A Modified Neutral Point Method for Kernel-based Fusion of Pattern-Recognition Modalities with Incomplete Data Sets

Maxim Panov[1], Alexander Tatarchuk[2], Vadim Mottl[2], and David Windridge[3]

[1] Moscow Institute of Physics and Technology,
Institutsky Per. 9, Dolgoprudny, Moscow Region, 141700, Russia
panov.maxim@gmail.com
[2] Computing Center of the Russian Academy of Sciences,
Vavilov St. 40, Moscow, 119991, Russia
aitech@yandex.ru, vmottl@yandex.ru
[3] Center for Vision, Speech and Signal Processing, University of Surrey,
The Stag Hill, Guildford, GU2 7XH, UK
D.Windridge@surrey.ac.uk

Abstract. It is commonly the case in multi-modal pattern recognition that certain modality-specific object features are missing in the training set. We address here the missing data problem for kernel-based Support Vector Machines, in which each modality is represented by the respective kernel matrix over the set of training objects, such that the omission of a modality for some object manifests itself as a blank in the modality-specific kernel matrix at the relevant position. We propose to fill the blank positions in the collection of training kernel matrices via a variant of the Neutral Point Substitution (NPS) method, where the term "neutral point" stands for the locus of points defined by the "neutral hyperplane" in the hypothetical linear space produced by the respective kernel. The current method crucially differs from the previously developed neutral point approach in that it is capable of treating missing data in the training set on the same basis as missing data in the test set. It is therefore of potentially much wider applicability. We evaluate the method on the Biosecure DS2 data set.

1 Introduction

It is well-established that the classification performance of modality-specific classifiers can be improved by combining several different object-representation modalities within a single pattern-recognition procedure. This fusion may be performed at the early or late stage. In the former case of early fusion [1, 2], the growing overall dimensionality of the object representation with increasing the number of modalities can be reduced by incorporating some form of modality-selection within the final classification procedure [3, 4], thereby eliminating the danger of over-fitting. Such modality-selectivity is correlated with the generalization performance of the training process, so that, if performed ideally, the recognition system user is free to include object-representation modalities without constraint.

This freedom creates a new difficulty – the greater the number of modalities employed for comprehensive object representation, the more likely is the omission of some modality-specific feature in the available data.

The problem of missing features has been intensively studied in the pattern recognition literature. However, the aspect of combining diverse pattern-recognition modalities makes special demands on the method of handling blanks in object information.

In [1], three levels of fusing several biometric modalities are compared:

- sensor level, when what is fused are signals acquired immediately from sensors forming different initial object representations;
- classifier score level, that presupposes fusion of scores of multiple classifiers as preliminary decisions made from different modalities to be combined;
- decision level, implying fusion of final decisions made separately by single classifiers on the basis of each modality.

Practically all known methods of compensating for missing data tacitly address the latter two levels of combining modalities [5, 6] and boil down to replacing the missing features via some surrogate values or designing a fusion classifier for all possible combinations of observable features.

At the same time, it is noted in [1] that the sensor level of fusing modalities can potentially yield better results if it is possible to find an appropriate algorithm for combining signals of incomparable physical type. One such algorithm is given in [2] under the assumption that a kernel-based methodology is utilized to obtain a recognition rule for each particular modality, for which a discriminant hyperplane is specified in the linear space associated with each modality. In this case, the kernel trick [11, 12] transforms the problem of combining diverse modalities with missing data into that of appropriately treating blanks in the modality-specific kernel matrices when fusing them into a unified matrix.

Two types of incomplete data samples are to be distinguished – those in the training set, during the classifier learning stage, and those in the test set, when the classifier is already operational.

For the latter case, it was proposed in [8] to adopt the neutral point substitution (NPS) method originally developed in [7] as a means of kernel-based combining of disjoint multi-modal training data, i.e., when only one feature is known for each object. Important advantages of the NPS method are that it is implicitly incorporated into the SVM training framework and it is free from the necessity of inventing any heuristic surrogates for replacing the missing data. It is shown in [8] that the omission of features of the given object at the testing stage is theoretically equivalent, in the case of completely disjoint data sets, to the sum-rule fusion of classifiers within the available modalities [9].

However, the NPS method of treating missing object representations does not lend itself, in its original version, to immediate extension to the training stage (except in the degenerate case of completely disjoint data). The purpose of this paper is to fill in this gap while retaining the advantages of a strictly mathematical approach to the missing-data problem for the case of training sets with a more typical density of blanks.

As the data source for experiments, we use the publicly available biometric database Biosecure DS2 [10].

2 Inferring a modality-specific kernel-based classifier from an unbalanced training set

2.1 Modality-specific kernel functions

Let each real world object $\omega \in \Omega$ be represented by several characteristics (features) measured by respective sensors in sensor-specific scales $x_i(\omega)\colon \Omega \to \mathbb{X}_i$, $i \in I$, where $I = \{1, \ldots, n\}$ is the set of sensors. It is typical in the practice of data analysis that the signals of the initial sensors are of different physical natures and hardly lend themselves to joint treatment. We keep in this paper to the kernel-based approach to combining arbitrary object-representation modalities under the basic assumption that a modality-specific kernel function $K_i\big(x_i', x_i''\big)$ is defined in the output scale of each particular sensor [2].

A kernel is a symmetric two-argument function $K_i\big(x_i', x_i''\big)\colon \mathbb{X}_i \times \mathbb{X}_i \to \mathbb{R}$, which forms a positive semidefinite matrix $\big[K_i\big(x_i(\omega_j), x_i(\omega_l)\big); j, l = 1, \ldots, m\big]$ for each finite collection of objects $\{\omega_j, j = 1, ..., m\}$ [11]. Any kernel $K_i\big(x_i', x_i''\big)$ embeds the scale of the respective sensor $\mathbb{X}_i$ into a hypothetical linear space $\mathbb{X}_i \subseteq \tilde{\mathbb{X}}_i$, in which the null element and linear operations are defined in a particular way [12]:

$$\phi_i \in \tilde{\mathbb{X}}_i, \;\; x_i' + x_i''\colon \tilde{\mathbb{X}}_i \times \tilde{\mathbb{X}}_i \to \tilde{\mathbb{X}}_i, \;\; cx_i\colon \mathbb{R} \times \tilde{\mathbb{X}}_i \to \tilde{\mathbb{X}}_i.$$

The role of inner product is played by the symmetric kernel function itself, which is inevitably bilinear $K_i\big(\alpha' x_i' + \alpha'' x_i'', x_i\big) = \alpha' K_i\big(x_i', x_i\big) + \alpha'' K_i\big(x_i'', x_i\big)$.

The major convenience factor of the kernel-based approach to data analysis is its ability to provide the constructor of a data-analysis system with the possibility of working with objects of arbitrary nature in unified terms of linear functions $f(\omega) = f(x_i(\omega))\colon \Omega \to \mathbb{X}_i \to \mathbb{Y}$, where $\mathbb{Y}$ is any desired linear space. More strictly, the carrier of kernel-specific linear functions is not the feature scale $\mathbb{X}_i$ itself, but rather its linear closure $\mathbb{X}_i \subseteq \tilde{\mathbb{X}}_i \to \mathbb{Y}$.

However, it should be kept in mind that $\tilde{\mathbb{X}}_i$ is thus a hypothetical linear space deriving from the kernel trick, in contrast to its observable subset $\mathbb{X}_i \subseteq \tilde{\mathbb{X}}_i$ which is the output scale of a particular sensor associated with its respective feature $x_i(\omega) \in \mathbb{X}_i$ relating to the set of real-world objects $\omega \in \Omega$.

In particular, to determine a scalar linear function $f_i(x) : \tilde{\mathbb{X}}_i \to \mathbb{R}$, it is enough to specify a direction element (vector, in linear-space terms) $a_i \in \tilde{\mathbb{X}}_i$ and a numerical threshold $b_i \in \mathbb{R}$, then the function will be expressed by the formula $f_i(x|a_i, b_i) = K_i(a_i, x) + b_i$. The equation $f_i(x|a_i, b_i) = K_i(a_i, x) + b_i = 0$ defines a hyperplane which dichotomizes the hypothetical linear space $\tilde{\mathbb{X}}_i$ and, as a consequence, the feature scale $\mathbb{X}_i \subseteq \tilde{\mathbb{X}}_i$ along with the original set of objects:

$$f_i\big(x_i(\omega)|a_i, b_i\big) = K_i\big(a_i, x_i(\omega)\big) + b_i \gtrless 0. \tag{1}$$

The inequality (1) plays the role of a modality-specific kernel-based linear two-class classifier in the set of real-world objects of arbitrary kind.

Before discussing methods of combining diverse modalities of objects represented in a training set with missing measurements, we consider in the next Section the structure of a modality-specific classifier, and introduce the notion of neutral points in the linear closure of the feature scale $\tilde{\mathbb{X}}_i \supseteq \mathbb{X}_i$. This notion will be the main mathematical instrument for filling blanks in the training set.

2.2 A single modality-specific kernel-based classifier inferred from an incomplete training set

Let $\Omega^* = \{(\omega_j, y_j), j = 1, ..., N\}$ be the training set of real-world objects allocated by the trainer between two classes $y_j = y(\omega_j) = \pm 1$. In the case of training incompleteness, the partial set of training information for the subset of objects $\Omega_i^* \subset \Omega^*$, at which the ith modality $i \in I$ is acquired $x_{ij} = x_i(\omega_j) \in \mathbb{X}_i$, will consist of the matrix of available kernel values and class-indices:

$$\Omega_i^* \Rightarrow \{K_i(x_{ij}, x_{il}), y_j;\, \omega_j, \omega_l \in \Omega_i^*\}\,. \tag{2}$$

Perhaps the most widely adopted technique for finding a discriminant hyperplane (1) for a given training set of classified objects represented by a single kernel is the Support Vector Machine (SVM) [11]. The idea underlying the classical SVM for linearly separable training sets is that of finding the discriminant hyperplane which provides the maximum margin between the closest training points of both classes:

$$\begin{cases} K_i(a_i, x_{ij}) + b_i \geqslant \varepsilon_i,\ y_j = 1, \\ K_i(a_i, x_{ij}) + b_i \leqslant -\varepsilon_i,\ y_j = -1, \end{cases}\ \omega_j \in \Omega_i^*,\ 2\varepsilon_i \to \max_{K_i(a_i,a_i)=1} (a_i \in \tilde{\mathbb{X}}_i, b_i \in \mathbb{R}). \tag{3}$$

The attempt to maximize the overall margin between the classes $2\varepsilon_i \to \max$ is what has given rise to the terminology "Support Vector Machine", because the direction vector of the optimal discriminant hyperplane $\hat{a}_i$ obtained as the solution of the optimization problem (3) is completely determined (supported) by the projections of a few number of objects into the modality-specific feature space $\mathbb{X}_i \subseteq \tilde{\mathbb{X}}_i$.

In the more realistic case of a linearly inseparable training set, the normalized form of criterion (3) can be put as

$$\begin{cases} K_i(a_i, a_i) + C_i \sum_{\omega_j \in \Omega_i^*} \delta_{ij} \to \min(a_i \in \tilde{\mathbb{X}}_i,\ b_i \in \mathbb{R}, \delta_{ij} \in \mathbb{R}), \\ y_j \big(K_i(a_i, x_{ij}) + b_i\big) \geqslant 1 - \delta_{ij}, \delta_{ij} \geqslant 0,\ \omega_j \in \Omega_i^*, \end{cases} \tag{4}$$

where coefficient $C_i > 0$ penalizes the shifts δ_{ij} of objects breaking the linear separability of classes [11]. The dual form of this criterion is a quadratic programming problem with respect to modality-specific Lagrange multipliers λ_{ij} at the inequality constraints:

$$\begin{cases} \sum_{\omega_j \in \Omega_i^*} \lambda_{ij} - (1/2) \sum_{\omega_j \in \Omega_i^*} \sum_{\omega_l \in \Omega_i^*} y_j y_l K_i(x_{ij}, x_{il}) \lambda_{ij} \lambda_{il} \to \max, \\ \sum_{\omega_j \in \Omega_i^*} y_j \lambda_{ij} = 0,\ 0 \leqslant \lambda_{ij} \leqslant C_i/2,\ \omega_j \in \Omega_i^*. \end{cases} \tag{5}$$

As the most essential result of training, the solution of the dual problem ($\hat{\lambda}_{ij} \geqslant 0, \omega_j \in \Omega_i^*$) picks out a subset of *support objects* within the modality-specific training set (2):

$$\hat{\Omega}_i = \{\omega_j \in \Omega_i^* : \hat{\lambda}_{ij} > 0\} \subseteq \Omega_i^*. \tag{6}$$

The positive Lagrange multipliers at the support objects ($\hat{\lambda}_{ij} > 0, \omega_j \in \hat{\Omega}_i$) completely determine the values of the variables which optimize (4), first of all, the direction vector and position of the hyperplane:

$$\hat{a}_i = \sum_{\omega_j \in \hat{\Omega}_i} y_j \hat{\lambda}_{ij} x_{ij} \in \tilde{\mathbb{X}}_i, \tag{7}$$

$$\hat{b}_i = \frac{\sum_{\omega_j \in \Omega_i^*, 0 < \hat{\lambda}_{ij} < C/2} \hat{\lambda}_{ij} K_i(\hat{a}_i, x_{ij}) + (C/2) \sum_{\omega_j \in \Omega_i^*, \hat{\lambda}_{ij} = C/2} y_j}{\sum_{\omega_j \in \Omega_i^*, 0 < \hat{\lambda}_{ij} < C/2} \hat{\lambda}_{ij}}. \tag{8}$$

As collateral solutions, the training problem (4) yields also the forced shifts $\hat{\delta}_{ij} \geqslant 0$ of the training objects, but for our purpose there is no need to compute these values.

The direction vector $\hat{a}_i \in \tilde{\mathbb{X}}_i$ of the modality-specific discriminant hyperplane is expressed in (7) as the sum in terms of the hypothetical linear operations defined in $\tilde{\mathbb{X}}_i$ by the modality-specific kernel by virtue of the kernel trick. However, there is no need to compute it explicitly. Substitution of the formal equality (7) into (1) and (8) yields the family of recognition rules immediately applicable to any new object $\omega \in \Omega$ under the only condition that the ith modality $x_i(\omega) \in \mathbb{X}_i$ is completely defined for it, i.e., kernel values $K_i\big(x_{ij}, x_i(\omega)\big)$ are known for all the objects of the training set:

$$\begin{gathered}
\hat{f}_i(\omega | \Omega_i^*, C_i, b_i) = \sum_{\omega_j \in \hat{\Omega}_i} y_j \hat{\lambda}_{ij} K_i\big(x_{ij}, x_i(\omega)\big) + \hat{b}_i \gtrless 0, \\
\hat{b}_i = \frac{\sum\limits_{\omega_j \in \Omega_i^*, 0 < \hat{\lambda}_{ij} < C/2} \hat{\lambda}_{ij} \sum\limits_{\omega_k \in \Omega_i^*, \hat{\lambda}_{ik} > 0} y_k \hat{\lambda}_{ik} K_i(x_{ij}, x_{ik}) + (C/2) \sum\limits_{\omega_j \in \Omega_i^*, \hat{\lambda}_{ij} = C/2} y_j}{\sum\limits_{\omega_j \in \Omega_i^*, 0 < \hat{\lambda}_{ij} < C/2} \hat{\lambda}_{ij}}.
\end{gathered} \tag{9}$$

2.3 Neutral points in the modality-specific linear space of object representation

Let the training set of object representations in terms of the ith modality Ω_i^* (2) be fixed. Suppose the training problem in terms of the ith modality (4)-(5) has been solved, namely, the Lagrange multipliers are known ($\hat{\lambda}_{ij}, \omega_j \in \Omega_i^*$).

This solution determines the optimal discriminant hyperplane in the hypothetical linear closure $\tilde{\mathbb{X}}_i$ of the modality-specific feature scale $\mathbb{X}_i$. Depending on the sign of the decision function (9), the ith modality votes for assigning a new object $\omega \in \Omega$ to the positive or negative class, but a firm decision will be impossible if the object maps exactly to the discriminant hyperplane. For this reason, we call the points of the discriminant hyperplane *neutral points*, using the special symbols $x_{\phi,i} \in \tilde{\mathbb{X}}_{\phi,i}$ to denote them:

$$\tilde{\mathbb{X}}_{\phi,i} = \Big\{ x_{\phi,i} \in \tilde{\mathbb{X}}_i : \sum_{\omega_j \in \hat{\Omega}_i} y_j \hat{\lambda}_{ij} K_i\big(x_{ij}, x_{\phi,i}\big) + \hat{b}_i = 0 \Big\} \subset \tilde{\mathbb{X}}_i. \tag{10}$$

It is clear that $\tilde{\mathbb{X}}_{\phi,i}$ is a set of continuum cardinality. All the neutral points $x_{\phi,i} \in \tilde{\mathbb{X}}_{\phi,i}$ possess the same property of ambiguous class membership (10), but, in what follows, it will be convenient for us to distinguish one of them having the minimum norm:

$$\hat{x}_{\phi,i} = \arg\min_{x_{\phi,i} \in \tilde{\mathbb{X}}_{\phi,i}} K_i(x_{\phi,i}, x_{\phi,i}). \tag{11}$$

In terms of the linear operations in $\tilde{\mathbb{X}}_i$, this point is proportional to the direction vector of the optimal discriminant hyperplane (7) $\hat{x}_{\phi,i} = c_i\hat{a}_i = c_i\sum_{\omega_j\in\hat{\Omega}_i} y_j\hat{\lambda}_{ij}x_{ij}$. The coefficient $c_i \in \mathbb{R}$ is given by the equation $K_i(\hat{a}_i, c_i\hat{a}_i) + \hat{b}_i = c_iK_i(\hat{a}_i, \hat{a}_i) + \hat{b}_i = 0$, whence it follows that $c_i = -\hat{b}_i/K_i(\hat{a}_i, \hat{a}_i)$, and, with respect to (7),

$$\hat{x}_{\phi,i} = \frac{\hat{b}_i}{\sum_{\omega_j\in\hat{\Omega}_i}\sum_{\omega_k\in\hat{\Omega}_i} y_jy_kK_i(x_{ij}, x_{ik})\hat{\lambda}_{ij}\hat{\lambda}_{ik}}\ \sum_{\omega_j\in\hat{\Omega}_i} y_j\hat{\lambda}_{ij}x_{ij}\ \in\ \tilde{\mathbb{X}}_i. \quad (12)$$

The neutral points (12) (or more exactly, the coefficients of their representation as linear combinations of object features in the hypothetical linear spaces $\tilde{\mathbb{X}}_i$), are additional results of training from the incomplete modality-specific training sets $(\Omega_i^*, i \in I)$, that contain only those objects in the entire training set Ω^* for which the respective modality is defined. The central idea behind harnessing such values for joint training with respect to all of the modalities is using $\hat{x}_{\phi,i}$ instead of missed actual values of the respective modality-specific features for incompletely represented objects. Such a strategy of replacing missed feature values is then free of arbitrary assumptions regarding the nature of the original natural data set.

3 Fusing pattern-recognition modalities at the training stage for incomplete data

3.1 The principle of additive kernel fusion

We will call the union of all modality-specific training sets $\Omega^* = \bigcup_{i\in I}\Omega_i^*$ (2) over all the available modalities $I = \{1, \ldots, n\}$ the *unified training set*. We shall say the unified training set Ω^* is full if each object $\omega_j \in \Omega^*$ is represented by all modality-specific signals $\big(x_{ij} = x_i(\omega_j) \in \mathbb{X}_i, i \in I\big)$, i.e., all the kernel-specific training sets coincide $\Omega_1^* = ... = \Omega_n^*$.

A full training set Ω^* allows for immediate combination of the various modalities by kernel fusion. It is enough to define an appropriate combined kernel (inner product) $K(\boldsymbol{x}', \boldsymbol{x}'')$, $\boldsymbol{x} = (x_i, i \in I) \in \tilde{\mathbb{X}}$, in the Cartesian product $\tilde{\mathbb{X}} = \tilde{\mathbb{X}}_1 \times ... \times \tilde{\mathbb{X}}_{n=|I|}$ of the linear spaces $\tilde{\mathbb{X}}_i \supseteq \mathbb{X}_i$ defined by the respective kernels. In particular, the sum of the initial kernels $K(\boldsymbol{x}', \boldsymbol{x}'') = \sum_{i\in I} K_i(x_i', x_i'')$ will be a kernel in $\tilde{\mathbb{X}}$. From this point of view, any choice of a point $\boldsymbol{a} = (a_i \in \tilde{\mathbb{X}}_i, i \in I) \in \tilde{\mathbb{X}}$ and real number $b \in \mathbb{R}$ yields a discriminant hyperplane $\hat{f}(\omega|\Omega^*) = K\big(\boldsymbol{a}, \boldsymbol{x}(\omega)\big) + b = \sum_{i\in I} K_i\big(a_i, x_i(\omega)\big) + b \gtrless 0$ with direction vector $\boldsymbol{a}$ in the Cartesian product $\tilde{\mathbb{X}}$, and produces, thereby, a kernel fusion technique. However, just as in the case of a single kernel, there is no need to implicitly evaluate the hypothetical direction vector which exists only in terms of the kernel trick.

The straightforward application of the SVM training principle (4)-(12) to the Cartesian product of the particular linear spaces $\boldsymbol{x}_j = (x_{ij}, i \in I) \in \tilde{\mathbb{X}} = \tilde{\mathbb{X}}_1 \times ... \times \tilde{\mathbb{X}}_n$, $\omega_j \in \Omega^*$, results in the dual training problem, in which $C > 0$ is the penalty coefficient on the shifts of objects that break the linear separability of the training set in $\tilde{\mathbb{X}}$:

$$\begin{cases} \sum_{\omega_j \in \Omega^*} \lambda_j - (1/2) \sum_{\omega_j \in \Omega^*} \sum_{\omega_l \in \Omega^*} y_j y_l \Big(\sum_{i \in I} K_i(x_{ij}, x_{il}) \Big) \lambda_j \lambda_l \to \max, \\ \sum_{\omega_j \in \Omega^*} y_j \lambda_j = 0,\ 0 \leqslant \lambda_j \leqslant C/2,\ \omega_j \in \Omega_i^*. \end{cases} \tag{13}$$

This quadratic programming problem over Lagrange multipliers ($\lambda_j, \omega_j \in \Omega^*$) has the same structure as that for a single modality (5). The only difference is that the training set occurs in (13) through kernels $K(\boldsymbol{x}_j, \boldsymbol{x}_l) = \sum_{i \in I} K_i(x_{ij}, x_{il})$ in the unified linear space $\tilde{\mathbb{X}}$ instead of single modality-specific kernels $K_i(x_{ij}, x_{il})$ (5) in the particular spaces $\tilde{\mathbb{X}}_i$.

Let the training set be not full, i.e., such that each object ω_j is, in general, represented by only a fraction of the modalities $x_{ij} \in \mathbb{X}_i$, $i \in I_j \subseteq I$. Then, there are objects ω_j not represented by certain of the modalities (x_{ij}, $i \notin I_j$), and the respective kernel values adjacent to these objects are unknown in the dual criterion (13) $\big(K_i(x_{ij}, x_{il}) =?,\ i \notin I_j,\ \omega_l \in \Omega^*\big)$.

The central idea for overcoming the problem of incomplete data at the training stage outlined in the next Section, is that of substituting the neutral values $\hat{x}_{\phi,i}$ for the missing modalities x_{ij}.

3.2 Neutral point substitution for missing representations of training objects

Let the SVM be applied within each modality-specific partial training set Ω_i^* (Section 2.2). Then, the sets of support objects $\hat{\Omega}_i$ along with Lagrange multipliers ($\hat{\lambda}_{ij}, \omega_j \in \Omega_i^*$) and biases of discriminant hyperplanes $\hat{b}_i$ are found for all the modalities $i \in I$ in accordance with (5), (6) and (12). As a result, the hypothetical neutral points $\hat{x}_{\phi,i} \in \tilde{\tilde{\mathbb{X}}}_i$ are defined by (12) as linear combinations of modality-specific object features.

Thus, it is possible to compute the neutral-point substitutes for missing values of kernels in (13):

$$\begin{aligned} &K_i(x_{ij}, x_{il}) \Leftarrow K_i(\hat{x}_{\phi,i}, x_{il}),\ i \notin I_j,\ \omega_l \in \Omega^*; \\ &K_i(\hat{x}_{\phi,i}, x_{il}) = \frac{\hat{b}_i \sum_{\omega_k \in \Omega_i^*} y_k \hat{\lambda}_{ik} K_i(x_{ik}, x_{il})}{\sum_{\omega_k \in \hat{\Omega}_i} \sum_{\omega_q \in \hat{\Omega}_i} y_k y_q K_i(x_{ik}, x_{iq}) \hat{\lambda}_{ik} \hat{\lambda}_{iq}}. \end{aligned} \tag{14}$$

The Lagrange multipliers ($\lambda_j, \omega_j \in \Omega^*$), found as a solution of the dual problem (13) after such a substitution, determine, on the full application of the sequence of computations (6)-(12), first, the set of support objects $\hat{\Omega} = \{\omega_j \in \Omega^* : \hat{\lambda}_j > 0\}$, then the bias of the discriminant hyperplane $\hat{b}$, and finally yield the decision rule

$$\hat{f}(\omega | \Omega^*, C) = \sum_{\omega_j \in \hat{\Omega}} y_j \hat{\lambda}_j \sum_{i \in I} K_i\big(x_{ij}, x_i(\omega)\big) + \hat{b} \gtrless 0, \tag{15}$$

which is the result of fusing the available pattern-recognition modalities taking into account the final imbalance of the incomplete training set Ω^*.

4 Experiments: Biometric-based identity authentication from incomplete data

To demonstrate the above principle experimentally, we employ the Biosecure database [10], derived from a European project whose aim is to integrate multidisciplinary research efforts in biometric-based identity authentication.

We randomly chose a total of 333 different individuals from the database, with distinct identities $Z = \{z = 1, ..., 333\}$. Each of them is represented by four time-spaced measurements $v_i^t(z) \in \mathbb{V}_i$, $t = 0, 1, 2, 3$, of eight modalities $i \in I = \{1, ..., n\}$, $n = 8$, where $\mathbb{V}_i$ is the scale for measuring the ith modality:

- two versions of the frontal face image (high- and low-resolution ones from, respectively, professional and web camera), $v_i^t(z) \in \mathbb{V}_i$, $i = 1, 2$,
- six fingerprints of the right hand (optical and thermal imprints of the index finger, middle finger and thumb) $v_i^t(z) \in \mathbb{V}_i$, $i = 3, ..., 8$.

The Cartesian product of all the modality-specific measurement scales will be denoted as $\mathbb{V} = \mathbb{V}_1 \times ... \times \mathbb{V}_n$. However, not all of potential measurements $\{v_i^t(z), z \in Z, i \in I, t = 0, 1, 2, 3\}$ are available in the data base. Approximately one fourth of them have missing constituents, but for each of the chosen persons $z \in Z$ at least one of the measurement sets, let it be $t = 0$, is full, i.e., all the modalities $(v_i^0(z), i \in I)$ are properly represented in the data base, and neither of the remaining sets $(v_i^{1,2,3}(z), i \in I)$ is completely missed.

In the experiments, we used this full set $\boldsymbol{v}^0(z) = (v_i^0(z), i \in I) \in \mathbb{V}$ as the personal template of person $z \in Z$, whereas the remaining three sets

$$\boldsymbol{v}^{1,2,3}(z) = (v_i^{1,2,3}(z), i \in I) \in \mathbb{V}, \tag{16}$$

some of whose elements may be missing, served as his/her independent representation in the experiments. Let symbols $V_i(z) = \{v_i^t(z), t = 1, 2, 3\} \subset \mathbb{V}_i$ and $V_i = \bigcup_{z \in Z} V_i(z) \subset \mathbb{V}_i$ stand, respectively, for the set of the representations of person z in terms of the ith modality and the total set of such representations of all the persons involved in the experiments.

Thus, we distinguish here between the people's identities $z \in Z$ and the three times greater number of their multi-modal computer representations $\boldsymbol{v}^t(z) = (v_i^t(z), i \in I) \in V = \bigcup_{z \in Z} V(z) \subset \mathbb{V}$, $V(z) = V_1(z) \times ... \times V_n(z) \subset \mathbb{V}$, $t = 1, 2, 3$.

To constitute the total set of real-world pattern-recognition objects $\omega \in \Omega$, we choose the set of pairs

$$\Omega = \{\omega = (\boldsymbol{v}^t(z), \tilde{z})\} = \mathbb{V} \times \mathbb{V} \times \mathbb{V} \times Z, \tag{17}$$

where $\boldsymbol{v}^t(z) = (v_i^t(z), i \in I)$ is one of the three received representations of a person $t = 1, ..., 3$, and $\tilde{z}$ is its claimed identity, which may be true or false. We shall say that object $\omega = (\boldsymbol{v}(z), \tilde{z})$ belongs to the class of clients $y = 1$ if the identity claim is correct $z = \tilde{z}$, and to the class of impostors $y = -1$ in the case of a fraudulent claim $z \neq \tilde{z}$.

For each modality $i \in I$, a real-valued similarity measure $S_i(v_i', v_i'') : \mathbb{V}_i \times \mathbb{V}_i \to \mathbb{R}$ is defined in the Biosecure database. It appears natural to measure the credibility of the identity claim $\tilde{z}$ in the received pair $\omega = (\boldsymbol{v}(z), \tilde{z}) = ((v_i, i \in I), z)$ from the viewpoints of different modalities $i \in I$ as the real-valued modality-specific features $x_i(\omega) = S_i(v_i(z), v_i^0(\tilde{z})) \in \mathbb{R}$. In this case, the natural modality-specific kernel is dot product of feature values $K_i(\omega', \omega'') = K_i(x_i(\omega'), x_i(\omega'')) = K_i(x_i', x_i'') = x_i' x_i''$.

We thus used information on 333 persons (person identities) $Z = \{z = 1, ..., 333\}$, each of which is represented by three independent sets of multi-modal

measurements $\boldsymbol{v}^{1,2,3}(z)$ in accordance with (16). All in all, we have $999 \times 333 = 332667$ pairs of person representations and person identities $\omega = (\boldsymbol{v}(z), \tilde{z})$, which is the size of the full set of objects $|\Omega| = 332667$ (17) in the experiments.

From the part of the full data set Ω, which contains only complete person representations $\left(v_i^t(z), i \in I\right)$, we chose the fixed test set consisting of 20962 objects, namely, pairs <complete person representation/claimed identity>. From the rest of Ω, containing complete as well as incomplete person representations, we further randomly chose 500 training sets each consisting of 200 pairs with the correct claimed identity $y = 1$ and 800 incorrectly claimed pairs $y = -1$. On average, one fourth of 1000 objects in each of the random training sets were incompletely represented, i.e., about 250 of them had at least one missing value in the feature vector.

The goal of the experiment is to show that filling-in blanks in the multi-modal training sets by the generalized neutral-point technique improves generalization performance of the inferred recognition rule in comparison with other methods of imputation. We compared the SVM-NPS technique outlined in this paper with the following five SVM-based methods of treating blanks in the training data:

- SVM handling only objects represented by all the features, in our case, about 3/4 of the training set (SVM-Full);
- sum-rule of combining single SVM-based modality-specific classifiers, inferred each from the partial training subset containing only objects for which the respective modality is known (SVM-SumRule);
- SVM handling all the objects with replacing the unknown features by their averaged known values over the entire training set (SVM-OverallMean);
- the same with replacing the unknown features by their averaged known values over the objects of the same class (SVM-ClassSpecificMean);
- the same with replacing the unknown features by their averaged known values over 5 nearest neighboring objects in the feature space (SVM-5NN).

The interpretation of training results was based on computing the Equal Error Rate of the direction vector of the discriminant hyperplane inferred by each of the six techniques under comparison from each of the 500 random training sets and applied to the test set. The EER value of the respective technique was further averaged over all the training sets.

The following table summarizes the averaged EERs in percentages for the imputation methods under comparison starting with our SVM-NPS:

SVM-NPS	SVM-Full	SVM-SumRule	SVM-OverallMean	SVM-ClassSpecificMean	SVM-5NN
1.07	1.93	1.93	1.38	1.92	1.85

As we can see, the SVM-NPS approach shows almost two times better performance than the SVM-based learning from the training set consisting only of objects with the complete set of features and the SVM-based sum-rule of combining modality-specific classifiers. All other imputation methods are also far outperformed.

5 Conclusions

In this paper, we have set out to generalize the previous neutral point method for accommodating missing data within multi-modal kernel fusion problems in

order to accommodate arbitrary amounts of missing *training* (as opposed to test) data. By using imbalance-sensitive SVM methods, we have shown that the SVM-NPS approach to multi-modal pattern-recognition with incomplete data displays exceptionally good generalization performance as compared to the sum rule fusion of modality-specific classifiers, and to the known SVM-based methods of missing-data imputation. Future experimental study will set out to determine the full bounds of its practical applicability.

Acknowledgements The research leading to these results has received funding from the Russian Foundation for Basic Research, Grant No. 08-01-00695-a. We also gratefully acknowledge the support of EPSRC through grant EP/F069626/1.

References

1. A. Ross, A.K. Jain. Multimodal biometrics: An overview. *Proc. of the 12th European Signal Processing Conference (EUSIPCO)*, 2004, Vienna, Austria, pp. 1221-1224.
2. V. Mottl, A. Tatarchuk, V. Sulimova, O. Krasotkina, O. Seredin. Combining pattern recognition modalities at the sensor level via kernel fusion. *Proc. of the 7th International Workshop on Multiple Classifier Systems, Prague, Czech Republic, May 23-25, 2007*. Lecture Notes in Computer Science, Vol. 4472, 2007, pp. 1-12.
3. A. Tatarchuk, V. Sulimova, D. Windridge, V. Mottl, M. Lange. Supervised selective combining pattern recognition modalities and its application to signature verification by fusing on-line and off-line kernels. *Proc. of the 8th International Workshop on Multiple Classifier Systems, Reykjavik, Iceland, June 10-12, 2009*. Lecture Notes in Computer Science, Vol. 5519, 2009, pp. 324-334.
4. A. Tatarchuk, E. Urlov, V. Mottl, D. Windridge. A support kernel machine for supervised selective combining of diverse pattern-recognition modalities. *Proc. of the 9th International Workshop on Multiple Classifier Systems, Cairo, Egypt, April 7-9, 2010*. Lecture Notes in Computer Science, Vol. 5997, 2010, pp. 165-174.
5. K. Nandakumar, A.K. Jain, A. Ross. Fusion in multibiometric identification systems: What about the missing data? *Proc. of the 3rd IAPR/IEEE International Conference on Biometrics, Alghero, Italy, June 2009*.
6. N. Poh, T. Bourlai, J. Kittler, at all. Benchmarking quality-dependent and cost-sensitive score-level multimodal biometric fusion algorithms. *IEEE Trans. on Information Forensics and Security*, 2009, Vol, 4, No. 4, pp. 849866.
7. D. Windridge, V. Mottl, A. Tatarchuk, A. Eliseyev. The neutral point method for kernel-based combination of disjoint training data in multi-modal pattern recognition problem. *Proc. of the 7th International Workshop on Multiple Classifier Systems, Prague, Czech Republic, May 23-25, 2007*. Lecture Notes in Computer Science, Vol. 4472, 2007, pp. 13-21.
8. N. Poh, D. Windridge, V. Mottl, A. Tatarchuk, A. Eliseyev. Addressing missing values in kernel-based multimodal biometric fusion using neutral point substitution. *IEEE Trans. on Information Forensics and Security*, 2010, Vol. 5, Issue 3, 461-469.
9. J. Kittler, M. Hatef, R. P. W. Duin, J. Matas. On combining classifiers. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1998, Vol. 20, pp. 226-239.
10. N. Poh, T. Bourlai, J. Kittler. A multimodal biometric test bed for quality-dependent, cost-sensitive and client-specific score-level fusion algorithms. *Pattern Recognition*, Vol. 43, Issue 3, 3/2010, pp. 1094-1105, http://www.biosecure.info/.
11. V. Vapnik. *Statistical Learning Theory*. John-Wiley & Sons, Inc., 1998.
12. V. Mottl. Metric spaces admitting linear operations and inner product. *Doklady Mathematics*, 2003, Vol. 67, No. 1, pp. 140–143.