Clustering of Concept Drift Categorical Data Using Our-NIR Method

S Viswanadha Raju, N Sudhakar Reddy, K V N Sunitha, H Venkateswara Reddy, and G Sreenivasulu

Abstract—**In the clustering using Ming-Syan Chen NIR method has deficiency that is importance of data labeling and outlier detection. The Our-NIR method introduced to improve Ming-Syan Chen method. In this paper the newly introduced method is taken for comparison to improve the cluster efficiency. To improve the efficiency of clustering by the sampling techniques. However, with sampling applied, those sampled points that are not having their labels after the normal process. Even though there is straight forward method for numerical domain and categorical data. But still it has a problem that is how to allocate those unlabeled data points into appropriate clusters in efficient manner. In this paper the concept-drift phenomenon is studied, and we first propose an adaptive threshold for outlier detection, which is a playing vital role detection of cluster. Second, probabilistic approaches for detection of cluster are proposed using Our-NIR method.**

Index Terms—**Clustering, concept-drift, threshold, weather prediction**

I. Introduction

Extracting Knowledge from large amount of data is difficult which is known as data mining. Clustering is a collection of similar objects from a given data set and objects in different collection are dissimilar. Most of the algorithms developed for numerical data may be easy, but not in Categorical data [1], [2], [3], [4]. It is challenging in categorical domain, where the distance between data points is not defined. It is also not easy to find out the class label of unknown data point in categorical domain. Sampling techniques improve the speed of clustering and we consider the data points that are not sampled to allocate into proper clusters. The data which depends on time called time evolving data. For example, the buying preferences of customers may change with time, depending on the current day of the week, availability of alternatives, discounting rate etc. Since data evolve with time, the underlying clusters may also change based on time by the data drifting concept [5]. The clustering time-evolving data in the numerical domain [1], [6], [7], [8] has been explored in the previous works, where as in categorical domain not that much. Still it is a challenging problem in the categorical domain.

As a result, our contribution in modifying the frame work which is proposed by Ming-Syan Chen in 2009[9] utilizes any clustering algorithm to detect the drifting concepts. We adopted sliding window technique [10] and initial data (at time t=0) is used in initial clustering. These clusters are represented by using Our-NIR [11], where each attribute value importance is measured. We find whether the data points in the next sliding window (current sliding window) belongs to appropriate clusters of last clustering results or they are outliers. We call this clustering result as a temporal and compare with last clustering result to drift the data points or not. If the concept drift is not detected to update the Our-NIR otherwise dump attribute value based on importance and then reclustering using clustering techniques.

The rest of the paper is organized as follows. In section II discussed related work, in section III basic notations and concept drift, in section IV new methods for decision function discussed and also contains results with comparison of Ming-Syan Chen method and our method, in section V discussed distribution of clustering and finally concluded with section VI.

Manuscript received March 20, 2011; revised June 12, 2011.

S.Viswanadha Raju is with Department of Computer Science and Engineering JNT University, Hyderabad. INDIA.

II. Related Work

In this section, we discuss various clustering algorithms on categorical data with cluster representatives and data labeling. We studied many data clustering algorithms with time evolving. Cluster representative is used to summarize and characterize the clustering result, which is not fully discussed in categorical domain unlike numerical domain. In K-modes which is an extension of K-means algorithm in categorical domain a cluster is represented by 'mode' which is composed by the most frequent attribute value in each attribute domain in that cluster. Although this cluster representative is simple, only use one attribute value in each attribute domain to represent a cluster is questionable. It composed of the attribute values with high co-occurrence. In the statistical categorical clustering algorithms [12], [13] such as COOLCAT and LIMBO, data points are grouped based on the statistics. In algorithm COOLCAT, data points are separated in such a way that the expected entropy of the whole arrangements is minimized. In algorithm LIMBO, the information bottleneck method is applied to minimize the information lost which resulted from summarizing data points into clusters.

However, all of the above categorical clustering algorithms focus on performing clustering on the entire dataset and do not consider the time-evolving trends and also the clustering representatives in these algorithms are not clearly defined.

The new method is related to the idea of conceptual clustering [14], which creates a conceptual structure to represent a concept (cluster) during clustering. However, NIR only analyzes the conceptual structure and does not perform clustering, i.e., there is no objective function such as category utility (CU) [15] in conceptual clustering to lead the clustering procedure. In this aspect our method can provide in

better manner for the clustering of data points on time based.

The main reason is that in concept drifting scenarios, geometrically close items in the conventional vector space might belong to different classes. This is because of a concept change (drift) that occurred at some time point.

Our previous work [16] addresses the node importance in the categorical data with the help of sliding window. That is new approach to the best of our knowledge that proposes these advanced techniques for concept drift detection and clustering of data points. In this regard the concept drifts handling by the headings such as node importance and resemblance. In this paper, the main objective of the idea of representing the clusters by above headings. This representation is more efficient than using the representative points.

After scanning the literature, it is clear that clustering categorical data is un touched many ties due to the complexity involved in it. A time-evolving categorical data is to be clustered within the due course hence clustering data can be viewed as follows: there are a series of categorical data points D is given, where each data point is a vector of q attribute values, i.e., $p_j=(p_j^1,p_j^2,...,p_j^q)$. And $A = \{A_1, A_2,..., A_q\}$, where A_a is the a^{th} categorical attribute, $1 \leq a \leq q$. The window size N is to be given so that the data set D is separated into several continuous subsets S^t, where the number of data points in each S^t is N. The superscript number t is the identification number of the sliding window and t is also called time stamp. Here in we consider the first *N data points of data set D* this makes the first data slide or the first *sliding window* S^0. C^i_j or *Cij* is representing for the cluster, in this the j indication of the cluster number respect to sliding window *i.* Our intension is to cluster every data slide and relate the clusters of every data slide with previous clusters formed by the previous data slides. Several notations and representations are used in our work to ease the process of presentation:

III. Concept Drift Detection

Concept drift is an sudden substitution of one sliding window S1(with an underlying probability distribution ΠS1),with another sliding window S2 (with distribution ΠS2).As concept drift is assumed to be unpredictable, periodic seasonality is usually not considered as a concept drift problem. As an exception, if seasonality is not known with certainty, it might be regarded as a concept drift problem. The core assumption, when dealing with the concept drift problem, is uncertainty about the future - we assume that the source of the target instance is not known with certainty. For successful automatic clustering data points we are not only looking for fast and accurate clustering algorithms, but also for complete methodologies that can detect and quickly adapt to time varying concepts. This problem is usually called "concept drift" and describes the change of concept of a target class with the passing of time.

As said earlier in this section that means detects the difference of cluster distribution between the current data subset S^t (i.e. sliding window 2)and the last clustering result $C^{[tr,t-1]}$ (sliding window 1)and to decide whether the resulting is required or not in S^t. Hence the upcoming data points in the slide S^t should be able to be allocated into the corresponding proper cluster at the last clustering result. Such process of allocating the data points to the proper cluster is named as "labeled data". Labeled data in our work even detects the outlier data points as few data points may not be assigned to the cluster, "outlier detection".

If the comparison between the last clusters (C11 and C12)and the temporal clusters availed from the new sliding window data labeling in fig 2, produce the enough differences in the cluster distributions, then the latest sliding window is considered as a concept-drifting window as per the equation 3. A re-clustering is done on the latest sliding window. This includes the consideration of the outliers that are obtained in the latest sliding window, and forming new clusters which are the new concepts that help in the new decisions. The above process can be handled by the following headings such Node selection, Our-NIR, Resemble method and threshold value. This is new scenario because of we introduced the Our-NIR method compared with existing method and also published in [16]

A. Node Selection

In this category, proposed systems try to select the most appropriate set of past cases in order to make future clustering. The work related to representatives of the categorical data with sliding window technique based on time. In sliding window technique, older points are useless for clustering of new data and therefore, adapting to concept drift is synonym to successfully forgetting old instances knowledge. Examples of this group can be found in [17], [18], [19].

B. Node Importance

In this group, we assume that old knowledge becomes less important as time goes by. All data points are taken under consideration for building clusters, but this time, new coming points have larger effect in the model than the older ones. To achieve this goal, we introduced a new weighting scheme for the finding of the node importance with probability and also published in [10], [20].

C. Resemblance Method

The main aim of this method is to have a number of clusters that are effective only on a certain concept. It has importance that is to find label for unlabeled data points and store into appropriate cluster.

1) Maximal resemblance

All the weights associated with a single data point corresponding to the unique cluster forms the resemblance. This can be given with the equation:

$$R(P_j,C_i)= \sum_{r=1}^{q} W(Ci,N_{[i,\,r]}) \quad \rightarrow \quad (1)$$

Here a data point P_j of the new sliding window and the Our-NIR of the data point with all the clusters are calculated and are placed in the table. Hence resemblance $R(P_j,C_i)$ can be obtained by summing up the Our-NIR of the cluster c_i. This just gives the measurement of the resemblance of the node with cluster. And now these measurements are used to find the maximal resemblance. i.e, if data point P_j has maximum resemblance R (P_j,C_i),towards a cluster C_i, then the data point is labeled to that cluster.

If any data point is not similar or has any resemblance to any of the cluster then that data point is considered to be the

outlier. We even introduce the threshold to simplify the outlier detection. With the threshold value the data points with small resemblance towards many clusters can be considered as the outlier if the resemblance is less than the threshold.

IV. Value of Threshold

In this section, we introduce the decision function that is to find out the threshold, which decides the quality of the cluster and the number of the clusters. Here we have to calculate the threshold (λ) for every cluster can be set identical, i.e., $\lambda_1=\lambda_2=\ldots=\lambda_n=\lambda$. Even then we have a problem to find the main λ(threshold) that can be found comparing with all the clusters. Hence an intermediate solution is chosen to identify the threshold (λ_i) the smallest resemblance value of the last clustering result is used as the new threshold for the new clustering. After data labeling we obtain clustering results which are compared to the clusters formed at the last clustering result which are base for the formation of the new clusters. This leads to the "Cluster Distribution Comparison" step.

A. Labeling and Outlier Detection Using Adaptive Threshold

The data point is identified as an outlier if it is outside the radius of all the data points in the resemblance methods. Therefore, if the data point is outside the cluster of a data point, but very close to its cluster, it will still be an outlier. However, this case might be frequent due to concept- drift or noise, As a result, detecting existing clusters as novel would be high. In order to solve this problem, here we adapted the threshold for detecting the outliers/labeling. The most important step in the detection of the drift in the concept starts at the data labeling. The concept formation from the raw data which is used for the decision making is to be perfect to produce proper results after the decision; hence the formation of clustering with the incoming data points is an important step for the comparison of the incoming data point with the initial clusters generated with the previous available data gives rise to the new clusters.

If a data point P_j is the next incoming data point in the current sliding window, this data point is checked with the initial cluster C_i, for doing so the resemblance R (C_i, P_j) is measured, and the appropriate cluster is the cluster to which the data point has the maximum similarity or resemblance. Our-NIR is used to measure the resemblance. Maximal Resemblance was discussed in C.1 section**.**

$$Label = \begin{cases} C_i^*, & if \max R(p_j, c_i) \geq \lambda_i, where \quad 1 \leq i \leq k, \\ outliers, & otherwise. \end{cases} \qquad (2)$$

Example 1: Consider the data set in fig 1 and the Our-NIR of C^1 in fig 2 now performing the labeling based on second sliding window data points and the thresholds $\lambda_{1=}\ \lambda_2=1.83$ and the first data point P7 = {A, K, D} in S^2 is decomposed into three nodes they are { $[A_1$ = A], $[A_2$=K],$[A_3$=D]} the resemblance of P7 is C_1^1 is 1.33 and in C_2^1 is zero. Since the maximal resemblance is less than or equal to threshold λ_1, so the data point is considered in outlier**.** The next data point of current sliding window P8 {Y, K, P} is C_1^1 is zero and in C_2^1 is 1.33 and the maximal resemblance value is less than or equal to threshold λ_2, so the data point is considered in outlier. Similarly for the remaining data points in the current sliding window that are p9 and P10 are in outlier, P11 in C_1^1 and P12 in C_1^2. All these values shown in figure 3 are temporal clusters. Here the ratio of number of outliers is 4/6 > 0.5 there the concept drift occurred in this regard need to apply reclustering that is shown in fig. 4.

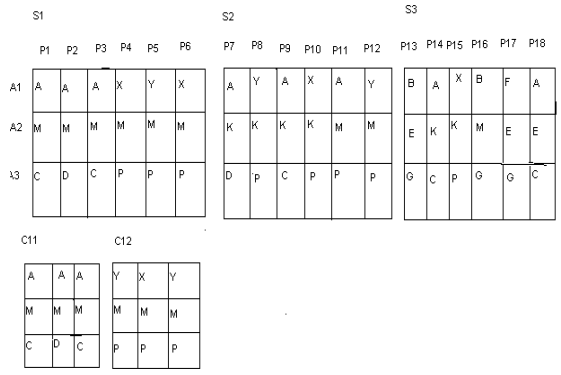

Fig. 1. Data set with sliding window size 6 where the initial clustering is performed

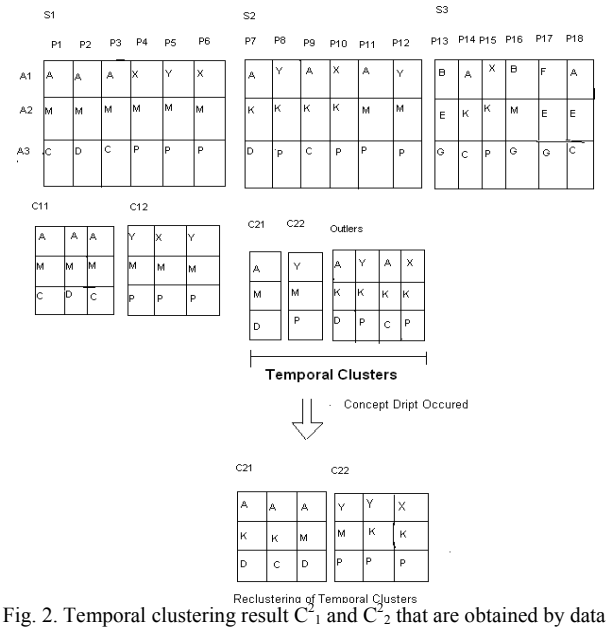

Fig. 2. Temporal clustering result C^2_1 and C^2_2 that are obtained by data labeling

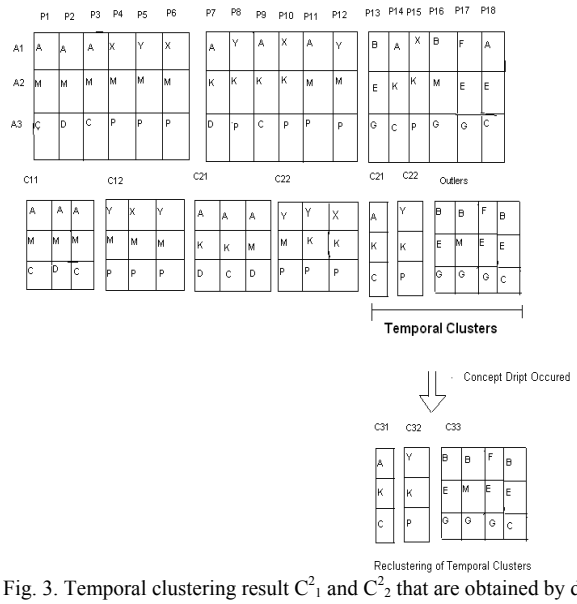

Fig. 3. Temporal clustering result C^2_1 and C^2_2 that are obtained by data labeling

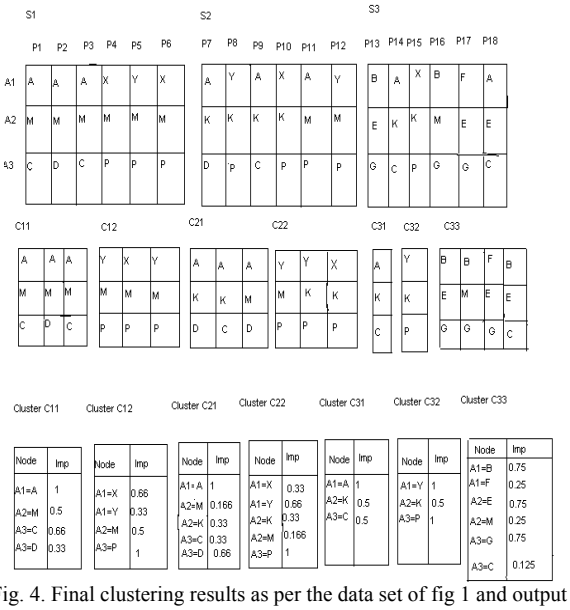

Fig. 4. Final clustering results as per the data set of fig 1 and output Our-NIR Results

The decision making here is difficult because of the calculating values for all the thresholds the simplest solution to fix the constant identical threshold to all the clusters. However it is difficult still, to define a single value threshold that is applied on all clusters to determine the data point label. Due to this we use the data points in last sliding window that construct the last clustering result to decide the threshold

V. Cluster Distribution Comparison

To detect the concept drift by comparing the last clustering result and current clustering result obtained by data points. The clustering results are said to be different according to the following two criteria's:

1. The clustering results are different if quite a large number of outliers are found by the data labeling.
2. The clustering results are different if quite a large number of clusters are varied in the ratio of data points.

In the previous section outliers detected during the data labeling/outlier detection ,but there may be many outliers which are not able to be allocated to any of the cluster, that means the existing concepts are not applicable to these data points. But these outliers may carry a concept within themselves this gives the idea of generating new clusters on the base of the number of the outliers formed at the latest clustering. In this work we considered two types of measures such outlier threshold and cluster difference threshold.

Here we introduced the outlier threshold that is OUTTH can be set so as to avoid the loss of existing concepts. If the numbers of outlier are less it can restricts the re-clustering by the OUTTH otherwise re-clustering can be done. If the ratio of outliers in the current sliding window is larger than OUTTH then the clustering results are said to be different and re-clustering is to be performed on the new sliding window. The ratio of the data points in a cluster may change very drastically following a concept drift, this is another type of concept drift detection. The difference of the data points in an existing cluster and new temporal cluster is high that indicates the drastic loss in the concept of the cluster, this can be disastrous when it comes to the decision making with new clusters available. Hence *cluster variance threshold (ϵ)* is introduced which can check the amount of variation in the cluster data points, finally it helps to find the proper cluster. The cluster that exceeds the cluster variation threshold is seen as a different cluster and then the count the number different clusters that number compared with other threshold named cluster difference threshold. It the ratio of the different cluster is large than the cluster difference threshold the concept is said to be drift in the current sliding window. The cluster process as shown in equation (3)

$$\begin{cases} yes, \quad if \ \dfrac{\#\ outliers}{N} > \theta \\ yes, \quad if \left\{ \dfrac{\sum_{i=1}^{k^{[t_e,t-1]}} d\left(c_i^{[t_e,t-1]}, cc_i^{'t}\right)}{k^{[t_e,t-1]}} > \eta \right\}, \\ Where\ d\left(c_i^{[t_e,t-1]}, cc_i^{'t}\right) \\ \begin{cases} 1, if \left| \dfrac{m_i^{[t_e,t-1]}}{\sum_{x=1}^{k^{[t_e,t-1]}} m_x^{[t_e,t-1]}} - \dfrac{m_i^t}{\sum_{x=1}^{k^{[t_e,t-1]}} m_x^t} \right| > \epsilon \\ 0, \text{otherwise} \end{cases} \\ \text{No, otherwise} \end{cases} \tag{3}$$

Example 2: Consider the example shown in fig 2. The last clustering result C^1 and current temporal clustering result C_1^2 is compared with each other by the equation (3). Let us take the threshold OUTH is 0.4, the cluster variation threshold (ϵ) point is 0.3 and the cluster threshold difference is set to 0.5. In fig 2 there are 4 outliers in C_1^2 , and the ratio of outliers in S^2 is 4/6=6.66>OUTH, so that the s^2 is considered as concept drift and is going to be reclustering.

Example 3: Suppose the result of performing reclustering on S^2 and data labeling on S^3 is shown in fig 3. The equation (3) is applied on last clustering result C^2 and current temporal clustering result C_1^3 . There is four outliers in C_1^3 , and the ratio of outliers in S^3 is 4/6<=0.4 however the ratio of the data points between clusters are satisfied as per the condition given in equation (3) and the ratio of different clusters are also satisfied so therefore the S3 is considered as concept drift occurred. Finally, reclustered the temporal clusters and updated Our-NIR shown in fig. 3.

If the current sliding window t considered that the drifting concept happens, the data points in the current sliding window t will perform re-clustering. On the contrary, the current temporal clustering result is added into the last clustering result is added into the last clustering result and the clustering representative Our-NIR is updated.

If the current sliding window t considered that the drifting concept happens, the re-clustering process will be performed. The last clustering result $C^{[t_e,t-1]}$ represented in Our-NIR is first dumped out with time stamp to show a steady clustering result that is generated by a stable concept from the last concept-drifting time stamp t_c to t-1. After that, the data points in the current sliding window t will perform re-clustering, where the initial clustering algorithm is applied. The new clustering result C^t is also analyzed and represented by Our-NIR. And finally, the data points in the next sliding window S^2 and the clustering result C^t are input to do the

DCD algorithm. If the current sliding window t considered that the stable concept remained, the current temporal clustering result C^t that is obtained from data labeling will be added into the last clustering result $C^{[te,t-1]}$ in order to fine-tune the current concept. In addition, the clustering representative Our-NIR is also needed to be updated. For the reason of quickly updating the process, not only the importance but also the counts of each node in each cluster are recorded. Therefore, the count of the same node in $C^{[te,t-1]}$ and in C^{1t} is able to be summed directly, and the importance of each node in each of the merged clusters can be efficiently calculated by node importance.

A. Time Complexity of DCD

All the clustering results C are represented by Our-NIR, which contains all the pairs of nodes and node importance. inverted file structure and hashing for better execution efficiency, among these two we chosen the hashing can be applied on the represented table, and the operation on querying the node importance have a time complexity of O(1). Therefore the resemblance value of the specific cluster is computed efficiently in data labeling shown in algorithm 1 by the sum of the each node importance through looking up the Our-NIR hash table only q times and the entire time complexity of data labeling is O(q×k×N). DCD may occur on the reclustering step when the concept drifts on the updating Our-NIR result step when the concept does not drift. When updating the NIR results. We need to scan the entire data hash table for the calculate their importance reclustering performed on S^t. the time complexity of most clustering algorithms is $O(N^2)$.

VI. Conclusion

In this paper, a frame work proposed by Ming-Syan Chen in 2009[9] which is modified by new method that is Our-NIR to find node importance. We analyzed by taking same example in this we find the differences in the node importance values of attributes in same cluster which plays an important role in clustering. The representative of the clusters help improving the cluster accuracy and purity and hence the Our-NIR method performs better than the CNIR method. In this aspect the class label of uncluttered data point and therefore the result demonstrates that our method is accurate. The future work cluster distribution based on Pour-NIR method, cluster relationship based on the vector representation model and also it improves the performance of precision and recall of DCD by introducing the leader-sub leader algorithm for reclustreing.

Acknowledgment

We would like to thank Director of SIT, JNTUHyderabad, INDIA for his support and encouragement. We also would like to thank our teacher Dr. Vinay Babu for his advice of clearing the doubts. Last but not the least we thank to our friends who helped us in doing this work.

References

[1] C. Aggarwal, J. Han, J. Wang, and P. Yu, "A Framework forClustering Evolving Data Streams," *Proc. 29th Int'l Conf.Very Large Data Bases (VLDB)* ,2003.

[2] C. C. Aggarwal, J.L. Wolf, P. S. Yu, C. Procopius, and J. S. Park, "Fast Algorithms for Projected Clustering," *Proc. ACM SIGMOD"* pp. 61-72, *1999*.

[3] A. J. M. Murthy and P. J. Flyn "Data Clustering: A Review," *ACM Computing Survey,* 1999

[4] O. Narsoui and C. Rojas, "Robust Clustering for Tracking Noisy Evolving Data Streams" SIAM I*nt. Conference Data Mining,* 2006.

[5] G. Hulton and Spencer, "Mining Time-Changing Data Streams" *Proc. ACM SIGKDD,* 2001.

[6] F. Cao, M. Ester, W. Qian, and A. Zhou, "Density-Based Clustering over an Evolving Data Stream with Noise," *Proc Sixth SIAM Int'l Conf. Data Mining (SDM),* 2006.

[7] D. Chakrabarti, R. Kumar, and A. Tomkins, "Evolutionary Clustering," *Proc. ACM SIGKDD"* pp. 554-560, *2006*.

[8] M. Gaber and P. S Yu "Detection and Classification of Changes inEvolving Data Streams," *International. Journal of Information Technology and Decision Making,* v5 no 4, 2006.

[9] H.-L. Chen, M.-S. Chen, and S-U Chen Lin "Frame work for clustering Concept –Drifting categorical data," *IEEE Transaction Knowledge and Data Engineering v21 no 5,* 2009.

[10] C. E.Shannon,"A Mathematical Theory of Communication, " Bell System Technical J., 1948.

[11] S. Viswanadha Raju, H.Venkateswara Reddy, and N. Sudhakar Reddy" Our-NIR Node Importance Representation of Clustering Categorical Data," IJCST July 2011.

[12] P. Andritsos, P. Tsaparas, R. J. Miller, and K.C. Sevcik, Limbo: Scalable Clustering of Categorical Data," *Proc Ninth Int'l Conf. Extending Database Technology (EDBT)* 2004.

[13] D. Barbará, Y. Li, and J. Couto, "Coolcat: An Entropy-Based Clustering," *Proc. ACM Int'l Conf. Information and Knowledge Management (CIKM),* 2002.

[14] D. H. Fisher, "Knowledge Acquisition via Incremental Conceptual Clustering," Machine Learning, 1987.

[15] M. A. Gluck a n d J . E . Corter, "Information Uncertainty and the Utility of Categories," Proc. Seventh Ann. Conf. Cognitive Science Soc., pp. 283- 287, 1985.

[16] S. Viswanadha Raju, H. Venkateswara Reddy and N. Sudhakar Reddy," A Threshold for clustering Concept – Drifting Categorical Data", IEEE Computer Society, ICMLC 2011.

[17] Fan, W. Systematic data selection to mine concept-drifting data streams. in Tenth ACM SIGKDD international conference on Knowledge Discovery and Data Mining. Seattle, WA, USA: ACM Press: p. 128-137, 2004..

[18] R. Klinkenberg, *Learning Drifting Concepts: Example Selection vs. Exam- ple Weighting* Intelligent Data Analysis, Special Issue on Incremental Learning Systems Capable of Dealing with Concept Drift, 8(3): p. 281, 2004.

[19] H.-L. Chen, K.-T. Chuang, and M.-S. Chen, "Labeling Unclustered Categorical Data into Clusters Based on the Important Attribute Values," *Proc. Fifth IEEE Int'l Conf.Data Mining (ICDM),* 2005.

[20] S. Viswanadha Raju, N.Sudhakar Reddy and H.Venkateswara Reddy," POur-NIR: Node Importance Representation of Clustering Categorical Data", IJCSIS. 2011.

Dr.S.Viswanadha Raju being M.C.A from Osmania University, Hyderabad, INDIA, M.Tech. in Computer Science and Engineering from Jawaharlal Nehru Technological University, Hyderabad, INDIA was awarded Ph.D. in Computer Science and Engineering by Acharya Nagarjuna University, Guntur INDIA. Presently he is working as Professor of CSE, SIT, JNT University, Hyderabad, Kukatpally, Hyderabad. He was Head, Dept. of Computer Science and Engineering JNT University, Hyderabad from Sept.2010 to Jan. 2011. Director, MCA (Accredited by NBA), Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, from July 2009 to July 2010 and also HOD of CSE/MCA from July 2003 to April 2009. Convener, International Conference on Advanced Computing Technologies (ICACT 2008) at GRIET, Hyderabad. Member Governing body in GRIET from April 2009 to July 2010. He has attended many national international Conferences/ seminars and presented papers. He published many research articles in various national and international journals.