The Neutral Point Method for Kernel-Based Combination of Disjoint Training Data in Multi-Modal Pattern Recognition

David Windridge[1], Vadim Mottl[2], Alexander Tatarchuk[2], Andrey Yeliseyev[3]

[1] University of Surrey, Guildford, UK
D.Windridge@surrey.ac.uk

[2] Computing Center of the Russian Academy of Sciences, Moscow, Russia
vmottl@yandex.ru, aitech@yandex.ru

[3] Moscow Institute of Physics and Technology, Moscow, Russia
andrey_eliseev@7ka.mipt.ru

Abstract. Multiple modalities present potential difficulties for kernel-based pattern recognition in consequence of the lack of inter-modal kernel measures. This is particularly apparent when training sets for the differing modalities are disjoint. Thus, while it is always possible to consider the problem at the classifier fusion level, it is conceptually preferable to approach the matter from a kernel-based perspective. By interpreting the aggregate of disjoint training sets as an entire data set with missing inter-modality measurements to be filled in by appropriately chosen substitutes, we arrive at a novel kernel-based technique, the *neutral-point method*. On further theoretical analysis, it transpires that the method is, in structural terms, a kernel-based analog of the well-known sum rule combination scheme. We therefore expect the method to exhibit similar error-canceling behavior, and thus constitute a robust and conservative strategy for the treatment of kernel-based multi-modal data.

1 Introduction

In data analysis and, in particular, pattern recognition, it is common practice to employ the term "modality" when speaking about a specific kind of mathematical computer-perceptible object representation. In terms of the measured modality, the hypothetical set of "all" real-world objects $\omega \in \Omega$ is represented by the output of the respective sensor $x(\omega) \in \mathbb{X}$ in the form of signals, images, or, in relatively rare simple cases, in the form of one-dimensional numerical features.

The essence of the training problem in supervised pattern recognition is extrapolation of the information contained in the finite training set of the accessible objects $\Omega^* = (X,Y) = \{x(\omega_j) \in \mathbb{X},\ y(\omega_j) \in \mathbb{Y} = \{1,...,m\},\ \omega_j \in \Omega^*\}$ onto the entire scale of the re-

This work is supported by the Russian Foundation for Basic Research, Grants 05-01-00679, 06-01-08042, 06-07-89249, and INTAS Grant 04-77-7347.

spective object representation $\hat{y}\left(x_1(\omega),...,x_n(\omega)\right): \mathbb{X}_1 \times ... \times \mathbb{X}_n \to \mathbb{Y}$. The intention of increasing the generalization performance of the resulting recognition rule has led to the concept of multimodal systems, which combine several object representation modalities $\left\{x_i(\omega) \in \mathbb{X}_i, i=1,...,n\right\}$ into a unified recognition procedure $\hat{y}\left(x_1(\omega),...,x_n(\omega)\right)$: $\mathbb{X}_1 \times ... \times \mathbb{X}_n \to \mathbb{Y}$.

In the overview of multimodal biometrics given in [1]two principle levels of combining modalities are distinguished: the *signal level*, when, prior to training the classifier $\hat{y}(\omega)$, a unified representation of objects is formed $x(\omega)=\varphi\left(x_i(\omega), i=1,...,n\right)$ to combine all the particular modalities $\hat{y}(\omega)=\hat{y}\left(x(\omega)\right)$, and the *classifier level*, when what is to be combined are the classifiers $\hat{y}(\omega)=\gamma\left[\hat{y}_i\left(x_i(\omega)\right), i=1,...,n\right]$, each trained individually by a single modality.

Until recently, most attention had been paid in the literature to principles of classifier fusion [2,3], because it was assumed that combining modalities of different character (real numbers and labels, for example) is not straightforward. However, recent achievements in the methodology of kernel fusion [4,5,6,7,8] have cleared the way for combining any number of modalities at the signal level.

The aim of this paper is to consider relationships between the two approaches to multimodal machine learning, kernel fusion and classifier fusion, under the specific assumption that the problem to be solved is that of two-class pattern recognition, and that, in addition, the kernel-based approach is applied within each modality.

Before closely scrutinizing the relationship between kernel and classifier fusion, we consider the specificity of a single modality-specific kernel-based classifier. As applied to the kernel-based approach, the principle of classifier fusion implies combining several recognition rules inferred from modality-specific data. In this paper, on the basis of the kernel fusion methodology considered in [8], we propose a unified view on the seemingly different principles of combining modalities at the signal and classifier level by, respectively, kernel and classifier fusion.

2 The modality-specific kernel-based classifier

A two-argument symmetric function $K_i(x_i', x_i'') = K_i(x_i, x_i')$ defined in the output scale of a particular sensor $\mathbb{X}_i = \left\{x_i(\omega), \omega \in \Omega\right\}$ is said to be kernel function in $\mathbb{X}_i$ if it forms positive semidefinite matrices $\left[K_i\left(x_i(\omega_j), x_i(\omega_l)\right); j,l=1,...,k\right]$ for all finite subsets of this set [9]. Any kernel $K_i(x_i', x_i'')$ embeds the scale of the respective sensor $\mathbb{X}_i$ into a hypothetical linear space with inner product $\tilde{\mathbb{X}}_i \supseteq \mathbb{X}_i$, in which the null element $\phi_i \in \tilde{\mathbb{X}}_i$ and linear operations $x_i' + x_i'' : \tilde{\mathbb{X}}_i \times \tilde{\mathbb{X}}_i \to \tilde{\mathbb{X}}_i$ and $\alpha x_i : \mathbb{R} \times \tilde{\mathbb{X}}_i \to \tilde{\mathbb{X}}_i$ are defined in a special way. The role of the inner product is played by the kernel function

$K_i(x_i', x_i)'$ [10] which will be linear with respect to its arguments $K_i(\alpha' x_i' + \alpha\ x_i'', x_i)' = \alpha\, K(x_i', x_i)' + \alpha\ K(x_i'', x_i)$.

Thus, in terms of a single modality, a training set $\Omega_i^* = \{\omega_j,\ j=1,...,N_i\}$ is completely represented by the kernel matrix and class-indices of objects $y_j = y(\omega_j) = \pm 1$:

$$\Omega_i^* \Rightarrow \left\{\mathbf{K}_i = \left[K_i\left(x_i(\omega_j), x_i(\omega_l)\right),\ \omega_j, \omega_l \in \Omega_i^*\right],\ y(\omega_j),\ \omega_j \in \Omega_i^*\right\}. \tag{1}$$

In addition, it is required to uphold the ability to compute the kernel values $K_i\left(x_i(\omega), x_i(\omega_j)\right)$ for any new real-world object $\omega \in \Omega$ and all the objects $\omega_j \in \Omega_i^*$ represented in the training set.

A commonly adopted kernel-based approach to the two-class pattern recognition problem is widely known under the name of Support Vector Machine (SVM) [9]. The main concept of this approach is that of the optimal discriminant hyperplane in the linear space $\tilde{\mathbb{X}}_i$ produced by the respective kernel $\hat{y}_i\left(x_i(\omega)\right) = K_i\left(\vartheta_i, x_i(\omega)\right) + b_i \gtrless 0$. In our terms, the discriminant hyperplane is defined by a hypothetical element of this linear space $\vartheta_i \in \tilde{\mathbb{X}}_i$ and by the threshold $b_i \in \mathbb{R}$. The SVM training criterion follows from the idea of maximizing the margin between the points of two classes in $\tilde{\mathbb{X}}_i$:

$$\begin{cases} K_i(\vartheta_i, \vartheta_i) + C\sum_{\omega_j \in \Omega_i^*} \delta_j \rightarrow \min\left(\vartheta_i \in \tilde{\mathbb{X}}_i, b \in \mathbb{R}, \delta_j \in \mathbb{R}\right), \\ y_j\left[K\left(\vartheta_i, x_i(\omega_j)\right) + b\right] \geq 1 - \delta_j,\ \delta_j \geq 0,\ \omega_j \in \Omega_i^*. \end{cases} \tag{2}$$

where $C > 0$ is sufficiently large coefficient. The dual form of this criterion is a quadratic programming problem with respect to the nonnegative Lagrange multipliers $\lambda_j \geq 0$ for the inequality constraints:

$$\begin{cases} \sum_{\omega_j \in \Omega_i^*} \lambda_j - (1/2)\sum_{\omega_j \in \Omega_i^*}\sum_{\omega_l \in \Omega_i^*}\left[y_j y_l K_i\left(x_i(\omega_j), x_i(\omega_l)\right)\right]\lambda_j \lambda_l \rightarrow \max, \\ \sum_{\omega_j \in \Omega_i^*} y_j \lambda_j = 0,\ 0 \leq \lambda_j \leq C/2,\ \omega_j \in \Omega_i^*. \end{cases} \tag{3}$$

The direction vector of the optimal discriminant hyperplane is the linear combination of the training-set objects with coefficients defined by the Lagrange multipliers found as the solution of this problem $\hat{\vartheta}_i = \sum_{\omega_j \in \Omega_i^*} y_j \hat{\lambda}_{i,j} \omega_j$. It must be kept in mind that the training-set objects occur in this linear combination as elements of the hypothetical linear space $\tilde{\mathbb{X}}_i \supseteq \mathbb{X}_i$ in accordance with the specific linear operations produced by the kernel $K_i(x_i', x_i)'$.

The objects $\omega_j \in \Omega_i^*$ whose Lagrange multipliers are positive in the solution of the dual problem $\hat{\lambda}_{i,j} > 0$ make the subset of support objects in the full training set:

$$\hat{\Omega}_i = \left\{\omega_j \in \Omega_i^* : \hat{\lambda}_{i,j} > 0\right\} \subseteq \Omega_i^*. \tag{4}$$

Only the support objects will form the direction vector of the optimal discriminant hyperplane

$$\hat{f}_i\left(x_i(\omega)\right)=\sum_{j:\omega_j\in\hat{\Omega}_i} K_i\left(\hat{\vartheta}_i, x_i(\omega)\right)+\hat{b}_i \gtrless 0 \Rightarrow \hat{y}_i\left(x_i(\omega)\right)=\pm 1,\ \hat{\vartheta}_i=\sum_{\omega_j\in\hat{\Omega}_i} y_j\hat{\lambda}_{i,j}\omega_j\ , \quad (5)$$

and only the kernel matrix at the support objects will affect the recognition rule inferred from the training set of the respective modality:

$$\begin{aligned}&\hat{f}_i\left(x_i(\omega)\right)=\sum_{j:\omega_j\in\hat{\Omega}_i} y_j\hat{\lambda}_{i,j}K_i\left(x_i(\omega_j), x_i(\omega)\right)+\hat{b}_i \gtrless 0,\\ &\hat{b}_i=-\left(\sum_{j:\omega_j\in\hat{\Omega}_i}\hat{\lambda}_{i,j}\sum_{l:\omega_l\in\hat{\Omega}_i} y(\omega_l)\hat{\lambda}_{i,l}K_i\left(x_i(\omega_j), x_i(\omega_l)\right)\right)\Big/\sum_{j:\omega_j\in\hat{\Omega}_i}\hat{\lambda}_{i,j}\Big).\end{aligned} \quad (6)$$

So, the result of training within the bounds of a single modality is completely represented by the subset of support objects and the positive values of Lagrange multipliers at them (4).

We introduce here a new notion, which will be especially important for the comparison of kernel fusion and classifier fusion. If a new object maps into a point strictly at the discriminant hyperplane $\hat{y}_i\left(x_i(\omega)\right)=0$ (6), it cannot be attributed to any one of the two classes. All these point will be said to be neutral points produced by the training set and denoted them by special symbol $\hat{x}_{\phi,i}$. It is obvious that there exists a continuum of neutral points for each modality $\tilde{\mathbb{X}}_{\phi,i}$ in the respective space $\tilde{\mathbb{X}}_i$:

$$\hat{x}_{\phi,i}\in\tilde{\mathbb{X}}_{\phi,i},\ \ \tilde{\mathbb{X}}_{\phi,i}=\left\{x_i\in\tilde{\mathbb{X}}_i : K_i(\hat{\vartheta}_i, x_i)+\hat{b}_i=0\right\},\ \ \hat{b}_i=-K_i(\hat{\vartheta}_i, x_{\phi,i})\,. \quad (7)$$

3 Kernel fusion: Combining modalities at the signal level from a full training set by kernel fusion

Let, at least, one kernel be defined in the output scale of each of several sensors $K_i(x_i', x_i'')$, $x_i', x_i''\in\mathbb{X}_i$, $i=1,...,n$. The union of all the modality-specific training sets $\Omega^*=\bigcup_{i=1}^n\Omega_i^*$ (1) will be called 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 the modality-specific signals $\boldsymbol{x}(\omega_j)=\left(x_i(\omega_j)\in\mathbb{X}_i, i=1,...,n\right)$, i.e., all the kernel-specific training sets coincide $\Omega_1^*=...=\Omega_n^*$.

A full training set Ω^* allows for immediate combination of several modalities by kernel fusion. All the known kernel fusion techniques are based on the idea of constructing an appropriate combined kernel (inner product) $K(\boldsymbol{x}', \boldsymbol{x}'')$, $\boldsymbol{x}=(x_1,...,x_n)\in\tilde{\mathbb{X}}$, in the Cartesian product $\tilde{\mathbb{X}}=\tilde{\mathbb{X}}_1\times...\times\tilde{\mathbb{X}}_n=\left\{\boldsymbol{x}=(x_i\in\tilde{\mathbb{X}}_i, i=1,...,n)\right\}$ of the linear spaces $\tilde{\mathbb{X}}_i\supseteq\mathbb{X}_i$ defined by the respective kernels. The sum of the initial kernels $K(\boldsymbol{x}', \boldsymbol{x}'')=\sum_{i=1}^n K_i(x_i, x_i')$ will retain all the properties of inner product, i.e., be a kernel in $\tilde{\mathbb{X}}$. From this point of view, any choice of a point

$\boldsymbol{\vartheta} = (\vartheta_i \in \tilde{\mathbb{X}}_i, i = 1,...,n) \in \tilde{\mathbb{X}}$ and real number $b \in \mathbb{R}$ yields a discriminant hyperplane with direction vector in the Cartesian product $\tilde{\mathbb{X}}$

$$\hat{f}(\boldsymbol{x}(\omega)) = \hat{f}(x_i(\omega), i = 1,...,n) = K(\boldsymbol{\vartheta}, \boldsymbol{x}(\omega)) + b = \sum_{i=1}^{n} K_i(\vartheta_i, x_i(\omega)) + b \gtrless 0, \quad (8)$$

and produces, thereby, a kernel fusion technique.

It is apparent that if the norm $\sqrt{K_i(\vartheta_i, \vartheta_i)}$ of a component of the direction vector $\vartheta_i \in \tilde{\mathbb{X}}_i$ is small in its linear space, the respective kernel $K_i(x_i', x_i')'$ will little affect the recognition rule (8).

The straightforward application of the SVM training principle to the Cartesian product of the particular linear spaces $\tilde{\mathbb{X}} = \tilde{\mathbb{X}}_1 \times ... \times \tilde{\mathbb{X}}_n$ [9], namely, finding the optimal discriminant hyperplane in $\tilde{\mathbb{X}}$ with respect to the full training set Ω^*, results in the training criterion

$$\begin{cases} \sum_{i=1}^{n} K_i(\vartheta_i, \vartheta_i) + C \sum_{\omega_j \in \Omega^*} \delta_j \to \min\left(\vartheta_i \in \tilde{\mathbb{X}}_i, b \in \mathbb{R}, \delta_j \in \mathbb{R}\right), \\ y_j \left[\sum_{i=1}^{n} K_i(\vartheta_i, x_i(\omega_j)) + b\right] \geq 1 - \delta_j, \ \delta_j \geq 0, \ \omega_j \in \Omega^*. \end{cases} \quad (9)$$

This optimization problem leads to the dual quadratic programming problem of the analogous structure as the usual SVM dual problem (3):

$$\begin{cases} \sum_{\omega_l \in \Omega^*} \lambda_j - (1/2) \sum_{\omega_j \in \Omega^*} \sum_{\omega_l \in \Omega^*} \left(y_j y_l \sum_{i=1}^{n} K_i(x_i(\omega_j), x_i(\omega_l))\right) \lambda_j \lambda_l \to \max, \\ \sum_{\omega_j \in \Omega^*} y_j \lambda_j = 0, \ 0 \leq \lambda_j \leq C/2, \ \omega_j \in \Omega^*. \end{cases} \quad (10)$$

The Lagrange multipliers obtained for the set of support objects

$$\hat{\Omega} = \left\{\omega_j \in \Omega^* : \hat{\lambda}_j > 0\right\} \subseteq \Omega^* \quad (11)$$

yield the optimal recognition rule:

$$\begin{aligned} &\hat{f}(\boldsymbol{x}(\omega)) = \hat{y}(x_1(\omega),...,x_n(\omega)) = \sum_{\omega_j \in \hat{\Omega}} y_j \hat{\lambda}_j \sum_{i=1}^{n} K_i(x_i(\omega_j), x_i(\omega)) + \hat{b} \gtrless 0, \\ &\hat{b} = -\left(\sum_{\omega_j \in \hat{\Omega}} \hat{\lambda}_j \sum_{\omega_l \in \hat{\Omega}} y_l \hat{\lambda}_l \sum_{i=1}^{n} K_i(x_i(\omega_j), x_i(\omega_l))\right) \Big/ \left(\sum_{\omega_j \in \hat{\Omega}} \hat{\lambda}_j\right). \end{aligned} \quad (12)$$

This is the simplest but not the only possible way of kernel fusion. The quasi-statistical approach to the signal-level modality combination considered in [8] covers the main kernel fusion principles known at present.

With the objective function in (9) as $\sum_{i=1}^{n} (1/r_i) K_i(\vartheta_i, \vartheta_i) + C \sum_{\omega_j \in \Omega^*} \delta_j \to \min\left(\vartheta_i \in \tilde{\mathbb{X}}_i, r_i \in \mathbb{R}, b \in \mathbb{R}, \delta_j \in \mathbb{R}\right)$ under additional constraint $\prod_{i=1}^{n} r_i = 1$, the training criterion displays a tendency to suppressing the weights at the "redundant" kernels $\hat{r}_i \to 0$ along with emphasizing $\hat{r}_i \gg 0$ the kernels which are "adequate" to the trainer's data, and, so, results in soft extraction of a relatively small number of most adequate kernels without full suppression of the others. Due to this property, this training technique is called in [8] the Relevance kernel Machine (RKM).

If $\sum_{i=1}^{n}\sqrt{K_i(\vartheta_i,\vartheta_i)}+C\sum_{\omega_j\in\Omega^*}\delta_j\rightarrow\min\left(\vartheta_i\in\tilde{\mathbb{X}}_i, b\in\mathbb{R}, \delta_j\in\mathbb{R}\right)$ is taken as the objective function in (9), the training technique selects a subset of support kernels $\hat{I}\subseteq\{1,...,n\}$ with positive norms of the direction vectors $\left(K_i(\hat{\vartheta}_i,\hat{\vartheta}_i)>0, i\in\hat{I}\right)$ in contrast to the others which get completely suppressed $\left(K_i(\hat{\vartheta}_i,\hat{\vartheta}_i)=0, i\notin\hat{I}\right)$ [8]. Because only the support kernels $i\in\hat{I}$ participate in the recognition rule, this is a kind of Support Kernel Machine (SKM) first considered in [5].

All of these approaches to kernel fusion techniques are closely related to the problem of studying the relationship between kernel and classifier fusion as alternative strategies for combining pattern recognition modalities at, respectively, signal and classifier level. However, in this paper we restrict our consideration only to the simplest kernel fusion technique (9).

4 The neutral point method of combining modalities from disjoint training sets

It is common practice that particular modalities are employed by different expert groups, which hence derive their training sets independently of each other. If it is so, the training set $\Omega^*=\left\{\omega_j,\ j=1,...,N\right\}$ will consist of disjoint subsets $\Omega^*=\bigcup_{i=1}^{n}\Omega_i^*$, $\Omega_i^*\bigcap\Omega_l^*=\varnothing$, such that the output signals of only one modality-specific sensor $\left(x_i(\omega_j),\omega_j\in\Omega_i^*\right)$ are captured within the bounds of each of them.

With respect to this notation, the kernel fusion criterion (9) may be put in the following equivalent form:

$$\begin{cases}\sum_{i=1}^{n}\left(K_i(\vartheta_i,\vartheta_i)+C\sum_{\omega_j\in\Omega_i^*}\delta_j\right)\rightarrow\min\left(\vartheta_i\in\tilde{\mathbb{X}}_i, b\in\mathbb{R}, \delta_j\in\mathbb{R}\right),\\ \left[y_j\left(K_i\left(\vartheta_i,x_i(\omega_j)\right)+\sum_{l=1,l\neq i}^{n}K_l\left(\vartheta_l,x_l(\omega_j)\right)+b\right)\geq 1-\delta_j,\ \delta_j\geq 0,\ \omega_j\in\Omega_i^*\right], i=1,...,n.\end{cases} \tag{13}$$

Here, in each group of constraints at the training-set objects $\left\{\omega_j\in\Omega_i^*, i=1,...,n\right\}$, for any value of the abstract variable $\vartheta_i\in\mathbb{X}_i$, only one of n summands is defined, namely, $K_i\left(\vartheta_i,x_i(\omega_j)\right)$, whereas the other summands $K_l\left(\vartheta_l,x_l(\omega_j)\right)$ are not, because the sensor signals $x_l(\omega_j)$ are unknown for $l\neq i$ due to the assumption that the particular training sets are disjoint.

We hence propose a new method of combining modalities in supervised kernel-based pattern recognition in the case when the training sets for different modalities are disjoint. The idea consists in treating the problem (13) as that of learning with incomplete data and filling-up the unknown actual values of the sensor signals corresponding to other modalities $x_l(\omega_j)$, $l\neq i$, by one common value being the arbitrary neu-

tral point $\hat{x}_{\phi,i} \in \tilde{\mathbb{X}}_{\phi,i}$ (7) of the i th linear space. After this substitution, the problem (13) takes the following form:

$$\begin{cases} \sum_{i=1}^{n}\left(K_i(\vartheta_i,\vartheta_i)+C\sum_{\omega_j\in\Omega_i^*}\delta_j\right)\to\min\left(\vartheta_i\in\tilde{\mathbb{X}}_i, b\in\mathbb{R}, \delta_j\in\mathbb{R}\right), \\ \left[y_j\left(K_i\left(\vartheta_i, x_i(\omega_j)\right)+\sum_{l=1,l\neq i}^{n}K_l\left(\vartheta_l,\hat{x}_{\phi,l}\right)+b\right)\geq 1-\delta_j,\ \delta_j\geq 0,\ \omega_j\in\Omega_i^*\right], i=1,\ldots,n. \end{cases} \tag{14}$$

Theorem. The solution of the optimization problem (13) is the totality of the optimal direction elements in the linear spaces of generalized features $\hat{\vartheta}_i \in \tilde{\mathbb{X}}_i$ (5) found as the solutions of the training problems (2) independently for each modality $i=1,\ldots,n$, along with the common threshold value equal to the sum of optimal thresholds for all modalities $\hat{b}=\sum_{i=1}^{n}\hat{b}_i$ (6).

Hence, replacement of the unknown actual values of sensor signals by the neutral points of the respective linear spaces leads to the discriminant function (5)

$$\hat{f}\left(x_i(\omega), i=1,\ldots,n\right)=\sum_{i=1}^{n}K_i\left(\hat{\vartheta}_i, x_i(\omega)\right)+\sum_{i=1}^{n}\hat{b}_i=\sum_{i=1}^{n}\left[K_i\left(\hat{\vartheta}_i, x_i(\omega)\right)+\hat{b}_i\right]\gtrless 0\,. \tag{15}$$

Here the expressions in brackets are nothing other than the discriminant functions built independently for each modality $\hat{f}_i\left(x_i(\omega)\right)$, thus,

$$\hat{f}\left(x_i(\omega), i=1,\ldots,n\right)=\sum_{i=1}^{n}\hat{f}_i\left(x_i(\omega)\right). \tag{16}$$

So, the approach to filling-in the missing values of sensor signals we have adopted leads to the indicated recognition rule which, in structural terms, is a technique for combining particular classifiers, namely, by summation of particular discriminant functions. The neutral point method should therefore exhibit the error-canceling properties associated with classifier combination and should hence be a robust and safe approach to kernel-based classification of disjoint data sets.

An analogous technique of combining classifiers is known in the literature under the name of Sum Rule [2]. The distinction consists in that the known Sum Rule method is based on the assumption of the probabilistic output of the particular classifiers in the form of posterior class-membership probabilities $p_i^{(k)}\left(x_i(\omega)\right)$, $\sum_{k=1}^{m}p_i^{(k)}\left(x_i(\omega)\right)=1$. The combination principle consists in computing the unified posterior probabilities of classes by way of summing over the particular posterior probabilities: $p^{(k)}\left(x_i(\omega), i=1,\ldots,n\right)=(1/n)\sum_{i=1}^{n}p_i^{(k)}\left(x_i(\omega)\right)$. When there are only two classes $m=2$, the posterior probabilities at the output of the i th classifier are completely determined by the posterior probability of one of the classes:

$$p\left(x_i(\omega), i=1,\ldots,n\right)=(1/2)\sum_{i=1}^{n}p_i\left(x_i(\omega)\right).$$

The analogy between our classifier fusion rule (16) and the Sum Rule is immediately apparent.

5 Discussion

At first inspection, it is hardly possible to encompass a generic way of combining modalities because of the vast variety of possible object representations. However, the kernel-fusion approach converts, in a natural way, different modalities into a unified mathematical language of inner products in linear spaces, which in fact makes such a comparison realistic. This is so even if the original modalities are not themselves vector or scalar quantities: the only scalar constraint is that of the kernel itself. This is hence particularly necessary in situations in which only relative distance measures are available, such as Genomics.

Thus in the simplest case, the sensor signals might have the form of scalar numerical features, i.e. be real numbers. What we have customarily done when learning in a multidimensional linear space, which is the Cartesian product of several real-valued axes, is thus nothing other than combining several modalities via a form of kernel fusion.

However, any purely kernel-based fusion method exhibits difficulties when the differences in modality are accompanied by differences in training set composition: in this case straightforward kernel-fusion will not suffice. This difficulty is also apparent for conventional classification: in fact it is clear that application of the classifier fusion principle is an inescapable necessity in the case of disjoint training subsets contained within disjoint modalities, since multiple decision confidences are the only quantities available for combining in a meaningful manner.

We have hence, by making certain conservative assumptions about the 'missing' kernel values, derived a *neutral point method* for addressing the above difficulty in a Kernel-based context. However, it transpires that the neutral point method has *itself* the exact structural form of a classifier combination scheme (in fact the Sum Rule decision scheme).

At its purest level, though, the principle of combining modalities with disjoint training sets via classifier fusion is based on the assumption that the modalities are *independent* (that is, for decision problems in which the individual modalities cannot be straightforwardly taken to define a composite Cartesian product space in which classification can take place). The principle of kernel-fusion, on the other hand, is not attached to this assumption: the fact that it becomes equivalent to one particular combination scheme under the neutral point assumption for missing data should not therefore be taken as significant for combination in general, but rather for the Sum Rule, *specifically*.

The fact that the Sum Rule combination scheme also exhibits ideal error-canceling properties [2] is thus a significant bonus, and a considerable further reason for advocating the neutral point method.

6 Conclusions

We have set out to address the difficulties that multiple modalities and disjoint training sets represent for kernel-based pattern recognition due to their absence of intra-modal kernel information. Though possible to consider the problem at the classifier fusion level, we have motivated our work on the basis of the conceptual preferability of addressing the issue from a purely kernel-specific perspective. Hence, by interpreting the aggregate of disjoint training sets as complete data-sets with missing inter-modality measurements that can be substituted by appropriately-chosen values, we have arrived at a novel classification technique, which we have named the *neutral-point method.* We proceeded to theoretically demonstrate that the neutral-point method is a kernel-based analog of the well-known sum rule combination scheme. It is thus capable of error-cancellation, and gives strong backing for our assertion that the neutral-point choice of replacements for inter-modality measurements is a conservative and safe one.

References

1. Ross A., Jain A.K. Multimodal biometrics: An overview. *Proceedings of the 12th European Signal Processing Conference (EUSIPCO), 2004*. Vienna, Austria, pp. 1221-1224.
2. Kittler J., Hatef M., Duin R., Matas J. On combining classifiers. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1998, Vol. 20, No. 3, pp. 226-239.
3. *Multiple Classifier Systems. Proceedings of the 1st - 6th International Workshops:* Lecture Notes in Computer Science, Springer, 2001, 2002, 2003, 2004, 2005.
4. Lanckriet G.R.G., Cristianini N., Ghaoui L.E., Bartlett P., Jordan M.I. Learning the kernel matrix with semidefinite programming. *J. Machine Learning Research*, 5, 2004, pp. 27–72.
5. Bach F.R., Lankriet G.R.G., Jordan M.I. Multiple kernel learning, conic duality, and the SMO algorithm. *Proceedings of the 21th International Conference on Machine Learning*, Banff, Canada, 2004.
6. Sonnenburg S., Rätsch G., Schäfer C. A general and efficient multiple kernel learning algorithm. *Proceedings of the 19th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, December 5-8, 2005.
7. Mottl V., Krasotkina O., Seredin O., Muchnik I. Principles of multi-kernel data mining. *Proceedings of the 4th International Conference on Machine Learning and Data Mining in Pattern Recognition*, Leipzig, 2005, 52-61.
8. Mottl V., Tatarchuk A., Seredin O., Krasotkina O., Sulimova V. Combining pattern recognition modalities at the sensor level via kernel fusion. *Submitted to the 7th International Workshop on Multiple Classifier Systems.*
9. Vapnik V. *Statistical Learning Theory*. John-Wiley & Sons, Inc. 1998.
10. Mottl V. Metric spaces admitting linear operations and inner product. *Doklady Mathematics*, 2003, Vol. 67, No. 1, pp. 140–143.