www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 1 January, 2014 Page No.3778-3783

A Review of Method of Stream data classification through Optimized Feature Evolution Process

Archana Bopche[1], Malti Nagle[2], Hitesh Gupta[3]

[1]Department of Computer Science & Engineering
PIES, Bhopal, India
archanabopche@gmail.com

[2]Professor, Faculty of P.G.Depatrment of Computer science & Engineering,
PCST, Bhopal, India
nagle.malti083@gmail.com

[3]Professor, Faculty of P.G.Depatrment of Computer science & Engineering
PCST, Bhopal, India
hitesh034@gmail.com

Abstract: ***Dynamic changing nature of stream data are induced a difficulty of training pattern and process of class labeling in classification. The stream data classification has some difficulty such as feature evaluation, data drift, concept evaluation and infinite length. The infinite length and feature evaluation is more realistic problem in stream data classification technique. Different authors used different method such as data miner and tree based approach for reduced such types of issues. In this paper we discuss the stream data classification process and method of these classification techniques. Optimized feature evaluation process reduces the feature evaluation using ensemble technique. The feature evaluation process generates a new class label for classification. Some authors used optimization technique such as neural network, ANT colony optimization and genetic algorithm. The feature optimization technique improved the performance of stream data classification. Optimization technique for feature evaluation process also discussed in this paper.***

Keywords: Stream data, classification technique and optimization technique.

1. Introduction

Data streams have several unique properties due to which data stream classification is more difficult than classifying stationary data. Primary, data streams are implicit to have infinite length, which makes it unfeasible to store and use all the historical data for training. Therefore, habitual multi-pass learning algorithms are not directly applicable to data streams. Second, data streams sense concept-drift, which occurs when the underlying concept of the data changes over time. In order to deal with concept-drift, a classification model must endlessly get a feel for itself to the most recent idea. Third, data streams also detect concept-evolution, which occurs when a novel class appears in the stream. In order to handle with concept-evolution, a classification model must be able to mechanically detect novel classes when they appear, before being trained with the labeled cases of the novel class. Finally, high speed data streams go through with insufficient labeled data. Data stream classifiers may either be single model incremental approaches, or ensemble techniques, in which the classification output is a function of the predictions of Different classifiers. Ensemble techniques have been more popular than their single model counterparts because of their simpler implementation and higher efficiency [2]. Most of these ensemble techniques use a chunk-based approach for learning [4] in which they divide the data stream into chunks, and train a model from one chunk. The data points in the stream may or may not have a fixed feature set. If they have a fixed feature set, then we simply use that feature set. Otherwise, we apply a feature extraction and feature selection technique. Note that we need to select features for the instances of the test chunk before they can be classified by the existing models, since the classification models require the feature vectors for the test instances. However, since the instances of the test chunk are unlabeled, we cannot use supervised feature selection (e.g. information gain) on that chunk. With the advent of advanced data streaming technologies [5], we are now able to continuously collect large amounts of data in various application domains, e.g., daily fluctuations of stock market, traces of dynamic processes, credit card transactions, web click stream, network traffic monitoring, position updates of moving objects in location-based services and text streams from news etc [1]. Due to its potential in industry applications, data stream mining has been studied intensively in the past few years. In particular, much research has been focused on classifying data streams. Many features might be irrelevant and possibly is not favorable to classification. Also, redundancy among the features is common [7]. The presence of irrelevant and redundant features not only slows down the learning algorithm but also confuses it by

causing it to over fit the training data. In other words, ignoring (or) removing irrelevant and redundant features make the classifier's design simple, improves its prediction performance and its computational efficiency [8].Multiclass miner is combination of OLINDDA and FAE approach. This combination work with dynamic feature vector and detect novel class. OLINDDA is used to detect the novel class and FAE classifies the data chunks. MCM detects outlier classification and also used in recognizing the novel class instances. MCM is the fastest method in all datasets. MCM is roughly 25% faster than Mine Class. The reason for faster running time is it uses the dynamic thresholding and Gini coefficient analysis; MCM filters out the majority of the outliers and reduces the cost of novel class detection because it is proportional to the number of outliers. Genetic algorithm is also advantageous to the data streaming. This genetic algorithm is used to rectify the problem of segmenting the data stream. Properly segmented streams can be better arranged and reused. They provide points of access that facilitate browsing and retrieval. Data stream classification has many approaches[6]. These approaches fall into two categories: single model and ensemble classification. Single model classification techniques uphold and incrementally renew a single classification model and effectively react to concept-drift. The above section discuss introduction of stream data classification and feature optimization. In section II we describe related work of stream data classification. In section III method of stream data classification. In section IV discuss problem in stream data classification and our approach and finally conclude in section V.

2. RELATED WORK

In this section explain related work of stream data classification using various techniques such as multi-class miner and data miner for minimized the problem of infinite length and feature evaluation problem. Feature evaluation decides the process of concept evaluation for generation of new class for classification purpose. The process of data labeling for classification purpose also suffered from problem of data drift. all these process discuss here.

[1] In this paper author describes a stream data which is based on support vector machine, and the details are we propose a novel data stream clustering algorithm, termed SV Stream, which is based on support vector domain description and support vector clustering. In the proposed algorithm, the data elements of a stream are mapped into a kernel space, and the support vectors are used as the summary information of the historical elements to construct cluster boundaries of arbitrary shape. To adapt to both dramatic and gradual changes, multiple spheres are dynamically maintained, each describing the corresponding data domain presented in the data stream. By allowing for bounded support vectors (BSVs), the proposed SVStream algorithm is capable of identifying overlapping clusters. A BSV decaying mechanism is designed to

[2] In this paper author describes a stream data classification, and the details are we propose an ensemble classification framework, where each classifier is equipped with a novel class detector, to address concept-drift and concept evolution. To address feature-evolution, we propose a feature set homogenization technique. We also enhance the novel class detection module by making it more adaptive to the evolving stream, and enabling it to detect more than one novel class at a time. Comparison with state-of-the-art data stream classification techniques establishes the effectiveness of the proposed approach. We propose a classification and novel class detection technique for concept-drifting data streams that addresses four major challenges, namely, infinite length, concept-drift, concept-evolution, and feature evolution. The existing novel class detection techniques for data streams either do not address the feature-evolution problem or suffer from high false alarm rate and false detection rates in many scenarios.

[3] In this paper author describes a challenges for stream data classification, and the details are We present ActMiner, which addresses four major challenges to data stream classification, namely, infinite length, concept-drift, concept-evolution, and limited labeled data. Most of the existing data stream classification techniques address only the infinite length and concept-drift problems[3]. Our previous work, MineClass, addresses the concept-evolution problem in addition to addressing the infinite length and concept-drift problems. Concept-evolution occurs in the stream when novel classes arrive. However, most of the existing data stream classification techniques, including MineClass, require that all the instances in a data stream be labeled by human experts and become available for training. This assumption is impractical, since data labeling is both time consuming and costly. Therefore, it is impossible to label a majority of the data points in a high-speed data stream. This scarcity of labeled data naturally leads to poorly trained classifiers. ActMiner actively selects only those data points for labeling for which the expected classification error is high. Therefore, ActMiner extends MineClass, and addresses the limited labeled data problem in addition to addressing the other three problems. It outperforms the state-of-the-art data stream classification techniques that use ten times or more labeled data than ActMiner.

[4] In this paper author studies on classification and the details are how to devise PU learning techniques for the data stream environment. Unlike existing data stream classification methods that assume both positive and negative training data are available for learning, we propose a novel PU learning technique LELC (PU Learning by Extracting Likely positive and negative micro-Clusters) for document classification. LELC only requires a small set of positive examples and a set of unlabeled examples which is easily obtainable in the data stream environment to build accurate classifiers. Experimental results show that LELC is a PU learning method that can effectively address the issues in the data stream environment with significantly better speed and accuracy on capturing concept drift than the existing state-of-the-art PU learning techniques.

[5] In this paper author describes a feature selection and extraction techniques, the data points in the stream may or may not have a fixed feature set. If they have a fixed feature set, then we simply use that feature set. Otherwise, We apply a feature extraction and feature selection technique. We predict the features of the test instances without using any of their information, rather we use the past labeled instances to predict the feature set of the test instances. There are two Different approaches: single model classification, and ensemble classification. The single model classification techniques apply

some form of incremental learning to address the infinite length problem, and strive to adapt themselves to the most recent concept to address the concept-drift problem. Ensemble classification techniques maintain a fixed-sized ensemble of models, and use ensemble voting to classify unlabeled instances. These techniques address the infinite length problem by applying a hybrid batch-incremental technique.

[6] In this paper author review a feature selection process, and the details are Feature selection is an effective technique for dimension reduction and an essential step in successful data mining applications. It is a research area of great practical significance and has been developed and evolved to answer the challenges due to data of increasingly high dimensionality. Its direct benefits include: building simpler and more comprehensible models, improving data mining performance, and helping prepare, clean, and understand data. We first brifley introduce the key components of feature selection, and review its developments with the growth of data mining. We then overview FSDM and the papers of FSDM10, which showcases of a vibrant research field of some contemporary interests, new applications, and ongoing research efforts. We then examine nascent demands in data-intensive applications and identify some potential lines of research that require multidisciplinary efforts.

[7] In this paper author presents an adaptive feature selection process for multi class classification, we propose an adaptive feature selection method for multiclass classification task. With our method, the "siren pitfall" could be avoided, the selected features could be re-produced, the feature selection scheme does not rely on any earlier knowledge, and the corresponding calculation cost is less. The Experimental results shows the effectiveness of our adaptive feature selection process. the strongly predictive features for the few "easy" classes rank before the weakly predictively features for the remaining "difficult" classes. As a result, the features that are necessary for discriminating "difficult" classes would be ignored by traditional feature scoring methods. This problem is called the "siren pitfall".

[8] Author proposes here a nearest neighbor classification of data, the details are we propose a taxonomy based on the main characteristics presented in prototype selection and we analyze their advantages and drawbacks. Empirically, we conduct an experimental study involving different sizes of data sets for measuring their performance in terms of accuracy, reduction capabilities, and runtime. The results obtained by all the methods studied have been verified by nonparametric statistical tests. Several remarks, guidelines, and recommendations are made for the use of prototype selection for nearest neighbor classification. To propose a complete taxonomy based on the main properties observed using the PS methods. The taxonomy will allow us to know the advantages and drawbacks from a theoretical point of view. To make an empirical discussion for analyzing the methods in terms of accuracy, reduction capabilities, and time complexity. The objective is to an identify the best methods in each family and to stress the relevant properties of each one.

[9] In this paper Author introduce an adaptive fuzzy neural network framework for classification of data stream using a partially supervised training algorithm. The structure consists of an evolving granular neural network capable of processing non stationary data streams using a one-pass incremental algorithm. The granular neural network evolves fuzzy hyper boxes and uses null norm based neurons to classify data. The learning algorithm performs structural and parametric adaptation whenever environment changes are reflected in input data. It needs no prior statistical knowledge about data and classes. Computational experiments show that the fuzzy granular neural network is robust against different types of concept drift, and is able to handle unlabeled examples efficiently.

[10] In This paper author illustrates a novel web usage mining method, depend on the sequence mining technique used to user's navigation behavior, to find pattern in the routing of websites. Three critical contributions are made in this paper: using the footstep graph to visualize the user's click-stream data and any interesting pattern can be detected more easily and quickly; illustrating a novel sequence mining approach to identify pre-designated user navigation patterns automatically and integrates back-propagation network (BPN) model smoothly; and applying the empirical research to indicate that the proposed approach can predict and categorize the users' navigation behavior with high accuracy. The initial design for identifying the user's prior knowledge for specific products. Our method is based on the customer's on-line navigation behaviors by analyzing their navigation patterns through web mining and constructing artificial neural networks to predict potential customers' need in the future.

[11] In this paper author proposed work a novel approach which uses an Intuitionist fuzzy version of k-means has been introduced for grouping based interdependent features. Genetic algorithm reinstate which is the variation of traditional genetic algorithm is then applied to appraise whether the measured feature is independent of class labels, therefore it lead to remove unrelated clusters to classification process and progress the selection of features subset. The proposed approach achieves improvement on classification accuracy and perhaps to select less number of features which show the way to simplification of learning task to a huge amount. The Implementation results have been demonstrated by the good performance and also find good enough subset features of this method on using UCI benchmark datasets that are for data mining methods such as Breast Cancer, Sensor and Iris Records.

[12] In this paper author describe the classification method, they would like to address diversity in ensemble approaches and propose an effective ensemble approach by considering further diversity in genetic programming. A set of classification rules was generated by genetic programming, and then diverse ones were selected from among them in order to construct an ensemble classifier. In contrast to the conventional approaches, diversity was measured by matching the structure of the rules based on the interpretability of genetic programming. An effective ensemble approach that does use diversity in genetic programming technique is proposed. This multiplicity is calculated by comparing the structure of the classification rules instead of output-based diversity estimating.

[13] In this paper author presents the classifiers to satisfy accuracy performance necessities, as functional to a real world classification problem. The optimal assessment fusion

approach is found to perform significantly better than the conventional classifier fusion methods, i.e., traditional decision level fusion and averaged sum rule, then parameters of the selected classifiers are optimized so that the specified performance is met. For certain real-world classification problems, this single classifier design approach may fail to meet the desired performance even after all parameters/architectures of the classifier have been fully optimized.

3. METHOD OF STREAM DATA CLASSIFICATION

In this section gives comparative table for method used in stream data classification. The method also compared with classification accuracy and problem oriented method adopted.
The table 3.1 shows the method and method demerits used in stream data classification technique. In all these technique the multi-class miner with genetic algorithm is perform better performance in comparisons of all technique. The method table gives the little bit information of feature optimization in feature evaluation process. The optimized feature resolves the problem of feature evaluation method and infinite length problem.

4. APPROACH USED FOR MULTI-CLASS MINER

Data stream classifier conciliation with selection of optimal features selection. The selection of optimal features a problem of fixed number of features generation using multi-class miner. The fixed number of features technique induces a problem of data imbalance in classification and pattern recognition. The unbalance data ratio issue creates a problem of minority and majority voting principle. The selection of features in classification is chock point. An another approach to producing diverse binary classifiers involving random sampling over the feature space was proposed in [12]. In this work each base classifier was generated with a randomly selected subset of features. The final ensemble with a combining voting technique was able to improve performance in comparison with the binary base classifiers. An alike method based on this concept was presented by [13]. In this work each binary model was built by applying a different subset of features. Each feature had a weight assigned to reflect its relevance to the problem being considered. A weighted Euclidean distance metric was applied while picking out neighbors. To make the final decision a voting scheme was applied among the instances selected by all the binary in the ensemble. Subsequent evaluations it was found that the proposed approach based on sampling over the feature space provided a significant improvement in comparison with a single binary classifier. Some current problem found in review process.

1. Unbalanced ratio of train and test data [10]
2. Selection of optimal features for ensemble classifier [1, 3]
3. Diversity of feature selection process [12]
4. Boundary value of features [11]
5. Outlier data treat as noise [9].

Features oriented data classification well knows method for stream data classification. In features oriented ensemble classifier is suffered from a selection of optimal number of features selection technique. The selection of optimal number of features enhanced the process of features oriented ensemble classification for data classification. The optimality of features is selected by Meta function. For this process we used PSO technique. The PSO is an efficient global optimizer for continuous

Table 3.1

Method	Approach	Demerits	Accuracy (%)
MCM(multi-class miner)	Ensemble technique	Optimal selection of cluster for classification.	91
DX Miner	Dynamic feature evaluation concept is used.	Ambiguous feature space are generated and data conversion loss are occurred	92
SVSTREAM(support vector clustering)	Cluster labeling technique are used	Noise outlier and boundary value decrease the performance of cluster	90
Associative classification	Rule mining technique is used.	Lower value of data are not considered	87
Kernel-Based Selective Ensemble	modeling of structured data in learning algorithms	Variance of kernel function degraded the performance of classifier	89
MCM-GA(multi-class miner with genetic algorithm)	Optimization process are done for new feature evaluation technique	Population of genetic algorithm are work in limited constraints of data.	93
ANNCAD	using small classifier ensembles ensemble size	Concept evaluation are not considered	89

Variable problems .The advantages of the PSO are very few parameters to deal with and the large number of processing elements, so called dimensions, which enable to fly around the solution space effectively, more over PSO is meta-heuristic function inspired by real biological animals. The fitness constrains of PSO is multiple. Using particle of swarm optimization we maintain the selection process of feature technique and noise removal of boundary base class. Noise reduction and selection of optimal number of features in ensemble classifier used features index selection process using ant colony optimization technique. We introduce a new feature sub set selection method for finding similarity matrix for features without alteration of ensemble classifier. The proposed features index selection method based on ant colony optimization, PSO technique find the most similar features index for ensemble of classifier. In this method we introduced continuity of PSO for similar features and dissimilar features collect into next node. In that process PSO find optimal selection of features index. Suppose ants find features of

similarity in continuous root. Every ant of features compares their property value according to initial features set.

5. CONCLUSION AND FUTURE SCOPE

In this paper we review a various method of stream data classification and handling a problem during stream classification, such as infinite length of data, feature evaluation, concept evaluation and data drift of stream data. 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 common purpose programming and problastic reduced data drift and infinite length problem. Also all these ways take a process of optimization for the solution of feature evaluation problem. Mining data streams is still in its early life state. Handled along with open issues in data stream mining are discussed in this paper. Further developments would be realized over the next few years to address these problems. Having these systems that address the above research issues developed, 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.

References:-

[1] Chang-Dong Wang, Jian-Huang Lai, Dong Huang, Wei-Shi Zheng "SVStream: A Support Vector-Based Algorithm for Clustering Data Streams" TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING IEEE Vol- 25, 2013. pp. 1410-1424.

[2] Mohammad M. Masud,Qing Chen,Latifur Khan,Charu C. Aggarwal "Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams" TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE 2011. pp. 1-14.

[3] Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, Bhavani Thuraisingham "Classification and Novel Class Detection in Data Streams with Active Mining" Springer 2010. pp. 311-324.

[4] [4] Xiao-Li Li, Philip S. Yu, Bing Liu ,See-Kiong Ng "Positive Unlabeled Learning for Data Stream Classification" SIAM 2010. pp. 259-270.

[5] Mohammad M. Masud, Qing Chen, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham "Classification and Novel Class Detection of Data Streams in a Dynamic Feature Space" springer 2010. pp. 337-352.

[6] Huan Liu, Hiroshi Motoda, Rudy Setiono, Zheng Zhao "Feature Selection: An Ever Evolving Frontier in Data Mining" Fourth Workshop on Feature Selection in Data Mining, 2010. pp. 1-10.

[7] Xin Xu, Wei Wang, Guilin Zhang, Yongsheng Yu "An Adaptive Feature Selection Method for Multi-class Classification" IEEE 2011. pp. 225-233.

[8] Salvador Garcıa, Joaquın Derrac "Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study" IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 34, 2012. pp. 417-435.

[9] Daniel Leite, Pyramo Costa Jr., Fernando Gomide "Evolving Granular Neural Network for Semi-supervised Data Stream Classification" IEEE World Congress on Computational Intelligence, 2010. pp. 1877-1885.

[10] Pao-Hua Chou , Pi-Hsiang Li , Kuang-Ku Chen , Menq-Jiun Wu "Integrating web mining and neural network for personalized e-commerce automatic service" Expert Systems with Applications, 2010. pp. 2898-2910.

[11] S. Senthamilarasu, M. Hemalatha " A Genetic Algorithm Based Intuitionistic Fuzzification Technique for Attribute Selection" Indian Journal of Science and Technology, 2013. pp. 4336-4346.

[12] Jin-Hyuk Hong, Sung-Bae Cho "The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming" Artificial Intelligence in Medicine, 2006. pp. 43-58.

[13] Kalyan Veeramachaneni, Weizhong Yan, Kai Goebel, Lisa Osadciw "Improving Classifier Fusion Using Particle Swarm Optimization" IEEE Symposium on Computational Intelligence in Multicriteria Decision Making, 2007. pp. 128-135.