Study and Analysis of Multi-Label Classification Methods in Data Mining

Shubhangi R. Khade
Dept. of Computer Engineering
Modern education society
college of engineering.
University of Pune, pune-01

Suraj R. Balwan
Dept. of Computer Engineering
Modern education society
college of engineering.
University of Pune, pune-01

ABSTRACT

Multi-label classification is major research problem in machine learning domain. Multi-label classification is nothing but the variants of classification problem in which different target labels should be allocated to every instance. Multi-label classification is different from the multiclass classification. In general, multi-label classification is defined as problem of searching model which maps the input to binary vectors, rather than outputs in scalars. Basically there are two different techniques for handling the multi-label classification problem such as techniques of problem transformation and techniques of algorithm adaptation. In problem transformation approaches, multi-label classification problem is transformed to binary classification problems set and this can be further processed through single class classifiers. In algorithm adaptation approaches, algorithms are adapted in order to perform the multi-label classification directly. In this paper, different multi-label classification algorithms are studied and evaluated with current research problems. Methods such as binary relevance (BR), high-order approaches, hierarchical tree based algorithms, and the most recent method called ML-Forest are studied and evaluated with different real time datasets such as medical, emotions, yeast etc.

Keywords

BR, HOMER, TSA, ML-Forest, Multi-label classification, datasets, accuracy

1. INTRODUCTION

The conventional single-label classification problem is associated with learning from examples set those are related to single label from the disjoint labels set. If number of labels is two, then learning problem is known as binary classification problem or filtering classification problem in data mining. If number of labels is more than 2, then it comes under the problem of multi-class classification. The aim of multi-label classification is to predict the absence or presence of certain labels of particular example which is related to different classes. As the real world objects frequently contains the multiple semantic objects, multi-label classification is more general. The example of multi-label classification such as real world image is basically associated with multiple categories depending on various contexts like ship, water etc. Also text document can be divided into different set of topics like sports, news etc. Since from last 15 years, multi-label classification problem is widely studied by various researchers in different domains such as computer vision, bioinformatics as well as text categorization. The methods of multi-label classification techniques requirement is growing now days in different application like protein function classification, music categorization, semantic scene classification etc.

The BR (binary relevance) is nothing but the straightforward multi-label classification method. BR method decomposes the main problem into a single-label multi-class sets sub-problems. According to this approach, multi-class classifiers are leant and proceed for the prediction. BR is most simple approach for multi-label classification, but this approach completely rejects the dependencies between multiple labels. Practically, the different objects in examples like images may have the possibility of strong dependencies or relations. If the category of ship is available in image, then it is sure that water category is also available in that image. Such label dependency exploitation can improve the performance of prediction significantly for problem of multi-label classification. There are number of methods proposed for exploiting the label dependency in order to enhance the prediction performance recently. But such methods are suffering of number of limitations such as how to explicitly effectively model the label dependency, over-fitting problems etc.

Recently, these problems overcome by method proposed Ml-Forest. This approach proposed for explicitly exploits the label dependency to perform multi-label classification. The goal of this paper is to study Ml-Forest method with algorithmic representation and comparative analysis against earlier methods such as BR, HOMER etc. In section II, the statistics and performance metrics to evaluate any multi-label classification technique are discussed. In section III, the different methods of multi-label classification are discussed. In section IV, the recent ML-Forest method is presented in form of algorithms. In section V, the practical evaluation and results discussed.

2. MULTI-LABEL CLASSIFICATION METRICS

The extent to which a dataset is multi-label can be captured in two statistics such as:

1. Label cardinality is the average number of labels per

$$\frac{1}{2}\sum_{i=1}^{N} |Y_i|;$$

Example in the set:

2. Label density is the number of labels per sample divided by the total number of labels, averaged over the samples: $\frac{1}{N}\sum_{i=1}^{N}\frac{|Yi|}{|L|}$

Where

Evaluation metrics for multi-label classification performance are inherently different from those used in multi-class (or binary) classification, due to the inherent differences of the classification problem. If *T* denotes the true set of labels for a given sample, and *P* the predicted set of labels, then the following metrics can be defined on that sample:

Hamming loss: the fraction of the wrong labels to the total

$$\frac{1}{|N|}\sum_{i=1}^{|N|}\frac{\sum_{j=1}^{|N|}\mathrm{XOR}\,(y_{i,j},Z_{i,j})}{|L|}$$

Number of labels, i.ewhere $y_{i,j}$ is the target and $z_{i,j}$ is the prediction. This is a loss function, so the optimal value is zero.

The closely related **Hamming score**, also called accuracy in the multi-label setting, is defined as the number of correct labels divided by the union of predicted and true labels, $\frac{|T\cap P|}{|T\cup P|}$

Precision, recall and F1 score: precision is $\frac{|T\cap P|}{|P|}$ recall is $\frac{|T\cap P|}{|T|}$ and is their harmonic mean.

Exact match: is the strictest metric, indicating the percentage of samples that have all their labels classified correctly.

3. LITERATURE REVIEW

Since from the last 15 years there are number of methods for multi-label classification designed in different areas such as categorization of text [2] [3], computer vision [4] [5], and bioinformatics [6] [7]. These works claimed that dependency exploitation among various labels is important for improving the performance of multi-label classification. Below different methods studied.

H. Blockeel et.al (2000)

In [8], author introduced predictive clustering trees in order to make multi-label classification or multi-target prediction. Author adopted the approach of clustering the basic top-down induction of decision trees method towards clustering. To this aim, it employs the principles of instance based learning. The resulting methodology is implemented in the TIC (Top down Induction of Clustering trees) system for first order clustering. The TIC system employs the first order logical decision tree representation of the inductive logic programming system Tilde. Author conducted several experiments in order to illustrate the type of tasks TIC is useful for. For this method PCT method has been referred in this paper.

G. Tsoumakas et.al. (2007)

In [9], the BR based approach introduced by author for multi-label classification problem. The study and evaluation of different BR methods with different datasets and compared in terms of precision, hamming score, recall etc.

G. Tsoumakas et.al. (2008)

In [10], author introduced a novel method for effective and computationally efficient multi-label classification in domains with large label sets *L*. The HOMER algorithm constructs Hierarchy of Multi-label classifiers, each one dealing with a much smaller set of labels compared to *L* and a more balanced example distribution. This leads to improved predictive performance along with linear training and logarithmic testing complexities with respect to |*L*|. Label distribution from parent to children nodes is achieved via a new balanced clustering algorithm, called balanced *k* means. HOMER followed the divide-and-conquer paradigm of algorithm design. The main idea was the transformation of a multi-label classification task with a large set of labels *L* into a tree-shaped hierarchy of simpler multi-label classification tasks, each one dealing with a small number *k* << |*L*| of labels.

W. Cheng et.al. (2010)

In [11], author aimed to provide a formal setting that allows for a more thorough analysis of multi-label classification in general and label dependence in particular. They presented approach to distinguish two types of label dependence, conditional and unconditional, and then they focused on the former. They proposed a probabilistic framework that suggests looking at the problem from the point of view of risk minimization and Bayes optimal prediction. Concretely, author analyzed three types of loss functions and, based on the results, raised the following conjecture: While considering conditional label dependence can indeed be useful for certain loss functions, there are others that are less likely to benefit. A second important contribution of this paper was a new method for multi-label classification, called probabilistic classifier chains (PCC).

J. Read et.al (2011)

In [12], classifier chains for Multi-label Classification proposed. They presented the advantages of BM-based methods and present their classifier chains method Classifier Chains (CC), which overcomes disadvantages of the basic binary method. Further they introduced an ensemble framework for classifier chains called Ensembles of Classifier Chains (ECC). The Classifier Chain model (CC) involves |L| binary classifiers as in BM. Classifiers are linked along a chain where each classifier deals with the binary relevance problem associated with label lj € L. The feature space of each link in the chain is extended with the 0/1 label associations of all previous links. Recall the notation for a training example (x, S), where S belongs L is represented by binary feature vector (l1, l2…ln) and x is an instance feature vector. Then in ECC, author trained *m* CC classifiers C1, C2…Cm. Each Ck is trained with: a random chain ordering (of L); and a random subset of D.

G. Madjarov et.al. (2012)

In [13], author proposed another approach for multi-label classification problem. They introduced a Two Stage Architecture (TSA) for efficient multi-label learning. They analyzed three implementations of this architecture such as Two Stage Voting Method (TSVM), the Two Stage Classifier Chain Method (TSCCM) and the Two Stage Pruned Classifier Chain Method (TSPCCM). Eight different real-world datasets were used to evaluate the performance of the proposed methods. The performance of their methods was compared with the performance of two algorithm adaptation methods (Multi-Label k-NN and Multi-Label C4.5) and five problem transformation methods (Binary Relevance, Classifier Chain, Calibrated Label Ranking with majority voting, the Quick Weighted method for pair-wise multi-label learning and the Label Power set method). Overall objective of TSA was to reduce the number of classifiers that are needed to be consulted in the prediction phase of the CLR algorithm and increase the predictive accuracy.

Limitations

Above discussed techniques suffered from either or all of below mentioned problems:

- BR methods do not support for exploiting the label dependencies among multiple objects.
- Some methods required prior knowledge for exploiting label dependencies.

Some methods leading to over-fitting the problems.

4. CURRENT SOLUTION

The above problems are recently solved by new method called ML-Forest [1]. The ML-FOREST proposed to build an ensemble classifier. In ML-FOREST, constructing a set of hierarchical trees that is able to automatically exploit the label correlation, and develop a label transfer mechanism which identifies the relevant labels hierarchically. ML-FOREST models the label dependency as a hierarchical scheme and performs the multi-label classification on this tree structure as a hierarchical decision process. As a result, ML-FOREST can have more discriminating ability than the first-order multi-label classification methods which only transform a multi-label problem into multiple separate and independent binary problems. In this section, document presenting the main algorithms designed for ML-Forest method.

Algorithm 1: ML-FOREST
Input: A training data set D, the number of trees K
Output: A forest of tree classifiers F

```
F = _
for i = 1 to K do
    prepare the training set Di = bootstrap (D)
    build tree classifier Ti = ML-TREE (D, none) // Calling Algorithm 2 for tree construction
    F = F [Ti]
end for
return F
```

Algorithm 2: ML-TREE
Input: A training data set D, and a relevant label vector b = none
Output: A hierarchical multi-label tree

```
(b, h, P) = SPLITTEST (D; b) // Algorithm 3 is called
if h 6= none ^ Acceptable (P) then
    for Di 2 P do
        treei=ML-TREE (Di, b)
    end for
    return node (h, b, [iftreeig)
else
    return leaf (h, b)
end if
```

Algorithm 2: SPLITTEST
Input: A training data set D, a relevant label vector bp from parent
Output: A classifier h, a new relevant label vector b, and a partition P for current node

```
compute p
compute b
(h, P) = (none, _)
h = build classifier on D for those labels which have not been identified according to b
if h 6=none then
    P= partition D using h
end if
return (b, h, P)
```

5. PRACTICAL ANALYSIS

The designing and implementation of ML-Forest method is performed using yeast dataset and its performance in terms of precision, recall and hamming loss is compared against other 4 methods. Table 1 is showing the comparative analysis for precision rate.

Table 1: Precision Rate Evaluation

METHODS	BR	CC	PCT	TSA	ML-FOREST
YEAST	65.31%	64.29%	59.78%	65.98%	66.72%

Table 2: Recall Rate Evaluation

METHODS	BR	CC	PCT	TSA	ML-FOREST
Yeast	54.3 %	55.43 %	52.39 %	57.44 %	56.43 %

Table 3: Hamming Loss Evaluation

Methods	BR	CC	PCT	TSA	ML-FOREST
Yeast	0.267	0.2563	0.2933	0.263	0.213

Above tables showing that, ML-Forest methods achieve the best performance for hamming loss and precision rate against other methods. This claims that method ML-Forest exploiting the label dependencies efficiently.

6. CONCLUSION AND FUTURE WORK

This paper aimed to present the study on multi-label classification problem and its different methods. In this paper, first introduce the problem of multi-classification, and then discussed the different performance metrics to measure the efficiency of multi-classification solutions. The review of different solutions for multi-label classification for various applications is presented in this paper. The current problems and current solution with their algorithm design and results is introduced in this paper. This paper is nothing but roadmap for future research in machine learning domain prepared by us.

7. REFERENCES

[1] Qingyao Wu, Mingkui Tan, Hengjie Song, "ML-FOREST: A Multi-label Tree Ensemble Method for Multi-Label Classification", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, MAY 2016.

[2] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers, "Statistical topic models for multi-label document classification," Machine learning, vol. 88, no. 1-2, pp. 157–208, 2012.

[3] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang and Y.-F. Li, "Multi-instance multi-label learning," Artificial Intelligence, vol. 176, no. 1, pp. 2291–2320, 2012.

[4] M. Liu, Y. Luo, D. Tao, C. Xu, and Y. Wen, "Low-rank multi-view learning in matrix completion for multi-label image classification," in Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[5] F. Sun, J. Tang, H. Li, G.-J. Qi, and T. S. Huang, "Multi-label image categorization with sparse factor representation," Image Processing, IEEE Transactions on, vol. 23, no. 3, pp. 1028–1037, 2014.

[6] X. Xiao, P. Wang, W.-Z. Lin, J.-H. Jia, and K.-C. Chou, "iamp-2l: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types," Analytical biochemistry, vol. 436, no. 2, pp. 168–177, 2013.

[7] K.-C. Chou, “Some remarks on predicting multi-label attributes in molecular biosystems,” Molecular Biosystems, vol. 9, no. 6, pp. 1092–1100, 2013.

[8] H. Blockeel, L. De Raedt, and J. Ramon, “Top-down induction of clustering trees,” arXiv preprint cs/0011032, 2000.

[9] G. Tsoumakas and I. Katakis, “Multi-label classification: An overview,” International Journal of Data Warehousing & Mining, vol. 3, no. 3, pp. 1–13, 2007.

[10] G. Tsoumakas, I. Katakis, and I. Vlahavas, “Effective and efficient multilabel classification in domains with large number of labels,” in Proc. ECML/PKDD’08 Workshop on Mining Multidimensional Data, 2008, pp. 30–44.

[11] W. Cheng, E. H¨ ullermeier, and K. J. Dembczynski, “Bayes optimal multilabel classification via probabilistic classifier chains,” in Proceedings of ICML’10 the 27th International Conference on Machine Learning, 2010, pp. 279–286.

[12] J. Read, B. Pfahringer, G. Holmes, and E. Frank, “Classifier chains for multi-label classification,” Machine learning, vol. 85, no. 3, pp. 333–359, 2011.

[13] G. Madjarov, D. Gjorgjevikj, and S. Dˇzeroski, “Two stage architecture for multi-label learning,” Pattern Recognition, vol. 45, no. 3, pp. 1019–1034, 2012.