ISSN: 2277-9655
Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Survey on Pattern Optimization for Novel Class in MCM for Stream Data Classification

Vinay Singh[*1], **Divakar Singh**[2]
[*1,2] Department of Computer Science & Engineering, BUIT, Bhopal, M.P, India
vinay.cse5@gmail.com

Abstract

The classification of stream data is somehow difficult. Existing data stream classification techniques assume that total number of classes in the stream is fixed. Therefore, instances belonging to a novel class are misclassified by the existing techniques. Because data streams have endless length, conventional multi pass learning algorithms are not appropriate as they would require infinite storage and training time. Concept-drift occurs in the stream when the underlying concept of the data changes over time. Thus, the classification model must be updated continuously so that it reflects the most recent concept. In this paper we are presenting some efficient research approaches suggested by numerous scholars.

Keywords: Data stream, multi-class miner

Introduction

Stream data classification faced a problem of new class generation during process of pattern evaluation. The evaluation process of pattern raised new pattern of data for classification. The evolving new pattern mismatches the assigned class for stream data classification, now the generated pattern creates new class for classification process. For this process of handling multi- class miner process are used. But the multi-class miner failed a process of new pattern evaluation mechanism. For the generation of pattern and optimized pattern process for improving of multi-class miner used pattern optimization technique using genetic algorithm. The optimized pattern settled the new class and improved the efficiency multi-class miner. Pattern optimization plays an important role in stream data classification. The feature evaluation process of stream data induced a problem for classification such as infinite length. So it is not possible to store the data and use it for training. Infinite length, concept-evolution and concept-drift are major challenges in data streaming.

The data stream is infinite amount of data; data continuous arrived and can only be read for one or a few times. So the faster method of data stream mining need to be updated. Data-stream mining is a technique which can find valuable information or knowledge from a great deal of primitive data. Unlike mining static databases, mining data streams poses many new challenges [1].

Data stream has different characteristics of data collection to the traditional database model. Such as the date of data stream continuous generation with time progresses and the data stream is dynamic and the arrival of the data stream cannot be controlled by the order. The data of data stream can be read and process based on the order of arrival. The order of data cannot be changed to improve the results of treatment. Therefore, the processing of the data stream requires first, each data element should be examined almost one time, because it is unrealistic to keep the entire stream in the main memory. Second, each data element in data streams should be processed as fast as possible. Third, the memory usage for mining data streams should be bounded even though new data elements are continuously generated. Finally, the results generated by the online algorithms should be instantly available when user requested [1].

Data stream compared with traditional data collection, the data stream is a real-time, continuous, orderly, time-varying, infinite tulle data stream has the following distinctive features such as orderly, Cannot Reproduce, High-Speed, Infinite, High Dimensional and Dynamic. Learning, like intelligence covers such a broad range of processes that it is difficult to define precisely. A dictionary definition includes phrases such as “to gain knowledge, or understanding of, or skill in, by study, instruction, or experience”, and “modification of a behavioural tendency by experience”. Certainly, many techniques in Machine Learning derive from the efforts of psychologists to make more precise their theories of animal and human learning through computational models. It seems likely also that the concepts and techniques being explored by researchers in Machine Learning may illuminate certain aspects of biological learning [2].

Machine Learning usually refers to the changes in systems that perform tasks associated with artificial intelligence (AI). Such tasks involve recognition, diagnosis, planning, robot control, prediction, etc. The “changes” might be either enhancements to already performing systems or abolition synthesis of new systems [2]. So, Machine Learning is a subfield of artificial intelligence that is concerned with the design and development of algorithms and techniques that allow computers to "learn". Learning can be classified broadly into Supervised Learning, Unsupervised Learning, Semi-supervised Learning, and Reinforcement Learning [3]. Support Vector Clustering is one of the kernel-based learning methods. SVC algorithm has two main steps a) SVM Training and b) Cluster Labelling. SVM training step involves construction of cluster boundaries and cluster labelling step involves assigning the cluster labels to each data point.

a. Supervised learning

Supervised learning is a learning technique for deducing a function from training data. The training data consist of pairs of input objects and desired outputs. The output of the function can be a continuous value called classification. The task of the supervised learner is to predict the value of the function for any valid input object. After having seen a number of training examples i.e. Pairs of input and target output. The learner has to generalize from the presented data to unseen situation in a reasonable way. In this every input pattern that is used to train the network is associated with an output pattern. A teacher is assumed to be present during the learning process, when a compression is made between the networks’ computed output and correct expected output, to determine the error. The error can then be used to change network.

b. Unsupervised learning

Unsupervised learning involves no target values. Unsupervised learning is very useful for data visualization .Unsupervised learning is used in a wide variety of fields such as cluster analysis. That minimizes the same error function as an auto-associative network with a linear hidden layer, trained by inputs data set. It means there is no teacher to present the desired pattern and hence the system learns by discovering and adapting to structural feature in the input pattern.

c. Semi-supervised learning

Semi-supervised learning is a class of learning techniques that make use of both labelled and unlabeled data for training, typically a small amount of labelled data with a large amount of unlabelled data. Semi supervised learning falls between unsupervised learning without any labelled training data and supervised learning with completely labelled training data. Many Machine Learning researchers have found that unlabelled data when used in conjunction with a small amount of labelled data that can produce considerable improvement in learning accuracy. The acquisition of labelled data for a learning problem often requires a skilled human agent to manually classify training examples.

d. Reinforcement learning

In this method, a teacher though available, does not present the expected answer but only indicates if the computed output is correct or incorrect. The information provided helps the network in its learning process. A reward is given for correct answer computed and a penalty for a wrong answer. Cluster analysis is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity. Intuitively, patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster [4].

Several incremental learners have been proposed to address stream data classification problem [5], [6]. In addition, concept-drift occurs in the stream when the underlying concepts of the stream change over time. A variety of techniques have also been proposed in the literature for addressing concept-drift [2], [7], [8] in data stream classification. However, there are two other significant characteristics of data streams, such as concept evolution and feature evolution that are ignored by most of the existing techniques. Concept-evolution occurs when new classes evolve in the data. For example, consider the problem of intrusion detection in a network traffic stream. If consider each type of attack as a class label, then concept-evolution happen when a completely new kind of attack take place in the traffic.

In the classification process of a developing data stream, either the short-term or long-term behaviour of the stream may be more important, or it often cannot be known a priori as to which one is more significant. How do determine the window or horizon of the training data to use so as to obtain the best classification accuracy? While techniques namely decision trees are useful for one-pass mining of data streams, these cannot be easily used in the context of an on-demand classifier in a developing environment. This is because such a classifier needed rapid variation in the horizon selection process due to data stream evolution. In this respect, nearest-neighbour classifiers tend to be more amenable to quick horizon adjustments because of their easiness. However, it is still too costly to keep track of the entire history of the data in its original precise granularity.

Concept-drift is a common phenomenon in data streams, which occurs as a result of changes in the underlying concepts. Or data changes over time concept-evolution occurs as a result of new classes evolving in the stream or new type of data may evolve in the stream that has not been seen before. For example, consider the problem of intrusion detection in a network traffic stream. If consider each type of attack as a class label, then concept-evolution occurs when a completely new kind of attack occurs in the traffic. Another example is the case of a text data stream, such as that occurring in a social network such as Twitter. In this case, new topics (classes) may frequently emerge in the underlying stream of text messages.

MCM

The multi-class miner algorithm basically consists of ensemble technique of clustering and classification [2]. The main idea in detecting multiple novel classes is to construct a graph, and identify the connected components in the graph. The number of connected components determines the number of novel classes. A data point should be closer to the data points of its own class (cohesion) and farther apart from the data points of other classes (separation). If there is a novel class in the stream, instances. For example, if there are two novel classes, then the separation among the different novel class instances should be higher than the cohesion among the same-class instances. The main idea in detecting multiple novel classes is to construct a graph, and identify the connected components in the graph. The number of connected components determines the number of novel classes.

Background

Data stream classification poses many challenges to the data mining community. In this address four such major challenges, namely, infinite length, concept-drift, concept-evolution, and feature-evolution. Since a data stream is theoretically infinite in length, it is impractical to store and use all the historical data for training. Stream data classification plays important role in the field of data mining. The need and requirement of online transaction of data is stream classification, due to stream classification save time of computation and storage area of network. For the purpose of stream data classification various machine learning algorithm are applied, such as clustering, classification, and regression. Two of the most critical and well generalized problems of data streams are its infinite length and concept-drift. Since a data stream is a fast and continuous event, it is assumed to have infinite length. Therefore, it is difficult to store and use all the historical data for training. The most discover alternative is an incremental learning technique.

Related Work

In year 2012, Mohammad M. Masud et al describe a process of stream data classification by Concept-Evolution in Concept-Drifting Data Streams as Concept-evolution occurs as a result of new classes evolving in the stream. This method addresses concept-evolution in addition to the existing challenges of infinite-length and concept-drift. The concept-evolution phenomenon is studied and the insights are used to construct a superior novel class detecting techniques. Firstly, suggest an adaptive threshold for outlier detection, which is a vital part of novel class detection. Secondly suggest a probabilistic approach for novel class detection using discrete gini Coefficient and this prove its effectiveness both theoretically and empirically [1].

Finally, address the issue of simultaneous multiple novel class occurrence and give a refined solution to detect more than one novel class simultaneously. They also consider feature evolution in text data streams which occurs because new features (i.e., words) evolve in the stream data classification. Comparison with state of the art data stream classification techniques establishes the effectiveness of the propose approach proposes an improved technique for outlier detection by defining a dynamic slack space outside the decision boundary of each classification pattern. Secondly suggest a better alternative for identifying novel class instances using discrete Gini Coefficient. Finally propose a graph-based approach for distinguishing among multiple novel classes. They apply technique on several real data streams that experience concept-drift and concept-evolution, and achieve significant performance improvements over the existing techniques [1].

Mohammad M. Masud, et al describe a process of stream data classification by novel class detection In Concept-Drifting Data Streams Under Time Constraints as Novel class detection problem becomes more challenging in the presence of concept drift, when the underlying data distributions develop in streams. In order to determine whether an instance belongs to a Novel class, the classification models sometimes require waiting for more test instances to discover similarities among those instances. To show how to make fast and accurate classification decisions under these constraints and apply them to real benchmark data. Comparing with state of the art stream classification techniques proves the superiority of this approach. Existing data stream classification techniques assume that total number of classes in the stream is set. Therefore instances belonging to a novel class are misclassified by the currently techniques. Now show how to detect novel classes automatically even when the classification model is not trained with the novel class instances. Novel class detection becomes more challenging in the presence of concept-drift [2].

Clay Woolam et al describe a process of stream data classification by Evolving Stream Data with Few Labels as It is practical to assume that only a small fraction of instances in the stream are tagged. A more practical assumption would be that the labeled data may not be independently distributed among all train documents. How can ensure that a good classification model would be built in these scenarios, considering that the data stream also has changing nature? In previous work apply semi-supervised clustering to build classification models using limited amount of labelled train data. However, it assumed that the data to be labelled should be chosen randomly [3].

Aggarwal, Charu C. et al offered a method for On-Demand Classification of Evolving Data Streams. This model indicates real-life situations effectively, since it is desirable to classify test streams in real time over an evolving training and test stream. The objective here is to make a classification system in which the training model can adapt quickly to the changes of the underlying data stream. In order to achieve this goal, they propose an on-demand classification process which can dynamically select the appropriate window of past training data to build the classifier. The empirical results show that the system maintains high classification accuracy in a developing data stream, while providing an efficient solution to the classification task [5].

In year of 2010, Valerio Grossi et al proposed a process of stream data classification by Kernel-Based Selective Ensemble Learning as Kernel methods enable the modelling of structured data in learning algorithms, still they are computationally demanding. Both efficacy and efficiency of the proposed approach are assessed for different models by using data sets exhibiting different levels and types of concept drift. Kernel methods provide a powerful tool for modelling structured objects in learning algorithms. Unfortunately, they require a high computational complexity to be used in streaming environments. This work is the first that demonstrates how kernel methods can be employed to define an ensemble approach able to quickly react to concept drifting and guarantees an efficient kernel computation [6].

Yan-Nei Law and Carlo Zanily describe a process of stream data classification by adaptive nearest classification as the algorithm achieves excellent performance by using small classifier ensembles where approximation error bounds are guaranteed for each ensemble size. The very low update cost of incremental classifier makes it highly suitable for data stream applications. ANNCAD is very suitable for mining data streams as its update speed is very quick. Also, the accuracy compares favourably with existing algorithms for mining fact streams. ANNCAD adapts to concept drift efficaciously by the exponential bury approach. However, the very detection of sudden concept drift is of interest in many applications. The ANNCAD framework can also be extended to detect concept drift, for example changes in class label of blocks is a good indicator of possible concept drift [9].

Li Su et al describe a process of stream data classification by Associative classification (AC) as Associative classification (AC) which is based on association rules has shown great promise over many other classification techniques on static dataset. Meanwhile, a new challenge has been proposed in that the increasing prominence of data streams arising in a wide range of advanced application. This technique describes and evaluates a new associative classification algorithm for data streams which is based on the estimation mechanism of the lossy Counting (LC) and landmark window model. And this technique was applied to mining several datasets obtained from the UCI Machine Learning Repository and the result show that the algorithm is effective and efficient [10].

Conclusion

The method of stream data classification generates a drift in case of stream. The garneted drift discovers a problem of computational efficiency and rate of classification. The method such as general purpose programming and probabilistic reduced the infinite length and drift problem. Furthermore evolution would be realized over the next few years to address these problems. Having these systems that address the above research issues create, that would speed up the science discovery in physical and astronomical applications in addition to business and financial ones that would improve the real-time decision making process. A multi-class miner algorithm modified using genetic algorithm. The empirical evaluation of modified algorithm is better in compression of MCM algorithm. The error rate of modified algorithm decreases in comparison of MCM algorithm. It also improved the rate of Fnew and Mnew for evolution of result. The error rate reduced 20% in comparison of MCM instate of that classification of data are improved.

References

[1] Masud M., Mohammad, Qing Chen, Khan Latifur, Aggarwal Charu C. , Jing Gao, Jiawei Han, Srivastava Ashok and Oza Nikunj C. “Classification and Adaptive Novel Class

Detection of Feature-Evolving Data Streams", in IEEE TRANSACTION -2012.

[2] Masud M. M., Gao J., Khan L., Han J., and Thuraisingham B. M. "Classification and novel class detection in concept drifting data streams under time constraints". *IEEE Trans. Knowl. Data Eng*, 23(6):859–874, 2011.

[3] Woolam Clay, Masud Mohammad M., and Khan Latifur "Lacking Labels in the Stream: Classifying Evolving Stream Data with Few Labels" in *I.J.Modern Education and Computer Science*, 2011.

[4] Bhowan Urvesh, Johnston Mark, Zhang Mengjie and Yao Xin "Evolving Diverse Ensembles using Genetic Programming for Classification with Unbalanced Data" in IEEE Transaction, 2010.

[5] Aggarwal Charu C. , Han Jiawei, Wang Jianyong, Philip S. Yu "A Framework for On-Demand Classification of Evolving Data Streams" in ECML PKDD 2010, Part II, LNAI 6322, pp. 337–352, 2010.

[6] Grossi Valerio, Sperduti Alessandro "Kernel-Based Selective Ensemble Learning for Streams of Trees", in Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence 2010.

[7] Masud Mohammad M., Chen Qing, Khan Latifur, Aggarwal Charu, Gao Jing , Han Jiawei and Thuraisingham Bhavani "Addressing Concept-Evolution in Concept-Drifting Data Streams", in IEEE Transaction 2010.

[8] Katakis I., Tsoumakas G., and Vlahavas I. "Tracking recurringcontexts using ensemble classifiers: an application to email filtering", *Knowledge and Information Systems*, vol. 22, pp. 371–391, 2010.

[9] Law Yan-Nei and Zanily Carlo "An Adaptive Nearest Neighbor Classification Algorithm for Data Streams" in PKDD 2005, LNAI 3721, pp. 108–120, 2005.

[10] Su Li, Liu Hong-yan, Song Zhen-Hui. "A New Classification Algorithm for Data Stream", International Journal of Modern Education and Computer Science, Vol. 4, pp. 32-39, 2011.