IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

Novel Sequence Clustering Approach for Biological Data

Sunila[1], Rishipal[2], Sanjeev Kumar[3]

PhD Scholar, Dept of CSE, Guru Jambheshwar University of Science & Technology, Hisar[1]

Prof, Dept of CSE, Guru Jambheshwar University of Science & Technology, Hisar[2]

Abstract: Clustering is one of the unsupervised learning technique in which a set of basics is separated into uniform groups. It is more hard job as compared to supervised classification where classes are already known for training the system. This dilemma becomes most awful when sequential data are to be measured. Hidden Markov Models (HMM) comprise a tool for sequential data modeling. In this paper a scheme for HMM based sequential clustering is proposed and compared with K-Means using machine learning tool WEKA. In this approach proximity based methods are used, in which the main endeavor of the clustering process is in formulating similarity or distance measures between sequences. Proposed K-Means is a useful tool for identifying co-expressed genes, biologically relevant groupings of genes and samples. Experimental results demonstrate that HMM based K-Means outperforms K-Means in terms of accuracy. But Proposed K-Means has an intense computational load.

Keywords: Data mining, Clustering, K-Means Clustering, HMM, Distance measure.

I. INTRODUCTION

Diagnostic decision support till today is a talent for physicians in their practices due to lack of quantitative tools. A medical diagnostic DSS is a computer program that contains all significant medical domain knowledge and generates a betterl diagnosis on the root of individual patient results. A medical diagnostic DSS may be tremendously helpful because it is capable to develop the convenience of expert knowledge and patient information resulting in quality improvement of the diagnostic process, increase of competence and cutback of costs [1].

Machine Learning helps to take decisions faster and with greater degree of confidence by lowering the degree of indecision in decision process. Clustering is a method of un-supervisory learning and a common technique used for biological data analysis, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. Clustering is a division of data into groups of similar objects. Each group, called a cluster, consists of objects that are similar between them- selves and dissimilar compared to objects of other groups. Clustering is extremely important technology in Data Mining. It divides the datasets into several important clusters to replicate the data sets' natural structure. Cluster is aggregation of data objects with common characteristics based on the amount of some kind of information. Clustering is more complicated job as compared to supervised classification where classes are already known for training the system.

This difficulty becomes worst when sequential biological data are to be considered. This is due to lack of ability to infer sequences of different lengths. HMMs are widely used tool for sequence modeling. Clustering of sequences is extensively used in Biological data mining and DNA genome modeling. HMMs can model protein sequences in many traditions, depending on what features of the protein are represented by the Markov states [9]. For protein structure prediction, states have been selected to signify homologous sequence positions, local or secondary structure types [10], or transmembrane locality. The consequential models can be used for prediction of common ancestry, secondary or local structure, or membrane topology by applying one of the two standard algorithms for comparing a sequence. Clustering of sequences, partitions of sequences is based on some measure of dissimilarity between members of the same cluster.

The rest of the paper is organized as follows: Section 2 reviews the previously work done in related field. Section 3 describes HMM. Section 4 describes K-Means method. Section 5 describes HMM using K-Means. Section 6 describes datasets the performance evaluation of Proposed K-Means. Conclusions are remarked in section 7.

II. RELATED WORK

Sunila Godara et al. [1] proposed a method based on modified Euclidian distance for making the K-Means algorithm more useful and proficient so as to obtain better clustering with reduced complexity for medical domains.

Yujing Zeng et al. [3] Proposed a narrative hidden Markov model (HMM) and clustering algorithm for the analysis of gene expression The, proposed model is designed to capture dynamic nature of gene expression profiles, which is ignored by many clustering techniques. In this model, gene expression dynamics are represented by a special set

DOI 10.17148/IJARCCE.2016.57151

of paths, with each path characterizing a pattern. The profile-HMM is trained to contain the most likely set of patterns given the dynamic microarray data, and the clustering result is obtained by grouping together the most related pattern. The resulting performance is confirmed on biological data sets.

Adil M. Bagirov et al. [8] developed a new version of the global k-means algorithm, the modified global k-means algorithm. Clustering algorithms based on global optimization techniques are not applicable to even relatively large data sets. Algorithms which are applicable to such data sets can locate only local minima of the function and these local minima can differ from global solutions significantly as the number of clusters increases. The number of clusters, as a rule, is not known in advance. So an incremental approach used to locate a local solution gave solution close to global one.

Smyth. et al. [9] presented an approach consists in two steps: first, it devises a pair wise distance between observed sequences, by computing a symmetries similarity. This similarity is obtained by training an HMM for each sequence, then log-likelihood (LL) of each model can be computed. This information is used to construct an LL matrix which is used to cluster the sequences in K groups, using a hierarchical algorithm. In the second step, one HMM is trained for each cluster; the resulting K models are then merged into a global HMM, where each HMM is used to design a disjoint part of this global model. As a result, a global HMM modeling all the data is obtained.

Law et al. [10] proposed an HMM-based method for sequence clustering where HMMs are used as cluster prototypes. The clustering was obtained by employing competitive learning algorithm.

Li al. et al. [11] presented an approach in which the clustering problem is addressed by spotlighting on the model selection problem, i.e. the search for the HMM topology best representing data, and the clustering structure issue, i.e. finding the number of clusters.

Li al. et al. [12] presented model selection issue with Bayesian Network. They extended further by Bayesian Model merging approach.
Panuccio, A et al. [14]. HMMs were engaged to compute similarities between sequences, using dissimilar approaches, and standard pair wise distance matrix-based approaches were then used to obtain clustering.

III. HIDDEN MARKOV MODEL (HMM)

This model can be thought of as a transition diagram with N nodes called state and edges representing transition between those states. Each state node contains initial state distribution and observation probability at which a given symbol is to be observed. An edge maintains a transition probability with which a state transition from one state to another state is made [7].

A hidden Markov model can be considered a generalization of a mixture model where the hidden variables (or latent variables), which control the mixture component to be selected for each observation, are related through a Markov process rather than independent of each other.

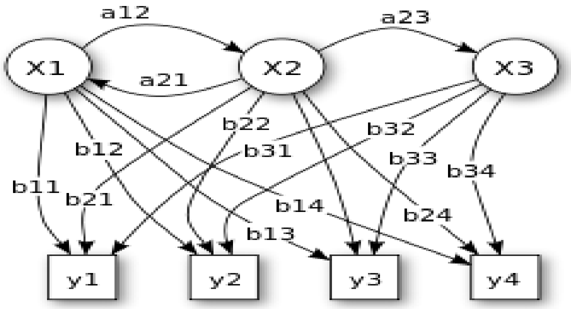

Figure 2. Probabilistic parameters of a hidden Markov model

x — states
y — possible observations
a — state transition probabilities
b — output probabilities

In its discrete form, a hidden Markov process can be visualized as a generalization of the Urn problem (where each item from the urn is returned to the original urn before the next step)[9]. Consider this example; in a room that is not visible to an observer there is a genie. The room contains urns X1, X2, X3, ... each of which contains a known mix of balls, each ball labeled y1, y2, y3, ... The genie chooses an urn in that room and randomly draws a ball from that urn. It then puts the ball onto a conveyor belt, where the observer can observe the sequence of the balls but not the sequence of urns from which they were drawn. The genie has some procedure to choose urns; the choice of the urn for the n-th ball depends only upon a random number and the choice of the urn for the (n − 1)-th ball. The choice of urn does not directly depend on the urns chosen before this single previous urn; therefore, this is called a Markov process. It can be described by Figure 2[9].

Given an input sequence {T=O1,O2 ,…….,ON}. HMM can model it with its own probability parameters using Markov process though state transition process cannot be seen outside. Once a model is built, the probability with which a given sequence is generated from the model can be evaluated the probability with which the sequence is generated from the model can be calculated by summing the probabilities of all the possible state transitions.

IV. K-MEANS CLUSTERING TECHNIQUES

A. K-Means algorithm:
Clustering technique in data mining has received a significant amount of attention from machine learning community in the last few years and become one of the fundamental research areas. Among the vast range of clustering algorithms, K-means is one of the most popular

clustering algorithms. The basic principle of the K-means algorithm is to know how different distance measure is defined. It is a critical issue for K-means users. For example, how can we select a unique distance measure method for an optimum clustering task of biological data [1].

K-means algorithm follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). The main idea is to define k centroids, one for each cluster. The simple K-means algorithm chooses the centroid randomly from the data set. The next step is to take each data belonging to a given data set and associate it to the nearest centroid. The K-means clustering partitions a data set by minimizing a sum of squares cost function. Flow chart of K-means algorithm is given below3 [1]]:

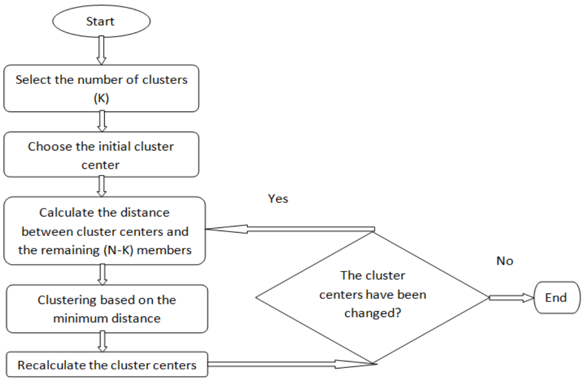
Start
Select the number of clusters (K)
Choose the initial cluster center
Calculate the distance between cluster centers and the remaining (N-K) members
Clustering based on the minimum distance
Recalculate the cluster centers
Yes
The cluster centers have been changed?
No
End

Figure 3: K-means algorithm steps.

V. PROPOSED K-MEANS USING HMM

A. Proposed K-Means using HMM

The design of the Proposed K-Means is to construct a new space, using the similarity values between sequences obtained by means of the HMMs, and to perform the clustering in that space. Flow of proposed work is given below in Fig 4.

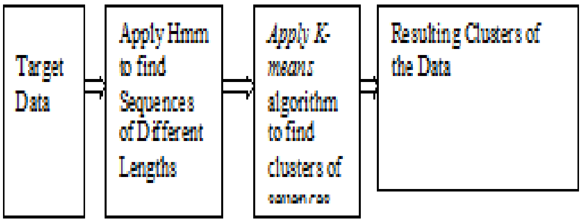
Target Data
Apply Hmm to find Sequences of Different Lengths
Apply K-means algorithm to find clusters of
Resulting Clusters of the Data

Figure 4: Flowchart of proposed approach.

In Proposed K-Means sequence clustering approach based on HMM , proximity based measure is used .In this proposed approach each sequence is represented by a vector of similarities to set of refrence sequences. Standard point clustering Kmeas method is conceded out on these sequences. After this, sequence clustering work is converted into more convenient task i.e clustering of points. Similarity values al-low discrimination, fit into the same group, and low for objects of dissimilar clusters. Consequently, we can understand the similarity measure D(O,Oi) between a sequence O and another reference sequence Oi as a feature of the sequence O. This information proposes the building of a feature vector for O by taking the similarities between O and a set of reference sequences R = {Ok}, so that O is described by a pattern i.e. {D(O,Ok), Ok € R}.

Proposed K-Means can be briefly described as follows:

Suppose N sequences {T=O1,O2 ,…….,ON} to be clustered .

R = {P1, ...,PR} be a set of R “reference” or “representative” objects. R may be subset of N.

Train one HMM for value λi for each sequence Oi

The distance DR(Oi ,Pi,) between two sequences Oi and Pi of equal length, each having k points is defined as mean distance between two.

DR(Oi ,Pi,) = Dmean(Oi ,Pi,) = 1/k $\sum_{i=0}^{k-1}$ d (Oi , Pi,)

Where d()is Euclidean distance between two points. If sequences are of unequal length than Oi is ith point of sequence O and Pi is ith point of sequence P. Distance DR(Oi ,Pi,) is minimum Euclidean distance between two hyper rectangles that bound all point in each sequence. Than DR(Oi ,Pi,) is shorter than the distance between any pair of points in sequence P and Q respectively.

Similarity measure between sequences is presented below:

$$\mathcal{D}_{\mathcal{R}}(\mathbf{O}_i) = \begin{bmatrix} \mathcal{D}(\mathbf{O}_i, \mathbf{P}_1) \\ \mathcal{D}(\mathbf{O}_i, \mathbf{P}_2) \\ \vdots \\ \mathcal{D}(\mathbf{O}_i, \mathbf{P}_R) \end{bmatrix}$$

3. Proposed distance formula is given in equation(1) below:

$$d = \mathcal{D}_{\mathcal{R}}(\mathbf{O}_i) \qquad (1)$$

4 The value of function E given in equation (2) below must be minimized.

$$\mathbf{E} = \sum_{i=1}^{k} \sum_{x \in C_i} d \qquad (2)$$

5 Likelihood values given by equation (3) must be maximized. Which signifies sequence will be placed in cluster where sequences of its maximum similarities are placed.

DOI 10.17148/IJARCCE.2016.57151

$$\mathcal{D}_{\mathcal{R}}(\mathbf{O}_i) = \frac{1}{T_i}\begin{bmatrix} \log P(\mathbf{O}_i|\boldsymbol{\lambda}_1) \\ \log P(\mathbf{O}_i|\boldsymbol{\lambda}_2) \\ \vdots \\ \log P(\mathbf{O}_i|\boldsymbol{\lambda}_R) \end{bmatrix} \quad (3)$$

Apply k means clustering method to perform clustering. Repeat above steps till error reduces to predefined level.

VI. DATASETS & RESULTS

A. Data sets Used:
Pima Indian Diabetes data set:
This data set is downloaded from archive.ics.uci.edu/ml/datasets.html and has 9 attributes and 768 instances.

Haberman data set:
This data set is downloaded from archive.ics.uci.edu/ml/datasets.html and contains 4 attributes and 306 instances.

Lympography data set:
This data set is downloaded from archive.ics.uci.edu/ml/datasets.html and contains 19 attributes and 148 instances.

Segment-test data set:
This data set is downloaded from archive.ics.uci.edu/ml/datasets.html and contains 20 attributes and 810 instances.

Liver Disorder data set:
This data set is downloaded from archive.ics.uci.edu/ml/datasets.html and contains 7 attributes and 345 instances.
Breast Cancer Wisconsin data set: This data set is downloaded from archive.ics.uci.edu/ml/datasets.html and has 11 attributes and 699 instances.

Hepatitis data set: This data set is downloaded from archive.ics.uci.edu/ml/datasets.html and contains 20 attributes and 155 instances.
Cleveland Heart data set: This data set is downloaded from archive.ics.uci.edu/ml/datasets.html and contains 14 attributes and 303 instances.

B. 6.2 Results:

WEKA HMM tool is used for this experiment.

Knowledge flow using WEKA HMM for proposed approach is shown below in Fig 5. Here data is loaded with CSV loader then it is given to cross validation. 10v fold cross validation is used for proposed work. Then it is given to HMM. Output of HMM is given to Prediction Appender. Output of Prediction appender further provides input data to Kmeans .Results of proposed K Means are shown using text viewer.

The K-Means and Proposed K-Means using HMM clustering are applied on Biological data sets and their results are compared with respect to time complexity and accuracy. With help of analysis, it is shown that K-Means using HMM has taken some more time to make cluster on Biological datasets. But it has more accuracy then K-Means.

The table1 shows the accuracy in terms of correctly clustered instances by the above algorithms to make clusters. The given figure no.5 shows respective results in case of accuracy. Fig 6 shows graphical representation of accuracy on various data sets. The table2 shows the time taken by the above algorithms to make clusters. The given figure no.7 shows respective results in case of time complexity.

Finally, the generated results by Proposed K-Means using HMM outperforms K-Means in terms of, accuracy. But takes some more time to build clusters .Further it is observed that performance of Proposed K-Means using HMM depends on dimensions of data sets. As dimensionality increases Proposed Approach takes more time and has less enhancement in accuracy. So proposed Approach suffer from dimensionality. By using appropriate dimensionality reduction techniques we can enhance performance of proposed approach.

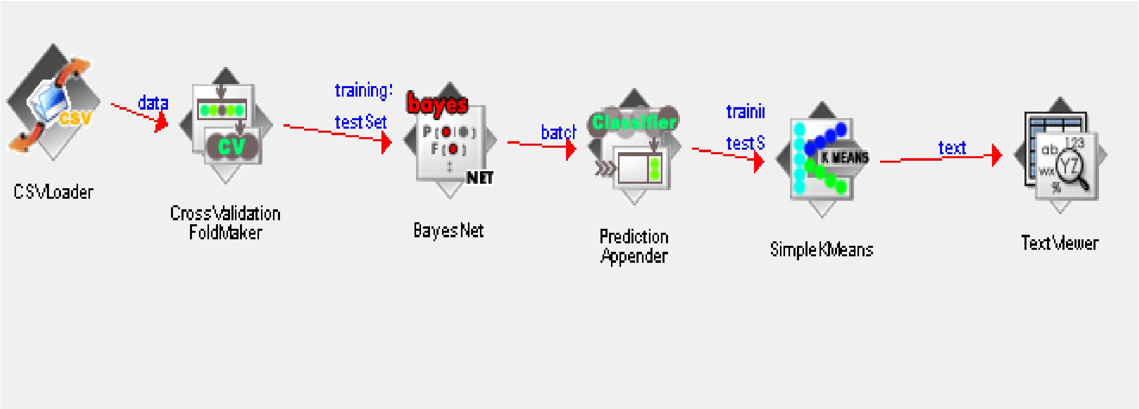

Figure 5: Knowledge Flow Diagram using WEKA HMM for proposed approach

TABLE 1. ACCURACY OF VARIOUS CLUSTERING ALGORITHMS

	Accuracy for Pima Indian diabetes data set	Accuracy for Haberman data set	Accuracy for Lymphography data set	Accuracy for segment-test data set	Accuracy for Cleveland Dataset	Accuracy of Breast Cancer Dataset	Accuracy for Hepatitis Dataset	Accuracy for Liver Disorder data set
K-Means	64.8	54.11	53.26	59.26	72.18	80.97	70.1	46.3
Proposed K-Means using HMM	66.55	58.72	54.41	60.71	73.12	81.5	71.5	50.1

TABLE 2 TIME COMPLEXITY OF VARIOUS CLUSTERING ALGORITHMS FOR GIVEN DATA SETS

	Time for Pima Indian diabetes data set	Time for Haberman data sets	Time taken for Lymphography data set	Time taken for segment-test data set	Time taken for cleveland Dataset	Time taken for Breast Cancer Dataset	Time taken for Hepatitis Dataset	Time taken for Liver Disorder data set
K-Means	0.11	0.02	0.03	0.27	0.052	0.75	0.05	0.025
Proposed K-Means using HMM	0.16	0.022	0.11	0.39	0.072	0.88	0.15	0.011

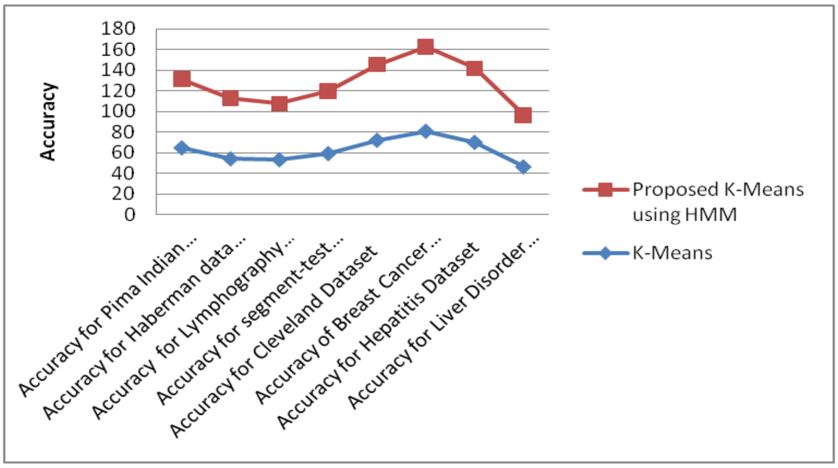

Figure 6: shows respective results in terms of accuracy.

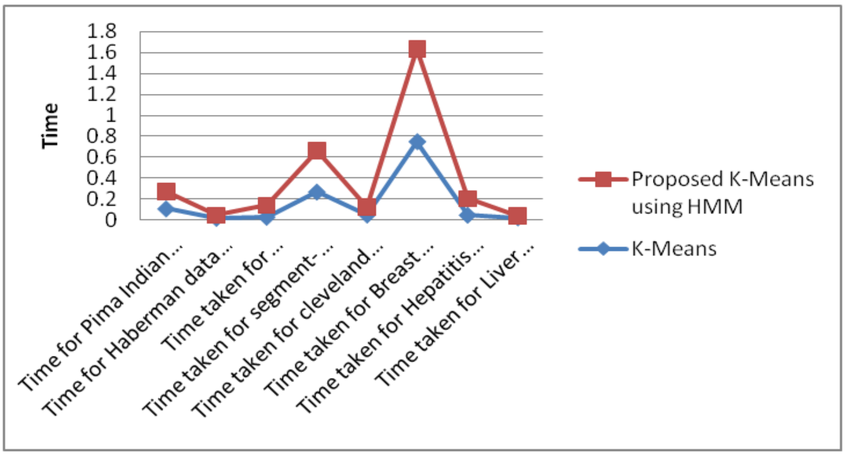

Figure 7: shows respective results in terms of time complexity

VII. CONCLUSION

In this paper, an approach for sequence clustering based on HMM using proximity based measure was proposed. The performance of proposed algorithm is tested across eight real world datasets using WEKA HMM data mining tool and the results are quite encouraging and have established the effectiveness of the proposed algorithm. The proposed work can also be explored by use of various dimensionality reduction algorithms for data preprocessing. Which will not only improve its cluster accuracy but also its efficiency.

REFERENCES

1. Sunila Godara , Rishipal Singh, “An Efficient method to Improve Performance of K- Means Clustering Algorithms for Medical Domains”, 7th International conference on Advanced Computing and Communication Technologies, APIIT SD, India Panipat, Nov 16, 2013.
2. Sunila Godara, Amita Verma, “Analysis of Various Clustering Algorithms”, International Journal of Innovative Technology and Exploring Engineering, Vol 3, No 1, June 2013.
3. Yujing Zeng ,Javier Garcia-Frias,” A novel HMM-based clustering algorithm for the analysis of gene expression time-course data”, Computational Statistics & Data Analysis,Volume 50, Issue 9, 1 May 2006, Pages 2472–2494
4. A.K. Jain, M.N. Murty, P.J. Flynn, “Data Clustering: A Review”, ACM Computing Surveys, vol. 31, no. 3, pp. 264-323,Sep. 1999.
5. D.Napoleon,S.Pavalakodi,”A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Set,” International Journal of Computer Applications (0975– 8887),vol. 13, no.7, pp.41-46, Jan 2011.
6. Mu-Chun Su and Chien-Hsing Chou, “A Modified Version of the K-Means Algorithm with a Distance Based on Cluster Symmetry,”IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp.674-680, Jun. 2001.
7. Nivedita Rao,Sunila Godara,”Modelling of Protins Sub –Cellular Sites using Hidden Markov Model: A Review;” ;” International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol 4, No 8,pp308-314,June 2014.
8. Adil M. Bagirov Karim Mardaneh,” Modified global k-means algorithm for clustering in gene expression data sets”, Workshop on Intelligent Systems for Bioinformatics (WISB2006), Hobart, Australia,vol 73,2006.
9. Smyth, P.,”Clustering sequences with hiddenMarkov models. InMozer,M., Jordan,M., Petsche, T., eds.:” Advances in Neural Information Processing. Volume 9., MIT press 1997.
10. Law, M., Kwok, J.: Rival penalized competitive learning for model-based sequence.In: Proc. Int. Conf. Pattern Recognition. Volume 2. (2000) 195–198
11. Li, C.: A Bayesian Approach to Temporal Data Clustering using Hidden Markov Model Methodology. PhD thesis, Vanderbilt University 2000.
12. Li, C., Biswas, G.: Clustering sequence data using hidden Markov model representation.In: Proc. of SPIE’99 Conf. on Data Mining and Knowledge Discovery: Theory, Tools, and Technology. (1999) 14–21
13. TapasKanungo,David M.Mount ,NathanS. Netanyahu, Christine D.Piatko, Ruth Silverman, and Angela Y.Wu,” An Efficientk-Mean Clustering Algorithm: Analysisan Implementation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881-891, Jul 2002.
14. Panuccio, A., Bicego, M., Murino, V.: A Hidden Markov Model-based approach to sequential data clustering. In Caelli,
15. T Amin, A., Duin, R., Kamel, M., de Ridder,D., eds.: Structural, Syntactic and Statistical Pattern Recognition. LNCS 2396,Springer (2002) 734–742
16. Ian Davidson S. S. Ravi” Using Instance-Level Constraints in Agglomerative Hierarchical Clustering: Theoretical and Empirical Results,” International Journal of Data Mining and Knowledge Discovery”, Springer, vol 18,257-282, June 2008.

BIOGRAPHIES

Sunila Godara received MSc degree in Computer Science & Engg from Guru Jambheshwar University of Science & Technology, HISAR. Nowadays she is an Assistant Professor in Deptt of Computer Sc. & Engg, Guru Jambheshwar University of Science & Technology, HISAR. Her research areas are Data Mining and machine learning.

Rishipal Singh received his Ph.D degree from J.N.U Delhi. Nowadays he is a Professor in Guru Jambheshwar University of Science & Technology, HISAR His major research interests are in high speed communications, wireless sensors networks and machine learning.

DOI 10.17148/IJARCCE.2016.57151