ij} of A(I, J) represents the expression level of customer i under customer pain point j. For a customer subset I□ ⊆ I and customer pain point subset J□ ⊆ J, A(I□, J□) denotes the sub-matrix (bi-cluster) of A(I, J) that contains only the elements a_{ij} satisfying i Є I□ and j Є J□ [5]. The goal is to find a subset of customers that are related to the reference customer. When the reference customer is not known, that can enumerate all customers in the matrix or randomly select a number of customers as the reference customers.

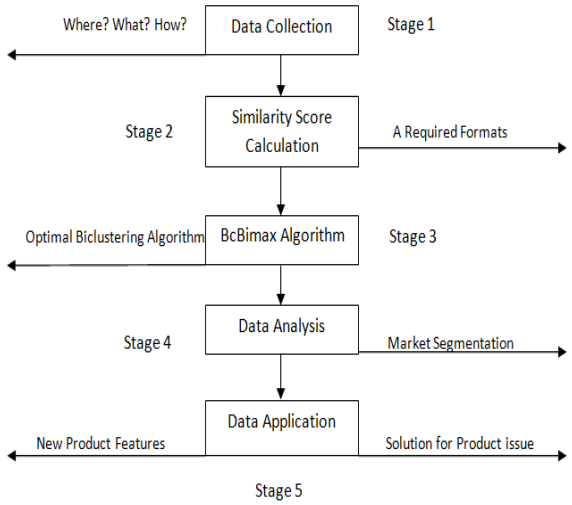

Figure 5 Flowchart in market segmentation.

3.3 Constant Biclusters and Additive Biclusters

Let A(I, J) be an m x n customer matrix and i* Є I a reference customer. A bicluster A(I□, J□) with I□⊆ I and J□⊆ J is a constant bicluster for reference customer i* if for any i Є I□ and any j Є J□, $a_{ij} = a_{i*j}$. A sub-matrix A(I□, J□) with set of rows I□ and set of columns J□ is an additive bicluster for reference customer i* if for any i Є I□ and any j Є J□, $a_{ij} \subseteq a_{i*j} = c_i$, where c_i is a constant for any row i. First, a similarity score to measure the similarity between the reference customer and any other customers can be defined.

Similarity score between customers

For an element a_{ij} of expression matrix A(I, J) and a reference customer i* Є I, define $d_{ij} = | a_{ij} \subseteq a_{i*j} |$. When finding constant biclusters, it is necessary to ignore elements with big d_{ij}. So a threshold is assigned $\alpha \cdot d_{avg}$, where

$$d_{avg} = \frac{\sum i \varepsilon I \& j \varepsilon J d_{ij}}{|I||J|}$$

This is the average distance value of all elements in A(I, J). If $d_{ij} \subseteq \alpha \cdot davg$, it is believed that the two elements a_{ij} and a_{i*j} are not similar and set the similarity s_{ij} to be 0. Otherwise, the similarity score is

$$1 - \frac{d_{ij}}{\alpha.d_{avg}} + \beta$$

where β is the bonus for small d_{ij}. The purpose for using β is to further enlarge the similarity score for small d_{ij} and ignore d_{ij}'s that are greater than the threshold. That is defined as

$$S_{ij} = \begin{cases} 0 & if\, d_{ij} > \alpha.d_{avg} \\ 1 - \frac{d_{ij}}{\alpha.d_{avg}} + \beta & Otherwise. \end{cases} \quad (1)$$

When $d_{ij} \leq a \cdot d_{avg}$, this includes $\frac{d_{ij}}{\alpha.d_{avg}} \leq 1$. Thus, s_{ij} is always greater than or equal to 0. S(I, J) to denote the m x n similarity matrix containing the set of rows I and the set of columns J with every element s_{ij} computed as in (1).

Similarity score for a bicluster

Let S(I, J) be an m x n similarity matrix and S(I□, J□) be a bi-cluster (submatrix) of S(I, J). For row i Є I□, the similarity score of row i in S(I□, J□) is $s(i, J□) = \sum_{j \in J□} s_{ij}$. For column j Є J□, the similarity score of column j in S(I□, J□) is $s(I□, j) = \sum_{i \in I□} s_{ij}$. The similarity score of s(I□, J□) is $s(I□, J□) = \min\{\min_{i \in I□} s(i, J□), \min_{j \in J□} s(I□, j)\}$.Consider a constant bi-cluster S(I□, J□). If the similarity score of row i Є I□ in S(I□, J□) is high, customer i has similar expression values with the reference customer i* under the column subset J□. If the similarity score of column j Є J□ in S(I□, J□) is high, the expression values in column j of all customers in I□ are similar to that of the reference customer i*. Thus, to find a constant bicluster, it is necessary to find a sub-matrix S(I□, J□) with the highest similarity score s(I□, J□).

3.4 BCBimax Algorithm

Biclustering algorithm choosing is an problem dependent. An appropriate algorithm should be selected to match with the characteristics of the dataset. For example, in the research, the poll data of customer pain points will be transformed into a binary matrix by using similarity score, the biclustering algorithm to be selected should be capable of detecting all the subgroups who suffer from only the same pain points. Hence, the BCBimax algorithm is suitable for discovering market segments in the poll data of customer pain points.

3.5 Data Analysis

Once the algorithm is executed, it can be used to perform data analysis. Results will be further evaluated and refined, which is a complicated process. Whether the number of biclusters is too large or too small, identifying the most potential or appealing segment markets and their key common pain points are of great significance. In addition, some small segment markets can be merged into a larger segment market based on their affiliations if necessary.

3.6 Data Application

According to the outputs of Stage 4, companies ought to define the best new product features to be adopted in product design, or figure out "antidotes" to customer's pains in distinguished market segments when updating or improving product.

4. CONCLUSION

This method can assist companies to take full advantage of customer knowledge for product development. Before the biclustering technique being introduced, companies determine the importance of customer pain points in terms of frequency statistics and have not found proper tools to extract the valued customer knowledge hidden inside the data. Because of no discretization there is no loss of data and overlapping problem can also be solved by using similarity score measures. This offers a managerially attractive solution for utilizing the poll data of customer's pain points effectively. In practice, companies can also carry out market segmentation or customize individual products or service to target markets by analyzing other online customer knowledge.

5. ACKNOWLEDGMENT

I respectfully submit all the credit and thanks to great for showering the blessing upon me and give me the necessary wisdom for accomplishing this project. I take immense pleasure to thank my guide Mr.R.Balamurugan, Assistant Professor, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam for his guidance to do this project. I would like to pronounce special thanks to my friends, teaching and non-teaching staff who have directly and indirectly contributed to the success of this project. Last, but not least I take it a great privilege to express my deep sense of gratitude to my beloved parents and relations for their dedications and support.

REFERENCES

[1] Brodie, Ilic, Juric and Hollebeek 2013, 'Consumer engagement in a virtual brand community: an exploratory analysis', Journal of Business Research, vol. 66 , no. 1, pp. 105–114.

[2] Chan, Kwong and Hu 2012, 'Market segmentation and ideal point identification for new product design using fuzzy data compression and fuzzy clustering methods', Applied Soft Computing, Vol. 12, no. 4, pp. 1371–1378.

[3] Homburg and Furst 2007, 'See no evil, hear no evil, speak no evil: a study of defensive organizational behaviour towards customer complaints', Academy of Marketing Science, Vol. 35, no. 4, pp. 523–536.

[4] Kluger, Basri, Chang and Gerstein 2003, 'Spectral biclustering of microarray data: co-clustering genes and conditions', Genome Research, Vol. 13, no. 4, pp. 703–716.

[5] Liu and Wang 2007, 'Computing the maximum similarity bicluster of gene expression data', Bioinformatics, Vol. 23, no. 1, pp. 50-56.

[6] Moorthy 1984, 'Market segmentation, self-selection, and product line design', Marketing Science. Vol. 3, no. 4, pp. 288–307.

[7] Nambisan and Baron 2010, 'Different roles, different strokes: organizing virtual customer environments to promote two types of customer contributions', Organization Science, Vol. 21, no. 2, pp. 554–572.
[8] Sawhney, Verona and Prandelli 2005, 'Collaborating to create: the internet as a platform for customer engagement in product innovation', Journal of Interactive Marketing, Vol. 19, no. 4, pp. 4–17.
[9] Wee-Chung Liew 2012, 'Biclustering analysis for pattern discovery: current techniques, comparative studies and applications', Current Bioinfomatics, Vol. 7, no. 1, pp. 43–55.
[10] Wang, Miao, Zhao, Jin and Chen 2015, 'A biclustering-based method for market segmentation using customer pain points', Engineering Applications of Artificial Intelligence, doi:10.1016/j.engappai.2015.06.005.
[11] Yankelovich and Meer 2006, 'Rediscovering market segmentation', Harvard Business Review, Vol. 84, no. 2, pp. 122–133.
[12] Zhao, Chan, Cheng and Hong 2007, 'A new geometric biclustering algorithm based on the Hough transform for analysis of large-scale microarray data', Journal of Theoretical Biology, Vol. 251, no. 2, pp. 264–274.