International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Survey of Adaptive Novel Class Detection and Classification of Feature-Evolving Data Streams

Punam D. Dhande[1], Dr. A. M. Dixit[2]

[1,2]Savitribai Phule Pune University, PVPIT, Pune, Maharashtra, India

Abstract: ***In the data mining communities, Data stream classification causes number of difficulties. There are four major difficulties present. Those are: 1. Infinite Length 2. Concept-Drift. 3. Concept-Evolution. 4. Feature-Evolution. Since a data stream has hypothetically infinite length, it is unreasonable to store and utilize all the past data for training. Concept-drift is a regular incident in data streams, which takes place as a consequence of modification in the core concepts. Concept-evolution takes place as a consequence of new classes developing in the stream. Feature-evolution is an often happening process in number of streams, for example, text streams, in which new features that is words or expressions, show up as the stream advances. Most of present data stream classification methods tackle merely the initial two difficulties, and disregard the last two. To tackle concept-drift and concept-evolution, an ensemble classification technique can be implemented, in which every classifier is equipped by a novel class detector. A feature set homogenization method can be implemented for tackling feature-evolution. Also the novel class identification module can be improved by making it more versatile to the advancing stream, and empowering it to detect number of novel class at once.***

Keywords: Data stream classification, Infinite Length, Concept-Drift, Concept-Evolution, Feature-Evolution.

1. Introduction

In last few years, large amount of research has been done on Data stream classification. The dynamic and developing nature of data streams requires effective and successful methods that are not quite the same as static data classification methods. Two of the most difficult and generally considered attributes of information streams are its infinite length and concept-drift. As a data stream is a quick and consistent incident, it is considered to have endless length. Along these lines, it is not practically possible to store and utilize all the data for training. The most obvious option is an incremental learning procedure. A number of incremental learners have been proposed to attend this issue [1] [2]. Likewise, concept-drift takes place in the stream when the core concepts of the stream change along with time. Different types of methods have additionally been proposed in the past for handling concept-drift [3] [4] [5] in data stream classification. Nonetheless, there are two other critical qualities of information streams, in particular, concept-evolution and feature-evolution that are disregarded by the greater part of the current systems.

Concept-evolution takes place when new classes advance in the data. For instance, consider the issue of intrusion detection. If every sort of attack is considered as a class name, then concept-evolution takes place when a totally new sort of attack takes place in the traffic. An alternate case is the situation of a text data stream. For this situation, new classes might regularly rise in the core stream of text messages. The issue of concept-evolution is addressed in very existing data stream classification systems.

In past, study [6] has attended the novel class detection issue in the vicinity of concept-drift and infinite length. In that method, an assembly of models is utilized to categorize the unlabeled data, and identify novel classes. The novel class detection methodology contains three steps. 1) A decision limit is made in between training. 2) Test points falling outside the decision limit are announced as anomalies. 3) The anomalies are examined to check whether there is sufficient cohesion among themselves (i.e., among the exceptions) and separation from the current class examples. This study did not address the feature-evolution issue. The feature-evolution issue is attended in [7], which additionally attended the concept-evolution issue. On the other hand, both [6] and [7] have two disadvantages. 1) The false alarm rate that is recognition of existing classes as novel is more for few data sets. 2) If more than one novel class is present, they are not able to recognize among them.

2. Literature Review

Aim of A. Bifet et. al [8] was to construct a experimental schema for data streams like the WEKA framework, with the goal that it will be simple for researchers to execute experimental data stream benchmarks. Novel bagging techniques were exhibited: ASHT Bagging utilizing trees of distinctive sizes, and ADWIN Bagging utilizing a change detector to choose when to remove ensemble members which are underperforming. These systems compared positively in a widespread cross-method comparison. Data stream assessment is generally three-dimensional. These comparisons, given your particular resource restrictions, show the technique for preference. Case in point, on the SEA Concepts and Forest Covertype datasets the best performing system over each of the three measurements are apparently HT DDM and HT EDDM, as they are just about the quickest, and very nearly the most exact and, by at any rate a order of magnitude, effectively the most memory-proficient techniques.

S. Hashemi [9] proposed to utilize the one-vs-all (OVA) classification plan to classify streaming data. Not the same as traditional ensembles of multi-class classifiers, OVA includes an ensemble of binary classifiers every one capable of tackling a class of issues. They studied OVA's hypothetical complexity, benefits, difficulties and solutions for data stream classification. It recommends the accompanying. 1) OVA's segment classifiers are generally

uncorrelated in error. Hence the ensemble's general classification precision can be enhanced as its segments have higher multiplicity. 2) OVA can respond quickly to concept changes. After getting each one labeled instance *I*, OVA just needs to overhaul two segment classifiers, the one *I* really have a place with and the one likely to misclassify *I*. 3) nourishing each one labeled instance to just two classifiers is a smart under-sampling plan, which can ease the imbalanced training data issue of OVA learning. It can expand OVA's proficiency of training and updating.

Experiments are led to assess OVA on a vast suite of benchmark data sets. OVA's learning productivity and classification precision are evaluated against those of representative state-of-the-art data stream classification algorithms including CVFDT (single multi-class classifier), SEA and WCE (ensemble of multi-class classifiers), and UFFT (all-vs-all classifiers). Experimental proof supports their hypothetical examination. OVA has the capacity of distributing quick and precision classification for data streams most of the time, and establishes itself as a solid candidate for data stream classification. Utilizing OVA to arrange data streams is another issue and there are different fascinating issues to further explore.

Katakis et. al [10] proposed a technique that dynamically builds and keeps up a ensemble of classifiers for streaming data that are described by concept drift and particularly by repeating settings. After the change of data into another reasonable representation model, incremental clustering was manipulated with a specific end goal to find concepts in the stream and keep up a classifier for every one of these concepts. Keeping in mind the end goal to assess the technique they created an instantiation and developed two true email streams with concept drift. For the purpose of examination, they included in their investigations five methods for managing differing concepts in data streams. Experimental results demonstrate that the instantiation of the CCP system effectively detects concepts in the stream outflanking whatever is left of the systems. This exploratory conclusion with the expansion of the performance of a benchmark variant of their technique upholds the instinct of concept specific classifiers.

X. Li et. al [11] had proposed a novel PU learning strategy called LELC (PU Learning by Extracting Likely positive and negative micro-Clusters). As a PU learning technique, LELC is basically not quite the same as existing data stream classification strategies which consider both positive and negative training data are accessible for learning. Actually, their LELC strategy just obliges a little set of positive documents and a set of unlabeled documents to construct precise classifiers. This is essential for the data stream environment where it is regularly the case that the negative training illustrations are missing, as well as the quantity of positive illustrations accessible for learning can be genuinely constrained.

Their proposed LELC method functions admirably since it can remove dependable negative documents more viably than existing procedures, as demonstrated by their experimental results. Furthermore, LELC strategy has been intended to concentrate great positive and negative micro-clusters from unlabelled information. The principle development is in the procedure for selecting likely positive and negative occasions from the unlabeled set, which manipulates the attributes of data streams, that is data points near one another change their labels together under concept drift instead of each one transforming its label arbitrarily. As opposed to the current PU learning techniques, which extricate likely negative/positive cases point by point, the new system extricates them by micro-clusters. The micro-clusters are structured by using names in the past streams to deliver great grouping.

Since LELC can consequently extricate high-quality positive and negative micro-clusters from data streams, the limits connected with the first positive set P, for example, its constrained size, does not have an immense effect on their algorithm. Augmented by the high quality likely positive set LP and likely negative set LN that obtained, their LELC algorithm is consequently capable to construct a vigorous classifier for data stream classification. The experimental results demonstrated that LELC can be utilized as a part of the data stream environment considerably more successfully than the existing PU learning methods, with essentially better speed and precision.

Masud et. al [12] proposed a few developments over the current classification and new class detection strategy. Firstly, they proposed an enhanced strategy for outlier identification by characterizing a dynamic slack space outside the decision boundary of every classification model. Secondly, they proposed a superior option for recognizing novel class instances utilizing discrete Gini Coefficient. In the end, they proposed a graph-based methodology for recognizing among numerous novel classes. They applied their procedure on a number of real data streams that experience concept-drift and concept-evolution, and accomplish huge performance enhancements over the current systems.

B.M. Thuraisingham et. al [13] have tended to a number of real world issues identified with data stream classification. They had presented a solution for the concept-evolution issue, which has been overlooked by many current data stream classification strategies. Current data stream classification methods consider that aggregate number of classes in the stream is settled. So, occurrences relating a novel class are misclassified by the current systems. They demonstrated how to identify novel classes consequently actually when the classification model is not prepared with the novel class occurrences. Novel class identification gets more and more difficult in the vicinity of concept-drift.

3. Proposed System

There are four developments of proposed work of this paper.

1) An adaptable decision limit for outlier detection is presented by permitting a slack space outside the decision boundary. This space is managed by a threshold, and the threshold is adjusted constantly to lessen the threat of false alarms and missed novel classes.
2) A probabilistic methodology is applied to identify novel class occasions utilizing the discrete Gini Coefficient. With this methodology, recognition of diverse reasons

for the occurrences of the outliers, specifically; noise, concept-drift, or concept-evolution is possible. An explanatory threshold for the Gini Coefficient is inferred, which distinguishes the situation where a novel class shows up in the stream.

3) A graph-based methodology is applied to recognize the occurrences of more than one novel class at the same time, and separate the instances of one novel class from the others.
4) Proposed methodology attends the feature-evolution issue on top of the improvements examined previously. This the first work that presents these enhanced methods for novel class detection and classification in data streams and locations characteristic development. Proposed method is applied on various benchmark data streams.

4. Conclusion

A classification and novel class detection strategy is presented for concept-drifting data stream that addresses four most important difficulties, to be specific, infinite length, concept-drift, concept-evolution, and feature-evolution. The current novel class detection procedures for information streams either do not tackle the feature-evolution or experience the ill effects of high false alarm rate and false detection rates in numerous situations. The feature space conversion procedure is discussed to attend feature-evolution issue. At that point, two key systems of the novel class discovery procedure are distinguished, in particular, outlier detection, and identifying novel class occasions, as the prime reason for high error rates for past methodologies. To tackle this issue, an enhanced method for outlier detection is presented by characterizing a slack space outside the decision bounds of every classification model, and adaptively changing this slack space focused around the normal for the developing data.

Reference

[1] G. Hulten, L. Spencer, and P. Domingos, “Mining Time-Changing Data Streams,” Proc. ACM SIGKDD Seventh Int’l Conf. Knowledge Discovery and Data Mining, pp. 97-106, 2001.
[2] W. Fan, “Systematic Data Selection to Mine Concept-Drifting Data Streams,” Proc. ACM SIGKDD 10th Int’l Conf. Knowledge Discovery and Data Mining, pp. 128-137, 2004.
[3] H. Wang, W. Fan, P.S. Yu, and J. Han, “Mining Concept-Drifting Data Streams Using Ensemble Classifiers,” Proc. ACM SIGKDD Ninth Int’l Conf. Knowledge Discovery and Data Mining, pp. 226-235, 2003.
[4] J. Gao, W. Fan, and J. Han, “On Appropriate Assumptions to Mine Data Streams,” Proc. IEEE Seventh Int’l Conf. Data Mining (ICDM), pp. 143-152, 2007.
[5] J. Kolter and M. Maloof, “Using Additive Expert Ensembles to Cope with Concept Drift,” Proc. 22nd Int’l Conf. Machine Learning (ICML), pp. 449-456, 2005.
[6] M.M. Masud, J. Gao, L. Khan, J. Han, and B.M. Thuraisingham, “Integrating Novel Class Detection with Classification for Concept-Drifting Data Streams,” Proc. European Conf. Machine Learning and Knowledge Discovery in Databases (ECML PKDD), pp. 79-94, 2009.
[7] 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,” Proc. European Conf. Machine Learning and Knowledge Discovery in Databases (ECML PKDD), pp. 337-352, 2010.
[8] A.Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavalda`, “New Ensemble Methods for Evolving Data Streams,” Proc. ACM SIGKDD 15th Int’l Conf. Knowledge Discovery and Data Mining, pp. 139-148, 2009.
[9] S. Hashemi, Y. Yang, Z. Mirzamomen, and M. Kangavari, “Adapted One-versus-All Decision Trees for Data Stream Classification,” IEEE Trans. Knowledge and Data Eng., vol. 21, no. 5, pp. 624-637, May 2009.
[10] Katakis, G. Tsoumakas, and I. Vlahavas, “Tracking Recurring Contexts Using Ensemble Classifiers: An Application to Email Filtering,” Knowledge and Information Systems, vol. 22, pp. 371-391, 2010.
[11] X. Li, P.S. Yu, B. Liu, and S.-K. Ng, “Positive Unlabeled Learning for Data Stream Classification,” Proc. Ninth SIAM Int’l Conf. Data Mining (SDM), pp. 257-268, 2009.
[12] 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,” Proc. IEEE Int’l Conf. Data Mining (ICDM), pp. 929-934, 2010.
[13] M.M. Masud, J. Gao, L. Khan, J. Han, and B.M. Thuraisingham, “Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints,” IEEE Trans. Knowledge and Data Eng., vol. 23, no. 6, pp. 859-874, June 2011.