ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 12, July 2017

Classification of Yeast Dataset by Transductive Multi Label Learning via Label set Propagation

Girisha G S [1], Raghavendra C K [2], Deepak Hanumanthaiah[3]

Associate Professor, Department of Information Science & Engineering, B N M I T, Bangalore, India[1]

Assistant Professor, Department of Computer Science & Engineering, B N M I T, Bangalore, India[2]

Department of Information Science and Engineering, B N M I T, Bangalore, India[3]

ABSTRACT: The problem of multilabel classification has attracted great interest in the last decade, where each instance can be assigned with a set of multiple class labels simultaneously. It has a wide variety of real-world applications, like automatic image annotations and gene function analysis. In these applications, the labelling of multilabeled data is extremely expensive and time consuming and there are abundant unlabeled data available. In this paper, we propose an algorithm called Transductive Multilabel Classification (TRAM), to effectively assign a set of multiple labels to each instance. We estimate the label sets of the unlabeled instances effectively by utilizing the information from both labeled and unlabeled data. The results have been compared with the supervised settings algorithm.

KEYWORDS: Data mining, Machine learning, Multilabel learning, Transductive learning, unlabeled data.

I. INTRODUCTION

Traditional single-label classification is concerned with learning from a set of examples that are associated with a single label l from a set of disjoint labels L, |L| > 1. If |L| = 2, then the learning problem is called a binary classification problem (or filtering in the case of textual and web data), while if |L| > 2, then it is called a multi-class classification problem.

In multi-label classification, the examples are associated with a set of labels $Y \subseteq L$. In the past, multi-label classification was mainly motivated by the tasks of text categorization and medical diagnosis. Text documents usually belong to more than one conceptual class example, one news article can cover multiple aspects of an event, thus being assigned with a set of multiple topics, such as economics and politics [1]. An effective classification model for these real-world data should be able to adopt the multiple labels of each training example and predict a label set, instead of one single label, for each testing example.

There are many different approaches available for solving multilabel problem that uses supervised settings which require a sufficiently large amount of labeled examples in order to train an accurate model. Creating a sufficiently large training data set where each example is labeled with a set of multiple labels within the candidate classes is usually infeasible in practice because the labelling of multilabel data in many real world applications is extremely expensive and time consuming as it requires time, efforts, and excessive resources to manually assign each example with all its labels and hence only a limited amount of labeled examples are available. Moreover, there are often large amounts of unlabeled data available and it is desired that these unlabeled data can be effectively utilized together with the limited amount of labeled data to improve the multilabel classification performances. Transductive learning [2] is a type of approach which makes use of unlabeled data in the classification process. It assumes that the testing data are available, and the goal is to achieve better performances on these testing data by exploiting the unlabeled testing data in the classification process. It has been shown useful in many single-label classification tasks [2].

Transductive multilabel classification problem [3] corresponds to predicting the label sets of a group of unlabeled instances simultaneously by utilizing the information from both labeled and unlabeled data. Transductive learning is

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 12, July 2017

particularly challenging in multilabel settings. The reason is that, in the single-label case, conventional transductive learning methods can be applied to propagate class labels among the unlabeled data and predict each unlabeled instance with the class label which has the highest confidence. But in multilabel cases, each instance contains multiple label concepts and the transductive classification task corresponds to finding a label set for each unlabeled instance within the space of label sets, i.e., the power set of all labels. The number of possible label sets is exponential to the number of candidate labels, which is extremely large even with a small number of candidate labels.

II. RELATED WORK

Multilabel Classification

Several methods have been proposed for multi label classification. These methods are derived from traditional learning techniques. One famous approach proposed by Schapire and Singer, BOOSTEXTER [4], is extended from the popular ensemble learning method ADABOOST [5]. In the training phase, BOOSTEXTER maintains a set of weights over both training examples and their labels, which will be incrementally enlarged if examples or labels are hard to be predicted correctly. Elisseeff and Weston [6] presented a kernel method RANK-SVM for multilabel classification, by minimizing a loss function named ranking loss. Experimental results on the Yeast gene functional classification problem demonstrate its effectiveness. Zhang and Zhou [7] extended the lazy learning algorithm, kNN, to a multilabel version, Ml-KNN. It employs label prior probabilities gained from each example's k nearest neighborsand use Maximum a Posteriori (MAP) principle to determine labels. Unlike the previous works that only consider the correlations among different categories, Liu et al. [8] present a semi-supervised multilabel classification method to exploit unlabeled data as well as category correlations. This approach is based on constrained non-negative matrix factorization. Generally, in comparison with supervised methods, semi-supervised methods can efficiently make use of the information provided by unlabeled instances.

Transductive Learning

In this paper, we focus on transductive learning. Transductive learning was proposed by Vapnik [2] in the 1990s where all unlabeled points belong to the testing set. Many transductive learning approaches have been proposed. One famous approach is Transductive SVMs, introduced by [2] and applied to text classification. They exploit the structure in both training and testing data for better positioning the maximum margin hyperplane. Another type of approaches are graph-based methods, which define a graph with the nodes representing both labeled and unlabeled instances, and edges reflect the similarity of instances. Graph-based approaches usually assume label smoothness over the graph.

III. METHODOLOGY

Consider a data set having both labeled and unlabeled instances. The data set is divided into training data which is training sample matrix, testing data which is testing sample matrix, training label and testing label which are matrices where each element in the matrix has a value of +1 or -1. +1 indicates that the corresponding label is assigned to the instance and -1 indicates that the label is not assigned to the instance.

IV. ALGORITHM

Input: Training data matrix, Training label matrix, Testing data matrix

Step 1: Construct the kNN graph.We build a kNN graph G(V,E) for both labeled and unlabeled instances where V is the set of instances and E is the corresponding set of edges between the vertices. An edge $e_i$between vertices v_i and v_j denotes that v_i is among the k nearest neighbors of v_j. Initially we construct a kd-tree to perform the kNN search for both the labeled and unlabeled instances as it helps to reduce the computational cost of the kNN search. As the number of dimensions increases while using kd-tree it will ultimately degenerate to a linear search therefore a multilabel dimensionality reduction approach called MDDM is implemented before constructing the kd-tree which reduces the number of dimensions required to construct the kNN graphs.

Step 2: Find the similarities among the k-neighboring instances by constructinga n x n sparse matrix which contains the ratio of Euclidian distance and the average distance between the instances, and normalize these values to 1.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 12, July 2017

Step 3: For finding optimal alpha value, first we have to find out the smoothness. Smoothness is, label concept composition within the labelset will be similar for similar instance, i.e, If an unlabeled instance is similar to a labeled instance then the alpha values should also be similar.
Step 4: Compute the sorted label list for each unlabeled instances by storing the optimal alpha values in descending order.

V. DATASETS

We experiment with yeast dataset which contains 2417 instances of genes and 14 class labels as shown in Fig. 1.

name	domain	instances	nominal	numeric
yeast	biology	2417	0	103

labels	cardinality	density	distinct
14	4.237	0.303	198

Fig. 1. Yeast dataset

VI. RESULTS

The result we get for yeast dataset is as shown in Fig. 2.

Instance 1	Instance 2	Instance 3	Instance 4	Instance 5
0.0748	0.0511	0.0544	0.0999	0.0481
0.0701	0.1291	0.0947	0.0358	0.0803
0.0813	0.1845	0.1310	0.1919	0.1260
0.0829	0.1258	0.0604	0.2425	0.0802
0.0542	0.0492	0.0974	0.0591	0.0492
0.0796	0.0290	0.1031	0.0048	0.0520
0.0575	0.0181	0.0292	0.0023	0.0124
0.0584	0.0187	0.0737	0.0122	0.0314
0.0037	0.0020	0.0532	0.0101	0.0245
0.0014	0.0363	0.0313	0.0074	0.0619
0.0055	0.0388	0.0332	0.0075	0.0865
0.2158	0.1585	0.1188	0.1625	0.1737
0.2148	0.1585	0.1187	0.1624	0.1735
3.8374e-05	2.3268e-04	8.6449e-04	0.0014	2.6675e-04

Here, each row represents the composition of the labelset, and each column represents the instances. Since the yeast dataset is partitioned into 1500 train data and 917 test data among 2417 yeast dataset, the answer result we get is for 917 test data which contains 14 class labels. The summation of the composition of 14 labelset will makes up to exactly 1 that is 100%

REFERENCES

[1] A. McCallum, "Multi-Label Text Classification with a Mixture Model Trained by EM," Proc. Working Notes Am. Assoc. Artificial Intelligence Workshop Text Learning (AAAI '99), 1999.
[2] V.N. Vapnik, Statistical Learning Theory. Wiley, 1998.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 12, July 2017

[3] Xiangnan Kong, Michael K. Ng, and Zhi-Hua Zhou, Fellow TransductiveMultilabel Learning via Label Set Propagation" ieeetransactions on knowledge and data engineering, vol. 25, no. 3, March 2013

[4] R.E. Schapire and Y. Singer, "BoosTexter: A Boosting-Based System for Text Categorization," Machine Learning, vol. 39, nos. 2/3, pp. 135-168, 2000.

[5] Y. Freund and R.E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,"J. Computer and System Sciences, vol. 55, no. 1, pp. 119-139, 1997.

[6] A. Elisseeff and J. Weston, "A Kernel Method for Multi Labelled Classification," Advances in Neural Information Processing Systems14, T.G. Dietterich, S. Becker and Z. Ghahramani, eds., pp. 681- 687, MIT Press, 2002.

[7] M.-L. Zhang and Z.-H. Zhou, "ML-kNN: A Lazy Learning Approach to Multi-Label Learning," Pattern Recognition, Vol. 40, no. 7, pp. 2038-2048, 2007.

[8] Y. Liu, R. Jin, and L. Yang, "Semi-Supervised Multi-Label Learning by Constrained Non-Negative Matrix Factorization," Proc. 21st Nat'l Conf. Artificial Intelligence, pp. 421-426, 2006.