International Journal of Soft Computing 9 (3): 138-142, 2014
ISSN: 1816-9503

Behaviour Analysis Model for Social Networks using Genetic Weighted Fuzzy C-Means Clustering and Neuro-Fuzzy Classifier

[1]P. Indira Priya, [2]D.K. Ghosh, [3]A. Kannan and [4]S. Ganapathy
[1]Tagore Engineering College, Chennai, Tamil Nadu, India
[2]VSB Engineering College, Karur, Tamil Nadu, India
[3]Anna University, Chennai, Tamil Nadu, India
[4]S.M.K. Fomra Institute of Technology, Chennai, Tamil Nadu, India

Abstract: Genetic algorithms are helpful to make effective decisions using suitable fitness functions. They can be used to perform both clustering and classification. However, Clustering algorithms enhanced only with genetic operators are not sufficient for making decision in many critical applications. In this study, researchers propose a new user behaviour analysis model by combining Genetic algorithm with Weighted Fuzzy C-Means Clustering Algorithm (GNWFCMA) for effective clustering. The proposed clustering algorithm is used to improve the classification accuracy by providing initial groups. In addition, researchers use a five factor analysis also for effective clustering. Finally, researchers use a neuro-fuzzy classifier for classifying the data. The experimental results obtained from this study shows that the clustering results when combined with classification algorithm provides better classification accuracy when tested with Weblog dataset.

Key words: Clustering, Genetic algorithms, global optimization, Weighted Fuzzy C-Means algorithm, unsupervised learning

INTRODUCTION

Social network services allow the internet users to create virtual personalities using their profiles and to establish social connections. The user profile is gathered to collect the user details and user's interests on web pages which the user bookmark or any other information that matches the type of Social Networking System. The social connections are represented by explicit friendship links or they are extracted from various forms of people interaction. The social services allow the users to access the information about people with similar interests. Moreover, these groups are formed to indicate the social relations in terms of the communities they make. The information about community membership is valuable to provide social network services and to make relations more natural than the basic friendships.

Clustering helps to group a collection of objects that are either "similar" with one another or "dissimilar" from the objects of other clusters (Lipczak and Milios, 2009). In a Clustering algorithm, the distance measure between data points is an important component. If the components of the data instance vectors are present in the same group then simple Euclidean distance metric can be used to successfully find such similar data instances. However, the Euclidean distance may be misleading in certain instances. Therefore, different distance measures can be used to form clusters. Clustering algorithms are classified as exclusive clustering, overlapping clustering, hierarchical clustering and probabilistic clustering. In exclusive clustering grouped in an exclusive way and hence if certain data item belong to a particular cluster then it cannot be included in any other cluster. Here, the separation of points is performed by making a straight line on the two dimensional plane. On the other hand in the overlapping clustering, uses uncertainty is applied to cluster data so that, each point can be belong to more than one cluster with different degrees of membership. In the hierarchical clustering algorithm on the union between the two nearest clusters is considered to form a cluster. With a few iterations this algorithm provides the clusters necessary. Finally, the probability value is used to form clusters in the probabilistic clustering approach.

Genetic Algorithms (GAs) are heuristic search techniques that work based on the principles of evolution and natural genetics. GAs perform heuristic search in complex, large and multimodal landscapes and it provides near-optimal solutions using a fitness function in an optimization problem. In GAs, the components of the search space are represented in the form of strings (chromosomes). A collection of such strings is termed as a population (Krishna and Murty, 1999). Initially, a

Corresponding Author: P. Indira Priya, Tagore Engineering College, Chennai, Tamil Nadu, India

random population is created to represent different points in the search space. A fitness function is associated with each string which provides the degree of suitability of the string. Based on the principle of survival of the fittest, the least strings are selected from the population and they are sent for reproduction. Operators like cross-over and mutation are applied on these strings in order to form a new generation of strings. The process of selection, crossover and mutation continues for a fixed number of generations or till a termination condition is satisfied.

In this study, researchers propose a user behaviour analysis model which by combining Genetic algorithm with Weighted Fuzzy C-Means Clustering algorithm to form groups. These groups are further classified to form more interesting groups using a neuro-fuzzy classifier. The combination of clustering and classification provides effective groups with respect to areas of interest.

LITERATURE REVIEW

Lipczak and Milios (2009) proposed an Agglomerative Clustering Genetic Algorithm (ACGA): a population of clusters evolves from the initial state in which each cluster represents one user to a high quality clustering solution. Each step of the evolutionary process is performed locally, engaging only a small part of the social network limited to two clusters and their direct neighbourhood. This makes the algorithm practically useful independently of the size of the network. Evaluation on two social network models indicates that ACGA is potentially able to detect communities with accuracy comparable or better than two typical centralized Clustering algorithms seven though ACGA works under much stricter conditions.

As the traditional spectral clustering uses K-Means algorithm to cluster the data objects, K-Means algorithm itself is sensitive to the initialization and easy to fall into the local optimum. Combined with the global search ability of genetic algorithm, a Genetic Spectral Clustering algorithm is proposed by Wang *et al*. (2011). Zhou *et al*. (2004) proposed to combine the mutual information criterion and traditional distance criteria such as the Euclidean distance and the fuzzy membership metric in designing the Clustering algorithm. A novel hybrid Genetic Algorithm (GA) proposed by Krishna and Murty (1999) to specified number of clusters according to the given data and finds a globally optimal partition. This hybrid GA circumvent expensive crossover operations by using a Classical Gradient Descent algorithm used in clustering viz., K-Means algorithm. In Genetic K-means Algorithm (GKA), K-means operator was defined and used as a search operator instead of crossover. GKA also define a biased mutation operator specific to clustering called distance-based-mutation. Using finite Markov chain theory, it was proved that the GKA converges to the global optimum. GKA searches faster than some of the other evolutionary algorithms used for clustering. One of the important problems in partition clustering is to find partition of the given data with a specified number of clusters which minimizes the Total Within Cluster Variation (TWCV). Problem of minimization of TWCV was handled in GKA by Krishna and Murty (1999).

Fast Genetic K-means Algorithm (FGKA) (Lu *et al*., 2004a, b) was inspired by GKA but features several improvements over GKA. Experiments indicate that while K-means algorithm might converge to a local optimum both FGKA and GKA always converge to the global optimum eventually but FGKA runs much faster than GKA. Incremental Genetic K-means Algorithm (IGKA) (Lu *et al*., 2004a, b) was an extension to previously propose Clustering algorithm, the Fast Genetic K-means Algorithm (FGKA). IGKA outperforms FGKA when the mutation probability was small. The main idea of IGKA was to calculate the objective value Total Within Cluster Variation (TWCV) and to cluster centroids incrementally whenever the mutation probability was small. IGKA inherits the salient feature of FGKA of always converging to the global optimum. Lin *et al*. (2005) have proposed a GA-based unsupervised clustering technique that selects cluster centers directly from the data set allowing it to speed up the fitness evaluation by constructing a look-up table in advance, saving the distances between all pairs of data points and by using binary representation rather than string representation to encode a variable number of cluster centers. More effective versions of operators for reproduction, crossover and mutation were introduced.

Clustering Genetic Algorithm (CGA) proposed by Kudova (2007), the K-Means algorithm has some tasks. These tasks are well formed and separated clusters to find the optimizing number of clusters. The framework was the same as in Genetic algorithm while the individual building blocks of the algorithm were modified and adopted for the clustering task. In the field of data mining, it is often encountered to perform cluster analysis on large data sets with mixed numeric and categorical values. However, most existing clustering algorithms were only efficient for the numeric data rather than the mixed data set. A Hybrid Genetic Based Clustering algorithm called HGA-clustering was proposed by Liu *et al*. (2004) to explore the proper clustering of data sets. This algorithm with the cooperation of tabu list and aspiration criteria has achieved harmony between population diversity and

convergence speed. K-Modes algorithm has been developed for clustering categorical objects by extending from the K-Means algorithm.

Genetic Weighted K-Means Algorithm (GWKMA) which was a hybridization of a Genetic Algorithm (GA) and a Weighted K-Means Algorithm (WKMA) proposed by Wu (2008). GWKMA encodes each individual by a partitioning table which uniquely determines a clustering and employs three genetic operators (selection, crossover, mutation) and a WKMA operator. The superiority of the GWKMA over the WKMA and other GA-Clustering algorithms without the WKMA operator was demonstrated.

Functional Link Artificial Neural Network (ISO-FLANN) Model is proposed for the task of classification. Using the weight value obtained by IPSO and the set of optimized trigonometric basis functions chosen for the expansion of the feature vector, the method overcomes the non-linearity of the classification problem. Further the self adaptively Gaussian and Cauchy mutation can further fine-tune the solutions found by the proposed algorithm. Experimental study demonstrated that the performance of ISO-FLANN for classification task is promising. In most cases, the result obtained with the EFLSNN Model proved to be as good as or better than the best results found by the MLP, SVM, FLANN with gradient decent and FSN. The architectural complexity of the ISO-FLANN Model is quite less compare to MLP whereas it is the same or less as FLANN with gradient descent and FSN. This property of ISO-FLANN can attract the researches working in data mining for classification task (Dehuria *et al*., 2012).

SYSTEM ARCHITECTURE

The architecture of the system proposed in this research consists of six major components namely, weblog data set, user interface module, behaviour learning module, clustering module, classification module and response module as shown in Fig. 1.

Weblog data set: Weblog dataset is a offline dataset of the old online database like face book, twitter, etc. This dataset retrieved from internet using the default program.

User interface module: The user interface module collects the weblog dataset from internet. This data are sent to the behaviour learning module for checking the similarity between the different data. Users can interact with the system through this user interface. User interface gives the request to the model and gets the result.

Behaviour learning module: The behaviour learning module detects the dissimilar data from the given dataset using the five factor analysis. This module distinguishes

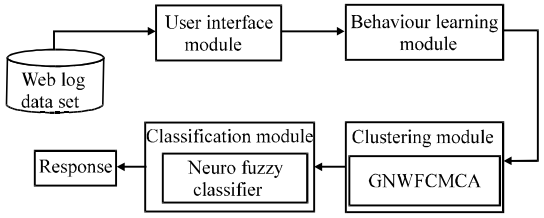

Fig. 1: System architecture

the dissimilar data from similar data using an intelligent technique for obtaining better result in similarity measurement.

Clustering module: The clustering module detects the dissimilar data from the given dataset using the five factor analysis. This clustering module distinguishes the dissimilar from similar data using the proposed Clustering algorithm called GNWFCMCA.

Classification module: The classification module classifies the dissimilar data from the given data using neuro-fuzzy classifier. The classified results are used to form most related interest group.

PROPOSED RESEARCH

Seven factor analysis: In the past, social networks were formed using a five factor analysis. In that model, five factor parameters namely frequency, duration, friends, gender and qualification were considered. Researchers used some ranges for all factors for clustering the data. Table 1 shows the list of parameters and ranges used by them. They have used the qualification factor also for easy identification in the member of data. However, finding friends with more similarities with respect to culture, food habits and climate are also essential to form better groups. Hence, two more new factors namely age and area are also included in this paper to provide a new seven factor analysis.

Genetic new weighted fuzzy C-Means Clustering algorithm: A Genetic based Weighted Fuzzy C-Means Algorithm (GNWFCMA) is proposed in this study for solving high-dimensional multiclass problems. In the existing FWCM, weighted means are calculated based on all the sample points whereas in the proposed NWFCM weighted mean is calculated based on cluster centers and the rest of sample points. As the weighted mean is calculated based on the cluster centers, this proposed algorithm is less computationally exhaustive than the existing FWCM.

Table 1: Seven factors analysis

Parameters	Description	Range of values
Frequency	Daily sessions	1 = one, 2 = two, three, 3 = 4-6, 4 = more
Duration	Typical length of a session	1 = few minutes, 2 = up to 1 h, 3 = 1-3, 4 = >3, 5 = always online
Friends	Number of friends	1 = <10, 2 = 10-20, 3 = 20-30, 4 = 30-50, 5 = 50-80, 6 = 80-100, 7 = 100-200, 8 = 200-400, 9 = 400
Gender	Male or female	1 = M, 2 = F
Qualification	Arts or engg.	1 = Arts, 2 = Engg.
Age	Age group	18-35 = Young, >35 = Senior
Area	Continent	1 = Asia, 2 = Erope, 3 = Africa, 4 = North America, 5 = South America, 6 = Autralia

Genetic Weighted Fuzzy C-Means Clustering algorithm:

Step 1: Set the parameters: population size N, the maximum number of iteration T, the number of clusters C, etc.

Step 2: Generate m chromosomes randomly, a chromosome represents a set of initial cluster centres to form the initial population.

Step 3: According to the initial cluster centers showed by every chromosome, compute weights to perform the weighted fuzzy C-means clustering. Calculate the chromosome fitness in line with clustering result using the activation function:

$$\text{Fitness} = \alpha \times \left(\frac{1}{\text{Count of ones}}\right) + \beta \times \text{Sensitivity} + \gamma \times \text{Specificity}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \quad \text{Specificity} = \frac{TN}{TN + FP}$$

Step 4: For each group, to carry out selection, crossover and mutation operator to produce a new generation of the group.

Step 5: To determine whether the conditions meet the genetic termination conditions, if meet then withdraw genetic operation and proceed to step 6, otherwise go to step 3.

Step 6: Calculate the fitness of the new generation of group; compare the fitness of the best individual in current group with the best individual's fitness so far to find the individual with the highest fitness.

Step 7: Carry out new weighted fuzzy C-means clustering according to the initial cluster center represented by the chromosome with the highest fitness and then output clustering result.

Neuro-fuzzy classifier: In this research, ANFIS (Jang, 1993) is used to perform effective classification of the data available in each classifier.

EXPERIMENT SETUP

In this study, researchers present the data collected from social media for evaluation and the baseline methods for comparison. Two benchmark data sets by Tang and Liu (2009) and Uchida and Shibata (2006) are used to examine this proposed model for collective behavior learning. The first data set is acquired from BlogCatalog3, the second data set is very famous social media dataset is Weblog dataset. Moreover, to examine the scalability, researchers also include a mega-scale network6 crawled from YouTube7. The following metrics are used for evaluating the earlier datasets:

$$\text{Precision} = \left(\frac{TP}{TP+FN}\right) \times 100$$

$$\text{Recall} = \left(\frac{TP}{TP+FP}\right) \times 100$$

Table 2: Performance evaluation of the clustering algorithms

	IGA-FKKMM (%)		IGA-NWFCM (%)		GNWFCMCA (%)	
Datasets	Precision	Recall	Precision	Recall	Precision	Recall
BlogCatalog3	96.98	96.35	98.13	97.43	98.65	97.92
Weblog	96.27	95.93	97.23	97.23	97.92	97.34
YouTube7	96.34	96.20	96.73	96.54	97.05	96.93

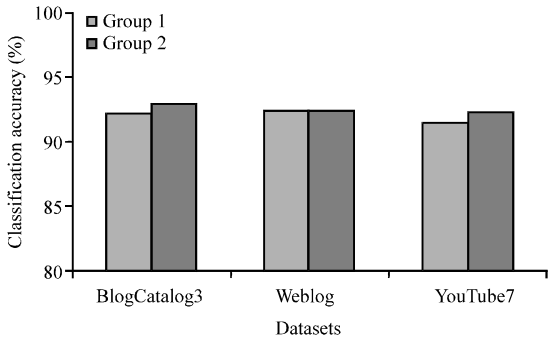

Fig. 2: Performance evaluations comparisons between group 1 and 2

Table 2 provides the comparison of recall and precision values of the proposed algorithm with the existing algorithms. From Table 2, it can be seen that the proposed Clustering algorithm shows better performance in terms of precision and recall values when it is compared with the existing systems.

Figure 2 shows the performance analysis of the neural classifier for the various types of datasets. From Fig. 2, it can be observed that the classification accuracy is more for group 2 when it is compared with group 1 due to the behaviour of the dataset.

Figure 3 shows the performance analysis of the neural classifier for the various types of datasets. From Fig. 3, it can be observed that group 4 classification accuracy shows better than group 3. This is due to the fact that more parameters including age and area have been considered in group 4.

Figure 4 shows the overall performance analysis of the neural classifier for the various types of datasets with proposed Clustering algorithm and without Clustering algorithm.

From Fig. 4, it can be observed that the proposed model provides better classification accuracy than the without clustering of neural classifier.

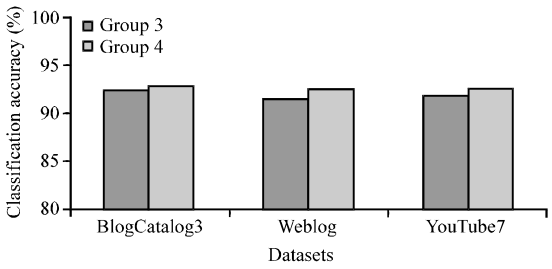

Fig. 3: Performance evaluations comparisons between group 3 and 4

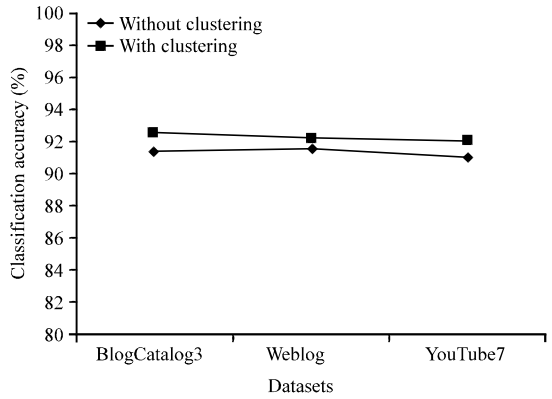

Fig. 4: Overall performance comparisons between with clustering and without clustering for all groups

CONCLUSION

In this study, a user Behaviour Analysis Model has been proposed by combining Genetic based Weighted Fuzzy C-Means Clustering Algorithm (GWFCMCA) and neural networks. The main advantage of this proposed model is that it uses clustering to identify the behavioural difference between two features of the records. The Genetic Weighted Fuzzy C-means Clustering algorithm proposed in this research build the system by grouping the records having significant difference. It also classified the other records exactly using neural classifier. The proposed model provides better classification accuracy than the existing systems. Future research in this direction can be the use of intelligent agents for enhancing the classification accuracy further by providing intelligent decision support.

REFERENCES

Dehuria, S., R. Royb, S.B. Choc and A. Ghosh, 2012. An improved swarm optimized functional link artificial neural network (ISO-FLANN) for classification. J. Syst. Software, 85: 1333-1345.

Jang, J.S.R., 1993. ANFIS: Adaptive-network-based fuzzy inference system. IEEE Trans. Syst. Man Cybern., 23: 665-685.

Krishna, K. and M.N. Murty, 1999. Genetic K-means algorithm. Trans. Syst. Man Cybern, Part B: Cyber., 29: 433-439.

Kudova, P., 2007. Clustering genetic algorithm. Proceedings of the 18th International Workshop on Database and Expert Systems Applications, September 3-7, 2007, ACM New York, USA., pp: 138-142.

Lin, H.J., F.W. Yang and Y.T. Kao, 2005. An efficient GA based clustering technique. Tamkang J. Sci. Eng., 8: 113-122.

Lipczak, M. and E. Milios, 2009. Agglomerative genetic algorithm for clustering in social networks. Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation, July 8-12, 2009, Montreal, QC., Canada, pp: 1243-1250.

Liu, Y.G., K.F. Chen and X.M. Li, 2004. A hybrid genetic based clustering algorithm. Proceedings of International Conference on Machine Learning and Cybernetics, Volume 3, August 26-29, 2004, Chaina pp: 1677-1682.

Lu, Y., S. Lu, F. Fotouhi, Y. Deng and S.J. Brown, 2004a. Incremental genetic K-Means algorithm and its application in gene expression data analysis. J. BMC Bioinform., Vol. 5. 10.1186/1471-2105-5-172.

Lu, Y., S. Lu, F. Fotouhi, Y. Deng and S.J. Brown, 2004b. FGKA: A fast genetic K-Means Clustering algorithm. Proceedings of the ACM Symposium on Applied Computing, October 10-16, 2004, ACM New York, USA., pp: 622-623.

Tang, L. and H. Liu, 2009. Relational learning via latent social dimensions. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, June 28-July 1, 2009, Paris, France, pp: 817-826.

Uchida, M. and N. Shibata, 2006. Extracting and visualization of an emerging topic from the blogspace. Proceedings of the 20th Annual Conference of the Japanese Society for Artifical Inteligence, ACM New York, USA., June 7-9, 2006.

Wang, H., J. Chen and K. Guo, 2011. A genetic spectral clustering algorithm. J. Comput. Inform. Syst., 7: 3245-3252.

Wu, F.X., 2008. Genetic weighted K-Means algorithm for clustering large-scale gene expression data. BMC Bioinform., Vol. 9. 10.1186/1471-2105-9-S6-S12.

Zhou, X., X. Wang, E.R. Dougherty, D. Russ and E. Suh, 2004. Gene clustering based on clusterwide mutual information. J. Comput. Biol., 11: 147-161.