A Meta-Top-Down Method for Large-Scale Hierarchical Classification

Technofist,
YES Complex, 19/3&4, 2nd Floor, Dinnur Main Road, R.T.Nagar,Bangalore-560032
Ph:080-40969981, Website:www.technofist.com. E-mail:technofist.projects@gmail.com

ABSTRACT

Recent large-scale hierarchical classification tasks typically have tens of thousands of classes on which the most widely used approach to multiclass classification one-versus-rest—becomes intractable due to computational complexity. The top-down methods are usually adopted instead, but they are less accurate because of the so-called error-propagation problem in their classifying phase. To address this problem, this paper proposes a meta-top-down method that employs metaclassification to enhance the normal top-down classifying procedure. The proposed method is first analyzed theoretically on complexity and accuracy, and then applied to five real-world large-scale data sets. The experimental results indicate that the classification accuracy is largely improved, while then increased time costs are smaller than most of the existing approaches.

INTRODUCTION

MULTICLASS classification—classifying samples into multiple predefined classes—is a fundamental task in both machine learning and data mining domains .For example, each document in the Reuters-215781 corpus is assigned one or more labels from 120 predefined classes such as business, sports, and military.

The ensemble method of one-versus-rest is the most widely adopted solution for multiclass classification .First a binary-class classifier, named base classifier, is trained for each class ci to predict whether an input sample x belongs to this class; the thresholding strategies are employed to decide the predicted labels according to the confidence scores of the base classifiers. Two commonly used thresholding strategies are score-cut (S-cut) that accepts the classes whose scores are larger than a predefined threshold, and rank-cut (R-cut) that accepts the classes whose scores are among the top-r (r is a predefined integer) directories of the Open Directory Project or Yahoo! The top down methods which organize base classifiers hierarchically are widely adopted for large-scale hierarchical .They classify a sample x by filtering it down a tree of base classifiers from the root node f1. For each parent node where this sample arrives, those child nodes whose confidence scores (produced by the base classifiers) pass the thresholding strategies will carry it on. The bottom leaf nodes where the sample terminates are the predicted labels The top-down method employing the S-cut strategy, named S-cut topdown (ScutTD), is usually used in multilabeled classification ,m and the one employing the R-cut strategy (r ¼ 1), named R-cut top-down (RcutTD), is usually used in single labeled classification. The computational complexity of the top-down methods is the logarithm of the number of classes [8],], while that of the one-versus-rest method is linear to the number of classes. As an evidence, in a classification experiment on 492,617 training documents, 275,364 test documents, and 132,199 categories of Yahoo!, the former costs only 2.1 hours for training and 0.12 hours for classifying, while the latter costs 310 hours for training and 54 hours for classifying .