Research and Analysis of Fast Training in SVM-based Audio Classification

Shilei Zhang, Hongchen Jiang, Shuwu Zhang, and Bo Xu
Institute of Automation, Chinese Academy of Sciences, Beijing, 100080
{slzhang, hcjiang, swzhang, xubo}@hitic.ia.ac.cn

Abstract. In this paper, we propose a new method to choose the effective samples for support vector machines (SVM) training in audio classification task. The objective is to reduce the training time of SVM by choosing effective examples from the training set of binary classes. We test the performances of our new method on a dataset composed of about 6-hour audio data which illustrate that the computation time can be significantly reduced without a significant decrease in the prediction accuracy.

1 Introduction

With the rapid increase of multimedia information, the problem of distinguishing audio signals into these different audio types is thus becoming increasingly significant. Content-based audio classification and representation is broadly used in speech recognition, audio archive management, audio searching and indexing etc. Various methods for audio discrimination have been proposed for the needs of different applications [1, 2]. SVM is a very effective classifier algorithm for audio classification. Support vector machines are derived from the idea of the generalized optimal hyperplane with maximum margin between the two classes and this idea implements the structural risk minimization (SRM) principle in the statistical learning theory. Various training algorithms have been proposed to speed up the SVM training, including chunking, decomposition method, and Platt’s Sequential Minimal Optimization (SMO) [3]. Although these algorithms have been proven to accelerate the training, they do not scale well with the size of the training data.

We can learn about standard SVM training has $O(m^3)$ time and $O(m^2)$ space complexities, where $m$ is the training set size. Thus, when handling a large amount of data in machine learning, it is important to reduce the computation complexity and memory requirement without degrading the prediction accuracy. Research in this field has gained a lot of attention in these past few years [5–8]. In this paper, we present a new tree-based effective training samples selection method. In other words, regression tree is built through a process known as binary recursive partitioning where each decision node in the tree contains a training subset from the whole dataset; then we choose and balance the most qualified and effective samples for binary classes via data-driven algorithm.
2 Basic theory of SVM and LIBSVM package

The SVM is a discriminative classifier that is simple in concept but has some extensions that make it very powerful. Here we focus on the C-SVM applied to the two-class pattern recognition problem. Given \( n \) training patterns \( x_i \) and their associated classes \( y_i \in \{+1, -1\} \), the SVM decision function is:

\[
 f\{x\} = \sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + b.
\]

The coefficients \( \alpha \) are obtained by solving a quadratic programming problem:

\[
 \max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i y_i \alpha_j y_j K(x_i, x_j) \quad (1)
\]

\( \forall i, 0 \leq \alpha_i \leq C, \text{and} \sum_{i=1}^{n} \alpha_i y_i = 0. \) This optimization yields three categories of training samples depending on \( \alpha \). Samples corresponding to \( \alpha_i = C \) are called boundary support vectors (SVs). The set of boundary SVs includes all training samples misclassified by the SVM; Samples corresponding to \( 0 < \alpha_i < C \) are called standard SVs; Samples corresponding to \( \alpha_i = 0 \) play no role in the SVM decision function. Retraining after discarding these examples of non-SVs would still yield the same SVM decision function.

Our SVM implementation is based on the LIBSVM [4], a library for SVM classification and regression. LIBSVM adopts an SMO-type method for SVM
training strategy of solving quadratic problems. Radial basis function (RBF) is used as kernel, which is define as \( K(x_i, x_j) = exp(-\gamma ||x_i - x_j||^2), \gamma > 0 \). Fig. 1 shows a two-dimensional version of such a kernel. A smaller value of \( \gamma \) will give a smoother decision surface and more regular decision boundary. Choosing accurate model parameters is very important to SVM training. Model selection in this class of SVM involves two hyper-parameters: the penalty parameter \( C \) and the kernel width \( \gamma \). The \( \gamma \) in the RBF kernel controls the shape of the kernel and \( C \) controls the tradeoffs between margin maximization and error minimization. We perform a grid-search on \( C \) and \( \gamma \) using 5 fold cross-validation [4]. Basically pairs of \((C, \gamma)\) are tried and the one with the best cross-validation accuracy is picked. Fig. 2 shows an example of SVMs trained with different pairs \((C, \gamma)\). We can see the SVM is strongly linked to the two hyper-parameters and the performances are very different with different parameters. We can select a proper range of good parameter pairs by empiristic analysis and reject bad parameters by coarse grid search to accelerate the model selection for a given classification task.

3 Framework of SVM-based audio classification

We uniformly segment the audio signal into non-overlapping long clips, then the clip is further divided into non-overlapping 25ms long frames. Then various frame-level features are extracted from each clip and the means and standard deviations in this clip are computed to get clip-based feature to represent it. These clips are used as classification units. In this paper, multi-class audio classification classifies audio clips into one of four classes: pure speech, non-pure speech, music, noise. We adopt hierarchical classification structure for the SVMs in a multi-class pattern recognition task. Silence segments are first detected and removed using a simple energy-based algorithm. Then the non-silence clips are classified into speech and non-speech by the SVM1 classifier. Next, speech signals are classified into pure speech and non-pure speech by the SVM2 classifier, while non-speech signals are further classified into music and environment noise by the SVM3 classifier, respectively.

Fig. 3. Probability distribution curves of Normalized RMS Variance and 4Hz modulation energy
Nineteen kinds of audio Frame-level features are considered in this work. The detail description of these features can be found in the references [2, 9]. For different audio classes, the effectiveness and robustness of these features are not identical, therefore we select different subsets of available features for different classification spaces. Fig. 3 shows the probability distribution curves of Normalized RMS Variance and 4Hz modulation energy and illustrates the effectiveness for classifying music and noise. Based on experiments and analyses, we construct three groups of feature sets for the three SVM classifiers respectively, as shown in Table 1. This Framework will be applied to our following testing experiments. More details can be referred to [10].

Table 1. Three groups of feature sets

<table>
<thead>
<tr>
<th>Feature Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM1</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>SVM2</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>SVM3</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

4 Choice of effective samples based on distributional properties of the training data

The training data points which are adjacent to the boundary between the two classes tend to support the decision boundary and can be chosen as effective samples for SVM training [11]. The samples which are far from the boundary are easy to be classified correctly and can be considered as approximate non-support vectors. In other words, by exploiting the spatial distribution of the samples, the whole training dataset can be portioned into disjoint clusters, each of which consists of either samples belonging to two classes (i.e. samples adjacent to the boundary) or samples belonging to only one class (i.e. approximate non-support vectors). The new family of training data is constructed by choosing the clusters containing samples with different labels and replacing the clusters containing only non-support vectors by representative.

4.1 Tree-based clustering

Regression trees are attractive due to their simplicity in model interpretation, and are particularly suited for effective data mining. One of the important attributes of tree-based regression is its capability to generalize input-output mapping from the limited set of training samples. A regression tree is a binary tree
constructed by repeatedly splitting (sub)sets of learning cases into two descendant subsets. It is reasonable to assume that the regression tree corresponds to the acoustic space of the training data and describes the spatial distribution of the training data. Each node is assigned an acoustic class represented by the mean vector, covariance matrix and mixture weight, while every level can be viewed as a Gaussian Mixture Model (GMM). Next we will choose the effective samples within a proper tree level.

4.2 Effective samples choice

In this step, we select the number of clusters contained within a certain level based on the Bayesian Information Criterion (BIC) according to the training data. For the selected level, all nodes can be divided into two types: the ones consisting of samples with different class labels and the ones consisting of samples with the same labels. After replacing the latter with those representative mean vectors, a new training set for a given SVM classifier can be obtained by collecting all node clusters within the selected level. For some classification tasks, the above steps tend to produce a final training set with very different numbers of samples for both classes. Specific step to alleviate this problem is required. We added data points from the clusters which had already been replaced by the centroid vectors in the smaller class to achieve balance.

To further increase the training speed, we can reject obvious bad parameters by trained with the respective means of clusters for binary classes within the above selected tree level. Fig. 4 shows the training process with the means of clusters for two classes using the same data of Fig. 2. We can see they get the similar trend of performance.

![Fig. 4. SVMs trained with different pairs(C,γ) based on the means of tree-based clusters](image)

5 Experimental results

5.1 Dataset and experimental condition

It could be seen from Fig. 5 that, in general, the accuracy of classification decreases with the decrease of training data for binary classification. These experimental results indicate that 20 minutes of training set for each class can keep relative robust and good classification performance, so the original training data for each class in our experiments are about 20 minutes. The data are collected from real TV programs, which are about 343 minutes in total. 94 minutes of data are used for training, and 249 minutes of data are used for testing. The training set consists of 25 minutes of pure speech, 25 minutes of non-pure speech, 25 minutes of music and 19 minutes of environment noise. Pure speech and non-pure speech can be combined into speech class, while music and noise can be combined into non-speech class.

Fig. 6 shows the test clip accuracy, training time and the numbers of SVs with the growing size of testing unit on SVM3. It can be seen that 1s is a good tradeoff between classification and training time. In our experiments, we set 1s as a test unit. For LIBSVM, we try exponentially growing sequences for pairs of \((C, \gamma)\) \((C = 2^{-1}, 2^0, \ldots, 2^{13}; \gamma = 2^{-4}, 2^{-3}, \ldots, 2^3)\) and the one with the best cross-validation accuracy is picked as the optimal parameters.

![Fig. 5. Classification accuracy versus the size of training set on a) SVM2 and b) SVM3](image1)

![Fig. 6. Performance comparison on SVM3 with different testing unit (s). a) Test Accuracy; b) Training time and number of SVs](image2)
5.2 Performance evaluation

The training and testing results based on the original and the new training set are shown in Fig. 7. Results of effective samples choice are shown in Fig. 7 a). For SVM1 and SVM3, we could obtain a relatively small and balanced number of training data points according to our choice process. For SVM2, since the ratio between the number of support vectors and the total number of training data points is high, we can predict the data is relatively highly unseparable. So we cannot achieve effective data reduction using the proposed method. Comparison of the total number of SVs, training time and test accuracy is shown in Fig. 7 b), c) and d), respectively. From the experiments, it can be seen that the running time of the proposed method is greatly shorter than that of the original method, and the testing accuracy of the presented method is almost the same as those of the original method. We should notice that our algorithm uses fewer support vectors and keeps good generalization performance.

5.3 Visualization of the samples distribution

In order to visualize data points in the geometrical space, we need to project high-dimensional data to a 2-dimensional space. The linear discriminant analysis (LDA) finds a projection which minimizes the within-cluster scatter and maximizes between-cluster scatter and is used as an analysis tool to understand the natural distribution in this section. In Fig. 8, LDA was used to draw the distribution of training data set and support vectors of SVM3 in 2-dimensional space. We can see that effective samples chosen by our proposed algorithm are compact representations of original training set. Effective samples and new SVs can keep the distributional properties of the original data. Generally speaking, we can say that our algorithm has chosen very compact data points that maintain the original classification performance.

![Fig. 7. Performance comparison on the original and the new training sets. a) the size of training data; b) number of SVs; c) Training time(hour); d) test accuracy](image-url)
Fig. 8. Data distribution of data points of SVM3 in 2-dimensional space. a) Original training data; b) Effective samples; c) SVs trained with original data; d) SVs trained with effective samples

6 Discussion and Conclusions

In this paper, we implement the process of effective samples choice based on regression tree by exploiting the distributional properties of the training data, that is, the natural clustering of the training data and the overall layout of these clusters relative to the decision boundary of support vector machines. Experimental results show that our proposed method dramatically improves the speed of SVM training without reducing the generalization performance of SVM. For SVM2 in our experiments, we can try our algorithm in high dimensional feature space by projection in future work.

7 Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (NSFC) under the grant No. 60475014.

References