Knowledge-Based Systems 24 (2011) 740–748

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A novel virtual sample generation method based on Gaussian distribution

Jing Yang [a,*], Xu Yu [a], Zhi-Qiang Xie [a,b], Jian-Pei Zhang [a]

[a] College of Computer Science and Technology, Harbin Engineering University, Harbin, China
[b] College of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China

ARTICLE INFO

Article history:
Received 15 January 2010
Received in revised form 20 December 2010
Accepted 25 December 2010
Available online 31 December 2010

Keywords:
Virtual sample
Regularization theory
Cost-sensitive learning
Gaussian distribution
Prior knowledge

ABSTRACT

Traditional machine learning algorithms are not with satisfying generalization ability on noisy, imbalanced, and small sample training set. In this work, a novel virtual sample generation (VSG) method based on Gaussian distribution is proposed. Firstly, the method determines the mean and the standard error of Gaussian distribution. Then, virtual samples can be generated by such Gaussian distribution. Finally, a new training set is constructed by adding the virtual samples to the original training set. This work has shown that training on the new training set is equivalent to a form of regularization regarding small sample problems, or cost-sensitive learning regarding imbalanced sample problems. Experiments show that given a suitable number of virtual sample replicates, the generalization ability of the classifiers on the new training sets can be better than that on the original training sets.

1. Introduction

Classification is one of the most active fields of data mining. For many years, researchers in areas including machine learning, pattern recognition, and statistics are contributing to this field, and have proposed many classification methods, such as neural network [2,7,25], support vector machine (SVM) [17,19,20], and decision tree [27]. Practice shows that these technologies are all with good generalization ability if the training samples are sufficient. But the classification problem is far from being solved, because traditional classification technologies are not able to process the noisy, imbalanced, and small sample data set effectively.

Since noise is inevitable, how to learn and improve the performance of the classifiers under noise is an important problem. If the training set is large enough, the influence caused by noise is little, and the current learning algorithms can find a good classification rule to avoid over-fitting the data. However, in many real-world problems, not enough samples can be used, and moreover the sample set is always imbalanced, so over-fitting and generalization problem arise easily. Thus it is important and urgent to improve the learning ability of classifiers on a noisy, imbalanced, and small sample data set.

The rest of this paper is organized as follows. Section 2 introduces the research status of learning algorithms on noisy, imbalanced, and small sample data set. Section 3 presents a novel virtual sample generation (VSG) method based on Gaussian distribution. Section 4 explores the reasons why it can work. Section 5 reports on experiments. Section 6 summaries the main contribution of this paper and discusses the issues related to the proposed method.

2. Research status

It is a general perception that more data usually provide more information to a training system and can make the classifiers reach higher learning accuracy. However, researchers have tried to find effective ways to acquire knowledge from noisy, imbalanced, and small sample data set when a large data set is unavailable. Some of the related studies are introduced in the following.

2.1. Previous VSG methods

The concept of virtual samples was first introduced by Poggio and Vetter [14], and has been applied in many fields. Usually, virtual sample can be seen as additional training samples created from the current set of examples by utilizing specific knowledge about the task at hand. The idea of virtual samples is a possible way of incorporating prior information in classification learning problems.

From the introduction of virtual sample, many VSG methods have been proposed in different machine learning research fields. Broadly speaking, VSG methods can be classified into two categories by their generation thought. One is to generate virtual samples by extracting the nontrivial prior knowledge hidden in the

* Corresponding author. Address: M-2-1503, Culture Home, 258 NanTong Street, NanGang District, Harbin City, 150001 Heilongjiang Province, China. Tel.: +86 13624508906; fax: +86 0451 82519602.
E-mail address: yangjing@hrbeu.edu.cn (J. Yang).

0950-7051/$ - see front matter © 2011 Published by Elsevier B.V.
doi:10.1016/j.knosys.2010.12.010

question being solved. The other is to generate virtual samples by the idea of perturbing the original samples. Detailed descriptions are given in the following.

2.1.1. Generating virtual samples by extracting the nontrivial prior knowledge

Examples of this category first came from Poggio and Vetter. They created virtual samples by the application of prior knowledge to improve recognition ability in the field of pattern recognition. The method is, given a 3D view of an object, to create new images from any other angles through mathematical transformations. The new images generated are called virtual samples. In most real-world pattern recognition fields, extracting the prior knowledge and creating virtual samples are highly nontrivial. Wen et al. [23] proposed another VSG method of this category, which generated virtual samples using prototype faces. Besides the application in image recognition, VSG methods of this category have also been applied in many other fields, such as handwritten number recognition [15], text recognition [16,22], and noise source recognition [24].

2.1.2. Generating virtual samples by the idea of perturbing the original samples

Several noise replication methods can be considered as this category. Among them, Lee [10] proposed a method, which generated virtual samples by adding small normal noise to the original samples. But he did not give an approach to determine the parameters of Gaussian distribution. Moreover, he did not give any theoretical analysis to prove the effectiveness of the method. Li and Fang [11] proposed a non-linear VSG method, which combined a unique group discovery technique with a VSG method. Wang and Yang [21] proposed a perturbance-based VSG method, which added a small constant to every dimension of the p-dimensional training sample. Thus every training sample can generate p virtual samples. Zhang and Chen [26] introduced another VSG method, which firstly divided the training samples of the rare class into p groups by k-nearest-neighbor algorithm, then generated virtual samples by averaging every two samples of each group, and finally kept the labels unchanged.

As is mentioned above, the first category is to generate virtual samples by extracting the nontrivial prior knowledge, so the rationality can be assured, but the adaptability is very low. The second category is to generate virtual samples by perturbing, so the adaptability can be assured, but the rationality is very low. Thus the previous VSG method is either with a low rationality or with a low adaptability. Detailed definitions on rationality and adaptability are given in Section 3.

2.2. The addition of noise to the input data

The addition of noise to the input data can lead to improvements in generalization performance. Bishop [3] and An [1] have proven that training with noise is equivalent to Tikhonov regularization [18] if the standard deviation of the noise is little. Bishop [3] also found that the coefficient of the regularization was related with the standard deviation of the noise. This method can improve the generalization ability of the learning methods to some degree, but it cannot expand the sample set effectively. Thus the generalization ability improved by this method is limited.

2.3. Regularization theory

Regularization theory is another approach to solve over-fitting and generalization problem on small sample problems. One of the central issues in classification is to determine the optimal degree of complexity for the model. A model which is too limited will not capture enough of the structure in the data, while one which is too complex will model the noise on the data (the phenomenon of over-fitting). In either case the performance on new data, that is the ability of the classification to generalize, will be poor [3].

One technique that is often used to control the over-fitting phenomenon in such cases is that of regularization, which involves adding a penalty term to the error function. The technique of regularization makes use of a relatively flexible model, and then controls the variance by modifying the error function by the addition of a penalty term $\Omega(y)$. Thus the total error function becomes

$$E_t = E + \lambda\Omega(y) \tag{1}$$

where λ is a positive number that is usually called the regularization coefficient and $\Omega(y)$ is a cost function that constrains the space of possible solutions according to some form of prior knowledge. In effect λ now controls the effective complexity of the model and hence determines the degree of over-fitting. Although the introduction of regularization terms can control over-fitting for models with many parameters, this raises the question of how to determine a suitable value for the regularization coefficient λ.

2.4. Cost-sensitive learning

Cost-sensitive learning is an effective approach to solve over-fitting and generalization problem on imbalanced sample problems. The class imbalanced datasets occur in many real-world applications where the class distributions of data are highly imbalanced. For imbalanced sample problems, traditional classifiers usually lead to a poor learning accuracy on the rare class. One fundamental assumption in the traditional classifiers is that the goal of the classifiers is to maximize the accuracy. Under this assumption, in the case of imbalanced sample problems, predicting everything as the prevalent class is often the right thing to do. But this always leads to a poor effect for the rare class. Thus a natural thought to solve this question is to raise the misclassification cost of the rare class. For example, the misclassification cost of the rare class can be set to 10 times that of the prevalent class. Cost-sensitive learning is a type of learning in data mining that takes the misclassification costs into consideration. Thus, it is an effective approach to solve the imbalanced sample problems.

The goal of this type of learning is to minimize the total cost. The key difference between cost-sensitive learning and cost-insensitive learning is that cost-sensitive learning treats the different misclassifications differently. Cost-insensitive learning does not take the misclassification costs into consideration. The goal of this type of learning is to pursue a high accuracy of classifying examples into a set of known classes. Detailed descriptions on cost-sensitive learning can be found in Kukar and Kononenko [9] and Ling and Sheng [13].

3. VSG method based on Gaussian distribution

As is mentioned above, the previous VSG methods are hard to take both rationality and adaptability into consideration. In this section, a novel VSG method based on Gaussian distribution is proposed, generating virtual samples by utilizing the most common prior knowledge, smoothness. Theoretical justification for this method will be provided in Section 4.

3.1. Relevant definitions and theories

As Poggio and Vetter did not give an explicit definition on virtual sample, this paper made a summary of this concept and gave an explicit definition by the following. Along with the emergence