Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 54 (2015) 405 – 411

Procedia Computer Science

Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)

Multi-Label Learning with Class-Based Features Using Extended Centroid-Based Classification Technique (CCBF)

P. R. Suganya Devi*, R. Baskaran and S. Abirami

Department of Computer Science and Engineering, College of Engineering Guindy, Anna University, Chennai

Abstract

Real world applications, such as news feeds categorization deal with multi-label classification problem, where the objects are associated with multiple class labels and each object is represented by a single instance (feature vector). In this paper, a new algorithm adaptation method called centroid-based multi-label classification using class-based features (CCBF) algorithm has been proposed to tackle the multi-label classification problem. It includes class-based feature vectors generation and local label correlations exploitation. In the testing stage, centroid-based classification algorithm is extended for multi-label classification problem. Experiments on reuters multi-label dataset with 103 labels demonstrate the performance and efficiency of CCBF algorithm and the result is compared with those obtained using other multi-label classification algorithms. The CCBF algorithm obtains competitive F measures with respect to the most accurate algorithms.

Peer-review under responsibility of organizing committee of the Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)

Keywords: Class-based features; Cure clustering algorithm; Document classification; Label correlations; Multi-label learning.

1. Introduction

Multi-label learning deals with objects having multiple class labels and each object is represented by one single instance. The task is to learn a model which can predict a set of possible labels for an unseen object. For example, given class labels Asia, N. America, S. America, Europe and Australia, a news article about U.S troops in Bosnia may be labeled with both *N*. America and Europe classes. Multi-label learning has been applied to a variety of domains, such as text classification, image annotation, video annotation, social network and music categorization into emotions, bioinformatics, etc.

A common approach to multi-label classification is problem transformation, in which a multi-label problem is transformed into one or more single-label problems. The alternative to problem transformation is algorithm adaptation which modifies an existing single-label classification algorithm for multi-label classification. The common strategy adopted by existing approaches is that all the class labels are discriminated based on identical feature representation of the object. Also, existing approaches consider global label correlations only. However, using identical feature representation is inadequate to discriminate different class labels, as different class labels in the label space may carry

*Corresponding author.
E-mail address: suganya.rvks5@gmail.com

Peer-review under responsibility of organizing committee of the Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)
doi:10.1016/j.procs.2015.06.047

specific characteristics of their own. Also, different examples may share different label correlations. The main goal of this work is to improve the classification accuracy by considering class-based features and local label correlations.

The rest of the paper is organized as follows. Section 2 presents existing work on multi-label learning and further relevant literature. Section 3 presents the proposed framework for multi-label learning. The experimental results are discussed in section 4. Section 5 is a brief conclusion of this work.

2. Related Work

In the past decades, many well-established methods have been proposed to solve multi-label learning problems in various domains. All these methods can be divided into two categories: Problem Transformation Methods (fitting data to algorithm) and Algorithm Adaptation Methods (fitting algorithm to data).

Problem Transformation Methods convert the multi-label problem into a set of binary classification problems. Binary relevance (BR) method[10] takes each class label as an independent binary problem. Dependent binary relevance (DBR) learning[8] combines properties of both, chaining and stacking. The limitation of these binary relevance methods is that the computational complexity is worst. EPS method[9] is concentrated on the concept of treating sets of labels as single labels. It achieves better performance, and trains much faster than other multi-label methods. Label powerset (LP) method[2] considers label correlation by combining the unique set of class labels. But it is usually unfeasible for practical application, because it generates a huge number of class labels.

Algorithm Adaptation Methods modify traditional single label learning algorithms for multi-label learning, which can handle multi-label data directly. Ricardo[12] proposed a method composed of an online procedure. Documents are classified using statistics computed from labeled instances. The limitation is that context is not considered in the feature vector. ML-kNN[7] is Multi-Label k Nearest Neighbor which is extended from the standard kNN algorithm. Tsoumakas[1] implements BRkNN algorithm and compares different multi-label classification algorithms.

Tan[13] proposed a novel batch-updated approach which takes advantage of errors to update the model by batch. But it can be applied only to classic train/test problems. Yu[15] proposed two novel multi-label classification algorithms, called multi-label classification using rough sets (MLRS) and MLRS using local correlation (MLRS-LC). They achieve promising performance when compared with other multi-label learning algorithms. Ren[11] introduced class-indexing-based term-weighting approach, in which the inverse class frequency (ICF) is incorporated to generate more informative terms. Vale[14] proposed a class-based feature selection method. This method chooses the attributes that are important for a specific class. Qian[16] proposed CURE-NS (CURE with new shrinking scheme), which uses CURE clustering algorithm and uses the difference of density values of the representative points to determine the direction and distance of shrinking.

3. CCBF Multi-Label Learning

This paper proposes a strategy to learn from multi-label data, where class-based features and local label correlations are exploited. Finally, centroid-based classifier is extended for multi-label classification problem. Class-based features are considered to benefit the discrimination of different class labels. They are generated by performing clustering on positive examples and on negative examples for each class label. Cluster centroids are building blocks for generating modified class-based feature vectors are computed. For performing clustering analysis, CURE clustering algorithm is used. Then class-based feature vectors, exclusively for multi-label classification, are generated for each data point. Local label correlations are considered to improve the performance. For this, possible label subsets are recognized and training documents are grouped to the corresponding label subset. Using the modified class-based feature vectors and labels of each data point, label space is partitioned. Then, prototypes for each partition will be computed. A map function is used to automatically accommodate the incoming document in a region of the partition using Euclidean metric. Then, centroid-based classifier for multi-label classification is used to output a set of labels, according to the region. The overall architecture is shown in the following Fig. 1.

Processes involved:

- Cure clustering analysis.
- Class-based feature vectors construction.