J. Vis. Commun. Image R. 55 (2018) 720–728

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Content-based image retrieval model based on cost sensitive learning ☆

Cong Jin [a,*], Shu-Wei Jin [b]

[a] School of Computer, Central China Normal University, Wuhan 430079, PR China
[b] Département de Physique, École Normale Supérieure, 24, rue Lhomond 75231, Paris Cedex 5, France

ARTICLE INFO

Article history:
Received 10 December 2017
Revised 9 July 2018
Accepted 11 August 2018
Available online 13 August 2018

Keywords:
Content-based image retrieval
Distance metric learning
Cost sensitive learning
Classification performance
Misclassification cost
Class imbalance

ABSTRACT

How to retrieve the desired images quickly and accurately from the large scale image database has become a hot topic in the field of multimedia research. Many content-based image retrieval (CBIR) technologies already exist, but they are not always satisfactory. In many applications, the CBIR model based on machine learning relies heavily on the distance metric between samples. Although the traditional distance metric methods are simple and convenient, it is not always appropriate for CBIR tasks. In this paper, a novel distance metric learning (DML) method based on cost sensitive learning (CSL) is studied, and then it is used in a large margin distribution learning machine (LDM) to replace the traditional kernel functions. The improved LDM also takes into account CSL, and which is called CS-DLDM. Finally, CS-DLDM model is applied to CBIR tasks for implementation classification. We compare the proposed CS-DLDM model with other classifiers based on CSL. The experimental results show that the proposed CS-DLDM model not only has satisfactory classification performance but also the lowest misclassification cost, can effectively avoid the class imbalance of sample.

1. Introduction

Recently, massive attempts have been devoted to designing descriptive and discriminative visual features for accurate image representation in image retrieval [1,2]. However, the metric space, where the pairwise features are matched for similarity measure, is usually underplayed and thus leads to the suboptimal retrieval accuracy [3,4]. Although some distance metrics enable characterizing the feature distances in some cases, they are defined empirically without sufficiently capturing the distribution statistics in the feature space. Therefore, they fail to provide desirable retrieval performance in a variety of learning tasks.

Many machine learning techniques were applied in content-based image retrieval (CBIR). The performance of machine learning techniques is often heavily dependent on the distance metrics between samples, however some existing distance metrics, including Euclidean distance, Mahalanobis distance and Hausdorff distance and so on, are sometimes difficult to accurately measure the similarity between samples, thus a large number of other distance metrics are used in machine learning for various applications. Usually, these existing distance metrics lack universality, which perform well in some applications and poorly in other applications [4]. How to more effectively measure the distance between samples has become a key research problem in CBIR.

Over the past decades, researchers have spent a lot of effort designing different distance metrics [5,6], until now finding an appropriate distance metric has always been an open challenge for CBIR tasks. In recent years, one solution of addressing this challenge is distance metric learning (DML) [7,8] by using training samples and their label information. The goal of DML is to obtain better distance metric, which can describe the relationship between samples by learning known training samples [9]. Under this metric, the distance between the same class samples is smaller, and the distance between different class samples is larger. However, it is popular for the class imbalance of samples [10,11] in CBIR tasks, *i.e.*, most of training samples are not relevant to query, belonging to the majority class, only a few are relevant to query, belonging to the minority class. The recognition of the minority class is often important and even more important because the goal of CBIR is correctly identifying those images relevant to query, and therefore it is not appropriate to use the existing DML algorithms directly for CBIR.

In fact, cost sensitive learning (CSL) and class imbalance of samples are closely related [12]. For class imbalance classification issue, the misclassification cost of a minority class sample is usually more expensive than a majority class sample [13,14]. CSL not only examines the different type misclassification costs, but also studies how to minimize the total cost of misclassifications. Classifier based on CSL can reduce the misclassification cost caused

☆ This paper has been recommended for acceptance by Dr. Dacheng Tao.

* Corresponding author.
E-mail address: jinc26@aliyun.com (C. Jin).

https://doi.org/10.1016/j.jvcir.2018.08.009
1047-3203/

by class imbalance of samples, and therefore existing CSL algorithms were mainly designed for classification. We notice that, although there were many studies on DML, the influence of class imbalance of samples on DML is rarely considered. In other words, the existing CSL algorithms were seldom designed for DML.

In this paper, a novel DML based on CSL is studied. Specifically, the weights of the different classes are respectively determined by the sizes of the classes in order to overcome the influence of class imbalance of samples. Then, the CSL is also used to large margin distribution machine (LDM) for CBIR tasks. Usually, the misclassification cost factors are given by experts, which ignore the influence of the size of the class [13,15]. And our method is determined by the ratio of number of samples in the majority and minority classes, which is very intuitive and simple.

Moreover, many kernel functions play an important role of the distance between samples [16], thus the proposed cost sensitivity DML is applied to LDM in order to replace the general kernel functions. Our cost sensitive CBIR model is called CS-DLDM. Where, the meaning of the first letter "D" is a distance obtained by the proposed cost sensitive DML. The main contributions of this work are as follows:

(1) In the proposed cost sensitive DML, the class imbalance of sample is adequately considered. The samples of the majority and minority classes are arranged respectively with different weights so that the samples of the majority and minority classes have different importance, which can significantly reduce the influence of class imbalance.
(2) In existing cost sensitive classifiers [17,18], only two metrics of the minority class were often discussed, one is the margin mean and the other is the misclassification penalty cost. In this paper, apart from the above two metrics, the margin variance is also considered, which makes the results of the statistics more comprehensive and sufficient.
(3) Usually, the kernel function can explore the intrinsic relationship between samples, and however the results are relatively rough. In the proposed CS-DLDM, the kernel functions of LDM are replaced by the proposed cost sensitive DML, which not only can fully use the characteristics between samples, but also significantly reduce the misclassification cost of CS-DLDM model.

The remainder of this paper is organized as follows. The related work of DML and CSL are introduced in Section 2. The proposed CS-DLDM is presented in Section 3 including the proposed cost sensitive DML, the proposed CS-DLDM model and its optimization solution method. Section 4 provides experimental setup. The experiment results and analysis are in Section 5. Finally, conclusions are given in Section 6.

2. Related work

2.1. DML

For CBIR based on machine learning, the similarity of the images depends on the distance metric between their feature vectors, and therefore distance metric is very important for machine learning and CBIR [19–21]. How to get an appropriate distance metric through learning has attracted more and more researchers' attention. DML uses the information provided by the labels and features of training samples to automatically learn from the image dataset and get the distance metric for satisfying the specific requirements.

Many DML algorithms have been proposed [22–27] mainly including four categories [4]. The first category is supervised DML, which contains supervised global DML, local adaptive supervised DML, neighborhood component analysis (NCA) [26], relevant components analysis (RCA) [22] and so on. The second category is unsupervised DML, contains linear (e.g., principal component analysis (PCA) [28]), nonlinear embedding methods (e.g., locally linear embedding (LLE) [23]) and so on. The third category is maximum margin DML, includes the large margin nearest neighbor (LMNN) [11], semi-definite programming [24] and so on. The fourth category is kernel DML, includes kernel alignment (KA) [25], information-theoretic metric learning (ITML) [29] and so on.

Although many DML algorithms have been proposed, the class imbalance problem of the samples has not been discussed. In fact, the performance of DML algorithms is also affected by class imbalance of samples. The DML algorithm usually optimizes a distance metric loss function [11,23]. When the sample classes are unbalanced, the smaller sample number of the minority class, the greater corresponding loss, which lead to the performance of existing DML algorithms will tend to the majority class and ignore the minority class. However, the minority class samples are more important in the DML process. Therefore, in this paper, different from existing DML algorithms, we apply CSL to DML so that the samples of different classes have different weights according to the sizes of different classes, *i.e.*, samples of different classes have different importance for DML.

2.2. CSL

Costs are very important for pattern classification and it has been widely concerned [15]. CSL is a learning method in data mining that takes into account misclassification costs and possibly other types of costs. The goal of CSL is to accurately classify samples into a known class with a minimum total misclassification cost. The key difference between CSL and cost-insensitive learning is that CSL handles different misclassifications differently [30–32], and assigns different levels of misclassification penalty to each class. Currently, CSL has been applied to intrusion detection [32], software defect prediction [33], remote sensing and so on [34].

In many existing CSL models, the misclassification cost is usually either given by the expert in advance or given a constant factor as a weight for each type of cost. Currently, there are roughly two different costs [13], one is based on the category and the other is based on the sample. Since the goal of DML is to reflect the relationship between samples, in this paper, the cost based on samples is applied to the proposed cost sensitive DML. Similarly, because the goal of the classifier is to achieve classification of the sample set, the cost based on the category is used in the proposed CS-DLDM. Different from the existing CSL methods, in this paper, the misclassification cost of the proposed CS-DLDM factors are calculated according to the ratio of number of samples in the majority and minority classes.

3. Our approach

3.1. Block diagram of the proposed approach

In this paper, we first introduce the block diagram of the proposed cost sensitive CBIR model and its implementation algorithm, and then introduce its details in Sections 3.2 and 3.3. Fig. 1 displays main implement processes of the proposed cost sensitive CBIR model.

In the proposed cost sensitive CBIR model, CS-DLDM classifies the test image database into two subsets, namely one image subset relevant to query and another image subset not relevant to the query, and maximizes the margin between these two image subsets. Obviously, the image subset relevant to query belongs to