21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Discriminative Indefinite Kernel Classifier from Pairwise Constraints and Unlabeled Data

Hui Xue[1] Songcan Chen[2] Jijian Huang[1]

[1] *School of Computer Science and Engineering, Southeast University, Nanjing, P. R. China*
[2] *College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, P. R. China*
E-mail: { hxue@seu.edu.cn, s.chen@nuaa.edu.cn, huangjijian2008@seu.edu.cn}

Abstract

Semi-supervised classification from pairwise constraints is a challenge in pattern recognition, since the constraints just represent the relationships between data pairs rather than the definite labels. In the last few years, several methods have been proposed, however, they still utilize either the discriminability within the constraints or the abundant unlabeled data insufficiently. In this paper, we present a novel discriminative indefinite kernel classifier. We first transform the constrained data pairs into newly-labeled samples by an outer product transformation, and then introduce an indefinite discriminative regularizer in the transformed space in order to further embed the discriminative and structural information involved in the newly labeled and unlabeled samples into the classifier design. We validate that such classifier naturally lies in the more general Reproducing Kernel Kreĭn Space rather than the common Reproducing Kernel Hilbert Space. Experiments show the superiority of our method.

1. Introduction

Practical pattern recognition is often confronted with the situations where few labeled data together with abundant unlabeled data are available [1, 2]. Furthermore, in some tasks, it is possible to obtain discriminative information in the form of pairwise must-link and cannot-link constraints which are more general than labels [1]. This problem has been widely investigated in many research fields in the last decade, such as dimensionality reduction [3], clustering [4], image segmentation [5].

However, the study on classification with pairwise constraints is relatively scarce due to the difficulty in extracting the discriminative information from the constraints, which just represent the relationships between data pairs rather than the labels. Consequently, the constraints are hard to be directly incorporated into common classifier models. Recently, some solutions have been proposed. Zhang and Yan first combined the pairs into some single samples and then designed a least square classifier based on the value of the constraints as the labels for these samples [6]. Yan et al. incorporated the constraints into a margin-based learning framework through defining another new loss function in order to model the decision boundary [7]. Although these methods have shown better classification performance, they more likely extract the prior information from the constraints insufficiently, such as discriminative and structural information. Furthermore, they basically neglect the utilization of unlabeled data, whose distribution is vital in semi-supervised learning [8].

In this paper, we propose a novel discriminative indefinite kernel classifier. In order to convert the classification with pairwise constraints and unlabeled data into a common semi-supervised problem that can be solved more conveniently, we transform the pairs into some newly-labeled samples by outer product due to the attractive properties of the product to guarantee solution consistency [6], as well as the unlabeled data. Then we introduce an indefinite discriminative regularizer into the classifier instead of the traditional smoothness regularizer, which embeds both discriminative and structural information involved in the new samples simultaneously. We further validate that such classifier is naturally in the generalized Reproducing Kernel Kreĭn Space (RKKS) [9] induced from the so-defined indefinite regularizer.

The rest of the paper is organized as follows. Section 2 presented the proposed classifier. The corresponding indefinite kernel analysis is derived in

Section 3. In Section 4, the experimental comparisons are given. Some conclusions are drawn in Section 5.

2. Discriminative Indefinite Kernel Classifier (DIKC)

Given the pairwise constraints $\{(\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \tilde{y}_i)\}_{i=1}^{l}$, where $\tilde{y}_i \in \{1, -1\}$ indicates the must-link and cannot-link constraints respectively, $\boldsymbol{x}_i \in \boldsymbol{R}^m$. Moreover, $\{\boldsymbol{x}_j\}_{j=l+1}^{n}$ are the unlabeled data. For each pair, we define a new single vector [6]

$$\boldsymbol{z}_i = vech(\boldsymbol{x}_{i1} \circ \boldsymbol{x}_{i2}) \quad (1)$$

where $\boldsymbol{x}_{i1} \circ \boldsymbol{x}_{i2}$ is the outer product. The operator *vech* [6] returns the upper triangular elements of the matrix $\boldsymbol{x}_{i1} \circ \boldsymbol{x}_{i2}$ in order of row to construct $\boldsymbol{z}_i$ whose dimension is $(m+1) \times m/2$. As a result, the constraints are converted into some labeled samples $\{(\boldsymbol{z}_i, \tilde{y}_i)\}_{i=1}^{l}$ in the transformed space $\widetilde{\boldsymbol{X}}$.

For the unlabeled data, we make the similar transformation

$$\boldsymbol{z}_j = vech(\boldsymbol{x}_j \circ \boldsymbol{x}_j) \quad (2)$$

Up to now, the classification problem with pairwise constraints has been changed to design a semi-supervised binary classifier in $\widetilde{\boldsymbol{X}}$. Sugiyama et al. [8] have pointed out that both the global structure of unlabeled data and class information brought by the labeled data are important for classification. So we further introduce a discriminative regularizer instead of the smoothness regularizer into the classifier to embed such prior information sufficiently.

Assume that the classifier has linear form

$$f(\boldsymbol{z}) = \boldsymbol{w}^T \boldsymbol{z} \quad (3)$$

We use the total scatter measure to reflect the global structure [8]

$$\begin{aligned} S_t &= \sum_{i=1}^{n} \left\| f(\boldsymbol{z}_i) - \frac{1}{n}\sum_{j=1}^{n} f(\boldsymbol{z}_j) \right\|^2 \\ &= \frac{1}{2}\boldsymbol{w}^T \sum_{i=1}^{n}\sum_{j=1}^{n} \Phi_{i,j}^{t} (\boldsymbol{z}_i - \boldsymbol{z}_j)(\boldsymbol{z}_i - \boldsymbol{z}_j)^T \boldsymbol{w} \\ &= \frac{1}{2}\boldsymbol{w}^T \boldsymbol{S}_t \boldsymbol{w} \end{aligned} \quad (4)$$

where $\Phi_{i,j}^{t} = 1/n$.

Then we define the local discriminative structure of the labeled data by the improved intra-class and inter-class scatter measures [8]

$$\begin{aligned} S_{lw} &= \sum_{k=1}^{2}\sum_{i=1}^{l_k} \left\| \tilde{f}(\boldsymbol{z}_i^{(k)}) - \frac{1}{l_k}\sum_{j=1}^{l_k} \tilde{f}(\boldsymbol{z}_j^{(k)}) \right\|^2 \\ &= \frac{1}{2}\boldsymbol{w}^T \sum_{i=1}^{l}\sum_{j=1}^{l} \Phi_{i,j}^{lw} (\boldsymbol{z}_i - \boldsymbol{z}_j)(\boldsymbol{z}_i - \boldsymbol{z}_j)^T \boldsymbol{w} \\ &= \frac{1}{2}\boldsymbol{w}^T \boldsymbol{S}_{lw} \boldsymbol{w} \end{aligned} \quad (5)$$

$$\begin{aligned} S_{lb} &= \sum_{k=1}^{2} l_k \left\| \frac{1}{l_k}\sum_{i=1}^{l_k} \tilde{f}(\boldsymbol{z}_i^{(k)}) - \frac{1}{l}\sum_{j=1}^{l} \tilde{f}(\boldsymbol{z}_j) \right\|^2 \\ &= \frac{1}{2}\boldsymbol{w}^T \sum_{i=1}^{l}\sum_{j=1}^{l} \Phi_{i,j}^{lb} (\boldsymbol{z}_i - \boldsymbol{z}_j)(\boldsymbol{z}_i - \boldsymbol{z}_j)^T \boldsymbol{w} \\ &= \frac{1}{2}\boldsymbol{w}^T \boldsymbol{S}_{lb} \boldsymbol{w} \end{aligned} \quad (6)$$

where $\Phi_{i,j}^{lw} = \begin{cases} \Psi_{i,j}/l_k & if \ \tilde{y}_i = \tilde{y}_j = k \\ 0 & if \ \tilde{y}_i \neq \tilde{y}_j \end{cases}$

$$\Phi_{i,j}^{lb} = \begin{cases} \Psi_{i,j}(1/l - 1/l_k) & if \ \tilde{y}_i = \tilde{y}_j = k \\ 1/l & if \ \tilde{y}_i \neq \tilde{y}_j \end{cases}$$

$$\Psi_{i,j} = \begin{cases} \exp\left(-\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2/\sigma^2\right) & if \ \boldsymbol{z}_j \in ne(\boldsymbol{z}_i) \ or \ \boldsymbol{z}_i \in ne(\boldsymbol{z}_j) \\ 0 & otherwise \end{cases}$$

and $ne(\boldsymbol{z}_i)$ denotes the k nearest neighbors of $\boldsymbol{z}_i$.

We bridge the two kinds of structural information into two new scatter measures [8]

$$\boldsymbol{w}^T \boldsymbol{S}_{rlw} \boldsymbol{w} = \frac{1}{2}\boldsymbol{w}^T [(1-\gamma)\boldsymbol{S}_{lw} + \gamma \boldsymbol{I}] \boldsymbol{w} \quad (7)$$

$$\boldsymbol{w}^T \boldsymbol{S}_{rlb} \boldsymbol{w} = \frac{1}{2}\boldsymbol{w}^T [(1-\gamma)\boldsymbol{S}_{lb} + \gamma \boldsymbol{S}_t] \boldsymbol{w} \quad (8)$$

where γ is the regularization parameter that regulates the relative significance of such information, $0 \leq \gamma \leq 1$, and $\boldsymbol{I}$ is the identity matrix to avoid the ill-conditioned $\boldsymbol{S}_{lw}$.

Then we define the discriminative regularizer as

$$R_{disreg}(\boldsymbol{f}, \eta) = \boldsymbol{w}^T [\eta \boldsymbol{S}_{rlw} - (1-\eta)\boldsymbol{S}_{rlb}] \boldsymbol{w} \quad (9)$$

where η is the regularization parameter, $0 \leq \eta \leq 1$.

The final optimization function of DIKC in $\widetilde{\boldsymbol{X}}$ can be formulated as

$$\min_{f \in \widetilde{K}} \ \frac{1}{l}\sum_{i=1}^{l} \left(\tilde{y}_i - f(\boldsymbol{z}_i)\right)^2 + R_{disreg}(\boldsymbol{f}, \eta) \quad (10)$$

In order to classify unseen samples in the original space, we should apply an inverse operation of *vech* to the discriminative vector $\boldsymbol{w}$ obtained in $\widetilde{\boldsymbol{X}}$ [6]

$$\boldsymbol{\theta} = vech^{-1}(\boldsymbol{w}) \quad (11)$$

Then we perform the eigen-decomposition to the matrix $\boldsymbol{\theta}$, and select the largest eigenvalue s_1 and corresponding eigenvector $\boldsymbol{u}_1$ as the sign-insensitive estimator $\boldsymbol{\beta} = \sqrt{s_1}\boldsymbol{u}_1$ of $\widehat{\boldsymbol{w}}$ [6], which is the discriminative vector in the original space.

The real sign of $\widehat{\boldsymbol{w}}$ can be determined by few labeled examples $(\boldsymbol{x}_1, y_1)$, ..., $(\boldsymbol{x}_p, y_p)$ [6]

$$sign(\widehat{\boldsymbol{w}}) = \begin{cases} +1, & \sum_{i=1}^{p} I(y_i \boldsymbol{\beta}^T \boldsymbol{x}_i \geq 0) \geq \lceil m/2 \rceil \\ -1, & otherwise \end{cases} \quad (12)$$

where $I(\cdot)$ is the set indicator function and $\lceil m/2 \rceil$ is the ceil function.

For an unseen sample $\boldsymbol{x}$, the predicted class label is

$$y = sign(\widehat{\boldsymbol{w}})\widehat{\boldsymbol{w}}^T \boldsymbol{x} \quad (13)$$

3. Justification

The discriminative regularizer embeds the global and local discriminative structures of the samples simultaneously, however, it is obviously indefinite. In this section, we will validate that such regularizer actually satisfies a generalized inner product definition $\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\widetilde{K}} = \langle \boldsymbol{f}_+, \boldsymbol{g}_+ \rangle_{H_+} - \langle \boldsymbol{f}_-, \boldsymbol{g}_- \rangle_{H_-}$ in the RKKS, which admits the inner product indefinite and thus is more general than the common Reproducing Kernel Hilbert Space (RKHS) [9].

Proposition 1. *The discriminative regularizer can be formulated as an inner product in the RKKS, that is,*

$$R_{disreg}(\boldsymbol{f},\eta) = \langle \boldsymbol{f},\boldsymbol{f}\rangle_{\widetilde{K}_{disreg}} \quad (14)$$

Proof. Recall that

$$R_{disreg}(\boldsymbol{f},\eta) = \boldsymbol{w}^T[\eta \boldsymbol{S}_{rlw} - (1-\eta)\boldsymbol{S}_{rlb}]\,\boldsymbol{w}$$

Decompose the joint matrix $\eta \boldsymbol{S}_{rlw} - (1-\eta)\boldsymbol{S}_{rlb}$ into

$$\eta \boldsymbol{S}_{rlw} - (1-\eta)\boldsymbol{S}_{rlb} = \boldsymbol{Z}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{Z}^T \quad (15)$$

So
$$\begin{aligned} R_{disreg}(\boldsymbol{f},\eta) &= \boldsymbol{w}^T\boldsymbol{Z}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{Z}^T\boldsymbol{w} \\ &= \boldsymbol{f}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{f}^T \\ &= \boldsymbol{f}\left(\textstyle\sum_{\lambda_i>0}\lambda_i \boldsymbol{u_i}\,\boldsymbol{u_i}^T + \sum_{\lambda_j<0}\lambda_j \boldsymbol{u_j}\,\boldsymbol{u_j}^T\right)\boldsymbol{f}^T \quad (16) \end{aligned}$$

Let $\boldsymbol{\Gamma}_+ = \boldsymbol{U}_+\boldsymbol{\Lambda}_+\boldsymbol{U}_+^T$, $\boldsymbol{\Gamma}_- = \boldsymbol{U}_-\boldsymbol{\Lambda}_-\boldsymbol{U}_-^T$, obviously,

$$\boldsymbol{\Gamma}_+^T\boldsymbol{\Gamma}_- = \boldsymbol{U}_+\boldsymbol{\Lambda}_+\boldsymbol{U}_+^T\boldsymbol{U}_-\boldsymbol{\Lambda}_-\boldsymbol{U}_-^T = \boldsymbol{0}$$

that is, $\boldsymbol{\Gamma}_+$ and $\boldsymbol{\Gamma}_-$ are orthogonal.

Decompose $\boldsymbol{f} = \boldsymbol{f}_+ + \boldsymbol{f}_-$, where $\boldsymbol{f}_+ \in Hilbert(\boldsymbol{\Gamma}_+)$, $\boldsymbol{f}_- \in Hilbert(-\boldsymbol{\Gamma}_-)$, then

$$\begin{aligned} R_{disreg}(\boldsymbol{f},\eta) &= \boldsymbol{f}[\boldsymbol{\Gamma}_+ - (-\boldsymbol{\Gamma}_-)]\boldsymbol{f}^T \\ &= (\boldsymbol{f}_+ + \boldsymbol{f}_-)\boldsymbol{\Gamma}_+(\boldsymbol{f}_+ + \boldsymbol{f}_-)^T \\ &\quad - (\boldsymbol{f}_+ + \boldsymbol{f}_-)(-\boldsymbol{\Gamma}_-)(\boldsymbol{f}_+ + \boldsymbol{f}_-)^T \\ &= \boldsymbol{f}_+\boldsymbol{\Gamma}_+\boldsymbol{f}_+^T - \boldsymbol{f}_-(-\boldsymbol{\Gamma}_-)\boldsymbol{f}_-^T \quad (17) \end{aligned}$$

Let

$$\boldsymbol{f}_+\boldsymbol{\Gamma}_+\boldsymbol{f}_+^T = \langle \boldsymbol{f}_+,\boldsymbol{f}_+\rangle_{\boldsymbol{H}_+},\ \boldsymbol{f}_-(-\boldsymbol{\Gamma}_-)\boldsymbol{f}_-^T = \langle \boldsymbol{f}_-,\boldsymbol{f}_-\rangle_{\boldsymbol{H}_-} \quad (18)$$

We can formulate the discriminative regularizer as

$$\begin{aligned} R_{disreg}(\boldsymbol{f},\eta) &= \langle \boldsymbol{f}_+,\boldsymbol{f}_+\rangle_{\boldsymbol{H}_+} - \langle \boldsymbol{f}_-,\boldsymbol{f}_-\rangle_{\boldsymbol{H}_-} \\ &= \langle \boldsymbol{f},\boldsymbol{f}\rangle_{\widetilde{K}_{disreg}} \end{aligned}$$ □

Consequently, DIKC is actually an inherent indefinite kernel classifier in RKKS, which induces the indefinite kernel from the regularizer itself. The corresponding solution $\boldsymbol{f}^*$ admits the generalized Representer Theorem in RKKS and thus the optimization boils down to the general smoothing problem which can be solved analytically [9].

4. Experiments

We conduct experiments on the IDA database to evaluate the proposed DIKC by comparing with the two methods mentioned in the Introduction section: Pariwise Kernel Logistic Regression (PKLR) [7] and classifier On the Value of Pairwise Constraints (OVPC) [6]. We also select regularized least square classifier as the Baseline. The database consists of thirteen datasets, which all contain two classes. We use the training and testing sets offered by the database. The constraints are randomly created depending on whether the class labels of the pairs are same or not, whose number is changed from 10 to 50. In DIKC, the number of the k nearest neighbors is selected from {5,10,15,20}. The regularization parameters in PKLR and OVPC are chosen from $\{2^{-10}, 2^{-9},\ldots,2^{9}, 2^{10}\}$. And the parameters in DIKC are selected in [0,0.1,…,0.9,1]. All the selections are done by cross-validation. Since labeled samples are only used to determine the sign, we only select one sample from each class.

Figure 1 shows the average classification accuracies of the compared methods. The accuracies of PKLR, OVPC and DIKC are basically improved with the step-by-step increased constraints. However, DIKC outperforms the other methods almost in all datasets due to the further utilization of the discriminative and structural information involved in the constraints and unlabeled data.

5. Conclusion

In this paper, we propose a novel indefinite kernel classifier DIKC from pairwise constraints and unlabeled data. DIKC first transforms the constraints into some single samples and then designs a discriminability-driven regularizer in order to fully capture the discriminative and structural information in the data. Experimental results demonstrate the effectiveness of DIKC.

Acknowledgement

This work was supported in part by National Natural Science Foundations of China (Grant Nos. 60905002 & 60973097), and the Scientific Research Startup Project of New Doctorial Faculties of Southeast University. Furthermore, the work was also supported by the Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education.

References

[1] M.S. Baghshah and S.B. Shouraki. Efficient kernel learning from constraints and unlabeled data. In: the 20th Intl. Conf. on Pattern Recognition (ICPR): 3364-3367, 2010.

[2] J.J. Pan, S.J. Pan, J. Yin, L.M. Ni, and Q. Yang. Tracking mobile users in wireless networks via semi-supervised colocalization. IEEE Trans. on Pattern Analysis and Machine Intelligence, 34(3): 587-600, 2012.

[3] D. Zhang, Z.-H. Zhou, and S. Chen. Semi-supervised dimensionality reduction. In: Proc. 7th SIAM Intl. Conf. on Data Mining (SDM), 629-634, 2007.

[4] O. Chapelle, B. Schölkopf, and A. Zien. Semi-Supervised Learning. MIT Press, Cambridge, MA, USA, 2006.

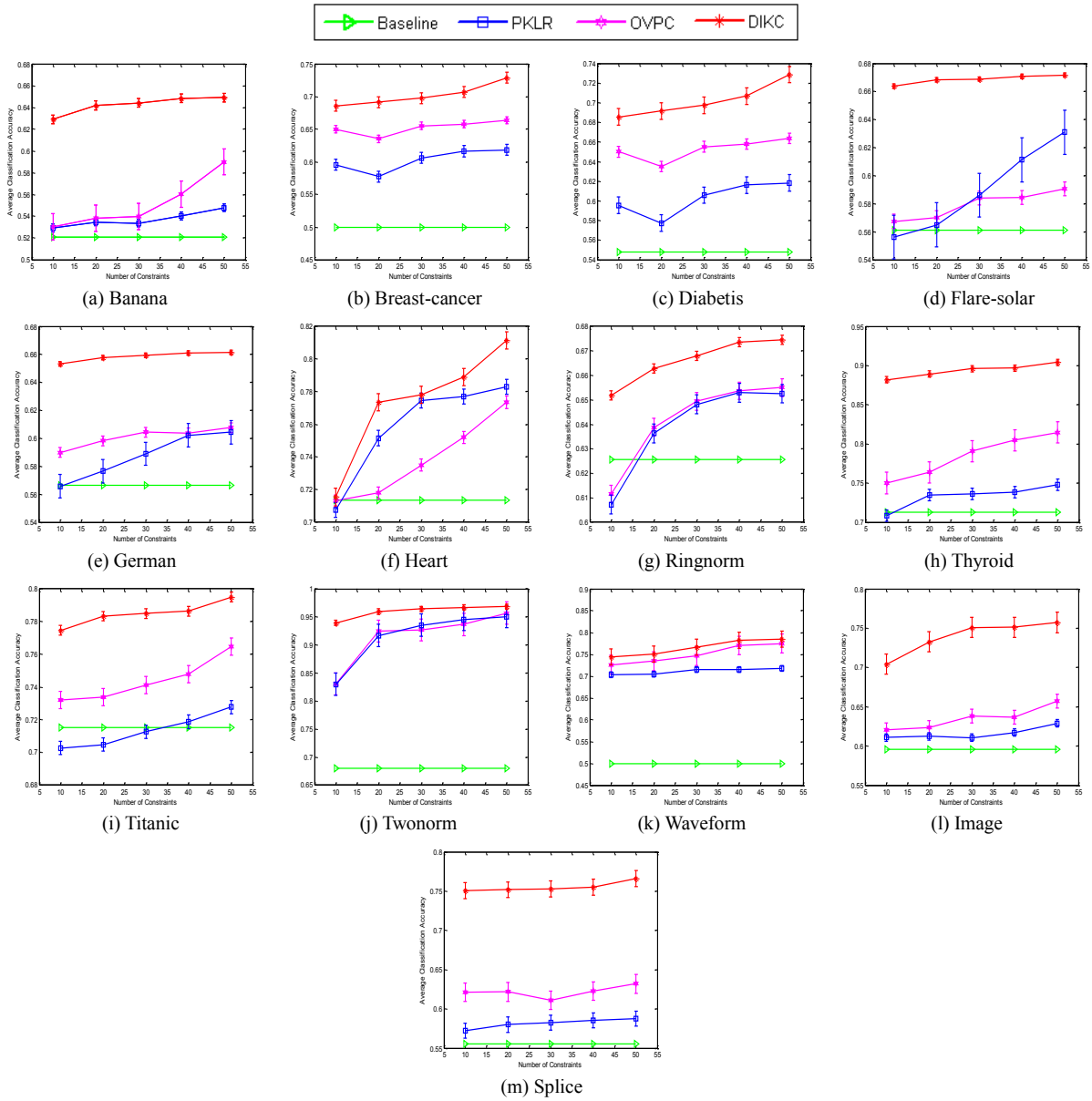
[5] S. Yu and J. Shi. Grouping with directed relationships. In M. Figueiredo, J. Zerubia, and A.K. Jain (Eds.), Energy Minimization Methods in Computer Vision and Pattern Recognition, 283-297, Springer, Berlin, LNCS 2134, 2001.

[6] J. Zhang and R. Yan. On the value of pairwise constraints in classification and consistency. In: Proc. of the 24th Intel. Conf. On Machine Learning (ICML), Corvallis, OR, 1111-1118, 2007.

[7] J.J. Pan, Q. Yang, H. Chang, and D.-Y. Yeung. A manifold regularization approach to calibration reduction for sensor-network based tracking. In: Proc. of the 21st National Conference on Artificial Intelligence (AAAI), 988-993, 2006.

[8] M. Sugiyama, T. Idé, S. Nakajima, and J. Sese. Semi-supervised local fisher discriminant analysis for dimensionality reduction. Machine Learning, 78(1-2): 35-61, 2010.

[9] C.S. Ong, X. Mary, S. Canu, and A.J. Smola. Learning with non-positive kernels. In: Proc. of the 21st Intl. Conf. on Machine Learning (ICML), 2004.



(a) Banana (b) Breast-cancer (c) Diabetis (d) Flare-solar

(e) German (f) Heart (g) Ringnorm (h) Thyroid

(i) Titanic (j) Twonorm (k) Waveform (l) Image

(m) Splice

Figure 1. Classification performance comparisons of Baseline, PKLR, OVPC and DIKC in the IDA datasets