Multi-label Classification Trending Challenges and Approaches

Pooja Pant, A. Sai Sabitha, Tanupriya Choudhury and Prince Dhingra

Abstract Multi-label classification (MLC) has caught the attention of researchers in various domains. MLC is a classification which assigns multiple labels to a single instance. MLC aims to train the classifier for modern applications such as sentiment classification, news classification, and text classification. MLC problem can be solved by either converting into a single-label problem or by extending machine learning methods for solving it. In this paper, the challenges faced during training the classifier which includes label space dimensionality, label drifting, and incomplete labeling are considered for review. This paper also shows the newly emerged data analysis methods for multi-label data.

Keywords Multi-label classification · Active learning · Label drifting
Hierarchal MLC

1 Introduction

Supervised learning aims at developing a learning model from a set of instances. Considering X as an instance and L as the label space, the aim of supervised learning is to build a function which can map F (X, L): X → L. Traditional classifications were used with the assumption that each instance is assigned to a single class label from a set of labels (L), L > 1. If $|L| = 2$, the learning problem is called binary classification problem [1]. Many machine learning algorithms [2] had

P. Pant · A. Sai Sabitha · T. Choudhury (✉) · P. Dhingra
Amity University Uttar Pradesh, Noida, Uttar Pradesh, India
e-mail: tchoudhury@amity.edu; tanupriya1986@gmail.com

P. Pant
e-mail: ashima81k@gmail.com

A. Sai Sabitha
e-mail: assabitha@amity.edu

P. Dhingra
e-mail: princedhingra52@yahoo.com

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_51

been developed to deal with binary classification problem or traditional classification problem. Recently, multi-label classification (MLC) had captured a lot of attention in the research community. In real-world data analysis, there is a need to assign multiple labels to a single label. This approach is called multi-label classification (MLC) and it has the potential requirement in many real-life applications. For example, in a medical field, a patient can have multiple symptoms in a single disease. The news data can be classified under various labels like entertainment, education, and public awareness. MLC problems are quite challenging because of the exponential size of the label space and the dependencies of these labels on each other. Thus, there is a need to understand and address this problem. This paper discusses the various problem transformations and adapted algorithm methods for solving MLC problems. It also discusses the various challenges faced during analyzing multi-label data, the new data analysis approaches emerged from MLC data. There are two types of classification techniques, multi-class classification and multi-label classification. Multi-class classification (MCC) is based on the assumption that every instance could be assigned to only one class label. For example, an employee of an IT can belong to one and only one department; he cannot belong to finance and programming department simultaneously [3]. The multi-label classification aims at assigning a set of labels to every instance. For example, an employee can have marketing, finance, computer, and database skills.

2 Systematic Review

Before preceding with the systematic review, three research questions (RQ) were framed related to multi-label classification. The solutions of the RQ discussed in this paper have been proposed by various researchers. The RQ questions are as follows:

RQ1. What are the various machine learning algorithms/methods that are used for multi-label classification?
RQ2. To identify the challenges faced during classification of multi-label data?
RQ3. What are the trending paradigms for multi-label data classification?

3 Search Strategies for Preliminary Studies

A number of papers related to MLC from various sources were considered for the research work. Boolean OR and Boolean AND strategies were used to search terms with similar meaning and restricting the research.

Table 1 Research paper collection from different sources

S. no.	Source	No. of result retrieved	No. of relevant papers identified
1	IEEE	19	10
2	Springer	10	5
3	Conferences	20	16
4	Other journals	33	10
5	CiteSeerX	9	3
6	Others	14	10

Information Collection

A total of 105 papers are collected using the abovementioned strategies for review purpose. All the duplicate papers were not considered. Fifty-four papers were found to be relevant and are considered for research. Multi-label classification papers published before 2000 are not considered for this research (Table 1).

4 Result

RQ1: What are the various machine learning algorithms/methods that are used for multi-label classification?

Problem transformation and adapted algorithm are the two methods for solving multi-label classification.

Problem transformation method decomposes the given problem into single-label problems in order to train the classifier [4]. In the adapted algorithm, traditional single-label classification methods are applied to multi-label data [5]. The most common way to perform multi-label classification is by problem transformation method.

4.1 Problem Transformation (PT) Method

In problem transformation method, a multi-label problem is transformed into a single-label problem and the classifier is trained accordingly. The result of the single-label transformation is again transformed to multi-label. There are a number of well-defined multi-label methods like PT1, PT2, etc.

This is explained using an example given below (Refer Table 2).

A. **PT methods**

PT1, PT2, PT3 Method

PT1 is the first problem transformation based on single-label classification. This method randomly chooses a single label of the instance and discards the rest as

Table 2 Example of multi-label dataset

Instance	L1	L2	L3	L4	L5
1	X	X			
2	X	X	X		
3				X	
4	X	X			X
5		X		X	

Table 3 PT1 and PT2 problem transformation method

Instance	L1	L2	L3	L4	L5		Instance	L1	L2	L3	L4	L5
1	X					⇒	1	X				
2			X				2			X		
3				X			3				X	
4	X						4	DISCARDED				

shown in Table 3. This can lead to loss of important information and affects the accuracy of the result. PT2, the second problem transformation method, uses the result of PT1 and discards the similar instances from the resulted dataset as shown in Table 3.

PT3 problem transformation uses the initial dataset and considers every possible label set combination (Refer Table 4). PT3 takes into consideration the relationship or the dependencies among labels in the label set. It is capable of taking distinct label sets or labels occurred in the label space. PT3 considers frequently occurring, rarely occurring and exceptional label sets with the same preference resulting in unbalancing the single-label classification. Thus, the PT3 approach works slower as compared to PT1 and PT2 (Table 5).

B. Binary Relevance

Binary relevance (BR) is the most researched work on multi-label transformation method. It transforms a multi-label data problem to “L” binary problems where each binary classifier considers each label independently (Refer Table 6).

C. Label Power Set

Label power set eliminates the label independency problem of binary relevance method. It considers each unique combination of labels. This technique uses label ranking to calculate the ranking of the labels [6] (Refer Table 7). Label power set performance can be calculated using the probability of occurrence of the labels.

Table 4 PT3 problem transformation

Instance	L4	L1 ∩ L2	L2 ∩ L4	L1 ∩ L2 ∩ L3	L1 ∩ L2 ∩ L5
1		X			
2				X	
3	X				
4					X
5			X		

Table 5 PT4 problem transformation

Instance	L1	⌐L1	L2	⌐L2	L3	⌐L3	L4	⌐L4	L5	⌐L5
1	X		X			X		X		X
2	X		X		X			X		X
3		X		X		X	X			X
4	X		X			X		X	X	
5		X	X			X	X			X

Table 6 Binary relevance

Ins	Labels	Label	Label	Label	Label
1	L1	L2	⌐L3	⌐L4	⌐L5
2	L1	L2	L3	⌐L4	⌐L5
3	⌐L1	⌐L2	⌐L3	L4	⌐L5
4	L1	L2	⌐L3	⌐L4	L5
5	⌐L1	L2	⌐L3	L4	⌐L5

Table 7 Label power set

Ins	P(c/x)	L1	L2	L3	L4	L5
L1, L2	0.4	1	1	0	0	0
L1, L2, L3	0.2	1	1	1	0	0
L4	0.1	0	0	0	1	1
L1, L2, L5	0.0	1	1	0	0	1
L2, L4	0.3	0	1	0	1	0
	$\sum P(c/x)L_c$	0.6	0.9	0.2	0.4	0.0

4.2 Adapted Algorithm

A. **Multi-label decision tree**

Multi-label decision tree adapts the principle of simple decision tree for classification in the field of genomics using a modified C4.5 algorithm. This requires entropy to split the tree. They recursively use the sum of entropy of the labels for creating a decision tree. Decision tree classification of multi-label instances paved way for the concept of hierarchal multi-label classification.

For a multi-label instance,

$$\mathrm{T} = \{(\mathrm{xi},\ \mathrm{li})\ 0 < \mathrm{i} \leq \mathrm{n}\} \tag{1}$$

The entropy (Ent) of the instance is

$$\mathrm{Ent}\,(\mathrm{T}) = \sum \mathrm{p(l)}\log(\mathrm{p(l)}) + (1 - \mathrm{p(l)})\log(1 - \mathrm{p(l)}) \tag{2}$$

where p(l) is the probability that instance has a label l and l ε L.

B. **TREE-Based Boosting**

TREE-based boosting is based on AdaBoost. It aims to reduce the hamming loss by covering and improving the loss function [7]. Weights are assigned to instances after each iteration. The hypothesis is defined by Adaboost.MR. This method is able to detect and remove the outliers, but it is susceptible to noise. AdaBoost-M1, AdaBoost-M2, AdaBoost-MH, and AdaBoost-MR are boosting techniques which can be used for multi-label data.

C. **Multi-label K-Neural Network**

Multi-label K-neural network (ML k-NN) provides a method to deal with multi-label problem using lazy learning and k-NN. This method follows the concept of error function in backpropagation. K-NN is used for traditional classification or single-label classifications.

The basic concept of k-NN is used for multi-label problems. For a training set “S”, let “E” be the training error and “Ei” be the error on (xi, li) and Cij is the actual output on jth label [8].

$$\mathrm{S} = \{(\mathrm{xi}, \mathrm{Li}) | 0 < \mathrm{i} \leq \mathrm{m}\} \tag{3}$$

$$\mathrm{Ei} = \frac{1}{|\mathrm{Li}||\mathrm{Li}|} \sum \exp\left(-\left(\mathrm{C}^{\mathrm{i}}_{\mathrm{k}} - \mathrm{C}^{\mathrm{i}}_{\mathrm{l}}\right)\right) \tag{4}$$

$$\mathrm{E} = \sum_{\mathrm{I}=1}^{\mathrm{M}} \mathrm{Ei} \tag{5}$$

D. **Rank SVM**

SVM has been considered the most successful binary classification technique which uses the concept of maximum margin strategy rank [9]. Research algorithms had been proposed for MLC, using popular classification techniques like neural network, evolutionary algorithms like gnetic algorithms.

5 Challenges of Multi-label Data Classification

RQ2. To identify the challenges faced during classification of multi-label data?

With new multi-label learning methods, new challenges are emerging. Many solutions are provided in the literature and are based on assumptions. The accuracy obtained is also based on the nature of the dataset.

A. Dimensionality

As the label space increases, the dimension of the instance space also increases. A dimensionality reduction algorithm aims at removing irrelevant and noisy data [10]. When there exists a high-dimensional label space, dimensionality reduction needs to be applied as a separate data preprocessing step. Dimensionality reduction, for single-label data, is the most intensively researched. It is based on feature extraction and feature selection methods. Dimensionality reduction can be applied to high-dimensional data using a number of unsupervised methods without modifying the data [11], Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Support Vector Machines (SVM), Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA), etc. [11, 12]. PCA is the mostly applied on feature extraction dimensionality reduction method.

B. Data cleaning

There are several issues in data cleaning step which needs to be addressed. For example, in a job recruitment application, data of job seekers are asked to mention their skills. Every seeker enters his skills, irrespective of which domain his skills belongs to. This type of multi-label data requires text analytics, tokenization, stemming, and stop word removal. This can result in loss of important information and degrade the quality of data. Malik and Bhardwaj [13] had provided a way of finding label errors in their research work.

C. Label Dependency

Label dependency or label correlation focuses on the nature of the occurrence and combinations of labels. Many proposed methods only lead to decomposing the multi-label problem to either pair label problem or binary label problem. These approaches had reduced the quality of data [14]. Creating various label combinations may further result in exponentially increased data. Simultaneously, learning from rarely occurring labels becomes a challenge.

There exist two types of label dependencies, namely, conditional label dependency and marginal label dependency [15]. Conditional dependency tends to capture the dependency of able for a given instance. Marginal dependency is said to exist when product of marginals is not equal to as given in the equation [p(y2) != p (y2|y1) p(y1) p(y2) != p(y1, y2)].

Marginal dependency can be considered as expected dependency. Multi-label classifier processes the multi-label simultaneously, and thereby it affects the performance of the classifier. Hamming loss and subset 0/1 loss are used for calculating the performance matrix. Dembczyński et al. [16] proposed a way of linking label dependencies and loss minimization, in viewing a problem with three different perspectives and minimizing the multi-label loss functions.

D. Label Uncertainty

Uncertainty of labels in real-world applications like recruitment dataset, e.g., the key skills, is to be filled by the users. This results in creation of massive new labels. Every skill can have different names henceforth adding unnecessary new labels. The same label can be repeatedly generated also. In this scenario, the most trending challenge is gaining the knowledge of all the skills of each available domain and using stemming and sub-categorization process for each label.

E. Drifting

As single instance of multi-label data has a number of labels, the interest on labels starts drifting as it is hidden conceptually. For instance, a job seeker changes his job domain or marks a new domain; this is called as "Conceptual drift". Drifting sometimes occurs suddenly and sometimes with a slow rate. Drifting of an instance can be analyzed using instance selection methods. Spyromitros [17] considered these drifting issues as the toughest challenges for any classifier.

F. Data imbalance

It is a very common problem in decision tree and SVM method that some class labels with less frequency have greater importance than frequently occurring labels. According to Min-Ling [18], Charte [19], and Xioufis [20], every multi-label dataset has few labels which tend to be more relevant than most of the commonly occurring labels. This causes label imbalance and has a wide impact on the performance of the classifier. To solve label imbalance problem, methods have been proposed by Wang [21]. His approach could reduce hamming loss but was not able to completely eliminate it. In big data gathering, most of the data are generated by sensors, which are mainly uncertain and imbalance. Extensive research works are being carried out to identify the MLC methods to address these challenges.

RQ3. What are the trending paradigms for multi-label data classification?

Manual labeling of instances is time-consuming. It becomes laborious and impractical when labeling multiple data. Active learning aims at reducing the labeling cost and minimizes the efforts required to label the instances. There is less research work on active learning of multi-label data, and limited research articles on multi-label active learning are available.

Mostly active learning is used for labeling single-label data. It considers the label that provides the most valuable information. An active learner system comprises the raw input data and an active learner. These learners may include a classifier and an expert system. The expert system does the analyses for determining the labels [22]. The learner can request supervision of his own choice, and the three active learning methods are query synthesis, pool-based [22, 23], and stream-based method.

In the query synthesis, the learner can query the unlabeled instance and the query generated by the synthesis. In the pool-based active learning method, the learner can evaluate all the instances (unlabelled) before selecting the label. The most informative instance is chosen from the unlabelled pool of instances, which is then converted into a query and sent to the expert. The expert labels the query and sends it to the label training set which in turn is sent to the pool after processing by the learner. In stream-based active learning method [24], the learner decides whether the unlabeled instances should be discarded or sent to the learner for further processing. Most of the methods check all the labels in the label space for an unlabelled instance. This results in very high processing cost. Recently, Rai [24] proposed a framework which selects an instance and pairs it with labels; the expert decides which label instance pair is of the more relevant. AURO-r approaches are able to separate relevant and irrelevant labels, and provide a better label ranking technique as compared to other multi-label active learning techniques.

MIML describes multiple instances and multiple labels linked together. In MIML, instances can be associated with multiple class labels. MIML is the most widely seen in real-life classification, for example, an image classification problem, where an image is segmented based on various instances like semantic instances as shown in Eq. (3). In the past decade, many MIML frameworks have been proposed [25, 26] (Zhang 2010). MIML-SVM converts a single instance into multi-label problem in order to solve it. MIML-boot converts a multi-instance–multi-label problem into multiple instance problems for a single label. MIML-NN framework was built for reducing the loss function. Many multi-instance–multi-label frameworks have been proposed, but they are unable to analyze large volumes of data. Many MIML frameworks have been proposed in order to reduce the loss, but these frameworks are built for either single-instance or single-label problems. MIML-fast has outperformed many MIML problems as it uses linear mapping of labels and label optimization via supervised learning models.

Multi-view multi-label learning:With huge multi-label data available, there has emerged a need of analyzing data from different views. The data for these views can be obtained from different data sources. For example, a human can be identified by fingerprint, iris scan, or lip scan. Many learning algorithms tend to concatenate multi-views of a single instance into a single-instance view. The multi-view learning methods can be classified into three broad categories: co-training, subspace learning, and multiple kernel learning. According to Sun [27], co-training is the first proposed concept for multi-view analysis. The method works on three assumptions: (a) Every view is sufficient in classifying and identifying the instance individually. (b) The view is conditionally independent irrespective of the label. (c) Target function of different predicts the same label.

Subspace learning: As the data comes from different sources, the scale of data to be processed becomes large and complex to manage. Subspace learning represents large-scale data in comparatively lower dimensionality such that accurate reconstruction is possible. Subspace learning in multimodal data aims at finding the conditional independence in order to improve the accuracy of the result [28].

Machine kernel learning: Machine Kernel Learning (MKL) is widely used for multi-view problems which require a set of kernel. $K(x_i, x_j)$ is the kernel function used for calculating the similarity between examples x_i and x_j. These kernel functions are used for calculating the dot product in the feature space such that the nonlinear mapping is performed in the input space. Thus, there is a need for multiple kernels. MKL provides multiple machine kernel learning algorithm which enables user to combine different predefined learning methods for each source. Many MKL algorithms have been developed for supervised and unsupervised learning. According to Gonen and Alpaydın [29], all the MKL algorithm for finding the kernel functions falls into five categories: (a) Fixed rules, (b) Heuristic approaches, (c) Optimization approaches, (d) Bayesian approaches, and (e) Boosting approaches.

Hierarchical Multi-label Classification (HMC) and Hierarchy of Multi-label Classifiers (HOMER):

Many multi-label classification methods are now effective in classification of data. Few instances belong to number of classes, and such data needs to be represented in a hierarchal format. Hierarchical multi-label classification is similar to other classification except that an instance can belong to two or more classes simultaneously and that a subclass automatically belongs to the superclass.

Many approaches have been proposed in recent years. Single-Label Classification (SLC) approach or SC approach transforms an HMC into SLC problem by performing classification on every individual class. This results in large creation of data as the hierarchy is not considered and the relevance for classes is ignored. Many learners may have skewed distribution of classes. The second approach is Hierarchal Single-Label classification (HSC) which applies the single-label result in a hierarchal way. In Hierarchical MLC (HMC), the classifier predicts all classes at once. HMC has been widely used for text classification. HOMER uses extended meta-labels learning and balanced distribution of labels. It creates a hierarchal tree using top-down and depth-first approach. The HMC method includes a layer of meta-labels or subset of label space, which leads to increased processing performance.

6 Conclusion and Future Scope

In this review paper, a number of methods for addressing MLC problems had been discussed. The challenges faced during multi-label data processing are discussed in detail. This paper highlights the new methods which emerged as a need for better analysis in MLC. Many data mining techniques like classification and association are identified that are used to solve these issues. Ensemble techniques like AdaBoost and bagging can be used to solve MLC. The research work needs to understand multi-label data preprocessing for big data analysis, as the classification can become very complicated since the real-world data is incomplete and imbalanced. Data reduction for large dimensional dataset and classifying multi-instance data is also a challenging task.

References

1. Zhang M, Zhou Z (2014) A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng 26(8):1819–1837
2. Chrystal B, Joseph S (2015) Multi-label classification of product reviews using structured SVM. Int J Artif Intell Appl 6:61–68
3. Dekel O, Shamir O (2010) Multiclass-multilabel classification with more classes than examples. In: Proceedings of the 13th international conference on artificial intelligence and statistics (AISTATS), pp 137–144
4. Liu H, Li X, Zhang S (2016) Learning instance correlation functions for multilabel classification. IEEE Trans Cybern 1–12
5. Read J (2010) Scalable multi-label classification. PhD Thesis, University of Waikato
6. Cherman EA, Monard MC, Metz J (2011) Multi-label problem transformation methods: a case study. CLEI Electron J 14(1), Paper 4
7. Amit Y, Dekel O, Singer Y (2007) A boosting algorithm for label covering in multilabel problems. In: Proceedings of the eleventh international conference on artificial intelligence and statistics (AISTATS-07), pp 27–34
8. Zhang M-L, Zhou Z-H (2007) ML-kNN: a lazy learning approach to multi-label learning. Pattern Recogn 40(7), 2038–2048
9. Ahuja Y, Yadav SK (2012) Multiclass classification and support vector machine. Glob J Comput Sci Technol Interdiscip 12(11), Version 1.0
10. Ji S, Ye J (2009) Linear dimensionality reduction for multi-label classification. In: IJCAI'09 Proceedings of the 21st international joint conference on artificial intelligence, pp 1077–1082
11. Sorower M (2010) A literature survey on algorithms for multi-label learning, Citeseerx.ist. psu.edu. http://citeseerx.ist.psu.edu/viewdoc/versions?doi=10.1.1.364.5612 (2016)
12. Varghese N (2012) A survey of dimensionality reduction and classification methods. Int J Comput Sci Eng Surv 3(3):45–54
13. Malik H, Bhardwaj V (2011) Automatic training data cleaning for text classification. In: 2011 IEEE 11th international conference on data mining workshops, pp 442–449
14. Read J, Puurula A, Bifet A (2014) Multi-label classification with meta-labels. In: 2014 IEEE international conference on data mining
15. Dembczynsk J, Waegeman W, Cheng W, Hullermeier E (2010) On label dependence in multi-label classification. In: International Workshop on Learning from Multi-Label Data
16. Dembczyński K, Waegeman W, Cheng W, Hüllermeier E (2012) On label dependence and loss minimization in multi-label classification. Mach Learn 88(1–2):5–45

17. Spyromitros E (2011) Dealing with concept drift and class imbalance in multi-label stream classification, thesis
18. Min-Ling Z, Li Y-K, Liu X-Y (2015) Towards class-imbalance aware multi-label learning. In: Proceedings of the 24th international joint conference on artificial intelligence (IJCAI'15)
19. Charte F, Rivera A, Jose del Jesus M, Herrera F (2013) A first approach to deal with imbalance in multilabel datasets. In: Hybrid artificial intelligent systems, pp 150–160. Springer
20. Xioufis ES (2011) Dealing with concept drift and class imbalance in multi-label stream classification. PhD thesis, Department of Computer Science, Aristotle University of Thessaloniki
21. Wang H (2016) Towards label imbalance in multi-label classification with many labels. In: Arxiv.org. https://arxiv.org/abs/1604.01304
22. Sheng-Jun H, Chen S, Zhou Z-H (2015) Multi-label active learning: query type matters. In: Proceedings of the twenty-fourth international joint conference on artificial intelligence, pp 946–952
23. Cherman EA, Grigorios T, Monard MC (2016) Active learning algorithms for multi-label data Volume 475 of the series IFIP advances in information and communication technology, pp 267–279
24. Rai P (2016) Active learning, 1st edn., pp 1–24. https://www.cs.utah.edu/~piyush/teaching/10-11-print.pdf
25. Briggs F, Fern X, Raich R (2012) Rank-loss support instance machines for MIML instance annotation. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, pp 534–542
26. Nguyen, C-T, Zhan D-C, Zhou Z-H (2013) Multimodal image annotation with multi-instance multi-label LDA. In: Proceedings of the twenty-third international joint conference on artificial intelligence, pp 1558–156
27. Sun S (2013) A survey on multi-view machine learning. Neural Comput Appl 23(7):2031–2038
28. White M, Yu Y, Zhang X, Schuurmans D (2012) Convex multi-view subspace learning. In: Advances in neural information processing systems (NIPS)
29. Gonen L, Alpaydın E (2011) Multiple kernel learning algorithms. J Mach Learn Res 12:2211–2268
30. Multiple kernel learning (2016) In: En.wikipedia.org. https://en.wikipedia.org/wiki/Multiple_kernel_learning. Accessed Sept 2016