EXPERIMENTAL INVESTIGATION ON HDSC CLASSIFICATION MODEL

Dr.S.Jayanthi.
Professor,
Department of Computer Science and Engineering,
Samskruti College of Engg and Technology,
Ghatkesar – 501 301, Hyderabad, India.
nigilakash@gmail.com

ABSTRACT - Data Stream Classification is an inevitable task in almost digitized sectors. However, data stream classification task confronts many issues by which the efficacy of the classification task is getting flawed. This research work expounds a novel Hybrid Data Stream Classifier (HDSC) which combines the features of Support Vector Machine, fuzzy logic and Lagrangian interpolation method. Support Vector Machine (SVM) is a prominent classifier for performing supervised classification on static data. Its accuracy in handling linear as well as non-linear data attracts many researchers in probing whether, and if so, to what extent, the SVM algorithm acts upon the dynamic data stream classification process, if optimized.

The proposed research work investigates if and how the ensemble of fuzzy logic, SVM and Lagrangian interpolation method, which formulates HDSC, contributes in enriching the data stream classification process. The proposed classifier approach is deployed in a real time video server providing video services through the web and there by its efficacy is accentuated over various classifier evaluation metrics in a data stream classification process.

Key words: Fuzzy Logic, Hybrid Data Stream Ensemble Classifier, K-Means Classifier, Data Stream Classifier

I. INTRODUCTION

SVM is highly resistant to noisy data and has the most powerful generalization capability on yet to-be seen linear and nonlinear data. However, it becomes flabby when it is directly applied for data stream classification due to its intrinsic nature. In a bid to devoid this issue, this research work is intended to develop and investigate a novel SVM based HDSC which makes a series of optimization on SVM by synergizing it with Lagrangian interpolation and fuzzy logic to pep up the data stream classification process amid of all the upcoming issues arises on it.

In precise, the proposed Data Stream Classifier is expounded with the fusion of fuzzy logic, SVM and Lagrangian interpolation method which is chosen and synergized scrupulously by analyzing the coherence among them in achieving the data stream classification task

Nevertheless, the SVM is well known for its accuracy and robustness, it is not widely used for online data stream classification due to the following reasons:

- ➢ It needs more memory space to store all possible support vectors, where support vectors are the points that are closest to the boundary between the instances of the target classes.
- ➢ It consumes more processing time in deciding maximum margin and hyper planes between the instances of different classes.

In light of addressing the above said deficits of SVM, this research work is meant to equip SVM to contend against the incidence of concept drifts and novel classes in the online data stream classification.

This paper is organized into seven segments. Following the introductory part, the state of the art of data stream classification and related work of the proposed system are explored in segment two and three respectively. The contributions of the proposed system in data stream classification are spotted in segment four. System requirements and performance evaluation are illustrated in the fifth and sixth segment respectively, after a deliberate experimental research on video data stream classification. This paper is concluded in the seventh segment by enlightening the directions for future enhancements of the proposed system.

II. STATE OF THE ART

The SVM has rarely been used for mining data streams because of the challenges coupled with incrementally updating the SVM classification model with the increasing number of instances. Some SVM based methods such as SVMLight and SVMPerf have been proposed to

lifting up SVM classification for data stream classification. Nonetheless, these classification methods are not developed for the real time streaming model.

Knowing the high generalization feature of SVM classifiers, it is intuitively assumed that its contribution will achieve greatness in real time data streams. Moreover, SVM based methods use a quadratic programming formulation with much constraint. This incurs high computational complexity in solving data streaming problems. In contrast, in the case of kernel based SVM methods, the size of the kernel matrix scales up with the square of the number of data points. In data streams, the number of instances constantly increases with respect to time. Hence, it is evident that such methods cannot be used directly in the data streaming context.

The research work concerned on equipping SVM for data stream classification can be conducted out in three different proportions, either individually or in combination (Aggarwal *et al.*, 2009):

1. Tune up SVM as incremental so that the classification model can be adjusted efficiently as new instances come in without having to learn from scratch.
2. Combine incremental SVM with the learning on a window of instances. In this approach, the window size should be adapted according to the varying intensity of concept drifts in data stream classification which is a quite challenging job in the data streaming scenario.
3. Equip SVM by combining with other suitable algorithms to achieve optimality in classifying the data streams.

The scantiness of the related works and other comparative algorithms in addressing concept drifts and concept evolution during the data stream classification is discussed in the subsequent sections of this paper. The limitations observed in the existing approaches prompt the proposed HDSC classifier on resolving and to further advancing the data stream classification task.

III RELATED WORK

SVM is a supervised classifier which is widely adopted to perform off-line batch analysis and binary classification on static data. In contrast, it is rarely used for real time data stream classification due to its incompatible intrinsic nature with incremental data. Since data in data streams evolve over time, SVM struggles to perform classification on this evolving nature of data streams, as it is incepted to perform classification on static data.

The extensive survey explored on data stream classification reveals that a large number of research works have been focused on equipping SVM to fit for data stream classification. Buc *et al*, 2005, presented an SVM based incremental learning algorithm that employs the locality of Radial Basis Function (RBF) that re-learns only the weights of training instances lying in the vicinity of the new incremental data. However, this approach didn't regard the incidence of concept drift and has been deployed in a concept drift free data streams.

Inspired by these research works, the proposed research work is intended to equip SVM with a series of synergization with fuzzy logic, Lagrangian interpolation method and parallel genetic algorithm.

However a plenty of methods have been attended in finding solutions for confronting the issues in data stream classification process, the most contemporary methods such as, Accuracy Weighted Ensemble (AWE), Accuracy Updated Ensemble (AUE), E-TREE and Classification and Regression Tree (CART) are taken up for comparison.

A. Accuracy Weighted Ensemble (AUE)

Accuracy Weighted Ensemble (AUE) is an incremental online ensemble approach that deals with concept drifts by incrementally selecting and updating the component classifiers in the ensemble. This model selects the component classifiers by weighing its accuracy so as to remove the inaccurate classifiers and to keep accurate classifiers. The drawback of this approach is that the classification model struggles to retain diversity among the component classifiers during the longer distribution stability, (Brzezinski *et al*., 2014).

B. E-TREE

An Ensemble Tree, E-TREE, is a height balanced tree which has an efficient indexing structure and treats the ensemble model as a

spatial database. It arranges all the base classifiers of the ensemble in the nodes of E-TREE which helps to achieve sub linear prediction time. In addition, it also has two more components, R-tree and a table structure which facilitates to classify each incoming data stream, to insert new base classifiers and delete the outdated ones respectively from the E-Tree. However, this method confronts the high complexity in constructing E-TREE when applied to real time data stream classification task (Zhang *et al*., 2011).

C. Accuracy Weighted Ensemble (AWE)

Accuracy Weighted Ensemble (AWE) combines the response of base classifiers by using a weighted-majority voting approach. The weight of the base classifiers depends on the accuracy obtained by them when classifying the instances from the current training data chunk. Upon the arrival of a new data chunk, it includes a newly trained classifier into the ensemble and removes the least performing classifier from the ensemble membership. The inference made on AWE is that its efficiency depends on both the size of the data chunk and the intensity of the concept drift. That is, it adapts to gradual changes, but it entails trouble in adapting to abrupt concept drifts.

D. Classification and Regression Trees (CART)

Classification and Regression Tree (CART) is a decision tree based machine learning algorithm which is applied as an umbrella term for performing both classification and prediction. The accuracy of this algorithm turns down at the incidence of unlabeled instances (Leo Breiman *et al*, 1984).

The proposed data stream classifier is deployed in the scenario of classifying videos in an online environment where the incidence of concept drift and novel class is indispensable. That is, the proposed classification model is applied in analyzing and predicting the users' interest in viewing the videos in a video player so as to confront the drift in their interest of viewing different category of videos at different time. It is a highly challenging task, since the interest in viewing the videos online is chaotic and transient over time.

For instance, the user who is interested in viewing the industry related videos may not be in the same category of videos during his/her spare time. In addition, due to the ever growing technology, new category of videos may emerge at any instant. Consequently the users also might be captivated on viewing the latest trendy category of videos. To confront this scenario, a novel HDSC has been proposed.

IV PROPOSED SYSTEM

This section highlights the strength of the proposed system HDSC which is acquired by synergizing the complementary strength of fuzzy logic, SVM and Lagrangian interpolation method.

In the proposed approach, Fuzzy logic, which is good in generalization and fault tolerance, is applied with SVM to accelerate the speed of the classification task. Since the proposed system is intended for analyzing transient data streams, Lagrange's interpolation method has been chosen as an additional constituent algorithm due to its inherent nature of efficiency in handling transient data.

As the interpolation method is primarily concerned about minimizing the misclassification of data while finding the maximum margin between the separating hyper planes, it is deployed to handle the situations when the features of a data set fall outside the decision boundary of support vector machine. Herein, Multiclass SVM which combines the output of the multiple SVM is used to classify the various categories of videos.

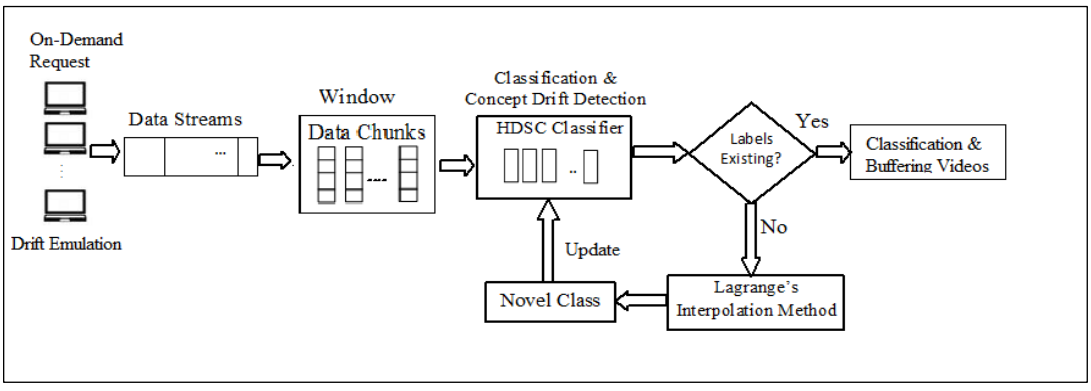

Figure.1 System Architecture of HDSC

Procedure	Accuracy	Precision	Recall	F1-Measures	Memory	Average Response Time	Average Processing Time
AWE	83.13	83.09	83.20	83.17	**5984820**	**884**	**1439**
AUE	83.43	83.29	83.00	83.17	5989444	**884**	**1442**
E-TREE	85.26	85.22	85.33	85.30	6347870	956	1523
CART	86.86	86.82	86.93	86.90	6332472	946	1512
HDSC	**87.96**	**88.02**	**88.03**	**88.05**	6271020	916	1486

Table.1 Average performance report of HDSC

The proposed method HDSC is applied for drift detection on analyzing the request generated for video streaming. Before classifying the request on videos, the classification model inspects for concept drift and if no concept drift is detected, the classification model can be used for classifying the classes without any intact. If any drift is detected, it might also lead to novel class occurrence. To ensure the incidence of novel class, consequently upon the incidence of concept drift, the results are again inspected using Lagrangian interpolation method.

The processing steps of HDSC are given below:

1. Initially, the data streams are split into fixed sized data chunks.
2. They are buffered and scanned for fixing errors and noise in data chunks by adopting fixed sized sliding window approach.
3. Fuzzy logic and SVM which are the components of HDSC are applied to perform data stream classification in both the normal and concept drifting scenario.
4. If any novel class is found, then the steps 5 to 9 are executed.
5. In HDSC, Lagrange's Interpolation method is enabled in case of novel class incidence.
6. If similarity between the instances of the data streams is found high, then the instances will be declared as novel classes.
7. Else they will be tracked for some time, to find similarity with the upcoming instances.
8. If the similarity is found high, then they will be declared as a novel class, else will be discarded as outliers.
9. Upon identifying the novel class, the HDSC model is updated to learn the newly arrived class.

In case of concept drift and novel class incidence, the data stream classifier HDSC is updated to sustain its efficiency on classifying the upcoming dataset. Upon knowing the interest of the users, their request can be classified for pre buffering of the interesting videos on their video player.

V HARDWARE AND SOFTWARE REQUIREMENTS

The proposed system has been implemented in an enterprise STD dedicated server having the configuration HP DL 160 G8 Series, 1 X Intel Hexa Core Xeon Processor, 15 M Cache. This model is tested by generating video requests from personal computers having the configuration, Intel core I5 processor, Windows 8.1 (64 bit) operating system and Visual Studio 2013.

Data stream classification model is trained with the dataset obtained from Filmsodm video service provider, and is tested in real time server which facilitates the data stream classification by analyzing the video browsing history of the users in terms of eighteen attributes such as, protocol, IP address, source and destination port, source and destination data header, service count, type, login status, number of failed logins, total data size, duration, service count, type, bit rate, resolution, repeat status, and audio quality.

The classification model is trained to classify the twelve different categories of video classes, namely, do it yourself, drama, sport, movie, funny, technology, entertainment, vector, short film, 2D-Animation, 3D-Animation and tutorial. When a new kind of video category is found, it is classified as a novel video category.

VI PERFORMANCE EVALUATION

The efficiency of the proposed system on HDSC has been empirically tested and compared with four contemporary data stream classification techniques, such as, CART, ETREE, AWE, and AUE over various parameters. The average performance report of varying sizes of data chunks is illustrated in Table 1.1.

VII CONCLUSION

The important aspects of this research work are corroborated briefly in this section.

- SVM is widely preferred to classify nonlinear data due to its high accuracy and generalization ability. However it entails sluggish performance and high memory requirement in case of data stream classification.
- Hence, the proposed HDSC has been designed and investigated with the fusion of fuzzy reasoning and Lagrangian interpolation method so as to tune up SVM for data stream classification.
- Here fuzzy logic, which is good in generalization and fault tolerance, is applied with SVM to accelerate the speed of the classification task.
- Lagrange's interpolation method is good in handling transient data and also it controls the tradeoff between the dual objectives of maximizing the margin of separation and minimizing the misclassification error where the SVM struggles to classify non linear data.
- Hence, the Lagrange interpolation method has been chosen as an additional constituent algorithm due to its inherent nature of efficiency in handling transient data.
- The exhaustive experiments carried out on the proposed system evince that that the proposed HDSC is the most accurate of all comparative methods.
- It is ascertained with the results obtained not only by calculating accuracy, but also precision, recall, and F1-measure on all varying sizes of data chunks.
- However, its performance negligibly drops down in terms of memory utilization, processing and response time with respect other comparative algorithms. This might be caused due to inherent feature of the SVM which is used as the base classifier in the HDSC.
- It is also assumed that the deteriorated performance in processing and response time shall be caused by the attempt of making SVM classifier in multi label data stream classification.
- It is planned to enhance the efficiency of the proposed model in all aspects by deploying the revised composition of the model which includes the parallel genetic algorithm to expedite the processing of data streams in the subsequent phase of this research work.

REFERENCES

1. Albert Bifet, Bernhard Pfahringer, Jesse Read, and Geoff Holmes, "Efficient Data Stream Classification via Probabilistic Adaptive Windows", ACM, Proceeding SAC, 2013, pp.801-806.
2. Albert Bifet, Geoff Holmes and Bernhard Pfahringer, "New Ensemble Methods For Evolving Data Streams", Proceeding 15th ACM SIGKDD, 2009, pp. 139-148.
3. Bernhard Pfahringer, Geoffrey Holmes, and Richard Kirkby, "New Options for Hoeffding Trees", Proceeding AI'07, Springer-Verlag Berlin, Heidelberg, 4830, 2007, pp.90-99.
4. Dariusz Brzezinski, and Jerzy Stefanowski,. "Combining block-based and online methods in learning ensembles from concept drifting data streams", Information Sciences, 265, 2014, pp.50–67.
5. Gama, J., Medas, P., Castillo, G., Rodrigues, P, "Learning with drift detection", Lecture Notes in Computer Science, 2004.
6. Haixun Wang, Wei Fan, Philip S. Yu, and Jiawei Han, "Mining concept-drifting data streams using ensemble classifiers", Proceedings of the 9th ACM SIGKDD, 2003, pp. 226–235.
7. Indre Zliobaite, 2010. "Learning under Concept Drift: an Overview", CoRR abs/1010.4784.
8. J. Zico Kolter, and Marcus A. Maloof, "Dynamic Weighted Majority: An Ensemble Method for Drifting Concepts", Journal of Machine Learning Research, 8, 2007, pp. 2755–2790.
9. Jesse Read, Albert Bifet, Geo Holmes, and Bernhard Pfahringer, "Scalable and Efficient Multi-label Classifcation for Evolving Data

Streams", Machine Learning, 88, 2012, pp.243-272.

10. Jesse Read, Albert Bifet, Bernhard Pfahringer, and Geo Holmes, "Batch-incremental versus instance-incremental learning in dynamic and evolving data", Proceeding, Springer-Verlag Berlin, 2012, pp. 313-323.

11. Joao Gama, Raquel Sebastiao, and Pedro Pereira Rodrigues, "On evaluating stream learning algorithms, Machine Learning", vol. 90(3), 2013, pp. 317-346.

12. Kenneth O. Stanley, "Learning Concept Drift with a Committee of Decision Trees", Technical Report UT-AI-TR-03-302, 2003.

13. Kyosuke Nishida, "Learning and Detecting Concept Drift, Graduate School of Information Science and Technology", Hokkaido University, Dissertation, 2008.

14. Leandro L. Minku, Allan P. White, and Xin Yao, "The Impact of Diversity on Online Ensemble Learning in the Presence of Concept Drift", IEEE Transactions on Knowledge And Data Engineering, 22(5), 2010, pp.730-742.

15. Littlestone and M. Warmuth, "The weighted majority algorithm", Inf. Comput., 108(2): 1994, pp.212–261.

16. Manuel Baena-Garcia, Jose del Campo-Avila, Raul Fidalgo, Albert Bifet, Ricard Gavalda, and Rafael Morales-Bueno, "Early Drift Detection Method", ECML PKDD'06, 2006.

17. Michael D. Muhlbaier, Apostolos Topalis, and Robi Polikar, "Learn++.NC: Combining Ensemble of Classifiers With Dynamically Weighted Consult-and-Vote for Efficient Incremental Learning of New Classes", IEEE Transactions on Neural Networks, 20(1): 2009, pp.152-168.

18. Mohammad M. Masud, Clay Woolam, Jing Gao, Latifur Khan, Jiawei Han, Kevin W. Hamlen, and Nikunj C. Oza, "Facing the reality of data stream classification: coping with scarcity of labeled data", Knowledge and Information Systems, 2011.

19. Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham, "Integrating Novel Class Detection with Classification for Concept-Drifting Data Streams", *ECML PKDD*, Springer-Verlag Berlin Heidelberg, 2009, pp.79–94.

20. Mohammad M. Masud, Member, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham, "Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints", IEEE Transactions on Knowledge And Data Engineering, 23(6), 2010, pp. 859-874.

21. Mohammad M. Masud, Qing Chen, Latifur Khan, Charu C. Aggarwal, Jing Gao, Jiawei Han, Ashok Srivastava, and Nikunj C. Oza, "Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams", IEEE Transactions On Knowledge And Data Engineering, 25(7): 2013, pp.1484-1497.

22. Peng Zhang, Xingquan Zhu, Yong Shi, and Li Guo, Xindong Wu, "Robust Ensemble Learning for mining noisy data streams", Decision Support System, 2011, pp.469-479.

23. Sarah Jane Delany, "Using Case-Based Reasoning for Spam Filtering", Dublin Institute of Technology, dissertation, 2006.

24. Sobolewski P., Wozniak M., "Concept drift detection and model selection with simulated recurrence and ensembles of statistical detectors", Journal of Universal Computer Science, 2013, (19)4: pp.462-483.

25. W. Nick Street, and Yong Seog Kim, "A Streaming Ensemble Algorithm (SEA) for Large Scale Classification", Proceedings of the seventh ACM SIGKDD, 2001, pp. 377-382.

26. X. Wei, Y. Li, D. Liu, and L. Zhan, Mahalanobis, "Support Vector Machines Made Fast and Robust - New Advances in Machine Learning", InTech, ISBN: 978-953-307-034-6, China, 2010, pp.227-250.

Dr.S.Jayanthi was born in the year 1981. She received her Bachelor degree in Computer Science from Bharathidasan University, India in 2002, and Master degrees in Computer Applications, and Computer Science and Engineering from Bharathidasan University, in 2005, and from Anna University of Technology, India in 2009, respectively. She obtained her Ph.D from Karpagam University, Coimbatore, India. She has 8 years of teaching experience. Her area of research includes Data Mining, Neural Networks and Big Data analysis. She has worked as Assistant professor and Head of the department in the department of Computer Science and Engineering for 7 years in Srinivasan Engineering College, Perambalur, Tamilnadu, India. At present,

She is working as professor and Dean(R&D) in Samskruti College of Engineering and Technolgy, Hyderabad, Telangana, India. So far she has published 30 papers in various national, International conferences and journals.