International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Optimal Biclustering using Hybrid Swarm Intelligence for Web usage Mining

Kavitha

Abstract*: Web usage mining is used to analyze the user browsing behavior among the web pages which can be further utilized in other applications like recommender system, personalized web pages, providing insight for better business functionality. Since this type of mining does not only depends on the user or web pages, conventional clustering techniques may not suit very well for the analysis. Biclustering techniques are used to discover the subset in the form of submatrices as objects and attributes of objects are considered symmetrically. Finding optimal biclusters is a critical research issue. This research proposes a hybrid swarm intelligence-based method having Particle Swarm Optimization combined with Leader Clustering method along with Uniform Crossover operator. The experimental study shows that the proposed method performs well than traditional biclustering techniques in terms of evaluation metrics.

Index Terms*: Biclustering, Web Usage Mining, Particle Swarm Optimization, Leader Clustering, Hybrid Swarm Intelligence.

I. INTRODUCTION

Cluster examination is one of the obviously used techniques in Machine Learning, Pattern Recognition, Data Mining and Exploratory Data Analysis (EDA). In Pattern Recognition, clustering techniques can be adopted to discover instinctive coalitions in the data. It can then be used to develop representative outlines for objects which could be used for classification [1]. In clustering, data are apportioned into clusters so that, the within cluster similarities are maximized and disparities amongst the clusters are maximized by minimizing an objective function that can reduce the within cluster dissimilarity.

Objects and characteristics of objects are treated in a different way in conventional clustering. However, in some real-world applications, they should be defined in a way such that the exploratory analysis comprises probing data sets for sub matrices that exhibits the exclusive patterns as clusters. Biclustering allows revealing localized groupings within data matrix by expelling the constriction that associated objects must perform likewise throughout the whole set of attributes. It deliberates only a subsection of attributes when observing for resemblance amongst objects. The aim of biclustering is to find the good sub matrices in the dataset, as subsets of objects and subsets of features, where the subset of objects shows noteworthy similarity within the subset of features [2]. They are grouped together to have a high relevance to each other. Unlike traditional clusters based on row-wise or column-wise clustering, biclusters may overlay. Subsections of features and subsets of objects may vary in different biclusters. Biclusters may have some alike objects and features, and few objects might not have its place to any bicluster at all. Because of this flexibility, biclustering became prevalent in many scientific and research fields as a data exploration analysis [3].

Biclustering techniques were first proposed to analyze gene expression data. But, at the present time, it is mainly used in recommender systems. In these systems, data are collected from customers' reviews on product items and utilizes the data to recommend the most interested product items to customers.

Customer clusters that have similar preferences or purchase patterns can be discovered. By using that cluster, recommendations can be made in two directions. One way, the company can commend those products to a set of new consumers who are comparable to the customers in the cluster. The other way of using it is, the company can recommend to consumer about the brand-new products that are more or less similar to those involved in the cluster.

Alike to the above scenario, biclustering delivers a way to discover the users' browsing behavior, and thereby needed web services can be advised to other users with lowest cost. The usage data with respect to web is composed from the web server in terms of log file. The explored and analyzed insight from those files is utilized by the companies and corporations to render the best consumer association by providing them precisely what they require [4]. Moreover, this analysis proposes appreciated information on to improvise the structure and construction of both website and web page to augment the strategies for target marketing which is the base for the recommender system.

Since, generation of the biclusters can be viewed as an optimization-based problem, swarm based heuristic techniques can be considered to solve it. In this study, a swarm intelligence-based Particle Swarm Optimization (PSO) algorithm combined with Leader Clustering (LC) approach is proposed to find the optimal biclusters in Web Usage Mining. Moreover, Uniform Crossover Genetic Operator is embedded to offer good population as particle in each iteration of PSO. Section 2 provides the literature review needed for this study and Section 3 gives an overview of Biclustering analysis.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Kavitha*, School of Computer Science and Applications, REVA University, Bangalore, India.

The proposed approach (PSO-LC) is described in Section 4 and the next section provides the experimental results there by performance evaluation of the proposed approach. The final section concludes with the summary of study and future enhancements that can be made.

II. LITERATURE REVIEW

Web usage mining is a significant research area to examine the user browsing patterns from web usage data. The analysis can be applied to recommender systems. In recommender system, target marketing is the excellent and efficient way to converse with the customers and consumers to merchandise [5].

At the foremost, conventional clustering, a data mining technique has been used for analyzing the web data to form target markets. It is used to cluster the users who have comparable browsing behavior upon various multiple pages of a web site or to cluster the pages based on the alike browsing attention of the users. Here, it is presumed that the entire associated users behave more or less similarly across all set of pages of a website. But, most of the time, the scenario is different where the users of web perform similarly only on a certain subsection of pages and their browsing behavior is not associated on the remaining set of conditions. Therefore, conventional clustering techniques will not recognize those kinds of users' groups. To address the challenges of conventional clustering, the biclustering method [5] was presented to analyze web usage data. Biclustering analysis is a different and very effective methodology to un reveal the local comprehensible patterns hidden in a data matrix. Those biclusters expose the browsing outlines of associated users along with connected set of pages of a website.

Biclustering was initially used for gene expression analysis. Biclustering problem has been proposed as an optimization problem, defining a score for each possible bicluster and heuristics to solve the constrained optimization problem defined by that score function [6,7]. Biclustering based on fuzzy approach is experimented. The web data elements are allocated to different biclusters with varied membership levels [8]. The couple two-way clustering which uses the hierarchical approach for clustering and it is applied separately to each dimension and they describe the process to combine both results [9]. But the time complexity of this method is Exponential.

Biclustering is an exciting computational problem. Hence, it is considered as one of the NP hard problem [10]. This has inspired the exploration towards optimal and stochastic algorithms such as evolutionary computation techniques and heuristic search techniques [11-13]. The multi objective methods are proposed because, the biclustering problem has some objectives exist, which are in conflict with each other [2]. Also, it is difficult to uncover biclusters since; there will be overlapping patterns between biclusters [14]. They can have a complicated coherent pattern. The types and structure of the biclusters plays a crucial role in the evaluation of those biiclusters [15,16].

III. BICLUSTERING

The objects are clustered according to the attribute values in conventional cluster analysis. But, nowadays, in most of the applications, objects and characteristics are defined in a symmetric way, where the analysis involves in finding the data matrices for submatrices that shows exclusive patterns as clusters. This type of clustering technique is called biclustering.

Biclustering techniques were proposed to overcome the challenges that exists in analyzing the gene expression data. Biclustering is useful not only in bioinformatics, but also in other applications as well [3].

Let a dataset has 'M' number of samples and 'N' number of features as a data matrix D. Usually, a bicluster is a subset of rows that shows resembling behaviors across a subset of columns and subset of columns which is resembling towards a subset of rows. The bicluster BC = (X, Y), appears as a sub matrix of 'D' with around similar patterns, where X = {M1, …, Mx} ⊆ R and Y = {N1, …, Ny} ⊆ C are separate subsets of Rows and Columns, respectively [6].

The lower the mean squared residue, the stronger the coherence exhibited by the bicluster and the quality of the bicluster is high. Average Correlation Value (ACV) [5] is the commonly used evaluation measure for a bicluster. It evaluates the correlation homogeneity of a bicluster. A high ACV indicates the high similarities among the rows or columns. ACV can be used as a metric function to find correlated/coherent biclusters.

Madeira and Oliveira discussed about the different types of biclusters which are normally presumed as (i) Bicluster with constant values, (ii and iii) Bicluster with constant values in rows or columns, (iv and v) Bicluster with coherent values including additive or multiplicative models [7].

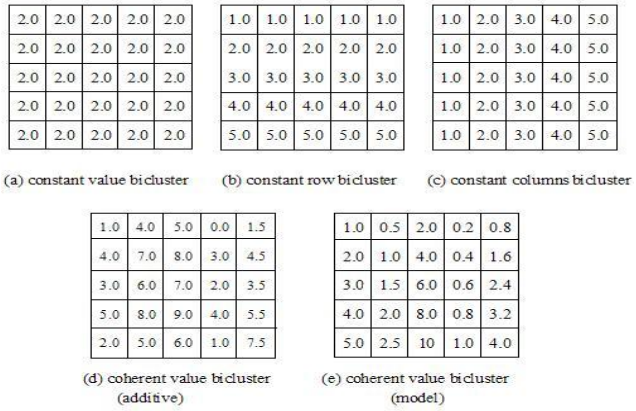

Fig. 1. Types of Biclusters

Since, biclusters are varied from conventional clusters, different coherence measures are there to check the quality of a bicluster. Some of them are Variance, Mean Squared Residue (MSR), Scaling Mean Squared Residue (SMSR), Relevance Index (RI), Average Correlation Value (ACV).

IV. PROPOSED PSO-LC METHOD

Biclustering is considered as an optimization task. The objective function that can be used here is of generating set of biclusters having high Average Correlation Value. As Particle Swarm Optimization is one of the best swarm intelligence based optimization algorithms, it is used in the proposed method along with Leader Clustering algorithm.

The web log data has to be pre-processed and transformed to a format which is suitable for the analysis. The quality of the final biclusters directly associates on the quality of the input that has been given to PSO. Hence, initial biclusters are generated using modified Coupled Two-Way Clustering (CTWC).

By taking the rows of users and columns of page category, the Leader Algorithm is applied and clusters are generated. The combinations of clusters which are having less Mean Squared Residue alone has been taken as the Initial BiClusters. It is encoded into binary form to make available for PSO.

PSO algorithm is applied by taking Initial BiClusters as particles. As it is an iterative approach, each time, the personal best position is calculated based on the ACV. At the end of each iteration, Uniform Crossover from genetic algorithm is applied to get a better population for the next iteration. Accordingly, the global best positions are updated.

// Transform the User Click Stream Access Data

1. $UserAccessMatrix\ UAM = (U.P)_{n\ x\ m}$

Where U = 'n' set of Users,
P = 'm' set of Pages

$$UAM_{ij} = \begin{cases} Hits\ (U_i, P_j) & if\ U_i\ visited\ P_j \\ 0 & Otherwise \end{cases}$$

// Initial Biclusters

2. $UserCluster\ U_c = Leader(U)$

3. $PageCluster\ P_c = Leader(P)$

Let $N_u = Number\ of\ User\ Clusters$

$N_p = Number\ of\ Page\ Clusters$

$$V_{i,j=1}^{N_u\ x\ N_p} MSR =$$

4. $MeanSquaredResidue\ (U_c^i, P_c^j)$

$MSR_{Avg} = Average\ (MSR)$

5. $InitialBiCluster\ IBC = \{\}$

6. $for\ i,j \leftarrow 1\ to\ Nu\ x\ Np$

$If\ MSR(U_c^i, P_c^j) < MSR\_Avg$

$IBC = IBC\ U\ \{U_c^i, P_c^j\}$

// Biclustering

Initial Particles = IBC

Initialize Velocity and Position, Personal Best pBest of each particle 'p'

7. $while\ termination\ condition\ not\ satisfied\ do$

$for\ each\ particle\ p\ do$

$CurrentFitness\ CF = ACV\ (p)$

$if\ CF > fitness\ (pBest(p))$

$Update\ pBest\ with\ current\ position$

$Update\ gBest, Velocity\ and\ Position$

Where $gBest$ is the Global Best

$Apply\ UniformCrossover\ (p)$

$return\ gBest\ as\ the\ Global\ Optimal\ Bicluster$

V. PERFORMANCE EVALUATION

The performance of the proposed method PSO-LC is evaluated by implementing the algorithm on the benchmark clickstream dataset 'MSNBC'. There are 989,818 users and 17 web page categories. The pages are observed as URL categories rather than as the individual web page. Hence, the dimensionality of the data is reduced. The page category is counted and it is stored for a user in a session. The longer records with less frequent visit of each page category is discarded from the analysis. The proposed PSO-LC method is implemented using Python. To check the performance of the proposed method, it is compared with a conventional Coupled Two-Way Clustering method. The following table shows the comparison of those two methods in terms of ACV. It clearly shows that the proposed PSO-LC method outperforms well than CTWC. There is an excellence in the performance because the quality of input has been improvised using Leader algorithm. Also, the solution has been envisaged using Crossover operator.

TABLE 1. Comparison of ACV

No. of Iterations	CTWC	Proposed PSO-LC Method
10	0.5291	0.5418
25	0.6370	0.7049
50	0.7644	0.8437
75	0.8292	0.8950
100	0.8316	0.8974

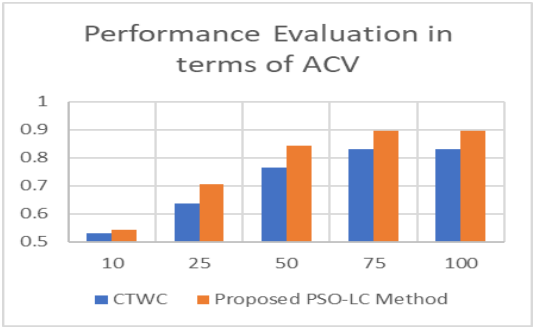

Fig 1. Performance Comparison in terms of ACV

The Table 2 shows the various evaluation metrics exclusively available for biclustering. It is apparently evident that the proposed method performs efficiently since there is high value for ACV and Relevance Index, minimal value for Variance, and other squared residue error measures.

TABLE 2. Evaluation measures for the proposed method

Evaluation Metric	Measure Value
ACV	0.8437
Variance	0.1639
Mean Squared Residue	0.1921
Scaling Mean Squared Residue	0.0483
Relevance Index	0.9207

Table 3 shows the optimal web page categories for the given user click stream data. A subset of coherent users had improved and associated browsing notice for the identified subset of pages.

TABLE 3. Optimal Web Page Categories

Id	**Category**
2	News
3	Tech
4	Local
11	Business
12	Sports

Hence, it will be now effective to provide more focus and target towards similar users. Thus the crucial association between users and web pages is efficiently achieved by means of the proposed method.

VI. CONCLUSION

Generating optimal biclusters is one of the significant research challenges in many domains such as recommender systems, gene expression analysis, etc., A hybrid method PSO-LC has been proposed by embedding Crossover operator in Particle Swarm Optimization. Moreover, the input data is improvised for quality by applying Leader Clustering algorithm rather than applying a conventional clustering method. The performance of the proposed method is evaluated in comparison with the traditional Coupled Two-way Clustering method in terms of ACV measure. The experimental study shows that the proposed method performs well. In future, any other meta heuristic will be embedded to check for the increase in efficiency.

REFERENCES

1. Xu, R. and Wunsch, D.I., Survey of clustering algorithms, IEEE Transactions on Neural Networks, Vol. 16, No. 3, (2005) pp.645–678.
2. Federico Divina, Jesus S. Aguilar-Ruiz, A Multi-Objective Approach to Discover Biclusters in Microarray Data, ACM GECCO „07, pp. 385-392
3. Busygin S, Prokopyev O and Pardalos PM. Biclustering in Data Mining, Computers & Operation Research (2008) pp. 2964-2987.
4. Jaideep Srivastava, Robert Cooley, Mukund Despande and Pnag-Ning Tan, Web Usage Mining: Discovery and Applications of Usage from Web Data, ACM SIGKDD, (2000), Volume 1, Issue 2, pp. 12-23
5. Rathipriya, Thangavel and Bagyamani, Binary Particle Swarm Optimization based Biclustering of Web usage data, International Journal of Computer Applications, Vol. 25, No. 2, (2011), pp. 43-49.
6. Hongya Zhao, Alan Wee-Chung Liew, Doris Z. Wang, and Hong Yan, Biclustering Analysis for Pattern Discovery: Current Techniques, Comparative Studies and Applications.
7. Madeira SC and Oliveira AL. Biclustering Algorithms for Biological Data Analysis: a Survey. IEEE/ACM Trans. Computational Biology & Bioinformatics (2004) Vol. 1 No. 1, pp.24-45.
8. Koutsonikola, V.A. and Vakali, A, A fuzzy bi-clustering approach to correlate web users and pages, International Journal of Knowledge and Web Intelligence, vol. 1, no. 1/2, (2009) pp.3–23.
9. Rajesh, M. & Gnanasekar, J.M. Wireless Pers Commun (2017) 97: 1267. https://doi.org/10.1007/s11277-017-4565-9
10. Busygin S, Prokopyev O and Pardalos PM. Biclustering in Data Mining. Computers & Operation Research (2008) Vol. 35, pp.2964-2987.
11. Bleuler S, Prelić A and Zitzler E. An EA framework for biclustering of gene expression data.Proceedings of Congress on Evolutionary Computation (2004), pp. 166-173.
12. Divina F and Ruiz JA. Biclustering of expression data with evolutionary computation. IEEE Trans. Knowledge & Data Engineering (2006) Vol. 18, pp. 590-602.
13. Mitra S and Banka H. Multi-objective evolutionary biclustering of gene expression data. Pattern Recognition (2006), Vol. 39 (12) pp. 2464-2477.
14. Cheng KO, Law NF, Siu WC and Liew AWC. Identification of Coherent Patterns in Gene Expression Data Using an Efficient Biclustering Algorithm and Parallel Coordinate Visualization. BMC Bioinformatics (2008) ; 9.
15. Prelic A, Bleuler S and Zimmermann P, A Systematic Comparison and Evaluation of Biclustering Methods for Gene Expression Data. Bioinformatics (2006) Vol. 22, pp.1122–1129.
16. Lee Y, Lee JH and Jun CH. Validation Measures of Bicluster Solutions. Industrial Engineering & Management Systems (2009), Vol. 8, pp. 101-108.

AUTHORS PROFILE

Dr. Kavitha has done her doctorate in Computer Science. She has published more than 30 papers in her research area of interest in Data Mining, Swarm Intelligence. She is a member in GSTF, IDES, UACEE, IACSIT and AIRCC. She is reviewer for Journals and Technical Programme Committee Member for many international conferences. She has given key note speech in various conferences, seminars and workshops. Her research interest includes Data Mining, Data Analytics and Swarm Intelligence.