OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

E-MORES: EFFICIENT MULTIPLE OUTPUT REGRESSION FOR STREAMING DATA USING DECISION TREE AND RANDOM FOREST ALGORITHMS

Priyanka Konde[1], Prof. Varsha Dange[2]

PG Students, Department of Computer Engineering, Dhole Patil College of Engineering, Wagholi, Near Eon IT park., Pune-412207, Maharashtra, India[1]

Assistant Professor, Department of Computer Engineering, Dhole Patil College of Engineering, Wagholi, Near Eon IT park., Pune-412207, Maharashtra, India[2]

Abstract: **Online efficient multiple-output Regression is an important machine learning technique for modeling, predicting, and compressing multi-dimensional correlated data streams. This proposed system introduced, a novel online efficient multiple-output Regression method, called E-MORES, for streaming data. E-MORES can dynamically learn the structure of the Regression coefficients to facilitate the models continuous refinement. Considering that limited expressive ability of Regression models often leading to residual errors being dependent, E-MORES intends to dynamically learn and leverage the structure of the residual errors to improve the prediction accuracy. This system also introduce RandomForest and DecisionTree to predict (classify) the next event type that will occur during the transition time, that is growing, continuing, shrinking, dissolving, merging or splitting. The conducted experiments suggest that the RandomForest and DecisionTree classifiers usually provide more accurate results. Moreover, system introduce three modified covariance matrices to extract necessary information from all the seen data for training, and set different weights on samples so as to track the data streams' evolving characteristics. Furthermore, an efficient algorithm is designed to optimize the proposed objective function, and an efficient online eigen value decomposition algorithm is developed for the modified covariance matrix. Finally, this system analyze the convergence of EMORES in certain ideal condition. Experiments carried out on two synthetic datasets and three real-world datasets validate the effectiveness and efficiency of E-MORES.**

Keywords: *Decision Tree, Dynamic Relationship Learning, For-getting Factor, Lossless Compression, Online Efficient Multiple –Output regression method.*

I INTRODUCTION

Multi-output Regression, also known as multi-target, multi-variate, or multi-response Regression, aims to simultaneously predict multiple real valued output/target variables. When the output variables are binary, the learning problem is called multilabel classification. However, when the output variables are discrete (not necessarily binary), the learning problem is referred to as multidimensional classification. Several applications for multi-output Regression have been studied. They include ecological modeling to predict multiple target variables describing the condition or quality of the vegetation, chemo metrics to infer concentrations of several analytes from multivariate calibration using multivariate spectral data, channel estimation through the prediction. Data streams arise in many scenarios, such as online transactions in the financial market, Internet traffic and so on. Unlike traditional datasets in batch mode, a data stream should be viewed as a potentially infinite process collecting data with varying update rates, as well as continuously evolving over time.

In the context of data streams, although many research issues, such as classification, clustering, active learning, online feature selection, multi-task learning, change point detection, etc., have been extensively studied over the last decade, little attention is paid to multiple output Regression. However, multiple-output Regression also has a great variety of potential applications on data streams,

including weather forecast, air quality prediction, etc. In batch data processing, the purpose of multiple-output Regression is to learn a mapping from an input space to an output space on the whole training dataset. A basic assumption in multiple-output Regression is that there is related information among multiple outputs, and learning such information can result in better prediction performance. We propose a novel Efficient Multiple-Output Regression method for Stream data, named as E- MORES. E-MORES works in an incremental fashion. Specifically, when a new training sample arrives, we transform the update of the Regression coefficients into an optimization problem.

II LITERATURE SURVEY

Svstream: A support vector-based algorithm for clustering data streams

In this a novel data stream clustering algorithm is proposed, termed SVStream, which is based on support vector domain description and support vector clustering[1].

Active learning with drifting streaming data

This paper presents a theoretically supported framework for active learning from drifting data streams and develops three active learning strategies for streaming data that explicitly handle concept drift. They are based on uncertainty, dynamic allocation of labeling efforts over time, and randomization of the search space[2].

Online feature selection and its applications

This paper investigated a new research problem, Online Feature Selection (OFS), which aims to select a small and xed number of features for binary classication in an online learning fashion[3].

Kernelized bayesian matrix factorization

Kernelized matrix factorization is extended with a full-Bayesian treatment and with an ability to work with multiple side information sources expressed as different kernels. Kernels have been introduced to integrate side information about the rows and columns, which is necessary for making out-of-matrix predictions[4].

Multivariate Regression with calibration

In this a new method named calibrated multivariate Regression (CMR) for fitting high dimensional multivariate Regression models is proposed. Compared to existing methods, CMR calibrates the regularization for each Regression task with respect to its noise level so that it is simultaneously tuning insensitive and achieves an improved nite sample performance[5].

Online multitask learning for policy gradient methods

Policy gradient algorithms have shown considerable recent success in solving high-dimensional sequential decision making tasks, particularly in robotics. However, these methods often require extensive experience in a domain to achieve high performance[6].

Online sketching hashing

This paper proposes a novel approach to handle these two problems (Streaming data And Huge Dataset) simultaneously based on the idea of data sketching. A sketch of one dataset preserves its major characters but with significantly smaller size. With a small size sketch, our method can learn hash functions in an online fashion, while needs rather low computational complexity and storage space[8].

Online multi-task learning via sparse dictionary optimization

This paper develops an efficient online algorithm for learning multiple consecutive tasks based on the KSVD algorithm for sparse dictionary optimization. The K-SVD algorithm is explored (Aharon et al. 2006) in the lifelong machine learning setting[7].

Modeling and Predicting Community Structure Changes in Time-Evolving Social Networks

This paper proposes a sliding window analysis from which authors develop a model that simultaneously exploits an auto regressive model and survival analysis techniques. The auto regressive model is employed here to simulate the evolution of the community structure, whereas the survival analysis techniques allow the prediction of future changes the community may undergo[10].

Dynamic Structure Embedded Online Multiple-Output Regression for Streaming Data

This paper introduced three modified covariance matrices to extract necessary information from all the seen data for training, and set different weights on samples so as to track the data streams' evolving characteristics[11].

III SYSTEM ARCHITECTURE

In this proposed system, I propose a novel online efficient multiple-output Regression method, called E-MORES, for streaming data. E-MORES can dynamically learn the structure of the Regression coefficients to facilitate the models continuous refinement. E-MORES intends to dynamically learn and leverage the structure of the residual errors to improve the prediction accuracy. I also introduce Random Forest and Decision Tree to predict (classify) the next event type that will occur during the transition time, that is growing, continuing, shrinking, dissolving, merging or splitting.

IV CONCLUSION AND FUTURE WORK

In proposed system, an efficient algorithm is introduced called E-MORES for streaming data. The proposed system will apply on streaming data to find the regression coefficient dynamically. Also proposed system will find the critical events by using decision tree and random forest algorithm. In experiment result will show effectiveness of proposed system to find the regression coefficient and prediction of critical events.

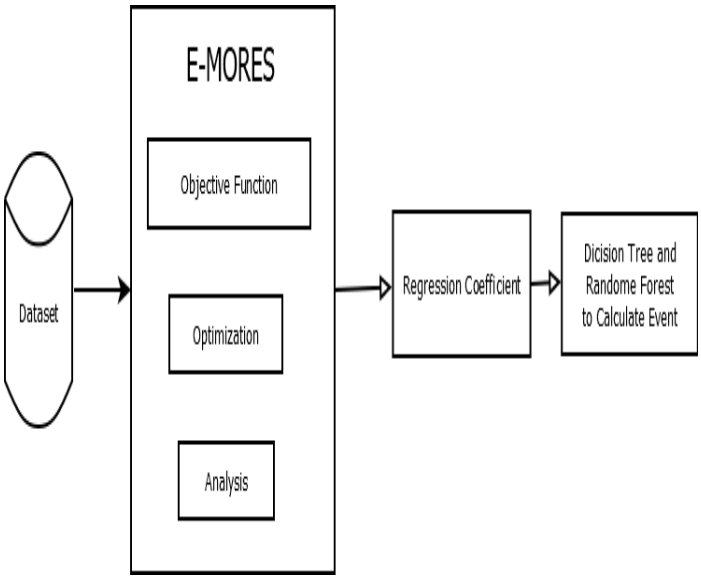

Figure 1 System Architecture

REFERENCES

1) C. D. Wang, J. H. Lai, D. Huang, and W. S. Zheng, "Svstream: A support vector-based algorithm for clustering data streams", IEEE Trans. on Knowledge and Data Engineering, vol. 25, no. 6, pp. 14101424, 2013.
2) I. Zliobaite, A. Bifet, B. Pfahringer, and G. Holmes, "Active learning with drifting streaming data", IEEE Trans. on Neural Networks and Learning Systems, vol. 25, no. 1, pp. 2739, 2014.
3) J. Wang, P. Zhao, S. Hoi, and R. Jin, "Online feature selection and its applications", IEEE Trans. on Knowledge and Data Engineering, vol. 26, no. 3, pp. 698710, 2014.
4) M. Gonen and S. Kaski, "Kernelized bayesian matrix factorization", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 36, no. 10, pp., 2014.
5) H. Liu, L. Wang, and T. Zhao, "Multivariate Regression with calibration", in Advances in Neural Information Processing Systems (NIPS), 2014.
6) H. B. Ammar, E. Eaton, P. Ruvolo, and M. Taylor, "Online multitask learning for policy gradient methods", in Proceedings of the 31st International Conference on Machine Learning (ICML), 2014.
7) P. Ruvolo and E. Eaton, "Online multi-task learning via sparse dictionary optimization", in Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI), 2014.
8) C. Leng, J. Wu, J. Cheng, X. Bai, and H. Lu, "Online sketching hashing", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
9) C. Li, Q. Liu, W. Dong, X. Zhu, J. Liu, and H. Lu, "Human age estimation based on locality and ordinal information", IEEE transactions on cybernetics, vol. 45, no. 11, pp. 25222534, 2015.
10) Etienne Gael Tajeuna, Mohamed Bouguessa, and Shengrui Wang, "Modeling and Predicting Community Structure Changes in Time-Evolving Social Networks", IEEE Transactions on Knowledge and Data Engineering, 2018.
11) Changsheng Li, Fan Wei, Weishan Dong, Xiangfeng Wang, Qingshan Liu, Senior Member and Xin Zhang, "Dynamic Structure Embedded Online Multiple- Output Regression for Streaming Data", IEEE, 2018.