Literature Review:

Multi-label Text Classification

By

KUSHAL GOHIL

Rutgers, the State University of New Jersey,

School of Computer Research and Education,

Professional Master's Program

Proposal for a literature review on Multi-label Text Classification

Proposal,

With vast amounts of textual data being generated every second around the globe from various sources like News Media, Social Networks, Blogs, Journals, etc., there is a lot of important information that is being overlooked, which otherwise would have significant impact. Such information is highly important, it can be a highly important factor in various fields like, defence and Security, where it could be used to detect and predict security threats, in manufacturing industry, to gain a sense of real customer insight, in Educational institutes to gather and organize information, and in many other industries. The field of Machine Learning deals with such problems. There is a lot of independent research going on in this discipline but there is no consolidated source of information that summarises them.

This problem can be solved by the consolidation of individual research findings into a single resource could help greatly enhance the understanding of the field and support further research. With my academic background in the field of Computer Sciences and Machine Learning, and my experience in building machine learning models, I would like to help consolidate these resources into a Literature Review of all the major advances in the field of Text Classification in particular for Multi-Label Text Classification.

The literature review will summarise 8 of the most fundamental research articles on Text Classification from reputed journals like Science Direct, and of the latest articles on the topic of Multilabel Classification form top IEEE journals like IEEE Transactions on Pattern Analysis and Machine Intelligence.

In order to carry out the research for the literature review a sum of $5000 would be required to cover various costs like the journal costs which are around $200, computing facility charges around $800, the review would require 4 Research assistants for 10 hours who charge $100 per hour.

I would be able to deliver the first draft of the literature review by 15th of December and the final literature review by the 20th of December. With a literature review in place, the consolidation of independent researches will enhance our current knowledge on the topic of textual classification and help bolster research on Multi-label text classification.

Abstract

With large amount of unlabelled textual data being generated every second, it is becoming extremely important to have technologies to sort them and extract meaningful information from them. We explore the concept of text classification, in particular the concept of Multi-Label text classification and some of its applications. In this paper we review different types of multi-label learning techniques and explore some of their applications.

Introduction

With the advent if the 'Information Age', a vast amount of textual data being generated every second around the globe from various sources like News Media, Social Networks, Blogs, Journals, etc. With such a huge corpus of ever growing data, there is a lot of important information that is being overlooked, which otherwise would have significant impact. Such information is highly important, it can be a highly important factor in various fields like, defence and Security, where it could be used to detect and predict security threats, in manufacturing industry, to gain a sense of real customer insight, in Educational institutes to gather and organize information, and in many other industries. The field of Machine Learning, text mining in particular, deals with such problems. Traditional text classification systems allocate a label to a text by classifying it into one of the classes, it assumes that there is no correlation in between the class labels. The Multi-Label approach to text classification problem allows the text to belong to more than one class by allocating multiple labels to it. It functions on the assumption that the classes are mutually related either by generalization or by specialization operators. In this paper we will review the concept of Text Classification, we will further look into the field of Multi-Label classification. We will look into the conventional approach to the problem of Multi-Label text classification and review them. We will also look into the varied applications of the concept of Multi-Label learning in text classification, in particular, we will explore the new Hierarchical approach for multilabel text classification.

Methods

The papers for this literature review were found from the Rutgers Digital Library with the inclusion criteria that the papers selected were peer-reviewed from reputed academic journals. The search for the papers was divided into three broad stages. Firstly we searched for papers on the topic of text classification using the keyword "Text Classification" between the years 2000 to 2010, 9,347 results were returned, we selected a paper from this that included the broad overview and application of text classification on large scale datasets. In the second stage we searched for topics on multi-label classification with the keywords, "review on Multi Label learning", "Study on Multi-label classification", between the period 2010 and 2015, we found 36 papers.

There were 3 papers selected for this stage which were either a study or a review on the topic of multi-label classification. Lastly in the final stage we searched for advances in multi-label classification using the keywords, "hierarchical multi-label classification" from the period of 2005 to 2015, more than 100 papers were found, 5 papers were selected which were either a study of Hierarchical Multi-Label Classification, or its applications in active learning and exploiting dependencies within its labels.

Results

Jeske, D. R., et al., 2007 – In this paper, the authors have proposed a data mining methodology to mine textual data from large scale datasets. The overall methodology is divided into 3 key components, first is to extract features from the text data and develop a system to analyse text data using text classification. Second is to generate relevant tracking statistic to identify useful features from the above extracted feature set. Third is to validate the inferences made by the selected tracking statistics by accounting for miss-classification errors that occurred. They formulate the methodology using the data provided by aviation safety report repository Program Tracking and Reporting Subsystem (PTRS) generated by the Federal Aviation Administration (FAA). The FAA database has a huge amount of formatter or free-style textual data on inspections of aviation entities like air carriers and flight training schools. The data generation and feature extraction was done using keyword selection as shown in the figure 1. They used Naive Bayes classifier to predict the class of success, failure and skip as shown in the figure 2.After doing a 80:20 split of the data in trainig:testing data sets, the model was trained and tested and the results of the cross validation were retrieved as shown in table (figure 3).

Zhang, M. L., et al., 2014 - This paper gives us a review on multi-label learning algorithms. With real-world objects becoming increasingly complex where each meaning simultaneously has multiple semantic meaning, the need for multi-label learning emerged. This paper goes through this concept in three section, the first part is the fundamentals on multi-label learning, the second is a review of eight different multi-label algorithms and their analyses and the third and the final part is summary of it all.

The approach of multi-labelled learning was met with many key challenges, the first being the overwhelming size of the multi-label output space that grew exponentially with addition of new labels that generated even more combinations of the labels. In order to address this challenge, the correlation between labels were identified, this helped remove unrelated combinations reducing the output space. The degree or order of correlations is described in three parts, First order learning, where the correlations are entirely ignored, the model is very simple but not effective. The second part is the Second-order learning, where pairwise relations between labels

are explored and the concept is learnt based on the interaction between pairs of labels, it is more effective that the first-order and is a better performer. The third part is the Higher-Order learning, where higher order relations are considered and the influence of each label on all other labels is accounted for. It computationally demanding and less scalable, but is highly effective. In the second section the author talks about the evaluation criteria for evaluating the performance of an algorithm. An algorithm basically tries its best to optimize a type of evaluation matrix. There are two ways to evaluate an algorithm, first is based on examples or instances, whether the examples are predicted right or not, and the second is based on label, whether the label is predicted right or not. Both these types of matrix are evaluated for classification or ranking using various matrices as shown in figure 9. In the third and the final section the author reviews algorithms used for Multi-label classification. With many algorithms implemented for this, the author chooses eight highly cited and famous algorithms that are representative of the broad spectrum of available algorithms covering a variety of design strategies, having a number of related algorithms researched along those lines. The algorithms are broadly classified into two types, the first is the Problem Transformation method, which transform the problem statement to fit established machine learning algorithms for fist-order, second-order and higher-order learning. The second being Algorithm adaptation method, where the algorithms tackle this problem by adapting popular machine learning techniques to deal with multi-label data as shown in figure 10.The results of the analysis of eight multi-label learning algorithms include their basic idea, label correlations, computational complexity, tested domains, and optimized (surrogate) metric as shown in figure 11.

Purvi,P., et al., 2012- In this paper the author explores the concept of hierarchical Multilabel classification. When a problem contains instances can be identified with more than one class, such a problem is handled by multi-label classification. The authors explore the two ways of data representation in the case of multi-label data, the flat structure and the hierarchical structure. The flat structure represented all the labels at the same level, in the hierarchical structure the labels are organized in a hierarchy with respect to their correlations. The authors explore the concept of hierarchical multi-label classification, which combines the concepts of hierarchy and multi-label classes, where the multiple labels are stored in the form of a tree or a Directed Acyclic Graph (DAG). The paper talks about two approaches to this problem, the k-Binary Classifier approach and the Single Classifier approach. The authors use the C4.5H algorithm to implement the concept of multi-label classification using k-Binary approach. They train the model on the data which is in the form of a hierarchy where the relationships between labels are stored in the form of trees as shown in figure 4. The authors used the Predictive Clustering Tree (PCT) model to implement the concept of single

classifier approach. PCT is constructed with the standard "Top-Down Induction of Decision Tree" method where the top node corresponds to a cluster containing all the data and is recursively partitioned into smaller clusters forming a tree structure as shown in figure 5. The authors used the yeast dataset containing gene information where each gene is annotated with one or more classes. The results from the tests as shown in figure 6, showed that PCT or cluster based single tree model (Clus) performed better in terms of precision and the C4.5H or the multi-tree approach performed better in terms of the coverage or recall as seen in figures 7 and figures 8.

Li, T., et al., 2006 – In this paper the authors identify the problem that in many real world applications an instance can be associated with multiple classes or labels. They presented a study of various approaches to solve the multi-label classification problem. In their study they used the Gene Dataset and the Scene dataset with multiple classes. They experimented with many algorithms like SVM-Binary, C4.8-Binary, Multi-label ADTree, and tried to solve the problem of multiple labels by creating individual classifiers for each class. The performance with this approach was not good as the data the highly segmented data was not sufficient to build reliable models and the associations among the segmented data were not considered. So, they came up with a meta-learning approach which solved the above problem with a combination of SVM Binary model and an ADTree model, where the classifiers were constructed using SVM and the result was fed to ADTree to consider their internal associations. Although this approach did not provide a major improvement in every aspect, it provided a good insight into solving the multi-label problem.

Vens, C., et al., 2008 – In this paper the authors present several approaches to a variant of classification for multi-class data that are organized in a hierarchical fashion. They explore the differences in a set of Single Class (SC) classification trees or regular trees with a single Hierarchical Multiclass Classification tree (HMC) which associates multiple classes together. Unlike SC, Hierarchical Single Class (HSC) structure takes advantage of associations between different classes. They compare the approaches for SC, HSC and the HMC on 24 yeast data sets for single-class and Gene Ontology dataset for multi-class problem. It was observed that the HMC model outperforms the HSC and the SC models in terms of predictive performance. It was also observed that the precision-recall behaviour of SC was lagging compared to HSC and HMC. The size of the HMC was the smallest as compared to HSC or the combined sizes of different SC's, this also meant the HMC model took the least amount of time for training as compared to the other models.

Alaydie, N., et al., 2012 – In this paper the authors talk about the hierarchical approach to multi – label classification. They propose the HiBLADE or Hierarchical multi-label Boosting with LAbel

DEpendency algorithm that takes advantage of not only the pre-established hierarchical taxonomy of the classes, but also effectively exploits the hidden correlation among the classes that is not shown through the class hierarchy, thereby improving the quality of the predictions. They used pre-defined hierarchical taxonomy of labels to decide upon the training set for each classifier, and then they captured and analysed the dependencies of the children for each label in the hierarchy using Bayes method and instance-based similarity. Their experimental results on several real-world bimolecular datasets show that the proposed method can improve the performance of hierarchical multi-label classification.

Santos, A., et al., 2014 – In this paper the authors proposed and evaluated the use of semi-supervised learning methods in three different types of hierarchical multi-label classification methods, comparing both the supervised(HMC-BR, HMC-LP and HMC-RAkEL) and the semi-supervised methods(HMC-SSBR, HMC-SSLP and HMC-SSRAkEL). For each method of each group, they applied five different classification methods that were used as base classifiers in the multi-label methods, namely, Support Vector Machine (SVM), k-Nearest Neighbour (k-NN), Naive Bayes (NB), Decision Tree (DT) and Repeated Incremental Pruning to Produce Error Reduction (RIPPER). On evaluation with the help of 5 different data sets the results showed that the use of semi-supervised learning had similar performance of the corresponding supervised versions effect in almost 75% of the cases, and having a superior result in some cases.

Levatic, J., et al., 2015 - In this paper the author starts by introducing the problem of multi-label datasets and provides a comparative study on the solutions for the same. The authors broadly categorise the solution into two broad categories based on the scope, the first being Local solution, where individual models are created for predicting each component and all their models were combined to get the overall model. The second type is the Global solution, where a single model predicts the entire structure. This paper like the paper by Vens, C., et al., 2008, further explore different algorithms that belong to both these groups, namely, SC, HSC, MLC and HMLC, and reports on findings based on eight different datasets using tree learning methods like PCT tree as shown in paper by Purvi,P., et al., 2012, and ensemble learning methods like Bagging and Random Forests.

Discussion

This paper takes us through the significant advances in the field of textual data classification. This paper starts with an overview of text classification and its application for large scale datasets, by Jeske, D. R., et al., 2007, it takes us through the fundamental procedures of text classification. The study by Zhang, M. L., et al., 2014 is a broad overview of

different types of multi-label learning techniques. After we explore the different techniques of multi-label learning, we explore some of its applications. The paper by Vens, C., et al.,2008, introduces the concept of hierarchy in multi-label classification and its evaluation. The paper by Li, T., et al., 2006, is another implementation of exploiting the associations between labels using a tree structure or hierarchy. The paper by Purvi,P., et al., 2012, was similar to the study by Vens, C., et al.,2008, where the topic of Hierarchical Multi-label classification was explored. The study by Alaydie, N., et al.,2012, explores the inter label dependencies in between hierarchical models leading to better results. Lastly, the paper by Santos, A., et al., 2014, explores an advanced application of multi-label classification in the form of semi-supervised learning methods.

Literature Cited

- Jeske, D. R., & Liu, R. Y. (2007). Mining and Tracking Massive Text Data: Classification, Construction of Tracking Statistics, and Inference under Misclassification. *Technometrics*, (2). 116.
- Zhang, M. L., & Zhou, Z. H. (2014). A review on multi-label learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on*, 26(8), 1819-1837.
- Purvi, P., Amit, T., & Amit, G. (2012). A Comprehensive and Comparative Study on Hierarchical Multi Label Classification. *International Journal Of Engineering And Advanced Technology*, (3), 110.
- Li, T., Zhang, C., & Zhu, S. (2006, November). Empirical Studies on Multi-label Classification. In *IcTAI* (Vol. 6, pp. 86-92).
- Vens, C., Struyf, J., Schietgat, L., Džeroski, S., & Blockeel, H. (2008). Decision trees for hierarchical multi-label classification. *Machine Learning*,*73*(2), 185-214.
- Alaydie, N., Reddy, C. K., & Fotouhi, F. (2012). Exploiting label dependency for hierarchical multi-label classification. In *Advances in Knowledge Discovery and Data Mining* (pp. 294-305). Springer Berlin Heidelberg.
- Santos, A., & Canuto, A. (2014). Applying semi-supervised learning in hierarchical multi-label classification. *Expert Systems with Applications*, *41*(14), 6075-6085.
- Levatic, J., Kocev, D., & Dzeroski, S. (2015). The importance of the label hierarchy in hierarchical multi-label classification. *Journal Of Intelligent Information Systems*, 45(2), 247-271.

Appendix

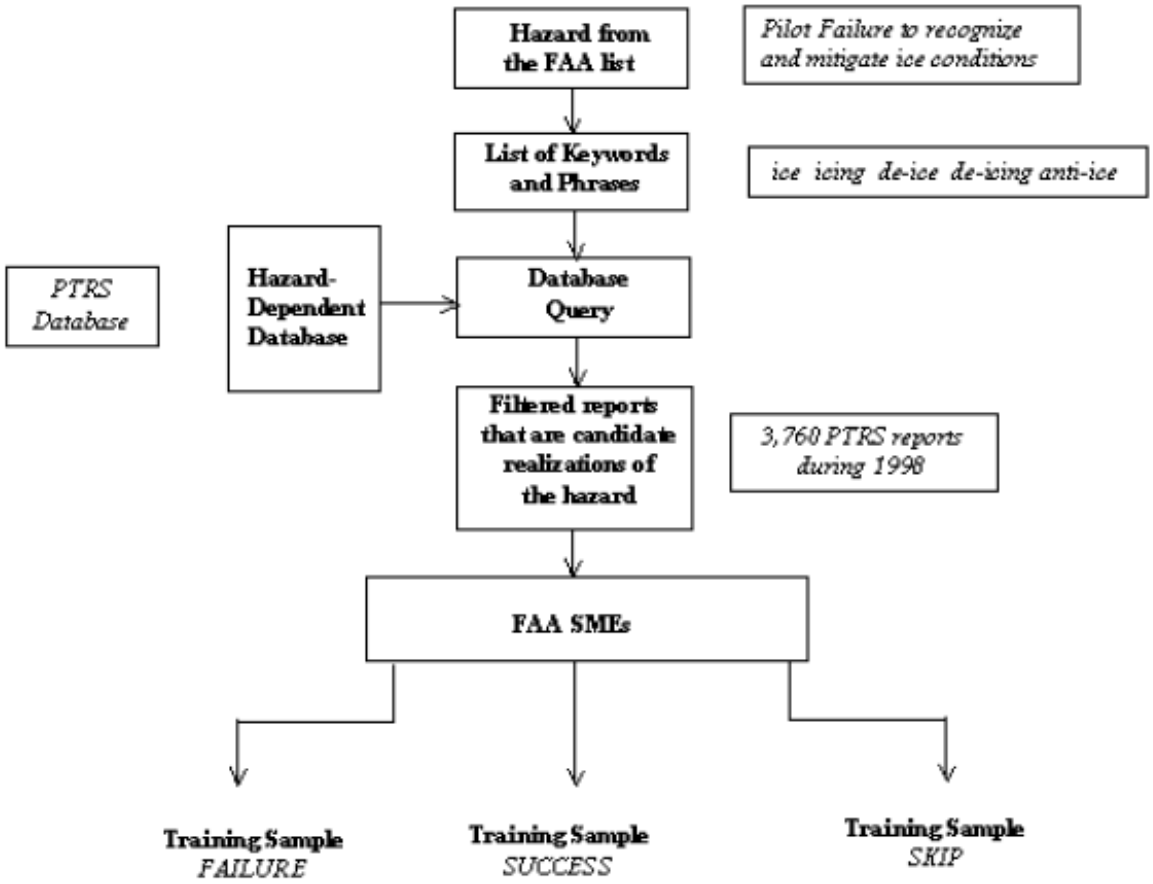

Figure 1: Processing of the PTRS Training Dataset (Jeske, D. R., et al., 2007)

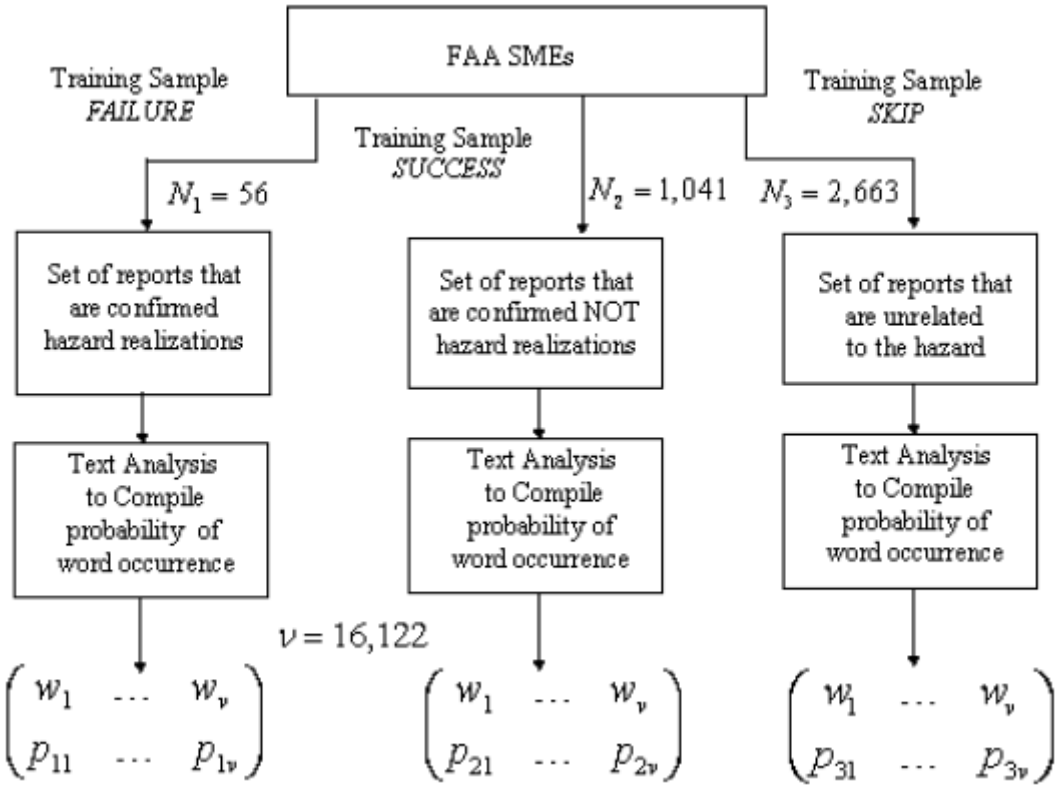

Figure 2 : Creating Word Distributions–(Jeske, D. R., et al., 2007)

Classified class	*True class*		
	FAILURE	*SUCCESS*	*SKIP*
FAILURE	42.26%	2.39%	.83%
SUCCESS	42.26%	64.18%	16.67%
SKIP	15.48%	33.43%	82.50%

Figure 3: Results of Cross-Validation–(Jeske, D. R., et al., 2007)

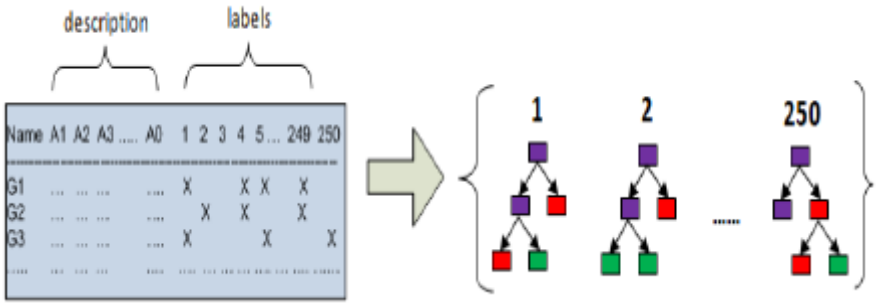

Figure 4: Decision Tree for each individual label, According to above figure, for 250 labels separate decision tree will be created. – (Purvi, P., et al., 2012)

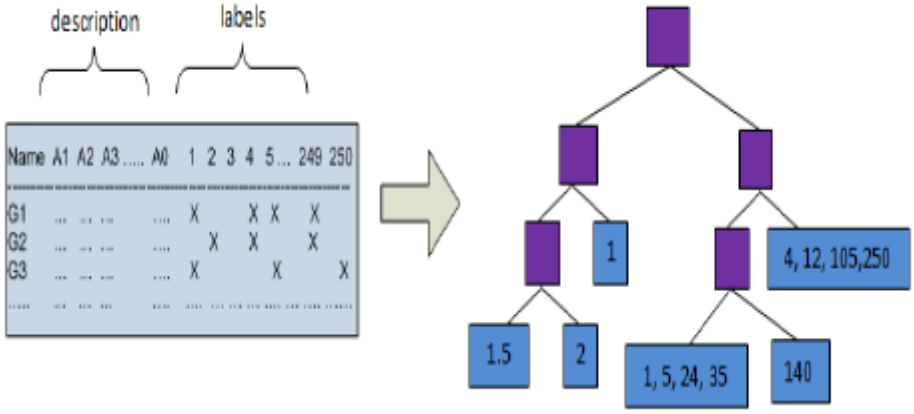

Figure 5: multiple prediction tree for multiple labels. – (Purvi, P., et al., 2012)

	Precision		Coverage	
Name	**Clus**	**C4.5H**	**Clus**	**C4.5H**
Seq	61	71	80.18	14.16
Phenol	67	68	3.09	3.26
Struc	68	58	29.71	2.05
Hom	64	55	81.41	12.06
Cellcycle	82	54	0.86	71.34
Church	75	53	8.18	58.64
Derisi	77	61	2.90	8.39
Eisen	88	48	5.73	37.63
gasch1	67	38	15.77	47.24
gasch2	96	60	3.09	64.06
Spo	79	46	3.70	12.82
Expr	77	75	27.28	5.56

Figure 6: Average precision and coverage for yeast data sets. – (Purvi, P., et al., 2012)

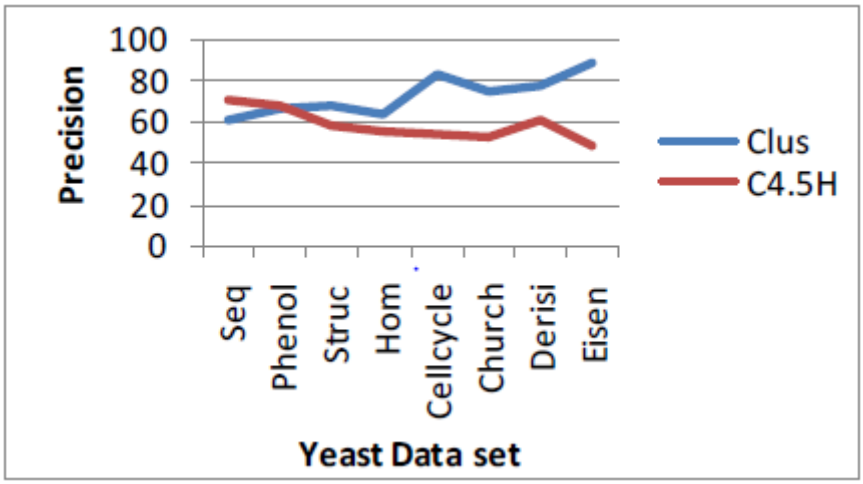

Figure 7: Average precision for yeast data sets. – (Purvi, P., et al., 2012)

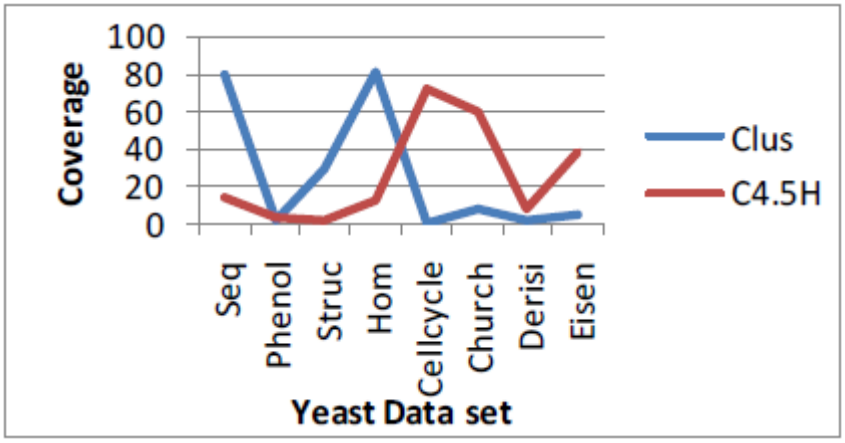

Figure 8: Average recall or Coverage for yeast data sets. – (Purvi, P., et al., 2012)

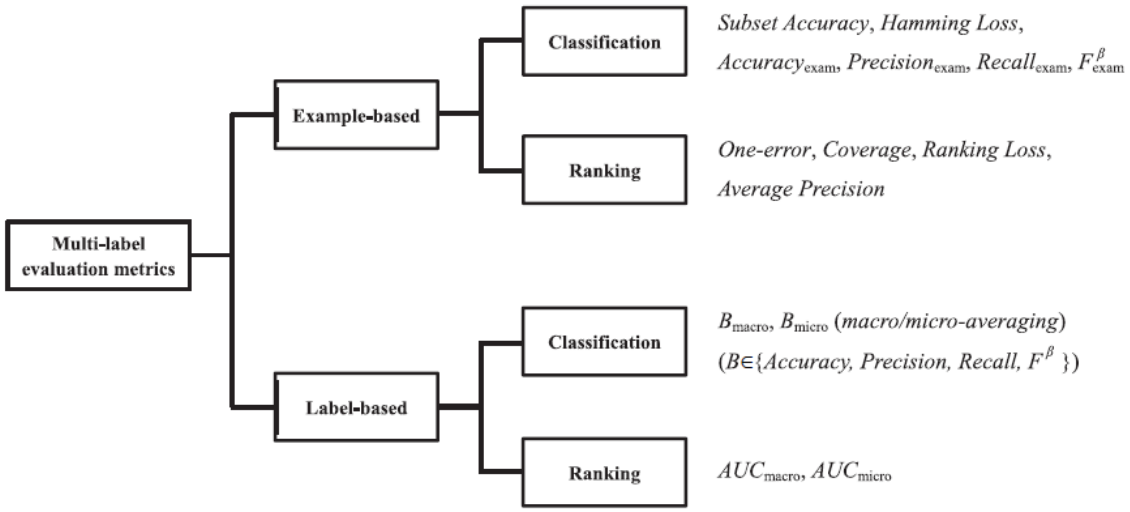

Figure 9: Summary of major multi-label evaluation metrics. – (Zhang, M. L.,et al. - 2014)

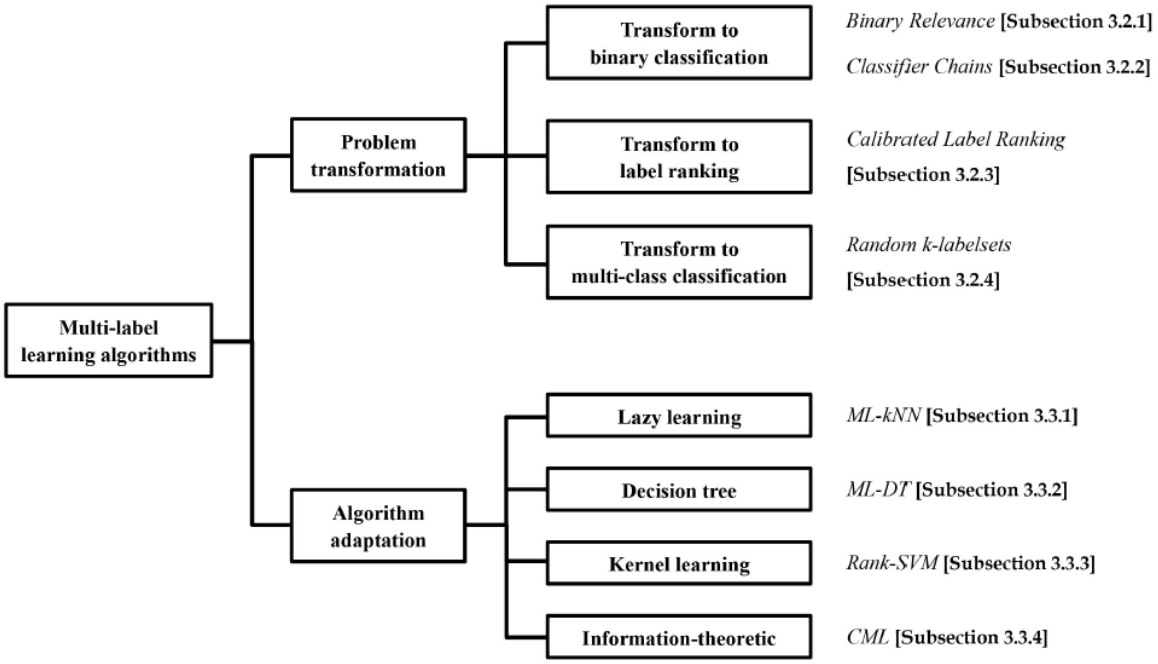

Figure 10: Categorization of representative multi-label learning algorithms being reviewed. – (Zhang, M. L.,et al. -2014)

Algorithm	Basic Idea	Order of Correlations	Complexity [Train/Test]	Tested Domains	Optimized Metric
Binary Relevance [5]	Fit multi-label data to q binary classifiers	first-order	$\mathcal{O}(q \cdot \mathcal{F}_{\mathcal{B}}(m,d))/$ $\mathcal{O}(q \cdot \mathcal{F}'_{\mathcal{B}}(d))$	image	classification (hamming loss)
Classifier Chains [72]	Fit multi-label data to a chain of binary classifiers	high-order	$\mathcal{O}(q \cdot \mathcal{F}_{\mathcal{B}}(m,d+q))/$ $\mathcal{O}(q \cdot \mathcal{F}'_{\mathcal{B}}(d+q))$	image, video text, biology	classification (hamming loss)
Calibrated Label Ranking [30]	Fit multi-label data to $\frac{q(q+1)}{2}$ binary classifiers	second-order	$\mathcal{O}(q^2 \cdot \mathcal{F}_{\mathcal{B}}(m,d))/$ $\mathcal{O}(q^2 \cdot \mathcal{F}'_{\mathcal{B}}(d))$	image, text biology	Ranking (ranking loss)
Random k-Labelsets [94]	Fit multi-label data to n multi-class classifiers	high-order	$\mathcal{O}(n \cdot \mathcal{F}_{\mathcal{M}}(m,d,2^k))/$ $\mathcal{O}(n \cdot \mathcal{F}'_{\mathcal{M}}(d,2^k))$	image, text biology	classification (subset accuracy)
ML-kNN [108]	Fit k-nearest neighbor to multi-label data	first-order	$\mathcal{O}(m^2d + qmk)/$ $\mathcal{O}(md + qk)$	image, text biology	classification (hamming loss)
ML-DT [16]	Fit decision tree to multi-label data	first-order	$\mathcal{O}(mdq)/\mathcal{O}(mq)$	biology	classification (hamming loss)
Rank-SVM [27]	Fit kernel learning to multi-label data	second-order	$\mathcal{O}(\mathcal{F}_{\mathrm{QP}}(dq + mq^2, mq^2)$ $+q^2(q+m))/\mathcal{O}(dq)$	biology	Ranking (ranking loss)
CML [33]	Fit conditional random field to multi-label data	second-order	$\mathcal{O}(\mathcal{F}_{\mathrm{UNC}}(dq + q^2, m))/$ $\mathcal{O}((dq + q^2) \cdot 2^q)$	text	classification (subset accuracy)

Figure 11: Summary of Representative Multi-Label Learning Algorithms Being Reviewed. – (Zhang, M. L.,et al. -2014)