

AN ALGORITHM FOR CO-TRAINING IN MEDICAL IMAGE RETRIEVAL

HUI LIU¹, CAIMING ZHANG², XIUHUA JI¹ AND YUNFENG ZHANG¹

¹Department of Computer Science & Technology
Shandong Economic University
ErHuan Road 7366#, Jinan 250014, P. R. China
liuh_lh@126.com; jane@sdu.edu.cn; qyhjs@163.com

²Department of Computer Science & Technology
Shandong University
ShunHua Road, Jinan 250101, P. R. China
czhang@sdu.edu.cn

Received July 2008; revised December 2008

ABSTRACT. *In CBIR study, different from traditional supervised learning which only makes use of labeled data, semi-supervised learning makes use of both labeled and unlabeled data. For the importance of unlabeled data, it is necessary to define conditions to utilize the unlabeled examples enough. A new SVM-based co-training algorithm that defining two learners is presented in this paper, where both the learners are re-trained after every relevance feedback round, and then each of them gives every image in a rank. In relevance feedback, the user may label several images according to whether they are relevant or not to a query, and images with the top result-size ranks are returned. Finally, to test the performance of our methodology, a database consisting of 500 chest CT images were used, PR-graphs and BEP-graphs all illustrate that the performance of our method is better.*

Keywords: Co-training, Unlabeled data, SVM, PR-graphs, BEP-graphs

1. Introduction. In recent years, some researches have introduced semi-supervised learning in Content Based Image Retrieval (CBIR). Different from traditional supervised learning which only makes use of labeled data, semi-supervised learning makes use of both labeled and unlabeled data. Generally, it's difficult to collect vast amounts of labeled data and easy for unlabeled data, moreover, the unlabeled data can be used to better describe the intrinsic geometrical structure of the data space, and hence improve the image retrieval performance. For instance in computer aided medical image analysis, we can obtain a great deal of medical images as training dataset, but it's very unpractical for medicinal experts to distinguish all focus from these images. In fact, in many real-world applications labeled training examples are fairly expensive to obtain while unlabeled examples are abundantly available, especially in some online application problems.

So, learning from unlabeled examples has become a hot topic during the past few years, there are two main machine learning paradigms for this purpose: semi-supervised learning and active learning [1]. In semi-supervised learning there is a kind of co-training algorithm, which impliedly utilizes the clustering hypothesis or manifold hypothesis, and two or more learners. During the course of training, these learners select some unlabeled samples with high confidence to label each other, for updating the model continually.

In this paper, we introduce a new co-training algorithm for image retrieval problem, which test in medical image analysis system. Our algorithm is intrinsically based on Support Vector Machines (SVM), and researches an applied method that makes use of a