A clusterwise supervised learning procedure based on aggregation of distances applied on Energy data

Sothea HAS

sothea.has@lpsm.paris

June 24, 2020

Motivation

- Modeling is a common tool used in many real-world prediction problems especially in the domain of Energy.

Motivation

- Modeling is a common tool used in many real-world prediction problems especially in the domain of Energy.
- Building an accurate model with generalization capabilities is not an easy task and may require information of unknown data structure mostly hard to recover.

Motivation

- Modeling is a common tool used in many real-world prediction problems especially in the domain of Energy.
- Building an accurate model with generalization capabilities is not an easy task and may require information of unknown data structure mostly hard to recover.
- With the aim to automatically combine efficiently clustering and modeling, we propose the KFC procedure to effectively solve this problem.

Motivation

- Modeling is a common tool used in many real-world prediction problems especially in the domain of Energy.
- Building an accurate model with generalization capabilities is not an easy task and may require information of unknown data structure mostly hard to recover.
- With the aim to automatically combine efficiently clustering and modeling, we propose the KFC procedure to effectively solve this problem.
- Excellent performances of the KFC procedure were obtained on many real datasets especially in the Energy domain for air compressor and wind turbine.

Outline

A. Introduction

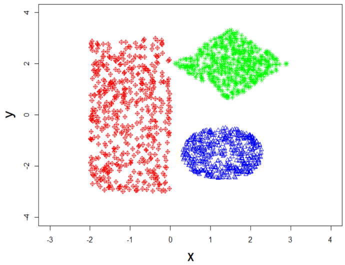
B. KFC procedure

1. K-step: K-means algorithm with Bregman divergences
2. F-step: Fitting Candidate Models
3. C-step: Consensual Aggregation

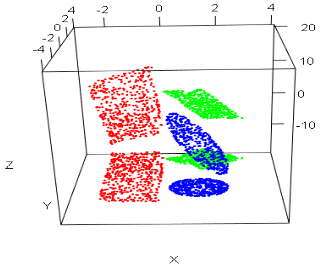
C. Applications on the Energy domain

1. Air compressor
2. Wind turbine

Consider an example...



Input data with 3 clusters



Different model on each cluster

x	y	z
x_1	y_1	z_1
x_2	y_2	z_2
...	...	...
x_n	y_n	z_n

Introduction

Setting:

- $(X, Z) \in \mathcal{X} \times \mathcal{Z}$: input-out data.
  - $\mathcal{X} = \mathbb{R}^d$: input space.
  - $\mathcal{Z} = \begin{cases} \mathbb{R} & \text{: regression} \\ \{0, 1\} & \text{: binary classification} \end{cases}$
- $\mathcal{D}_n = \{(x_i, z_i)_{i=1}^n\}$: *iid* learning data.

Introduction

Setting:

- $(X, Z) \in \mathcal{X} \times \mathcal{Z}$: input-out data.
  - $\mathcal{X} = \mathbb{R}^d$: input space.
  - $\mathcal{Z} = \begin{cases} \mathbb{R} & : \text{regression} \\ \{0, 1\} & : \text{binary classification} \end{cases}$
- $\mathcal{D}_n = \{(x_i, z_i)_{i=1}^n\}$: *iid* learning data.

Objective:

Construct a good predictive model for regression or classification.

Introduction

Setting:

- $(X, Z) \in \mathcal{X} \times \mathcal{Z}$: input-out data.
  - $\mathcal{X} = \mathbb{R}^d$: input space.
  - $\mathcal{Z} = \begin{cases} \mathbb{R} & \text{: regression} \\ \{0, 1\} & \text{: binary classification} \end{cases}$
- $\mathcal{D}_n = \{(x_i, z_i)_{i=1}^n\}$: *iid* learning data.

Objective:

Construct a good predictive model for regression or classification.

Assumption:

- X is composed of more than one group or cluster.
- The number of clusters K is available.
- There exists **different underlying models** on these clusters.

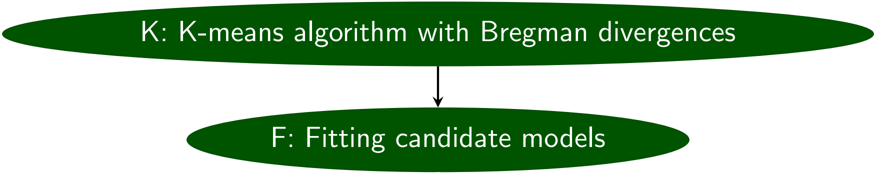
KFC procedure

KFC procedure consists of 3 important steps:

K: K-means algorithm with Bregman divergences

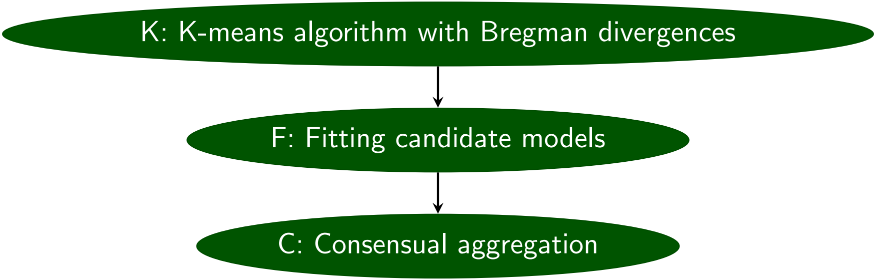
KFC procedure

KFC procedure consists of 3 important steps:



KFC procedure

KFC procedure consists of 3 important steps:



Bregman divergences (BD) [Bregman, 1967]

$\phi : \mathcal{C} \subset \mathbb{R}^d \to \mathbb{R}$, strictly convex and of class $\mathcal{C}^1$ then for any $(x, y) \in \mathcal{C} \times int(\mathcal{C})$ (points of the input space $\mathcal{X}$),

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle$$

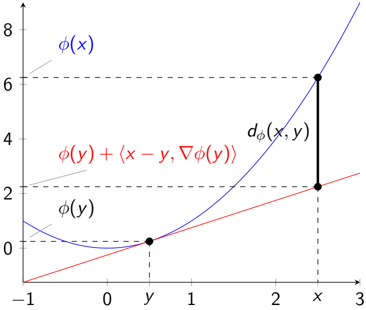


Figure: Graphical interpretation of Bregman divergences.

K-step: K-means Algorithm with Bregman Divergences

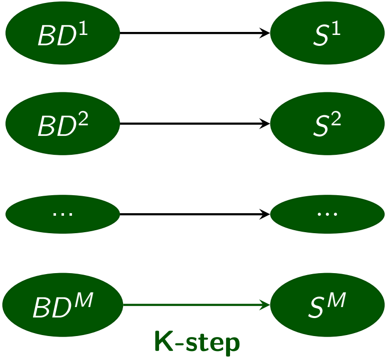
- Perform K-means algorithm with M options of Bregman divergences.

K-step: K-means Algorithm with Bregman Divergences

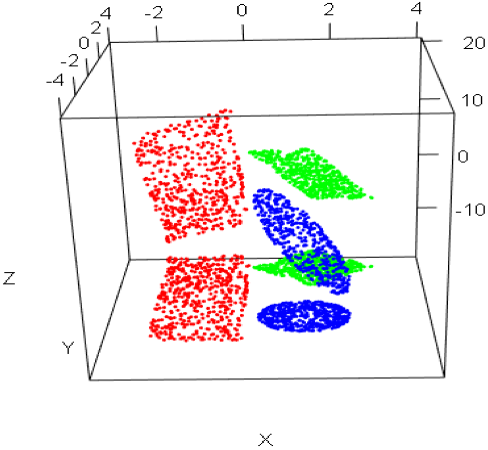
- Perform K-means algorithm with M options of Bregman divergences.
- Each BD^ℓ gives an associated partition cell $S^\ell = \{S_k^\ell\}_{k=1}^K$.

K-step: K-means Algorithm with Bregman Divergences

- Perform K-means algorithm with M options of Bregman divergences.
- Each BD^ℓ gives an associated partition cell $S^\ell = \{S_k^\ell\}_{k=1}^K$.



Recall something...



Here, we need 3 local models to explain Z.

F-step: Fitting Candidate Models

- Suppose that $\forall \ell, k : S_k^\ell \in S^\ell$ contains enough data points.

F-step: Fitting Candidate Models

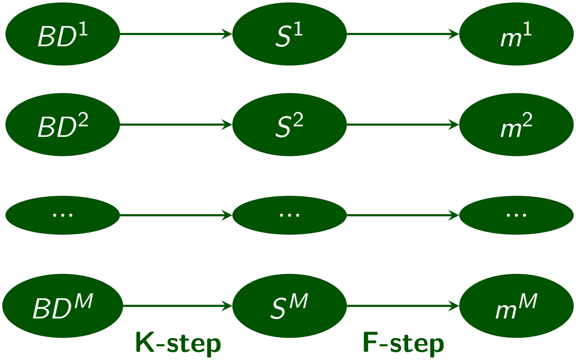
- Suppose that $\forall \ell, k : S_k^\ell \in S^\ell$ contains enough data points.
- $\forall \ell, k$: construct an estimator m_k^ℓ on S_k^ℓ.

F-step: Fitting Candidate Models

- Suppose that $\forall \ell, k : S_k^\ell \in S^\ell$ contains enough data points.
- $\forall \ell, k$: construct an estimator m_k^ℓ on S_k^ℓ.
- $m^\ell = \{m_k^\ell\}_{k=1}^K$ is the candidate estimator associated to DB^ℓ.

F-step: Fitting Candidate Models

- Suppose that $\forall \ell, k : S_k^\ell \in S^\ell$ contains enough data points.
- $\forall \ell, k$: construct an estimator m_k^ℓ on S_k^ℓ.
- $m^\ell = \{m_k^\ell\}_{k=1}^K$ is the candidate estimator associated to DB^ℓ.



C-step: Consensual Aggregation

Note that

C-step: Consensual Aggregation

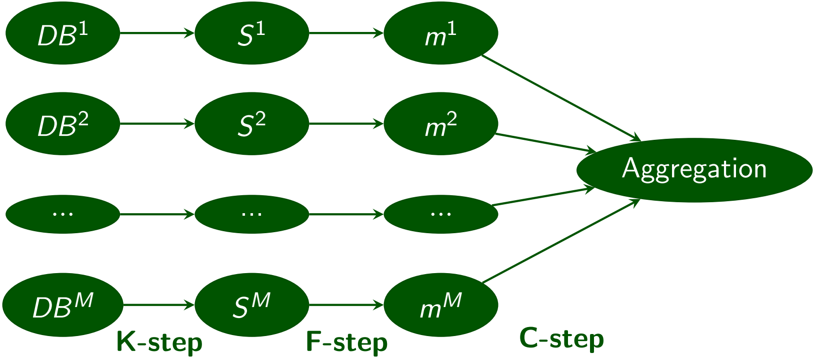
Note that

- neither the distribution nor the clustering structure of the input data is available.
- it is not easy to choose the "best" one among $\{m^\ell\}_{\ell=1}^M$.

C-step: Consensual Aggregation

Note that

- neither the distribution nor the clustering structure of the input data is available.
- it is not easy to choose the "best" one among $\{m^\ell\}_{\ell=1}^M$.



Classification

Example:

- Suppose we have 4 classifiers: $\mathbf{m} = (m^1, m^2, m^3, m^4)$
- An observation x with predictions: $(1, 1, 0, 1)$.

ID	m^1	m^2	m^3	m^4	z
1	1	1	0	1	1
2	0	0	0	1	0
3	1	1	0	1	0
4	1	0	1	1	1
5	1	1	0	1	1

Table: Table of predictions.

Classification

Based on the following works:

1. [Mojirsheibani, 1999]: Classical method.

$$Comb_1^C(x) = \mathbb{1}_{\left\{ \sum_{(x_i,y_i)\in\mathcal{D}_n}(2y_i-1)\mathbb{1}_{\{\mathbf{m}(x_i)=\mathbf{m}(x)\}}>0 \right\}}$$

2. [Mojirsheibani, 2000]: A kernel-based method, for any $h > 0$:

$$Comb_2^C(x) = \mathbb{1}_{\left\{ \sum_{(x_i,y_i)\in\mathcal{D}_n}(2y_i-1)K_h\left(d_{\mathcal{H}}(\mathbf{m}(x_i),\mathbf{m}(x))\right)>0 \right\}}, K(x) = e^{-\|x\|^2}$$

3. [Fischer and Mougeot, 2019]: MixCOBRA, for any $\alpha, \beta > 0$:

$$Comb_3^C(x) = \mathbb{1}_{\left\{ \sum_{(x_i,y_i)\in\mathcal{D}_n}(2y_i-1)K(\frac{x_i-x}{\alpha},\frac{\mathbf{m}(x_i)-\mathbf{m}(x)}{\beta})>0 \right\}}$$

Regression

The aggregation takes the following form:

$$Agg_n(x) = \sum_{i=1}^{n} W_{n,i}(x) z_i$$

Regression

The aggregation takes the following form:

$$Agg_n(x) = \sum_{i=1}^{n} W_{n,i}(x) z_i$$

1. [Biau et al., 2016]: with weight $0 - 1$ (COBRA).

$$W_{n,i}(x) = \frac{\prod_{\ell=1}^{M} \mathbb{1}_{\{|m^\ell(x_i) - m^\ell(x)| < \varepsilon\}}}{\sum_{j=1}^{n} \prod_{\ell=1}^{M} \mathbb{1}_{\{|m^\ell(x_j) - m^\ell(x)| < \varepsilon\}}}$$

2. Kernel-based method of COBRA (kernel-based weight): for any $h > 0$,

$$W_{n,i}(x) = \frac{K_h\Big(\mathbf{m}(x_i) - \mathbf{m}(x)\Big)}{\sum_{j=1}^{n} K_h\Big(\mathbf{m}(x_j) - \mathbf{m}(x)\Big)}$$

for some kernel function K with $K_h(x) = K(x/h)$.

3. [Fischer and Mougeot, 2019]: MixCOBRA.

Applications on the Energy domain

Bregman divergences

- Euclidean: For all $x \in \mathcal{C} = \mathbb{R}^d$, $\phi(x) = \|x\|_2^2 = \sum_{i=1}^d x_i^2$,
  $d_\phi(x, y) = \|x - y\|_2^2$
- General Kullback-Leibler (GKL): $\phi(x) = \sum_{i=1}^d x_i \log(x_i)$, $\mathcal{C} = (0, +\infty)^d$,
  $d_\phi(x, y) = \sum_{i=1}^d \left[x_i \log\left(\frac{x_i}{y_i}\right) - (x_i - y_i) \right]$
- Logistic: $\phi(x) = \sum_{i=1}^d [x_i \log(x_i) + (1 - x_i) \log(1 - x_i)]$, $\mathcal{C} = (0, 1)^d$,
  $d_\phi(x, y) = \sum_{i=1}^d \left[x_i \log\left(\frac{x_i}{y_i}\right) + (1 - x_i) \log\left(\frac{1-x_i}{1-y_i}\right) \right]$
- Itakura-Saito: $\phi(x) = -\sum_{i=1}^d \log(x_i)$, $\mathcal{C} = (0, +\infty)^d$,
  $d_\phi(x, y) = \sum_{i=1}^d \left[\frac{x_i}{y_i} - \log\left(\frac{x_i}{y_i}\right) - 1 \right]$
- Polynomial: $\phi(x) = \sum_{i=1}^d |x_i|^p$, $\mathcal{C} = \mathbb{R}^d, p \geq 1$,
  $d_\phi(x, y) = \sum_{i=1}^d (|x_i|^p - |y_i|^p) + p \sum_{i=1}^d (-1)^{\mathbb{1}_{\{y_i < 0, p \text{ is odd}\}}} (x_i - y_i) y_i^{p-1}$

K-means with BD on some simulated datasets

$M = 4$ et $K = 3$.

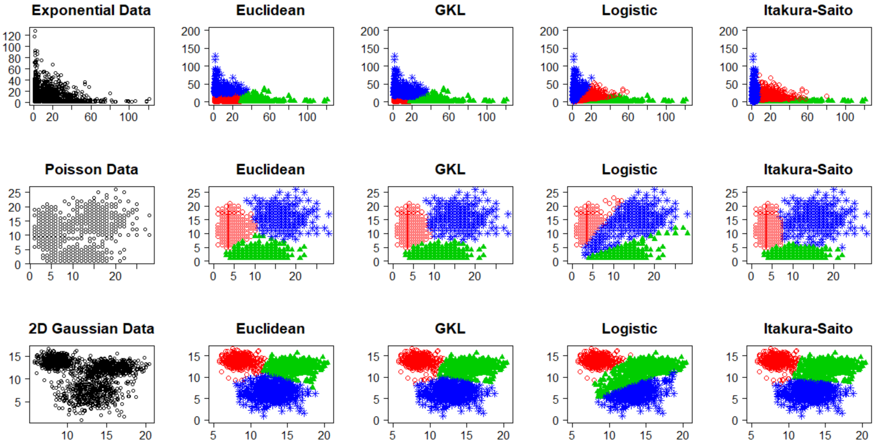


Figure: K-means with Bregman divergences on some simulated data.

Air compressor data

- Provided by [Cadet et al., 2005].
- Six predictors: air temperature, input pressure, output pressure, flow and water temperature.
- Response variable: power consumption.

Air compressor data

- Provided by [Cadet et al., 2005].
- Six predictors: air temperature, input pressure, output pressure, flow and water temperature.
- Response variable: power consumption.
- In real-world problems, K is usually **not available** and it is the case here!

Performance on air compressor

K	**Euclid**	**GKL**	**Logistic**	**Ita**	**KFC$_1$ (Gaussian)**	**KFC$_2$ (Gaussian)**
2	158.85 (6.42)	158.90 (6.48)	159.35 (6.71)	158.96 (6.41)	153.34 (6.72)	**116.69** (5.86)
3	157.38 (6.95)	157.24 (6.84)	156.99 (6.65)	157.24 (6.85)	153.69 (6.64)	**117.45** (5.55)
4	154.33 (6.69)	153.96 (6.74)	153.99 (6.45)	154.07 (7.01)	152.09 (6.58)	**117.16** (5.99)
5	153.18 (6.91)	153.19 (6.77)	152.95 (6.57)	152.25 (6.70)	151.05 (6.76)	**117.55** (5.90)
6	151.16 (6.91)	151.67 (6.96)	151.89 (6.62)	151.75 (6.57)	150.27 (6.82)	**117.74** (5.86)
7	151.08 (6.77)	150.99 (6.84)	152.81 (7.11)	151.85 (6.61)	150.46 (6.87)	**117.58** (6.15)
8	151.27 (7.17)	151.09 (7.01)	152.07 (6.65)	150.90 (6.96)	150.21 (7.03)	**117.91** (5.83)

Table: Performances of the KFC procedure.

Multiple LR	**22-NN**	**RF (500)**	**Boosting (500)**
178.67 (5.18)	292.08 (9.17)	217.14 (9.80)	158.92 (4.33)

Table: Performances of alternative models.

Performance on air compressor

K	Euclid	GKL	Logistic	Ita	KFC_1 (Gaussian)	KFC_2 (Gaussian)
2	158.85 (6.42)	158.90 (6.48)	159.35 (6.71)	158.96 (6.41)	153.34 (6.72)	**116.69** (5.86)
3	157.38 (6.95)	157.24 (6.84)	156.99 (6.65)	157.24 (6.85)	153.69 (6.64)	**117.45** (5.55)
4	154.33 (6.69)	153.96 (6.74)	153.99 (6.45)	154.07 (7.01)	152.09 (6.58)	**117.16** (5.99)
5	153.18 (6.91)	153.19 (6.77)	152.95 (6.57)	152.25 (6.70)	151.05 (6.76)	**117.55** (5.90)
6	151.16 (6.91)	151.67 (6.96)	151.89 (6.62)	151.75 (6.57)	150.27 (6.82)	**117.74** (5.86)
7	151.08 (6.77)	150.99 (6.84)	152.81 (7.11)	151.85 (6.61)	150.46 (6.87)	**117.58** (6.15)
8	151.27 (7.17)	151.09 (7.01)	152.07 (6.65)	150.90 (6.96)	150.21 (7.03)	**117.91** (5.83)

Table: Performances of the KFC procedure.

Multiple LR	22-NN	RF (500)	Boosting (500)
178.67 (5.18)	292.08 (9.17)	217.14 (9.80)	158.92 (4.33)

Table: Performances of alternative models.

∗ Even though K is not available, the KFC procedure still performs well on this dataset.

Wind turbine

- Provided by Maïa Eolis (see [Fischer et al., 2017]).
- Six predictors: wind speed (real part, imaginary part, and strength), wind direction (sine and cosine) and temperature.
- Response variable: power.

Wind turbine

- Provided by Maïa Eolis (see [Fischer et al., 2017]).
- Six predictors: wind speed (real part, imaginary part, and strength), wind direction (sine and cosine) and temperature.
- Response variable: power.
- And again, we don't know K.

Performance on wind turbine

K	Euclid	Poly	KFC_1 (Gaussian)	KFC_2 (Gaussian)
2	62.15 (3.01)	62.74 (2.78)	38.73 (2.05)	**36.09** (1.11)
3	62.54 (4.03)	64.21 (7.01)	38.88 (2.62)	**37.18** (3.09)
4	59.73 (4.15)	61.73 (6.08)	38.79 (2.81)	**36.49** (2.11)
5	54.52 (5.98)	56.74 (2.31)	38.68 (2.55)	**36.62** (2.02)
6	53.25 (2.69)	57.19 (7.71)	39.05 (2.81)	**36.83** (2.37)
7	51.34 (4.00)	55.67 (5.91)	38.61 (2.60)	**36.78** (2.28)
8	49.76 (5.31)	55.94 (7.21)	38.76 (2.56)	**36.55** (2.22)

Table: Performances of the KFC procedure.

Multiple LR	7-NN	RF (500)	Boosting (500)
69.46 (3.295)	40.30 (1.447)	37.26 (1.316)	41.65 (1.424)

Table: Performances of alternative models.

Performance on wind turbine

K	**Euclid**	**Poly**	**KFC$_1$ (Gaussian)**	**KFC$_2$ (Gaussian)**
2	62.15 (3.01)	62.74 (2.78)	38.73 (2.05)	**36.09** (1.11)
3	62.54 (4.03)	64.21 (7.01)	38.88 (2.62)	**37.18** (3.09)
4	59.73 (4.15)	61.73 (6.08)	38.79 (2.81)	**36.49** (2.11)
5	54.52 (5.98)	56.74 (2.31)	38.68 (2.55)	**36.62** (2.02)
6	53.25 (2.69)	57.19 (7.71)	39.05 (2.81)	**36.83** (2.37)
7	51.34 (4.00)	55.67 (5.91)	38.61 (2.60)	**36.78** (2.28)
8	49.76 (5.31)	55.94 (7.21)	38.76 (2.56)	**36.55** (2.22)

Table: Performances of the KFC procedure.

Multiple LR	**7-NN**	**RF (500)**	**Boosting (500)**
69.46 (3.295)	40.30 (1.447)	37.26 (1.316)	41.65 (1.424)

Table: Performances of alternative models.

$*$ Similarly, the KFC procedure also performs well in this case, even without the knowledge of K.

Conclusion

- Several simulations carried out on different simulated and real data have shown that the KFC procedure provides remarkable responses in many prediction problems involving clustering and modeling.

Conclusion

- Several simulations carried out on different simulated and real data have shown that the KFC procedure provides remarkable responses in many prediction problems involving clustering and modeling.
- In particular, we obtain its excellent performances on the domain of Energy for air compressor and wind turbine.

References I

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. (2005).
Clustering with Bregman divergences.
Journal Machine Learning Research, 6:1705–1749.

Biau, G., Fischer, A., Guedj, B., and Malleye, J. D. (2016).
COBRA: a combined regression strategy.
Journal of Multivariate Analysis, 146:18–28.

Bregman, L. M. (1967).
The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming.
USSR Computational Mathematical and Mathematical Physics, 7:200–217.

Cadet, O., Harper, C., and Mougeot, M. (2005).
Monitoring energy performance of compressors with an innovative auto-adaptive approach.
In *Instrumentation System and Automation -ISA- Chicago*.

Fischer, A., Has, S., and Mougeot, M. (2018).
a clusterwise supervised learning procedure based on aggregation of distances.

Fischer, A., Montuelle, L., Mougeot, M., and Picard, D. (2017).
Statistical learning for wind power: A modeling and stability study towards forecasting.
Wiley online library, 20.

Fischer, A. and Mougeot, M. (2019).
Aggregation using input-output trade-off.
Journal of Statistical Planning and Inference, 200:1–19.

References II

Mojirsheibani, M. (1999).
Combined classifiers via disretization.
Journal of the American Statistical Association, 94(446):600–609.

Mojirsheibani, M. (2000).
A kernel-based combined classification rule.
Journal of Statistics and Probability Letters, 48(4):411–419.

Thank you!