Lietuvos matematikos rinkinys
Proc. of the Lithuanian Mathematical Society, Ser. A
DOI: 10.15388/LMR.A.2016.02

ISSN 0132-2818
Vol. **57**, 2016
pages 7–11

Hilbert–Schmidt component analysis

Povilas Daniušis, Pranas Vaitkus, Linas Petkevičius

Faculty of Mathematics and Informatics, Vilnius University
Naugarduko str. 4, LT-03225 Vilnius, Lithuania
E-mail: povilas.daniusis@gmail.com, vaitkuspranas@gmail.com
E-mail: linas.petkevicius@mif.vu.lt

Abstract. We propose a feature extraction algorithm, based on the Hilbert–Schmidt independence criterion (HSIC) and the maximum dependence – minimum redundancy approach. Experiments with classification data sets demonstrate that suggested Hilbert–Schmidt component analysis (HSCA) algorithm in certain cases may be more efficient than other considered approaches.

Keywords: feature extraction, dimensionality reduction, HSCA, Hilbert–Schmidt independence criterion, kernel methods.

1 Introduction

In many cases the initial representation of data is inconvenient, or even prohibitive for further analysis. For example, in image analysis, text analysis and computational genetics, high-dimensional, massive, structural, incomplete, and noisy data sets are common. Therefore, feature extraction, or the revelation of informative features from raw data is one of the fundamental machine learning problems.

In this article we focus on supervised feature extraction algorithms, that use dependence-based criteria of optimality. The article is structured as follows. In Section 2 we briefly formulate an esimators of a Hilbert–Schmidt independence criterion (HSIC), proposed by [5]. In Section 3 we propose a new algorithm, Hilbert–Schmidt component analysis (HSCA). The main idea of HSCA is to find non-redundant features which maximize HSIC with a dependent variable. Finally, in Section 4, we experimentally compare our approach with several alternative feature extraction methods. Therein we statistically analyze the accuracy of k-NN classifier, based on LDA [4], PCA [6], HBFE [2, 10], and HSCA features.

2 Hilbert–Schmidt independence criterion

The Hilbert–Schmidt independence criterion (HSIC) is a kernel-based dependence measure proposed and investigated in [5, 8]. Let $T := (x_i, y_i)_{i=1}^m$ be a supervised training set, where $x_i \in \mathcal{X}$ are inputs, $y_i \in \mathcal{Y}$ – corresponding desired outputs, and $\mathcal{X}$, $\mathcal{Y}$ are two sets. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be two positive definite kernels [7], with corresponding Gram matrices $\mathbf{K}$, and $\mathbf{L}$. There are proposed two

empirical estimators of HSIC (see [5, 8])[1]:

$$\widehat{HSIC}_0(X,Y) := (m-1)^{-2}\mathrm{Tr}(KHLH), \tag{1}$$

and

$$\widehat{HSIC}_1(X,Y) := \frac{1}{m(m-3)}\left(\mathrm{Tr}\widetilde{K}\widetilde{L} + \frac{\mathbf{1}^T\widetilde{K}\mathbf{1}\mathbf{1}^T\widetilde{L}\mathbf{1}}{(m-1)(m-2)} - \frac{2}{m-2}\mathbf{1}^T\widetilde{K}\widetilde{L}\mathbf{1}\right). \tag{2}$$

The (1) is biased with an $O(m^{-1})$ bias, and the (2) is an unbiased estimator of HSIC [5, 8].

3 Hilbert–Schmidt component analysis (HSCA)

In this section we suggest an algorithm for Hilbert–Schmidt component analysis (HSCA), which is based on the HSIC dependence measure. The choice of HSIC is motivated by its neat theoretical properties [5, 8], and promising experimental results achieved by various HSIC-based feature extraction algorithms [2, 3, 5, 10].

Suppose we have a supervised training set $T := (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^m$, where $\mathbf{x}_i \in \mathbb{R}^{D_x}$ are observations, and $y \in \mathbb{R}^{D_y}$ are dependent variables. Let us denote the data matrices $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m]$, and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m]$, and assume that the kernel for the inputs is linear (i.e. $\mathbf{K} = \mathbf{X}^T\mathbf{X}$).

In HSCA we iteratively seek $d \leqslant D_x$ linear projections, which maximize the dependence with the dependent variable y and simultaneously minimize the dependence with the already computed projections. In other words, for the t-th feature we seek a projection vector $\mathbf{p}$, which maximizes the ratio

$$\eta_t(\mathbf{p}) = \frac{\widehat{HSIC}(\mathbf{p}^T\mathbf{X}, \mathbf{Y})}{\widehat{HSIC}(\mathbf{p}^T\mathbf{X}, \mathbf{P}_t^T\mathbf{X})}, \tag{3}$$

where $\mathbf{P}_t = [\mathbf{p}_1, \ldots, \mathbf{p}_{t-1}]$ are projection vectors extracted in previous $t-1$ steps, and $\widehat{HSIC}$ is an estimator of HSIC. Note that, at the first step, only $\widehat{HSIC}(\mathbf{p}^T\mathbf{X}, \mathbf{Y})$ is maximized.

For example, plugging (1) estimator into (3), we have to maximize the following generalized Rayleigh quotient

$$\eta_t(\mathbf{p}) = \frac{\mathrm{Tr}(\mathbf{X}^T\mathbf{p}\mathbf{p}^T\mathbf{X}\mathbf{H}\mathbf{L}\mathbf{H})}{\mathrm{Tr}(\mathbf{X}^T\mathbf{p}\mathbf{p}^T\mathbf{X}\mathbf{H}\mathbf{L}_f\mathbf{H})} = \frac{\mathbf{p}^T\mathbf{X}\mathbf{H}\mathbf{L}\mathbf{H}\mathbf{X}^T\mathbf{p}}{\mathbf{p}^T\mathbf{X}\mathbf{H}\mathbf{L}_f\mathbf{H}\mathbf{X}^T\mathbf{p}}, \tag{4}$$

where the kernel matrix of features $L_f(i,j) = l(\mathbf{P}_{t-1}^T\mathbf{x}_i, \mathbf{P}_{t-1}^T\mathbf{x}_j)$. The maximizer is principal eigenvector of the generalized eigenproblem

$$\mathbf{X}\mathbf{H}\mathbf{L}\mathbf{H}\mathbf{X}^T\mathbf{p} = \lambda\mathbf{X}\mathbf{H}\mathbf{L}_f\mathbf{H}\mathbf{X}^T\mathbf{p}. \tag{5}$$

The case of unbiased HSIC estimator (2) may be treated in the similar manner. Well known kernel trick [7] allows to extend HSCA to arbitrary kernel case, however we ommit the details due to space restrictions.

[1] In (2) $\widetilde{K}$ and $\widetilde{L}$ are corresponding Gram matrices with zero diagonals.

4 Computer experiments

In this section we will analyze twelve classification data sets, eleven of them are from the UCI machine learning repository [1], and the remaining *Ames* data set is from chemometrics.[2]

We are interested in the performance dynamics of the k-NN classifier, when the inputs are constructed by several feature extraction algorithms: unsupervised PCA [6], supervised LDA [4], HBFE [2, 10] and HSCA.

The measure of efficiency we will analyze therein is the accuracy of k-NN classifier, calculated over the testing set. The following procedure was adopted when conducting experiments. Fifty random partitions of the data set into training and testing sets of equal size was generated, and feature extraction was performed using all the above-mentioned methods. The projection matrices of the feature extraction methods were estimated using only the training data. The features generated from the testing set then were classified using k-NN classifier. The feature dimensionality was selected using a training data and 3-fold cross validation. Wilcoxon's sign rank test [9] with the standard p-value threshold of 0.05 was applied to the samples of corresponding classification accuracies. The following comparisons were made, indicating the statistically significant cases in the table:

1. $HBFE_1$ with $HBFE_0$, and $HSCA_1$ with $HSCA_0$ (better one indicated in **bold** text);
2. The most efficient method with the remaining ones (statistically significant cases are reported in underlined text);
3. HSCA with HBFE (data sets where HSCA was more efficient are indicated with $\bullet$, and $\circ$ means that it turned out to be less efficient);
4. The most efficient HSIC-based algorithm (i.e. $HBFE_0$, $HBFE_1$, $HSCA_0$ or $HSCA_1$) with the remaining ones ($\diamond$ means that HSIC-based algorithm outperformed other ones, and $\star$ means that PCA, LDA or unmodified inputs were more efficient).

The results in Table 1 show that HSCA approach may allow to achieve slightly better classification accuracy for some data sets.

5 Conclusions

Suggested HSCA (Hilbert–Schmidt component analysis) algorithm (Section 3) optimizes ratio of feature relevancy, and feature redundancy estimates. Both estimates are formulated in terms of HSIC dependence measure. Optimal features are solutions of generalized eigenproblem (4). In section 4 we statistically compared HBFE with several alternative feature extraction methods, analysing classification performance (accuracy) as the measure of feature relevance. The results of the conducted experiments demonstrate practical usefulness of HSCA algorithm.

[2] Details about *Ames* data set is not available due to agreement with the provider.

Table 1. Classification accuracy comparison.

Dataset	Full	$HBFE_1$	$HBFE_0$	$HSCA_1$	$HSCA_0$	PCA	LDA
1-NN classifier and linear kernel							
Ames • ◇	0.7753	0.7589	**0.7765**	0.7826	**0.8012**	0.7786	0.7714
Australian	0.7933	0.7987	0.8045	0.8093	0.8095	0.7868	0.8114
Breastcancer	0.9558	0.9553	0.9543	0.9553	0.9595	0.9566	0.9562
Covertype ◇	0.6868	**0.6956**	0.6732	0.7086	0.7014	0.6748	0.6756
Derm	0.9906	0.9973	0.9973	0.9973	0.9971	0.9949	0.9971
German	0.6698	**0.6841**	0.6730	0.6833	0.6910	0.6700	0.6851
Heart	0.7618	0.7600	0.7677	0.7612	0.7627	0.7520	0.7698
Ionosphere • ◇	0.8421	0.8555	**0.8683**	0.8686	0.8773	0.8581	0.8171
Sonar	0.8146	**0.7819**	0.7538	0.7427	**0.8046**	0.8146	0.6938
Spambase •	0.8975	0.8993	0.8979	0.9116	0.9056	0.9015	0.8680
Specft	0.6770	0.6690	0.7030	0.6730	0.7020	0.6630	0.5370
Wdbc •	0.9503	0.9355	**0.9455**	0.9476	**0.9570**	0.9506	0.9528
1-NN classifier and Gaussian kernel							
Australian •	0.7924	0.7900	0.7927	0.7878	**0.8185**	0.7807	0.8110
Breastcancer	0.9508	0.9505	0.9458	0.9472	0.9466	0.9522	0.9487
Covertype • ⋆	0.6932	0.6480	**0.6738**	0.6855	0.6860	0.6777	0.7091
Derm	0.9872	0.9985	0.9989	0.9993	0.9989	0.9935	0.9984
German • ◇	0.6717	0.6668	0.6784	**0.6956**	0.6797	0.6737	0.6802
Heart	0.7638	0.7689	0.7679	0.7575	0.7669	0.7588	0.7366
Ionosphere	0.8521	0.8895	**0.9128**	0.9025	0.9158	0.9083	0.8927
Sonar •	0.8358	0.6955	**0.7942**	0.7204	**0.8344**	0.8387	0.8258
Spambase • ⋆	0.8570	0.7994	0.7903	0.8119	0.8129	0.8471	0.8779
Specft	0.6778	0.7283	0.7317	0.7650	0.7400	0.6370	0.7370
Wdbc ⋆	0.9510	0.9148	**0.9425**	0.9320	0.9424	0.9510	0.9599

References

[1] A. Asuncion and D.J. Newman. *UCI Machine Learning Repository*. University of California, School of Information and Computer Science, 2007. Available from internet: http://archive.ics.uci.edu/ml/.

[2] P. Daniušis and P. Vaitkus. Supervised feature extraction using Hilbert–Schmidt norms. *Lecture Notes in Computer Science*, **5788**:25–33, 2009.

[3] P. Daniušis and P. Vaitkus. A feature extraction algorithm based on the Hilbert–Schmidt independence criterion. *Šiauliai Math. Sem.*, **4**(12):35–42, 2009.

[4] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Ann. Eug.*, **7**:179–188, 1936.

[5] A. Gretton, O. Bousquet, A. Smola and B. Schölkopf. Measuring statistical dependence with Hilbert–Schmidt norms. In *Proc. of 16th Int. Conf. on Algorithmic Learning Theory*, pp. 63–77, 2005.

[6] I.T. Jolliffe. *Principal Component Analysis*. Springer, Berlin, 1986.

[7] B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, 2002.

[8] L. Song, A. Smola, A. Gretton, K. Borgwardt and J. Bedo. Supervised feature selection via dependence estimation. In *Proc. Intl. Conf. Machine Learning*, pp. 823–830, 2007.

[9] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, **1**:80–83, 1945.

[10] Y. Zhang and Z.-H. Zhou. Multi-label dimensionality reduction via dependence maximization. In *Proc. of the 33rd AAAI Conf. on Artificial Intelligence*, pp. 1503–1505, 2008.

REZIUMĖ

Hilberto–Šmito komponenčių analizė

P. Daniušis, P. Vaitkus, L. Petkevičius

Straipsnyje pateikiamas naujas HSCA požymių išskyrimo metodas, kurio esmė yra maksimizuoti HSIC priklausomumo mato įvertinį tarp požymių ir priklausomo kintamojo, kartu siekiant eliminuoti požymių tarpusavio priklausomumą, metodas eksperimentiškai palygintas su klasikiniais bei naujais požymių išskyrimo metodais.

Raktiniai žodžiai: požymių išskyrimas, dimensijos mažinimas, HSCA, Hilbert–Schmidt nepriklausomumo kriterijus, branduolių metodai.