Unsupervised Kernel Function Building Using Maximization of Information Potential Variability

A.M. Álvarez-Meza, D. Cárdenas-Peña, and Germán Castellanos-Domínguez

Signal Processing and Recognition Group
Universidad Nacional de Colombia
Campus La Nubia, km 7 via al Magdalena, Manizales, Colombia
{amalvarezme,dcardenasp,cgcastellanosd}@unal.edu.co

Abstract. We propose a kernel function estimation strategy to support machine learning tasks by analyzing the input samples using Renyi's Information Metrics. Specifically, we aim to identify a Reproducing Kernel Hilbert Space spanning the most widely the information force among data points by the maximization of the information potential variability of Parzen-based pdf estimation. So, a Gaussian kernel bandwidth updating rule is obtained as a function of the forces induced by a given dataset. Our proposal is tested on synthetic and real-world datasets related to clustering and classification tasks. Obtained results show that presented approach allows to compute RKHS's favoring data groups separability, attaining suitable learning performances in comparison with state of the art algorithms.

1 Introduction

Kernel functions allow enhancing random data representation for supporting machine learning systems. Moreover, kernel-based methods are powerful tools for developing better performing solutions by adapting the kernel to a given problem, instead of learning data relationships from explicit raw vector representations. The kernel function is a very flexible container to express knowledge about the problem as well as to capture meaningful data relationships [1]. However, building suitable kernels requires some user prior knowledge about input data, which is not available in most of the practical cases; this situation becomes worse when handling unsupervised inferring tasks.

Among many feasible kernels, the Gaussian function is preferred since it aims to find a Reproducing Kernel Hilbert Space - RKHS with universal approximating capability [3]. However, its use highly relies on the appropriate selection of the kernel parameters that are not easy to fix when dealing with complex data structures. In fact, the Gaussian Kernel bandwidth (scale) must be accurately tuned as to estimate an RKHS that should hold the main data relationships; otherwise, an unappropriate scale value leads to distinct RKHS not fulfilling the learning task. To cope with this issue and specifically devoted to unsupervised tasks, authors in [11, 12] propose to adjust the kernel parameter by making use of local scales instead of a global one allowing to exploit the spatial-varying data properties. Yet, these methods do not guarantee the Mercer's properties required for building kernel functions [6].

E. Bayro-Corrochano and E. Hancock (Eds.): CIARP 2014, LNCS 8827, pp. 335–342, 2014.
© Springer International Publishing Switzerland 2014

Nonetheless, most of of kernel estimation approaches are limited to the conventional concepts of second order statistics (mainly L2 distances). Instead, some Information Theoretic Learning (ITL) frameworks have been developed based on information theoretic underpinnings, which more generally quantify data uncertainty. In fact, information-based approaches can improve interpretation of random data structures, making salient connections between information measures and RKHS [8]. In ITL methods, the kernel building task reduces to estimation of the probability density function (pdf) that is rarely known due to the only available information comes from data samples at hand. Here, the kernel estimator involves a symmetrical window sliding along a sequence with its weighted values being smoothed inside. In particular, author in [7] proposes to estimate the pdf using the Renyi's entropy along with a Gaussian kernel Parzen estimator. However, both the pdf estimation success and the learning performance are highly dependent on the kernel parameter, namely, the bandwidth value. Some ITL-based approaches have been also proposed to fix the kernel scale value by optimizing information quantities [4, 10, 13], nevertheless, supervised data is required.

We propose a new kernel function estimation strategy to build a suitable Gaussian kernel-based RKHS oriented towards clustering. To this end, we make use of the intrinsic information potential variations from a Parzen-based pdf estimator. Namely, we seek for a RKHS maximizing the global kernel parameter the whole information potential variability. As a result, we get a scale updating rule as a function of the information forces, which are induced by a kernel function applied over a finite sample set. We provide testing of our proposal on two classical machine learning tasks (clustering and classification) using both synthetic and real data. Obtained results show that presented approach allows building a RKHS kernel favoring data groups separability and reaching suitable clustering performance in comparison with other state-of-the-art algorithms.

2 Materials and Methods

2.1 Gaussian-Based Renyi's Information Metrics

The basis of the ITL framework is the Renyi's Information quadratic metric that for a random variable is plainly defined as follows:

$$\mathcal{H}_2(\boldsymbol{x}) = -\log \int_{\boldsymbol{x}\in\mathcal{X}} f^2(\boldsymbol{x})d\boldsymbol{x} \ , \tag{1}$$

where $f(\boldsymbol{x})$ is the pdf of the random variable $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^P$. Nevertheless, such a pdf is usually unknown. Hence, a method to estimate the Renyi's entropy directly from a set of N samples, $\boldsymbol{X}=\{\boldsymbol{x}_j : \forall j \in [1, N]\}$, can be achieved by using the Parzen's nonparametric pdf estimation for $\boldsymbol{x}$, defined as

$$f(\boldsymbol{x}) \approx p_X(\boldsymbol{x}|\theta) = \boldsymbol{E}\left\{\kappa\{\boldsymbol{x} - \boldsymbol{x}_j, \theta\} : \forall j \in [1, N]\right\} \ , \tag{2}$$

where $\kappa\{\cdot, \theta\} \in \mathbb{R}^+$ is a symmetric kernel function with parameter set θ, and notation $\boldsymbol{E}\{\cdot\}$ stands for averaging operator.

Provided the observation set $\boldsymbol{X}$ and substituting the Parzen's estimation of Eq. (2) into Eq. (1), we get the following estimator of the Renyi's quadratic entropy:

$$\mathcal{H}_2\left(\boldsymbol{x}\right) \approx H_2\left(\boldsymbol{X}\right) = -\log \sum\nolimits_{\boldsymbol{x}_i \in \boldsymbol{X}} p_{\boldsymbol{X}}^2(\boldsymbol{x}_i|\theta) = -\log \boldsymbol{V}(\boldsymbol{X}) \ , \tag{3}$$

where $\boldsymbol{V}(\boldsymbol{X})$ is the so termed information potential (IP) of the observation set $\boldsymbol{X}$.

Though there are many feasible functions, the Gaussian kernel that is defined as $\kappa\{\boldsymbol{x},\theta\}{=}g\{\boldsymbol{x},\sigma^2\} \triangleq (2\pi\sigma^2)^{-P/2}\exp(-\boldsymbol{x}^\top\boldsymbol{x}/(2\sigma^2))$ is preferred since it aims to find a Reproducing Kernel Hilbert Space - RKHS with universal approximating capability and with a single bandwidth parameter $\sigma \in \mathbb{R}^+$. Since the IP for the Gaussian kernel gets the following formula

$$\boldsymbol{V}(\boldsymbol{X}) = \boldsymbol{E}\left\{g\{\boldsymbol{x}_i - \boldsymbol{x}_j, \sigma^2\} : \forall i, j \in [1, N]\right\} \ , \tag{4}$$

we can infer that the IP yields an entropy estimate that is based on the summation of pairwise sample interactions through the Gaussian kernel function [4]. Also, the Information Force (IF), $F_i \in \mathbb{R}^P$, is defined as the force acting on particle $\boldsymbol{x}_i$ due to all other particles in $\boldsymbol{X}$ and is given by the derivative of the IP with respect to $\boldsymbol{x}_i$:

$$\begin{aligned} F_i &= \frac{\partial}{\partial \boldsymbol{x}_i}\boldsymbol{V}(\boldsymbol{X}) = -(N\sigma)^{-2}\sum\nolimits_{\boldsymbol{x}_j \in \boldsymbol{X}} g\{(\boldsymbol{x}_i - \boldsymbol{x}_j), \sigma^2\}(\boldsymbol{x}_i - \boldsymbol{x}_j) \\ &= \boldsymbol{E}\left\{F(\boldsymbol{x}_i|\boldsymbol{x}_j) : \forall j \in [1, N]\right\} \ , \end{aligned} \tag{5}$$

where $F(\boldsymbol{x}_i|\boldsymbol{x}_j) = (N\sigma^2)^{-1}g\{(\boldsymbol{x}_i - \boldsymbol{x}_j), \sigma^2\}(\boldsymbol{x}_i - \boldsymbol{x}_j)$ corresponds to the conditional IF acting on $\boldsymbol{x}_i$ due to $\boldsymbol{x}_j$. Generally, the IFs can be interpreted in light of inner products in a high dimensional feature space [2].

2.2 Kernel Function Estimation from Information Potential Variability

Two important facts have to be highlighted from Eq. (5). On one hand, given that $\boldsymbol{X}$ is fixed and the factor $(\boldsymbol{x}_i - \boldsymbol{x}_j)$ points towards $\boldsymbol{x}_i$, all IF directions are also fixed and attracting-natured. On the other hand, since F_i turns out to be dependent on the free parameter σ, the IP and all IF magnitudes become functions of the Gaussian kernel bandwidth. In fact, the IP follows a monotonically decreasing behavior over σ, while the conditional IF magnitude tends to zero as σ goes either to zero or infinite and reaching its maximum at some value in $\mathbb{R}^+$. Hence, the importance of an adequate Gaussian kernel bandwidth tuning becomes clear.

In this sense, we propose a novel kernel function estimation from the observed data $\boldsymbol{X}$, using the Gaussian Parzen estimate in Eq. (2). Namely, we seek for an RKHS maximizing the overall information potential variability with respect to the kernel bandwidth parameter, so that all IF magnitudes spread the most widely on $\mathcal{X}$. To this end, the variability of the estimated pdf $p_X(\boldsymbol{x}|\sigma)$ is maximized in terms of the kernel bandwidth parameter in the form:

$$\sigma^* = \arg\max_{\sigma} \operatorname{var}\{p_X(\boldsymbol{x}|\sigma)\} \ , \tag{6}$$

where $\operatorname{var}\{p_X(\boldsymbol{x}|\sigma)\} = \boldsymbol{E}\left\{(p_X(\boldsymbol{x}|\sigma) - \boldsymbol{E}\{p_X(\boldsymbol{x}|\sigma)\})^2 : \forall \boldsymbol{x}{\in}\boldsymbol{X}\right\}$.

Deriving with respect to σ, the optimal parameter value can be rewritten in terms of the before introduced Gaussian-based Renyi's Information Metrics as follows:

$$
\begin{aligned}
\frac{d}{d\sigma}\mathrm{var}\left\{p_X(\boldsymbol{x}|\sigma)\right\} &= \frac{2}{N^2\sigma^3}\left(1+\frac{1}{N}\right)\left(\sum_{i,j=1}^{N} g^2\{\boldsymbol{x}_i-\boldsymbol{x}_j,\sigma^2\}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2\right.\\
&\quad\left.-\left(\sum_{i,j=1}^{N} g\{\boldsymbol{x}_i-\boldsymbol{x}_j,\sigma^2\}\right)\left(\sum_{i,j=1}^{N} g\{\boldsymbol{x}_i-\boldsymbol{x}_j,\sigma^2\}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2\right)\right),\\
&= \frac{2(N^2+N)}{\sigma}\left(\sigma^2\sum_{i,j=1}^{N} F^2(\boldsymbol{x}_i|\boldsymbol{x}_j) - V(\boldsymbol{X})\sum_{i,j=1}^{N}\left(F(\boldsymbol{x}_i|\boldsymbol{x}_j)\right)^{\top}(\boldsymbol{x}_i-\boldsymbol{x}_j)\right)
\end{aligned}
$$

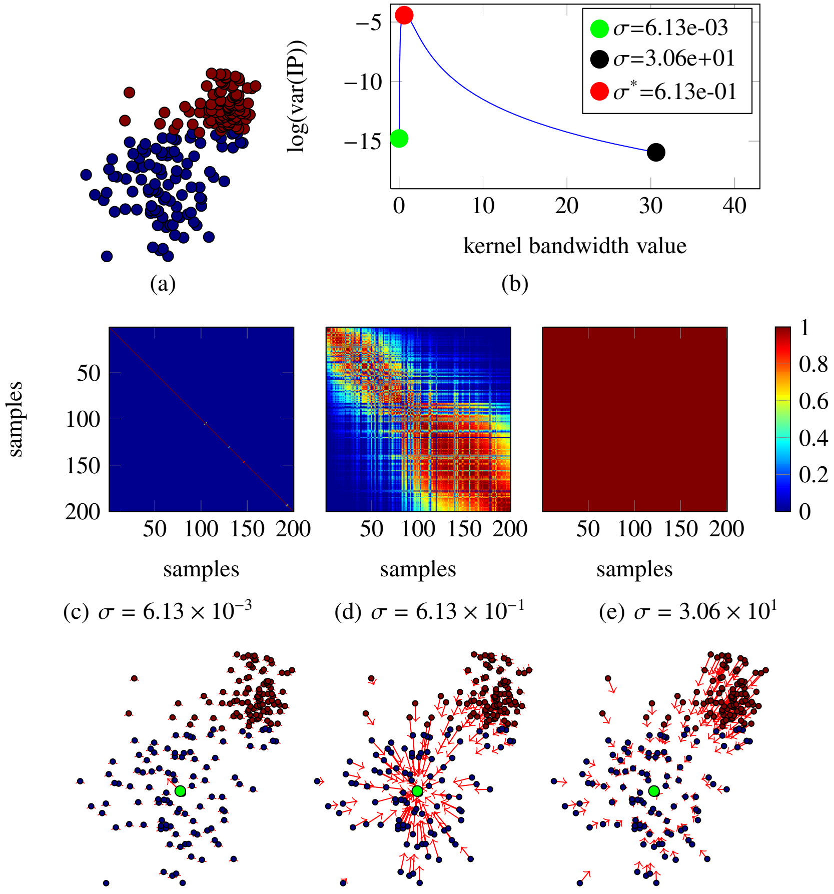
Lastly, equating the above equation to zero, the fixed point update rule becomes:

$$
\sigma^2(k+1) = \frac{V_k(\boldsymbol{X})\boldsymbol{E}\left\{(F_k(\boldsymbol{x}_i|\boldsymbol{x}_j))^{\top}(\boldsymbol{x}_i-\boldsymbol{x}_j) : \forall i,j\in[1,N]\right\}}{\boldsymbol{E}\left\{F_k^2(\boldsymbol{x}_i|\boldsymbol{x}_j) : \forall i,j\in[1,N]\right\}}, \tag{7}
$$

where $V_k(\boldsymbol{X})$ and $F_k(\boldsymbol{x}_i|\boldsymbol{x}_j)$ are the IP and conditional IF obtained when $\sigma = \sigma(k)$, respectively. As a result, we get a scale updating rule as a function of the IFs, which are induced by a kernel function applied over a finite sample set. Thereby, a Gaussian kernel-based RKHS coding the most spread out IF magnitudes can be estimated from Eq. (7), approach that we term as *Kernel Function Estimation from Information Potential Variability* - KEIPV.

3 Experimental Set-up and Results

We test the proposed KEIVP approach on both synthetic and real-world datasets for the concrete case of a clustering task. The former is a toy set holding two multivariate Gaussian distributions (see Fig. 1(a)): $f_1\,(x)=\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ and $f_2\,(x)=\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$, with parameters $\boldsymbol{\mu}_1 = \boldsymbol{0}$, $\boldsymbol{\mu}_2=\boldsymbol{1}$, $\boldsymbol{\Sigma}_1=0.5\boldsymbol{I}$ and $\boldsymbol{\Sigma}_2=0.25\boldsymbol{I}$, with $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2\in\mathbb{R}^2$ and $\boldsymbol{I}\in\mathbb{R}^{2\times2}$. To get the input sample set $\boldsymbol{X}\in\mathbb{R}^{200\times2}$, one hundred samples are randomly drawn from each of both simulated pdfs. As seen in Figs. 1(b) and 1(f) to 1(h), the IP variability cost function allows identifying different IF configurations. Particularly, for a narrow bandwidth value, particles are forced to apart each other due to the kernel function strongly reduces the scaling of the Euclidean-based distance between particles. Hence, low similarities between pair-wise samples and low magnitude IFs are estimated, as shown in Figs. 1(c) and 1(f). In contrast, employing a wide bandwidth value yields to an RKHS where all particles are attracted each other. Namely, the Euclidean distance scaling is strongly increased, which leads to a data representation space where all samples are closed similar, as seen in Fig. 1(e). Such a fact is shown in the IF distribution in Fig. 1(h), where red cluster particles are more attracted to the green particle. Note that low IP variability values are achieved for both narrow and wide bandwidths because, in either case, all the IFs tend to share the same magnitude regardless their direction. Therefore, the proposed KEIVP finds an RKHS where data samples share widely spread IF magnitudes, that is, close particles according to the Euclidean distance get high pairwise similarities and IFs while far ones have low similarities and IFs (see Figs. 1(d) and 1(g)).



(a) (b)

(c) $\sigma = 6.13 \times 10^{-3}$ (d) $\sigma = 6.13 \times 10^{-1}$ (e) $\sigma = 3.06 \times 10^{1}$

(f) $\sigma = 6.13 \times 10^{-3}$ (g) $\sigma = 6.13 \times 10^{-1}$ (h) $\sigma = 3.06 \times 10^{1}$

Fig. 1. KEIVP illustrative example. a) Multivariate Gaussian toy set. b) log of IP variability versus bandwidth. **2nd row**: Gaussian kernel for the toy set. **3rd row**: IFs acting on a fixed particle (green). Narrow (**1st column**), KEIVP (**2nd column**) and wide (**3rd column**) bandwidth values.

Also, to provide visual inspection on unsupervised clustering, three well-known synthetic databases are used that represent challenging clustering tasks due to their complex structures: *Bull's eyes*, *Circle with squares*, and *Noisy squares* (see Fig. 2 rows one, two, and three, respectively). Here, three baseline approaches for estimating the Gaussian kernel bandwidth parameter are considered: *i)* The *Sylverman's* rule criterion that computes the scale value as $\sigma_S = \sigma_X \left(4N^{-1}(2P+1)^{-1}\right)^{1/(P+4)}$, with $\sigma_X{=}\sum_{i\in N} s_{ii}$ and being s_{ii} the diagonal elements of the sample covariance matrix [9]. *ii)* The *Self-Tuning Spectral Clustering* (STSC) estimator that calculates a local scale parameter for each pair of samples $(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $i \neq j$, by considering nearest neighbor distances as: $\sigma_{sc}^{i,j} = \|\boldsymbol{x}_i - \boldsymbol{x}_K^i\|_2 \|\boldsymbol{x}_j - \boldsymbol{x}_K^j\|_2$, being $\boldsymbol{x}_K^i$ the K-th nearest neighbor of $\boldsymbol{x}_i$ in terms of the Euclidean distance and $\|\cdot\|_2$ stands for the L2-norm [11]. *iii)* The local density adaptive band-width is also tested, which computes a local scale parameter as function of *Common Near Neighbors* (CNN) between points $(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $i \neq j$, as: $\sigma^{i,j}_{cnn} = \sigma_o\left(\gamma\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + 1\right)^{1/2}$, where $\sigma_o \in \mathbb{R}^+$ and $\gamma\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \left|\Gamma_i \cap \Gamma_j\right|$, being $\Gamma_i = \{\boldsymbol{x}^i_k : k=1, \ldots, K\}$ the set holding the K nearest neighbors of $\boldsymbol{x}_i$ according to the Euclidean distance and $|\cdot|$ stands for the cardinality operator [12]. Here, $\sigma_o = \text{median}\{\sigma^{i,j}_{sc}\}$, $i < j$. For each of above introduced bandwidth selection approaches, namely Sylverman, STSC, CNN, and KEIVP, the resulting Gaussian kernel is employed to perform the unsupervised clustering learning by means of the well-known Spectral Clustering technique [5]. Additionally, the number of neighbors is fixed as $K = \sqrt{N}$ in cases of STSC and CNN. For concrete testing, the number of groups $C \in \mathbb{N}$ is fixed as three, three, and five, respectively. Furthermore, for fair comparison purposes, the KEIVP approach is calculated only considering data relationships (distances) belonging to connected samples according to a K-nearest graph. Figs. 2(b) to 2(d) show that both local scaling-based strategies (STSC and CNN) as well as the proposed KEIPV are able to deal with the *Bull's eyes* structure. Such approaches also correctly perform grouping of the *Noisy squares* dataset, as seen in Figs. 2(j) to 2(l). That is, local scaling-based techniques are able to approximate nonlinear structures from linear analysis over each sample neighborhood. Nonetheless, STSC performs wrong clustering for the *Circle with squares* (see Fig. 2(f)). These results can be explained by the fact that local scaling approximations lead to wrong cluster connections when dealing with data structures with highly varying densities. Similarly, CNN suffers of the same drawback, but the σ_o parameter can deal with it if properly fixed. Nonetheless, finding a suitable neighborhood size is a difficult task for the user, not mentioning that using different bandwidth values for each pair-wise sample similarity when estimating a Gaussian kernel does not guarantee a positive definite kernel function, violating the Mercer's conditions [6]. Regarding to the Sylverman-based estimation results, this method generally yields a biased RKHS due to its statistical assumptions, resulting in wrong clustering performances (see Figs. 2(a), 2(e) and 2(i)). In turn, KEIPV is able to find an RKHS coding widely spread IF magnitudes, allowing to close samples belonging to a similar structure while repelling distant points (see fourth column of Fig. 2).

Finally, the real-world databases from the Machine Learning UCI Repository[1] are tested as supervised clustering task (see Table 1). In this case, each computed kernel is used as similarity representation to learn a classification boundary based on the well-known k-nearest-neighbors classifier. A 10-folds-cross-validation strategy is carry out to validate the stability of each kernel function estimation approach. Furthermore, the k parameter is fixed from the set $\{1, 3, 5, 7, 9, 11\}$ according to the training error. As seen in Fig. 3, the proposed KEIPV allows to compute an RKHS favoring the cluster separability. STSC and CNN algorithms get competitive results in terms of classification accuracy. Nonetheless, they need a suitable graph representation, which practically can be difficult to estimate. Moreover, their local scaling approximation of the Gaussian kernel can not be correct theoretically as mentioned before. Again, the Sylverman's rule estimation suffers of biased kernel representations, particularly, when the input dimensionality P is considerably high (see obtained results by the `mnist` and `orl` datasets).

[1] `http://archive.ics.uci.edu/ml/`

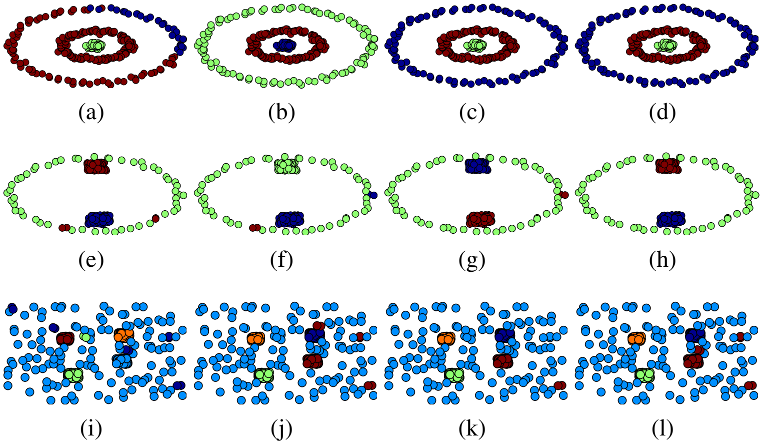


Fig. 2. Synthetic data sets clustering results. **First column**: Sylverman's rule. **Second column**: STSC. **Third column**: CNN. **Fourth column**: KEIPV.

Table 1. Employed UCI dataset description

Dataset	iris	sonar	mnist	orl	diabetes	breast	arrhythmia	ionosphere	heart	wine	glass
N	150	208	1000	400	768	699	420	351	303	178	214
P	4	60	784	10304	8	9	278	34	13	13	9
C	3	2	10	40	2	2	13	2	2	3	4

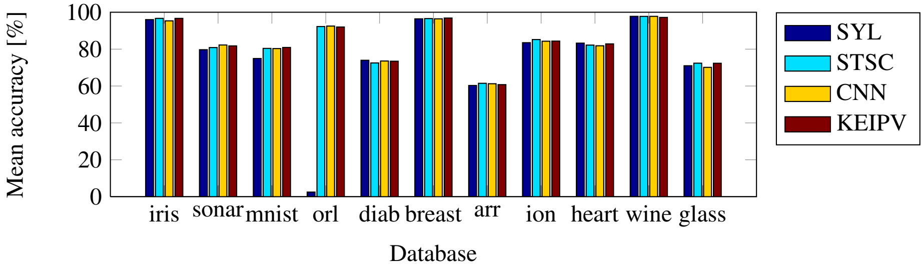


Fig. 3. Classification results using the fourth bandwidth selection approaches

4 Concluding Remarks

A new kernel function estimation based on an information potential variability framework is presented. Our approach, termed KEIPV, aims to estimate an RKHS to span the most widely information force magnitudes among data points. Particularly, KEIPV relates different kernel functions with the intrinsic information potential variations in Parzen-based pdf estimations [7]. Thereby, we seek for an RKHS that maximizes the overall information potential variability with respect to the global kernel parameter. As a case of interest, an updating rule for estimating the Gaussian kernel bandwidth parameter is proposed as a function of the forces induced by the distances among samples. Proposed strategy is tested on both unsupervised and supervised clustering tasks.

Performed results show that the presented approach allows computing RKHS's favoring data groups separability in comparison with other state-of-the-art alternatives.

Acknowledgments. This work was supported by *Programa Nacional de Formación de Investigadores "Generación del Bicentenario"*, 2011/2012 funded by COLCIENCIAS and the project "Desarrollo de un sistema automático de mapeo cerebral para el monitoreo y tratamiento de enfermedades neurodegenerativas" (22506) funded by Universidad Nacional de Colombia.

References

1. Belanche, L.: Developments in kernel design. In: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN, pp. 369–378 (2013)
2. Jenssen, R., Principe, J.C., Eltoft, T.: Information cut and information forces for clustering. In: NNSP, pp. 459–468 (September 2003)
3. Liu, W., Principe, J.C., Haykin, S.: Kernel Adaptive Filtering: A Comprehensive Introduction, vol. 57. John Wiley & Sons (2011)
4. Morejon, R.A., Principe, J.C.: Advanced search algorithms for information-theoretic learning with kernel-based estimators, 874–884 (2004)
5. Ng, A.Y., Jordan, M.I., Weiss, Y.: On Spectral Clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems 14 (2001)
6. Pokharel, R., Seth, S., Principe, J.C.: Mixture kernel least mean square. In: The 2013 International Joint Conference on Neural Networks (IJCNN), pp. 1–7. IEEE (2013)
7. Principe, J.C.: Information theoretic learning: Rényi's entropy and kernel perspectives. Springer (2010)
8. Giraldo, L.G.S., Principe, J.C.: Information theoretic learning with infinitely divisible kernels. arXiv preprint arXiv:1301.3551 (2013)
9. Silverman, B.W.: Density estimation for statistics and data analysis, vol. 26. CRC Press (1986)
10. Singh, A., Principe, J.C.: Kernel width adaptation in information theoretic cost functions. In: 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 2062–2065 (2010)
11. Zelnik-Manor, L., Perona, P.: Self-Tuning Spectral Clustering. Advances in Neural Information Processing Systems 17(2), 1601–1608 (2004)
12. Zhang, X., Li, J., Yu, H.: Local density adaptive similarity measurement for spectral clustering. Pattern Recognition Letters 32(2), 352–358 (2011)
13. Zhao, S., Chen, B., Principe, J.C.: An adaptive kernel width update for correntropy. In: The 2012 International Joint Conference on Neural Networks (IJCNN), pp. 1–5. IEEE (2012)