Entropy Reduction Using Hybrid Genetic Weighted K-Mean Clustering and Neural Network

Dr. Amit Verma
Professor and Head,
Department of Computer Science and Engineering
Chandigarh University, Gharuan, Punjab, India

Er. Parminder Kaur
Assistant Professor
Department of Computer Science and Engineering
Chandigarh University, Gharuan, Punjab, India

Sandeesh Kaur
Research Scholar
Department of Computer Science and Engineering
Chandigarh University, Gharuan, Punjab, India

Abstract- In data mining research area, clustering is supervised learning that has drawn a lot of attention for its importance in real world applications such as image segmentation, text mining, email processing, language identification etc. Clustering dispersion known as entropy which is disorderness that occurred due to dissimilar elements present in cluster. It can be reduced by combining clustering algorithms with classifier. Better clustering improves accuracy of search result and reduces retrieval time. This paper proposes hybrid technique in which three algorithms are combined to enhance the output like GWKM, FFBPNN. Weighted k-mean is used for making clusters whereas GA is used for optimization and FFBPNN as classifier to enhance the output. The results are evaluated by various parameters like entropy, precision, recall and accuracy. By this way, optimized result will have less entropy and accuracy is increased.

Keywords –Genetic Weighted K-mean, Feed Forward back propagation Neural Network, Entropy Reduction, Recall, precision, accuracy

I. INTRODUCTION

Data Mining is the process of automatically finding useful information from large data repositories [19, 9]. Its main purpose is to analyzing data that are already present in databases and deals with various operations like retrieval, classification and summarization. It involves six common tasks: Anomaly Detection, Association rules learning, Classification, Clustering and Regression [8]. Proper clustering and classification is a crucial step with the growth of huge amount of information like World Wide Web, electronics documents. In this paper we discussed mostly on clustering class. Clustering is the most important unsupervised-learning problem as every problem is of this type. The main purpose is finding a structure in a collection of unlabelled data.

Clustering is defined as the process to bring together similar characteristics data into one group and dissimilar into another one. Its main objective is to produce high superiority clusters with high intra-class similarity and low inter class similarity [18].Dividing objects in meaningful groups of objects or classes (cluster) based on common characteristic, play an important role in how people analyze and describe the world. In the field of understanding data we can say clusters are potential classes, and cluster analysis is a studying technique to find classes. If clusters are not proper then search result may contain lot of randomness and involves a lot of time. For this, it is necessary to reduce the entropy within the cluster [21].

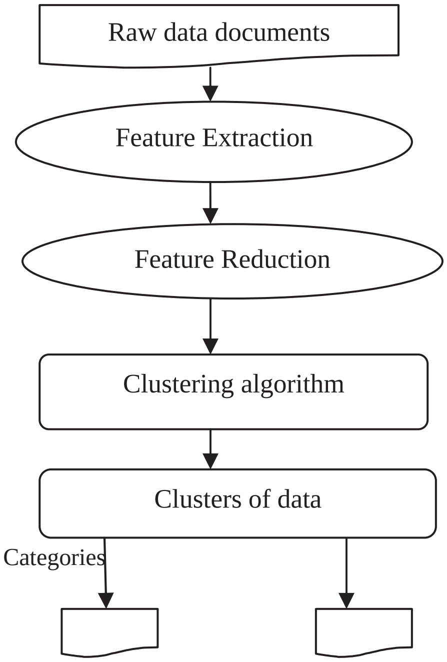

Figure 1. Clustering process

In Information theory, Entropy is the core concept that measure uncertainty [5,11]. Degree of uncertainty or dissimilar elements present in cluster indicates the entropy. Higher the dissimilar elements present in cluster which behave in unexpected way indicates the higher entropy value. It can be calculated by how much probability of elements belongs to particular cluster [].Supposing a discrete random variable X, which has x1, x2 ,..., xn , a total of n different values, the probability of xi appears in the sample is defined as P(xi), then the entropy of random variable X is[11]:

$$H(P) = -\sum p(x_i) \log p(x_i) \qquad \text{…Equation (1)}$$

If H (P) = 0 indicates the lower level of uncertainty, and the higher similarity in the sample. On the other hand, if H (P) value increases, it indicates the higher level of uncertainty, the lower similarity in the sample. By this way, when we detect the outliers from the cluster then the entropy decreases.

In this paper, comparative study of previous implemented GWKM with proposed GWKM +FFBPNN is done where Genetic Weighted k-mean is hybridization of weighted k-mean and Genetic algorithm. K-Means is arguably the most popular clustering algorithm; this is why it is of great interest to tackle its shortcomings. The drawback in the heart of this project is that this algorithm gives the same level of relevance to all the features in a dataset. Another issue of our concern is that K-Means results are highly dependent on the initial centroids. To address the issue of unequal relevance of the features we use a clustering algorithm called Weighted K-Means [10]. WKM is used for making clusters. In this, weights are assigned to objects for high dimensional data and objects which have higher weights are placed close to the centroid than the lower weight object. Genetic Algorithm is used to optimize the clusters that have been made by WKM so that search becomes efficient. It removes the unnecessary objects from the clusters. For every element of each cluster apply this algorithm and design the fitness function. FFBPNN is used as classifier to enhance the performance. It divides the optimized data into two sections 70% of data used for training and 30% used for testing section. In this, weight is updated again and again until expected output occurs and efficiently classified the data [4, 17]. Its response time is faster and performs well on large as well as small amount of data whereas Genetic algorithm works only on large data.

For this, leukemia dataset i.e. type of cancer is used which give information about the expression level of genes. The rows represent the expression level of genes and columns represent the samples. At last, results are evaluated by finding the parameters like entropy, accuracy, precision and recall after applying GWKM and GWKM+FFBPNN on same leukemia dataset. GWKM when merged with FFBPNN not only efficiently reduces the entropy but also improve the performance.

Remaining paper is organized as following sections. Section 2 describes the related work. Section 3 shows the problem statement. Section 4 shows the methodology to be used. Section 5 shows the results and discussion. Section 6 describes the conclusion and future work.

II. RELATED WORK

Ruiz et al. [2] designed automatic text categorization for solving manually organizing documents and recognizing problems. Mesh terms for particular document given the set of important words in the document and test it using MEDLINE dataset .The comparison of counter propagation network against back propagation neural network is done and observed that back propagation takes less training time and error rate is low then counter propagation network.

T.Kanungo et al. [3] represent the simple and efficient K-mean clustering algorithm. This algorithm requires kd-tree as the only major data structure. Firstly clusters are formed based on all data points assigned to nearest clusters and then determine cluster center. Its speed in making cluster is slow due to repeated two steps until all data is divided.

A.Selamat et al. [4] used WPCM method for classify web pages. For this, principle component analysis has been used to select the most relevant features and output of this is combined with the feature vector from class profile which contains the most regular words in each class. Then these are used as input to neural network for classification. It gives better classification accuracy with sport news database.

L.Jing et al. [10] done some extension to k-mean algorithm for clustering high dimensional data and remove sparsity problem that occur due to k-mean algorithm. For this, weight is calculate for each dimension in each cluster and uses this weight value to identify the important dimension of subset. This reduces the sparsity problem of high dimensional data and minimizes the cluster dispersion and entropy is reduced.

X.Bai et al. [11] done some advancement in k mean algorithm to remove its weakness like degree value is taken instead of sharp value i.e. determine how many data points belong to cluster. Entropy based soft k-mean algorithm is proposed in which the internal distance within cluster and external distance between clusters are calculated i.e. entropy and relative entropy. By this way, entropy value becomes smaller when value of stiffness function goes up and finally reaches a stable point.

F.Xiang et al. [12] proposed GWKM algorithm to overcome the limitations of k-mean and weighted k-mean like both are sensitive to initial partitions and its result is prone to local minima. For this, three genetic operators like selection, crossover, mutation and WKM operator used for clustering. Also rand index fitness function is used in genetic. Experiment result show that GWKM algorithm is better in terms of cluster quality and sensitive to initial partition.

Xiangjun et al. [13] described new relative rough entropy concept to deal with detects outlier in rough sets which is extended form of traditional entropy concept. The main aim is to consider the group of objects whose relative entropy high means that behave in unexpected way. By this way, an example shows that rough entropy measure is effective and suitable for evaluating outliers for high dimensional distance problems and reduces entropy.

P.Vishnu Raja et al. [14] developed new algorithm using genetic algorithm for outlier detection. It is better in computing the number of outliers in particular period of time. But this algorithm does not work on various data types and efficiency is low.

P.Patheja et al. [15] have done the comparison of efficient fitness function and rand index fitness function in GWKM algorithm to cluster gene expression data. An efficient fitness gives better result than rand index fitness function because variance in cluster can be improved using efficient fitness function i.e. minimize the internal features variance in cluster or maximize the variance between different clusters. But this algorithm only works for the large scale data of same type not for mixed type and also required to improve the performance and processing speed.

III. PROBLEM STATEMENT

The entropy of cluster is increased as the cluster gets dissimilar elements due to wrong selection of properties of the object. Therefore, the data which is wrongly classified produces entropy, i.e., disorderness or uncertainty of data object. If the cluster is not proper the searching may involve a lot of time and also the search results may not be exactly accurate. The main aim is to reduce the entropy which the user gets after searching.

This can be improved if clustering algorithm Genetic weighted k-mean and classifier feed forward back propagation neural network algorithm is used in combination. At last, results are evaluated by finding the parameters like entropy, accuracy, precision and recall after applying GWKM and GWKM+FFBPNN on same dataset.

IV. PROPOSED METHODOLOGY

The Proposed methodology works in two phases. At last, results are compared of these two phases:

Phase 1:

1. Upload the data sample.
2. Apply WKM algorithm for making clusters on basis of weight assigned to data.
3. Then apply GA to optimize the clusters.
4. Get the optimized results by evaluating parameters.

Phase 2:

5. Upload the data sample.
6. Repeat steps 2 & 3
7. Classify the clusters using FFBPNN.
8. Get the results by evaluating parameters.
9. At last, get a comparison between GWKM and GWKM with FFBPNN on basis of various parameters like entropy, accuracy, precision and recall.

A. *Pseudo Code*

The pseudo code of the proposed methodology Genetic weighted k-mean with Feed forward back propagation neural network are as follows:

Pseudo-code: GWKM+FFBPNN

```
Start
Do Upload Dataset
Then Features Extract
Do initialize clusters using weighted k-mean algorithm.
Apply genetic
Load genetic configurations
Evaluate fitness function
Decrease the cluster elements and optimize them
Initialize = cluster centre C and dimension weights W
If Data processing = TRUE
Then,
Randomly generate weights for n number of data = w1, w2, w3………..wn.
Then
Calculate distance from randomly generated centres = d1, d2, d3…dn.
Once all data in processed divide the data into k-mean clusters using Dn = W1 +w2 +w3……wn
Then apply Feed Forward Back Propagation Neural Network.
Initialize the network for the classification.
If (actual output = Desired output)
True
Classified data
If False
Then repeat steps
Stop
```

Figure 2. GWKM +FFBPNN algorithm

In this, weighted k-mean algorithm is the extension of k-mean algorithm. Its main importance to use is it overcomes the disadvantage of k-mean algorithm i.e. it works on numeric as well as text data and remove centroid problem. It is used for making clusters. In this, after initialize the cluster center, randomly weights are assigned to objects for high dimensional data and count distance from randomly generated centers and objects which have higher weights are placed close to the centroid than the lower weight object.

Then Genetic algorithm is used for optimization. It is commonly used in applications where search space is huge and the precise results are not very important. In this, chromosomes are generated with set of population called individuals. Individuals are evaluated using fitness function. For Selection, fitness function is used to evaluate the individuals from population and higher fitness individuals are selected for next population. The main components of GA are: crossover, mutation, and a fitness function [7]. If fs<=ft then return f=1 and if fs>ft then return f=0, where fs is the current element of the current dataset and ft is the overall optimization value.

At last back propagation neural network is applied to output of GA algorithm. The general strategy of FFBPNN is to optimal partition of the points into K classes and divides the optimized data into two sections, 70 % of the data for the training and 30 % of the data for the testing section. Initialize Feed Forward Back Propagation Neural Network with n Hidden Neurons and weight is updated again and again until excepted outcome occur [16]. It. efficiently classified the data and predicts accuracy.

V. RESULTS AND DISCUSSION

A. Experimental Setup

The proposed work is taken place in MATLAB environment. For this, Leukemia dataset i.e. type of cancer is used. This dataset represents the expression level of 4080 genes taken from 86 samples. In this, column represents the samples of one patient at different interval of time and rows represent the expression of corresponding genes. The main task is to classify the cancerous genes in one cluster and remaining to other ones.

B. Results

The results are evaluated by implementing both Genetic weighted k-mean and GWKM with Feed forward back propagation algorithms on same leukemia dataset. Results are evaluated on basis of various parameters values like accuracy, entropy precision and recall according different iterations. At last, comparison graphs are plotted to show GWKM with FFBPNN gives better result and show less entropy than the GWKM algorithm.

5.1. Optimization of data after Applying GA

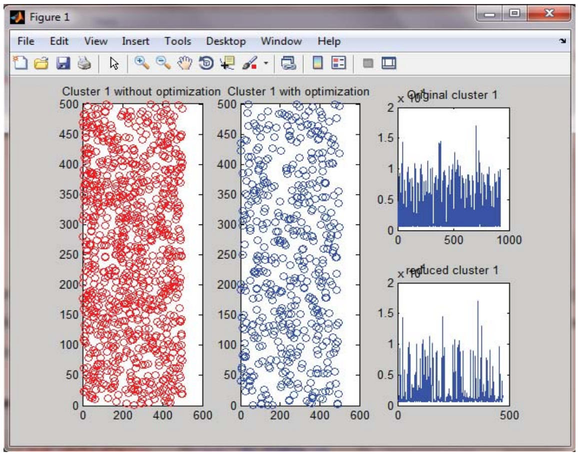

Figure 3. Comparison of cluster1 before and after applying GA

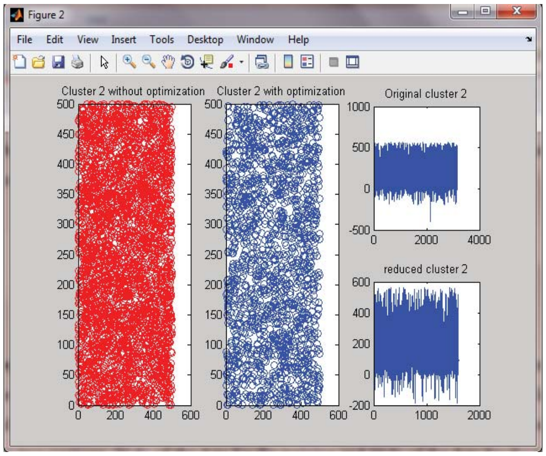

Figure 4. Comparison of cluster2 before and after applying GA

Both figures show the optimization of data after applying GA. Figure 3 show comparison of cluster1 before applying the GA to the after applying GA. It is seen numbers of genes are 954 of cluster1 but after applying GA these genes are reduced to 474 approximately. Figure 4 show Comparison of cluster2 before applying the GA to the after applying GA. It is seen numbers of genes are 3156 of cluster1 but after applying GA these genes are reduced to 1070 approximately.

5.2. Comparing GWKM and GWKM with FFBPNN

Comparison of different parameters of GWKM and GWKM with FFBPNN at different iterations is done to show GWKM with FFBPNN gives better result. In this, values of parameters are taken at different iterations.

Accuracy Graph

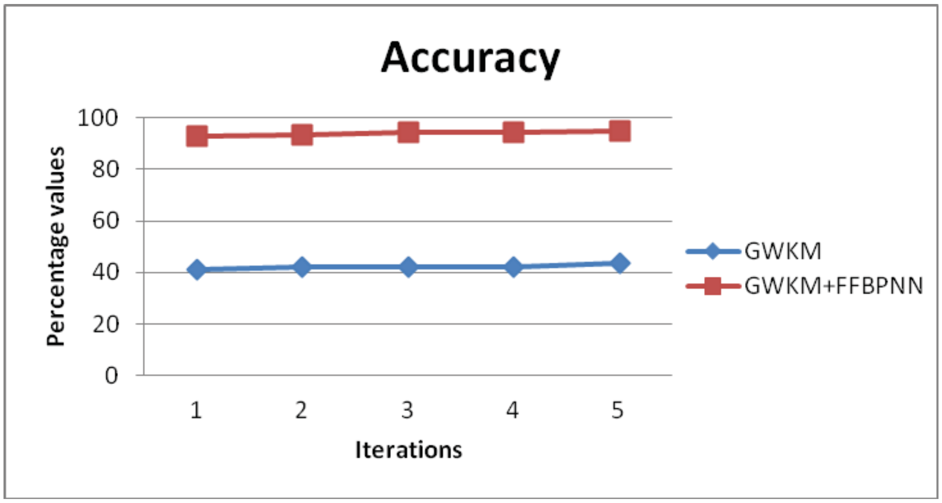

Figure 5. Accuracy graph between GWKM and GWKM with FFBPNN

Figure 5 show the comparison of accuracy parameter between GWKM and GWKM with FFBPNN. It is clearly seen from the graph that accuracy of GWKM with FFBPNN is greater at different iterations as neural network gives better performance in retrieved result.

Entropy Graph

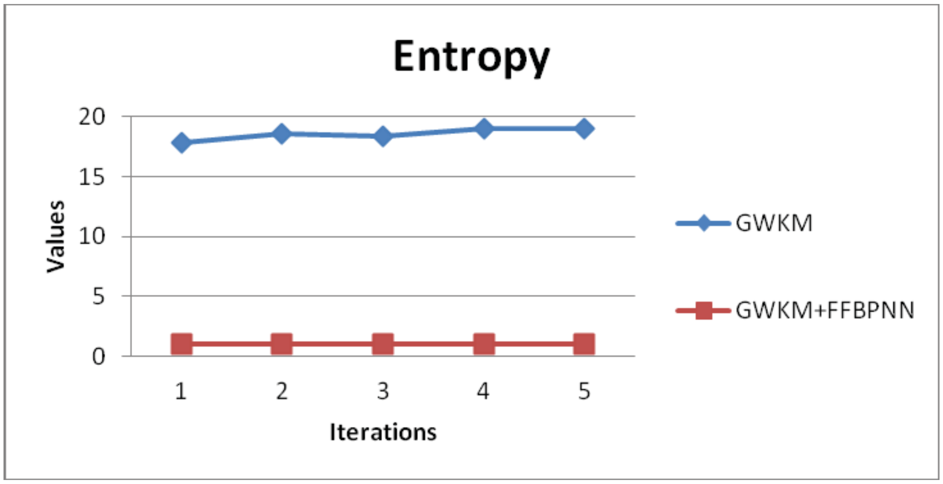

Figure 6. Entropy graph between GWKM and GWKM with FFBPNN

Figure 6 show the comparison of entropy parameter between GWKM and GWKM with FFBPNN. It is clear from the graph that entropy of GWKM with FFBPNN is lesser than GWKM at different iterations. Red line in graph show the lower uncertainty in the output means entropy is effectively reduced. Entropy is randomness in the retrieved result [20]. Entropy value increases indicates the higher level of uncertainty and decreases indicates the lower level of uncertainty.

Precision graph

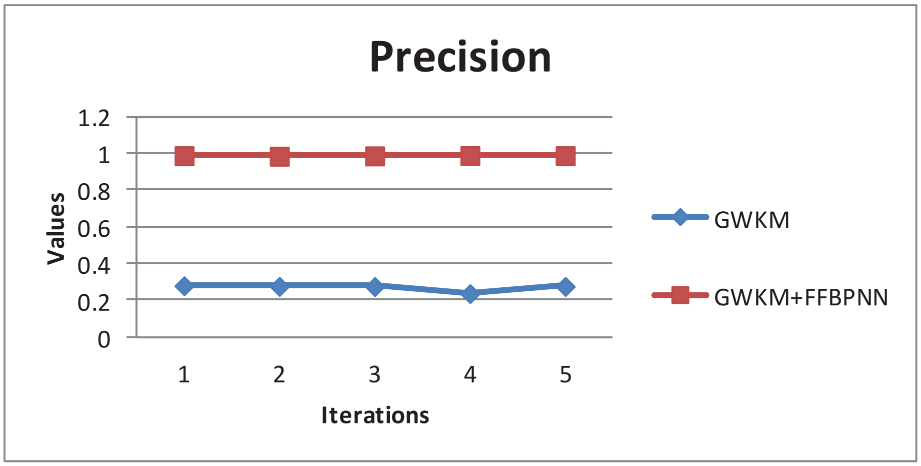

Figure 7. Precision graph between GWKM and GWKM with FFBPNN

Figure 7 show the comparison of precision parameter between GWKM and GWKM with FFBPNN. It is clearly seen from the graph that precision value of GWKM with FFBPNN is greater at different iterations.

Recall graph

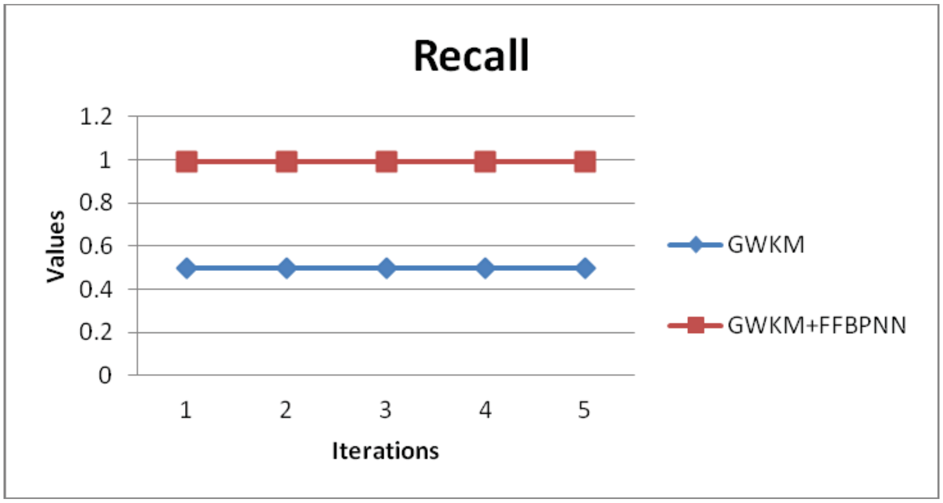

Figure 8. Precision graph between GWKM and GWKM with FFBPNN

Figure 8 show the comparison of precision parameter between GWKM and GWKM with FFBPNN. It is clearly seen from the graph that recall value of GWKM with FFBPNN is greater.

5.3. Comparing average values of parameters

Table 1: Parameters of GWKM and GWKM with FFBPNN

PARAMETERS	Accuray	Entropy	Precision	Recall
GWKM	42.2687	18.589	0.2769	0.497
GWKM+FFBPNN	93.9512	1.0416	0.9889	0.994

In this, average comparisons of parameters like accuracy, precision, recall and entropy of proposed methodology with GWKMA are given in form of values.

5.4. Comparing Average graphs of parameters

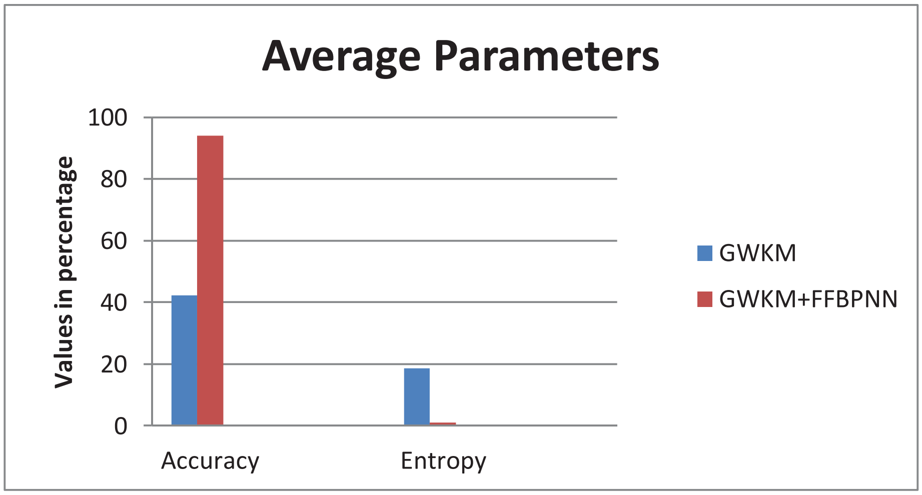

Figure 9. Average graph of parameters accuracy and entropy

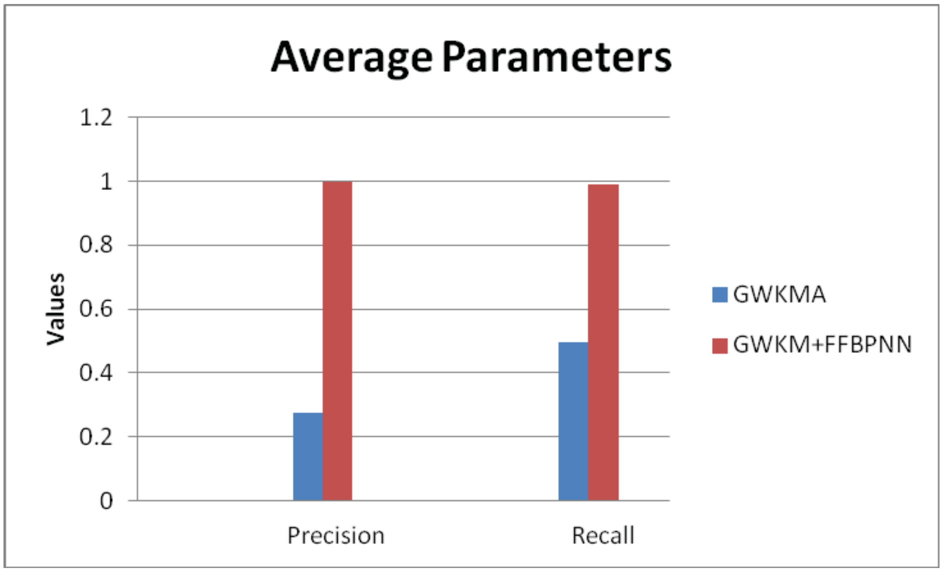

Figure 10. Average graph of parameters precision and recall

Figure 9 and 10 shows that GWKM with FFBPNN give greater accuracy, precision and recall value and less entropy value. So it gives better results than GWKM.

.

VI. Conclusion and Future Work

In this paper, we have presented clustering algorithm with hybrid GWKM/ FFBPNN classifiers. Weighted- k mean is used for clustering and GA is used for optimized the clusters and neural gives more accurate result because usage of many epochs. Weighted k-mean are suitable for high dimensional data and removes the sparsity problem. In order to label the unlabelled data, we have presented classification by NN because they can be effectively used for noisy data and it can also work on untrained data. Results of this hybrid technique are more optimized because GA works efficiently on large data set for optimization of cluster and neural network on both data sets i.e. small and large. Using this hybrid technique, entropy of the retrieved data can be reduced and to retrieval time, accuracy can be greatly enhanced.

The future aspect of this research work involves performing the clustering process on compound dataset to analyses the performance. Pyramidal NN can also be used for more accurate result so that result dependency on data can be reduced.

REFERENCES

[1] Marmelstein, Robert E. "Application of genetic algorithms to data mining." In Proceedings of 8th Midwest Artificial Intelligence and Cognitive Science Conference (MAICS-97), edited by E. Santos Jr., AAAI Press, pp. 58-65, 1997

[2] Ruiz, Miguel E., and Padmini Srinivasan. "Automatic text categorization using neural networks." In Proceedings of the 8th ASIS SIG/CR Workshop on Classification Research, pp. 59-72, 1998.

[3] Kanungo, Tapas, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. "An efficient k-means clustering algorithm: Analysis and implementation." Pattern Analysis and Machine Intelligence, IEEE Transactions on 24, no. 7, pp.881-892, 2002

[4] Selamat, Ali, Sigeru Omatu, Hidekazu Yanagimoto, Toru Fujinaka, and Michifumi Yoshioka. "Web page classification method using neural networks."IEEJ Transactions on Electronics, Information and Systems 123, no. 5, pp.1020-1026, 2003

[5] Li, Haifeng, Keshu Zhang, and Tao Jiang. "Minimum entropy clustering and applications to gene expression analysis." In Computational Systems Bioinformatics Conference, CSB 2004. Proceedings. 2004 IEEE, pp. 142-151, 2004

[6] Liang, Jiye, and Zhongzhi Shi. "The information entropy, rough entropy and knowledge granulation in rough set theory." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 12, no. 01, pp.37-46, 2004

[7] Sastry, Kumara, David Goldberg, and Graham Kendall. "Genetic algorithms." In Search methodologies, pp. 97-125, Springer US, 2005

[8] Singh, Yashpal, and Alok Singh Chauhan. "Neural networks in data mining."Journal of Theoretical and Applied Information Technology 5, no. 6, pp.36-42, 2005

[9] Sumathi, Sai, and S. N. Sivanandam. “Introduction to data mining and its applications.” Springer, Vol. 29, 2006

[10] Jing, Liping, Michael K. Ng, and Joshua Zhexue Huang. "An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data."Knowledge and Data Engineering, IEEE Transactions on 19, no. 8, pp.1026-1041, 2007

[11] Bai, Xue, Siwei Luo, and Yibiao Zhao. "Entropy based soft K-means clustering." In Granular Computing, IEEE International Conference on, pp. 107-110, IEEE, 2008

[12] Wu, Fang-Xiang. "Genetic weighted k-means algorithm for clustering large-scale gene expression data." BMC bioinformatics 9, no. 6, 2008

[13] Li, Xiangjun, and Fen Rao. "A rough entropy based approach to outlier detection." Journal of Computational Information Systems 8, no. 24, pp.10501-10508, 2012

[14] Raja, P. Vishnu, and V. Murali Bhaskaran. "An effective genetic algorithm for outlier detection." International Journal of Computer Applications 38, no. 6, pp. 30-33, 2012

[15] Patheja, P. S., Akhilesh A. Waoo, and Ragini Sharma. "Comparison of Efficient and Rand Index Fitness Function for Clustering Gene Expression Data." In Advances in Computer Science and Information Technology. Computer Science and Information Technology, pp. 160-167, Springer Berlin Heidelberg, 2012

[16] Patra, Anuradha, and Divakar Singh. "Neural Network Approach for Text Classification using Relevance Factor as Term Weighing Method." International Journal of Computer Applications 68, no. 17, pp.37-41, 2013

[17] Madasamy, B., and J. Jebamalar Tamilselvi. "Improving classification Accuracy of Neural Network through Clustering Algorithms." International Journal of Computer Trends and Technology, pp.3242-3246, vol.4, 2013

[18] Mann, Amandeep Kaur, and Navneet Kaur. "Review Paper on Clustering Techniques." Global Journal of Computer Science and Technology 13, no. 5, 2013

[19] Jain, Nikita, and Vishal Srivastava. "Data Mining techniques: A survey paper." IJRET: International Journal of Research in Engineering and Technology 2, no. 11, pp. 2319-1163, 2013

[20] Palwinder kaur1, Usvir kaur 2 ,Dr.Dheerendra Singh3. “Hybrid Clustering and Classification for Entropy Reduction: A Review.” International Journal of Innovative Research in Computer and Communication Engineering, vol.02, 2014

[21] Shalu Sharma , Sukhvinder Kaur and Ms. Jagdeep Kaur. “Hybrid Clustering and Classification.” International Journal of Advanced Research in Computer Science and Software Engineering, vol.05, 2015 Reddy