Volume 4, Issue 9, September 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Hybrid Clustering and Classification Using Weighted K Mean, Neural Networks and SVM

Palwinder Kaur
Computer Science, SGGSWU,
Fatehgarh Sahib, India

Usvir Kaur
Computer Science, SGGSWU,
Fatehgarh Sahib, India

Dheerendra Singh
Computer Science, SUS College,
Tangori, Mohali, India

Abstract-Clustering is very helpful in the analysis of raw data. As the clustering is unsupervised learning technique so in order to optimize the results we use classifier with clustering. This paper proposes hybrid technique in which weighted k mean is used for clustering whereas neural networks and SVM are used as classifiers .Weighted k mean results in unlabelled data by calculating the feature weights of clusters. Neural networks and SVM will label the unlabelled data as they have the ability to recognize the patterns. Optimized results will have the less entropy which is the randomness in the retrieved results. Also we have optimized the 8 performance measures of clustering and classification.

Keywords- Weighted k mean (WKM), Feed forward Neural networks (NN), Support vector machine (SVM), entropy reduction, Specificity, Sensitivity, F measure, G mean, precision, recall.

I. INTRODUCTION

In web mining , content process mining consist of the conversion of Web information like text, images, scripts and others into useful forms with the clustering, categorization ,classification of Web page titles. But this useful form is main concern as the search results contain a lot of randomness, inaccuracy in the retrieved results which are queried by the user.

In this paper, we are describing a hybrid technique in order to give better search results. In this hybrid method, first we are using weighted k mean for making clusters of text files and excel sheets. The output of weighted k mean which is the clusters is given as input to the neural networks. Neural networks by classifying the data in its internal hidden layers gives the optimized results with effective accuracy and reduced entropy.

Similarly, again we make the clusters with weighted k mean and give its output to SVM (support vector machine).SVM being classifier, classify the unlabelled data by applying labels and give us the optimized results. Results with both the classifiers are compared which shows that weighted k mean with neural networks gives more better results than weighted k mean with SVM.

II. TECHNIQUES

A. Weighted k mean for clustering-

By introducing weight vector, the weighted K Means algorithm not only decreases the effects of irrelevant attributes towards clustering results, but also reflects semantic information when clustering takes place [2].

Algorithm steps of weighted k mean

1. Begin
2. initialize Dataset D,N, K, C1, C2, . . . , CK, Current Pass=1; where D is dataset, N is size of data set, K is number of clusters to be formed, CW1, CW2, . . . , CWK are cluster centers and W is weight Assigned, Current Pass is the total no. of scans over the dataset.
3. do assign the n data points to the closest Ci; if CurrentPass%2==0 recomputed C1, C2, . . . , CK using weighted K Mean function; Else Increase Current Pass by one. Until no change in CW1, CW2 …. CWK;
4. Return CW1, CW2, . . . , CWK;
5. End [22].

B. Neural networks and SVM for the classification of data-

A neural network is a group of nodes that are interconnected for pattern recognition [3]. Here, each circular node represents an artificial neuron that processes the data to give the optimized classification results. Feed forward technique of neural is used in the proposed methodology. NN has the transparency and more epochs property which makes it better classifier.

SVM stands for support vector machine. It is very useful in Information retrieval (IR) for target recognition [11]. SVM is the classification function to distinguish between members of the two classes in the training data [9].Linear SVM technique is also used in our proposed work.

C. Genetic algorithm for more optimized results-

It is an evaluation function is used to calculate the "goodness" of each chromosome. During evaluation, two basic operators, crossover and mutation, are used to simulate the natural reproduction and Mutation of species. The selection of chromosomes for survival and combination is biased towards the fittest chromosomes. The selection of chromosomes for survival and combination is biased towards the fittest chromosomes [21].

D. Proposed methodology

The proposed methodology works in two phases. In first phase, WKM is combined with feed forward NN and in second phase WKM is combined with SVM. In the end, Results are compared between these two phases-.

Phase-1

1. Upload the data sample.
2. Apply WKM to get the clusters on the basis of weight assigned to data.
3. Classify the clusters using feed forward NN.
4. Get the optimized results.

Phase-2

5. Upload the data sample
6. Apply WKM to get the clusters on the basis of weight assigned to data features.
7. Classify the clusters using SVM.
8. Finally get a comparison between WKM with feed forward NN technique and WKM with SVM on the basis of various performances metric.

E. Parameters of Evaluation

1) Entropy-It is the randomness in retrieved results. Its value varies between 0 and 1.0 means less uncertainty in the sample so value must lie near 0 [4].

$$H(p) = -\sum_{i=1}^{n} p\,(xi) log p(xi) \qquad (2)$$

2) Efficiency-It is the percentage of the efficient results. Its value must be high.

3) Accuracy-It is the percentage of the accuracy with which the retrieved results are best coming. Its value must be high[16].

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \qquad (3)$$

Where TP= true positives, TN = true negatives, FP= false positives, FN= false negatives.

4) G mean-It represents the mean of the information represented. It is the normalized arithmetic mean form but less than arithmetic mean [18].

$$G\ mean = \sqrt{precision \times recall} \qquad (4)$$

5) F measure- It is used for the document classification and also for the query classification performance. Its value is best at 1 and worst at 0[14][15].

$$\text{F measure} = \frac{2 \times recall \times precision}{recall + precision} \qquad (5)$$

6) Recall-It is the ratio of hits to the number of URLs that are requested. Its value must be less for better results[14],[15].Where TP is True positive, FN is false negative.

$$recall = \frac{TP}{TP+FN} \qquad (6)$$

7) Precision-Ratio that hit the number of URLs prefetched. Its value must be higher and its range is in between 0 and 1[14],[15].

$$precision = \frac{TP}{TP+FP} \qquad (7)$$

8) Specificity-It is the true negative rate that is correctly rejected. Its value must be higher[17].

Where TN is true negative and FP is false positive.

$$Specificity = \frac{TN}{TN+FP} \qquad (8)$$

9) Sensitivity- It is the true positive rate that is correctly identified. Its value must be lower for better results. Its range is 0 and 1[17].

$$Sensitivity = \frac{TP}{TP+FN} \qquad (9)$$

III. RESULTS

A. Experimental setup

We have implemented the proposed scheme using MATLAB R2010a. Four datasets are used for the clustering and classification. Dataset 1 is apple.xls which contains the database of apple company.Dataset 2 is Job.xls which is the database of jobs. Dataset 3 is the lung cancer.txt, which contain the information regarding the symptoms of lung cancer. Dataset 4 is the breastcancer.txt, which give the symptoms of breast cancer.

Table 1- Datasets used

Dataset1	Apple.xls(Src. UCIrepository)
Dataset2	Jobs.xls(Src. UCI repository)

Dataset3	Lungcancer.txt(Src. UC repository)
Dataset4	Breastcancer.txt(Src. UCI repository)

First phase include WKM with feed forward NN- In first phase, WKM with feed forward NN is used. WKM make the clusters of datasets by assigning weights to the features of objects. Neural signals are tuned with weights. Neural has 3 layers; which are the input layer, hidden layers and output layers. In input layer we give classification set and training set as input. Classification set is the output of weighted k mean and training set is the data that is already trained by default in mat lab. Hidden layers contain epochs. In these hidden layers, FAR and FRR are created for the classification results.
Entropy reduction in WKM with NN-

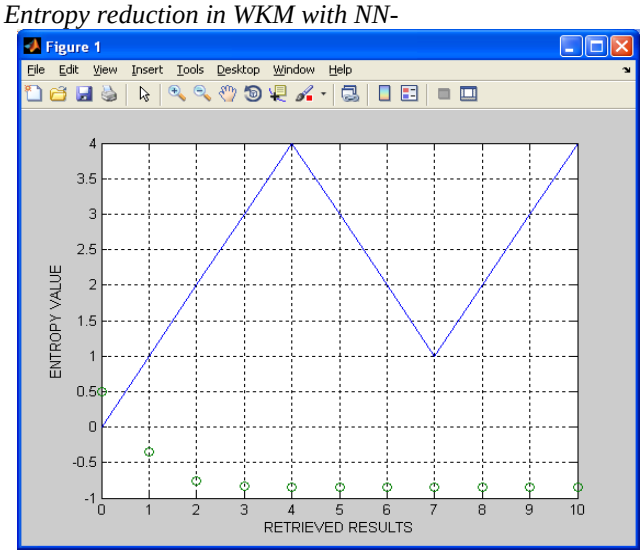

Figure 1- Entropy factor before optimization

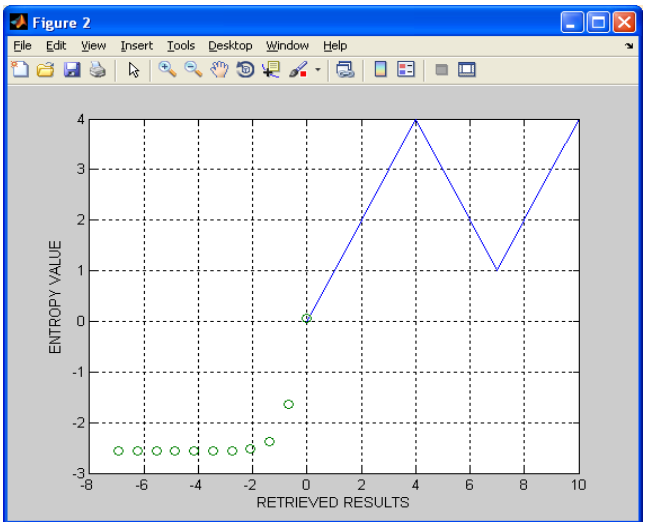

Figure 2- Entropy factor after optimization

In figure 1, green dots tell us the outliers detected in the retrieved results. They are scattered to the larger area. Outliers more in results leads to more entropy which is a major cause for worst results. In figure 2, green dots are the outliers. After optimization of WKM with feed forward NN we get the better results, as the entropy is reduced to the smaller area and after that we get the results without outliers which are shown by blue line.
Second phase using WKM with SVM – in second phase, WKM with SVM classifier is used. Linear classification is used. Linear SVM technique is used. Also for more optimization genetic algorithm has been used in both phases which gives the survival of fittest**.**

B. Comparing WKM with NN and WKM with SVM
Accuracy graph
Figure 3 shows that accuracy of neural networks is greater than SVM as neural networks give more accuracy in retrieved results. Accuracy depends upon the probability of detection and probability of false alarm.

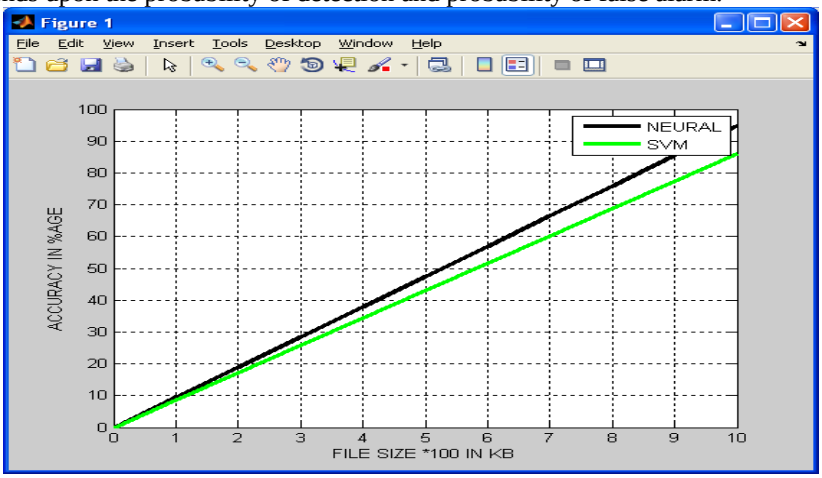

Figure 3 Accuracy graph between WKM with NN and WKM with SVM

Entropy graph
In figure 4, entropy of NN retrieved results is very less as compared to SVM. By hybrid WKM with feed forward NN entropy is effectively get reduced. Blue line of SVM shows the higher uncertainity in the outputs and black line proves that NN reduced the entropy. Entropy is the randomness in the retrieved results. If the uncertainity is lower then its value reaches 0 otherwise reaches 1. Entropy is reduced in case of feed forward NN as compared to SVM classifier because SVM has very less concept of transparency.

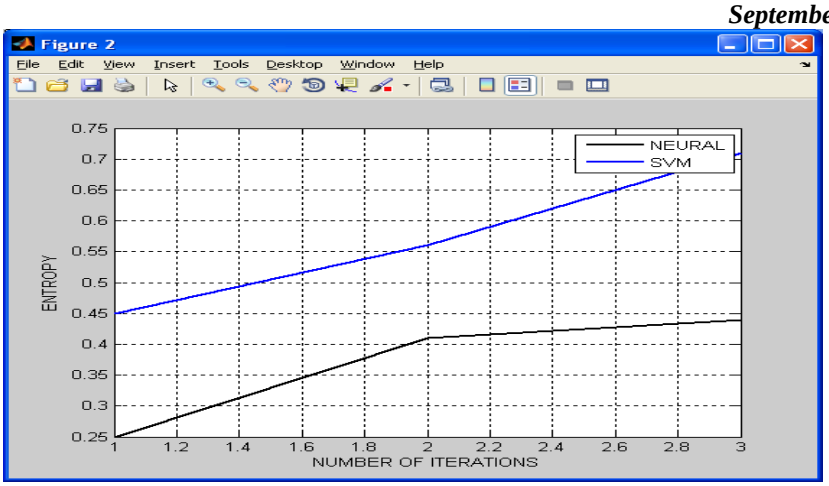

Figure 4 Entropy graph between WKM with NN and weighted k mean with SVM.

Sensitivity graph

In figure 5, Sensitivity for NN is less as compared to the SVM technique. Sensitivity is the true positive rate that can be correctly identified. The black line in graph proves that NN has less sensitivity and blue line proves the greater sensitivity in SVM.

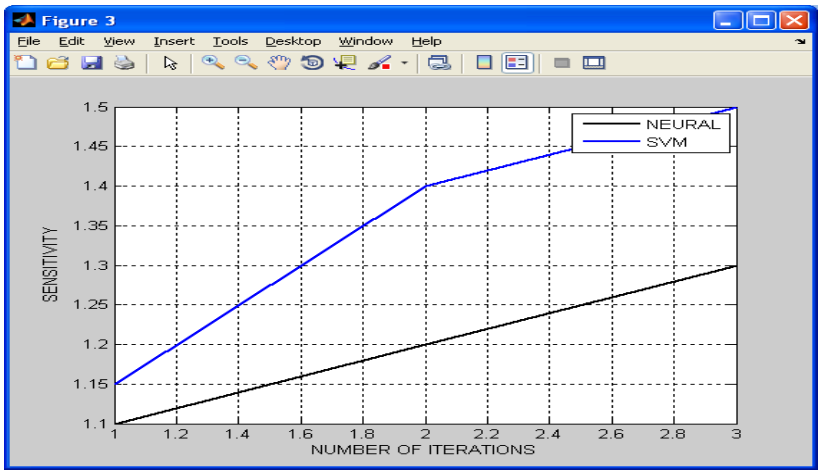

Figure 5 Sensitivity graph between WKM with NN and WKM with SVM

Specificity graph

In figure 6, Specificity for the neural is greater, as compared to SVM technique. Black line of neural in figure 6 shows that specificity is greatly enhanced in comparison to SVM. Blue line proves that the SVM results have less specificity values.

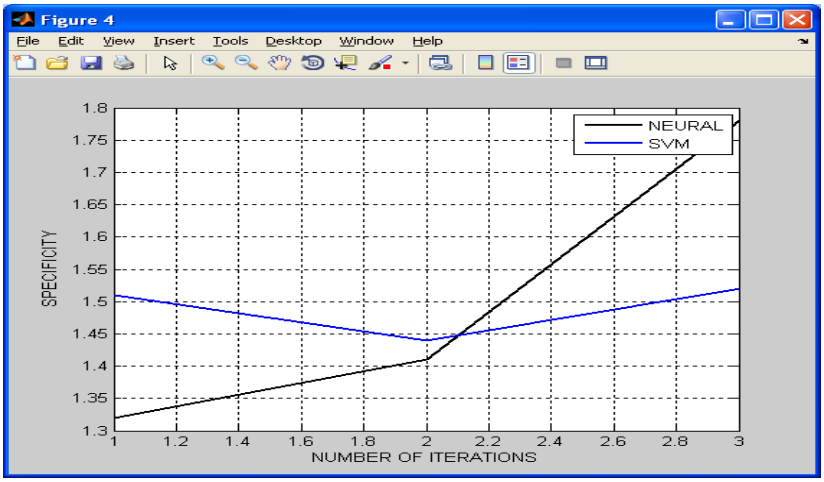

Figure 6 Specificity graph between WKM with NN and WKM with SVM

F measure graph

In figure 7, f measure value is high for NN as neural in combination with WKM makes better classification results. F measure is used for the query classification or document classification. F measure is best at its 1 value. Black line in figure 7 proves that f measure value for neural is better and blue line of SVM has comparatively less accurate results.

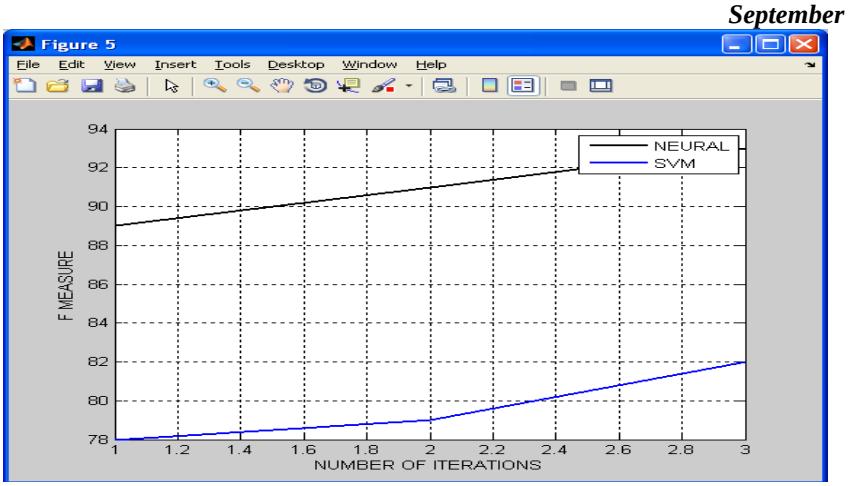

Figure 7- F measure graph between WKM with NN and WKM with SVM

C. Comparing the values of parameters in hybrid clustering and classification

Table 2 Parameters of weighted k mean and neural networks

	Weighted k mean + Neural networks				
	Ranges	Dataset1	Dataset2	Dataset3	Dataset4
Entropy	0-1(decimals)	0.1171	0.08701	0.01557	0.1197
Efficiency	0-100%	95.16	93.4293	92.4316	94.415
Accuracy	0-100%	90.99	92.51	92.981	91.90
G mean	0-1(decimals)	0.0503	0.0703	0.21253	0.1526
F measure	0-1(decimals)	0.1691	0.2113	0.1765	0.9339
Recall	0-1(decimals)	0.1303	0.1061	0.15066	0.2692
Precision	0-1(decimals)	0.9042	0.11664	0.93788	0.4752
Specificity	0-1(decimals)	0.2496	0.1161	0.2998	0.2693
Sensitivity	0-1(decimals)	0.1303	0.0426	0.15066	0.0865

Table 3 Parameters of weighted k mean and SVM

	Weighted k mean + SVM				
	Ranges	Dataset1	Dataset2	Dataset3	Dataset4
Entropy	0-1(decimals)	0.5227	0.7978	0.6227	0.4321
Efficiency	0-100%	82.318	82.21	76.168	87.227
Accuracy	0-100%	84.318	85.50	83.610	82.438
G mean	0-1(decimals)	0.17224	0.7890	0.6670	0.1677
F measure	0-1(decimals)	0.06259	0.0567	0.0646	0.16997
Recall	0-1(decimals)	0.1906	0.0144	0.0190	0.14424
Precision	0-1(decimals)	0.3181	0.4988	0.3890	0.90305
Specificity	0-1(decimals)	0.15564	0.3221	0.2337	0.9497
Sensitivity	0-1(decimals)	0.1906	0.0156	0.01903	0.14424

D. Comparing average graphs of parameters

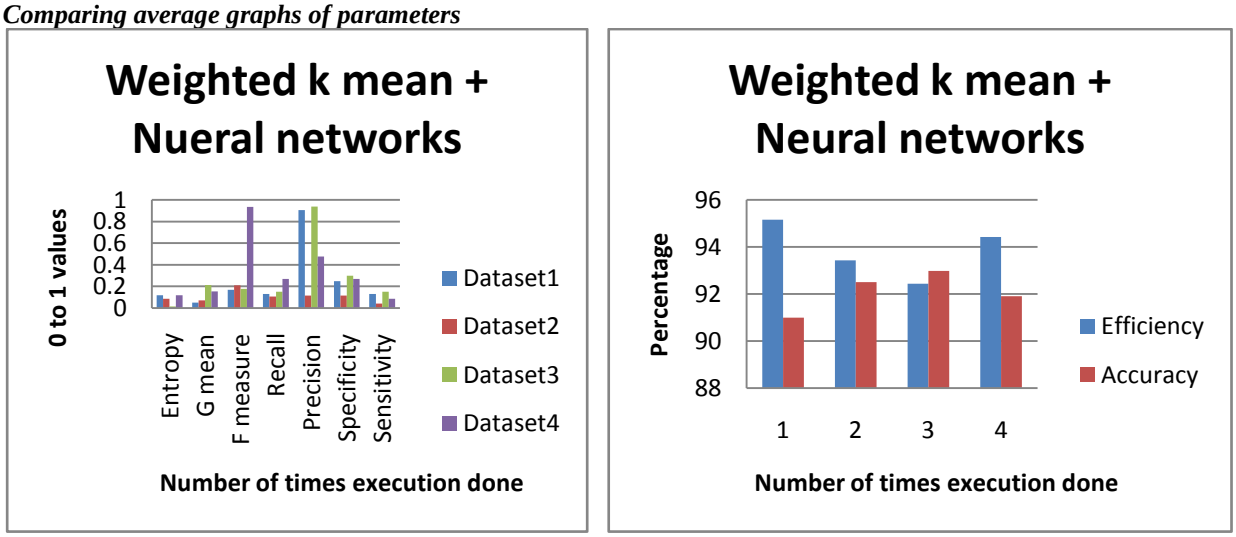

Figure 8 Average graph of parameters of weighted k mean with neural networks

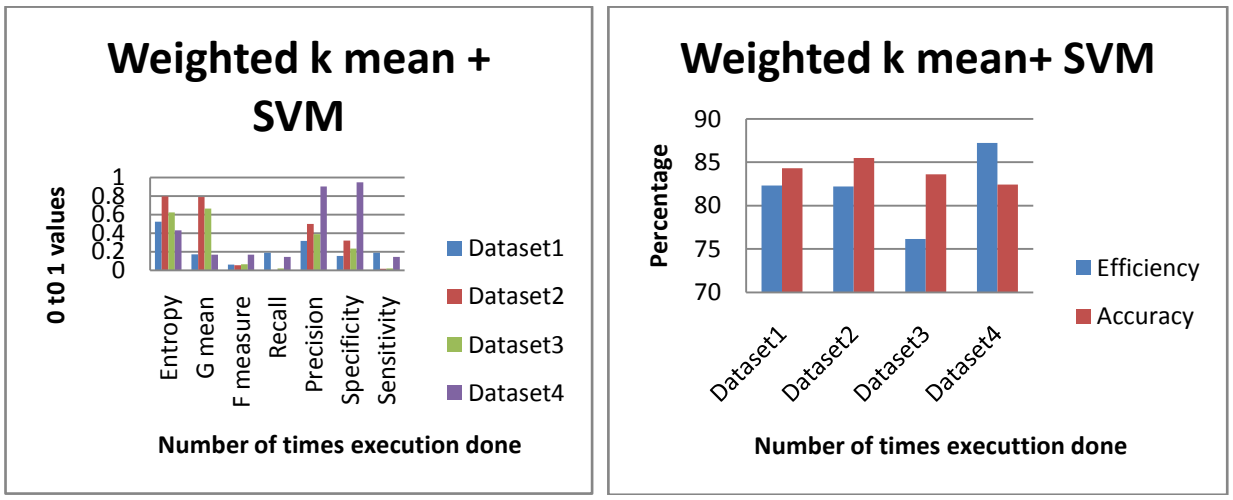

Figure 9 Average graph of parameters of weighted k mean and SVM

The Figure 8 shows that WKM with NN gives less entropy value, less G mean value, less recall value, less sensitivity value and gives greater f measure value, specificity value, precision value**,** but with more accuracy and efficiency. The Figure 9 shows that WKM with SVM gives more entropy value, less G mean value, less recall value, less sensitivity value and gives less f measure value, specificity value, precision value as compared to WKM with NN. Also it has less accuracy and less efficient results.

IV. CONCLUSION

In this paper, we have evaluated the hybrid clustering and classification technique. The evaluation us done in two phases; first is weighted k mean with neural networks; second is weighted k mean with SVM. The results in both the phases are compared with each other which show that neural classifier gives more optimized results with weighted k mean as compared to the SVM classifier because neural use many epochs to give better results. Weighted k mean with neural networks reduces the entropy, sensitivity, recall and increases effectively efficiency, accuracy, precision, specificity, f measure; which are the performance metrics for the retrieval process.

Future work-Results are depending on the data, so dependency can be reduced in future work. NN feed forward technique is used in the research work, pyramidal NN, back propagation can be used to obtain better results. This hybrid technique works on the numerical and categorical data, the future work can be done by using the time series data or spatial data.

REFERENCES

[1] Vipin Kumar, Himadri Chauhan and Dhiraj Panwar, "K-Means Clustering Approach to Analyse NSL-KDD Intrusion Detection Dataset", Vol.3, Issue-4, Sept 2013.

[2] Liping Jing,Michael k.Ng,Joshua Zhexue Huang, " An entropy weighting k-mean algorithm for subspace clustering of high dimensional sparse data*", IEEE transactions on knowledge and data engineering*,Vol.19,no.8,August 2007.

[3] Son lam Phung and Abdesselam bouzerdoum, "A pyramidal Nueral network for Visual pattern recognition",*IEEE transactions on neural networks*,vol.18,no.2,March 2007.

[4] Quan Qian, Tianhong Wang and Rui, Zhan, "Relative Network Entropy based clustering Algorithm for Intrusion detection", Vol.15, No. 1,pp.16-22, Jan,2013.

[5] Xiangjun Li and Fen Rao "An rough entropy based approach to outlier detection", *Journal computational information systems*, Vol. 8 ,pp. 10501-10508, 2012.

[6] J. Y. Liang, Z. Z. Shi., "The information entropy, rough entropy, knowledge granulation in rough set theory", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 12 (1), pp. 37 – 46, 2004.

[7] Velmurugan T., and Santhanam T., "Performance Evaluation of K-Means and Fuzzy C-Means Clustering Algorithms for Statistical Distributions of Input Data Points," *European Journal of Scientific Research*, vol. 46, no. 3, pp. 320-330,2010.

[8] Z. Deng, K. Choi, F. Chung, and S. Wang, "Enhanced Soft Subspace Clustering Integrating Within-Cluster and Between Cluster Information," *Pattern Recognition*, vol. 43, no. 3, pp. 767-781, 2010.

[9] Xindong Wu, Vipin Kumar ,J. Ross Quinlan , Joydeep Ghosh , Qiang Yang, Hiroshi Motoda , Geoffrey J. McLachlan, Angus Ng, Bing Liu,Philip S. Yu ,Zhi-Hua Zhou ,Michael Steinbach , David J. Hand ,Dan Steinberg , "Top 10 algorithms in data mining" ,*knwl inf syst*,Vol 14,pp.1-37,2008.

[10] Laura Auria and Rouslan A. Moro, "Support Vector Machines (SVM) as a Technique for Solvency Analysis",2008.

[11] Feng Wen-ge , "Application of SVM classifier in IR target recognition" ,*Physics procedia*,Vol.24.pp. 2138-2142,2012.

[12] Pablo Velarde Alvardo,Alberto f. Martinez Herra and Adalberto Iriarte Solis, "Using entropy spaces and mixture of gaussians distribution to characterize traffic anamolies*",procedia technology*,vol 13,P.no.97-108,2012.

[13] Kehar singh, Dimple malik, Naveen Sharma "Evolving limitations in k mean algorithm in data mining and their removal" *International journal of computational engineering and management*, Vol 12, April 2011.

[14] Hongyu Zhang, Xiuzhen Zhang, "Data mining static code attributes to learn defect predictors" *IEEE transactions on software engineering*, Vol 33, no. 3, September 2007.

[15] Shveta kundra Bhatia, VS dixit, "A propound method for the improvement of cluster quality" *International journal of computer science*, Vol 9, Issue 4, No. 2, July 2012.

[16] L.douglas baker, Andrew kachitus McCallum, "Distributional words for text classification".

[17] O¨. Johansson, W. Alkema, W. W. Wasserman, J. Lagergren, "Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm" *Bioinformatics.* Vol 19, 2003.

[18] Myoung-Jong Kim, "Geometric mean based boosting algorithm to resolve data imbalance problem" *The fifth international conference on advances in databases, knowledge and data applications*, 2013.

[19] Miao Wan, Arne J¨onsson, Cong Wang, Lixiang Li, "a random indexing approach for web user clustering and web prefetching".

[20] Guillemo N. Abras, Virginia L. "A weighted k mean applied to the brain tissue classification", *JCS&T*, Vol. 5, No. 3.,2006.

[21] Zorana Bankovic, Jose M. Moya, lvaro Araujo, Slobodan Bojanic and Octavio Nieto- Taladriz, "A Genetic Algorithm-based Solution for Intrusion Detection", *Journal of Information Assurance and Security*, Vol 4, P.No. 192-199,2009.

[22] Guillemo N. Abras, Virginia L. "A weighted k mean applied to the brain tissue classification", *JCS&T*, Vol. 5, No. 3.,2006.

