Journal of the Korean Statistical Society 41 (2012) 105–114

Contents lists available at SciVerse ScienceDirect

Journal of the Korean Statistical Society

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

The guaranteed estimation of the Lipschitz classifier accuracy: Confidence set approach

Andrey V. Timofeev *

Department of the Statistics, Speech Technology Center, 4 Krasutskogo str., St.-Petersburg, 196084, Russia

ARTICLE INFO

Article history:
Received 3 January 2011
Accepted 12 July 2011
Available online 30 July 2011

AMS 2000 subject classification:
62H12
62H30
62F25

Keywords:
Lipschitz classifiers
Confidence set
Classifier accuracy

ABSTRACT

This paper introduces an original method for the guaranteed estimation of the Lipschitz classifier accuracy in the case of a large number of classes. The solution was obtained as a finite closed set of alternative hypotheses, which contains an object of classification with probability of not less than the specified value. Thus, the classification is represented by a set of hypothetical classes. In this case, the smaller the cardinality of the discrete set of hypothetical classes is, the higher is the classification accuracy. This problem is relevant in practical biometrics, when the number of analyzed samples amounts to tens of thousands, and many of them are distinguished vaguely in the primary feature space.

1. Introduction

One of the main problems of stochastic sample classification is to estimate degree of similarity between a sample and those classes that are located close in the metric of the feature space. A question arises: what is the formal mechanism for the selection of classes, corresponding to a sample with a priori specified lower bound of the classification accuracy value? This issue is extremely important in practical biometrics, when the number of classes can be millions, and many of them can be poorly distinguishable in the feature space.

Typically, classifiers estimate the similarity of a sample $\mathbf{X}^{(\theta^*)}$ regarding to class $\theta \in \Theta$ (Θ is a set of classes, $\theta^* \in \Theta$ is a true index of the class to which the sample belongs) in a form of the so-called score-parameter $f\left(\theta|\mathbf{X}^{(\theta^*)}\right), f\left(\theta \mid \mathbf{X}^{(\theta^*)}\right) \in R^1$. We denote the score-parameter as a symbol $f\left(\theta \mid \mathbf{X}^{(\theta^*)}\right)$, and the class, to which actually the sample $\mathbf{X}^{(\theta^*)}$ belongs, as a symbol $\chi\left(\mathbf{X}^{(\theta^*)}\right)$. It is obvious that $\theta^* = \chi\left(\mathbf{X}^{(\theta^*)}\right)$. Let the classification decision be made as follows:

$$\tilde{\theta} = \operatorname{Arg}\operatorname*{Max}_{\theta\in\Theta}\left(f\left(\theta|\mathbf{X}^{(\theta^*)}\right)\right). \tag{1}$$

Since classifiers have a stochastic nature, their outputs are random variables. The output of any classifier can be described by the following statement:

$$f\left(\theta|\mathbf{X}^{(\theta^*)}\right) = \mathbf{E}_{\theta^*} f(\theta|\mathbf{X}^{(\theta^*)}) + \eta\left(\theta|\mathbf{X}^{(\theta^*)}\right).$$

Here,

* Tel.: +7 921 590 70 09; fax: +7 812 327 92 97.
E-mail address: timofeev.andrey@gmail.com.

1226-3192/$ – see front matter
doi:10.1016/j.jkss.2011.07.005

- $\theta^* \in \Theta$ is index of the target object class;
- $\theta \in \Theta$ is index of the testing hypothesis class;
- $\mathbf{E}_{\theta^*} f(\theta|\mathbf{X}^{(\theta^*)})$ is the expected value of a random function $f(\theta_k|\mathbf{X}^{(\theta^*)})$ with specified parameters $\theta^*, \theta \in \Theta$;
- $\eta(\theta|\mathbf{X}^{(\theta^*)})$ is a random function of the parameters $\theta^*, \theta \in \Theta$.

Of course, the target object class index $\theta^* \in \Theta$ is not explicitly observed.

But how can we estimate the classifier accuracy? The well-known methods are the ones that are based on exponential approximation of conditional a posteriori probability distribution of the solution function (Platt, 1999; Wahba, 1999). In this case, we consider a binary classification with classes $w_1 = 1$ and $w_2 = -1$. Let us denote the training set by $\mathbf{T} = \{(x_i, y_i) \,|i = 1, \ldots, N\}, y_i \in \{-1, 1\}$ and by $P(w_k \mid x_i)$ – posteriori probability of x_i corresponding to class $w_k, k \in \{1, 2\}$. As on Platt's approach (Zadrozny & Elkan, 2002), it is proposed to use sigmoid approximation of $P(w_k \mid x_i)$, assuming that this function is continuous at x. This approximation has the following representation: $P(class \mid input) = P(w_k \mid x) = (1 + \exp(Ax + B))^{-1}$. As it was described by Platt, the constants A and B could be found using the so-called "cross-entropy error function" (Bishop, 1995). Following Platt (1999), the constants A and B are determined by solving the following minimization problem:

$$\operatorname{Arg}\min_{A,B} \{J(A, B)\}.$$

Here, $J(A, B)$ is the so-called "cross-entropy function":

$$J(A, B) = -\sum_i \left(p_i \log\left((1 + \exp(Ax_i + B))^{-1}\right) + (1 - p_i) \log\left(1 - (1 + \exp(Ax_i + B))^{-1}\right)\right)$$

where $p_i = ((y_i + 1)/2)$. The peculiarity of this method is a lack of theoretical validity to select an approximating function $(1 + \exp(Ax + B))$. In Zadrozny and Elkan (2002) it was shown that Platt's method gives inadequate results for certain test datasets, e.g. for Adult and TIC (The Insurance Company Benchmark) datasets, which are available in UCI ML Repository (Blake & Merz, 1998). In case of a multi-class problem with a small cardinality of certain class j training set $\mathbf{T}^{(j)} = \left\{\left(x_i^{(j)}, j\right) |i = 1, \ldots, N_j\right\}$, Platt's method provides unacceptable results of a posteriori probability estimation (Bennett, 2000).

It is very important to realize that the quality of the classification decisions, ideally, should be evaluated in a probabilistic sense. Approach based on the Platt calibration has an empirical basis. Estimation of the conditional probability of the correct classification $P\left(\tilde{\theta} = \chi(\mathbf{X}^{(\theta^*)})|\theta^* = \operatorname{Max}_{\theta\in\Theta}(f(\theta|\mathbf{X}^{(\theta^*)}))\right)$, which results from the use of Platt approach, is entirely defined by the training set. If a training sample is unrepresentative (this case is very common in practical biometrics), estimation of the conditional probability $P(\omega|\theta^* = \operatorname{Max}_{\theta\in\Theta}(f(\theta \mid \mathbf{X}^{(\theta^*)})))$ of the event $\omega : \{\tilde{\theta} = \chi(\mathbf{X}^{(\theta^*)})\}$, following Platt's approach, may be ineffective.

Few interesting works, giving a detailed theoretical analysis of SVM-classifier errors, have recently appeared, for example Di-RongChen, QiangWu, YimingYing, and Ding-XuanZhou (2004). This work carries out a thorough research of regularization error values boundaries, as well as the boundaries of the corresponding approximation error values for reproducing kernel Hilbert spaces (in two classes separation case). The main focus of this paper is put on such method of regularization parameter value selection that gives the highest accuracy of classification.

In this work we propose a new method for probabilistic estimation of the classification decision accuracy. Assume that there exists a unique class $\theta^G \in \Theta$ such that

$$\theta^G = \operatorname{Arg}\operatorname{Max}_{\theta\in\Theta}(\mathbf{E}_{\theta^*} f(\theta|\mathbf{X}^{(\theta^*)})),$$

and in this case $\theta^G = \theta^*$. In fact, these constraints define the conditions that ensure the correct classification result in a deterministic case. In a nondeterministic case, which is more practical, the stochastic component of the classifier output gives an error in the classification decision. Let the absolute value of the stochastic component $\eta\left(\tilde{\theta} \mid \mathbf{X}^{(\theta^*)}\right)$ of $f(\tilde{\theta})$ in (1) be $|\eta(\tilde{\theta} \mid \mathbf{X}^{(\theta^*)})| > 0$. Informally, if for the certain class $\theta_1 \in \Theta$ the inequality $|f\left(\tilde{\theta} \mid \mathbf{X}^{(\theta^*)}\right) - f\left(\theta_1 \mid \mathbf{X}^{(\theta^*)}\right)| < |\eta\left(\tilde{\theta} \mid \mathbf{X}^{(\theta^*)}\right)|$ is true, then it is impossible to determine for sure which of these classes $\tilde{\theta}$ and θ_1 corresponds to the sample. Thus, these classes are distinguishable with respect to the sample, which was presented for the classification, only in probabilistic sense. We will call such classes the probabilistically distinguishable classes, or the *target sets*. From a practical standpoint, it is very important to define a formal procedure for the correct definition of these target sets.

It is imperative to ensure that the desired class $\theta^* = \chi(\mathbf{X}^{(\theta^*)})$ belongs to the target set with probability of not less than a given value of $P_c, P_c \in (0, 1)$. In this case, the classification results are yielded not as one class, but as a set of classes that together constitute the target set. The research subject of this paper is a rigorous solution to guarantee a certain closed set of classes (target set) to which the sample belongs with probability of not less than P_c. Naturally, the target set is defined for the previously given classifier.

It should be noted that the so-called Lipschitz classifiers (Luxburg & Bousquet, 2004) correspond to a very wide type of classifiers. This type includes such well-known and practically effective classifiers as SVM (Support Vector Machine), Linear Programming Machines and even NN (Nearest Neighbor) method. A variety of the Lipschitz classifiers and their broad application mainly caused the major focus on this type of classifiers within the frames of this paper. The presented method to

estimate the classification accuracy can be applied to any type of Lipschitz classifiers. The mechanisms of Lipschitz classifier construction and learning, however, are out of the scope of this research. Thus, within this article, Lipschitz classifiers are considered to be given, research on their accuracy only is being carried out.

The approach suggested in this paper allows estimation of the Lipschitz classifier accuracy by determining the closed confidence subset within a priori given set of classes. This set, having a specified confidence coefficient, contains an index of classification object. The classification accuracy is determined by the number of alternative hypotheses included in this confidence set (target set): the smaller the cardinality of the confidence set is, the higher the classification accuracy is. The suggested approach provides the guaranteed accuracy of estimation, i.e. how well we can expect the constructed confidence set to contain a true object of classification with priory specified confidence coefficient.

Within the content of this article, cardinality of set X is denoted as $|X|$. The symbol Diam (Z) denotes the diameter of the set Z and Diam $(Z) = \sup_{z_1, z_2 \in Z} \|z_1 - z_2\|$.

2. A brief note on the Lipschitz classifiers

As mentioned above, the Lipschitz classifiers (Luxburg & Bousquet, 2004) are now widely prevalent in a number of practical applications, such as biometrics, automatic text analysis, and in various mathematical economics applications. For example, widely-spread SVM classifiers, Linear Programming Machines, and even Nearest Neighbor method are all particular cases of Lipschitz classifiers. Let us discuss the concept of the Lipschitz classifiers in more detail. The fundamentals of class separation by means of large margin hyperplane go back to Vapnik's and Chervonenkis's works (1974). It was demonstrated that using the principle of margin maximization between sets of different classes elements from training data and a hyperplane separating these classes, allows to minimize the upper bound of the empirical risk and to increase the classifier's generalization ability. Note should be taken that marginal classifier construction is closely associated with problems of isomorphic and isometric embedding of metric spaces into Banach or Hilbert spaces. In this case, the feature parameters compact space (Z, d) (where d is a metric of that space, and Z is a set of feature parameters values) is isometrically embedded into the target Banach space $(B, \|\cdot\|)$ by means of the feature mapping $\phi : Z \to B$. The margin classifier is constructed in subspace of the space $(B, \|\cdot\|)$, where the convex sets are separated by hyperplane, as on Hahn–Banach separation theorem. Particularly, when a metric space is embedded into Hilbert space, the well known Support Vector Machine algorithm is used. Let us consider this topic in more detail. Generally speaking, a large margin classifier is a classification rule such that most of the training samples are "far away" from classification boundary (some hyperplane H in B), and the margin to the two classes is maximized. Let the set $\{(z_i, y_i) \,|i = 1, \ldots, n\} \subset Z \times \{\pm 1\}$ be the training data. The convex hull of set F is denoted by $C(F) = \left\{\sum_{i\in I} \alpha_i z_i \,\middle|\, \sum_{i\in I} \alpha_i = 1, \forall_{i\in I} \alpha_i > 0, z_i \in F, |I| < \infty\right\}$. Let F^+, F^- be training sets of the two classes "+" and "−", $C(F^+) \cap C(F^-) = \emptyset, B'$ − the dual space of B. Bennett and Bredensteiner (2000) showed that constructing the large margin classifier is equivalent to finding the distance $d(C(F^+), C(F^-))$ between $C(F^+)$ and $C(F^-)$. In this case, $d(C(F^+), C(F^-)) = \inf_{p^+\in C(F^+), p^-\in C(F^-)} \left\|p^+ - p^-\right\|$, and it was concluded that $d(C(F^+), C(F^-)) = \sup_{T\in B'} \inf_{p^+\in C(F^+), p^-\in C(F^-)} \left\langle T, p^+ - p^-\right\rangle \|T\|^{-1}$. Thereby, we have the equivalence optimization problem:

$$\inf_{T\in B', b} \|T\| \text{ subject to } \overset{n}{\underset{i=1}{\forall}} \left(y_i \left(\langle T, z_i\rangle + b\right) \geq 1\right).$$

As a result, large margin classifier is the solution of this optimization problem in a form of decision function $f(z) = \langle T, z\rangle + b$. In this case, the margin value is given by $\|T\|^{-1}$. Let us denote by $AE(Z)$ the Arens–Ells space (Arens & Eells, 1956), by $AE(Z)'$ − the dual space to $AE(Z)$ (i.e. the space of all continuous linear forms on $AE(Z)$), and the Lipschitz set of functions by:

$$\mathrm{Lip}(Z) := \{f \,|Z \to \mathbf{R}; \forall f \exists L]0, \infty[: \forall x, y \in Z : |f(x) - f(y)| \leq L\mathrm{d}(x, y)\}.$$

Here, $L(f)$ is the smallest constant L such that $|f(x) - f(y)| \leq L\mathrm{d}(x, y)$ for all $x, y \in Z$. Luxburg and Bousquet (2004) suggested embedding Z isometrically into the Banach space $AE(Z)$: $\left(\Phi : Z \to AE(Z), \Psi : \mathrm{Lip}(Z) \to AE(Z)'\right)$. Here, mapping Φ is an isometric embedding of Z into $AE(Z)$, mapping Ψ between $\mathrm{Lip}(Z)$ and $AE(Z)'$ is such that $\mathrm{Lip}(Z)$ is isometrically isomorphic to $AE(Z)'$. Also, it was suggested to construct large margin classifier by solving the following optimization problem:

$$\inf_{f\in \mathrm{Lip}(Z)} L(f) \text{ subject to } \overset{n}{\underset{i=1}{\forall}} \left(y_i f(x_i) \geq 1\right).$$

The solution of this problem is called hard margin Lipschitz classifier, and its margin is defined as $1/L(f)$ (Luxburg & Bousquet, 2004). On the other hand, soft margin Lipschitz classifier was constructed resulting from the solution of the following optimization problem:

$$\inf_{f\in \mathrm{Lip}(Z)} \lambda L(f) + n^{-1} \sum_i l\left(y_i f(z_i)\right), l\left(y_i f(z_i)\right) = \max\left\{0, 1 - y_i f(z_i)\right\},$$

where λ is a trade-off constant. Besides the isometric embedding (Z, d) into $AE(Z)$, there are many different isometric embeddings which could be used instead. For example, Hein and Bousquet (Hein & Bousquet, 2003) studied Kuratowski embedding of a metric space Z into the space of continuous function to construct a large margin classifier. The Lipschitz

classifier decision function has a small Lipschitz constant. This construction goes with regularization principal, which avoids using functions with a high variation.

The Lipschitz classifiers include a wide class of classification algorithms. That is why the results of this study deals specifically with the Lipschitz classifiers.

3. Problem statement

Let (Z, d) be a compact metric space of the feature parameters, d-a metric of this space, Z-a set of values of feature parameters. There exists a specified set of indices for classes $\Theta = \{\theta_k\} \subseteq R^1$, $|\Theta| < \infty$ such that $\theta_2 \neq \theta_1 \in \Theta : d(\theta_2, \theta_1) > 0$.

Classifier $f : Z \to \Theta$ is a function that divides the space (Z, d) into $m = |\Theta|$ classes.

Let us assume that the classifier f, denoted as $f(\theta \mid \mathbf{X}^{(\theta^*)})$, is Lipschitz margin classifier (Luxburg & Bousquet, 2004). Here, $f(\theta \mid \mathbf{X}^{(\theta^*)})$ is a stochastic function, dependent explicitly on $\theta \in \Theta$ and implicitly – on $\theta^* \in \Theta$, where $\theta \in \Theta$ is the hypothesis index. The classification decision is made according to (1).

The goal of this paper is to determine such a confidence set (target set) of indices $\Xi(\tilde{\theta}) \subseteq \Theta$, based on the observed data analysis and the result of point classification, for which the following statement is true: $\mathbf{P}(\theta^* \in \Xi(\tilde{\theta})) \geq P_c$. In this case, the confidence coefficient P_c is a priory specified.

4. Solution method

As earlier denoted, $f\left(\theta \mid \mathbf{X}^{(\theta^*)}\right)$ corresponds to margin classifier. Additionally, for the specified parameter $c \in (0, \infty)$, let us consider the following auxiliary set:

$$\Xi(\tilde{\theta} \mid c, \mathbf{X}^{(\theta^*)}) = \left\{\theta \in \Theta \,\middle|\, \left|f(\tilde{\theta} \mid \mathbf{X}^{(\theta^*)}) - f(\theta \mid \mathbf{X}^{(\theta^*)})\right| \leq c\right\}.$$

The main result of this paper is the following theorem:

Theorem 1. *Let the following conditions be true:*

1. $\forall \theta, \theta^* \in \Theta : \mathbf{E}_{\theta^*}\left(\eta\left(\theta \mid \mathbf{X}^{(\theta^*)}\right)\right) = 0$.
2. $\forall \theta^* \in \Theta : \mathrm{Max}_{\theta \in \Theta}(\mathbf{E}_{\theta^*} f(\theta \mid \mathbf{X}^{(\theta^*)})) = \mathbf{E}_{\theta^*} f(\theta^* \mid \mathbf{X}^{(\theta^*)}), |\{\theta^{\#} \in \Theta | \mathbf{E}_{\theta^*} f(\theta^{\#} \mid \mathbf{X}^{(\theta^*)}) = \mathrm{Max}_{\theta \in \Theta}(\mathbf{E}_{\theta^*} f(\theta \mid \mathbf{X}^{(\theta^*)}))\}| = 1$.
3. $\forall \theta_1, \theta_2, \theta^* \in \Theta \exists L \in]0, \infty[: \left\|f\left(\theta_1 \mid \mathbf{X}^{(\theta^*)}\right) - f\left(\theta_2 \mid \mathbf{X}^{(\theta^*)}\right)\right\| \leq L\mathrm{d}(\theta_1, \theta_2)$ *a.s.*

Then, if $c(P_c) = L \cdot \mathrm{Diam}(Z)(1 - P_c)^{-0.5} > 0$ *we have*

$$\forall \theta^* \in \Theta : \mathbf{P}_{\theta^*}\left(\theta^* \in \Xi\left(\tilde{\theta} \mid c(P_c), \mathbf{X}^{(\theta^*)}\right)\right) \geq P_c.$$

The proof of this theorem is presented in the Appendix.

The second condition of the theorem implies that, when no noise is present, the function reaches maximum at the classification object index. In fact, this is the necessary condition for proper classification by means of this discriminant function. At training stage, parameters of the classifier should be chosen in such a way that condition 2 of the Theorem 1 is met at least for the training set. In case if this condition is not fulfilled, it is reasonable to consider the following positive constant $\vartheta^*(\Theta)$:

$$\vartheta^*(\Theta) = \mathrm{Inf}\left\{\vartheta \in R^1 \,\middle|\, \forall \theta^* \in \Theta : \left(\mathrm{Max}_{\theta \in \Theta}\left(\mathbf{E}_{\theta^*} f\left(\theta \mid \mathbf{X}^{(\theta^*)}\right)\right) - \mathbf{E}_{\theta^*} f\left(\theta^* \mid \mathbf{X}^{(\theta^*)}\right) \leq \vartheta\right)\right\}.$$

The constant $\vartheta^*(\Theta)$ determines the minimum upper bound of $\left(\mathrm{Max}_{\theta \in \Theta}\left(\mathbf{E}_{\theta^*} f\left(\theta \mid \mathbf{X}^{(\theta^*)}\right)\right) - \mathbf{E}_{\theta^*} f\left(\theta^* \mid \mathbf{X}^{(\theta^*)}\right)\right)$ for all $\theta^* \in \Theta$. The value of $\vartheta^*(\Theta)$ is estimated numerically at the training stage of the classifier. It can be shown that in this case Theorem 1 would be true if $c(P_c) = \vartheta^*(\Theta) + L \cdot \mathrm{Diam}(Z)(1 - P_c)^{-0.5}$.

The third condition implies that there exists a Lipschitz constant for this discriminant function. The fourth condition presupposes absolute value limitedness of the discriminant function in the set Θ. Thus, if the conditions of Theorem 1 are satisfied, the discrete set of hypotheses $\Xi\left(\tilde{\theta} \mid c(P_c), \mathbf{X}^{(\theta^*)}\right)$ contains the target object index θ^* with probability of not less than P_c. The classification is represented by a set $\Xi\left(\tilde{\theta} \mid c(P_c), \mathbf{X}^{(\theta^*)}\right)$ of hypothetical classes. However, when the introduced method is applied in practice, it is crucial that for the given value of P_c the value of cardinality $|\Xi\left(\tilde{\theta} \mid c(P_c), \mathbf{X}^{(\theta^*)}\right)|$ would be relatively small. One important practical example will let us see whether this is satisfactory.

5. Practical example: voice biometric application

The output of biometric samples classification system always represents a set of classes that correspond to the tested sample with varying degrees of reliability. In this case, a possibility of bounding the cardinal number of this set from above is extremely important. For example, the problem of finding a human face in a database of photographic images may necessitate searching among the millions of samples. At the same time many face images differ very insignificantly, so there is always a high probability of making a false identification decision which may lead to very serious practical consequences. Therefore, it makes sense to choose an entire set of classes in stead of one class as the results of biometric samples classification. The chosen set of classes should be further studied, possibly involving expert methodology. Since expert research methods are very expensive, it is important to relevantly reduce the cardinality of the set chosen for further expert examination. Similar situation is also typical for biometric applications associated with identification of a person using a sample of his/her voice. Same as in facial recognition applications, search is carried out across the database containing hundreds of thousands and even millions of samples. The systems of automatic identification of a person that use a voice sample are now growing ever more widespread. Human voice as an object of biometric research is rather complex, comparatively to use of a facial image or a fingerprint for identification. This is the exact reason why voice biometrics is a good example to estimate the effectiveness of the proposed technology.

As an example of practical use of the suggested algorithm, evaluation of the target speaker identification accuracy was considered. In this case our goal is to determine which voice within the known voices group best matches the analyzed voice sample. Thus, we have a phonogram with a voice sample of a priori unknown speaker; however, it is known that this speaker is one of the D known speakers. The task is to determine the speaker to whom the voice sample on the phonogram belongs. This task has significant practical value. A large D leads to the issue of having the recognition system output in a form of quite long list of voice samples, each closely matching the target voice sample. However, most of the known classifiers do not provide theoretically justified outputs, and the voices similarity/dissimilarity degree is not theoretically justified either. So, to justify which voices of the "roughly similar" voices set should be disregarded in further identification, very expensive expert analysis methods are to be used. To effectively address this problem the confidence set approach appears to be quite promising.

Thus, we denote the cardinal number of the alternative speakers set as D. The computational investigation was based on the speech corpus which consisted of voice samples that belonged to 92 different speakers. Therefore, the number of classes (different speakers) D was equal to 92. Every speaker's speech was recorded twice with the same conditions. SNR value of ∼20 dB was guaranteed during recording. These two recordings of the same speaker's speech were made with a few days gap. Thus, the speech corpus that was used for research consisted of 184 phonograms. Phonograms were recorded in different communication channels, including GSM, CDMA and UMTS. In this case every speaker class was presented by only two phonograms of the same speaker containing at least 100 s of the pure speech. In voice biometrics this approach allows to partly compensate high variability of human voice (in different time slots speaker's voice can differ strongly). We introduce special notations for phonograms corresponding to the same speaker, but recorded in different time slots.

The set of these phonograms will be referred to as "speech corpus". The phonogram of speaker θ recorded first is denoted as $Ph_\theta^{(-)}$ and the phonogram recorded later is denoted as $Ph_\theta^{(+)}$. If phonogram Ph with a voice sample of speaker θ was used to calculate sample λ, we denote this as $\lambda\ (\theta \mid Ph)$.

Special voice models for each phonogram were developed a priori and were aimed to search for the target speaker in this speech corpus. The first stage of the voice models building is feature extraction stage. The second stage is the approximation of the probability distribution function of the feature vectors by semi-parametric multivariate probability distribution models, so-called Gaussian Mixture Models (GMM). The realization has taken place in the context of the open source ALIZE Toolkit (ALIZE, 0000) and SPRO (Gravier, 2003) (feature extraction stage). In this paper one kind of the well known speech signal acoustic features was specified: Linear-Frequency Spaced Filterbank Cepstrum Coefficients (LFCC) (Xing & Hansen, 2009). In our case these features are based on 19 linear filter-bank (from 300 to 3400 Hz) derived cepstra. Thus, 19 static and 19 first-order delta coefficients were used, giving the feature order $m = 38$.

Those coefficient vectors are computed on a 20 ms window with 10 ms shift. As previously stated, approximation of the feature vectors probability distribution is based on Gaussian Mixture Model, which has an ability to form smooth approximation to arbitrary-shaped densities. GMM is one of the principal methods of modeling speakers for text-independent speaker identification systems now. Practical use of GMM has proven its high efficiency in solving problems of speaker identification. GMM of speaker s feature vectors distribution is a weighted sum of J components densities and given by the equation:

$$P\left(x \mid \lambda_s\right) = \mathbf{w}_s \mathbf{B}_s^T(x),$$

where x is a random m-vector, $\mathbf{w}_s = \left(w_{s1}, \ldots, w_{sJ}\right) \in R^J$, $\mathbf{B}_s(x) = \left(B_{s1}(x), \ldots, B_{sJ}(x)\right) \in R^J$,

$$\underset{s,i}{\forall} B_{si}(x) = \left((2\pi)^{m/2} \left|\Sigma_{si}\right|^{1/2}\right)^{-1} \exp\left(-\frac{1}{2}(x-\mu_{si})^T \Sigma_{si}^{-1}(x-\mu_{si})\right), \qquad \lambda_s = \left\{(w_{si}, \mu_{si}, \Sigma_{si}) \mid i = 1, J\right\}.$$

In general, diagonal covariance matrices Σ_{si} are used to limit the model size. The model parameters λ_s characterize a speaker's voice in the form of a probabilistic density function. During training, those parameters are determined by the

well-known expectation maximization (EM) algorithm (Blimes & Gentle, 1998). In this experiment approximation value J was equal to 1024. Thus, for speaker identification each speaker is modeled by a GMM and is referred to as his model parameters λ.

The solution of the classification problem was obtained by using the multi-class Support Vector Machine (SVM) method. SVM is a state-of-the-art learning machine based on the structural risk minimization induction principle (Hearst, Dumais, Osman, Platt, & Scholkopf, 1998). The *one-against-all* method was used to decompose the D-class problem into a series of two-class problems and construct $D(D-1)/2$ binary SVM-classifiers, each of which separates one class from all the rest.

Thus, the i-th SVM is trained with all the training examples of the i-th class with positive labels, and all the others — with negative. In this case we have the training set $\Lambda = \{(\lambda_i, y_i)\}_{i=1}^{N}, y_i \in \{1, \ldots, D\}, N = 184$. For each class θ, $\theta \in \Theta$, $\Theta = \{1, \ldots, D\}$ convenient SVM classifier was developed with decision function:

$$f(\theta \mid \lambda) = w_\theta^T \phi(\lambda) + b_\theta = \sum_{i=1}^{N} \alpha_i^p \tilde{y}_i(\theta) K(\lambda_i, \lambda) + b_\theta,$$

where

$$\tilde{y}_i(\theta) = \begin{cases} 1, & \text{if } i = \theta, \\ -1, & \text{if } i \neq \theta. \end{cases}$$

The parameters w_θ^T, b_θ were defined so that

$$\underset{w_\theta, b_\theta, \xi}{\text{minimize}} : L(w_\theta, \boldsymbol{\xi}^\theta) = \begin{cases} 0.5 \left\| w_\theta \right\|^2 + C \sum_{i=1}^{N} \left\langle \xi^\theta \right\rangle_i \\ s.t.\ \tilde{y}_i(\theta) \left(w_\theta^T \phi(\lambda) + b_\theta \right) \geq 1 - \left\langle \xi^\theta \right\rangle_i, \quad \left\langle \xi^\theta \right\rangle_i > 0 \end{cases}$$

$\boldsymbol{\xi}^\theta = \left(\left\langle \boldsymbol{\xi}^\theta \right\rangle_1, \left\langle \boldsymbol{\xi}^\theta \right\rangle_2, \ldots, \left\langle \boldsymbol{\xi}^\theta \right\rangle_N \right) \in R^N$. Here, C is trade-off parameter, so-called soft margin parameter. The effectiveness of SVM is based on the choice of soft margin parameter C. Different values of C were tried and the one with the best cross-validation accuracy was picked. In this practical example, we used the Bhattacharyya divergence to measure the degree of similarity between two probability distributions (GMMs). Following (Chang Huai, Kong Aik, & Haizhou, 2009; Jebara & Kondor, 2003; Kailath, 1967) the Bhattacharyya-kernel $K\left(\lambda_s, \lambda_f\right)$ for two GMMs λ_s and λ_f is given by

$$K\left(\lambda_s, \lambda_f\right) = \frac{1}{8} \sum_{i=1}^{J} \left\{ \left(\mu_{si} - \mu_{fi}\right)^T \left[\frac{\left(\Sigma_{si} + \Sigma_{fi}\right)}{2} \right]^{-1} \left(\mu_{si} - \mu_{fi}\right) \right\} + \frac{1}{2} \sum_{i=1}^{J} \ln \left(\frac{\frac{\left(\Sigma_{si} + \Sigma_{fi}\right)}{2}}{\sqrt{\left|\Sigma_{si}\right| \left|\Sigma_{fi}\right|}} \right) - \frac{1}{2} \sum_{i=1}^{J} \ln \left(w_{si} w_{fi} \right).$$

At the stage of classification, the sample λ corresponded to $\tilde{\theta}(\lambda)$ class when, having $\theta = \tilde{\theta}(\lambda)$, there is a maximum value of function $f\ (\theta \mid \lambda)$ in the set $\Theta = \{1, \ldots, D\}$. That is

$$\tilde{\theta}(\lambda) = \operatorname{Arg} \underset{\theta \in \Theta}{\operatorname{Max}} \left(f(\theta \mid \lambda) \right).$$

Let us denote the index of the class of the sample λ as $\theta^*(\lambda)$. In this case the target set (confidence set) for the value $\theta^*(\lambda)$ will be denoted as $\Xi(\tilde{\theta}(\lambda) \mid c(P_c), \lambda)$. For the speech corpus, used in the experiment, the truth of conditions (1) and (2) of Theorem 1 has been verified by numerical analysis. The Lipschitz constant L was estimated numerically too: $L = \operatorname{Sup}_{p_1 \neq p_2, \lambda} \left(\left\| f(p_1 \mid \lambda) - f(p_2 \mid \lambda) \right\| \left\| p_1 - p_2 \right\|^{-1} \right)$. Thus, for the sample λ, the target set for $\theta^*(\lambda)$ is

$$\Xi(\tilde{\theta}(\lambda) \mid c(P_c), \lambda) = \left\{ \theta \in \{1, \ldots, D\} \mid \left| f(\tilde{\theta}(\lambda) \mid \lambda) - f(\theta \mid \lambda) \right| \leq c(P_c) \right\}.$$

For a class θ, $\theta \in \{1, \ldots, D\}$, the sample $\lambda\left(\theta \mid Ph_\theta^{(+)}\right) \in \Lambda$ corresponding to phonogram $Ph_\theta^{(+)}$ will be denoted as $\left(\lambda_\theta^{(+)}, y_\theta\right)$, and the sample $\lambda\left(\theta \mid Ph_\theta^{(-)}\right) \in \Lambda$ corresponding to phonogram $Ph_\theta^{(-)}$ will be denoted as $\left(\lambda_\theta^{(-)}, y_\theta\right)$. Thus, $\left(\lambda_\theta^{(+)}, y_\theta\right)$ and $\left(\lambda_\theta^{(-)}, y_\theta\right)$ were constructed using phonograms $Ph_\theta^{(+)}$ and $Ph_\theta^{(-)}$, containing voice of the same speaker θ. But all of these phonograms were recorded at different times and under different acoustic conditions. It is very important to note that the phonograms $Ph_\theta^{(+)}$ and $Ph_\theta^{(-)}$ may correspond to different psycho-physiological states of the speaker θ. This factor significantly complicates the speaker identification. Therefore, the problem of speaker identification is an excellent example to demonstrate the efficiency of the proposed algorithm for evaluating decision reliability of the classification problem.

Let the result of the classification be index of the class $\theta^*(\lambda)$. As it was mentioned earlier, the target set for the parameter $\theta^*(\lambda)$ would be denoted as $\Xi^{(-)}\left(\tilde{\theta}(\lambda) \mid c(P_c), \lambda\right)$ (this set was obtained as the result of this numerical experiment). It should be noted that the database used for the numerical study consisted of phonograms of real voices. Some of these voices were quite similar and weakly differed in the feature space. Conversely, some voices were very distinctive (e.g.: a

very high- and very low-pitched), and therefore greatly differed from other voices in the feature space. Roughly speaking, in this case the inequality $|f(\tilde{\theta} \mid \lambda) - f(\theta \mid \lambda)| \leq c(P_c)$ holds for a relatively small number of classes $\theta \in \{1, \ldots, D\}$, so the cardinal number of $\Xi^{(-)}(\tilde{\theta}(\lambda) \mid c(P_c), \lambda)$ is small. On the other hand, some voices differ slightly. Therefore, for these voices inequality $|f(\tilde{\theta} \mid \lambda) - f(\theta \mid \lambda)| \leq c(P_c)$ will hold for a relatively large number of classes $\theta \in \{1, \ldots, D\}$, and the value $|\Xi^{(-)}(\tilde{\theta}(\lambda) \mid c(P_c), \lambda)|$ will be relatively large. Thus, the target sets that were built as a result of numerical experiment for different voices (classes) with non-zero probability may have different cardinalities.

During the numerical experiment target sets were built for each GMM-vector $\lambda(\theta \mid Ph_\theta(k))$, where θ is a number of the class, k is a number of the particular phonogram $Ph_\theta(k)$ within the class θ. For example, the target set corresponding to the GMM-vector $\lambda(\theta \mid Ph_\theta(k))$ will be denoted as $\Xi(\theta \mid c(P_c), \lambda(\theta \mid Ph_\theta(k)))$.

Numerical experiments were carried out as follows:

- classifier training for θ class was based on the set $\Lambda^{(+)} = \{(\lambda_i, y_i)\}_{i=1}^{N} \setminus \left(\lambda_\theta^{(+)}, y_\theta\right), y_i \in \{1, \ldots, D\}$;
- θ class testing was based on the set $\Lambda^{(-)} = \{(\lambda_i, y_i)\}_{i=1}^{N} \setminus \left(\lambda_\theta^{(-)}, y_\theta\right)$, while the phonogram $Ph_\theta^{(+)}$ was partitioned into four phonograms $\left\{Ph_\theta^{(+)}(k) \mid k \in \{1, \ldots, S\}\right\}, S = 4$ of equal durations. For each class θ elements of the set $\boldsymbol{\lambda}(\theta) = \{\lambda(\theta \mid Ph_\theta(k)) \mid k = 1, 2, \ldots, S\}$ were used as test samples. Thus, for each class θ the corresponding target sets $\Xi(\theta \mid c(P_c), \lambda(\theta \mid Ph_\theta(k))), k \in \{1, 2, \ldots, S\}$ were numerically determined.

What could be the requirements to the target set $\Xi(\theta \mid c(P_c), \lambda(\theta \mid Ph_\theta(k)))$? First, the cardinality of the set $\Xi(\theta \mid c(P_c), \lambda(\theta \mid Ph_\theta(k)))$ must be as small as possible, and second, the true value of the class θ should be included in this set. Obviously, in theory it is possible that true value of θ-class would not be included into target set $\Xi(\theta \mid c(P_c), \lambda(\theta \mid Ph_\theta(k)))$ as a result of numerical experiment. In this case, the situation of $\theta \notin \Xi(\theta \mid c(P_c), \lambda(\theta \mid Ph_\theta(k)))$ is assumed absolutely unacceptable from a practical point of view. To accommodate this requirement in the process of numerical experiment, we consider the following function:

$$\chi(\theta, \Xi(\theta \mid c(P_c), \lambda(\theta \mid Ph_\theta(k)))) = \begin{cases} 1, & \text{if } \theta \in \Xi(\theta \mid c(P_c), \lambda(\theta \mid Ph_\theta(k))), \\ 0, & \text{if } \theta \notin \Xi(\theta \mid c(P_c), \lambda(\theta \mid Ph_\theta(k))). \end{cases}$$

Suppose that, as a result of numerical experiment, all of target sets for N classes are built. In this case the parameter $X(N) = 100 \cdot \left(\sum_{\theta=1}^{N} \left(\sum_{k=1}^{S(\theta)} \chi(\theta, \Xi(\theta \mid c(P_c), \lambda(\theta \mid Ph_\theta(k)))) S(\theta)^{-1}\right) N^{-1}\right)$ characterizes the percentage of such cases when true value of a class is included in the appropriate target set. Here, $S(\theta) = |\boldsymbol{\lambda}(\theta)|$.

The greater the value of $X(N)$ is, the better is the quality of the classification accuracy estimation algorithm. Obviously, the ideal value of $X(N)$ is equal to 100%. Furthermore, the parameter

$$\left|\Xi_\theta^{(-)}(c(P_c))\right| = \sum_{k=1}^{S(\theta)} |\Xi(\theta \mid c(P_c), \lambda(\theta \mid Ph_\theta(k)))| \, (S(\theta))^{-1}$$

represents the average power value of $\Xi(\theta \mid c(P_c), \lambda(\theta \mid Ph_\theta(k)))$ target set; the power values correspond to all GMM-vectors $\lambda(\theta \mid Ph_\theta(k))$ of θ-class target set. Besides, the value

$$\left|\boldsymbol{\Xi}^{(-)}(c(P_c))\right| = (N-1)^{-1} \sum_{i=1}^{N-1} \left|\Xi_i^{(-)}(c(P_c))\right|$$

is the average power of $\Xi_\theta^{(-)}(c(P_c))$ sets over all classes $\theta \in \{1, 2, \ldots, D\}$.

The value of $\left|\boldsymbol{\Xi}^{(-)}(c(P_c))\right|$ characterizes the classification accuracy averaged over all classes. Indeed, the smaller the value of $\left|\boldsymbol{\Xi}^{(-)}(c(P_c))\right|$ is, the lower is the average power of target set, and the higher is the classification accuracy.

For a certain class θ let us consider the set $\{\Xi(\theta \mid c(P_c), \lambda(\theta \mid Ph_\theta(k))) \mid \lambda(\theta \mid Ph_\theta(k)) \in \boldsymbol{\lambda}(\theta)\}$, which consists of target sets, corresponding to different $\lambda(\theta \mid Ph_\theta(k)) \in \boldsymbol{\lambda}(\theta)$. In this case the actual number of matching elements of these sets will characterize the algorithm robustness to random variations of samples parameters. Let us denote percentage of matches in sets $\Xi(\theta \mid c(P_c), \lambda(\theta \mid Ph_\theta(i)))$ and $\Xi(\theta \mid c(P_c), \lambda(\theta \mid Ph_\theta(j)))$ by $\pi_\theta(i, j \mid P_c)$. Obviously, $\pi_\theta(i, j \mid P_c) \in (0, 100)$. An ideal situation, of course, is when $\pi_\theta(i, j \mid P_c) = 100\%$, i.e. the sets $\Xi(\theta \mid c(P_c), \lambda(\theta \mid Ph_\theta(i)))$ and $\Xi(\theta \mid c(P_c), \lambda(\theta \mid Ph_\theta(j)))$ are identical. On the contrary, the worst situation is when $\pi_\theta(i, j \mid P_c) = 0\%$. Now, the value of $\pi(P_c)$, characterizing the average uniform robustness over all classes, is to be determined:

$$\pi(P_c) = \sum_{\theta \in \{1, \ldots, D\}} \left(\sum_{i,j \in \{1,2,\ldots,S\}} \pi_\theta(i, j \mid P_c) \right) \left(D \cdot S^2\right)^{-1}.$$

During the numerical experiments we assumed that $\forall_i (S = |\boldsymbol{\lambda}(i)| = 4)$. We will name $\pi(P_c)$ averaged indicator of algorithm robustness.

Table 1 contains results of numerical analysis. Here, P_c is the confidence coefficient, $\left|\boldsymbol{\Xi}^{(-)}(c(P_c))\right|$ is the average cardinality (power) of the target sets over classes $\theta \in \{1, 2, \ldots, D\}$, $X(N)$ is the percentage of such events when true

Table 1
The dependence of the target set cardinality on the coefficient P_c.

P_c (confidence coefficient)	$\left\|\Xi^{(-)}(c(P_c))\right\|$ (average cardinality of the target set)	$X(N)$ (percentage of the events when true value of class indices is included into the corresponding target set) (%)	$\pi\,(P_c)$ (average indicator of algorithm robustness) (%)
0.95	12	100	87
0.90	8	100	89
0.85	5	100	92
0.75	3	100	94

Table 2
The dependence of the target set cardinality on the coefficient P_c.

P_c (confidence coefficient)	$\left\|\Xi^{(-)}(c(P_c))\right\|$ (average cardinality of the target set)	$X(N)$ (percentage of the events when true value of class indices is included into the corresponding target set) (%)
0.95	2	99
0.90	2	95
0.85	1	89
0.75	1	81

value of the θ-class index is included into the corresponding target set $\Xi(\theta \mid c(P_c), \lambda(\theta \mid k))$. In other words, $X(N)$ is the percentage of incidents when $\chi(\theta, \Xi(\theta \mid c(P_c), \lambda(\theta \mid k))) = 1$. The third column contains the values of the indicator $\pi\,(P_c)$.

Those results can be considered promising and encouraging for practical use. Indeed, if we have only 3 (in average) suspects with $P_c = 0.75$ it gives us a good possibility for final identification decision using a very expensive expert methods. Thus, we have a very impressive possibility to improve the efficiency level of speaker identification business processes in many practical cases. As it was expected, $\left\|\Xi^{(-)}(c(P_c))\right\|$ increases with increasing P_c. On the other hand, $\pi\,(P_c)$ decreases comparatively slowly with increasing P_c.

6. Simulation study

Let us consider numerical study of a relatively simple example of multiclass classification, where the number of classes $D = 10$. The training data had been generated by means of two-dimensional Gaussian distributions. In this case the training classes θ, $\theta \in \{1, 2, 3, \ldots, 10\}$, were independently generated from $N\,(m_\theta, \sigma I)$. Here, $m_\theta \in M$,

$$M = \{(0, 0)\,, (0, 1)\,, (0, -1)\,, (1, 0)\,, (-1, 0)\,, (-1, -1)\,, (1, -1)\,, (1, 1)\,, (2, 1)\,, (2, 0)\}\,,$$

and $\sigma = 1, I = \mathrm{diag}(1, 1)$. The sample sizes are as follows: 300 for the training data (100 for every class) and 900 for the testing data (300 for every class). We used Bhattacharyya Kernel-based SVM classifier. The results of this numerical study, shown in Table 2, once again illustrate the effectiveness of the proposed approach.

As we can see, in all of the numerical experiments the value $X(N)/100$ approximately corresponds to P_c. However, $X(N)/100 > P_c$. In other words, the experimentally obtained estimate $X(N)/100$ of probability $\mathbf{P}_{\theta^*}\left(\theta^* \in \Xi\left(\tilde{\theta} \mid c(P_c)\right)\right)$ is somewhat larger than the theoretically predicted value of P_c. This does not contradict the Theorem 1. The inequality $X(N)/100 > P_c$ is based on excessive caution in constructing the target sets $\Xi_i^{(-)}(c(P_c))$ of the proposed method. Following the method, only the first and second moments of the probability distribution of stochastic component $\eta(\cdot)$ were taken into account when constructing $\Xi_i^{(-)}(c(P_c))$. Hence, the experimentally obtained estimate $X(N)/100$ of the probability $\mathbf{P}_{\theta^*}\left(\theta^* \in \Xi\left(\tilde{\theta} \mid c(P_c)\right)\right)$ is significantly larger than the confidence coefficient P_c by magnitude, which is guaranteed by the proposed method. The value of $(X(N)/100 - P_c)$ can be significantly reduced when considering a priori probability distribution of the component $\eta(\cdot)$. In fact, taking into account this a priori information can lead to improved efficiency of the proposed method and therefore represents an interesting topic for further investigations.

7. Conclusion

The guaranteed estimates of the Lipschitz classifier accuracy, suggested in this paper, are primarily designed to be used in case of large number of classes. In this case the classification is represented by a set of hypothetical classes, and this set contains the object of classification with probability of not less than the specified value P_c. Suggested approach has a clear theoretical foundation and is easy to be understood by practitioners, therefore is promising for practical use. A simple rule for estimating the classification efficiency: "the smaller the cardinality of the target set is, the higher is the accuracy of classification" is comprehensible by a broad variety of users. A practical example specified in this paper clearly illustrates the perspectives of practical use of the proposed estimation.

Appendix

Proof of Theorem 1. "X implies Y" is to be denoted as $X \Rightarrow Y$. "event $\omega(1)$ results in event $\omega(2)$" is to be denoted as $\omega(1) \subset \omega(2)$. Additionally, for the specified parameters $\theta_1, \theta_2 \in \Theta$, let us consider the following auxiliary function:

$$F(\theta_1, \theta_2 \mid \theta_2) = \mathbf{E}_{\theta_2}\left(f(\theta_1 \mid \mathbf{X}^{(\theta_2)})\right) - \mathbf{E}_{\theta_2}\left(f(\theta_2 \mid \mathbf{X}^{(\theta_2)})\right).$$

Let us consider the following:

$$\begin{aligned} f(\tilde{\theta}) - f(\theta) &= \mathbf{E}_{\theta^*} f(\tilde{\theta} \mid \mathbf{X}^{(\theta^*)}) + \eta(\tilde{\theta} \mid \mathbf{X}^{(\theta^*)}) - \mathbf{E}_{\theta^*} f(\theta \mid \mathbf{X}^{(\theta^*)}) - \eta(\theta \mid \mathbf{X}^{(\theta^*)}) \\ &= \left(\mathbf{E}_{\theta^*} f(\tilde{\theta} \mid \mathbf{X}^{(\theta^*)}) - \mathbf{E}_{\theta^*} f(\theta \mid \mathbf{X}^{(\theta^*)})\right) + \left(\eta(\tilde{\theta} \mid \mathbf{X}^{(\theta^*)}) - \eta(\theta \mid \mathbf{X}^{(\theta^*)})\right) \\ &= F(\tilde{\theta}, \theta \mid \theta^*) + \left(\eta(\tilde{\theta} \mid \mathbf{X}^{(\theta^*)}) - \eta(\theta \mid \mathbf{X}^{(\theta^*)})\right). \end{aligned}$$

We denote $\Delta\eta\left(\tilde{\theta}, \theta\right) = \left(\eta(\tilde{\theta} \mid \mathbf{X}^{(\theta^*)}) - \eta(\theta \mid \mathbf{X}^{(\theta^*)})\right)$. Following condition 2, we have:

$$\tilde{\theta} \in \Theta : F(\tilde{\theta}, \theta^* \mid \theta^*) = \left(\mathbf{E}_{\theta^*} f(\tilde{\theta} \mid \mathbf{X}^{(\theta^*)}) - \mathbf{E}_{\theta^*} f(\theta^* \mid \mathbf{X}^{(\theta^*)})\right) \leq 0.$$

The value $|F(\tilde{\theta}, \theta^* \mid \theta^*)|$ cannot be zero when the random component $\eta(\tilde{\theta} \mid \mathbf{X}^{(\theta^*)})$ is not zero. In case if it is absent and based on condition 2, we have: $F(\tilde{\theta}, \theta^* \mid \theta^*) = 0$ and $\tilde{\theta} = \theta^*\left(\tilde{\theta} = \text{Arg Max}_{\theta \in \Theta}\left(\mathbf{E}_{\theta^*} f(\tilde{\theta} \mid \mathbf{X}^{(\theta^*)})\right)\right)$. Thus, the following implication is assumed:

$$\left(F(\tilde{\theta}, \theta \mid \theta^*) = 0\right) \Rightarrow \left(\theta = \theta^*\right). \tag{2}$$

The following interpretation is possible:

$$\Xi(\tilde{\theta} \mid c(P_c)) = \left\{\theta \in \Theta \mid \left|F(\tilde{\theta}, \theta \mid \theta^*) + \Delta\eta\left(\tilde{\theta}, \theta\right)\right| \leq c(P_c)\right\}.$$

We denote $\delta = c(P_c) - \Delta\eta\left(\tilde{\theta}, \theta\right)$ and consider the following events:

$$\omega_0 : \left\{\left|\Delta\eta\left(\tilde{\theta}, \theta\right)\right| \leq c(P_c)\right\}, \qquad \omega_1 : \left\{0 < \left|F(\tilde{\theta}, \theta \mid \theta^*)\right| \leq \delta\right\}, \qquad \omega_2 : \left\{F(\tilde{\theta}, \theta \mid \theta^*) = 0\right\},$$
$$\omega_3 : \left\{\theta^* \in \Xi(\tilde{\theta})\right\}.$$

Considering (2) in the set $\Xi(\tilde{\theta} \mid c(P_c))$, the following statement is true:

$$\omega_0 \subset \omega_1 \cup \omega_2 \subset \omega_3. \tag{3}$$

Based on condition 3, we can affirm that:

$$\begin{aligned} \mathbf{E}_{\theta^*}\left(\Delta\eta\left(\tilde{\theta}, \theta\right)\right)^2 &\leq \mathbf{E}_{\theta^*}\left(f\left([\blacksquare] \mid \mathbf{X}^{(\theta^*)}\right) - f\left([\blacksquare] \mid \mathbf{X}^{(\theta^*)}\right)\right)^2 \\ &\leq \int_{-\infty}^{\infty}\left(f\left([\blacksquare] \mid \mathbf{X}^{(\theta^*)}\right) - f\left([\blacksquare] \mid \mathbf{X}^{(\theta^*)}\right)\right)^2 \rho(x \mid [\blacksquare], \theta^*)\mathrm{d}x \\ &\leq \int_{-\infty}^{\infty} L^2 \mathrm{d}^2([\blacksquare]) \cdot \rho(x \mid [\blacksquare], \theta^*)\mathrm{d}x \leq L^2 \mathrm{d}^2[\blacksquare]\int_{-\infty}^{\infty}\rho(x \mid [\blacksquare], \theta^*)\mathrm{d}x \\ &\leq L^2\left(\text{Diam}\,(Z)\right)^2. \end{aligned} \tag{4}$$

Here, $\rho(x \mid [\blacksquare], \theta^*)$ is probability density distribution of the random value $\left(f([\blacksquare] \mid \mathbf{X}^{(\theta^*)}) - f([\blacksquare] \mid \mathbf{X}^{(\theta^*)})\right)$ with fixed parameters [$\blacksquare$] $\theta^* \in \Theta$. Considering condition 1, we have:

$$\mathbf{E}_{\theta^*}\left(\Delta\eta\left(\tilde{\theta}, \theta\right)\right) = 0.$$

In this case, taking into account Chebyshev inequality and item (4), for any $C > 0$, we have:

$$\forall\left(\tilde{\theta}, \theta, \theta^* \in \Theta, C > 0\right) : \mathbf{P}_{\theta^*}\left(\left|\Delta\eta\left(\tilde{\theta}, \theta\right)\right| \leq C\right) > 1 - \frac{\mathbf{E}_{\theta^*}\left(\Delta\eta\left(\tilde{\theta}, \theta\right)\right)^2}{C^2} > 1 - \frac{L^2\left(\text{Diam}\,(Z)\right)^2}{C^2}.$$

Therefore,

$$\forall\left(\tilde{\theta}, \theta, \theta^* \in \Theta\right) : \mathbf{P}_{\theta^*}\left(\left|\Delta\eta\left(\tilde{\theta}, \theta\right)\right| \leq c(P_c)\right) > 1 - \frac{L^2\left(\text{Diam}(Z)\right)^2}{c(P_c)^2}.$$

It is obvious that if

$$c(P_c) = L \cdot \mathrm{Diam}\,(Z)\,(1 - P_c)^{-0.5}$$

the following equality is correct:

$$1 - \frac{L^2\,(\mathrm{Diam}\,(Z))^2}{c(P_c)^2} = P_c.$$

In this case

$$\forall\left(\tilde{\theta}, \theta, \theta^* \in \Theta\right) : \mathbf{P}_{\theta^*}\left(\left|\Delta\eta\left(\tilde{\theta}, \theta\right)\right| \le c(P_c)\right) \ge P_c,$$

and considering (3), we have:

$$\mathbf{P}_{\theta^*}\,(\omega_3) = \mathbf{P}_{\theta^*}\left(\theta^* \in \Xi(\tilde{\theta} \mid c(P_c))\right) \ge P_c.$$

Theorem is proved. □

References

ALIZE: open tool for speaker recognition. Software available at http://www.lia.univ-avignon.fr/heberges/ALIZE/.

Arens, R., & Eells, J. (1956). On embedding uniform and topological spaces. *Pacific Journal of Mathematics*, *6*, 397–403.

Bennett, P. N. (2000). Assessing the calibration of naive Bayes's posterior estimates. *Technical report* CMU-CS-00-155. School of Computer Science, Carnegie Mellon University, 1–8.

Bennett, K. P., & Bredensteiner, E.J. (2000). Duality and geometry in SVM classifiers. In *ICML* (pp. 57–64).

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford, UK: Oxford Univ. Press.

Blake, C. L., & Merz, C. J. (1998). UCI repository of machine learning databases. Irvine: Department of Information and Computer Sciences, University of California. http://www.ics.uci.edu/mlearn/MLRepository.html.

Blimes, J.A., & Gentle, A 1998 Tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *Tech. rep.* 97–021. Berkeley CA: Int'l Computer Science Institute.

You, C. H., Lee, K. A., & Li, H. (2009). A GMM supervector Kernel with the Bhattacharyya distance for SVM based speaker recognition. In *IEEE International conference on acoustics, speech and signal processing* (pp. 4221–4224).

Chen, Di-Rong, Wu, Qiang, Ying, Yiming, & Zhou, Ding-Xuan (2004). Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research*, *5*, 1143–1175.

Gravier, G. (2003). SPRO: a free speech signal processing toolkit (version 4.0.1). http://gforge.inria.fr/projects/spro.

Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems*, *13*(4), 18–28.

Hein, M., & Bousquet, O. (2003). Maximal margin classification for metric space. In M. Warmuth, & B. Scholkopf (Eds.), *Proceedings of the 16th annual conference on computational learning theory* (pp. 72–86). Heidelbergh: Springer Verlag.

Jebara, T., & Kondor, R. 2003 Bhattacharyya and expected likelihood kernels. In *Proceedings 16th annual conference on learning theory*.

Kailath, T. (1967). The divergence and Bhattacharyya distance measures in signal selection. *IEEE Transactions on Communication Technology*, *15*(1), 52–60.

Luxburg, U., & Bousquet, O. (2004). Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research*, *5*, 669–695.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likehood methods. In A. Smola, P. Barlett, B. Scholkopf, & D. Schuurmans (Eds.), *Advances in large margin classifiers* (pp. 61–74). MIT Press.

Vapnik, V., & Chervonenkis, A. (1974). *Pattern recognition theory, statistical learning problems*. Moskva: Nauka.

Wahba, C. (1999). Multivariate functions and operator estimation based of smoothing splines and reproducing kernels. In M. Casdagli, & S. Eubank (Eds.), *Advances in neural information processing systems, Vol. 11*. Cambridge, MA: MIT Press.

Fan, X., & Hansen, J. H. L. (2009). Speaker identification with whispered speech based on modified LFCC parameters and feature mapping. In *IEEE international conference on acoustics, speech and signal processing* (pp. 4553–4556).

Zadrozny, B., & Elkan, C. 2002 Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th international conference on knowledge discovery and data mining* (pp. 694–699).