Active Learning on the Classification of Voice Pathologies

Carlos Lino Rengifo, Diego Andrés Alvarez, Ricardo Henao, Germán Castellanos, Jorge Eduardo Hurtado

National University of Colombia at Manizales

Abstract

In this article, it is studied the usefulness of the support vector machines (SVM) algorithm in the active classification of voice records into the sets normal and pathologic. In practice, each one of the samples employed on the classifier training must be manually labelled by an specialist, increasing in this way the training cost. Thus, it is imperative to obtain a classifier with a low generalization error, such that the number of training samples is as low as possible. A model selection technique, namely the Leave-One-Out criterion, was applied for the tuning of the appropriate parameters of the SVM. Also, a Radial Basis Function kernel was employed. The results obtained in the categorization of the aforementioned voice records showed that the number of tagged training samples can be reduced up to a 70% for the same testing error that the one obtained when the whole training set is labelled.

1. Introducción

In this article, it is studied the usefulness of the support vector machines (SVM) algorithm in the active classification of voice records into the sets normal and pathologic. In practice, each one of the samples employed on the classifier training must be manually labelled by a specialist, increasing in this way the training cost. Thus, it is imperative to obtain a classifier with a low generalization error, such that the number of training samples is as low as possible. A model selection technique, namely the Leave-One-Out criterion, was applied for the tuning of the appropriate parameters of the SVM. Also, a Radial Basis Function kernel was employed. The results obtained in the categorization of the aforementioned voice records showed that the number of tagged training samples can be reduced up to a 70% for the same testing error that the one obtained when the whole training set is labelled.

The learning phase of a classifier requires a training set which effectively represents the sets which must be discriminated. In general, there are two main techniques that can be used in the recognition of signal records: statistical- and machine-learning-based. In the former (includes classic discrimination methods like the Bayesian ones [1]), the feature selection must be made by means of a sequential methodology, demanding expertise and knowledge from the skilled person who processes the information. The second one (includes artificial neural networks [1], support vector machines [2, 3], etc.) exploits both the interpolation ability and the competence to generate nonlinear classifiers in a possibly multidimensional space, allowing the formation of non-connected data clusters. In this case, neither the probabilistic a priori information of each class nor the independence of the feature vectors is needed.

The support vector machine (SVM) algorithm has been successfully employed in classification tasks which involve feature vectors expressed in large dimensional spaces such as text categorization [4, 5, 6] and image recognition [7]. Also, a desirable feature of the SVM algorithm is that their generalization error is upper bounded by limits given by the statistical learning theory [8, 9, 10].

On the other hand, the automatic identification of voice signals (normal or pathologic) commonly uses two kinds of information parameters known as voice features: the acoustic features, which assess the vocal qualities and have a physical related sense, and the representational features, that correspond to calculated values of different methods of voice representation and, in general, do not have an associated physical sense, generally employing the coefficients resulting form the wavelet transform decomposition. In any case, the voice perceptual assessment includes methods which make deal of high cost and sophisticated instruments and techniques such as nasoendoscopy, videofluoroscopy, lateral encephalogram and nasometry among others. Also, the results obtained from the evaluation must be supported by a physician diagnostic given by the pediatric plastic surgeon and the phonolandysiologist. As a result, the labelling of each voice sample is a expensive duty in terms of time and money; therefore it is imperative to label the smallest possible set of patterns, preserving the discriminatory ability of the classifier.

Little research has been done in the area of active learning for speech understanding. In [11] and [12] two methods. One of them, certainty-based learning, selects the examples that the classifier is less confident about. The other, inspired by committee-based learning selects the examples that multiple classifiers do not agree on. They found that it is possible to reduce the human labelling effort at least by a factor of two.

The present paper shows a SVM classifier aimed to the recognition of voice signals (normal and pathologic) with acoustic and wavelet-based representation features, making special emphasis in the reduction of the number of samples employed in the design phase of the classifier. A model selection strategy was employed to select the appropriate parameters of the SVM, namely, the Leave-One-Out criterion. Also, a Radial Basis Function kernel was used. The presented article differs from other approaches in that the active learning technique employed tries to minimize as fast as possible the number of classifiers that effectively discriminate the pool of training data, in this way, increasing notably the performance of the classifier.

2. Active learning using SVM

The SVM algorithm corresponds to a family of functions of the form:

\[ f(x, \gamma) = w \cdot \Phi(x) \]  

where \( x \in \mathbb{R}^d \equiv L \) denotes the feature vector, \( w \) represents the weight vector, which is calculated from the training set, and \( \Phi \) symbolizes a mapping function from the input lower dimensional space \( L \) to a generally high dimensional Hilbert space \( \mathcal{H} \), \( \Phi : L \mapsto \mathcal{H} \), where the training set is supposed to be linearly separable. Once the SVM is trained, it is expected that the sign of the evaluation of a vector \( x \) into (1), provides the correct label.

This work was supported by the Research Office of the National University of Colombia at Manizales (DIMA). Projects ODS530, ODS472.
The weight vector of the SVM $\mathbf{w}$, is adjusted via classifier training. Given a training set, $\{\mathbf{x}_i\}$, with their corresponding tags $\{y_i\} \in \{-1, 1\}$ for $i = 1, \ldots, l$, the weight vector $\mathbf{w}$ can be expressed as the linear combination,

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \Phi(\mathbf{x}_i)$$

so that $\alpha_i$ is found by minimizing the quadratic function

$$L(\alpha) = \sum_{i=0}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$$

(2)

This expression (2) is subject to the constraint $C \geq \alpha_i \geq 0$ for $i = 1, \ldots, l$, where $C$ is a constant which makes the classifier tolerant to the lack of separability of the training set in the space $\mathcal{H}$. For example, an infinite value of $C$ corresponds to the case of a separable training set, also referred to as the hard margin case. The points whose corresponding $\alpha_i > 0$, are known as support vectors and are those points closest to the decision hyperplane in the space $\mathcal{H}$. The inner product $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ in $\mathcal{H}$ is carried out using a function known as the kernel of the SVM $K(\mathbf{x}_i, \mathbf{x}_j)$ which must satisfy the Mercer conditions [2].

A commonly used kernel is the Radial Basis Function (RBF):

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\gamma}\right)$$

where $\gamma$ denotes a parameter which requires tuning for an adequate performance of the SVM. When $\gamma$ is changed, automatically the $\mathcal{H}$ space varies. Model selection techniques [13] are employed for the appropriate selection of the parameter $\gamma$. This techniques aim towards the minimization of the generalization error,

$$R(\gamma) = \int \frac{1}{2} \|y - f(\mathbf{x}, \gamma)\| dP(\mathbf{x}, \gamma)$$

making the assumption that the elements of the training set are sampled identically and independently from the probability distribution function $P$. In practice, it is hard to know the function $P$; therefore, an upper bond on the generalization error is employed; one technique of model selection is the so called leave-one-out (LOO) method [2]. Given a training set $T = \{(\mathbf{x}_j, y_j), j = 1, \ldots, l\}$ the optimal $\gamma$ is calculated using the LOO technique as,

$$\gamma = \arg \min_{\gamma} \frac{1}{l} \sum_{k=1}^{l} \left[ -y_k \cdot \text{signo} \left( f^{+k}(\mathbf{x}_k, \gamma) \right) \right]$$

where $I$ stands for the indicator function,

$$I[x, y] = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$$

and $f^{+k}$ denotes a SVM trained on the set $T \setminus \{\mathbf{x}_k, y_k\}$.

As the weight vector $\mathbf{w}$ (see 1) has a unit norm induced by the RBF kernel, the family of points $\mathbf{w}$ has the shape of a unit radius hypersphere $\mathcal{W}$. The subspace of $\mathcal{W}$ whose elements effectively separate the training set is called the version space, $\mathcal{V}$ [14]: from the generalized form of the SVM (1), it can be seen that the elements of $\mathcal{W}$ correspond to hyperplanes in $\mathcal{H}$ and viceversa. The training phase of the SVM can be associated with the search of the largest hypersphere that can be described into the version space; this sphere has a radius $1/\|\mathbf{w}\|_{\mathcal{H}}$ and those hyperplanes representing the training vectors and that are tangent to the considered sphere are the aforementioned support vectors.

It can be seen that the minimization of the area $\mathcal{V}$ leads to a diminution in the number of classifiers that can effectively separate the training set, reducing, in this way, the possibility of overfitting of the SVM.

On the other hand, the active learning technique is employed in the classifier training such that, from a pool of unlabelled vectors, a small subset of it is chosen such that the training on the already labelled subset brings about a classifier with a high discriminant performance.

Given a labelling set $E$, an active learning machine is made up of three components: $\{X, f, q\}$: the first one is the pool of labelled and unlabelled vectors, the second is the classifier $f : X \mapsto E$, trained on the labelled subset of $X$, and the last component, $q$, is named the oracle function, and decides which element of $E$ is the tag of an element of $X$. Thus, the active learning machine returns a classifier $f$ after each query. When the SVM is used as an active learning machine, it is desired to reduce, as fast as possible, the area of the version space and, in this manner, reduce the number of vectors $\mathbf{w}$ that can accomplish the classification task. In this case, given an unlabelled data set, a vector is randomly selected from each class. From this initial set, a SVM is trained. In the following, this SVM will be referred to as $\text{SVM}_2$. The version space generated by this classifier $\mathcal{V}_2$ gives the basis for the selection of the following vector to be labelled; the next point to present to the oracle is an unlabelled vector of $X$, such that, it splits, as much as possible, the version space $\mathcal{V}_2$ into halves. Once the oracle function returns the correct label of the point, a new support vector machine, $\text{SVM}_3$, is trained. This SVM has a corresponding version space $\mathcal{V}_3$. The training process goes on in the form sketched before, with the selection of a new vector $X$ to be tagged, such that it halves the version space $\mathcal{V}_3$, as much as possible.

There are basically three learning strategies of active training using SVM [14]:

- **Simple margin.** This points selects as the next point to query the oracle the one which is closest to the decision hyperplane in the space $\mathcal{H}$. In this way, it is expected to find the closest point to the center of the largest hypersphere which is actually contained in the version space. It must bear in mind that this criterion is not adequate when the version space has an elongated shape.

- **MaxMin Ratio.** As mentioned before, the vector $\mathbf{w}$ of the SVM represents in the version space the center of the largest hypersphere enclosed into $\mathcal{V}$. Given an unlabelled point of $X$, it is desired that the new vector to be tagged does split as much as possible the version space in halves. A rough estimator of the area of the version space is the area of the largest sphere which fits into it. In this way, labelling one by one each unlabelled point as +1, then classifying them (so that a version space $\mathcal{V}_+^i$ with an approximate radius $r^+$) is generated, and making the same using the tag -1 (obtaining in this case a corresponding version space $\mathcal{V}_-^i$ and a radius $r^-$). It is clearly understood that, the following point to label is the one for which $\min(\text{Area}(\mathcal{V}_+^i), \text{Area}(\mathcal{V}_-^i))$ is maximum; As these areas are related by the sizes of the hyperspheres in $\mathcal{H}$, it can be concluded that the following point to label is the one for which the minimum of the radius of the hyperspheres of the SVM obtained with the labels +1 and -1 corresponds to the maximum radius.
Margin Ratio. Basically this method is similar to the MaxMin method, however, the selection criterion is different. In this case, the next point to label is the one for which $\min\left(\frac{r^-}{r^+}, \frac{r^+}{r^-}\right)$ is maximum.

While the simple margin method calls once the SVM training algorithm, for the whole set of already labelled vectors, both the MaxMin and Margin Ratio methods call twice the SVM training algorithm for each labelled training vector, so that the computational cost of these strategies is high, however, the performance of them is far superior [14].

3. Analysis of results and conclusions

The experiments performed where made using 80 samples of each of the following training sets:

1. Acoustic features ($AF$).
2. Representational features ($RF$).
3. Compound set of acoustic and representational features ($AF+RF$).

A binary classifier was constructed to discriminate between the classes normal and pathologic. These sets where built as shown in Table 1.

<table>
<thead>
<tr>
<th>Set</th>
<th>Observations</th>
<th>Size of the feature vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>$RF$</td>
<td>80</td>
<td>12</td>
</tr>
<tr>
<td>$AF$</td>
<td>80</td>
<td>5</td>
</tr>
<tr>
<td>$AF+RF$</td>
<td>80</td>
<td>17</td>
</tr>
</tbody>
</table>

As a preliminar result, a SVM was trained with constant parameters $C$ and $\gamma$ and the whole set of training samples in order to estimate the margin of the SVM, using it as a rough criteria of the separability of the classes in matter. Table 2 presents those margins.

The classifiers were constructed using a SVM for the hard margin case ($C = \infty$), using a LOO strategy for the appropriate selection of gamma value. In view of the fact that the training set was randomly selected, the training phase was performed ten times using different starting sets, in order to obtain a model with consistent results; in this way, the current figures show both the mean and the standard deviation of the validation error.

**Representational features set** This feature set was extracted using the two main components of each of the six levels of decomposition for a wavelet db8 employing only utterings of the /a/ vowel [15]. Figure 1 shows the history of the validation error variation with regard to the number of training samples.

<table>
<thead>
<tr>
<th>Table 1: Training sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>$RF$</td>
</tr>
<tr>
<td>$AF$</td>
</tr>
<tr>
<td>$AF+RF$</td>
</tr>
</tbody>
</table>

**Acoustic features set** The acoustic features have been extensively employed in the field of voice recognition. The feature set used in the present research corresponds to: pitch ($AF_1$), jitter ($AF_2$), Normalized Error Prediction (NEP - $AF_3$), Glottal to Noise Excitation (GNE - $AF_4$) and Harmonic to Noise Ratio (HNR - $AF_5$) of the dataset [15]. As in the $RF$ set, only the /a/ vowel was taken in consideration. In the present case, the input space vectors where normalized since the characteristics are not in comparable ranges. Figure 2 shows the history of the validation error variation with regard to the number of training samples for the AF training set. It can be seen that making a comparison between figures 1 and 2 the convergence of the active training using the RF set is faster.

**Compound set of acoustic and representational features** This compound set was formed by the aforementioned acoustic and representational feature sets: $AF+RF$, so that the new feature vector has a dimension of 17. Figure 3 shows the history of the validation error variation with regard to the number of training samples for the compound AF+RF training set. It

![](figure1.png)

Figure 1: History of the validation error variation with regard to the number of training samples for the RF training set

![](figure2.png)

Figure 2: History of the validation error variation with regard to the number of training samples for the RF training set

![](figure3.png)

Figure 3: History of the validation error variation with regard to the number of training samples for the RF training set

Table 2: Margin of the training sets ($C = \infty, \gamma = 0.1$).

<table>
<thead>
<tr>
<th>Set</th>
<th>Margin</th>
</tr>
</thead>
<tbody>
<tr>
<td>$RF$</td>
<td>0.1272</td>
</tr>
<tr>
<td>$AF$</td>
<td>0.0348</td>
</tr>
<tr>
<td>$AF+RF$</td>
<td>0.1279</td>
</tr>
</tbody>
</table>
can be seen that the convergence of this set is the fastest of the mentioned strategies.

From the obtained results one can infer that the acoustic features are closer in the feature space $\mathcal{H}$ than those obtained by wavelets methods. For this reason, the RF and AF+RF sets require a smaller number of training samples to converge.

On the other hand, it can be observed that the compound set AF+RF has a better performance related to the number of samples required and the obtained testing error, resulting in a better generalization capability.

4. References


