iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 Issue: IX Month of publication: September 2020

DOI: https://doi.org/10.22214/ijraset.2020.31386

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IX Sep 2020- Available at www.ijraset.com

Efficient Multiple-Output Regression for Streaming Data

Priyanka K. Kendre[1], Pornima M. Birajdar[2]
[1, 2]Department of Computer Science and Engineering, Swami Ramanand Teerth Marathwada University, Nanded

***Abstract:** An efficient multiple-output Regression is an important machine learning technique, this technique is used for modeling, forecasting, and compressing multi-dimensional data streams. The proposed system consists of an efficient multiple-output regression method for data streams, called E-MORES. E-MORES have ability to gain the structure of the regression coefficients that can be used to provide the model's continuous improvement. E-MORES can dynamically learn the structure of the residual errors that can be used to improve the prediction accuracy; it also leverages the structure of residual errors to increase prediction accuracy.*

This proposed system also introduces Random Forest, Decision Tree to predict (classify) the next event type that will happen during the modulation time, that is increasing, continuing, reducing, and splitting and ARIMA model is based on the idea that the information in the past values of the time series can alone be used to predict future values. The result of the random forest and the decision tree provide better accuracy than the other. In addition to that, the system also presents modified covariance matrices and forgetting factors. Covariance matrices are used to extract needed data for the system and the forgetting factor is used to weight the sample data. it also used to measure forecasting error and covariance matrices make use of Eigenvalue decomposition algorithm. Experiments execute on two synthetic datasets and three real-world datasets that validate the effectiveness and efficiency of EMORES.

Keywords*: Decision Tree, Dynamic Relationship Learning, Forgetting Factor, Lossless Compression, Online Efficient Multiple Output regression method, Random Forest.*

I. INTRODUCTION

Many machine learning techniques are used for predicting single numeric value for that simply used regression and to predict two or more numeric outputs multiple-output regression is used. Regression is a form of predictive modeling technique which investigates the relationship between the dependent and independent variable. Online efficient multiple-output regression is a significant machine learning system for demonstrating, foreseeing, and compacting multi-dimensional corresponded information streams. The main purpose of regression analysis is the prediction of one variable from the other.

The main goal of this work is to propose an efficient multiple-output regression method for streaming information. A basic assumption in multiple output regression is that there is related information among multiple outputs and learning such information gives better prediction performance.

It can progressively learn the structure of the regression coefficient to provide the model ceaseless achievement. This paper proposes E-MORES that can be used to find out regression coefficients. The system also introduces a decision tree, random forest, and ARIMA model to make better performance of the system.

Decision tree and random forest used for both regression and classification. A decision tree can be used visually and explicitly to represent decision and decision making. It uses a tree-like model of decisions In the decision tree, the branch nodes make decisions inquiries and the leaf nodes represent prediction. Basically in decision tree taking the data, analyze it on the basis of some condition, and finally it divided into various categories.

A random forest is a predictive modeling algorithm. It can be used for both regression and classification. Random forest makes the model simpler to interpret, it reduces computational cost and time of training data. It runs efficiently on a large database and reduces the variances. An ARIMA model is a forecasting algorithm based on the past values of the time series. An ARIMA model is also used to predict the future values of the data.

It is also based on the AR model and the MA model. An Efficient multiple-output regression system uses a decision tree, random forest, and ARIMA model for better prediction performance. In this way, it proposes an efficient multiple-output regression system to find out predictions.

II. LITERATURE REVIEW

A. A Support vector-based Algorithm for Clustering Data Streams

A support vector-based algorithm is used for clustering dat**a** streams. It presents SVStream algorithms for clustering which is based on the idea of support vector domain description and clustering of support vector [1]. The system uses data components of streaming information that can be used for clustering purpose and that data components are represented into kernel space. In algorithm support vector gives important information of historical components in the data stream that can be used for clustering. A support-vector based algorithm uses Bounded Support Vector (BSVs) to identify the overlapping clusters. A BSV method is used to find out unnecessary data and to remove the unnecessary data. This algorithm works on synthetic data streams and real data streams. It is also used to handle noise situations of data streams. The result of this algorithm presents the efficiency of the system.

B. Active Learning with Drifting Streaming Data

It presents a system for dynamically learn the concept of drift. It uses drifting information streams for active learning. It presents three learning methods to handle the concept of drift. These three learning methods work dynamically to learn the idea of drift. It depends on the vulnerability of data streams and labeling of data streams and randomization of search space. Labeling is allocated dynamically after some time and presents three dynamic learning methods for streaming information to handle drift. Active learning is an accurate predictive model that carefully selects a few labeled z instances for the learning model [2]. In a streaming environment data is stored in a streaming fashion is the challenge for active learning. Active learning presents the time to time changed data. The active learning method uses the most variable data and provides better prediction accuracy. It presents a theoretically supported framework for this method and presents three learning strategies for data streams. These learning strategies explicitly handle the concept of drift. The main goal of active learning is to focus on unnecessary data and randomly allocated space. It presents these learning strategies when data changes anywhere in the system.

C. Detecting changes in data streams

In a streaming environment, data streams are observed continuously. Observation of data points consists of similarity, the variability of the data streams. In this paper, it detects a change in data streams by using the exchangeability property of visited data and creates a data-generating model. The data-generating model may change as the data is streaming. It proposes an efficient approach known as Martingale's approach. It is a one-pass algorithm that is used for both classification and regression. It is a nonparametric approach used for the clustering of data-generating models and the result of this approach presents the effectiveness of the martingale method used for detecting the changes in the data points and data generating model for continuously changes data in streaming information [3]. It also proposes an adaptive support vector machine (SVM) that uses a martingale approach Which compares with other an adaptive support vector machine to detect a change in the vector machine.

D. Kernelized Matrix Factorization

It presents a factorization method named kernelized matrix factorization which is based on a full-Bayesian treatment. Factorization uses multiple side information sources that are communicated as different kernels. kernels collect various information about lines and segments. This information about lines and segments are used for generating better predictions. It consists of two main aspects: the first aspect consists of an efficient variation approximation scheme is used for factorization with the help of a novel fully conjugate probabilistic model and the second aspect is consists of the conjugation of matrix factorization and it also describes learning of multiple kernels to incorporate multiple side information sources for matrix factorization [4]. This model supports fully Bayesian treatment and it is computationally feasible than the other probabilistic model. It also consists of the Bayesian model selection strategy by using automatic relevance determination and semi-supervised learning is used for partially observed output cases. It explains the effectiveness and usefulness of this method on one toy dataset, two molecular biological interaction datasets, multilabel classification data sets and one yeast cell cycle data set.

E. Online Sketching Hashing

It presents a new aspect which works on the concept of sketching. Sketching handles the following two problems concurrently:

1) In a streaming environment, information comes continuously like in a streaming fashion but many of the hashing techniques use batch-based models.
2) In batch-based models when the dataset comprises a large amount of data, it is impractical to load all the data into memory and it increases memory complexity and computational complexity [5].

In this paper, the concept of the sketch utilizes hashing purposes. A sketch of one dataset consists of its major characters but the size of the sketch is small. This method presents hash functions with a sketch. It decreases the computational complexity and memory complexity by using the sketch in hash functions. In this system, a concept of a sketch is used with three covariance matrices. In a computing environment, A small size sketch is used for hashing and each type of sketch requires a different objective function. The concept of a sketch is also used in the k-means algorithm and k-median algorithm.

F. K-SVD algorithm

This paper proposes an algorithm for the acquisition of multiple sequential tasks which is based on the concept of the KSVD algorithm. It is used for sparse dictionary optimization [6]. The KSVD method presents the concept of a lifelong machine. A lifelong machine is a machine that gives a detailed description of the workflow of the system. The system adapts K-SVD for the lifelong machine. The lifelong machine is able to change the SVD steps in the original algorithm to other SVD step when new data components are presented and it selectively updates the components of the system. The lifelong machine provides detailed workflows of the data in the system. It also proposes ELLA-SVD that works better on problems where data distributions of input are similar distribution. For domains data distribution of input are different, it shows the original approach of the model. In this way, K-SVD describes the system by using a lifelong machine.

III. PROPOSED ARCHITECTURE

A. Architecture

1) Goals & Objectives

a) To propose a novel efficient multiple-output Regression method, this is called E-MORES, for streaming data.

b) To propose an E-MORES this can powerfully gain proficiency with the structure of the Regression coefficients to encourage the model's consistent refinement.

c) The objective of online performs multiple tasks learning is to mutually get familiar with the related in an online manner, to improve speculation overall assignments.

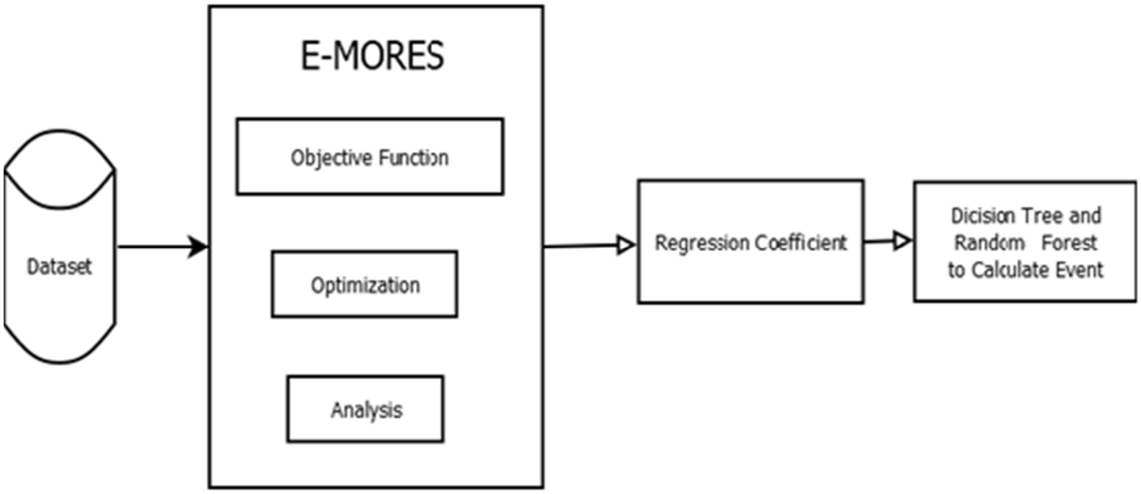

Fig .1 System architecture

In this proposed system, 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 model’s continuous refinement. E-MORES intends to dynamically learn and leverage the structure of the residual errors to improve prediction accuracy. The system also introduces 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.

B. Algorithms

1) Efficient Multiple-Output Regression for Streaming Data(E-MORES)[8]:

a) Input: Data streams { (x1,y1), (x2,y2), …. } that arrives one sample each time.

b) Parameters: α, β, η, and forgetting factor μ

c) Initialize: P0= 0dm, C0,XX= 0dd, C0, XY = 0dm, C0, YY = 0mm , and Ω0 = Γ_0 = Imm ;

```
Method
        Step 1 : for t = 1, 2,
        Step 2 : $C_{t,YY} = \mu C_{t1,YY} + y_t y_t^T;$
        Step 3: $C_{t,XY} = \mu C_{t1,XY} + x_t y_t^T;$
        Step 4 : $C_{t,xx} = \mu C_{t1,XX} + x_t x_t^T ;$
        Step 5 : Update $P_t$
                $P_t = U\bar{P}V^{-1}$
        Step 6 : Update $\Omega_t$
                $\Omega_t = \left( \frac{1}{\beta+\rho}\beta\Omega_{t-1}^{-1} + \rho I + M \right)$
        Step 7 : Update $\Gamma_t$
                $\Gamma_t = ( I + \frac{\eta}{\alpha} \mathrm{N})^{-1}$
        Step 8 : end.
    end method
```

d) Output : Regression coefficient matrix $P_t \in R^{md}$

Here, $\alpha, \beta, \eta, \rho$ are trade-off parameters whose value always greater than or equal to 0. C0,XX , C0, XY and C0, YY these are the 3 covariance matrices. Ω0, Γ_0 are positive semi-definite whose value always greater than or equal to 0 and t is the timestamp. The figure shows Ω_t in the step 6, it learns coefficient change from the current matrix and updated matrix. It uses to measure the divergence between the updated matrix and the current matrix. Γ_t is used to measure the total prediction error on all seen data. Eigen values of Ω_t , Γ_t are always bounded between 0 and 1 and P_t is updated matrix. Online multiple-output regression is a machine learning technique used for modeling, predicting, and compressing multi-dimensional correlated streaming information. An online multiple-output regression method is also called MORES, for streaming data. MORES can dynamically learn the structure of the regression coefficients to facilitate the model's continuous refinement. MORES aims to dynamically learn and leverage the structure of the forecasting errors to improve prediction accuracy. Moreover, here introduce three modified covariance matrices to extract necessary information from all the seen data for training, and set different weights on samples to track the data streams' evolving characteristics. In addition to that, an efficient algorithm designed to optimize the proposed objective function, and an efficient online eigenvalue decomposition algorithm is used for the modified covariance matrix. This method is an online multiple-output regression method for streaming data; this model can be always updated when new training data arrive. When new training data points are arriving, we will update the model based on step 5, step 6, and step 7.

2) Decision Tree Algorithm

```
GenDecTree (Sample S, Features F)
a)  If stopping_condition(S,F) = true  then
•   leaf = createNode()
•   leaf.label = Classify(S)
•   return leaf
b)  root = createNode()
c)  root.test_condition = findBestSplit(S,F)
d)  V = { v | v a positive outcome of root.test_condition }
e)  For each value v € V ;
•   Sv = { s | root.test_condition(s) = v and s € S };
•   child = TreeGrowth(Sv, F) ;
•   add child as a descent of root and label the edge ( root → child )
f)  return root
```

A decision Tree can be used visually and explicitly to represent decision and decision making. It uses tree-like a model of decisions. At the beginning of the decision tree, the whole training dataset is considered as root and then it determines the best attribute to split the remaining dataset and it proceeds recursively until only leaf nodes are created and no more split are possible It breaks down a dataset into a smaller dataset and it makes a decision on every node of the tree hence it is called as decision tree.

3) Random Forest Algorithm

```
To generate c bootstrap samples;
for  i = 1 to c  do
    Randomly sample the training data D with replacement to produce Di
    Create a root node, Ni containing Di
    Call BuildTree( Ni )
End for
BuildTree ( N )
If  N contains instances of only one class  then
return
else
    randomly select x%  of the splitting features in N
    Select the feature F with the highest information gain to split on
    Create f child nodes of N, N1, ……, Nf , where F has f possible values (F1,…., Ff)
    for i = 1 to f  do
        Set the contents of Ni to Di, where Di is all instances in N that match Fi
        Call BuildTree (Ni)
    End for
End if
```

Random forest is a predictive modelling algorithm. As the name indicates, it creates a forest using multiple decision trees randomly. It can be used for classification and regression. Random forest makes the model simpler to interpret and it reduces computational cost and time of training. It randomly picks elements from the dataset and create subsets and consider these subsets as different decision trees. Each decision tree has a different target output and the final prediction from the target values is calculated using majority voting. Random forest runs efficiently on the large datasets.

4) ARIMA Model

```
history = training dataset;
predictions = [ empty list ];
for  value € TestDataset   do
   model = ARIMA ( history , order = ( p,d,q ))
   model.fi() ;
   one_step_forecast = model.forecast();
   predictions.append( one_step_forecast );
   history.append( value )
End
Calculate MAE, RMSE
```

ARIMA model means Auto-Regressive Integrated Moving Average model. ARIMA is a forecasting algorithm. ARIMA is a combination of the autoregressive model and the moving average model. It is based on the idea that the information in the past values of the time series can alone be used to predict future values.

An ARIMA model is characterized by 3 terms (p,d,q), where p is the order of the AR term, q is the order of the MA term and d is the number of differences required to make time-series stationary. The AR term in the ARIMA represents the linear regression model that uses its own lags as predictors. The MA term in ARIMA model represents the number of lagged forecast errors and finally, it calculates MAE and RMSE for performance comparison.

IV. RESULT ANALYSIS

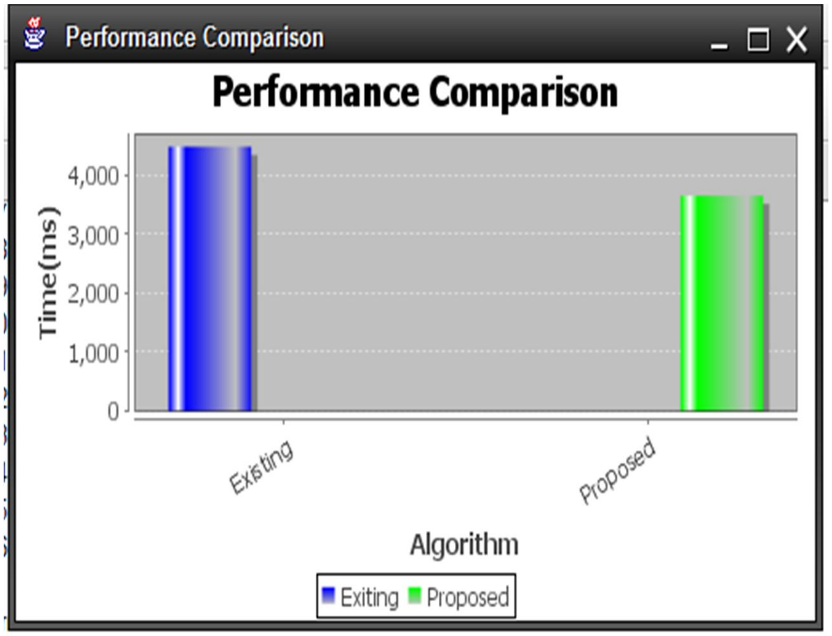

Fig.2 Performance comparison

Result analysis of an online efficient multiple-output regression for streaming data consists of performance comparison between the existing system and the proposed system. The system uses a water reservoir level dataset to perform different predictions. The figure shows the graph of the existing system and the proposed system and the time required to predict outputs in milliseconds. If the existing system gives a output of 5 days prediction and the proposed system gives a output of 10 days prediction then the time required for the existing system is greater than the time required for the proposed system so the proposed system is an efficient method than the existing system.

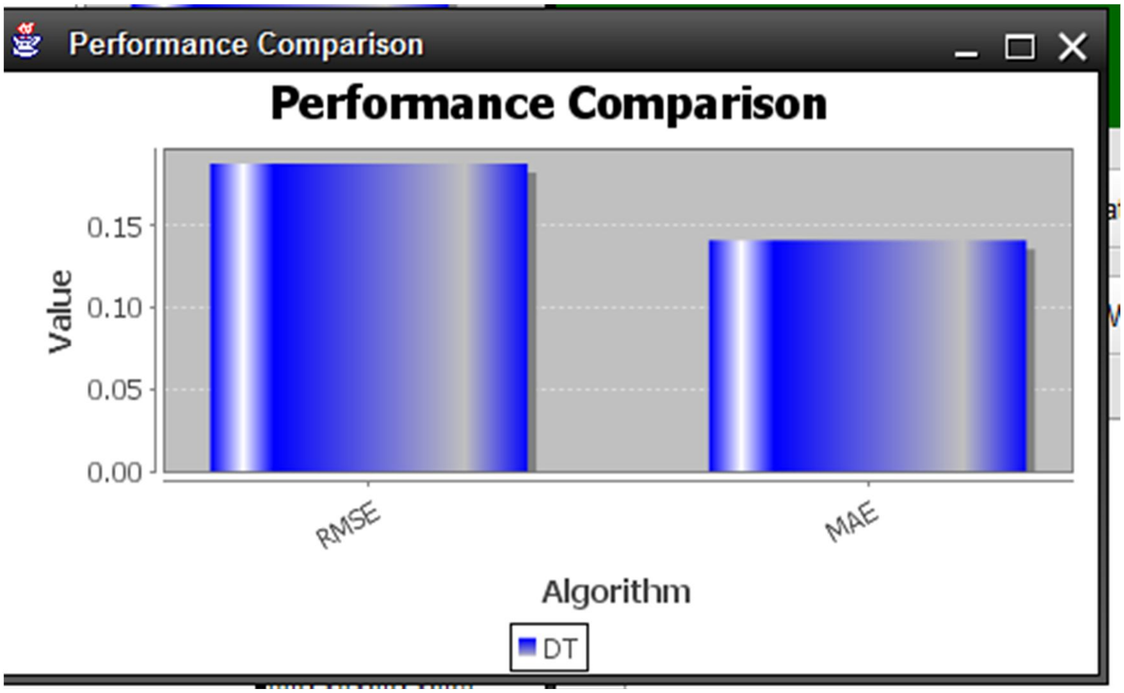

Fig. 3 RMSE and MAE in decision tree

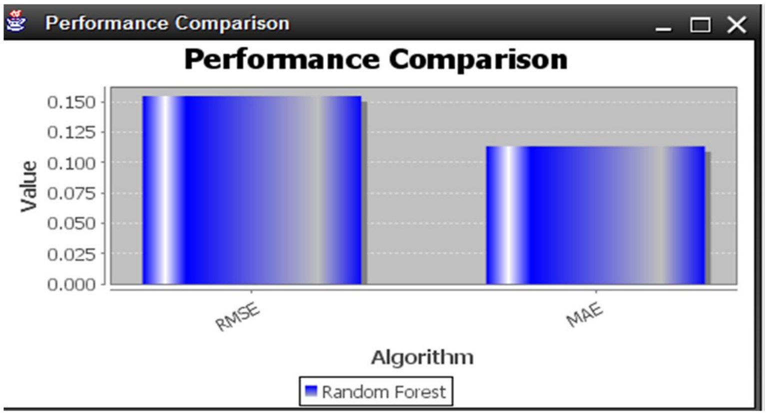

Fig. 4 RMSE and MAE in Random forest

Figure 3 and Figure 4 shows the performance comparison between the random forest and decision tree. MAE (Mean Absolute Error) is a measure of errors between paired observations. It is the arithmetic average of the absolute error. MAE is a common measure of forecast error in time series. It is simply the average absolute vertical or horizontal distance between each point in a scatter plot.MAE is known as scale-dependent accuracy measures. RMSE (Root Mean Square Error) is frequently used to measure the differences between values predicted by a model and the values observed. It represents the square root of the differences between the predicted value and observed value. RMSE is a measure of accuracy to compare forecasting errors of different models for a particular dataset and its value is always non-negative and a value of 0 indicates perfect fit of the data. A lower RMSE is better than a higher one. Figure 3 shows the MAE and RMSE present in the system when it uses a decision tree algorithm and figure 4 shows the MAE and RMSE present in the system when it uses random forest algorithm. The above figures show RMSE in the decision tree is greater than the random forest algorithm and MAE in the decision tree is also greater than the random forest algorithm. so lower RMSE and MAE is better than the higher in this way performance of the random forest algorithm is better than the decision tree.

V. CONCLUSIONS

In this paper proposed system works on the data which comes continuously like a streaming fashion. The main purpose of that system is to find out regression coefficients for that it proposes an efficient algorithm named as efficient multiple-output regression for streaming data (E-MORES). It also proposes a decision tree, random forest, an ARIMA model for the effectiveness of the system. Decision tree and the random forest are used to calculate the next event and ARIMA is a forecasting model that is used to find out the future values of the system. The experimental result shows the effectiveness of the proposed system. It can dynamically gain the structure of regression coefficients change and it also learns the structure of residual errors and makes use of this information to update the model continuously. At the same time system introduced the forgetting factor to weight the samples and perform lossless compression by calculating prediction error on all seen data. In this way, the experimental result of the proposed system provides efficiency and effectiveness of the proposed model.

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. 1410–1424, 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. 27–39, 2014.

[3] S.-S. Ho and H. Wechsler, “A martingale framework for detecting changes in data streams by testing exchangeability,” IEEE Trans. On Pattern Analysis and Machine Intelligence, vol. 32, no. 12, pp. 2113–2127, 2010.

[4] M. Gonen and S. Kaski, “Kernelized bayesian matrix factorization,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 36, no. 10, pp. 2047–2060, 2014.

[5] 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, pp. 2503 2511.

[6] P. Ruvolo and E. Eaton,“Online multi-task learning via sparse dictionary optimization,” in Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI), 2014.

[7] 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, pp. 1206– 1214.

[8] Changsheng Li, Fan Wei, Weishan Dong, Xiangfeng Wang, Qingshan Liu, and Xin Zhang, “Dynamic Structure Embedded Online Multiple-Output Regression for Streaming Data” Transactions on Pattern Analysis and Machine Intelligence, Volume: 41, Issue: 2, Feb. 1, 2019.