An Evolutionary K-means Clustering Approach to Recommender Systems

Kanupriya Bhao, Prof. Dharminder Kumar, Dr. Saroj
Dept. of Computer Science & Engineering
Guru Jambheshwar University of Science & Technology
Hisar, India

***Abstract*—The main strengths of k-means clustering, the most widely used clustering technique for recommender systems, is its simplicity and ease of applicability to practical problems. However, k-means clustering suffers from the drawbacks of falling in local optima and the quality of clusters is largely dependent on the initial cluster centers. An important contribution of this paper is a hybrid k–means clustering approach to recommender systems that combines 'outside the box' recommendation ability of collaborative filtering with k-means clustering and Genetic Algorithms. In this approach, genetic algorithm operators are used to pick up appropriate initial seeds for k-means clustering. This helps in improving cluster quality, thereby suggesting a new approach to recommender systems.**

The model focuses on identifying a set of users with similar liking for movies and accordingly making recommendations.

Keywords—Recommender Systems, Collaborative filtering, k-means clustering, Genetic algorithms

I. INTRODUCTION

The past one decade has observed a staggering growth in the amount of information available in the form of books, movies, news, advertisements and online information. But browsing through this enormous data pool to get the desired information is not easy. We need technology to help us wade through to find the items we really want and need [14]. Therefore, recommender systems have come to our rescue.

Recommender Systems apply knowledge discovery techniques to the problem of making personalized recommendations for information, products or services [13]. They recommend items to users based on their preferences. These systems are becoming very successful these days because of the numerous applications where they can be applied. To quote a few, they can help a user in searching for music from an online-music website, text documents on a search engine like Google, books on Amazon.com etc. Recommender systems can also help in increasing the sales of a business on an e-commerce website.

Various techniques have been proposed in the past but clustering techniques are most popular for making recommendations to the users. In addition, K-means clustering is the most frequently used recommender system technique because of its simplicity[15]. However k-means clustering suffers from problems like it can fall in local optima and the quality of clusters is largely dependent on the initial cluster centres. A lot of research has been done to overcome these problems and studies have showed that if k-means clustering is integrated with global search techniques like genetic algorithms, the quality of clusters generated can be improved[2], [5], [7], [9], [11], [12].

In this paper, k-means clustering has been used to recommend items to users. Also, a comparative study has been made to show how genetic algorithms can be used to improve its performance by picking up the initial seeds of k-means clustering. Data has been provided to validate the usefulness of using Genetic Algorithms with k-means clustering and improvement in the quality of clusters thus obtained.

The rest of the paper is organized as follows: the next section describes recommender systems along with k-means clustering and Genetic algorithms. Section III proposes the improvement in the performance of clusters of k-means clustering by using genetic algorithm approach followed by Section IV which describes the data and the experiments. In the final section, the conclusions and limitations of this study are presented.

II. BACKGROUND

A. Recommender Systems

Recommender systems recommend items to users based on user's needs and preferences. They are implemented in commercial and non-profit web sites to predict user preferences to draw user's attention and increase his satisfaction towards online information search results. They help users locate the items of their interest more quickly and can generate more accurate predictions which can lead to higher sales on an e-commerce website.

Recommender systems are classified into three categories on the basis of how the recommendations are made:

1. Content Based Recommender Systems
2. Collaborative Filtering Recommender Systems
3. Hybrid Recommender Systems

Content-based Recommender Systems recommend items to users based on correlation between the content of items and the user preferences [14]. In these systems, the user is recommended items similar to the items the user preferred in

the past [3]. Collaborative filtering recommender systems recommend items to users which people with similar tastes and preferences liked in the past [1]. The third category of recommender systems combines collaborative filtering and content based approaches [1].

Recommender systems often face a common issue of the user being limited to getting recommendationsfor items that are similar to those already rated. This problem is referred to overspecialization[1]. K-means hybrid approach proposed in this paper resolves this problem by identifying the users that have similar tastes as the target user. The target user is then recommended the items which have been liked by these similar users. This helps in adding some randomness.

B. Clustering Algorithms

K-means clustering is a method of cluster-analysis in which the initial k centroids (points in space that represent the centre of the cluster) are picked up randomly and each item is assigned to the cluster with the closest centroids. The number of clusters k is predetermined. Kim and Ahn highlighted that even though k-means clustering is simple and easy to use algorithm yet it suffers from some drawbacks [5]. The key limitation of k-means algorithm is that since the initial centroids are picked up randomly, so an inappropriate choice of clusters may yield poor results. Genetic algorithms because of their effectiveness in NP-complete global optimization problems can suggest a better solution to k-means clusters.

C. Genetic Algorithms

Genetic Algorithms (GAs) are stochastic search methods based on the mechanism of natural evolution and genetic inheritance. GAs work on the population of candidate solutions; each solution has a fitness value indicating its closeness to the optimal solution of the problem. The solutions having higher fitness values than others are selected and also survive to the next generation. GA then produces better offspring (i.e. new solutions) by the combination of selected solutions. The methods can discover, preserve and propagate promising sub-solutions [4], [6], [10].

The inductive nature of Genetic Algorithms has made it a popular technique for improving the quality of artificial intelligence techniques. Recently GAs have been mapped with k-means clustering to select its initial centroids.Lleti´, Ortiz, Sarabia, and Sa´nchezproposed a method in [8] where Genetic Algorithms could be used to pick up the initial variables of k-means algorithm thereby improving the quality of clusters. Other studies have tried to solve the problem of convergence tolocal minimum in k-means algorithm by mapping it with genetic algorithms [2], [5], [9], [11], [12].

III. K-MEANS HYBRID CLUSTERING APPROACH TO RECOMMENDER SYSTEMS

Central to the k-means hybrid approach is the notion that better quality clusters will yield better recommendations to the users. But as described in previous section, the quality of clusters generated by k-means clustering depends on the initial seeds. Selection of different initial seeds can produce huge differences in clustering results.

Picking up the right initial seeds is the pivot point of this approach because it will decide the quality of clusters and thus the quality of recommendations.

A. Framework of k-means hybrid approach

Fig. 1 describes the framework of k-means hybrid approach. The detailed explanation of k-means hybrid approach to recommender systems is as follows:

1. In the first step, initial random population consisting of random seed combinations, is generated for genetic algorithms to operate upon.

2. The second step involves using k-means clustering to generate clusters of like minded users.

3. At this step, Genetic Algorithm applies its selection, crossover and mutation operators on the current population of initial seeds to produce the next generation which is better fit then the current generation. Fitness of each generation is measured in terms of cluster quality it produces. Cluster quality is measured as the sum of distances of each point from their corresponding centroid.

Step 2 and 3 are repeated until the stopping conditions are satisfied.

The objective of our approach is to pick up optimal initial seeds to produce high quality recommendations.

B. Chromosome Encoding

The system aims at finding an optimal combination of initial seeds. To apply genetic algorithms to search for the optimal seeds, permutation encoding is used to represent each chromosome. It involves representing each chromosome by a string of k random integers in the range of 1 and maximum number of items. Thus each chromosome represents a combination of k seeds.

Length of each chromosome is equal to k, the number of clusters. So if we have 200 users, then for k=6, a single chromosome will look like:

[1, 13, 24, 49, 71, 160]

Here each integer represents a user which has been picked up from the dataset as a cluster center for k-means algorithm.Genetic algorithm over a series of generations picks up the right cluster centers that increase the quality of clusters produced by k-means clustering algorithm. This in turn results in generating better recommendations for the users.

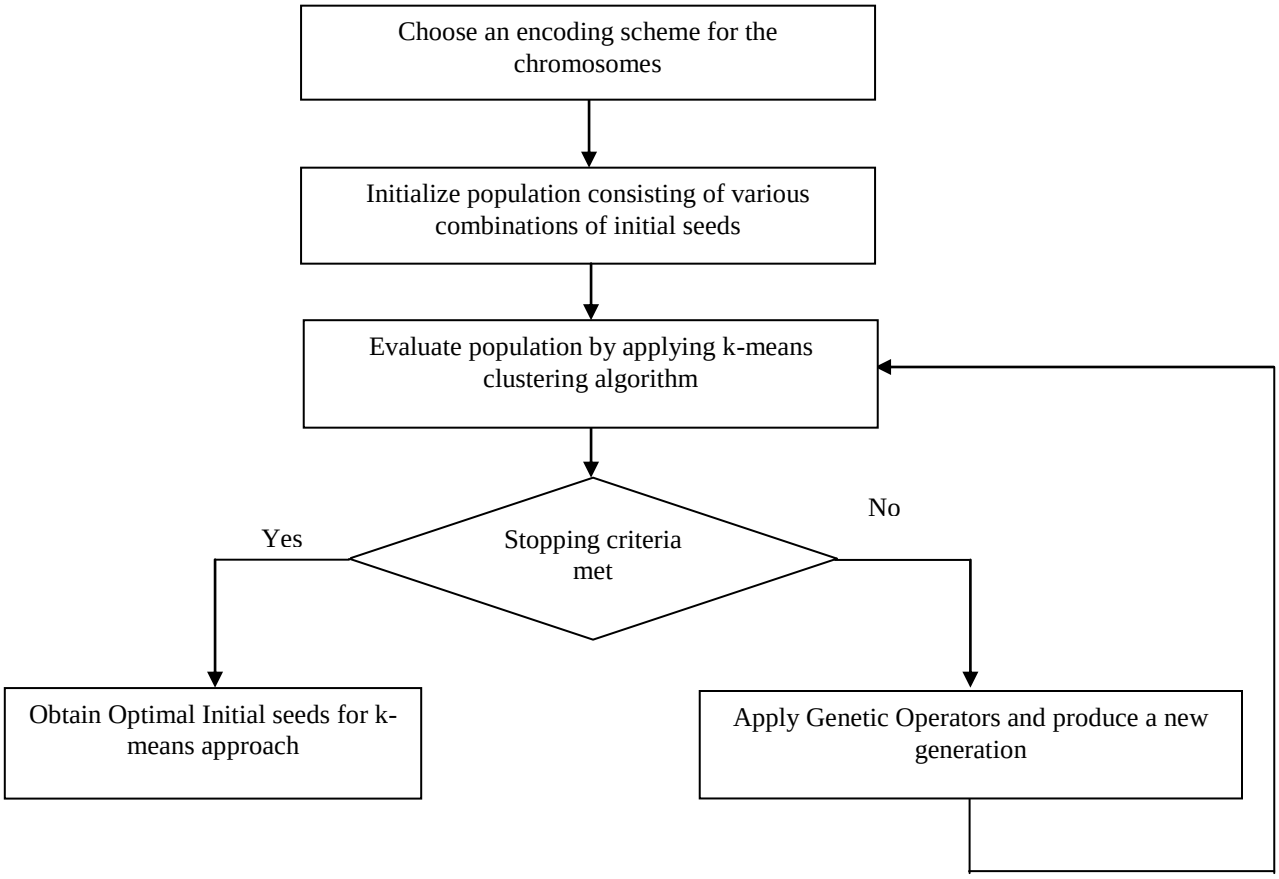

Fig.1. Overall framework of k-means hybrid approach

IV. EXPERIMENTAL DESIGN AND RESULTS

The performance of simple K- means algorithm is compared with the proposed hybrid model that combines GAs with k-means clustering for making recommendations. The study involves grouping users into 6 clusters. So k=6.

For controlling the parameters of Genetic Algorithm search, the population size is set at 100. The crossover rate is set at 0.1 and the mutation rate is set at 0.01. This study performs crossover operation by using one point crossover operator on permutation encoding. For carrying out mutation operation, a random position is identified, where the integer will be mutated by a random integer value which does not appear in that individual.

The k-means hybrid approach was conducted using MATLAB Version 7.2

A. Experimental Results

In this study, a data set from a movie recommender website named Movielens [17] is used for assessing the performance of the proposed model. A part of the data set consisting of 120 movieswas used. The dataset consisted of the ratings given by different users to these movies on the scale of 1 to 5 with 5 indicating the maximum liking for a movie and 1 indicating a strong dislike for the movie. 0 indicates that a user has not seen that movie. Consequently a recommender system has been built for these users.

As described earlier, the most common issue faced by recommender systems is that of overspecialization wherethe user is limited to being recommended items that are similar to those already rated by the user.

To overcome this problem, k-means clustering is used to group users into clusters with each cluster consisting of most similar users. So for suggesting movies to the target user, the cluster to which he belongs is identified and then he is recommended movies most liked by users in his cluster. As described in the Fig. 2, this entire process consists of 2 phases:

Phase1 consists of generating better quality clusters. This is done by using Genetic Algorithms to pick the initial seeds of k-means clustering algorithm.

Phase2 consists of recommending movies to the target user which were most liked by users in his cluster.

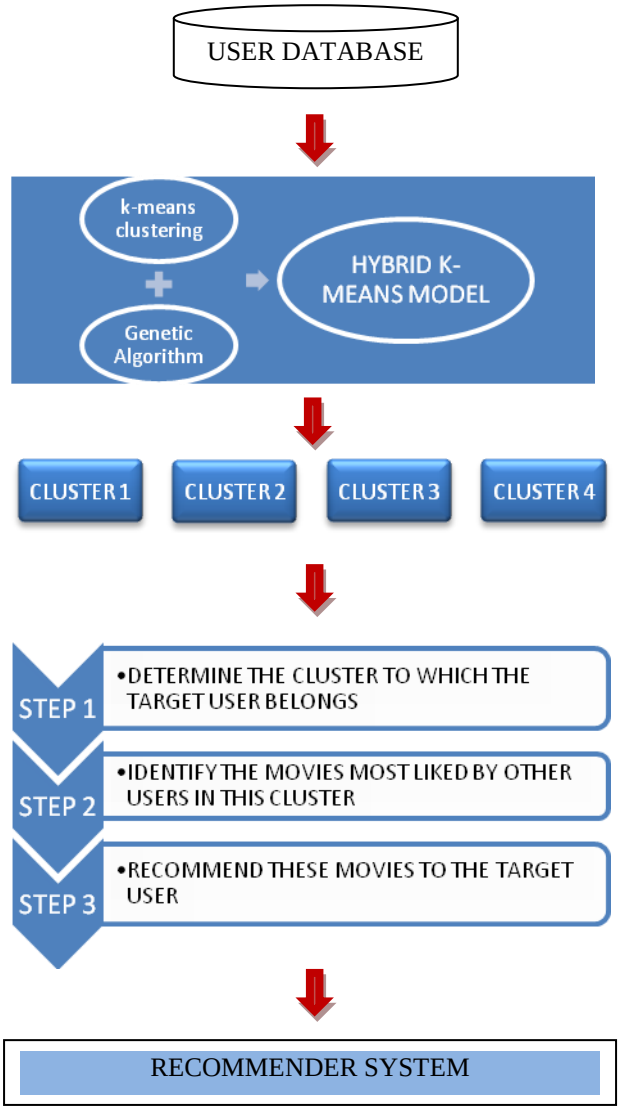

Fig 2: System Architecture of Recommender System

The collected data in this case included 200 customers. It consisted of the ratings given by different users to various movies on the scale of 1 to 5. K-means clustering algorithm used Euclidean distance to compute similarity between different users.

The results of simple k-means algorithm and the hybrid k-means approach using GAs were compared in terms of Silhouette plots.

Fig. 3 indicates the silhouette plot for the clusters obtained using simple k-means clustering whose seeds were picked randomly.

Fig. 4 indicates the improvement in the quality of clusters obtained using k-means hybrid approach whose seeds are optimized by genetic algorithms.

Fig. 5 indicates that the percentage of accurate predictions made by the hybrid approach is greater than simple k-means algorithm for majority of the users.

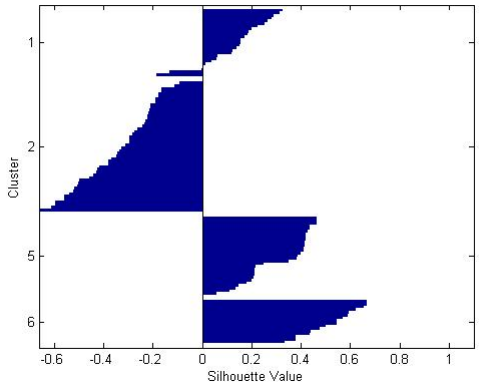

Fig 3: Silhouette plot for simple k-means clustering

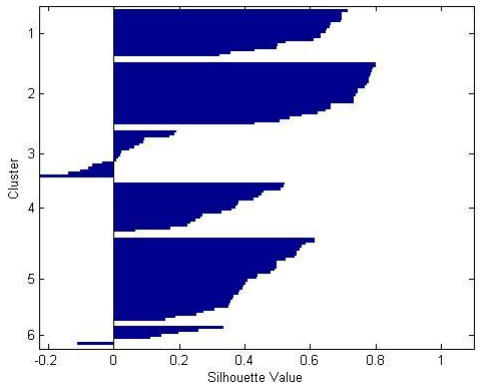

Fig 4: Silhouette plot for hybrid k-means approach

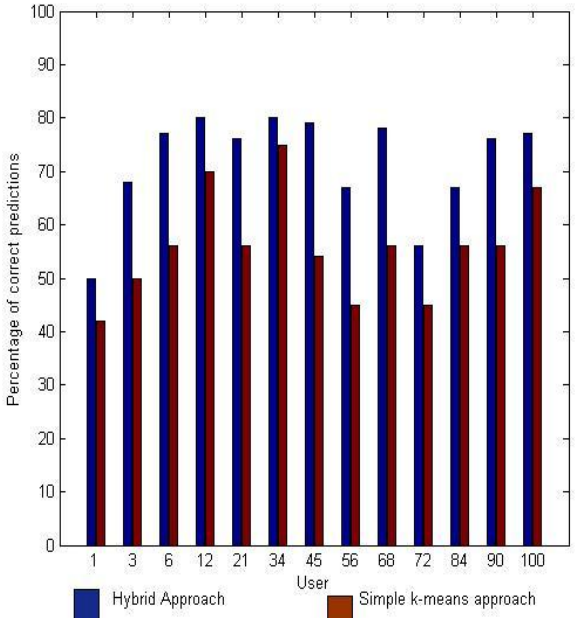

Fig. 5: Correct Predictions percentage for the users

V. CONCLUSIONS

This study suggests a new approach to recommender systems where k-means clustering can be mapped with genetic algorithms to improve the quality of clusters and there by provide better recommendations to the users.

However, this study has some limitations. Although this study considers collaborative filtering approach to recommender systems by identifying similar users, but here the characteristics of movies have not been considered for making recommendations. Consequently efforts to develop such a system should be made in future research.

Moreover, the number of clusters was arbitrarily set to six in this study. Unfortunately, a lot of work still needs to be done to propose a technique for identifying optimal number of clusters for k-means algorithm.

REFERENCES

[1] G. Adomavicius andA. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possibleextensions,"*IEEE Transaction on Knowledge and Data Engineering,*vol. 17, no. 6, pp. 734-749, 2005.

[2] G.P. Babu andM.N. Murty,"A near optimal initial seed value selection in K-means algorithm using a genetic algorithm,"*Pattern Recognition Letters*, vol. 14, no. 10, pp. 763-769, 1993.

[3] M. Balabanovic andY. Shoham, "Fab: Content-Based, Collaborative Recommendation,"*Communications of the ACM*, vol. 40, no. 3, pp. 66-72, 1997.

[4] D.E. Goldberg andJ.H. Holland, "Genetic Algorithm and Machine Learning,"*Machine Learning*, vol. 3, nos. 2 and 3, pp. 95-99, 1988.

[5] K. Kim andH. Ahn, "A recommender system using GA k-means clustering in an online shopping market,"*Proc. Expert Systems with Applications Journal*, vol. 34, no. 2, pp. 1200-1209, 2008.

[6] H. Kim, E. Kim, J. Lee andC. Ahn,"A Recommender System Based on Genetic Algorithm for Music Data,"*Proc. 2010 Second Conference on Computer Engineering and Technology*, vol. 6, pp. 414-417, 2010.

[7] R.J. Kuo, J.L. Liao andC. Tu,"Integration of ART2 neural network and genetic K-means algorithm for analyzing Web browsing paths in electronic commerce,"*Decision Support Systems*, vol. 40, no. 2, pp. 355-374, 2005.

[8] R. Lletı´, M.C. Ortiz, L.A. Sarabia andM.S. Sa´nchez, "Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes,"*Analytica Chimica Acta*, vol. 515, no. 1, pp. 87–100, 2004.

[9] U.Maulik andS. Bandyopadhyay, "Genetic algorithm - based clustering technique,"*Pattern Recognition*, vol. 33, no. 9, pp. 1455-1465, 2000.

[10] M. Mitchell,"An Introduction to Genetic Algorithm,"*MIT press*, 1998.

[11] C.A. Murthy andN. Chowdhury, "In search of optimal clusters using genetic algorithms,"*Pattern Recognition Letters*, vol. 17, no. 8, pp. 825-832, 1996.

[12] J.M. Pena, J.A. Lozano andP. Larranaga,"An empirical comparison of four initialization methods for the K-means algorithm,"*Pattern Recognition Letters*, vol. 20, no. 10, pp. 1027-1040, 1999.

[13] B. Sarwar, G. Karypis, J. Konstan andJ. Riedl, "Item Based Collaborative Filtering Recommendation Algorithm,"*Proc. 10th international conference on World Wide Web*, pp. 285-295, 2001.

[14] U. Shardanand andP. Maes, "Social Information Filtering: Algorithms for Automating 'Word of Mouth',"*Proc. conference on human factors in computing systems (CHI, 95)*, pp. 210-217, 1995.

[15] M. Wedel andW.A. Kamakura, "Market segmentation: concepts and methodological foundations,"*Boston Kluwer Academic Publishers*, 1998.

[16] http://www.amazon.com

[17] http://www.movielens.umn.edu