ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 10, March 2016

A Study on Clustering Data Stream Algorithm Evaluation Using MOA Framework

N.Sevugapandi[1] and Dr.C.P.Chandran[2]

[1]Research Scholar, Ph.D Part-Time, Category –B, Research and Development Center, Bharathiar University, Coimbatore, Tamilnadu
[2]Associate Professor of Computer Science, Ayya Nadar Janaki Ammal College, Sivakasi, Tamilnadu
sevugapandi1985[1]@gmail.com, [2]drcpchandran@gmail.com

Abstract: Nowadays the application scenario in evolving knowledge stream using some frameworks were unit omnipresent. Stream agglomeration algorithms were introduced to realize helpful information from these streams in time period. The standard of the obtained agglomeration, i.e. however smart they mirror the info, are often assessed by analysis measures. A large number of stream agglomeration algorithms and analysis measures for agglomeration were introduced within the literature; but, up to now there's no general tool for a right away comparison of the various algorithms or the analysis measures. In this study paper, it tends to gift a unique experimental framework for each task.

Keywords: data streams, clustering, MOA, visualization

I. INTRODUCTION

Data streams area unit present these days and a mess of algorithms exist for stream learning situations, e.g. stream classification or stream bunch. In most publications, recently planned algorithms area unit solely compared to little set or perhaps none of the competitive solutions, creating the assessment of their actual effectiveness robust. Moreover, the bulk of experimental evaluations use solely tiny amounts of knowledge. Within the context of knowledge streams this is often unsatisfying; as a result of to be really helpful the algorithms have to be compelled to be capable of handling terribly giant (potentially infinite) streams of examples. Demonstrating systems solely on tiny amounts knowledge of information doesn't build a convincing case for capability to resolve additional hard data stream applications.

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 10, March 2016

In ancient batch learning situations, analysis frameworks were introduced to address the comparison issue. one in all one amongst one in every of these frameworks is that the well-known data processing software package that supports adding new algorithms and analysis measures in a plug-and-play fashion [1], [2]. As knowledge stream learning may be a comparatively new field, the analysis practices don't seem to be nearly additionally researched and established as they're within the ancient batch setting. For this purpose, a framework for stream learning analysis was recently introduced, known as large on-line Analysis (MOA) [3]. So far, however, ratite bird solely considers stream classification algorithms. Consequently, no stream bunch analysis tool exists that provides a set of enforced stream bunch algorithms and analysis measures, though stream bunch is a lively field of analysis with several recent publications.

Besides scrutiny new algorithms to the state of the art, the selection of the analysis measures may be a second key issue for bunch performance on evolving knowledge streams. Most frequently ancient measures are utilized that don't replicate the errors that are specific to evolving knowledge streams, e.g. through moving or merging clusters. Therefore, the goal is to make AN experimental stream bunch system able to assess state-of-the art ways each relating to bunch algorithms and analysis measures.

II. FEATURES AND SYSTEM ARCHITECTURE

In this section we have a tendency to concisely describe the usage and configuration of this system also as a way to extend the framework. An in depth description are going to be obtainable within the manual and is on the far side the scope of this demo paper.

The goal is to make associate degree experimental framework for agglomeration information streams almost like the MOA framework, creating it straightforward for researchers to run experimental information stream benchmarks. The ratite framework offers such potentialities for classification algorithms on information streams. The new options of our ratite extension to stream agglomeration are:

- data generators for evolving streams (including events like novelty, merging clusters, etc. [4]),

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 10, March 2016

- associate degree protractile set of stream cluster algorithms,
- associate degree protractile set of analysis measures,
- ways to assess analysis measures underneath specific error eventualities
- image tools for analyzing results and scrutiny totally different settings.
- visualization tools for analyzing results and comparing different settings.

From the present ratite framework we have a tendency to inherit information generators that are most ordinarily found within the literature. MOA streams will be engineered exploitation artificial generators, reading ARFF files, change of integrity many streams, or filtering streams. They permit the simulation of a doubtless infinite sequence of information and embody construct drift simulation for classification tasks [5]. For stream clump we have a tendency to further new information generators that support the simulation of cluster evolution events like merging or disappearing of clusters [4].

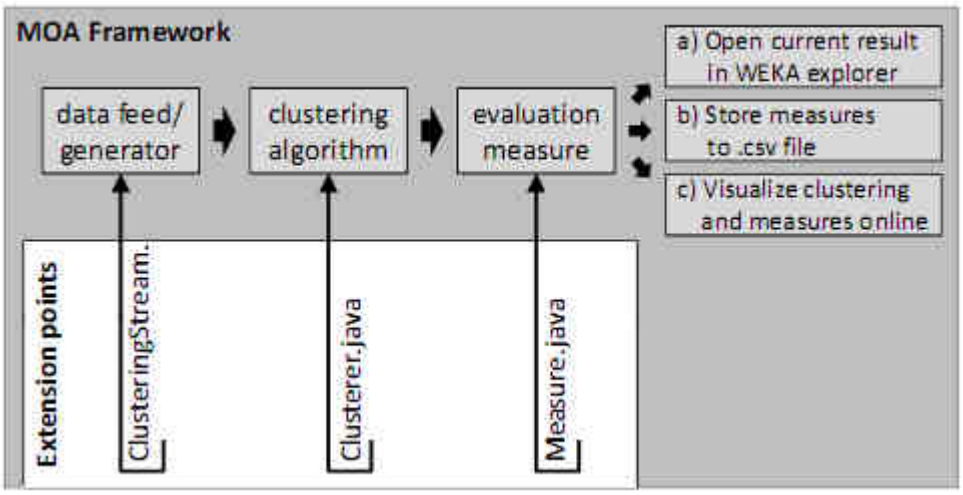

Figure 1. Extension points of the MOA stream clustering framework.

Both design and usage of this stream agglomeration framework follow an equivalent uncomplicated progress idea (Figure 1): First a knowledge feed is chosen and designed, then a stream agglomeration algorithmic program and its settings area unit fastened, and last a collection of analysis measures is chosen.

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 10, March 2016

III. ASSESSING ALGORITHMS

MOA contains many stream agglomeration ways like StreamKM++ [6], CluStream [7], ClusTree [8], Den- Stream [9], D-Stream [10], CobWeb [11] et al. ratite contains measures for analyzing the performance of the agglomeration models generated from each on-line and offline elements. The obtainable measures assess each the right assignment of examples [12] and also the internal structure of the ensuing agglomeration [13]. The mental image part (Figure 3) permits visualizing the stream moreover because the agglomeration results, selecting dimensions for multi dimensional settings, and comparison experiments with completely different settings in parallel.

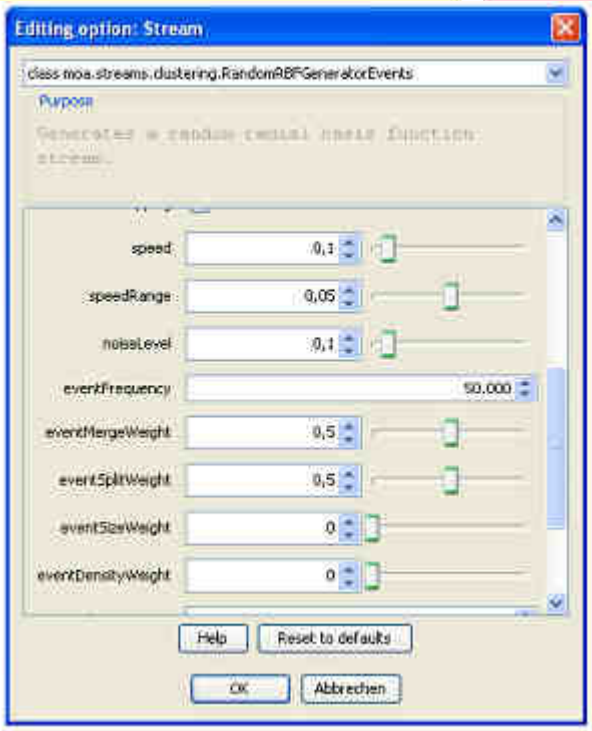

Figure 2. Option dialog for the RBF data generator.

Figure 2 shows a screenshot of the configuration dialog for RBF information generator with events. Usually the spatiality, number, and size of clusters will be set furthermore because the drift speed, decay horizon (aging), noise rate, etc. Events represent changes within the underlying information model like growing of clusters, merging of clusters or creation of latest clusters [4]. Victimization the event frequency and also the individual event weights, one will study the behavior and

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 10, March 2016

performance of various approaches on various settings. Finally, the settings for the information generators will be keeping and loaded, that offers the chance of sharing settings and thereby providing benchmark streaming information sets for repeatability and comparison.

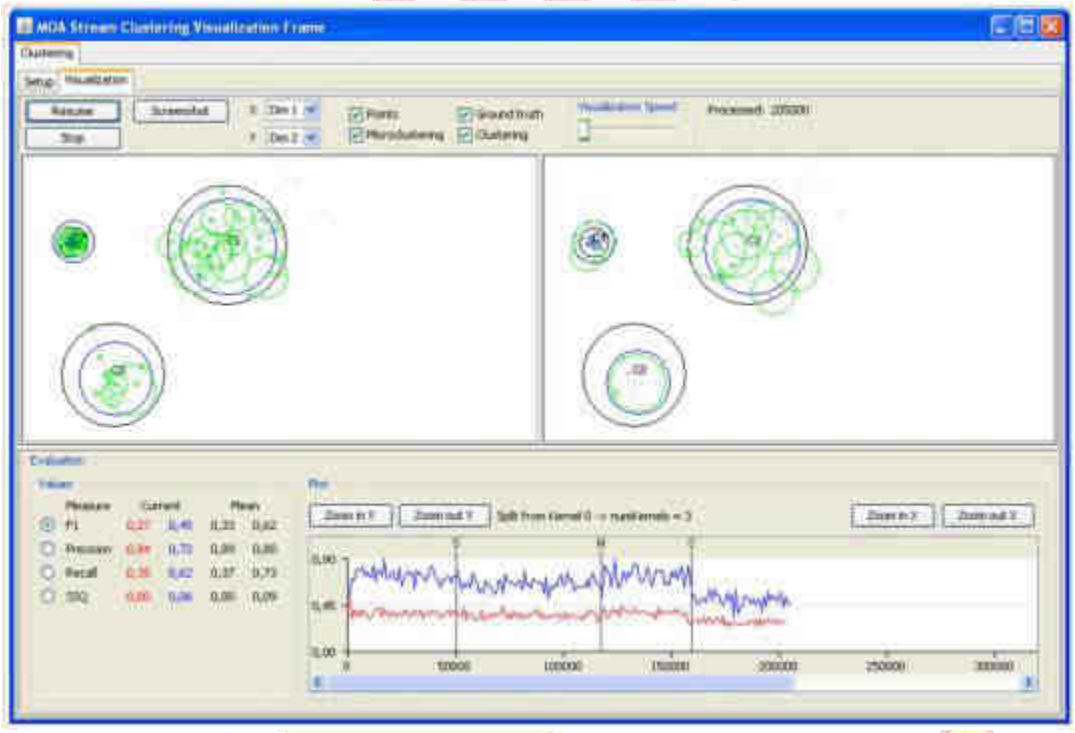

Figure 3. Visualization tab of the clustering MOA graphical user interface.

Figure 3 shows a screenshot of visualization tab. For this screenshot 2 completely different settings of the CluStream rule [7] were compared on identical stream setting (including merge/split events each 50000 examples) and 4 measures were chosen for on-line analysis (F1, Precision, Recall and SSQ). The higher a part of the user interface offers choices to pause and resume the stream, change the visualization speed, select the size for x and y in addition because the parts to be displayed (points, micro- and macro clump and ground truth). The lower a part of the user interface displays the measured values for each settings as numbers (left facet, together with mean values) and also the presently elect live as a plot over the arrived examples (right, F1 live during this example). For the given setting one will see a transparent visit the performance once the split event at roughly 160000 examples (event details ar shown once selecting the corresponding vertical line within the plot).

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 10, March 2016

IV. ASSESSING EVALUATION MEASURES

The results of the analysis of stream cluster algorithms extremely rely on the utilized analysis measures. These measures is classified into structural measures, known as internal and ground truth based mostly measures, known as external. In this demo provides the means that to assess the performance of such analysis measures by testing them in specific clustering error eventualities. We tend to acquire these eventualities by generating cluster out of artificial stream that reflects a desired error level. Then, we tend to assess the performance by testing whether or not they obtained qualities of the analysis measures mirror the error in these generated clusters. We will produce cluster center position errors, known as position offset errors, and radius errors. For position offset errors, the cluster centers are shifted far away from their ground truth position. The highest error level of one indicates that the bottom truth cluster and also the error cluster ar positioned next to every alternative. There are two forms of radius errors: the radius decrease error indicates that the generated error clusters have a radius that's smaller than the radius of the corresponding ground truth cluster. The highest error level of one states that the radius of the error cluster is zero; for a slip-up level of 0 the 2 radii are equal. The radius increase error is realized analogously: a slip-up level of one indicates that the radius of the error cluster has doubled. Moreover, cluster analysis measures are terribly sensitive to the overlap of clusters within the analyzed cluster. The overlap is extremely hooked in to the used aging of the clustered points; viewing a larger history of knowledge points leads to a lot of tail-like clusters, that successively yields the next overlap. In this demo, we will analyze the results of various aging eventualities and live the occurring overlap for elaborated analysis.

V. CONCLUSION

The study is to make associate experimental framework for cluster knowledge streams just like the data evolving framework, in order to that it'll be simple for researchers to run experimental knowledge stream clustering algorithms. The cluster framework provides a group of information generators, algorithms and analysis measures. Based on the MOA framework the visualization and other data evolving stream algorithms.

REFERENCES

[1] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," SIGKDD Explorations 11(1):10–18, 2009.

[2] E. Muller, I. Assent, S. Gunnemann, T. Jansen, and T. Seidl, "OpenSubspace: An open source framework for evaluation and exploration of subspace clustering algorithms in weka," in OSDM in conjunction with PAKDD, 2009, pp. 2–13.

[3] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "MOA: Massive Online Analysis http://sourceforge.net/projects/moa- datastream/," JMLR, 2010.

[4] M. Spiliopoulou, I. Ntoutsi, Y. Theodoridis, and R. Schult, "MONIC: modeling and monitoring cluster transitions," in ACM KDD, 2006, pp. 706–711.

[5] A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavalda`, "New ensemble methods for evolving data streams." in ACM KDD, 2009, pp. 139–148.

[6] M. R. Ackermann, C. Lammersen, M. Ma¨rtens, C. Raupach, C. Sohler, and K. Swierkot, "StreamKM++: A clustering algorithm for data streams," in SIAM ALENEX, 2010.

[7] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in VLDB, 2003.

[8] P. Kranen, I. Assent, C. Baldauf, and T. Seidl, "Self-adaptive anytime stream clustering," in ICDM, 2009, pp. 249–258.

[9] F. Cao, M. Ester, W. Qian, and A. Zhou, "Density-based clustering over an evolving data stream with noise," in SDM, 2006.

[10] L. Tu and Y. Chen, "Stream data clustering based on grid density and attraction," ACM TKDD 3(3):1–27, 2009.

[11] D. H. Fisher, "Knowledge acquisition via incremental conceptual clustering," Machine Learning 2(2):139–172, 1987.

[12] J. Chen, "Adapting the right measures for k-means clustering," in ACM KDD, 2009, pp. 877–884.

[13] G. W. Milligan, "A monte carlo study of thirty internal criterion measures for cluster analysis," Psychometrika, vol. 46, no. 2, pp. 187–199, 1981.