Trending Challenges in Multi Label Classification

Raed Alazaidah
School of Computing
Universiti Utara Malaysia (UUM)
Sintok-Kedah, Malaysia

Farzana Kabir Ahmad
School of Computing
Universiti Utara Malaysia (UUM)
Sintok-Kedah, Malaysia

***Abstract*—Multi label classification has become a very important paradigm in the last few years because of the increasing domains that it can be applied to. Many researchers have developed many algorithms to solve the problem of multi label classification. Nerveless, there are still some stuck problems that need to be investigated in depth. The aim of this paper is to provide researchers with a brief introduction to the problem of multi label classification, and introduce some of the most trending challenges.**

Keywords—Challenges; Correlations among labels; Multi Label Classification

I. Introduction

Classification is an important data mining task that could be defined as the prediction of class label for unseen instances as accurate as possible [1]. Most researchers are interested in single label classification, where the goal is to learn from a set of instances that are associated with a unique class label from a set of disjoint class labels. If the total number of disjoint classes equals two, then the problem is called binary classification, otherwise, the problem is a multi class classification. On the contrary of the previous problems, Multi-Label Classification (MLC) allows the examples (instances) to be associated with more than one class label at the same time. So, the goal of MLC is to learn from set of instances, where each instance belongs to one or more class labels at the same time [2].

MLC was motivated firstly by text categorization and medical diagnosis [3]. Recently, more researchers pay great attention toward the problem of MLC due to its importance in the real world problems [3]. In many domains where single label classification failed to solve the classification problem, MLC did. For example, single label classification may tag an email message as work or research project but not both, where the fact is, it could be tagged as both work and research project at the same time, which MLC does.

Nowadays, MLC is increasingly required by modern applications such as music categorization into emotions [4] , semantic video annotation [5], direct marketing [6], protein function classification [7] and semantic scene classification [8].

MLC is - by its nature- a challengeable problem due to many reasons such as the huge number of labels combinations that grows exponentially, high dimensionality, unbalanced data, and many other reasons [9]. This paper aims to pin point to the most trending challenges in MLC based on extensive study of many recent researches and articles. These challenges include but not limited to : exploiting correlations among labels from both types conditional and unconditional dependencies, features selection methods that are designed especially to handle multi label datasets, and having new stratification methods that are suitable to the nature of multi label datasets.

This paper is organized as follows. In the next section, we present some of the related work. In section 3, Trending challenges in the field of MLC are introduced. Finally, we conclude and present some of the future works.

II. Related Work

According to [1], there are two approaches that are widely used to handle the problem of MLC: Problem Transformation Methods (PTM) and Algorithm Adaptation Methods (AAM). The former transforms the multi label problem into one or more single label classification problems, that could be solved using any single label classification algorithm[9]. The latter extends a single label algorithm to directly handle a multi label data.

A. Problem Transformation Methods

An algorithmic independent methods that handle multi label datasets by transforming it to single label dataset or more as a preprocessing step, and then apply any single label classification algorithm. In fact, there are many transformation methods which could be grouped into two groups:

1) Simple Problem Transformation Methods

The most simple straightforward method is the *ignore* method, which ignores any multi label instances that exist in the dataset [9]. This naïve method is unacceptable, since it causes much of information loss. Other simple methods calculate the frequency of each label and then either select the *most frequent* label, *least frequent* label or *randomly* select any label as transformation criteria [10].

Transformation methods based on label frequency do not reflect any logic in solving the problem of MLC, and may cause different problems like increasing the complexity of the learning process when selecting the least frequent label or imbalance class distribution problem when selecting the most frequent label.

The last transformation method *copies* any multi label instance number of times equals to the number of labels it is associated to, with or without using a weight [11]. This method does not cause any information loss but it neglects the important correlations among labels and may increase the complexity of the learning process through increasing the number of single label instances in the dataset.

2) Complex Problem Transformation Methods

Roughly speaking, most complex problem transformation methods are based on or inspired by two famous methods : *Binary Relevance (BR)* and *Label Powerset (LP)*[12]. Each algorithm represents different approach in handling the problem of MLC.

BR divides the multi label dataset into q different datasets with each dataset contains all the positive and negative instances for specific label [12]. It then trains q classifiers for all datasets and merge the prediction of all these classifiers to get the final predictions. BR may considered to be simple method with linear complexity with respect to the total number of labels and has the advantage of being executed in parallel, but suffers from many limitations such as : It neglects any correlations among labels, and considers labels to be mutual exclusive, which is totally not correct when handling the problem of MLC. Another limitation for BR is the complexity of the method in the case of huge number of labels [11].

On the contrary of BR, LP considers correlations among labels as it treats every unique combination in the dataset as single class in multi class classification problem. LP exactly transforms MLC problem into multi class problem, and then trains any single label classifier [12]. LP suffers from several drawback as the problem of imbalance class distribution, especially when the number of distinct label sets is high compared to the number of instances in the dataset. Also, LP is capable to predict only those combinations that appeared in the training phase [12].

Although BR and LP are suffering from several limitations, but they inspired many researchers to design many algorithms based on their concepts, or try to do some enhancements to those basic transformation methods through overcoming their limitations. For example *Classifier Chains (CC)* tries to enhance BR through taking label correlations into account by training q classifier that are connected with each others in such a way that the prediction of each classifier is being added to the dataset as new feature, which is used to predict new labels [10]. CC suffers from one drawback that is related to the order of the chain. Different orders give different predictions which may influence the performance and the accuracy of the classifier. This problem has been solved by randomly ordering the classifier chains in new method called *Ensemble of classifier chains (ECC)* [13].

LP by itself has been studied intensively by many researchers, due to its simplicity and its great advantage of taking label correlations into account. The intensive studies of LP result in many algorithms that are based on LP or an enhancement of LP such as The *RAndom k-labELsets method (RAkEL)* [14] which solved the problem of imbalance class distribution of LP especially when having large number of labels. RAkEL trains an ensemble of LP classifiers, where each classifier is assigned to a small subset of label combinations of size k. RAkEL has the ability to predict combinations that are not exist in the training dataset. The bottle neck of RAkEL is to determine the optimal value for the combinations size (k); if k is large enough then it will suffer from the same shortcomings of LP, and if it is small enough then it will suffer from information loss especially in correlations among labels , in addition to having low accuracy and high complexity [12].

Pruned set (PS) is another transformation method that solved the problem of imbalance class distribution in LP by pruning instances that have frequency less that specific user defined threshold [13]. This technique reduces the high complexity of LP by considering only the important and frequent combinations of label sets. The price of this solution is to lose important information, and increase the probability of overfitting. An *Ensemble of Pruned Sets (EPS)* [13] enhanced the prediction of PS by considering the prediction of multiple classifiers obtaining by voting while increasing the complexity of the algorithm.

Different approach to solve the problem of MLC is based on Pairwise Methods. The Ranking by Pairwise Comparison (RPC) transformation method divides a dataset with q labels into q(q-1)/2 datasets for each pair of labels [15]. Then a binary classifier is trained for each dataset, and a final prediction is built based on counting the votes for each label. RPC was extended by adding a virtual label that has been used as split point between relevant and irrelevant labels. This transformation method is called *Calibrated Label Ranking (CLR)* [16].

B. Algorithm Adaptation Methods

The high efficiency of many algorithms in handling single label classification problems has inspired many researchers to adapt and enhance these algorithms to handle the problem of MLC. ML-C4.5 [17] adapted the popular algorithm C4.5 to handle multi label datasets. Two adaptations has been carried out: the first adaptation allowed the leaves to have multi labels, while the second adaptation was the modifying of the entropy definition in order to have enough information that determine to which classes an exact pattern belonged to.

Multi class Multi label Associative Classification (MMAC) is an algorithm that follows the concepts of Associative Classification (AC) [18]. Firstly, it transforms the multi label dataset into single label dataset using copy as problem transformation method. Then it trains single label associative classifier to predict a single label using if –then rules. Finally it merges the predictions of rules that have the same antecedent to form a rule with more than one label in the consequent of the rule. It is worth mentioning that all the datasets that have been used to evaluate MMAC are single label datasets and have never been tested against multi label datasets.

Rank-SVM is a multi label ranking algorithm that is based on SVM ranking [19]. This algorithm aims to optimize the ranking loss, but suffer from not taking the important correlations among labels into account, and never been tested against datasets with huge number of labels where it is expected to show very low performance.

Several algorithms are based on the popular K -Nearest Neighbors algorithm (KNN) that is based on the technique of lazy learning. *ML-KNN* [20] is an example of these algorithms. All of these algorithms share the same first step with KNN (retrieving the k nearest example) and distinguish

themselves on the aggregation of the label sets of these examples.

Back Propagation for Multi-Label Learning (BP-MLL) is an adaptation of the traditional feed-forward neural networks. It optimizes an error function that is similar to the ranking loss [21]. *Multilabel Multiclass Perceptron (MMP)* is also another algorithm that uses neural network to handle the problem of MLC [22]. It uses one perceptron for each label as in BR, and the final prediction is calculated using the inner products. MMP is an efficient algorithm especially for large datasets with many labels [9]. Figure 1 depicts a brief taxonomy of MLL methods.

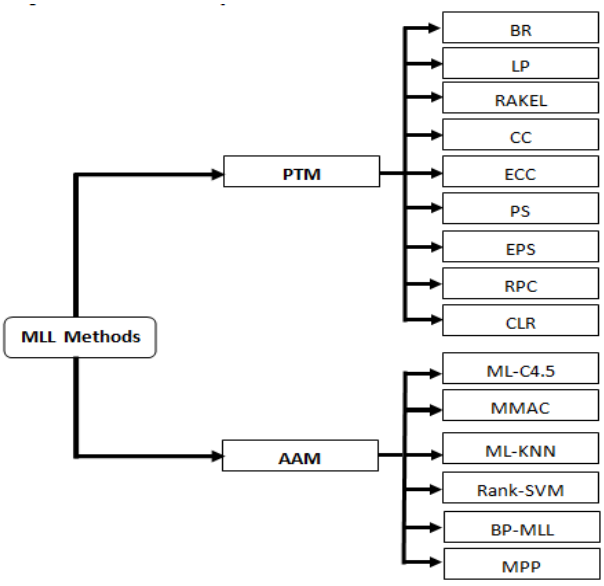

Fig. 1. MLL Methods Taxonomy

In addition to the previous way of categorizing MLC algorithms, there is another interesting way of categorization, which is based on the degree of correlations among labels that has been considered in the algorithms. Based on that, we can distinguish three types of MLC algorithms as shown in Table1.

TABLE I. CATEGORIZING MLC ALGORITHMS ACCORDING TO THE DEGREE OF CORRELATIONS AMONG LABELS

Type	Characteristics	Examples
First Order	• The task of MLL considers each label separately. • Ignore correlations with other labels. Simple and efficient. • Its results are usually suboptimal because of ignoring correlations among labels.	BR ML-KNN ML-C4.5
Second Order	• The task of MLL considers the pairwise relationships between labels like classifying labels into relevant and irrelevant labels. • Labels correlations are exploited to a limited degree.	RPC CLR BP-MLL
High Order	• The task of MLL considers the influence of every label on all other labels and finds a high order correlations among all labels or among random subsets of labels. • Demands more computations.	LP PS, EPS CC, ECC RAKEL

III. TRENDING CHALLENGES IN MLC

A. Exploiting correlations among labels to facilitate multi label learning

Multi label datasets usually have many features that do not exist in single label datasets such as high dimensionality, unbalanced data and the exponential growth of combinations of labels. These features, in addition to the core nature of multi label data; that is based on dependencies among labels, lead to an urgent need to exploit correlations among labels, in order to have additional knowledge that helps in facilitating the learning process [9]. Many algorithms [1] [11] [13] [14] [25] have tried to exploit the correlations among labels to enhance the accuracy of the multi label classifier, but most of these algorithms suffer from high complexity in the learning process [10]. Based on that, the true challenge is to exploit high order labels correlations locally and maintain a linear complexity at the same time [2].

B. Proposing new problem transformation methods based on correlations among labels

Transforming multi label datasets into one single label dataset or more is a basic step for most multi label algorithms that follow the approach of PTM. The selection of the transformation criteria is usually based on the frequency of a label. Some examples of transformation criteria are: Most Frequent Label (MFL), Least Frequent Label (LFL) or simply by selecting any label randomly [10] [11]. Since multi label datasets is based on a basic assumption which is; labels are not mutually exclusive, and they do have correlations and dependencies among them [9], it would make more sense if the transformation criteria will be based on correlations among labels [1].

C. Proposing new features selection methods that are suitable for the nature of multi label datasets

Features selection is a basic step in many data mining tasks that aims to define the relevant features in the dataset and eliminate irrelevant ones [23]. Labels in single classification are considered to be mutually exclusive, which is not completely true in MLC, and based on that, there is an urgent need to use suitable features selection methods that are designed especially to handle multi label data, and it would be even better if these features selection methods take into account the correlations among labels [23].

D. Hierarchical Multi Label Classification (H-MLC)

In some datasets, labels could be organized in a hierarchical way like "Yeast" dataset where labels are correlated to each others in a hierarchical way. Two types of structures could be used to represent the hierarchical nature of the multi label datasets: a tree or a Directed Acyclic Graph (DAG). In a tree structure a child have one and only one parent, while in DAG a child may have more than one parent at the same time [24]. It would be a nice and promising idea to design an algorithm that manage label correlations using a hierarchical structure with minimum complexity in the

learning process. Interesting approaches could be found in [24-25].

E. Proposing new stratification methods that are suitable for the nature of the multi label datasets

Stratification is a techniques that is used in sampling, and take into account the existence of all disjoint groups in the target population, so the chosen sample reflects the whole population in a representative way. In single label classification, stratification is easy since every instance is associated with only one label, and labels are mutually exclusive. Whereas in MLC, the task becomes more and more complicated as instances are usually associated with more than one label, and labels are not mutually exclusive. In [26] two stratification methods were proposed in the context of MLC, but much effort should be done to solve the problem of stratification in the field of MLC.

F. High dimensionality of label space in multi label datasets

High dimensionality is one of the most challengeable issue in MLC, and perhaps the main challenge. In MLC most labels are associated with a few number of training instances in comparison to the total number of instances in the dataset. This situation is similar to the problem of imbalance class distribution in single label classification. And the situation will be more worse when the number of labels in the dataset is very high (more than 100 labels). There is an urgent need to a simple yet fast algorithm that is capable of handling large number of labels that are associated with a few number of instances and maintaining a linear complexity at the same time. Example of such an algorithm could be found in [27] where the authors proposed new algorithm HOMER construct a hierarchy of ML classifiers where each classifier considers small subset of labels. This algorithm shows fair performance and good accuracy in only two datasets, and compared only against BR. HOMER needs to be investigated more in depth using larger datasets with a fair evaluation against other algorithms than BR.

IV. Conclusion and Future Work

In this paper, we have introduce a brief introduction to MLC. Also, we survey some of the most well known algorithms in the field of MLC. The main contribution of this paper is introducing some of the trending challenges in the domain of MLC. In the near future, we aim to investigate in depth about these trending challenges and propose new methods to exploit correlations among labels. Also, we are now evaluating new transformation methods that are based on the correlations among labels.

Acknowledgments

The first author would like to thank his family, especially his parent, for their continuous cheer on, patience and empathy. Deep thank to my great friends: Naela Alsalman and Hazem Nu'man.

References

[1] Raed Alazaidah, Fadi Thabtah and Qasem Al-Radaideh, "A Multi-Label Classification Approach Based on Correlations Among Labels" International Journal of Advanced Computer Science and Applications(IJACSA), 6(2), 2015. http://dx.doi.org/10.14569/IJACSA.2015.060208.

[2] Giorgio Corani , Mauro Scanagatta, Air pollution prediction via multi-label classification, Environmental Modelling & Software, Volume 80, June 2016, Pages 259-264.

[3] G. Tsoumakas, A. Papadopoulos, W. Qian, S. Vologiannidis, A. D'yakonov, A. Puurula, J. Read, J. Svec, and S. Semenov, "WISE 2014 challenge: Multi-label classification of print media articles to topics," in Web Information Systems Engineering - WISE 2014, Proceedings, Part II, 2014, pp. 541–548.

[4] Aiyesha Ma, Ishwar Sethi, and Nilesh Patel. 2009. Multimedia content tagging using multilabel decisiontree. In Proceedings of the 2009 11th IEEE International Symposium on Multimedia (ISM'09). 606– 611.

[5] Jingdong Wang, Yinghai Zhao, Xiuqing Wu, and Xian-Sheng Hua. 2010. A transductive multi-label learning approach for video concept detection. Pattern Recognition 44 (2010), 2274–2286.

[6] Yi Zhang, Samuel Burer, and W. Nick Street. 2006. Ensemble pruning via semi-definite programming. Journal of Machine Learning Research 7 (2006), 1315–1338.

[7] Kuo-Chen Chou, Zhi-Cheng Wu, and Xuan Xiao. 2011. iLoc-Euk: A multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. PLoS ONE 6, 3 (2011), e18258.

[8] Muhammad A. Tahir, Josef Kittler, Fei Yan, and Krystian Mikolajczyk. 2009. Kernel discriminant analysis using triangular kernel for semantic scene classification. In Proceedings of the 7th International Workshop, (CBMI'09). IEEE, Los Alamitos, CA, 1–6.

[9] E. Gibaja and S. Ventura. A tutorial on multilabel learning. ACM Computing Surveys, 47(3):Article 52, 2015.

[10] M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering, 26(8):1819– 1837, 2014.

[11] Min L. Zhang and Kun Zhang. 2010. Multi-label learning by exploiting label dependency. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10). ACM, New York, NY, 999–1008.

[12] Jesse Read. 2011. Advances in Multi-label Classification. Retrieved from http://users.ics.aalto.fi/jesse/talks/ Charla-Malaga.pdf.

[13] Jesse Read. 2008. A pruned problem transformation method for multi-label classification. In Proceedings of the NZ Computer Science Research Student Conference.

[14] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. 2010. Random k-labelsets for multi-label classification. IEEE Transactions on Knowledge and Data Engineering 23, 7 (2010), 1079–1089.

[15] Eyke H¨ ullermeier, Johannes F¨ urnkranz,Weiwei Cheng, and Klaus Brinker. 2008. Label ranking by learning pairwise preferences. Artificial Intelligence 172 (2008), 1897–1916.

[16] Klaus Brinker, Johannes F¨ urnkranz, and Eyke H¨ ullermeier. 2006. A unified model for multilabel classification and ranking. In Proceedings of the 17th European Conference on Artificial Intelligence. IOS Press, Amsterdam, The Netherlands, 489–493.

[17] Amanda Clare and Ross D. King. 2001. Knowledge discovery in multi-label phenotype data. In Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'01) (Lecture Notes in Computer Science), Vol. 2168. 42–53.

[18] Fadi A. Thabtah, Peter Cowling, and Yong hong Peng. 2004. MMAC: A new multi-class, multi-label associative classification approach. In Proceedings of the 4th IEEE International Conference on Data Mining (ICDM'04). 217–224.

[19] Aiwen Jiang, Chunheng Wang, and Yuanping Zhu. 2008. Calibrated rank-SVM for multi-label image categorization. In Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN'08). 1450–1455.

[20] Min-Ling Zhang and Zhi-Hua Zhou. 2005. A k-nearest neighbor based algorithm for multi-label classification. In Proceedings of the IEEE International Conference on Granular Computing (GrC'05). 718–721.

[21] Min-Ling Zhang and Zhi-Hua Zhou. 2006. Multilabel neural networks with applications to functional genomics and text categorization. IEEE Transactions on Knowledge and Data Engineering 18, 10 (2006), 1338–1351.

[22] Koby Crammer and Yoram Singer. 2003. A family of additive online algorithms for category ranking. Journal of Machine Learning Research 3 (March 2003), 1025–1058.

[23] Suping Xu, Xibei Yang, Hualong Yu, Dong-Jun Yu, Jingyu Yang, Eric C.C. Tsang. Multi-label learning with label-specific feature reduction. Knowledge-Based Systems. Volume 104, 15 July 2016, Pages 52–61.

[24] Celine Vens, Jan Struyf, Leander Schietgat, Saˇso Dˇzeroski, and Hendrik Blockeel. 2008. Decision trees for hierarchical multi-label classification. Machine Learning 73, 2, 185–214.

[25] Jaedong Lee, Heera Kim, Noo-ri Kim, Jee-hyong Lee. An Approach for multi-label classification by directed acyclic graph with label correlation maximization. Information Sciences. 2016.

[26] Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). On the stratification of multi-label data. Machine Learning and Knowledge Discovery in Databases, 145-158.

[27] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. 2008. Effective and efficient multilabel classification in domains with large number of labels. In Proceedings of the ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08).