International Journal of Master of Engineering Research and Technology, 2394-6172-(O)
Volume-5, Issue-4, April 2018

A Review of Stream Data Classification Based on Feature Optimization

Mamata Mishra
Department CSE
SVCST, Bhopal (MP)

Mr. Amit Thakur
Department CSE
SVCST, Bhopal (MP)

Abstract

The continuity of stream data is major challenge for the process of data classification and categorization. The continuity of data creates some problem such as infinite length, feature evaluation, data drift. The data drift raised the problem of mapping of class. in this paper present the review of stream data classification based on feature optimization. The process of feature optimization reduces the problem of data drift and improve the classification ratio of classifier. In this paper present the review of stream data classification using different feature optimization technique.

Keywords: - Stream data, Classification technique, and Optimization technique

INTRODUCTION

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 [10].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. 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. Several incremental learners have been proposed to address this problem [8], [5]. 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], [6], [7] 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. The feature optimization is the process of attribute reeducation process. The above section discusses 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.

II RELATED WORK

In this section describe 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

Paper ID:- 2018/04/IJMERT/5/321

classification purpose also suffered from problem of data drift. all these processes discuss here.

Srilakshmi Annapoorna P.V and Mirnalinee T.T Et al. [1] In this work, a novel algorithm has been implemented using Random Forest with stratified random sampling and Bloom filtering in order to reduce the training time and to handle high velocity data. Experimental results are shown by performing classification with sampling, classification with filtering and classification with sampling and filtering. This enhances the performance of the algorithm by decreasing the training time and testing time of the classifier with negligible compromise in accuracy of classification.

Mahardhika Pratama, Sreenatha G.Anavatti, Meng-joo Er and Edwin Lughofer Et al. [2] This paper presents a new online evolving classifier, termed Parsimonious classifier (pClass), where the major exposure is a formidable classifier, workable in the online real time situations. pClass discussed some new aspects of learning from streaming data, where new avenues to extract the fuzzy rules and to dwindle the rule base complexity from streaming data are discussed. Alongside with that, a rule recall mechanism is devised to overcome the contingency of the cyclic drift in the data distribution and a novel scheme of feature weighting is lodged, which is built upon the online FSC in the empirical feature space.

Poonam Sonar and Udhav Bhosle and Chandrajit Choudhury Et al. [3] In this paper machine learning based mammogram classification using modified hybrid SVM-KNN is discussed. The idea is to map the feature points to kernel space using kernel and find the K nearest neighbors among the training dataset for a given test data point. Doing this they narrow down the search for support vectors. Mammogram images are preprocessed and region of interest is extracted using Fuzzy C Means clustering and Active Counter technique. GLCM (grey level covariance matrix) based texture features are extracted from segmented ROI. These features are used to train modified hybrid SVM-KNN classifier discussed by authors.

Michal Wozniak,Pawel Ksieniewicz, Boguslaw Cyganek, Andrzej Kasprzak and Krzysztof Walkowiak Et al. [4] they may produce the model on the basis a few learning objects only and then they use and improve the classifier when new data comes. This concept is still vibrant and may be used in the plethora of practical cases. The paper presented the active learning of data stream classifier. The computer experiments on several benchmark data stream confirmed that discussed method can adapt to changes and their intuition has not let us down, that the semi-supervised learning (especially based on active learning) may return the similar results as the fully supervised approach.

Jonathan A. Cox, Conrad D. James and James B. Aimone Et al. [5] they evaluate the performance for three input spaces consisting of the power spectral density, byte probability distribution and sliding-window entropy of the byte sequence in a file. By combining all three, they trained a deep neural network to discriminate amongst nine common data types found on the Internet with 97.4% accuracy.

Heng Wang and Zubin Abraham Et al. [6] The paper presents a concept drift detection framework (LFR) for detecting the occurrence of a concept drift and identifies the data points that belong to the new concept. The versatility of LFR allows it to work with both batch and stream datasets, imbalanced data sets and it uses user-specified parameters that are intuitively comprehensible, unlike other popular concept drift detection approaches.

Nemanja Djuric, Hao Wu, Vladan Radosavljevic, Mihajlo Grbovic and Narayan Bhamidipati Et al. [7] they consider the problem of learning distributed representations for documents in data streams. The documents are represented as low-dimensional vectors and are jointly learned with distributed vector representations of word to-kens using a hierarchical framework with two embedded neural language models. They described a general unsupervised learning framework to model the latent structure of streaming documents, that learns low-dimensional vectors to represent documents and words in the same latent space.

Sara del Rıo, Victoria Lopez, Jose Manuel Benıtez and Francisco Herrera Et al. [8] In this work they have presented a linguistic fuzzy rule-based classification algorithm for big data problems called Chi-FRBCS-BigData. This algorithm obtains an interpretable model that is able to handle big collections of data providing a good accuracy and with fast response times. The performance of the Chi-FRBCS-BigData alter-natives is supported by an experimental study that is carried out over six classification big data problems. The results obtained show that the proposal is able to handle these problems providing competitive results.

Simon Fong, Raymond Wong and Athanasios V. Vasilakos Et al. [9] they discussed a novel lightweight

feature selection method by using Swarm Search and Accelerated PSO, which is supposed to be useful for data stream mining. The evaluation results showed that the incremental method obtained a higher gain in accuracy per second incurred in the pre-processing. The contribution of this paper is a spectrum of experimental insights for anybody who wishes to design data stream mining applications for big data analytics using lightweight feature selection approach such as Swarm Search and APSO.

Arati Kale and M.D. Ingle Et al. [10] Traditional data stream classifier only addresses Infinite Length and Concept Drift. In this paper they discussed ensemble classification framework where each classifier is equipped with novel class detector to address Concept Drift and Concept Evolution. Also increases accuracy of novel class detection techniques by using SVM based polynomial kernel.

Fan Zhang, Junwei Cao, Samee U. Khan, Keqin Li and Kai Hwang Et al. [11] In this paper, they discussed a task-level adaptive MapReduce framework. This framework extends the generic MapReduce architecture by designing each Map and Reduce task as a consistent running loop daemon. The beauty of this new framework is the scaling capability being designed at the Map and Task level, rather than being scaled from the compute-node level. This strategy is capable of not only scaling up and down in real time, but also leading to effective use of compute resources in cloud data center.

Nawel Yala, Belkacem Fergani and Anthony Fleury Et al. [12] This paper discussed an approach of human activity recognition on online sensor data. they present four methods used to extract features from the sequence of sensor events. their experimental results on public smart home data show an improvement of effectiveness in classification accuracy. Results show that the impact of the discussed methods compared to Baseline is reduced when is taken into account "other activity" class to learning, especially on Aruba dataset. There is a large confusion between "other activity" class and the different known activities.

Indre Žliobaite, Albert Bifet, Jesse Read, Bernhard Pfahringer and Geoff Holmes Et al. [13] This paper formalizes a learning and evaluation scheme of such predictive models. they theoretically analyze evaluation of classifiers on streaming data with temporal dependence. their findings suggest that the commonly accepted data stream classification measures, such as classification accuracy and Kappa statistic, fail to diagnose cases of poor performance when temporal dependence is present, therefore they should not be used as sole performance indicators.

III. OPTIMIZATION ALGORITHM COMPARATIVE STUDY

In this section present the comparative study of optimization algorithm. the comparative table focus on the process of optimization and selection of fitness constraints function.

Genetic Algorithm (GA)	Particle Swarm Optimization (PSO)	Ant Colony Optimization (ACO)
It is generally used to solve complex optimization problem as it can handle both detached and incessant variable, and nonlinear objective and constraint function without requiring the minute information. Simple genetic algorithm is given by: 1. Generate the population randomly 2. By using the fitness function, select parents 3. Apply crossover on the parent chromosomes 4. Mutate the offspring chromosomes	PSO, a heuristic search technique is inspired from the collaboration behavior of biological population or collective intelligence in biological population. PSO is similar to GA as they are evolutionary in nature and population-based methods. The basic set of steps of PSO is given by: 1. The swarm is initializing from the solution space 2. The fitness value of the	It is a population based general search technique which is used for difficult combinatorial problem, which is inspired by the pheromone trial laying behavior of real ant colonies. The ants, which are the search agents, search for a good solution to a given optimization problem. The basic steps in ACO are: 1. Represent the development of the solution by a construction graph 2. The parameters are initialized 3. From, each ant's

5. Append the offspring to the pool 6. Perform Elitism (Select parents)	individual particles is estimated 3. Modify gbest, pbest and velocity 4. Individual particles are moved to a new position 5. Go to Step 2, and repeat till the agreement or a stopping condition is satisfied	random walk, a random solution is generated 4. Update pheromone intensities 5. Go to step 3, and repeat until stopping condition is satisfied

IV. CLASSIFICATION ALGORITHM

Classification is a data mining technique used to predict group membership for data instances [4]. In this section, we present the basic classification techniques. Several major kinds of classification method including decision tree induction, Bayesian networks, rule based classifier, neural network, k-nearest neighbor classifier and support vector machine.

DECISION TREES

Decision trees are simple knowledge representation and they classify examples to a finite number of classes, the nodes are labeled with attribute names, the edges are labeled with possible values for this attribute and the leaves labeled with different classes. Objects are classified by following a path down the tree, by taking the edges, corresponding to the values of the attributes in an object.

BAYESIAN NETWORKS

A Bayesian Network (BN) is a graphical model for probability relationships among a set of variables features. The Bayesian network structure *S* is a directed acyclic graph (DAG) and the nodes in *S* are in one-to-one correspondence with the features X [6]. The arcs represent casual influences among the features while the *lack* of possible arcs in *S* encodes conditional independencies. Moreover, a feature (node) is conditionally independent from its non-descendants given its parents (*X*1 is conditionally independent from *X*2 given *X*3 if P (*X*1|*X*2, *X*3) =P (*X*1|*X*3) for all possible values of *X*1, *X*2, *X*3)

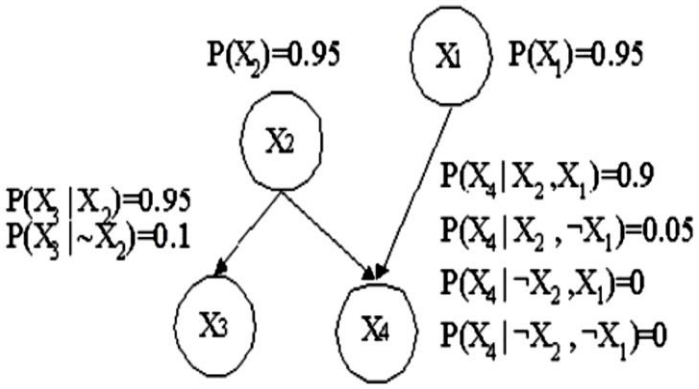

Figure 1: Shows that the calculation value of the attribute.

The structure of Bays networks typically, the task of learning a Bayesian network can be divided into two subtasks: initially, the learning of the DAG structure of the network, and then the determination of its Parameters. Probabilistic parameters are encoded into a set of tables, one for each variable, in the form of local conditional distributions of a variable given its parents.

NEURAL NETWORK

Neural networks are an approach to computing that involves developing mathematical structures with the ability to learn. The methods are the result of academic investigations to model nervous system learning. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques [10]. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyze. This expert can then be used to provide projections given new situations of interest and answer "what if" questions.

KNN CLASSIFIER

Nearest neighbor classifiers are based on learning by analogy. The training samples are described by n

dimensional numeric attributes. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space [14]. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample. "Closeness" is defined in terms of Euclidean distance, where the Euclidean distance, where the Euclidean distance between two points, X=(x1,x2,……,xn) and Y=(y1,y2,….,yn) is dist (X1,X2)=

The unknown sample is assigned the most common class among its k nearest neighbors. When k=1, the unknown sample is assigned the class of the training sample that is

IV. SUPPORT VECTOR MACHINES (SVM) ALGORITHM

A Support Vector Machine (SVM) separates the data into two categories of performing classification and constructing an N-dimensional hyper plane. These models are closely related to neural networks [5]. In fact, this model uses a sigmoid kernel function which is equivalent to a two-layer, perception neural network. These models are closely related to classical multilayer perception neural networks. By using a kernel function, these are an alternative training method for polynomial, radial basis function and multi-layer perception classifiers in which the weights of the network are found by solving a quadratic programming problem with linear constraints, rather than by solving a non-convex, unconstrained minimization problem as in standard neural network training. in the SVM literature, a predictor variable which is called an attribute and a transformed attribute that is used to define the hyper plane is called a feature[11]. Here, choosing the most suitable representation can be taken as feature selection. A set of features that describes one case (i.e., a row of predictor values) is called a vector. The goal of this modeling is to find the optimal hyper plane which separates clusters of vector in such a way that cases with one category of the target variable are on one side of the plane and cases with the other category are on the other size of the plane. The vectors near the hyper plane are the support vectors.

V. CONCLUSION AND FUTURE SCOPE

The stream data classification is critical data classification fields. For the classification of stream data used some optimization techniques. The optimization techniques control the process of feature evaluation in stream data classification. In this paper study of various paper related to feature optimization and classification of stream data. the optimization technique finds the new possibility of class for the miner of classification. The classification algorithm such as support vector machine and neural network gives better classification ratio in compression of KNN, Decision Tree and NB algorithm. The NB algorithm basically based on statically probability based algorithm. The estimation of probability is not good the classification process suffered. Instead of that the KNN algorithm based on similarity measure based on distance formula. In future used dynamic population-based glow worm optimization algorithm for the feature optimization.

References

[1] Srilakshmi Annapoorna P.V and Mirnalinee T.T "Streaming Data Classification", Trends in Information Technology, 2016, Pp 1-7.

[2] Mahardhika Pratama, Sreenatha G.Anavatti, Meng-joo Er and Edwin Lughofer "pClass : An Effective Classifier for Streaming Examples", IEEE, 2015, Pp 1-19.

[3] Poonam Sonar and Udhav Bhosle and Chandrajit Choudhury "Mammography Classification Using Modified Hybrid SVM-KNN", International Conference on Signal Processing and Communication, 2017, Pp 305-311.

[4] Michal Wozniak,Pawel Ksieniewicz, Boguslaw Cyganek, Andrzej Kasprzak and Krzysztof Walkowiak "Active Learning Classification of Drifted Streaming Data", Procedia Computer Science, 2016, Pp 1724–1733.

[5] Jonathan A. Cox, Conrad D. James and James B. Aimone "A Signal Processing Approach for Cyber Data Classification with Deep Neural Networks", Procedia Computer Science, 2015, Pp 349 – 354.

[6] Heng Wang and Zubin Abraham "Concept Drift Detection for Streaming Data", arXiv, 2015, Pp 1-9.

[7] Nemanja Djuric, Hao Wu, Vladan Radosavljevic, Mihajlo Grbovic and Narayan Bhamidipati "Hierarchical Neural Language Models for Joint

Representation of Streaming Documents and their Content", arXiv, 2016, Pp 1-8.

[8] Sara del Rıo, Victoria Lopez, Jose Manuel Benıtez and Francisco Herrera "A MapReduce Approach to Address Big Data Classification Problems Based on the Fusion of Linguistic Fuzzy Rules", International Journal of Computational Intelligence Systems, 2015, Pp 422-437.

[9] Simon Fong, Raymond Wong and Athanasios V. Vasilakos "Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data", IEEE, 2015, Pp 1-14.

[10] Arati Kale and M.D. Ingle "SVM based Feature Extraction for Novel Class Detection from Streaming Data", International Journal of Computer Applications, 2015, Pp 1-3.

[11] Fan Zhang, Junwei Cao, Samee U. Khan, Keqin Li and Kai Hwang "A task-level adaptive MapReduce framework for real-time streaming data in healthcare applications", Future Generation Computer Systems, 2015, Pp 149–160.

[12] Nawel Yala, Belkacem Fergani and Anthony Fleury "Feature extraction for human activity recognition on streaming data", IEEE, 2015, Pp 1-6.

[13] Indre˙ Žliobaite, Albert Bifet, Jesse Read, Bernhard Pfahringer and Geoff Holmes "Evaluation methods and decision theory for classification of streaming data with temporal dependence", Springer, 2015, Pp 1-28.