Sparse Support Matrix Machines for the Classification of Corrupted Data

Muhammad Imran Razzak

School of Computer Science
Faculty of Engg. & IT
University of Technology Sydney
NSW - 2019, Australia

Sparse Support Matrix Machines for the Classification of Corrupted Data

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy
in
Analytics

by

Muhammad Imran Razzak

to

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

November 2019

Author's declaration

I, *Muhammad Imran Razzak* declare that this thesis, submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy, in the *School of Computer Science*, *Faculty of Engineering and Information Technology* at the University of Technology Sydney, Australia, is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Production Note:

SIGNATURE: Signature removed prior to publication.

[Muhammad Imran Razzak]

DATE: 22nd November, 2019

PLACE: Sydney, Australia

Acknowledgments

I would like to express my sincere gratitude to Guandong Xu, for providing continuous support and motivation through out PhD.

Dedication

To my family ...

List of Publications

Published/Accepted

1. Imran Razzak, Raghib Abu Saris, Michael Blumenstein, and Guandong Xu, "Integrating Joint Feature Selection into Subspace Learning: A Formulation of 2DPCA for Outliers Robust Feature Selection." Neural Networks (Elsevier), doi.org/10.1016/j .neunet.2019.08.030, **[Core rank: A/JCR IF. 7.197/Q1]** [80] .

2. Imran Razzak, Michael Blumenstein, Guandong Xu, Multi-Class Support Matrix Machines by Maximizing the Inter-class Margin for Single Trial EEG Classification, IEEE Transactions on Neural Systems and Rehabilitation Engineering, doi: 10.1109/TNSRE.2019.2913142 **[Core rank: A*/JCR IF. 3.478/Q1]**[73].

3. Imran Razzak, Raghib Abu Saris, Michael Blumenstein, and Guandong Xu. "Robust 2D Joint Sparse Principal Component Analysis With F-Norm Minimization For Sparse Modelling: 2D-RJSPCA." International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1-7 **[Conference core rank: A]** [85].

4. Imran Razzak, Ibrahim A Hameed, Michael Blumenstein and Guandong Xu, "Sparse Representation and Support Matrix Machines for Epileptic EEG Signal Classification" Journal of Translational Engineering in Health and Medicine, doi:10.1109/JTEHM.2019.2942017 **[JCR IF. 2.05/Q2]** [74].

5. Imran Razzak, Muhammad Imran, Guandong Xu, Efficient Brain Tumor Segmentation with Multiscale Two-Pathway-Group Conventional Neural Networks, IEEE Journal of Biomedical and Health Informatics, 2018 Oct 4. doi: 10.1109/JBHI.2018.287403 **[Core rank: A*/JCR IF. 4.217/Q1]** [76].

6. Imran Razzak, Muhammad Imran, Guandong Xu, "Big Data Analytics for Preventive Medicine", Neural Computing and Application, DOI https://doi.org/10.1007/s00521-019-04095-y **[JCR IF. 4.664/Q1]** [82].

7. Zafar Saeed, Rabeeh Abbasi, Imran Razzak, Guandong Xu, "Event Detection in Twitter Stream using Weighted Dynamic Heartbeat Graph Approach", IEEE Computational Intelligence Magazine, 14(3): 29-38 (2019) **JCR IF. 5.857/Q1**[92]

8. Zafar Saeed, Rabeeh Ayaz Abbasi,Imran Razzak, Onaiza Maqbool, Abida Sadaf, Guandong Xu, "Enhanced Heartbeat Graph based Temporal Networks for Emerging Event Detection on Twitter" Vol, 136,pp 115,Äì132 , Expert System with Applications, **[JCR IF. 4.292/Q1]**[91]

9. Zafar Saeed, Rabeeh Abbasi, Imran Razzak, Guandong Xu, Text Stream to Temporal Network - A Dynamic Heartbeat Graph to Detect Emerging Events on Twitter, 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2018. **(Conference core rank: A)**[93]

10. Amina Naseer, Monahil Rani, Saeeda Naz, Muhammad Imran Razzak, Muhammad Imran, Guandong Xu, "Refining Parkinson's Neurological Disorder Identification Through Deep Transfer Learning" , Neural Computing and Applications", doi.org/10.1007/s00521-019-04069-0 **[JCR IF. 4.664/Q1]**[53]

11. Zafar Saeed, Rabeeh Ayaz Abbasi, Maqbool Onaiza, Sadaf Abida, Muhammad Imran Razzak, Daud Ali, Naif Radi Aljohani, Guandong Xu. "Whats Happening around the World? A Survey and Framework on Event Detection Techniques on Twitter", Journal of Grid Computing, Springer, doi.org/10.1007/s10723-019-09482-2 **[JCR IF. 3.288/Q1]** [90]

SUBMITTED/REVISION

12. Imran Razzak, Muhammad Imran, Michael Blumenstein, Guandong Xu, Robust Sparse Support Matrix Machines for Single Trial EEG Classification, Artificial Intelligence in Medicine *Minor Revision***[Core rank: A/JCR IF. 3.574/Q1]**[75]

13. Imran Razzak, Michael Blumenstein, Guandong Xu, Robust Support Matrix Machine for Classification of Corrupted Data, IEEE Transactions on Pattern Analysis and Machine Intelligence, *In Revision*, **[Core rank: A*/JCR IF. 17.730/Q1]** [78]

14. Imran Razzak, Muhammad Khurram, Muhammad Imran, Michael Blumenstein, and Guandong Xu, Randomized Nonlinear One-Class Support Vector Machines with Bounded Loss Function for Outliers Detection, Future Generation Computer Systems, *Minor Revision* **[Core rank: A/JCR IF. 5.768/Q1]** [33].

15. Imran Razzak, Raghib Abu Saris, Michael Blumenstein, Guandong Xu, Robust Two Dimensional Joint Sparse PCA with F-norm Minimization, IEEE Transactions on Image Processing **[Core rank: A*/JCR IF. 6.79/Q1]** [81]

16. Imran Razzak, Michael Blumenstein, Guandong Xu, Support Matrix Machine via Joint L2 and Nuclear Norm Minimization Under Matrix completion Framework for Classification of Corrupted Data, IEEE Transactions on Pattern Analysis and Machine Intelligence, **[Core rank: A*/JCR IF. 17.730/Q1]**[79]

17. Imran Razzak, Khurram Zafar, Michael Blumenstein, Guandong Xu, One-Class Support Tensor Machines with Bounded Hinge Loss Function for Classification of High-dimensional Data, Future Generation Computer Systems, *Minor Revision* **[Core rank: A/JCR IF. 5.768/Q1]** [77].

Table of Contents

List of Figures

Figure **Page**

LIST OF TABLES

TABLE **Page**

Abstract

Data acquisition has improved substantially over recent years, with devices acquiring data at faster rates and increased resolution. The interpretation process, however, has only recently begun to benefit from computer technology and still struggling especially for high dimensional and noisy data. We are still short of tools to convert all such data to useful information. Traditional support vector machines (SVMs) require data to reshape each matrix into a vectors, which ultimately results in losing the important structural information of the originally featured matrix. On the other-hand, the classification of high dimensional domains poses significant challenges. In contrast, modern classification approaches such as support matrix machine assume that all entities within each input matrix can serve as the explanatory features for its label. These methods are able to capture explanatory features by regularizing the regression matrix to be low-rank. However, in real-world, the data is noisy and most of the features may be redundant as well as may be useless, which in turn affect the classification performance. Thus it is important to perform robust feature selection under robust metric learning to filter out redundant features and ignore the noisy data points for more interpretable modelling. To overcome this challenge, in this work, we have adapted two different approaches. The first problem we address is the issue of dimensionality reduction. In our first approach, we introduce two-dimensional outliers-robust principal component analysis (ORPCA) by imposing the joint constraints on the objective function **(chapter 4)**. ORPCA relaxes the orthogonal constraints and penalizes the regression coefficient, thus, it selects most important features and in the meantime, it ignores the same features that have already been selected in other principal components. To overcome the data redundancy, we further extend ORPCA and introduced additional sparsity-inducing regularization that relaxes the orthogonal constraints resulting the joint features selection **(chapter 5)**. The introduced regularization terms penalizes all regression coefficients corresponding to single feature as a whole to features jointly. Hence, 2D-JSPCA approximates to high-dimensional data in flexible manner as it has more freedom to learn low-dimensional space efficiently.

Since the nuclear norm is the best convex approximation of the matrix rank over the unit ball of matrices, this makes it more tractable to solve the resulting optimization problem. Inspired by this, in our second approach, we propose a new model to address the classification problem of high dimensionality data by jointly optimizing the both regularizer terms ($||.||_{2,1}$ and $||.||_*$) and hinge loss. In our first approach **(chapter 6)**, we combine the hinge loss and regularization terms as spectral elastic net penalty. The regulariza-

tion term which promotes the structural sparsity and shares similar sparsity patterns across multiple predictors. It is a spectral extension of the conventional elastic net that combines the property of low-rank and joint sparsity together, to deal with complex high dimensional noisy data. Furthermore, it also leverages the structural information as well as the intrinsic structure of data and avoids the inevitable upper bound. The optimization problem for the RSMM is convex, non-smooth and non-differentiable, however, the combination of hinge loss, $\ell_{2,1}$-norm and nuclear norm makes the problem nontrivial to be solved directly. To tackle this issue, we split the problem into sub-problems with the *Generalized Forward-Backward* (GFB) splitting approach to solve the optimization problem efficiently.

Support matrix machine is fragile to the presence of outliers: even few corrupted data points can arbitrarily alter the quality of the approximation, What if a fraction of columns are corrupted? Combining the recovery along with feature selection and classification could significantly improve the performance. We assume that the data consists of a low rank clean matrix plus a sparse noise matrix. We extended our work and present support matrix machine **(chapter 7)** based on matrix recovery framework under the incoherence and ambiguity conditions and able to recover intrinsic matrix of higher rank and recover data with much denser corruption. We perform matrix recovery, feature selection and classification through joint minimization of $\ell_{2,1}$ and nuclear norm. We assume that the data consists of a low rank clean matrix plus a sparse noise matrix i.e. the data matrix can be decomposed as $X = L + S$. S is the column-sparse matrix that corresponds to corrupted columns, thus at most αn columns are non zeros, L corresponds to non corrupted matrix, thus $rank(L) = r$ and $(1-\alpha)n$ columns of matrix L are non zeros, corresponding to the outliers. Since the objective function is convex, non-smooth and non-differentiable, however, the combination of hinge loss, $\ell_{2,1}$-norm and nuclear norm makes the problem nontrivial to be solved directly. To decouple the hinge loss and nuclear norm with respect to W in SMMRe, we have introduced an *auxiliary variable, and applied Lagrange multiplier*.

Although, above both methods takes full advantage of low rank assumption to exploit the strong correlation between columns and rows of each matrix and able to extract useful features, however, are originally built for binary classification problems. To improve the robustness against data that is rich in outliers, we further extend this problem and present a novel multiclass support matrix machine **(chapter 8)** by utilizing the maximization of the inter-class margins (i.e. margins between pairs of classes). The proposed model is a combination of binary hinge loss for models fitting, and elastic net penalty as a regularization on regression matrix. The binary hinge loss uses C matrices to simulate one-vs-one classifier of all classes rather than $\frac{c(c-1)}{2}$ models. The optimization problem is convex but non-smooth and non-differentiable, thus, stochastic gradient descent and the Nesterov methods cannot be applied (i.e. in convex optimization setting, sub gradient of the nuclear norm function cannot be used in standard descent approaches and as a result solving it directly is difficult). Thus, an alternative approach is required to solve it, we devise an alternating direction method (*GFB splitting*) that can handle an arbitrary non-differentiable with a proximal operator.

Several non-convex and bounded loss function has been presented to substitute the hinge loss function in order to suppress the affect of outliers and improve the robustness of support vector machines. However, there is no work done for the improvement of one-class tensor machines. Furthermore, computational complexity of traditional support tensor machines is high and increases with the increase of training samples. Thus, it limits the applicability of OCSTM for large dataset. We consider one class support tensor machines and introduce a scalable algorithm for large dataset by replacing the traditional hinge loss with bounded loss function resulting in reduction of classification error caused by outliers **(chapter 9)**. For larger dataset, we further used randomized features rather than finding the optimized support tensors which results in not only improving the robustness against outliers as well as significantly reduces the training time. To solve the corresponding optimization problem, we have presented *half quadratic optimization* to transform the objective function to same like traditional OCSTM, followed by solving it like a typical OCSTM optimization problem.

We demonstrate the significance and advantage of our methods on different available benchmark datasets such as person identification, face recognition and EEG classification. Results showed that our methods achieved significantly better performance both in terms of time and accuracy for solving the classification problem of highly correlated matrix data as compared to state-of-the-art methods.