Review on Multilabel Classification Algorithms
Ms. Prajakta C. Chaudhari¹ Prof. Dr. S. S. Sane²
¹M.E. Student ²HOD & Vice Principal
¹,²Department of Computer Engineering
1,2KKWIEER, Nashik, Savitribai Pule Pune University, India

Abstract— Multilabel classification is a framework in which each input data in training data set can be related to more than one class labels simultaneously. The goal of multilabel classification is to produce set of labels for unseen instances by analyzing training dataset. This paper presents fundamentals of multilabel classification, some multilabel classification algorithms and evaluation metrics.

Key words: Training Dataset, Multilabel Classification Algorithms

I. INTRODUCTION
Traditionally in supervised learning each real world entity is represented by single instance and single label is assigned to this instance. Since only one label from a set of disjoint labels is assigned to the instance, this classification is known as single label classification. The basic assumption adopted by traditional supervised learning is that each object belongs to only one concept. However, there are several situations where real world object may be complicated and have multiple semantic meaning where the instance may belong to multiple classes. Above problem can be solved by using multilabel classification.

There exist various applications for multilabel classification such as text categorization [1, 2] in which document may consist of more than one semantics. Also multilabel classification used in bioinformatics, image annotation, direct marketing, Medical diagnosis, Tag recommendation, protein function prediction, Query categorization [3].

This paper presents a review of multilabel classification. Second section tells about definition, statistics and correlation strategies used for multilabel classification. Third section gives performance metrics. Fourth section presents summary of algorithms used for multilabel classification. Last section gives conclusion.

II. BACKGROUND
A. Definition
In multilabel classification input data instances are associated with multiple class labels. The multilabel classification aims to find the set of labels for unseen instances. This can be better understood mathematically using set theoretic notations as follows:

Let \( A = \{a_1, a_2, \ldots, a_m\} \) be the set of instances in training set.
Let \( P = \{p_1, p_2, \ldots, p_q\} \) be the set of \( q \) labels.
Let \( D = \{(a_1, b_1), (a_2, b_2), \ldots, (a_m, b_m)\} \) be a training set.

Given a training set \( (a_i, b_i) \), multilabel classification aims to produce a function \( f(a) \) which maps each instance \( a_i \) to labels set \( b_i \) (\( b_i \) contain one or more label), where \( a_i \in A \) and \( b_i \subseteq P \).

B. Statistics
In some applications class imbalance issue occur where some instances have large number of class labels than other. This situation can affect the performance of the different multilabel classification methods. Following statistical information obtained from multilabel data plays important role in such situation:

1) Label Cardinality
It measures the average amount of labels that are associated for each instance.

\[ LCard(D) = \frac{1}{m} \sum_{i=1}^{m} |b_i| \]

Where \( m \) denotes total number of instances in training dataset \( D \).
\( b_i \) = label set associated with instance \( a_i \).

2) Label Density
It is ratio of label cardinality to total number of labels.

\[ Ldensity = \frac{1}{|P|} LCard(D) \]

Where \( |P| \) = total number of labels in training dataset.

3) Label Diversity
It denotes the amount of multiple sets of labels present in the dataset.

\[ Label\ Diversity = \frac{|\{b|\exists a : (a,b) \in D\}|}{b} \]

\( b \) = label set associated with instance \( a \).

C. Correlation strategy for multilabel classification
Different strategies for label correlation are grouped into following three categories [4]:

1) First Order Strategy
First order strategy considers each label separately thus ignoring coexistence of the other labels. This strategy convert multilabel classification problem into different binary classification problem. Advantage of first order strategy found in its simplicity and high efficiency. On the other hand, drawback of this strategy is it does not consider correlation between labels.

2) Second Order Strategy
Second order strategy considers pair wise relations among the labels. These approaches can achieve good generalization performance than first order strategy. However, there are some real world applications where relationship between labels can go beyond the second order.

3) High Order Strategy
This strategy considers relation between multiple labels such as imposing all other labels’ influences on each or addressing relationship among random subsets of labels. High order strategy represents stronger correlation capabilities than first order strategies and second order strategies, while on the other hand computationally more complex and less scalable.

D. Performance Metrics
Performance metrics for multilabel classification system is not same as that of metrics for single label classification system [5].
For test set $\mathbf{T} = \{(a_1, b_1), (a_2, b_2), \ldots, (a_N, b_N)\}$, following multilabel classification metrics are used:

1. **Hamming Loss**
   
   It calculates number of times the <instance, label> pair is misclassified.
   
   $$hloss(h) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|q|} |h(a_i) \Delta b_i|$$
   
   where $\Delta$ denotes symmetric difference between two sets.

2. **Coverage**
   
   It calculates how many times the top ranked class labels are correctly predicted by classification system are not in the set of relevant labels. Smaller value of this metric represents better performance.
   
   $$Coverage(f) = \frac{1}{N} \sum_{i=1}^{N} \max_{p \in \text{label}} \text{rank}(f(a_i, p)) - 1$$

3. **One Error**
   
   It calculates how many times the top ranked class labels are incorrectly ordered for an instance. Smaller value has better performance.
   
   $$\text{One error}(f) = \frac{1}{N} \sum_{i=1}^{N} \left[ \left\lfloor \text{argmax}_{p \in \text{label}} f(a_i, p) \right\rfloor \notin b_i \right]$$

4. **Average Precision**
   
   It evaluates the average ratio of class labels that are top ranked than actual label which belongs to the set of ground truth labels. Larger value represents better performance.
   
   $$\text{Avgprec}(f) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|b_i|} \| \text{argmax} f(a_i, p) \|$$

III. **MULTILABEL CLASSIFICATION METHODS**

Tsoumakas and Katakis [6] summarized multilabel classification approaches into two categories that are problem transformation approach and algorithm adaption approach. Recently third category called Ensemble method is introduced and one of the popular ensemble method is Random k labelsets (RAKEL) [7].

A. **Problem Transformation Approach**

The problem transformation approach, which is algorithm independent, transforms the multilabel problem into a set of single label problem. Several problem transformation methods such as Label powerset [5], Binary relevance [5] are used in multilabel classification. In Binary relevance method distinct binary classifier is designed for each label. The original dataset converted into $|L|$ different datasets that includes all instances of original dataset. For each these dataset positively labels instances are belonging to label $p_i$ and otherwise negative. The output of all these binary classifier are collected to get the final result for unseen data. Drawback of this method is it does not consider the correlation between labels.

According to Grigorios Tsoumakas [5] in label powerset method each different combination of labels in the label space appearing in the training set considered as single class. In this way the Multilabel problem is converted into single label problem but the problem is number of class increases as number of labels increases.

B. **Algorithm Adaptation Approach**

The algorithm adaptation approach on the other hand updates or extends the available machine learning algorithms to make them appropriate for handling multilabel data. The several algorithm adaptation methods have been proposed such as multilabel k nearest neighbor (ML-kNN) [8], binary relevance k nearest neighbor (BR-kNN), Multilabel decision tree (ML-DT).

Zhang and Zhou [8] proposed a multilabel k nearest neighbor (ML-kNN) for Multilabel classification. ML-kNN extends k nearest neighbor (kNN) algorithm for multilabel data. First it identifies k nearest neighbor for each instance in training data. Then prior probabilities and frequency arrays for labels are calculated. After it identifies k nearest neighbor for unseen instance. For each label statistical information is calculated with the help of k nearest neighbors of unseen instances. After that set labels for an unseen instance is evaluated via MAP (Maximum a posterior) rule which is based on Bayes theorem. Advantage of ML-kNN is class imbalance issues can be reduced due to consideration of prior probabilities. ML-kNN is based on binary relevance classifier which does not consider label correlations.

E. Spyromitros, Tsoumakas, G., Vlahavas, I. [9] proposed binary relevance k nearest neighbor (BR-kNN) method which extends k nearest neighbor (kNN) machine learning algorithm for multilabel classification in conjunction with Binary Relevance problem transformation method. BR-kNN does not consider prior and posterior probabilities.

E. Spyromitros [9] also proposed two extension of BR-kNN algorithm i.e. BR-kNN-a and BR-kNN-b which are depend on value of confidence score for each class label from BR-kNN. Confidence is calculated by considering how many percentage of k nearest neighbor it includes. The first extension BR-kNN-a checks whether BR-kNN predicts the blank set of labels, due to none of the labels being found in at least half of the kNN. If condition holds BR-kNN-a predicts the label which have highest confidence. BR-kNN-b first calculates average size of the label sets of kNN and then predicts the labels with highest confidence. One of the limitation of BR-kNN is for datasets with low cardinality it outputs the empty set. So this problem can be solved with the help of BR-kNN-a. BR-kNN-b provides improvement in datasets which have larger cardinality.
BSVM [10] uses binary decomposition method to deal with multilabel classification problems. Initially the multilabel dataset is converted into single label data using one vs. all binary decomposition method. SVM is used as base classifier for each binary classification task. Final prediction is done combining predictions of all SVMs. BSVM considers all labels and instances are independent.

Multilabel decision tree (ML-DT) [11] is algorithm adaption method where decision tree technique is adapted to handle Multilabel data. Decision tree is obtained recursively by using information gain criteria based on multilabel entropy. ML-DT considers first order strategy which considers each label separately while evaluating multilabel entropy. One advantage of ML-DT has high efficiency due to incorporating decision tree framework for multilabel data.

Chunming Liu and Longbing Cao [12] proposed a new coupled k nearest neighbor algorithm for multilabel classification (CML-kNN) which is based on lazy learning and consider inner relationship between labels. CML-kNN extends ML-kNN algorithm so as to consider label correlation. Initially Coupled label similarity is calculated with the help of inter and intra coupling similarity. Then it identifies k nearest neighbor for each instance in training data. Then prior probabilities and frequency arrays for labels are calculated in which coupled label similarity is incorporated. After it identifies k nearest neighbor for unseen instance. For each label statistical information is calculated with the help of k nearest neighbors of unseen instances. After that set labels for an unseen instances are evaluated via MAP (Maximum a posterior) rule which is based on Bayes theorem. An advantage of CML-kNN is that it considers label correlations.

C. Ensemble Method

Ensemble method uses an ensemble of Algorithm Adaptation or Problem Transformation method to deal with multilabel classification problem. RAKEL [13] overcomes the drawback of LP by restricting the size of every labelset to k. Number of labelsets is selected randomly from the training set of labels. After that LP method is applied to training data set. RAKEL predicts labels for an unseen instance which is evaluated with the help of voting of the LP classifiers. RAKEL reduces amount of classes by restricting size of every labelset to k and it is possible to have more than one instance for particular class.

IV. CONCLUSION

In this paper study of multilabel classification is presented. We discussed need of multilabel classification, their application in various areas, multilabel statistics such as label cardinality, label diversity and label density, and their performance metrics. Two approaches of multilabel classifications that are problem transformation and algorithm adaptation are also discussed. The label correlation strategies used in multilabel classification plays an important role while developing multilabel classification algorithm. Classification task can be improved by using feature selection method.

REFERENCES