Integrated Intelligent Research (IIR)
International Journal of Data Mining Techniques and Applications
Volume: 06 Issue: 01 June 2017 Page No.12-16
ISSN: 2278-2419

Grey Wolf Optimizer Based Web usage Data Clustering with Enhanced Fuzzy C Means Algorithm

P. Selvaraju[1], B.Kalaavathi[2]
[1]AP/IT, Department of Information Technology, SSM College of Engineering, Namakkal, Tamilnadu, India
[2]Prof & Head, Department of Computer Science and Engineering, K.S.R. Institute for Engineering and Technology, Tiruchengode, Tamilnadu, India
Email: rpselvaraju@gmail.com

Abstract - Recommendation system plays a major role in web mining and it is applied to many applications such as e-commerce, e-government and e-library. The key challenges of recommendation system is to recommend the users based on their interest among more visitors and huge information. To make this challenge effective, there is a need for clustering algorithm to handle the data. Hence, this research focused on designing effective clustering algorithm to apply it in e-commerce applications. The grey wolf optimization based clustering is proposed to make an efficient clustering method for grouping the users based on their interest. To find the effective clustering, proposed a grey wolf optimization based fuzzy clustering algorithm, and made a comparison on Fuzzy C Means (FCM) based Genetic Algorithm (GA), Entropy based FCM and Improved Genetic FCM (FCM-GA). The experimental results proves that it performs better than traditional algorithms, at the same time the quality is improved.
Keywords - Recommender systems, E-commerce, Fuzzy C Means, Grey Wolf optimization, Fuzzy Clustering.

I. INTRODUCTION

Recent development of World Wide Web (WWW) leads to interface many information's on web and provides web based services such as support, service, shopping sites, etc. Many web based researches were made based on knowledge discovery represented in (Etzioni 1996) [1]. Generally, web mining may classified in to two types, such as web content mining and web usage mining. The web usage mining is a process of clustering the web users based on their common properties. It helps web masters to provide the more suitable and customized services to users in-terms of analyzing the characteristics of groups. In some applications, the recommendation processed by their reviews and purchase history. For further classification, it is necessary to consider some attributes, demographic data, items viewed and subjects interest. In online shopping, the recommendation algorithm is used to personalize the online store with each user interest. It helps to change the websites periodically based on the customer interests. In this research, the proposed method is elaborated in detail. Based on the several challenges represented below, the proposed E-Commerce recommendation is initiated.

- Normally, a retailer having a huge amount of data, millions of customer and several catalog items
- The information content of new user is very less and limited because of few purchases or due to product ratings
- The interaction between the algorithm and customer information must be fast, if it delays the data's will be erased because the customer data is volatile
- It works under the old customers who are all having thousands of purchases and ratings
- The recommendation process must be in real time basis to achieve best recommendation system.

Fu *et al*., (1999) [2] discussed the web users based on the access patterns. The clustering of web users are made based on their browsing history access patterns on web. Linden *et al* (2013) [3] presented an industry report about the recommendations in online shopping. They made a recommendation algorithm by finding set of customers rating and user's purchased history. The algorithm uses two types of versions namely, collaborative filtering and cluster models. The algorithm represents the recommendations based on the similarity between two customers namely A and B respectively. They represented a common way in vector format indicated in equation 1.1. They also stated collaborative filtering is computationally expensive.

$$Similarity\ (\vec{AB}) = \cos(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \cdot \|\vec{B}\|} \qquad (1.1)$$

Sarwar *et al* (2001) [4] analyzed different item-based recommendation generation. They used weighted sum versus regression model to obtain the recommendations from item-item correlation techniques and cosine similarities between each item vectors. Finally, traditional method are compared with the basic k-nearest neighbor approach and proved it is better than user based algorithms. Rakibe *et al* (2013) presented a K-Means based online market analysis. To manage the customer data as per requirement, the market segmentation is used. Market segmentation includes customer retention strategies, allocation of resources to advertise and segmenting the product price and identify the potential customers. Clustering divides the records presented in the data base, i.e., dividing the dataset into several subclasses. The remaining part of this paper is organized as follows, in section 2 the survey is made about the clustering in detail, it also tabulated similar clustering methods. In section 3, the problem identification is mentioned. In section 4, methodology describes about Fuzzy C-Means based Genetic Algorithm, entropy based FCM, improved genetic FCM and proposed grey wolf optimizer based clustering. Finally experimental results were made in section 5 and summarized in section 6.

II. LITERATURE SURVEY

Büchner and Mulvenna (1998) [6] described the combination of data mining techniques with internet data, in order to make correct action in electronic commerce. The data which is considered here is server data, web meta information, marketing data and marketing knowledge. Yuan and Cheng

International Journal of Data Mining Techniques and Applications
Volume: 06 Issue: 01 June 2017 Page No.12-16
ISSN: 2278-2419

(2004) [7] presented clustering based on the heterogeneous product recommendation in mobile marketing. Evolution of the contemporary similarity matrix is analyzed with proposed design named as Ontology Based Personalized Coupled Clustering (OBPC). The Recommender systems for automatic product recommendation acquire customer's preference and products. The products available in online varies with respect to the one to one basis with the factor price. Generally recommender system consists of content based and collaborative filtering (Sarwar *et al* 2000), [8-10] (Lawrence *et al* 2001). The content based recommendations made by matching customer interests with product attributes. But product recommendations in collaborative system overlap the preference ratings among customers.

Table I: Comparison of Related Clustering Work

Sl. No	Citation	Methodology	Function
1	Zhang and Zhou (2005) [13]	Latent Usage Information (LUI) Model for Clustering Web Transaction using Modified K-Means.	Clustering and Building User Profile
2	Gupta and Shrivastava (2012) [18]	Improved Genetic Fuzzy C-Means Algorithm and Entropy based FCM	Used to cluster the user sessions and overcomes the defects presents in FCM.
3	Su *et al* (2010) [19]	New Fuzzy Clustering Algorithm	Used to initialize cluster centers and to reduce the dependence on the initial cluster centers
4	Vellingiri and Pandian (2011).	Fuzzy possibilistic C-Means algorithm	Design to predict the user behavior
5	Zhu *et al* (2009)	Generalized Fuzzy C-Means Clustering Algorithm	To test the Clustering capability

Kim and Ahn (2008) [11] presented a K-Means clustering algorithm for system in online shopping market. It is stated as novel clustering algorithm with the genetic algorithms to segment effectively in online shopping market. The NP complete global optimization process is rectified by genetic algorithm. They compared K-Means and self-organizing maps (SOM) Kohonen and Somervuo (1998) [12]. Latent Usage Information Algorithm is discussed by (Zhang 2005) [13]. The various online user data clustering is reviewed and indicated in the table 1.1. Lang *et al* (2005) [14-15] search behavior of online shoppers, its main contribution is to track Search Behavior of user. Ahmeda *et al* (2015) discussed about consumer online shopping based on classification algorithm.

III. PROBLEM DEFINITION

From the survey it is noticed that the current clustering techniques do not address all the requirements concurrently. For example, real time problems occurs in e-commerce, e-government, etc. faces lot of problems in business wise requirements. GA and fuzzy clustering methods suffers from certain drawbacks due to the restriction such that the sum of membership values of a data points varies, hence it faces more issues in iterations. To manage all the data with large number of dimensions and large number of items becomes a problematic, due to time complexity. So to avoid time complexity and minimizes the error, the best algorithm is to be made for clustering.

IV. METHODOLOGY

A). The Fuzzy C-Means clustering (FCM)

Fuzzy C-Means clustering (FCM) is simple and most popular fuzzy clustering algorithm, its steps are explained in figure 1 Bezdek *et al* (1984) [22] described algorithm is considered. The FCM algorithm uses a set of unlabeled feature vectors and classifies them into C classes, where *C* is given by the user. It assumes that the number of clusters, is known a priori, and minimizes

$$J_{fcm} = \sum_{i=1}^{C}\sum_{k=1}^{N} u_{ik}^{m} d_{ik}^{2}, \qquad d_{ik} = \left\| x_k - v_i \right\| \tag{4.1}$$

with the constraint of:

$$\sum_{i=1}^{C} u_{ik} = 1 \qquad ; k = 1, \cdots, N$$

Here, $m > 1$ is known as the fuzzifier parameter.

The algorithm provides the fuzzy membership matrix 'u' and the fuzzy cluster center matrix 'v'. To apply C-Means algorithm the following criteria should consider.

Step 1: Initialize clusters by choosing 'C' data points
Step2: Find the nearest cluster center that is closest to each data point.
Step 3: Assign the data point to the corresponding cluster.
Step 4: Update the cluster centers in each cluster using the mean of the data points which are assigned to that cluster.
Step 5: Terminate if no changes else go to step 2.

In FCM two main deficiencies were associated, it is inability to distinguish outliers from non-outliers by weighting the memberships and attraction of the centroids towards the outliers.

B). Entropy based FCM (EFCM)

EFCM made by Vellingiri and Pandian (2011), they used information entropy to initialize the cluster centers to determine the number of cluster centers. The main of this method is to reduce some errors, and also improve the algorithm that introduce a weighting parameters, its steps are discussed in figure 2. After assigning the combination process made with the merger of ideas, and divide the large chumps into small clusters. Then merge various small clusters according to the merger of the conditions, finally the irregular datasets clustering were solved.

C). Improved Genetic FCM Algorithm

The fuzzy C-Means clustering is considered as input to this algorithm that follows several steps based on serial number coding by following the algorithm which is described by [23] Bandyopadhyay et al (2007) and it is shown in figure 3.

Step 1: In the algorithm 'T' is the input data with 'N' data points each of which has 'M' dimensions.
Step2: Calculate entropy for each X_i in T for i = 1. . . N.
Step 3: Choose X_{iMin} with least entropy.

International Journal of Data Mining Techniques and Applications
Volume: 06 Issue: 01 June 2017 Page No.12-16
ISSN: 2278-2419

Step 4: Remove X_{iMin} and the data points having similarity with X_{iMin} greater than p from T.
Step 5: If T is not empty go to Step 2.
Initially a random point is selected from each cluster to build a chromosome. For each chromosome the fitness value is calculated by using new fitness function as shown in figure 3. Based on this, genetic operators such as reproduction, crossover and mutation are applied to produce a new population

Step 1: Set the number of clusters c, population size 'N' evolutionary algebra 'T', crossover probability P_c and mutation probability P_m.
Step 2: Set the serial numbers to the individuals of cluster object.
Step 3: Randomly generate 'cxn' serial number and make chromosome with c serial number.
Step 4: calculate the fitness function with formula,

$$f = \frac{1}{1+J_m(U,P)}$$

Step 5: find smallest fitness value, judge the evolution if satisfied terminate else go to step 3.
Step 6: Decode the individual task and find cluster centers and get the final clusters results with FCM.
.

D). Proposed Grey Wolf Optimizer based Clustering

Grey Wolf Optimizer is a new Meta heuristic inspired by grey wolves, it is based on leadership hierarchy and the hunting mechanism of the grey wolves. The leadership hierarchy is proposed based on four types of wolves' like alpha, beta, delta and omega. In GWO algorithm, the hunting is guided by α, β and δ. The ω solutions follow these three wolves. During hunting, the grey wolves encircle prey. Mirjalili *et al* (2014) [24] work is considered for clustering with challenging problems. The mathematical model of the encircling behavior is given below.

$$D = |C.X_p(t) - A.X(t)|$$
$$X(t+1) = X_p(t) - A.D$$

where,'t' is the current iteration, A and C are coefficient vectors, X_p is the position vector of the prey, and X indicates the position vector of a grey wolf. The basic steps for grey wolf algorithm is shown in figure 4.

Step 1: Initialize the grey wolf population X_i (i = 1, 2... n)
Step 2: Initialize a, A, and C
Step 3: Calculate the fitness of each search agent
X_α=the best search agent
X_β=the second best search agent
X_δ=the third best search agent
Step 4: while (t < Max number of iterations)
For each search agent update the position of the current search agent.
Step 5: Update α, A and C, and Calculate the fitness value for all search
Step 6: Update $X_\alpha, X_\beta, X_\delta, make\ t = t+1;$

Grey wolf optimizer is selected for effectively solving the web usage data clustering, it is proposed for clustering web users or organizing objects into groups. The advantages of the proposed social hierarchy is, it assists GWO to save the best solutions (i.e.) retains the similar users in the same cluster or group recurrently. It also allows candidate solutions to locate the exact position of groups. The proposed flow diagram representation shown in figure 5, it explains the diagrammatic representation of proposed clustering.

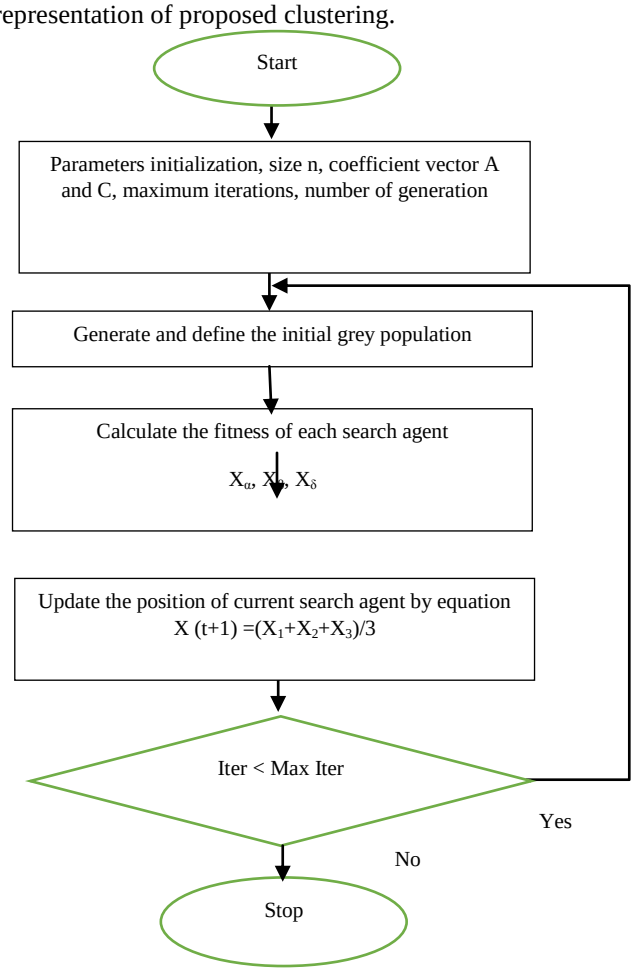

Figure 1: Flow chart representation of proposed clustering

V. EXPERIMENTAL RESULTS

This research considered real time data for analysis, the raw web logs between the date 20 September 2016 to 29 November 2016. The proposed grey wolf optimization based clustering algorithm applied to the different groups, to obtain the number of optimum clusters.

Table II: Experimental Results

Methods	Error Rate (%)	Threshold Value
Fuzzy C-Means clustering	4.9	0.2
	5.2	0.4
	5.6	0.6
Entropy based FCM (EFCM)	4.6	0.2
	4.8	0.4
	4.11	0.6
Improved Genetic FCM Algorithm	3.50	0.2
	3.90	0.4
	4.09	0.6
Proposed Grey Wolf Clustering	2.90	0.2
	3.20	0.4
	3.67	0.6

The proposed work compared with the Fuzzy C-Means clustering (FCM), Improved Genetic FCM algorithm and

International Journal of Data Mining Techniques and Applications
Volume: 06 Issue: 01 June 2017 Page No.12-16
ISSN: 2278-2419

Entropy based FCM (EFCM) based on sum of error and threshold. The sum of error is defined as $SE = \sum_i^c (X_i - Y_i)$, Where, C is the number of clusters, (i.e) in this problem is equal to number of classes, X_i is the number of instances that belong to cluster 'i'. Y_i is the numbers of instances from cluster 'i' that belong to a same class and also have maximum presence in cluster 'i'. From the table 1, it is observed that the proposed clustering has minimum error rate of 3.67 for maximum threshold value for 0.6.

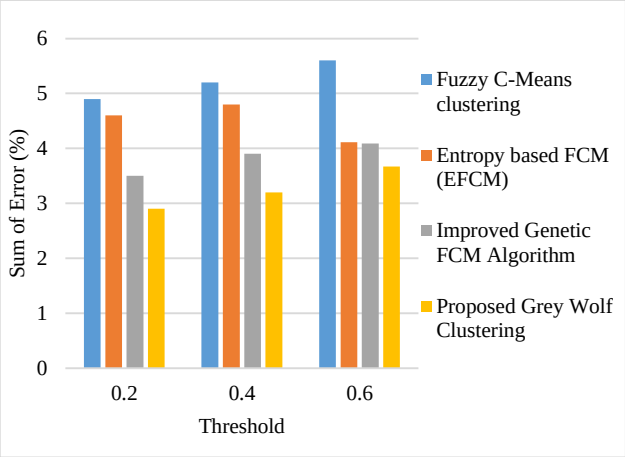

Figure 2: Comparison graph for sum of error and threshold

From the experimental results the proposed clustering method having error rate in least as shown in figure 6, if the threshold limit varies up to 0.6, the error rate increased to unit count. It is effective for large user data and also applicable for real time problems.

VI. CONCLUSION

The proposed Clustering model reduces the error rate while clustering the users based on their interests. The experimental results are confirmed that global distribution characteristics of clustering center found during the process of grey wolf optimizer based fuzzy clustering analysis. It determined the intrinsic grouping in a set of unlabeled data and the cluster effect is rational as well as conceptual clustering. It deals with large number of dimensions and large number of users in online to purchase or to review, hence the results were made with real time data and shows the effective complexity reduction. It gives detail separation of interested and non-interested users. In future the classification process is made for same process to find and made a new change in real time problems in e-commerce.

References

[1] Fu, Y., Sandhu, K., & Shih, M. Y. (1999, August). Clustering of web users based on access patterns. In Proceedings of the 1999 KDD Workshop on Web Mining. San Diego, CA. Springer-Verlag.

[2] Etzioni, O. (1996). The World-Wide Web: quagmire or gold mine. Communications of the ACM, 39(11), 65-68.

[3] Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet computing, 7(1), 76-80.

[4] Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001, April). Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th international conference on World Wide Web (pp. 285-295). ACM.

[5] Rakibe, M. P. L., Kalvadekar, P. N., & SRES, K. Web User Analysis Using Hierarchical and Optimized K-mean Algorithm for Online Market Analysis.

[6] Büchner, A. G., & Mulvenna, M. D. (1998). Discovering internet marketing intelligence through online analytical web usage mining. ACM Sigmod Record, 27(4), 54-61.

[7] Yuan, S. T., & Cheng, C. (2004). Ontology-based personalized couple clustering for heterogeneous product recommendation in mobile marketing. Expert systems with applications, 26(4), 461-476.

[8] Lawrence, R. D., Almasi, G. S., Kotlyar, V., Viveros, M., & Duri, S. S. (2001). Personalization of supermarket product recommendations. In Applications of Data Mining to Electronic Commerce (pp. 11-32). Springer US.

[9] Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2000, October). Analysis of recommendation algorithms for e-commerce. In Proceedings of the 2nd ACM conference on Electronic commerce (pp. 158-167). ACM.

[10] Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2000). Application of dimensionality reduction in recommender system-a case study (No. TR-00-043). Minnesota Univ Minneapolis Dept of Computer Science.

[11] Kim, K. J., & Ahn, H. (2008). A recommender system using GA K-Means clustering in an online shopping market. Expert systems with applications, 34(2), 1200-1209.

[12] Kohonen, T., & Somervuo, P. (1998). Self-organizing maps of symbol strings. Neurocomputing, 21(1), 19-30.

[13] Zhang, Y., Xu, G., & Zhou, X. (2005, July). A latent usage approach for clustering web transaction and building user profile. In International Conference on Advanced Data Mining and Applications (pp. 31-42). Springer Berlin Heidelberg.

[14] Kumar, N., Lang, K. R., & Peng, Q. (2005, January). Consumer search behavior in online shopping environments. In Proceedings of the 38th Annual Hawaii International Conference on System Sciences (pp. 175b-175b). IEEE.

[15] Lang, K. R., Peng, Q., & Kumar, N. Consumer Search Behavior in Online Shopping Environments.

[16] Ahmeda, R. A. E. D., Shehaba, M. E., Morsya, S., & Mekawiea, N. (2015, April). Performance study of classification algorithms for consumer online shopping attitudes and behavior using data mining. In Communication Systems and Network Technologies (CSNT), 2015 Fifth International Conference on (pp. 1344-1349). IEEE.

[17] Anand, M., Khan, Z., & Shukla, R. S. (2013). Customer relationship management using adaptive resonance theory. International Journal of Computer Applications, 76(6).

[18] Gupta, K., & Shrivastava, M. (2012). Web usage data clustering using improved genetic fuzzy C-Means algorithm. Int. J. Adv. Comput. Res, 2, 77-79.

[19] Su, X., WANG, X., Wang, Z., & Xiao, Y. (2010). A new fuzzy clustering algorithm based on entropy weighting. Journal of Computational Information Systems, 6(10), 3319-3326.

[20] Vellingiri, J., & Pandian, S. C. (2011). Fuzzy possibilistic C-Means algorithm for clustering on web usage mining to predict the user behavior. European Journal of Scientific Research, 58(2), 222-230.

International Journal of Data Mining Techniques and Applications
Volume: 06 Issue: 01 June 2017 Page No.12-16
ISSN: 2278-2419

[21] Zhu, L., Chung, F. L., & Wang, S. (2009). Generalized fuzzy C-Means clustering algorithm with improved fuzzy partitions. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39(3), 578-591.
[22] Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2-3), 191-203.
[23] Bandyopadhyay, S., Mukhopadhyay, A., & Maulik, U. (2007). An improved algorithm for clustering gene expression data. Bioinformatics, 23(21), 2859-2865.
[24] Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. Advances in Engineering Software, 69, 46-61.