Fuzzy logic application to data mining in sky surveys

Artem Poliszczuk

3rd Cosmology School

Cracow 2017

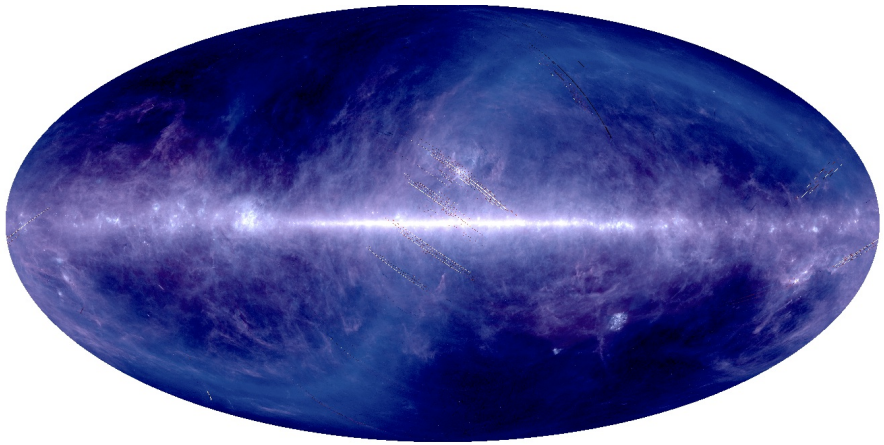
Sky Surveys and Big Data Challenge

- Provide statistical information about astronomical objects.
- Searching for the rare objects.
- One of the main tools of the observational cosmology.

Types

- spectroscopic
- photometric

- wide
- deep



Sky Surveys and Big Data Challenge

Sky Survey Projects	Data Volume
DPSS (The Palomar Digital Sky Survey) 1994	3 TB
2MASS (The Two Micron All Sky Survey) 1997	10 TB
GBT (Green Bank Telescope) 2001	20 TB
GALEX (The Galaxy Evolution Explorer) 2003	30 TB
SDSS (The Sloan Digital Sky Survey) 2005	40 TB
SkyMapper Southern Sky Survey 2008	500 TB
PanSTARRS ? (The Panoramic Survey Telescope Rapid Response System)	40 PB expected
LSST (The Large Synoptic Survey Telescope) 2021	200 PB expected
SKA (The Square Kilometer Array) 2030	4.6 EB expected

Tabela: Data volumes of different sky surveys.

źródło: Zhang, Zhao:*Astronomy in the Big Data Era*, Data Science Journal,2015.

New approach: machine learning algorithms (classification and regression tasks)

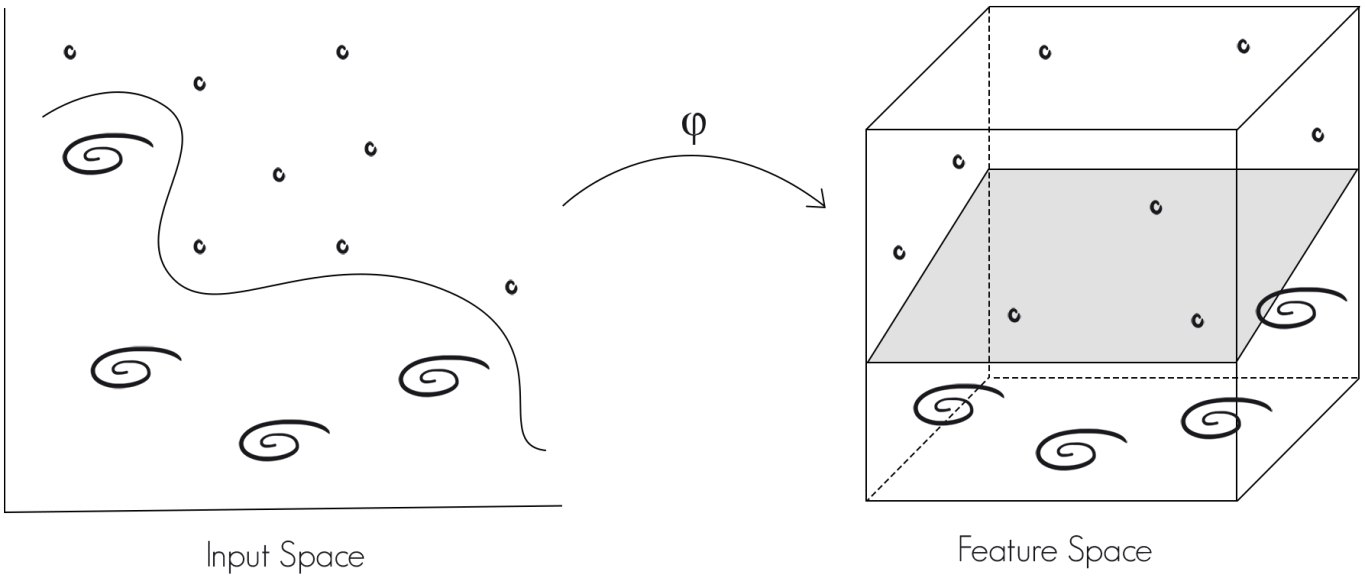
Support Vector Machine (SVM)

Support Vector Machines classification algorithm (V. Vapnik 1995).

- supervised learning algorithm (uses labeled data set for training).
- Higher efficiency than already existing algorithms.
- Powerful tool for solving classification and regression problems.

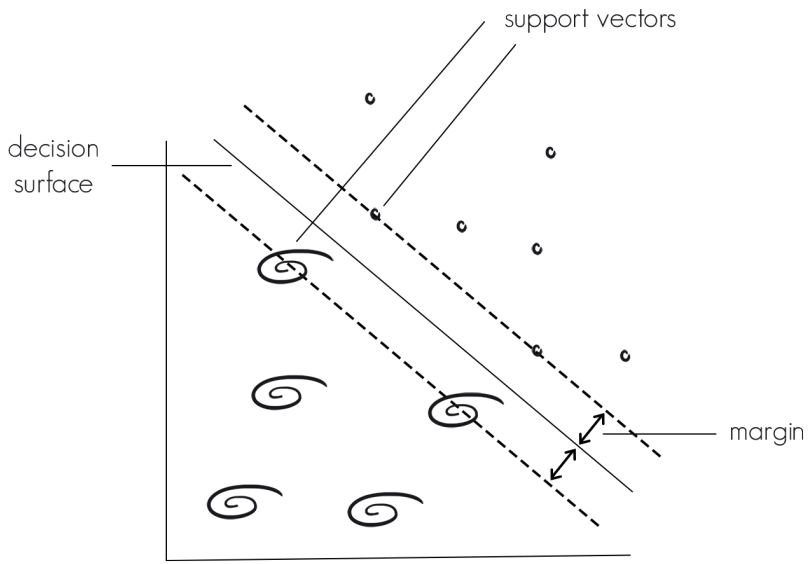
Support Vector Machine (SVM)

- Mapping data to the high dimensional space (feature space) using **kernel functions**.
- Construction of the hyperplane, which separates two classes.



Hyperplane construction

- SVM is trying to find an optimal hyperplane, which maximizes the distance (**margin**) between classes in the feature space.
- Only small amount of data is used for the hyperplane construction (**support vectors**).



Fuzzy logic SVM

Limitations of the classical SVM

- All objects are treated as equal.
- Every object belongs to a particular class.
- Only one free parameter.

Real measurements: presence of the background and measurement uncertainties.

Fuzzy SVM

- Fuzzy SVM (Lin, Wang; 2002). Every point is connected with a fuzzy membership (s_i) treated in the SVM as a weight.
- Number of the free parameters is equal to the number of the support vectors.

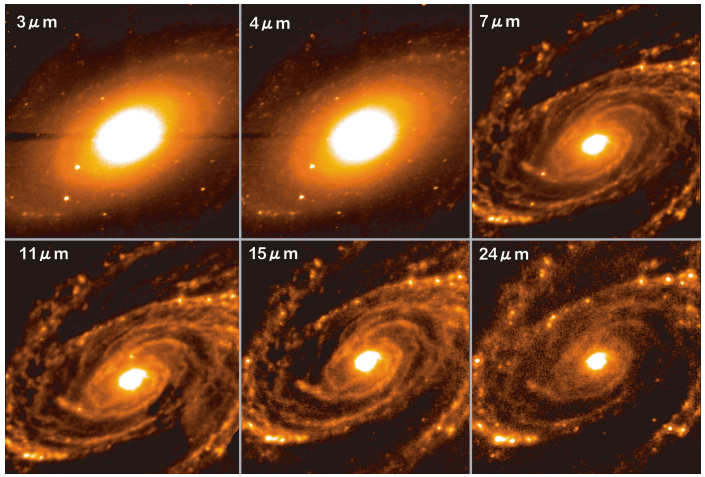
Application to the astronomy

- First attempt to apply measurement uncertainties to the SVM-based classification in astronomy.
- Basic problem: binary classification (star-galaxy separation)

AKARI Telescope

Space-based infrared telescope (2006-2011).

- Wide survey of the whole sky.
- Deep survey of the north ecliptic pole (NEP) 0.4 deg^2.
- Instruments: Far-Infrared Surveyor, Infrared Camera (IRC).
- IRC (near-, mid-IR): 9 passbands (2-24 μm).



Training

- **Training sample** : 513 galaxies, 241 stars.
- **Generalization sample** : 1808 objects.
- **Input space** : color values (4 parameters).

  $c_{\lambda_1\lambda_2} = m_{\lambda_1} - m_{\lambda_2} = -2.5 log_{10} \frac{F_{\lambda_1}}{F_{\lambda_2}}$
- **Kernels** : radial basis function (RBF), 4th degree polynomial.

  $k(\mathbf{x}_i, \mathbf{x}_j) = exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$

  $k(\mathbf{x}_i, \mathbf{x}_j) = (c_0 + \gamma \mathbf{x}_i \cdot \mathbf{x}_j)^4$

FSVM Versions

Error-based FSVM

Fuzzy membership based on the measurement uncertainties.

$$s_i = \frac{1}{e_i} \tag{1}$$

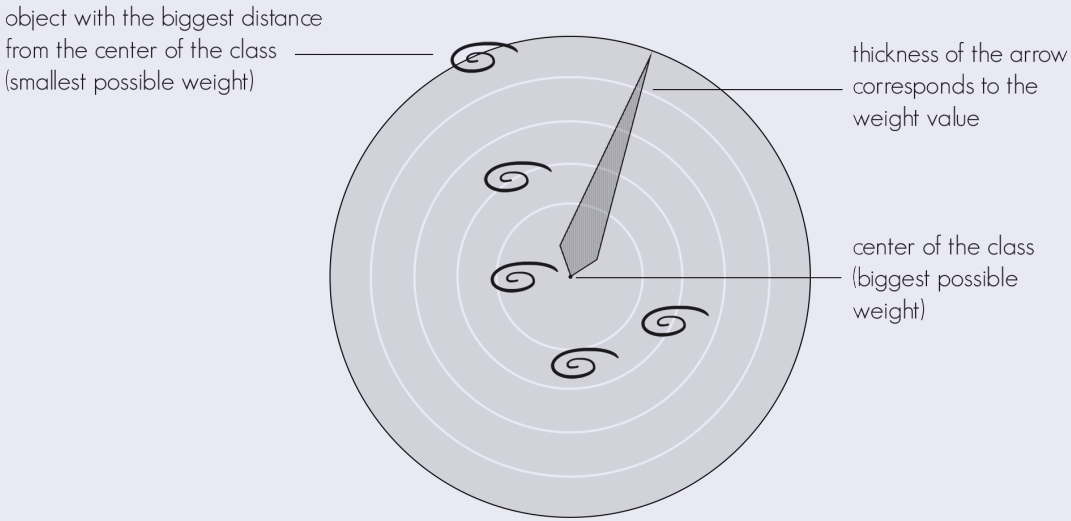
Distance-based FSVM

FSVM Versions

Error-based FSVM

Distance-based FSVM

Fuzzy membership based on the distance from the class center.

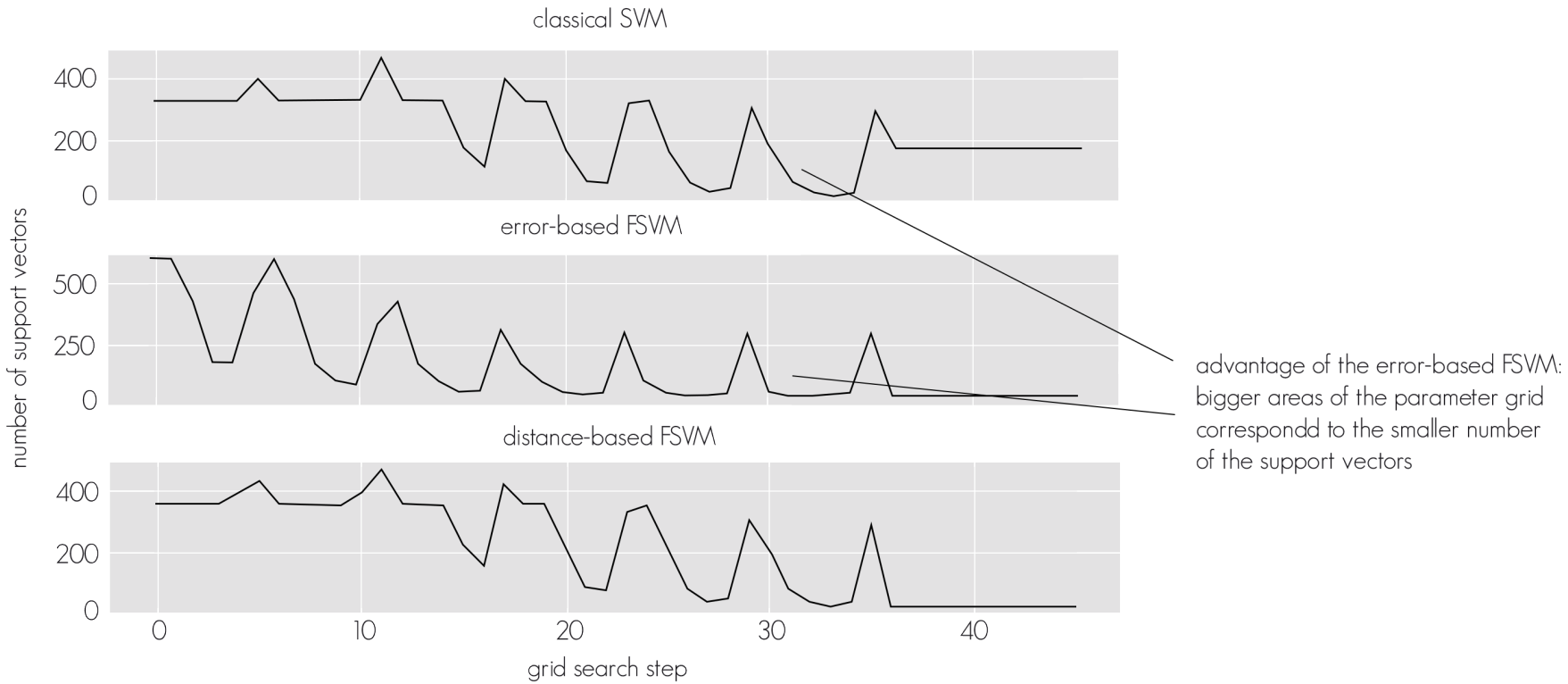


Grid Search Process

- **Grid Search** - searching for the best parameter combination.
- **Goal** : **number of the support vectors minimization**. Small number of the support vectors corresponds to the high generalization ablility and smaller probability of the **overfitting** (when algorithm is too precise in the attempt to properly reconstruct the training sample).

Grid Search Process. RBF Kernel

Error-based FSVM minimizes the number of the support vectors more effectively then the classical SVM.

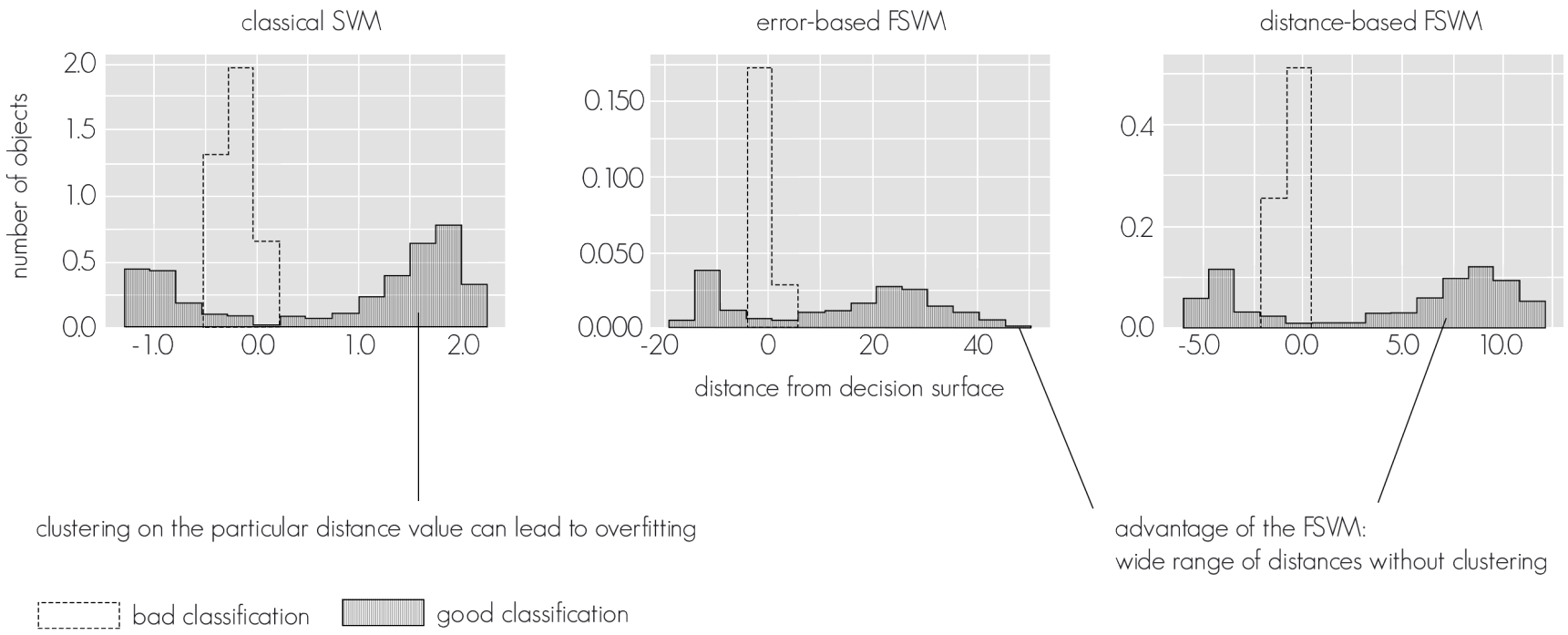


Distance Histograms

- **Distance histograms** - histograms of the distance of the points from the decision surface.
- Distance histograms should show a **wide range of the distances**. Clustering of the objects even at the big distances can lead to the overfitting.

Distance Histograms. RBF Kernel

Both error-based and distance-based FSVM overcome classical SVM.



Summary

- First attempt to apply measurement uncertainties to the SVM-based classification in astronomy.
- Both versions of the FSVM overcome classical SVM algorithm.
- However, they are focused on the different aspects of the data. Error-based FSVM: proves that the measurement uncertainties can be used as weights in order to improve classification. Distance-based FSVM is focused on the main populations and minimization of the effect of the outliers.
- Both methods should be used to study different features of the data.
- Future studies: bigger datasets (WISE, LSST), multiple class recognition (AGN, asteroids, different galaxy types).

Bibliography

- V. Vapnik, *Statistical Learning Theory*, Wiley 1998.
- V. Vapnik, C. Cortes, *Support-Vector Networks*, Machine Learning, **20**,273-297 (1995).
- C-F. Lin, S-D Wang, *Fuzzy Support Vector Machines*, IEEE Transactions on Neural Networks, vol **13**, 464-471 (2002).
- H. Murakami et al.:*The Infrared Astronomical Mission AKARI*,Publications of the Astronomical Society of Japan, 59:S369–S376, 2007.
- T. Takagi et al.:*The AKARI NEP-Deep survey: a mid-infrared source catalogue*,Astronomy & Astrophysics, 537, id.A24, 11 pp, 2012.
- A. Solarz et al.:*Star-galaxy separation in the AKARI NEP deep field*, Astronomy & Astrophysics, 541, id.A50, 2012.