Joint Diagonalization of Kernels for Information Fusion

Alberto Muñoz and Javier González

Universidad Carlos III de Madrid, c/ Madrid 126, 28903 Getafe, Spain
{alberto.munoz,javier.gonzalez}@uc3m.es

Abstract. Information Fusion is becoming increasingly relevant in fields such as Image Processing or Information Retrieval. In this work we propose a new technique for information fusion when the sources of information are given by a set of kernel matrices. The algorithm is based on the joint diagonalization of matrices and it produces a new data representation in an Euclidean space. In addition, the proposed method is able to eliminate redundant information among the input kernels and it is robust against the presence of noisy variables and irrelevant kernels.

The performance of the algorithm is illustrated on data reconstruction and classifications problems.

Keywords: Information Fusion, Approximate Joint Diagonalization, Kernel Methods, Support Vector Machines.

1 Introduction

Fusion information techniques are becoming increasingly relevant in different fields such as classifier combination [9] or image processing [17]. Data fusion processes combine different sources of information to feed some data processing algorithm. For instance, in the problem of kernel combination [4], there are several metrics available and the task is to produce a single kernel to increase the classification performance of Support Vector Machine algorithms. In image fusion [3], a typical problem considers different satellite pictures, with different resolutions and different color qualities, and the task is to produce a picture that has maximum resolution and the best color quality. In the field of Information Retrieval, the goal can be to classify a set of web pages [8], and the information that has to be combined lies in the co-citation matrix and in the terms-by-documents matrix.

In this paper we approach the problem of information fusion in the context of kernel methods. Consider a set of kernels $K_1, ..., K_t$. By the Mercer theorem [11] each positive-definite kernel K_i induces a transformation of the data set into a (possibly) high dimensional Euclidean space $\mathbb{R}^{n_i}$. Thus, each kernel induces a particular representation of the data set using some basis $\{v_i\}$ for $\mathbb{R}^{n_i}$. If we want to combine the information provided by a set of kernels, we will have to find some 'common' basis $\{v^*\}$ from the individual basis $\{v_i\}$, such that the inmersion

L. Rueda, D. Mery, and J. Kittler (Eds.): CIARP 2007, LNCS 4756, pp. 556–563, 2007.

of the data set in the resulting $\mathbb{R}^{n^*}$ contains all the relevant information from the individual kernels K_i.

Any technique to produce the desired combination basis needs to take into account the problem of information redundance. To illustrate this problem, let us consider a data set, and two representations given by two projections on two pairs of principal axes: (x, y) and (x, z), where the x variable is present in both representations. If we use the direct sum of the corresponding spaces as solution for the combination problem, we will have the representation (x, y, x, z). Thus, the weight of the x variable will be doubled when using the Euclidean distance and the results of the classification and regression algorithms will be distorted. In the general case the correlation between the variables will cause similar problems.

The Joint Diagonalization (JD) is a procedure that can be applied for fusion information purposes. The basis $\{v_i\}$ for the individual representation spaces are given by the eigenvectors of the K_i matrices. JD is able to produce a new basis $\{v_i^*\}$ from the $\{v_i\}$ basis and provides information to weight the new variables. Redundant kernel information can be removed during the process and the problem of overweighting variables avoided.

The paper is organized as follows. In Section 2 we review the simultaneous diagonalization process and introduce the case for more than two kernels. In Section 3 a new algorithm for kernel fusion is presented based on the joint diagonalization of matrices. Finally, in Section 4 the performance of the new data fusion methodology is tested using an ilustrative example.

2 Joint Diagonalization of Matrices

The calculus of eigenvalues is an usual task in many pattern recognition algorithms such as FDA [10], Kernel PCA [13,1], or Kernel Canonical Correlations [6] among others. Given a matrix $A \in \mathbb{R}^{n \times n}$ the diagonalization process seeks matrices V orthogonal and D diagonal such that $AV = VD$, or equivalently:

$$A = VDV^T. \tag{1}$$

When A is symmetric then a solution always exists and the elements of D are real numbers.

Some algorithms require the simultaneous diagonalization of two matrices. For instance, in FDA the *within-class* scatter matrix and the *between-class* scatter matrix have to be simultaneously diagonalized to find discriminative directions.

It is well known that exact simultaneous diagonalization is always possible [12]. This problem is referenced in the literature as the Generalized Eigenvalue Problem. Given two matrices $A, B \in \mathbb{R}^{n \times n}$ the problem is stated as finding $V \in \mathbb{R}^{n \times n}$, and two diagonal matrices D_1 and D_2 such that $AV = BVD$. In other terms,

$$\begin{aligned} V^T A V &= D_1 \\ V^T B V &= D_2. \end{aligned} \tag{2}$$

The base of vectors given by the columns of V is not necessarily orthonormal. This base is not unique and it is proven that V is orthogonal when the matrices A and B conmute, that is when $AB = BA$. If B is non-singular, the problem can be solved as and ordinary eigenvalue problem where the target matrix is $B^{-1}A$. See [7,5] and references therein for further details.

Next we afford the problem of diagonalization of more than two matrices at the same time.

2.1 Approximate Joint Diagonalization Algorithm

Given a set of matrices $S = \{A_1, ..., A_t\}$ it is not possible in general to achieve perfect joint diagonalization in a single step, unless $A_i A_j = A_j A_i \; \forall i, j \in \{1, ..., t\}$. These restrictions do not hold for most theoretical or practical problems. In practice we will have to find an orthonormal change of basis which makes the matrices in S 'as diagonal as possible' in a sense that will be detailed right away.

In this paper we consider the Approximate Joint Diagonalization (AJD) of symmetric matrices [14,2,15]. Given a square matrix A, we can measure the deviation of A from diagonality by defining

$$off(A) = \|A - diag(A)\|_F^2 = \sum_{i \neq j} a_{ij}^2, \tag{3}$$

where $\|A\|_F = \sum_i \sum_j a_{ij}^2$ is the Frobenius norm. If A is a diagonal matrix then $off(A) = 0$, while $off(A)$ will take small positive values when the off-diagonal values of A are close to zero.

Given the set S, the target is to find an orthonormal matrix V such that the departure from diagonality of the transformed matrices $D_i' = V^T A_i V$ is as small as possible $\forall i \in \{1, ..., t\}$. Therefore the goal will be to minimize

$$\begin{aligned} J(V) = & \textstyle\sum_{k=1}^{t} off(V^T A_k V) \\ s.t. & \\ & \|V^T V - I\|_F = 0 \\ & \|diag(V - I)\|_F = 0, \end{aligned} \tag{4}$$

where the restrictions have to be included to achieve orthonormality and to avoid the trivial solution $V = 0$. After solving (4) we will obtain quasi diagonal matrices $D_1', ..., D_t'$, where $D_i' = V^T A_i V \; \forall i \in \{1, ..., t\}$.

There is no closed solution for the problem in (4) and some type of numerical approach has to be adopted. We will apply the algorithm described in [2,16]. The idea is to generate a sequence of similarity transformations of the initial matrices that drive to zero the off-diagonal entries. The convergence of the algorithm is proven to be quadratic and the obtained eigenvalues and eigenvectors are robust against small perturbations of the data.

3 Fusion Joint Diagonalization Algorithm (FJDA)

As already mentioned, Approximate Joint Diagonalization involves the computation of a base of orthogonal vectors in which the set of kernels approximately diagonalize. We will obtain relevant information about the data structure by analyzing the resulting eigenvalues, or equivalently, the diagonal matrices obtained from the joint diagonalization procedure. The ideas are similar to that used in Principal Components Analysis, where the covariance matrix is diagonalized and the resulting eigenvalues can be interpreted as the weights of the new variables.

Let $\{v_1, ..., v_n\}$ be the column vectors of the matrix V obtained from the JD algorithm (the $\{v_i^*\}$ vectors in the introduction). These vectors constitute the basis where both kernels diagonalize and can be interpreted as the *average eigenspace* of the kernels. A detailed analysis of the kernels redundancy can be done in terms of the values of the diagonal matrices $D'_1, D'_2, ..., D'_t$ obtained. Given the kernel K_l, their components can be interpreted as follows:

- $D'_l(i,i) = 0$: the vector v_i is irrelevant for the kernel K_l. That is, the i-th variable v_i is in the null space of K_l.
- $D'_l(i,i) \neq 0$: in this case v_i is a relevant component for K_l.
- $D'_l(i,j)$: These values can be interpreted as the interactions among the new variables. Due to the JD operation, $D'_l(i,j) \approx 0$.

Given V and $D'_1, D'_2, ..., D'_t$, the straightforward sum of the kernel matrices can be reexpressed as:

$$\sum_{i=1}^{t} K_i = V^T \left(\sum_{i=1}^{t} D'_i \right) V \tag{5}$$

Given that the *off-diagonal values* of $\{D'_1, ..., D'_t\}$ are quite close to zero, $D'_l(i,i)$ can be interpreted as the weight that kernel K_l assigns to the i-th variable in the new basis. Since the new base is orthogonal, independent information is given by each component. The straightforward sum of kernels implies to include redundances in the operation and to overweight variables that appear in more than one kernel at the same time. In order to avoid these redundances, the sum of the quasi-diagonal matrices of expression (5) can be replaced by the function $F(D'_1, D'_2, ..., D'_t)$ defined as follows:

$$F(D'_1, D'_2, ..., D'_t) = \begin{cases} max\{D'_1(i,j), ..., D'_t(i,j)\} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{6}$$

The justification of this choice is as follows. The relevance of the $i-th$ variable in the basis induced by kernel K_l is given by $D'_l(i,i)$. The use of the max function guarantees that the i-th variable will be relevant in the resulting combined basis if this is the case for any of the individual representations. Thus, the weight of ith variable in the fusion kernel will be $max\{D'_1(i,i), ..., D'_t(i,i)\}$.

The final algorithm for kernel fusion is shown in Table 1 and it provides a global framework for kernel fusion. Notice that, since the matrix V is orthogonal and the diagonal matrices of $F(D'_1, D'_2, ..., D'_t)$ are positive, K^* is always a Mercer kernel matrix.

Table 1. Scheme of the Fusion Joint Diagonalization Algorithm in three steps

INPUT: Kernel matrices $K_1, ..., K_n$ OUTPUT: Kernel combination K^*
1.- $(V, D'_1, ..., D'_n) = AJD(K_1, ..., K_n)$ 2.- $D^* = F(D'_1, ..., D'_n)$ 3.- $K^* = V^T D^* V$

4 Experiments

In order to validate the effectiveness of the proposed methodology some experimental results are shown in this section. First, the algorithm is tested in a data reconstruction example where partial information about the data is given. Finally, the methodology is successfully tested in a real classification problem..

4.1 Simulated Example

In this example we illustrate the performance of the new JD algorithm in a data structure recovery task.

We consider two different one-dimensional random projections π_1 and π_2 of the spiral data in Figure 1 and calculate the kernel matrices K_1 and K_2 by applying the linear kernel $k(x, y) = x^T y$ to the projected data points, that is, $K_i(x, y) = \pi_i(x)^T \pi_i(y)$. We add a corrupted (random) representation of the data and calculate K_3 from this representation in the same way. K_3 plays the role of a non informative (non-related) piece of information in the system. This situation happens when the distance function is not appropiate for the data set under consideration or when we try to use irrelevant information to solve a problem. The task is to recovery the original data set from the three projections.

Two fusion schemes were compared in the experiment: The straightforward sum of kernels $K^{sum} = K_1 + K_2 + K_3$ and the combination K^* calculated with the Fusion Joint Diagonalization Algorithm. In Figure 2 the results are shown. It is clear that our procedure is able to recover the original data set structure while the straightforward sum of kernels fails on the task of recovering the data set structure.

4.2 Sonar Data

In this example we perform a study of classification of sonar signals [18]. The goal is to discriminate between two types of signals: those bounced off a metal cylinder and those bounced off a roughly cylindrical rock. The data set has 208 observations measured on 60 variables that take values in the interval $[0, 1]$. Each value represents the *energy within a particular frequency band, integrated over a certain period of time*. The goal is classify the objects as rocks or mines.

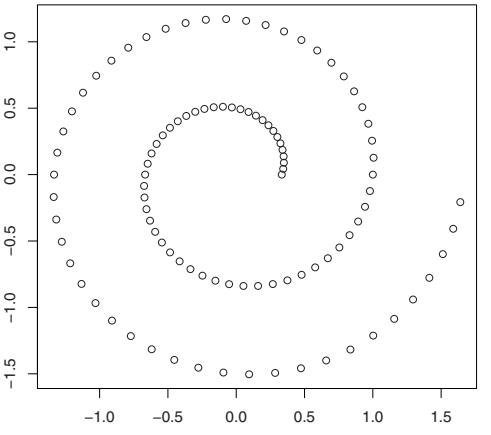
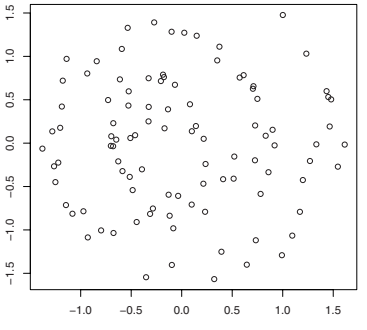
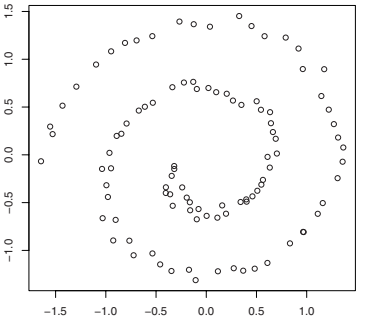


Fig. 1. Spiral data



(a) Direct fusion of kernels for the Spiral data



(b) Fusion Joint Diagonalization Algorithm applied to the Spiral data set

Fig. 2. Representations for recovered data structures after the direct combination of three kernels and after the Fusion Joint Diagonalization algorithm

We consider two Radial Basis Funtion kernels $K_i(x, y) = e^{-\gamma_i \|x-y\|^2}$, $i \in \{1, 2\}$, where $\gamma = 1$ and $\gamma = 0.1$. We want to combine K_1 and K_2 using the straightforward sum and the AJD fusion method. In order to evaluate the performance of both fusion approaches we will feed one SVM classifier with the resulting fusion kernels. The penalty value C is set to one in all the experiments. Table 2 shows the classification results for the SVM classifier using four different kernels: the individual kernels K_1 and K_2, and the two fusion kernels: Sum for the straightforward sum and KAJD for the AJD kernel.

It is apparent from the results that K_1 performs better than K_2. When the straightforward sum is considered, the performance of the SVM is worse than in the case of using the RBF kernel with $\gamma = 1$. It seems that the bad performance

Table 2. Percentage of missclassied data, and percentage of support vectors for the Sonar data set after 10 runs. Standard deviations in brackets.

Kernel	Train Error	Test Error	%SV
K_1 ($\gamma = 1$)	1.144 (0.460)	15.952 (0.372)	40.0 (0.0)
K_2 ($\gamma = .1$)	16.56 (0.077)	25.761 (0.170)	48.7 (0.0)
KSum	1.325 (0.516)	16.666 (0.380)	76.6 (1.8)
KAJD	0.783 (0.499)	15.238 (0.404)	82.9 (2.2)

of K_2 damages the performance of the straightforward sum approach. On the other hand, the kernel obtained by the AJD algorithm shows a better classification performance than the other fusion method and also than the individual kernels.

5 Conclusions and Future Work

In this work, we present a new framework for information fusion when the sources of information are given by a set of kernel matrices. The algorithm, based on the Approximate Joint Diagonalization of matrices, produces a new representation of the data set in a Euclidean space, where the basis is created from the representations induced by the individual kernels. In addition our method is able to eliminate redundant information from the individual kernels. The proposed fusion scheme has been tested in a couple of significative examples. Furthermore, the procedure is shown to be robust against the inclusion of noisy variables.

Future research will include the study of Joint Diagonalization Algorithms that take into account the label information in classification problems and also JD algorithms specific for regression problems.

References

1. Bach, F.R., Jordan, M.I.: Kernel Principal Components Analysis. Journal of Machine Learning Research 3, 1–48 (2002)
2. Cardoso, J.F., Souloumiac: A Jacobi Angles for Simultaneous Diagonalization. SIAM J. Mat. Anals. Applied 17(1), 161–164 (1996)
3. Choi, M., Young, R., Nam, M.-R., Kim, H.O.: Fusion of Multispectral and Panchromatic Satellite Images Using the Curvelet Transform. Geoscience and Remote Sensing Letters 2(2) (2005)
4. Martin de Diego, I., Moguerza, J.M., Muñoz, A.: Combining Kernel Information for Support Vector Classification. In: Roli, F., Kittler, J., Windeatt, T. (eds.) MCS 2004. LNCS, vol. 3077, pp. 102–111. Springer, Heidelberg (2004)
5. Epifanio, I., Gutierrez, J., Malo, J.: Linear transform for simultaneous diagonalization of covariance and perceptual metric matrix in image coding. Pattern Recognition 36, 799–1811 (2003)

6. Gretton, A., Herbrich, R., Smola, A., Bousquet, O., Schölkop, B.: Kernel methods for measuring independence. J. Machine Learning Research 6, 2075–2129 (2005)
7. Hua, Y.: On SVD estimating Generalized Eigenvalues of Singular Matrix Pencils in Noise. IEEE Transactions on Signal Processing 39(4), 892–900 (1991)
8. Joachims, T., Cristianini, N., Shawe-Taylor, J.: Composite Kernels for Hypertext Categorisation. In: Proceedings of the International Conference on Machine Learning, pp. 250–257 (2002)
9. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine intelligence 20(3), 226–239 (1998)
10. Mika, S., Ratsch, W.G., Scholkopf, B., Muller, K.-R.: Fisher Discriminant Analysis with Kernels. In: Proceedings of IEEE Neural Networks for Signal Processing Workshop, IEEE Computer Society Press, Los Alamitos (1999)
11. Moguerza, J., Muñoz, A.: Support Vector Machines with Applications. Statistical Science 21(3), 322–336 (2006)
12. Beresford, P.N.: The Symmetric Eigenvalue Problem. Classics in Applied Mathematics. SIAM (1997)
13. Schölkopf, B., Smola, A.J., Müller, K.R.: Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Computation 10, 1299–1319 (1998)
14. Wax, M., Sheinvald, J.: A least-squares approach to joint diagonalization. IEEE Signal Processing Lett. 4, 52–53 (1997)
15. Bunse-Gerstner, A., Byers, R., Mehrmann, V.: Numerical methods for simultaneous diagonalization. SIAM Journal on Matrix Analysis and Applications 14(4), 927–949 (1993)
16. Yeredor, A.: Non-Orthogonal Joint Diagonalization in the Least-Squares Sense With Application in Blind Source Separation. IEEE Transactions on Signal Processing. 50 (7), 1545–1553 (2002)
17. Piella, G., Heijmans, H.: Multiresolution Image Fusion Guided by a Multimodal Segmentation. In: Proceedings of ACIVS 2002, Ghent, Belgium (September 9-11, 2002)
18. Newman, D.J., Hettich, S., Blake, C.L. and Merz, C.J.UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science 1998, `http://www.ics.uci.edu/mlearn/MLRepository.html`