IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-8, August 2018 E-ISSN: 2347-2693

Analysis of Top-K Query for Data Stream Using Classification Adaptive Model

M.Nalini [1*], Anjali kuruvilla[2]

[1,2]Dep. of Computer Science, Rathinavel Subramaniam College of Arts and Science, Coimbatore,India

[*]*Corresponding Author: nalini.m@rvsgroup.com, Tel.: 8300160089*

Available online at: www.ijcseonline.org

Accepted: 15/Aug/2018, Published: 31/Aug/2018

***Abstract*:-** Data stream classification has been a wide studied detailed examination downside in recent years. The vigorous and evolving nature of knowledge streams needs economical and effective techniques that square measure considerably completely different from static data classification techniques. The foremost strenuous and well studied characteristics of information streams square measure its infinite length and concept-drift. Information stream classification poses several challenges to the info mining community. All through this paper, we enclose a affinity to talk four such major dispute namely, IL, CDI, CE, and FE. Since an information stream is in theory IL Model (Infinite long), it's impractical to store and use all the historical information for coaching. CD Model (Concept-Drift) could be a common development in information streams that happens as a result of changes within the underlying ideas. CE (Concept-Evolution) happens as a results of new categories evolving within the stream. Feature-evolution (FE) Model could be a oft occurring method in several streams, reminiscent of text streams, within which new options seem because the stream progresses. Most existing information stream classification techniques address solely the primary challenges, and ignore the latter two. The paper proposes associate degree ensemble classification framework, wherever every classifier is provided with a unique class detector, to handle CE.

Keywords: Outlier Detection, Big Data Mining, Concept Drift (CF)t, Concept Evaluation(CE),Feature Evaluation (FE), CP GraphModel.

I. INTRODUCTION

The use of huge information is changing into a vital means for leading firms to outmatch their peers. In most trade, established entrant alike can influence data-driven methods to introduce, compete, and capture price. Indeed, we tend to found early samples of such use of information in each sector we tend to examined. In attention, information pioneers square measure analyzing the health outcomes of prescription drugs once they were wide prescribed, and discovering edges and risks that weren't evident throughout essentially additional restricted clinical trials. alternative early adopters of huge information square measure exploitation information from sensors embedded in product from children's toys to industrial product to work out however these product are literally employed in the $64000 world. Such data then informs the creation of latest service offerings and therefore the style of future product. huge information can facilitate to form new growth opportunities and fully new classes of firms, comparable to people who mixture and analyse business information. several of those are firms that sit within the middle of enormous info flows wherever information concerning product and services, patrons and suppliers, client preferences and intent are often captured and analyzed.

MapReduce Model is cooperative in a very extensive selection of function, together with distributed pattern-based looking out, distributed sorting, internet link-graph reversal, Singular price Decomposition, internet access log stats, inverted index construction, document bunch, machine learning, and applied math AI. Also, MapReduce model have be made to order many computation setting like multi-core model and many-core systems model, desktop grids model, volunteer computing surroundings, dynamic cloud surroundings, and mobile environments. Next to Google, MapReduce model was accustomed utterly regenerate Google's index of the planet Wide internet. It replaced the recent unintended programs that updated the index and ran the varied analyses. Enlargement at Google enclose every since enraptured on to technologies comparable to coffeepot, Flume and water wheel that supply streaming operation and updates rather than instruction execution, to permit group action "live" search results while not reconstruction the entire index.

MapReduce's stable inputs and outputs ar sometimes hold on in an exceedingly distributed filing system. The transient information is sometimes hold on on native disk and fetched remotely by the reducers.

The existing system includes 3 major contributions in novel category detection for information streams. First, it proposes a versatile call boundary for outlier detection by permitting a slack house outside the choice boundary. This house is controlled by a threshold, and therefore the threshold is tailored unendingly to scale back the chance of false alarms and incomprehensible novel categories.

Second, it applies a probabilistic approach to sight novel category instances mistreatment the separate Gini constant. With this approach, it's ready to distinguish totally different causes for the looks of the outliers, namely, noise, concept-drift, or concept-evolution. It derives associate analytical threshold for the Gini constant that identifies the case wherever a completely unique category seems within the stream. Third, it applies a graph-based approach to sight the looks of quite one novel categories at the same time, and separate the instances of 1 novel category from the others.

The arrangement model is ensemble classification framework, wherever every categoryifier is supplied with a completely unique class detector, to handle concept-drift and concept-evolution. to handle feature-evolution, we tend to propose a feature set homogenization technique. The novel category detection module is employed by creating it additional adaptational to the evolving stream, and enabling it to sight quite one novel category at a time for any reference. Comparison with progressive information stream classification techniques establishes the effectiveness of the planned approach. The classification model additionally improve the novel category detection module by creating it additional adaptational to the evolving stream, and enabling it to sight quite one novel category at a time. additionally, idea drift approach is additionally applied.

II. RELATED WORKS

In this paper, they mentioned the main points of such an internet voice recognition system. For this purpose, we have a tendency to use our micro-clustering algorithms to style concise signatures of the target speakers. one in all the stunning and perceptive observations from our experiences with such a system I that whereas it had been originally designed just for efficiency, we have a tendency to later discovered that it had been conjointly a lot of correct than the wide used mathematician Mixture Model (GMM)

A renowned methodology for speaker classification and identification is that of mathematician Mixture Modeling (GMM) [1]. The first step is to extract multi-dimensional feature vectors so as to represent parts of sampled speech. during this methodology, it's assumed that every datum extracted from the speech segments from variety of known speakers square measure accustomed estimate the parameters of a GMM model.

In that Paper "A Framework for On-Demand Classification of Evolving information Streams" [2] describe a current models of the classification downside don't effectively handle bursts of explicit categories coming back in at completely different times. In actuality, the close to model of the classification draw back simply concentrates on ways in which for one-pass classification modeling of very big info sets.

In the paper, they developed such associate on-demand classifier. The on-demand classifier is intended by adapting the (unsupervised) micro clustering model 3] to the classification downside. Since micro clustering could be a information report technique, a number of the underlying ideas may be leveraged effectively for different issues, akin to classification, that utilize the mixture information behaviour over completely different time horizons.

In the paper "New Ensemble ways For Evolving information Streams" [4] the authors prince consort Bifet, Geoff Holmes, Bernhard Pfahringer, Richard Kirkby Ricard Gavaldà expressed that advanced analysis of information streams is quickly changing into a key space of information mining analysis because the range of applications hard to please such process will increase.

Online mining once such information streams evolve over time, that's once ideas drift or modification utterly, is changing into one in every of the core problems. once braving non-stationary ideas, ensembles of classier have many benefits over single classifier methods: they're straightforward to scale and put, they'll adapt to vary quickly by pruning under-performing elements of the ensemble, and that they so sometimes additionally generate additional correct thought descriptions.

In the paper "Using Additive skilled Ensembles to deal with thought Drift" [5] the authors Jeremy Z. Kolter and Marcus A. Maloof thought-about on-line learning wherever the target thought will modification over time. Skilled prediction model has predetermined the worst-case analysis on any subsequence of the coaching information relative to the act of the simplest skilled. However, as a result of these "experts" is also tough to implement, they took a additional general approach and sure performance relative to the particular performance of any on-line learner on this single subsequence.

III.METHODOLOGY

The planned system implements all existing system approach additionally with idea drift approach implementation. the fundamental steps in categoryification and novel class detection ar as follows. every incoming instance within the information stream is initial examined by a outlier detection module to see whether or not it's AN outlier. If it's not AN outlier, then it's categoryified as AN existing class victimisation majority selection among the classifiers within the ensemble. If it's AN outlier, it's briefly hold on in an exceedingly buffer.

When there ar enough instances within the buffer, the novel category detection module is invoked. If a completely unique category is found, the instances of the novel category ar labelled consequently. Otherwise, the instances within the buffer ar thought of as AN existing category and classified ordinarily victimisation the ensemble of models. The ensemble of models is invoked each within the outlier detection and novel category detection modules. The outlier detection method utilizes the choice boundary of the ensemble of models to make your mind up whether or not or not AN instance is outlier. This call boundary is constructed throughout coaching.

The novel category detection method computes the cohesion among the outliers within the buffer and separation of the outliers from the present categories to make your mind up whether or not a completely unique category has arrived. The planned system enhances the present novel category detection technique in 3 ways, which are

- Outlier detection victimisation adaptative threshold,
- Novel category detection and
- Simultaneous multiple novel category detection.

A. Outlier Detection

When the data arrived is more and the classes formed out of them increases the problem is termed as infinite length problem. This is to be avoided. Each incoming instance in the data stream is first examined by an outlier detection module to check whether it is an outlier. If it is not an outlier, then it is classified as an existing class using majority voting among the classifiers in the ensemble. If it is an outlier, it is temporarily stored in a buffer.

When there are more new classes formed, then the classes with less content are discarded so that the number of classes is maintained within a given limit and this avoids the infinite problem.

B. Concept Drift Identification (CDI)

The words and the category to which it belongs are added in the 'category' table. A client application is developed in which the text content is sent to the server application which updates the incoming message.

The words are extracted and the words fell in the given category are identified and counted. If there are more words in the category and the word count reduced in the successive incoming messages, then the concept is found to be reduced and when the number of words reduced to zero, the concept is said to be drifted. The number of observation time count is set so that when the number of word count is zero for that given number of time, then the concept is said to be drifted.

C. Novel Class Detection (NCD)

During the concept evolution phase, the novel class detection module is invoked. If a novel class is found, the instances of the novel class are tagged accordingly. Otherwise, the instances in the buffer are considered as an existing class and classified normally using the ensemble of models. The words occurred frequently but not matched with any of the category available, and then the word is considered to be fallen in new class.

Algorithm Used
Adjust-Threshold(x, OUTTH)
Input: x, OutTh

Which are most recent labelled instance and OutTh is current outlier threshold

Process:

i. Populate the class labels
ii. Check the incoming X data is matched with any of the class
iii. If not fallen in any of the class, then new class is said to be occurred. OutTh is increased with a slack variable Δ.

Output: OUTTH

New outlier OutTh threshold.

D. FEATURE EVOLUTION IDENTIFICATION

In this first phase, along with concept evolution, feature evolution is identified. The repeated patterns are identified in the received messages and if it is found that more number of received messages contains the patterns, then it is said that feature evolution occurs. The following output forms are available.

- Concept Evolution Identification
- New Feature Evolution Identification

i) Concept Evolution Identification:

In this phase, the words occurred frequently but not matched with any of the category available, and then the word is considered to be fallen in new class. A notifyicon is displayed when new concept is evoluted.

ii) New Feature Evolution Identification:

The repeated patterns are identified in the received messages and if it is found that more number of received messages contains the patterns, then it is said that feature evolution occurs. A notifyicon is displayed when new concept is evolves.

E. Frequent Pattern Mining Across Multiple Databases

In this phase frequent patterns that match the user-specified condition are mined. Mining sequential patterns across multiple databases with different domains has been addressed database. Assume two sequential databases D 1 and D 2, where D 1 and D 2 respectively have Xi and Xj at a given time. If Xi, Yi appears in more sequences than a user specified threshold, it is a frequent pattern. This is different from our problem, since they do not consider co-occurrence patterns

F. Frequent pattern mining across multiple streams.

The continuous mining model is co-occurrence patterns that appear in at least ρ streams, where ρ is a user-specified threshold. Their empirical studies show that mining co-ccurrence patterns across multiple streams is practically useful. The Seg-tree summarizes the valid transactions, and when a new transaction *t* arrives, *t* is inserted into the Seg-tree while merging the prefix nodes.

G. CP-Graph Model

In this CP-Graph start with the structure and to update the answer quickly, it is desirable that user can efficiently enumerate necessary closed co-occurrence patterns and compute their counts. The CP-Graph satisfies these requirements, and consists of *V* and *E*, where *V* (*E*) denotes the set of vertices (edges) at the current time-cycle *c now*. In a nutshell, each object *oi*, which appears in the valid transactions, is regarded as a vertex *vi*, and edges are created between vertices, to represent patterns on the window. We below introduce the details of vertices and edges, and describe edges first, for ease of presentation. The proposed system has following advantages.

- The DD [drift detection] issue is covered.
- Decision boundary for outlier detection is changing as the new data arrives.
- Uses any drift detection technique (DDT) to make the chunk size dynamic.
- Concept drift (CDI) approach is used and so models with less importance are eliminated and space is provided for new models.

IV. CONCLUSION

The paper identifies two key mechanisms of the novel class detection technique, namely, outlier detection, and identifying novel class instances, as the prime cause of high error rates for previous approaches.

To solve this problem, the paper proposes an improved technique for outlier detection by defining a slack space outside the decision boundary of each classification model, and adaptively changing this slack space based on the characteristic of the evolving data.

It also proposes a better alternative approach for identifying novel class instances using discrete Gini Coefficient and graph-based approach for multiple-novel class detection.

Through this paper, the drift detection issue is covered, DBRY (Decision boundary) for outlier detection is changing as the new data arrives; Uses any DDT (Drift Detection Technique) to make the chunk size dynamic; Concept drift approach is used and so models with less importance are eliminated and space is provided for new models.

We have added medical images captured in various time periods and the concept drift is studied, then disease severity level can be identified. This is a additional and application-oriented module we have included in our paper to increase its scope of usage and desirability.

V. SCOPE FOR FUTURE ENHANCEMENTS

At present, thought drift, evolution and have evolution is detected for the text information, if the appliance is meant as internet service, it are often integrated in several network applications. the appliance is developed such on top of aforesaid enhancements are often integrated with current modules. If one category split into many categories and once split, they occupy an equivalent feature house, meaning, the feature house they were covering before split is that the same because the union of the feature areas coated once split, none of the new categories are going to be detected as novel. However, if a part of one or each of the new categories occupies a brand new feature house, then those elements are going to be detected as novel. A motivating future work would be to spot this special case additional exactly to differentiate from the particular arrival of a unique category.

REFERENCES

[1] C. C. Aggarwal. On classification and segmentation of massive audio data streams. Knowl. and Info. Sys., 20:137–156, July 2009.
[2] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for on-demand classification of evolving data streams. IEEE Trans. Knowl. Data Eng, 18(5):577–589, 2006.
[3] A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavald. New ensemble methods for evolving data streams. In Proc. SIGKDD, pages 139–148, 2009.
[4] S. Chen, H. Wang, S. Zhou, and P. Yu. Stop chasing trends: Discovering high order models in evolving data. In Proc. ICDE, pages 923–932, 2008.
[5] W. Fan. Systematic data selection to mine concept-drifting data streams. In Proc. SIGKDD, pages 128–137, 2004.

[6] J. Gao, W. Fan, and J. Han. On appropriate assumptions to mine data streams. In Proc. ICDM, pages 143–152, 2007.
[7] S. Hashemi, Y. Yang, Z. Mirzamomen, and M. Kangavari. Adapted one-versus-all decision trees for data stream classification. IEEE Trans. Knowl. Data Eng, 21(5):624–637, 2009.
[8] G. Hulten, L. Spencer, and P. Domingos. Mining timechanging data streams. In Proc. SIGKDD, pages 97–106, 2001.
[9] I. Katakis, G. Tsoumakas, and I. Vlahavas. Dynamic feature space and incremental feature selection for the classification of textual data streams. In Proc. ECML PKDD, pages 102–116,2006.
[10] I. Katakis, G. Tsoumakas, and I. Vlahavas. Tracking recurring contexts using ensemble classifiers: an application to email filtering. Knowledge and Information Systems, 22:371–391, 2010.
[11] J. Kolter and M. Maloof. Using additive expert ensembles to cope with concept drift. In Proc. ICML, pages 449–456, 2005.
[12] D. D. Lewis, Y. Yang, T. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. Journal of Machine Learning Research, 5:361–397, 2004.
[13] X. Li, P. S. Yu, B. Liu, and S.-K. Ng. Positive unlabeled learning for data stream classification. In Proc. SDM, pages 257–268, 2009.
[14] M. M. Masud, Q. Chen, J. Gao, L. Khan, J. Han, and B. M. Thuraisingham. Classification and novel class detection of data streams in a dynamic feature space. In Proc. ECML PKDD, volume II, pages 337–352, 2010
[15] M. M. Masud, Q. Chen, L. Khan, C. Aggarwal, J. Gao, J. Han, and B. M. Thuraisingham. Addressing concept-evolution in concept-drifting data streams. In Proc. ICDM, pages 929–934,2010.

Authors Profile

Mrs.M.Nalini is currently pursuing Ph.D and currently working as Assistant Professor in Department of Computer Sciences, Rvs college of Arts and Science, Affiliated to Bharathiar University, Sulur, Coimbatore. Her main research work focuses on Data Mining, Computational Intelligence based education.

Ms.Anjali Kuruvilla is a Research Scholar and currently working as Assistant Professor in Department of Computer Science, Rathinam college of Arts and Science(Autonomous), Affiliated to Bharathiar University, Coimbatore. Her main research work focuses on Data Mining and Analysis of top K query for Data Stream.

.

