Scope International Journal of Science, Humanities, Management and Technology. ISSN : 2455-068X

Vol.3 Issue 4 (2017) 40 - 48. Submitted 01/11/2017. Published 10/12/2017

AN ACTIVE RE - CLUSTER BASED ON SELECTION PROCESSING ALGORITHM

[1]S.Anuradha, Assistant Professor, Department Of Computer Science, Shri Sakthikailassh Women's College, Salem.

[2]S.Surya, M.Phil Scholar, Department Of Computer Science, Shri Sakthikailassh Women's College, Salem

ABSTRACT

Clustering is a challenging problem in data streams domain. This is because the large volume of data arriving in a rivulet and evolving over time. Several clustering algorithms have be located developed for growing data streams. Besides inadequate memory, the Mother Nature of evolving data stream point toward roughly requirements for clustering. In this paper, analyze the requirements needed for clustering evolving data streams. Project reviews some of the latest algorithms in the literature and discourse how project meet the provisions. Feature pulling out is the exceptional arrangement of dimensionality saving somewhere characteristic variety is the subfield of attribute extraction. Piece mixture algorithms essentially have two basic criteria christened as, time prerequisite and quality. The crucial awareness of editorialmiscellanyprogression is increase precision level of classifier, reduce dimensionality; speedup the bundling task etc., they primarily focuses on Comparison of various techniques and algorithms for quality variety process. A resourceful re-cluster medley scheme is aproficientapproach to increase the accuracy of classifiers, dimensionality bargain, do not here within team work irrelevant and unnecessary data. A reasonable study of various article variety methods and algorithms. Considerations of individuals 'performances are mien over and the compensations in addition downsides of articlevarietymethodin addition algorithms are summarized. With unremittingly emergent data, clusters compatibly need to propagate once in a while to accommodate the increased ultimatum of data handing out. This is as a matter of course done by toting up of newer hardware, whose configuration might differ from the left over nodes. As an end result, constellations are pleasurable far removed from in Mother Nature. For plentiful material world contraption book learning in accumulation todocumentsdrawing out presentations, documents is suggested in the arrangement of leaflets.

1. INTRODUCTION

ECENT years, there is a dramatic increase in our ability to collect data from various sensors, devices, in different formats, from independent or connected applications. This data is generated continuously at high speed over time that can be considered as data streams. Financial applications, web application, sensor network data, monitoring environmental sensors, and security control in the networks are some examples of data streams. Clustering is a prominent data streams mining task. However, clustering in data stream environment needs some special requirements due to its characteristic such as data can be scanned only once, clustering in limited memory , and handling noisy data. Furthermore, since data streams evolve over time, the clustering result changes with time. Therefore, one of important feature of a clustering algorithm

is to handle evolving data streams. In this paper, explore the requirements that are needed for a clustering algorithm to process evolving data streams. There are different clustering algorithms such as partitioning and hierarchical, which are developed to find spherical-shape clusters. One of important class in clustering is density-based clustering which can discover the cluster of non-spherical shape and filter out the outliers. Furthermore, grid-based clustering is another method which has fast processing time which is independent from the number of data points. These clustering algorithm are adopted by researchers to use for evolving data streams in different way such as density micro clustering or grid based clustering as well as partitioning algorithm for data streams

2. RELATED WORK

This should not be viewed as an extensive study of all published clustering algorithms, but only the well-known algorithms using different summary method and cluster the evolving data streams. We study whether or not these algorithms meet the requirements established that.

A. CluStream CluStream

which is designed to cluster data streams over different time horizons in an evolving environment. It has online and offline component. In the online component, CluStream periodically stores summary information about data streams and offline component cluster the summary statistics. The CluStream produces k final clusters called macro clusters by executing the K-means algorithm on the summarized data.

• Summarization: it maintains statistical information about the data stream in terms of micro clusters. Micro-cluster, which was first introduced in, is a temporal extension of cluster feature vector.

• Processing: in order to cluster evolving data stream, CluStream uses titled time window model, which is called pyramidal time frame. It records snapshots of a set of micro clusters at the different level of granularity. CluStream has ability to cluster evolving data streams on the historical and current data stream.

• Outlier detection: in the online phase the algorithm does not have special method for detection outliers only if the arrival data point is not belong to any exsiting micro cluster is considerd as outlier. In the offline phase, CluStream uses k-means which is unable to detect noise and outliers.

B. D-Stream

D-Stream is density grid-based clustering. It has onlineoffline component. The online component maps the new data point into the density grids. The offline component forms final clusters by merging density grids.

• Summarization: for each density grids, D-Stream keeps summary about the corresponding grid in its characteristic vector.

• Processing: D-Stream uses fading window model. For each data points of data streams which is mapped to the grid, weight is considered based on fading function and density of the grid is defined by aggregation of the weight of data points. If no data, record is added to this grid, the density of grid decrease over the time.

• Outlier detection: D-stream introduces the dense, sparse, and sporadic grids to determine noises. It periodically checks the density of grids, if the density of sparse grid is less than the density threshold; it is considered as sporadic grid mapped to by outliers, and will remove from grid list.

C. DenStream

DenStream, is two phase algorithm: online-offline. Online phase keeps the summary information about data streams and the offline phase clusters the summarized data based on density-based clustering.

• Summarization: Den-Stream keeps the summary of data stream in micro clusters. The micro cluster is same as used by CluStream.

• Processing: it uses fading window model, which is consider for each data points in the micro cluster special weight that is exponential and reduce with time.

• Outlier detection: Den-Stream introduces potential and outlier micro cluster to make difference between the real data and outliers. Their difference is their weights therefore the algorithm removes the outlier micro clusters if their weight is less than density threshold of outlier micro cluster the micro cluster is a real outlier. In this way, DenStream will detect and remove the outliers in clustering algorithm.

D. MR-Stream

MR-Stream is an algorithm, which has the ability to cluster data streams at multiple resolutions. The algorithm partitions the data space in cells and a tree like data structure, which keeps the space partitioning. The tree data structure keeps the data clustering in different resolution

• Summarization: each tree node at specific height stores summary information about data point in that height.

• Processing: MR-Stream defines a weight for each data point in the tree node based on fading function.

• Outlier detection: MR-Stream checks the cluster weight, if the size and weight of cluster is less than threshold; it is considered as noise cluster and will be removed.

E. SWClustering

SWClustering is an approach for analyzing evolving data streams over sliding windows. The algorithm can capture the evolution of each clusters as well as distribution of recent records accurately.

• Summarization: SWClustering saves the summary in form of Exponential Histogram of Cluster Feature (EHCF), which is in fact an extension of micro clusters based on sliding window model. EHCF has bucket of TCF (Temporal Cluster Feature) which includes cluster feature with time stamp of most recent record.

• Processing: SWClustering algorithm uses sliding window model to process the most recent data.

• Outlier detection: The algorithm does not have any techniques to remove outliers. Furthermore, clusters are calculated from all EHCF synopses based on K-means, which cannot handle the outliers.

F. FlockStream

FlockStream is a density-based data stream clustering algorithm based on multi agent system, which has selforganization techniques to group similar data points. The algorithm merges online and offline phase.

• Summarization: FlockStream uses p-representative and o-representative in virtual space which are corresponds to a potential micro cluster or an outlier micro cluster, as used in DenStream, in the feature space.

• Processing: the damped window model is used to deal with cluster evolution.

• Outlier detection: The method is robust to noise. The algorithm considers the outliers as an o-representative.

DENGRIS-STREAM

DENGRIS-Stream has online and offline components. The online component reads data stream continuously, maps the data to the related grid cells. The offline component adjusts the clusters in each sliding window

• Summarization: each grid cell has a feature vector which keeps summary information about the grid.

• Processing: the algorithm uses sliding window model that emphasis on the recent data.

• Outlier detection: the grids with the density less than threshold are considered as sparse grid, which includes noises and removes from grid list.

3. EXISTING SYSTEM

Data stream clustering is typically done as a two-stage process with an online part which summarizes the data into many micro-clusters or grid cells and then, in an offline process, these micro-clusters (cells) are re-clustered/merged into a smaller number of final clusters. Since the re-clustering is an offline process and thus not time critical, it is typically not discussed in detail in papers about new data stream clustering algorithms. Most papers suggest using an (sometimes slightly modified) existing conventional clustering algorithm (e.g., weighted k-means in CluStream) where the micro-clusters are used as pseudo points. Another approach used in Den Stream is to use reach ability where all micro-clusters which are less than a given distance from each other are linked together to form clusters. Grid-based algorithms typically merge adjacent dense grid cells to form larger clusters (see, e.g., the original version of D-Stream and MR-Stream).

A PSM may be characterized as follows

• A PSM specifies which inference actions have to be carried out for solving a given task.

• A PSM determines the sequence in which these actions have to be activated.

• In addition, so-called knowledge roles determine which role the domain knowledge plays in each inference action.

These knowledge roles define a domain independent generic terminology. When considering the PSM Heuristic Classification in some more detail (Figure 1) we can identify the three basic inference actions abstract, heuristic match, and refine. Furthermore, four knowledge roles are defined: observables, abstract observables, solution abstractions, and solutions. It is important to see that such a description of a PSM is given in a generic way.

In the meantime various PSMs have been identified, like e.g. Cover-and-Differentiate for solving diagnostic tasks or Propose-and-Revise for parametric design tasks.

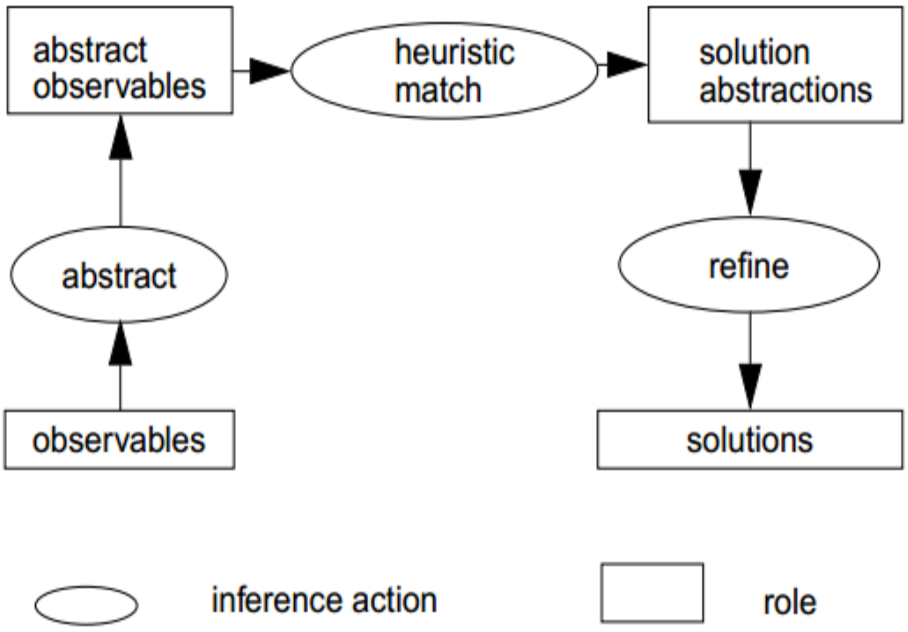

Fig. 1 The Problem-Solving Method Heuristic Classification

PSMs may be exploited in the knowledge engineering process in different ways: The Problem-Solving Method Heuristic Classification observables abstract solutions abstract refine observables heuristic match solution abstractions inference action role

• PSMs contain inference actions which need specific knowledge in order to perform their task. For instance, Heuristic Classification needs a hierarchically structured model of observables and solutions for the inference actions abstract and refine, respectively. So a PSM may be used as a guideline to acquire static domain knowledge.

• A PSM allows to describe the main rationale of the reasoning process of a KBS which supports the validation of the KBS, because the expert is able to understand the problem solving process. In addition, this abstract description may be used during the problem solving process itself for explanation facilities.

4. PROPOSED SYSTEM

With the aim of choosing a Re-Cluster subset of good features with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. While the efficiency concerns the time required to find a re-cluster subset of features, the effectiveness is related to the quality of the subset of features.

In this project, we address two different types of online feature selection tasks. For the first task, we assume that the learner can access all the features of training instances, and our goal is to efficiently identify a fixed number of relevant features for accurate prediction. In the second task, we consider a more challenging scenario where the learner is allowed to access a fixed small number of features for each training instance to identify the subset of relevant features. With the aim of choosing a Re-Cluster subset of good features with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility.

In proposed system develop and evaluate a new method to address this problem for micro-cluster-based algorithms. We introduce the concept of a shared density graph which explicitly captures the density of the original data between micro-clusters during clustering and then show how the graph can be used for re-clustering micro-clusters.

In this project, proposed Clustering based subset Selection algorithm uses minimum spanning tree-based method to cluster features. Moreover, our proposed algorithm does not limit to some specific types of data.

Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines. Thus, feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible. Moreover, "good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other."

In our proposed Cluster based subset Selection algorithm, it involves 1) the construction of the minimum spanning tree from a weighted complete graph; 2) the partitioning of the MST into a forest with each tree representing a cluster; and 3) the selection of representative features from the micro-clusters.

5. SYSTEM MODULES

A module is a part of a program. Programs are composed of one or more independently developed modules that are not combined until the program is linked. A single module can contain one or several routines.

1. Load Data and Convert Micro Data

Load the data into the process. The data has to be preprocessed for removing missing values, noise and outliers. Then the given dataset must be converted into the arff format which is the standard format for

WEKA toolkit. From the arff format, only the attributes and the values are extracted and stored into the database. By considering the last column of the dataset as the class attribute and select the distinct class labels from that and classify the entire dataset with respect to class labels.

2. Compute Density Value

Relevant features have strong correlation with target concept so are always necessary for a best subset, while redundant features are not because their values are completely correlated with each other. Thus, notions of feature redundancy and feature relevance are normally in terms of feature correlation and feature-target concept correlation.

To find the relevance of each attribute with the class label, Information gain is computed in this module. This is also said to be Mutual Information measure. Mutual information measures how much the distribution of the feature values and target classes differ from statistical independence. This is a nonlinear estimation of correlation between feature values or feature values and target classes.

3. Estimate Adjacent Relevance between Each Data

The relevance between the feature Fi € F and the target concept C is referred to as the T-Relevance of Fi and C, and denoted by SU(Fi,C). If SU(Fi,C) is greater than a predetermined threshold , we say that Fi is a strong T-Relevance feature.

$$SU(X,Y) = \frac{2 \times Gain(X|Y)}{H(X) + H(Y)}.$$

After finding the relevance value, the redundant attributes will be removed with respect to the threshold value.

4. Calculate Correlate and Remove Noise

The correlation between any pair of features Fi and Fj (Fi,Fj € ^ F ^ i ≠ j) is called the F-Correlation of Fi and Fj, and denoted by SU(Fi, Fj). The equation symmetric uncertainty which is used for finding the relevance between the attribute and the class is again applied to find the similarity between two attributes with respect to each label.

5. Heuristic MST Construction

With the F-Correlation value computed above, the heuristic Minimum Spanning tree is constructed. For that, we use heuristic algorithm which form MST effectively.

Heuristic algorithm is a greedy algorithm in graph theory that finds a minimum spanning tree for a connected weighted graph. This means it finds a subset of the edges that forms a tree that includes every vertex, where the total weight of all the edges in the tree is minimized. If the graph is not connected, then it finds a minimum spanning forest (a minimum spanning tree for each connected component).

6. Cluster Formation

After building the MST, in the third step, we first remove the edges whose weights are smaller than both of the T-Relevance SU(Fi`, C) and SU(Fj`, C), from the MST. After removing all the unnecessary edges, a forest Forest is obtained. Each tree Tj € Forest represents a cluster that is denoted as V (Tj), which is the

vertex set of Tj as well. As illustrated above, the features in each cluster are redundant, so for each cluster V (Tj) we choose a representative feature Fj R whose T-Relevance SU(Fj R,C) is the greatest.

CONCLUSION

In this Project present a FAST clustering-based feature subset selection algorithm for high dimensional data. The algorithm involves 1) removing irrelevant features, 2) constructing a minimum spanning tree from relative ones, and 3) partitioning the MST and selecting representative features. Experiments also show that shared-density reclustering already performs extremely well when the online data stream clustering component is set to produce a small number of large MCs. A heuristic algorithm used for solving a problem more quickly or for finding an approximate re-cluster subset selection solution. Minimum Redundancy Maximum Relevance selection used to be more powerful than the maximum relevance selection. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. It will provide effective way to predict the efficiency and effectiveness of the clustering based subset selection algorithm. The text data from the four different aspects of the proportion of selected features, run time, classification accuracy of a given classifier. Clustering-based feature subset selection algorithm for high dimensional data. For the future work, we plan to explore different types of correlation measures, and study some formal properties of feature space. In feature we are going to classify the high dimensional data.

BIBLIOGRAPHY

[1] S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan, "Clustering data streams," in Proceedings of the ACM Symposium on Foundations of Computer Science, 12-14 Nov. 2000, pp. 359–366.

[2] C. Aggarwal, Data Streams: Models and Algorithms, ser. Advances in Database Systems, Springer, Ed., 2007.

[3] J. Gama, Knowledge Discovery from Data Streams, 1st ed. Chapman & Hall/CRC, 2010.

[4] J. A. Silva, E. R. Faria, R. C. Barros, E. R. Hruschka, A. C. P. L. F. d. Carvalho, and J. a. Gama, "Data stream clustering: A survey," ACM Computing Surveys, vol. 46, no. 1, pp. 13:1–13:31, Jul. 2013.

[5] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in Proceedings of the International Conference on Very Large Data Bases (VLDB '03), 2003, pp. 81–92.

[6] F. Cao, M. Ester, W. Qian, and A. Zhou, "Density-based clustering over an evolving data stream with noise," in Proceedings of the 2006 SIAM International Conference on Data Mining. SIAM, 2006, pp. 328–339.

[7] Y. Chen and L. Tu, "Density-based clustering for real-time stream data," in Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2007, pp. 133–142.

[8] L. Wan, W. K. Ng, X. H. Dang, P. S. Yu, and K. Zhang, "Density based clustering of data streams at multiple resolutions," ACM Transactions on Knowledge Discovery from Data, vol. 3, no. 3, pp. 1–28, 2009.