Efficacious approach for clustering with percentage of drift technique for text frequency

WV Nkhila anjaneya Mtech CSE Visakha engineering college Email: nikhila45.wunnava@gmail.com
Guide: P.Pavithra Mtech

Abstract:

The attributed data will be the input for our new approach. The attributed data will be with 6rows in the document and random displancements(indexes) will be generated according the k value(example 4). These displacement values will be inputs for generation of clusters. Number of documents generation is k value. Once the indexed values are taken out from the main attributed data these values will be compared with the main attributed data(1-na) for attribute matchings. After we get the attributed matching numbers (in this case 4 numbers with example N{2,4,1,3}. So here we use one formula k- N(1)/k***(20)*** to get the values {0.5 , 0, 0.75,0.25}, so 2nd value is the least value in this case. So the main documented attributed nth line will be clustered into 2nd cluster. Like these 4 clusters will be generated. So total number of clustered data(4) will be equal to main attributed data document.

DCD is the proposal for calculating the drifting percentage for attributed data.

Literature survey:

The increasing volume of data in modern business and science calls for more complex and sophisticated tools. Although advances in data mining technology have made extensive data collection much easier, it is still always evolving and there is a constant need for new techniques and tools that can help us transform this data into useful information and knowledge.

Various steps in data mining involved:

- Pre-processing
- Clustering
- Stemming
- Cleaning
- Categorization
- Organizing

Introduction:

In this work mainly we take attributed data for clustering and prefix propagation. There are lot of clustering techniques are available but here we use one unique approach of drifting for clustering of data. Depends on the k value the random displacements will be generated. Those displancements values will be compared with attributed positions and gets the calculated values according to the above formula***(20)***.

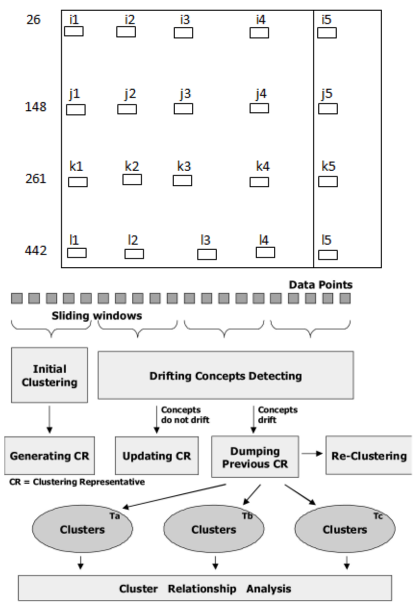
Once the minimum value finds and that position that compared document's data(line) will be placed in that clustered position. Ex: 0.3 is minimum value in the 3rd position then the data will be appended to the 3rd cluster. So total formed clusters data size will be exact size to main data.

(Note: Due to categorical data this work is limited to k value as 4 other k values is future work.)

A cluster represents the concept commonly shared by its data points. The change in the trend/concept is reflected by the data points that deviate from the existing clusters and the underlying clusters should be modified accordingly with time. Existing works on clustering categorical data focus on doing clustering on the entire data set and do not take the drifting concepts into consideration.

Comparable terms vectors:

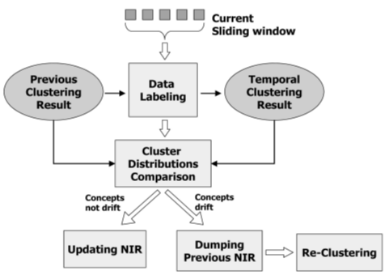
If k value is 4 and if randomly generated positions are {26,148,261,442} so at that offsets from main document



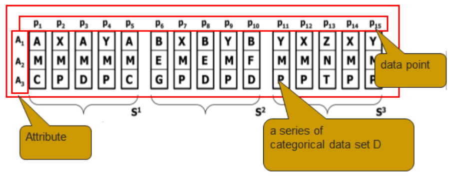
Comparable terms vectors:
26 -- {i1 i2 i3 i4 i5}
148 -- {j1 j2 j3 j4 j5}
261 -- {k1 k2 k3 k4 k5}
442 -- {l1 l2 l3 l4 l5}
Then above positions are compared with all the attributed lines in the main document. Then the formula applied for the generated comparable positions in sequence***(20)***.

Once the clusters are formed depends on the input terms sequence comparision on with λ value as 3 the values are framed in one document with propagation of final document.

Representing the clusters with categorical data:
An information theoretic metric, named "Node Importance Representative" (NIR) is used to extract concept of the cluster. NIR represents clusters by calculating the importance of each attribute-value pair in the clusters. NIR is considered as a better way compared to "Modes" for incremental clustering.
Drifting concept detection:
Based on NIR, we propose the "Drifting Concept Detection" (DCD).In DCD, the incoming categorical data points at the present sliding window are first allocated into the corresponding proper cluster at the last clustering result.If the distribution is changed (exceeding some criteria), the concepts are said to drift. The approach presented in this paper not only detects the drifting concepts in the attributed data but also explains the drifting concepts by analysing the relationship between clustering results at various times.
The main goal of the DCD algorithm is to detect the difference of cluster distributions between the current S_t data subset and the last clustering result $C^{[t_e, t-1]}$ and to decide whether the re clustering is required or not in



The problem of clustering the attributed time-evolving data is formulated as follows:



Algorithm:
Drift:
Input ←Ad(document with Attributes)
Output ←K Clusters

Initialization :

n ←Total no of data lines in Ad
k ←no of clusters
$\sum R_N \leftarrow 0$*// Random*
Cd←Clusters number vector
$\sum C_E$ *←Comparable elements*
$\sum V$ *←Values*
For j in 1 to m
C_E *=Ad {Element AT* R_N *(J)}*
End for

For l in 1 to n
Temp = Ad {l};
For m in 1 to k
V = compare (temp, C_E*{m})*
End for
Value =min (v);
P=position (, value);
C_d *{p}←temp*
End for

Propogation:
Input k- clusters
Base terms t_1 t_2 t_3
Final cluster output CD_0
For each di in k-1clulstersed documents
l=0
Loop start
F1<- di(n){1}
F2<-di(n){2}
F3<-di(n){3}
n=n+1
if(f1==t1)|if(f2==t2)|if(f3==t3)
CDo <- di(n)
I=i+1
end if
end loop
end for

Experimental results:

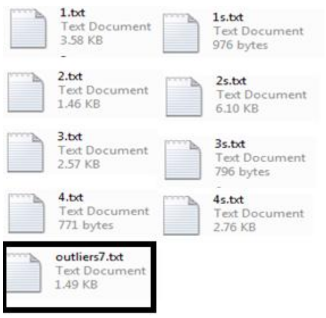
Data sets used:

```
vhigh vhigh 2 2 small med
vhigh vhigh 2 2 small high
vhigh vhigh 2 2 med low
vhigh vhigh 2 2 med med
vhigh vhigh 2 2 med high
vhigh vhigh 2 2 big low
vhigh vhigh 2 2 big med
vhigh vhigh 2 2 big high
vhigh vhigh 2 4 small low
vhigh vhigh 2 4 small med
vhigh vhigh 2 4 small high
vhigh vhigh 2 4 med low
vhigh vhigh 2 4 med med
vhigh vhigh 2 4 med high
vhigh vhigh 2 4 big low
```

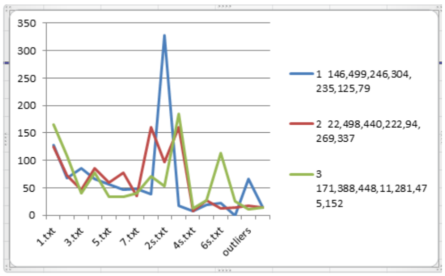
(data set1)

```
high vhigh 4 4 med high
high vhigh 4 4 big low
high vhigh 4 4 big med
high vhigh 4 4 big high
high vhigh 4 more small low
high vhigh 4 more small med
high vhigh 4 more small high
high vhigh 4 more med low
high vhigh 4 more med med
high vhigh 4 more med high
high vhigh 4 more big low
high vhigh 4 more big med
high vhigh 4 more big high
high vhigh 5more 2 small low
high vhigh 5more 2 small med
high vhigh 5more 2 small high
high vhigh 5more 2 med low
high vhigh 5more 2 med med
```

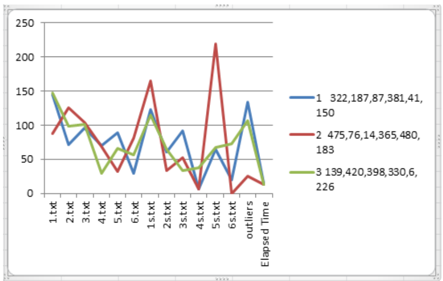
(data set2)



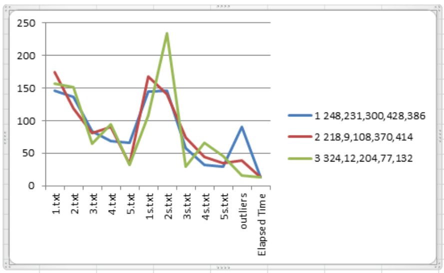
(Result)



K = 7



K = 6



K = 5

Conclusion: A framework to perform clustering on attributed data and time-evolving data. Finds the drifting concepts at different sliding window by DCD. Represents the relationship between clustering results pictorially. DCD can provide fine grained clustering results with correctly detected drifting concept result.

References:

[1] C. Aggarwal, J. Han, J. Wang, and P. Yu, "A Framework for Clustering 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. Procopiuc, and J.S. Park, "Fast Algorithms for Projected Clustering," Proc. ACM SIGMOD '99, pp. 61-72, 1999.

[3] 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.

[4] D. Barbara ´, Y. Li, and J. Couto, "Coolcat: An Entropy-Based Algorithm for Categorical Clustering," Proc. ACM Int'l Conf. Information and Knowledge Management (CIKM), 2002.

664 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 5, MAY 2009

Authorized licensed use limited to: National Central University. Downloaded on May 21, 2009 at 04:40 from IEEE Xplore. Restrictions apply.

[5] 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.

[6] D. Chakrabarti, R. Kumar, and A. Tomkins, "Evolutionary Clustering," Proc. ACM SIGKDD '06, pp. 554-560, 2006.

[7] 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.

[8] Y. Chi, X.-D. Song, D.-Y. Zhou, K. Hino, and B.L. Tseng, "Evolutionary Spectral Clustering by

Incorporating Temporal Smoothness," Proc. ACM SIGKDD '07, pp. 153-162, 2007.

[9] B.-R. Dai, J.-W. Huang, M.-Y. Yeh, and M.-S. Chen, "Adaptive Clustering for Multiple Evolving Streams," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 9, pp. 1166-1180, Sept. 2006.

[10] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Like- lihood from Incomplete Data via the EM Algorithm," J. Royal Statistical Soc., 1977.

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

[12] M.M. Gaber and P.S. Yu, "Detection and Classification of Changes in Evolving Data Streams," Int'l J. Information Technology and Decision Making, vol. 5, no. 4, pp. 659-670, 2006.

[13] V. Ganti, J. Gehrke, and R. Ramakrishnan, "CACTUS—Clustering Categorical Data Using Summaries," Proc. ACM SIGKDD, 1999.

[14] D. Gibson, J.M. Kleinberg, and P. Raghavan, "Clustering Categorical Data: An Approach Based on Dynamical Systems," VLDB J., vol. 8, nos. 3-4, pp. 222-236, 2000.

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

[16] S. Guha, R. Rastogi, and K. Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes," Proc. 15th Int'l Conf. Data Eng. (ICDE), 1999.

[17] E.-H. Han, G. Karypis, V. Kumar, and B. Mobasher, "Clustering Based on Association Rule Hypergraphs," Proc. ACM SIGMOD Workshop Research Issues in Data Mining and Knowledge Discovery (DMKD), 1997.

[18] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001.

[19] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values," Data Mining and Knowledge Discovery, 1998.