ISSN 2278-3091

Volume 9, No.1.1, 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse0991.12020.pdf

https://doi.org/10.30534/ijatcse/2020/0991.12020

Modified Ensemble of Pruned set for non-linear dataset

Oneil B. Victoriano[1], Arnel C. Fajardo[2]

[1]Technological Institute of the Philippines, Philippines, obvictoriano@addu.edu.ph

[2] Manuel L. Quezon University, Philippines, acfajardo2011@gmail.com

ABSTRACT

Current MLL studies show dataset characterization has effect on the performance of certain MLL algorithm. With MLL dataset characteristics as imbalance and with label dependency issues, it is the research hypothesis that dataset linearity characteristic affects algorithm performance. The study used Soil Test Report as the nonlinear dataset for Ensemble of Pruned set. EPS with different base classifier was modified to test the hypothesis. EPS with non-linear base classifier works better in MLL dataset with non-linear in character.

Key words: Ensemble of Pruned Set, Linearity, Multi-label learning.

1. INTRODUCTION

In simple classification problem, a single instance is associated with one label; the label can either be binary or multiclass values [1]. Multi-label learning or classification is considered a multi-output classification [2]. Common approaches to solved multi-label classification problem are tailored for specific classifiers and don't solved label dependencies or inter-label correlations [3].

Label Powerset (LP) considers each distinctive instance of set of labels in the training dataset as one group for newly modified dataset [4]. LP was extended to Pruned set (PS) by pruning away the label-sets that are occurring less time than a small user-defined threshold [5]. Ensemble of Pruned Set was the extension of PS, as ensemble learning improves ability of a learning system and reduce over fitting [6].

However, EPS complexity is not reduced with respect to LP as well as the problem of imbalance and label dependency issues [7] is not solved. Moreover, dataset analysis is highly important to select the most optimal algorithm depending on the characteristics of the dataset [8]. Multi-label Learning dataset characteristics are imbalance, high dimensional, and label relationship could hamper the performance of the algorithm; this leads to the hypothesis that dataset linearity issues might affect algorithms' performance.

The study modified EPS and improved prediction performance for the nonlinear dataset. First, the dataset was tested for linearly separability. Second, the dataset was known as nonlinear and was feed in original EPS with SVM base classifier. Third, the EPS base classifier was modified with J48, a nonlinear base classifier. Results shown that modified EPS was with better performance with nonlinear base classifier for nonlinear dataset, than the original EPS with a linear base classifier.

2. REVIEW OF RELATED LITERATURE

2.1 Multi-label Classification

Multi-label classification is the construction of a predictive model for each instance that may have many labels associated with it from the previously defined set of la-bels. MLC algorithms are categorized into three main groups: problem transformation [9], algorithm adaptation [10] and ensemble methods [6]. Problem Transformation methods examples are Label Powerset method (LP), Pruned Problem Transformation Methods or Pruned Set (PS), and Classifier Chains (CC). LP generates a single-label dataset from combinations of labels. LP takes into account label correlations problem but complex which leads to imbalance dataset [5] and make the learning process more difficult [8].

Multi-label learning has moderately attracted notable researches to diverse problems from automatic tagging for multimedia objects including images [11] [12], audio [13] [14] [15] [16], bioinformatics [17], document categorization [18] [19], information retrieval [20] [21], medical diagnosis [22] [23] [24] [25] [26] [28], rule mining [28], and web mining [29].

2.2 Ensemble of Pruned Set

Pruned Set was a modified Label Powerset transformation method less both complexity problems and not balanced class label by pruning less frequent labelsets. PS was extended to Ensemble of Pruned Set (EPS) to stabilize the negative issues between information loss. The EPS study shows superior options to other multi-label methods (CM, BM, RM, PS, RAkEL) over different multi-labelled datasets.

Multi-label ensembles can be homogeneous or heterogeneous, the former consist of a single-learning base algorithm, and latter consist of different and multi types of learning models. EPS is characterized as homogenous, fusion, and global [30]. In EPS, SVM is employed as the single label classifier [31]

2.3 Multi-label dataset characterization and linearity

A MLL datasets usually characterizes as imbalance, high dimensionality or its relation-ship its among labels. Dataset characterization [32] [33] using available tools like MLDA [34] and MEKA is essential for the different MLL algorithms performance. Performance of a classifier in one dataset may not the same in other datasets or problem domain [35].

MLL datasets are non-linear [36] due to label complexity and feature dependencies [27]. MLL classifier studies which focuses on linearity of datasets, like Ada-Boost.MH [37], BP-MLL [38], Join-SVM [39], KNN classifiers, LM-K [40], ML-kNN (Multi-label K-Nearest Neighbors) [41], One-verse-all OVA linear classifiers [42] [43], Perception in neural network [44], and RankSVM [45]

3. METHODOLOGY

3.1 Gather and preprocess the datasets

The study used the Department of Agriculture Region Field Office XI (DARFOXI) Regional Soils Laboratory Soil Test Report, an aggregated report from Soil Nutrient, Results of Analysis, Nutrient Requirements and Fertilizer Recommendation Reports. The dataset is made up of 31 features and 30+ labels as recommended fertilizers.

Due to islands of record that are not in the same format and with missing entries, the study manually searched, recoded, and retyped some information using MS Ex-cel. Besides, add and convert missing values, the original rows are being encoded with multiple crop values in the crops column. The numbers of rows were expanded to 5,870 rows from the original transaction of 3,978 rows.

3.2 Transform dataset to Label Powerset (LP)

The dataset was transformed into single labeled dataset using MLDA tool. The Label Powerset (LP) transformation method was used to transform the dataset into single labeled dataset.

3.3 Dataset linearity characteristics

Exploratory data analysis on linear separability characteristics using Weka was used in the transformed dataset. The SMO with linear kernel (SVM) and Logistic Regression were used to test dataset linearity characteristic.

4.4 Modify the EPS with non-linear base classifier

The original EPS is with SVM base classifier; the study modifies EPS base classifier with J48 for non-linear dataset.

5.5 Test the Modified EPS performance

To test the performance, the datasets were feed in EPS with SVM and J48 base classifier. Multi-label classification performance metrics: Accuracy, Jaccard Index, Hamming score, Exact match, Jaccard distance, Hamming loss, ZeroOne loss, Avg precision and F1 (micro averaged) were used in the study.

Dataset is the transactional soils test reports filed by the Department of Agriculture Region Field Office XI (DARFOXI) – Regional Soils Laboratory's encoder. The study used spreadsheet Soil Test Report, an aggregated report from Soil Nutrient, Results of Analysis, Nutrient Requirements and Fertilizer Recommendation Report.

4. RESULTS AND DISCUSSIONS

4.1 Dataset linear characteristics

A linear model was built using Logistic regression and SVM with a linear kernel in Weka. Simple Logistic Regression and SVM classifier shows Incorrectly Classified Instances is 68% and 80% respectively; and kappa statistic at 0.29 and 0.12 respectively. The initial method is to fit the data into a linear model but with higher incorrect classified instances and low kappa statistic (not close to 1), the linear model is not good. Hypothesis on using a non-linear base classifier is appropriate in EPS.

4.2 Linearity in EPS

The result shows that all performance metrics are with better performance on EPS with non-linear base classifier (J48) than that of using SVM base classifier.

Table I shows prediction performance results using EPS with linear and non-linear base classifier on the datasets. Metrics with better higher performance like Accuracy, Jaccard index, Hamming Score, Exact match, Average precision, and F1 (macro averaged) have average results of 0.822 in using J48, than 0.559 in using SVM. While, metrics with better performance if low scores in Jaccard distance, Hamming loss, and ZeroOne loss have an average results of 0.247 in J48, than 0.567 in SVM.

Table 1: Performance Metric table

Performance Metric	***Better performance***	***J48***	***SMO***
Accuracy	Higher	**0.817**	0.577
Jaccard index	Higher	**0.814**	0.587
Hamming score	Higher	**0.923**	0.713
Exact match	Higher	**0.725**	0.415
Jaccard distance	Lower	**0.279**	0.624
Hamming loss	Lower	**0.097**	0.298
ZeroOne loss	Lower	**0.365**	0.781
Avg precision	Higher	**0.818**	0.562
F1 (micro averaged)	Higher	**0.839**	0.503

5. CONCLUSION AND RECOMMENDATION

Multi-label classification is associated with multiple classes per instance. In current researches, MLL dataset characteristics are imbalance, high dimensional, and label relationship hampers the performance of a certain MLL algorithm. A hypothesis of MLL dataset linearity characteristic could affect MLL performance was discovered in this study. The study modifies EPS base classifier for nonlinear dataset. The results concluded that there is effect in performance metrics in using EPS with linear and non-linear base classifier on the dataset. The non-linearly separated dataset is with better performance in EPS with J48 a non-linear base classifier (J48 compared in EPS with a SVM linear base classifier. In the future, the study will look into embedding or dimensionality reduction in MLL to support further study on MLL dataset linearity characterization. Also, the study will look into the effects of other current MLL dataset characteristics compared with MLL dataset linearity characteristics.

REFERENCES

1. W. Zhang, F. Liu, L. Luo, and J. Zhang. **Predicting drug side effects by multi-label learning and ensemble learning**, BMC bioinformatics, 16, 365, doi:10.1186/s12859-015-0774-y, 2015.
2. D. Xu, Y. Shi, I. Tsang, Y. Ong, C. Gong, and X. Shen. **A Survey on Multi-output Learning**, Web published in Cornel University. https://arXiv.org/abs/190.1.00248, 2019.
3. L. Rokach, A. Schclar, and E. Itach. **Ensemble Methods for Multi-label Classification**, Expert Syst. Appl., 41, 7507-7523, 2014. https://doi.org/10.1016/j.eswa.2014.06.015
4. J. Nareshpalsingh, and H. Modi. **Multi-label Classification Methods: A Comparative study**, International Research Journal of Engineering and Technology (IRJET), Volume 4, Issue 12, 2017.
5. O. Gharroudi. **Ensemble multi-label learning in supervised and semi-supervised settings**. Apprentissage multi-label ensembliste dans le context supervisé et semi-supervisé, 2017.
6. S. Sharma, and S. Kumar. **Analysis of Ensemble Models for Aging Related Bug Prediction in Software Systems**, 13th International Conference on Software Technologies, 2018.
7. Z. Abdallah, M. Oueidat, and A. El-Zaart. **An Improvement of Multi-Label Image Classification Method Based on Histogram of Oriented Gradient**, International Journal of Computer, Electrical, Automation, Control and Information Engineering. Vol:11, 2017.
8. J. Moyano, E. Gibaja, K. Cios, and S. Ventura. **Review of ensembles of multi-label classifiers: Models, experimental study and prospects**, An international journal in Information Fusion, Science Direct, 2018. https://doi.org/10.1016/j.inffus.2017.12.001
9. J. Nareshpalsingh, and H. Modi. **Multi-label Classification Methods: A Comparative study**, International Research Journal of Engineering and Technology (IRJET), Volume 4, Issue 12, 2017.
10. A. Santos, A. Canuto, and A. Neto. **A comparative analysis of classification methods to multi-label tasks in different application domains**, International Journal of Computer Information Systems and Industrial Management Applications. ISSN 2150-7988 Volume 3, 2016, pp. 218-227.
11. M. Wang, X. Zhou, and T.S. Chua. **Automatic image annotation via local multi-label classification**, Proceedings of the 7th ACM International Conference on Image and Video Retrieval, Niagara Falls, Canada, 2015, pp. 17-26.
12. Q. Wang, N. Jia, and T. Breckon. **A Baseline for Multi-Label Image Classification Using Ensemble Deep CNN**, 2018.
13. H. Ahsan, V. Kumar, and C. V. Jawahar. **Multi-label annotation of music**, 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR), 1-5, 2015.
14. S. Oramas, O. Nieto, F. Barbieri, and X. Serra. **Multi-Label Music Genre Classification from Audio, Text and Images Using Deep Features**. ISMIR, 2017.
15. C. Sanden, and J. Z. Zhang. **Enhancing multi-label music genre classification through ensemble techniques,** Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Beijing, China, 2015, pp. 705–714.
16. K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas. **Multilabel classification of music into emotions**, Proceedings of the 9th International Conference on Music Information Retrieval, Philadephia, 2015, pp 325–330.
17. M. L. Zhang, and Z. H. Zhou. **Multilabel neural networks with applications to functional genomics and text categorization**, IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 10, 2016, pp. 1338–1351. https://doi.org/10.1109/TKDE.2006.162
18. K. Glinka, and Z. Danuta. **Effective Multi-label Classification Method with Applications to Text Document Categorization**, Information Systems in Managemen, Volume 5(1), 2016, pp. 24-35.
19. M. Pushpa, and S. Karpagavallia. **Multi-label Classification: Problem Transformation methods in Tamil Phoneme classification**, Procedia Computer Science, Volume 115, 2017, pp. 572-579.
20. S. Gopal, and Y. Yang. **Multilabel classification with meta-level features**, Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and

Development in Information Retrieval, Geneva, Switzerland, 2015, pp. 315–322.

21. S. Zhu, X. Ji, W. Xu, and Y. Gong. **Multi-labelled classification using maximum entropy method**, Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, 2015, pp. 274–281.
22. T. Baumel, J. Nassour-Kassis, M. Elhadad, and N. Elhadad. **Multi-Label Classification of Patient Notes a Case Study on ICD Code Assignment**, AAAI Workshops, 2018.
23. Z. Ceylan, and E. Pekel. **Comparison of Multi-Label Classification Methods for Prediagnosis of Cervical Cancer**, International Journal of Intelligent Systems and Applications in Engineering. 5. 10.18201/ijisae.2017533896, 2017.
24. R. Li, W. Liu, L. Yusong, H. Zhao, and C. Zhang. **An Ensemble Multilabel Classification for Disease Risk Prediction**, Journal of Healthcare Engineering Volume, Article ID 8051673, 2017.
25. P. Nigam. **Applying Deep Learning to ICD-9 Multi-label Classification from Medical Records**, 2016.
26. C. Salvatore, and I. Castiglioni. **A wrapped multi-label classifier for the automatic diagnosis and prognosis of Alzheimer's disease**. Journal of Neuroscience Methods, 302, 58-65, 2018.
27. D. Senthilkumar, and S. Paulraj. **Ensemble Deep Learning for Multi Label Classification in the Design of Clinical Decision Support System**, Asian Journal of Information Technology, Volume: 15, Issue: 15, DOI: 10.3923/ajit.2016.2632.2637, 2016, pp. 2632-2637.
28. A. Veloso, W. Meira, M. A. Gonçalves, and M. J. Zaki. **Multi-label lazy associative classification**, Lecture Notes in Artificial Intelligence 4702, J. N. Kok, J. Koronacki, R. L. de Mantaras, S. Matwin, D. Mladeniˇc, and A. Skowron, Eds. Berlin: Springer, 2016, pp. 605-612.
29. L. Tang, S. Rajan, and V. K. Narayanan. **Large scale multi-label classification via metalabeler**, Proceedings of the 19th International Conference on World Wide Web, Madrid, Spain, 2015, 2015, pp. 211–220.
30. S. Khanala, J. Fultonb, A. Klopfensteinb, N. Douridasc, and S. Shearerb. **Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield**, Computers and Electronics in Agriculture, 153(April), 2018, pp. 213-225. https://doi.org/10.1016/j.compag.2018.07.016
31. Y. Papanikolaou, G. Tsoumakas, M. Laliotis, N. Markantonatos, and I. Vlahavas. **Large-scale online semantic indexing of biomedical articles via an ensemble of multi-label classification models**. Journal of Biomedical Semantics, 2017.
32. L. Chekina, L. Rokach, and B. Shapira. **Meta-learning for Selecting a Multi-label Classification Algorithm**, 2011 IEEE 11th International Conference on Data Mining Workshops, 2011, pp. 220-227.
33. F. Pakrashi, D. Greene, and B. Mac Namee. **Benchmarking Multi-label Classification Algorithms**, 24th Irish Conference on Artificial Intelligence and Cognitive Science (AICS'16), At Dublin, Ireland, 2016.
34. F. Charte, A. J. Rivera, M. J. Jesús, and F. Herrera. **Addressing imbalance in multilabel classification: Measures and random resampling algorithms**, Neurocomputing, 163, 2015. pp. 3-16.
35. R. Venkatesan, M. J. Er, and M. Dave. **Evolving Systems**, Springer Berlin Heidelberg, https://doi.org/10.1007/s12530-016-9162-8, Online ISSN 1868-6486, 2017.
36. O. Luaces, J. Diez, J.J. del Coz, J. Barranquero, and A. Bahamonde. **Synthetic Datasets for Sound Experimental Evaluation of Multilabel Classifiers**, Artificial Intelligence Center, University of Oviedo at Gijon, Asturias, Spain, 2015.
37. P. Peng, Y. Zhang, Y. Wu, and H. Zhang. **An Effective Fault Diagnosis Approach Based On Gentle AdaBoost and AdaBoost.MH**, 2018 IEEE International Conference on Automation, Electronics and Electrical Engineering (AUTEEE), 2019.
38. M. Zhang, Y. Li, X. Liu, and X. Geng. **Binary relevance for multi-label learning: an overview**, Frontiers of Computer Science, 12(2), 2018, pp. 191-202.
39. H. Xiong, S. Szedmak, and J. Piater. **Implicit Learning of Simpler Output Kernels for Multi-Label Prediction**, Institute of Computer Science, University of Innsbruck Innsbruck, A-6020, Austria, 2015.
40. W. Liu, D. Xu, I. W. Tsang, and W. Zhang. **Metric Learning for Multi-Output Tasks**, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, 2019, pp. 408-422. https://doi.org/10.1109/TPAMI.2018.2794976
41. Z. Li, S. Cao, and H. Guo. **An Improved ML-kNN Multi-label Classification Model Based on Feature Dimensionality Reduction**, 2017.
42. X. Guan, J. Liang, Y. Qian, and Pang. **A multi-view OVA model based on decision tree for multi-classification tasks**, Knowledge-Based Systems 138, 2017, pp. 208-219.
43. H. Fang, M. Cheng, C. Hsieh, and M. P. Friedlander. **Fast Training for Large-Scale One-versus-All Linear Classifiers using Tree-Structured Initialization**, 2019.
44. A. Alali. **A Novel Stacking Method for Multi-label Classification**, Open Access Dissertations, no 584, 2016.
45. K. Chen, R. Li, Y. Dou, Z. Liang, and L. Qi. **Ranking Support Vector Machine with Kernel Approximation**, Computational Intelligence and Neuroscience, Volume 2017, Article ID 4629534, 2017. https://doi.org/10.1155/2017/4629534