Support Measure Data Description for Group Anomaly Detection

[1]Jorge Guevara Díaz,
[2]Stephane Canu,
[1]Roberto Hirata Jr.

[1]IME-USP, Brazil. [2]LITIS-INSA, France.

November 03, 2015.

Outline

Outline

Who am I?

Bachelor graduate project

Title	A Speech Recognition system for vocals recognition using a feedforward neural net.
Supervisor	Professor Ronald Leon
Description	We performed experiments with a feed-forward neural net with back-propagation algorithm to classify vocal sounds.

Informatics engineering thesis

Title	Feature Extraction using Wavelets for Speech Recognition
Supervisor	Professor Ronald Leon
Description	We developed an algorithm for feature extraction from speech audio using wavelet theory, we applied such algorithm in isolated speech recognition. Text in spanish.

- Three papers in local peruvian conferences.
- Lorito. A isolated word recognition from scratch, in JAVA. https://github.com/jorjasso

Who am I?

Masters Thesis

Title	Speech Recognition Framework for Information Retrieval
Supervisor	Professor Ronald Leon
Description	We explore speech recognition techniques for text retrieval from speech audios in spanish language. Text in spanish.

https://github.com/jorjasso

- Speech Miner. Speech recognition system using MFCC, HMM. using HTK
- SOM-TSP. SOM neural network to solve the TSP problem in JAVA.
- FNN. Neural network for digit classification in JAVA

Who am I?

Teaching Experience

2008–2009	**Speech Recognition**, *5rd year BSc*, National University of Trujillo, Peru.
2008–2009	**Computer Graphics**, *3rd year BSc*, National University of Trujillo, Peru.
2008–2009	**Image Processing**, *4rd year BSc*, National University of Trujillo, Peru.
2007–2008	**Numerical Computing**, *3rd year BSc*, National University of Trujillo, Peru.
2007–2008	**Artificial Intelligence**, *4rd year BSc*, National University of Trujillo, Peru.

Who am I?

Doctoral Thesis

Title	*Supervised Machine Learning using kernel methods, probability measures and fuzzy set theory*
Supervisor	Professor Roberto Hirata Junior, University of Sao Paulo, Brazil.
Internship Supervisor	Professor Stephane Canu, INSA-ROUEN, France.
Description	This thesis explored the idea of learning on training sets of points, where each individual point is itself a set. We treat each point-set as a realization of a fuzzy random variable or, as a realization of a random probability measure. We develop kernel algorithms to deal with such data.

Similarity between fuzzy sets using kernels

- link between fuzzy systems and kernels
- theory of positive definite kernels on fuzzy sets
- kernels induced by fuzzy distance

A data description model for set of distributions

Who am I?

Doctoral Thesis

Title	*Supervised Machine Learning using kernel methods, probability measures and fuzzy set theory*
Supervisor	Professor Roberto Hirata Junior, University of Sao Paulo, Brazil.
Internship Supervisor	Professor Stephane Canu, INSA-ROUEN, France.
Description	This thesis explored the idea of learning on training sets of points, where each individual point is itself a set. We treat each point-set as a realization of a fuzzy random variable or, as a realization of a random probability measure. We develop kernel algorithms to deal with such data.

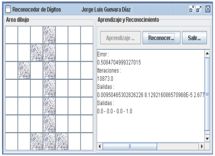
Similarity between fuzzy sets using kernels

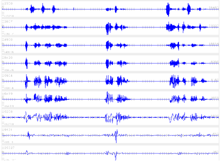
- link between fuzzy systems and kernels
- theory of positive definite kernels on fuzzy sets
- kernels induced by fuzzy distance

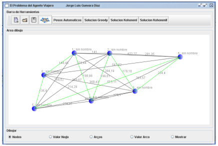
A data description model for set of distributions
`https://github.com/jorjasso`

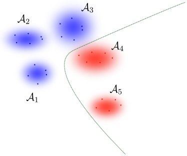
- Fuzzy kernel hypothesis testing
- TSK kernels on fuzzy sets, classification of low quality datasets.
- Group anomaly detection using SMDD.

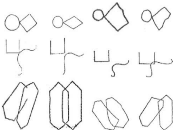
Who am I?

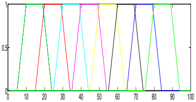

Reconocedor de Digitos
Jorge Luis Guevara Diaz
Area dibujo
Aprendizaje y Reconocimiento
Aprendizaje ...
Reconocer...
Salir...
Error :
0.5084704999327015
Iteraciones :
10873.0
Salidas :
Salidas :
0.0 - 0.0 - 0.0 - 1.0




El Problema del Agente Viajero
Jorge Luis Guevara Diaz


$\mathcal{A}_2$
$\mathcal{A}_3$
$\mathcal{A}_4$
$\mathcal{A}_1$
$\mathcal{A}_5$




1
0.5
0
0 10 20 30 40 50 60 70 80 90 100

Outline

Introduction

Definition (Anomaly)

An anomaly is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism[a]

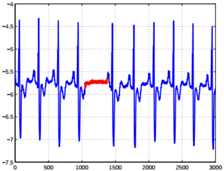
Figure : Rare starfish found. One in a million![b]

a) Hawkins D., Identification of Outliers, Chapman and Hall, 1980.

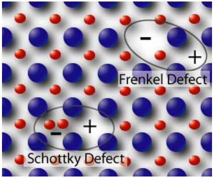
b) https://twitter.com/gotham3/status/421258659620855809

Some applications include[a]:

- Fraud detection. i.e., abnormal buying patterns;
- Medicine. i.e., Unusual symptoms, abnormal tests;
- Sports. i.e., Outstanding players;
- Measurement errors. i.e., abnormal values.
- Cyber-intrusion detection ;
- Industrial damage detection;
- Image processing,
- Textual anomaly detection



(a) Unusual contraction.



(b) Ceramic defects[b].

(c) Traffic jam recognition.

a)Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. ACM Comput. Surv.
b) `https://www.nde-ed.org/`

c) `https://www.youtube.com/watch?v=DAXUzWnsiQk`

Techniques:

- Generative models. i.e., HMM, GMM;
- Unsupervised methods. i.e., clustering, distance-based, density based;
- Discriminative models. i.e., SVM, neural networks;
- Information Theoretic Methods, Geometric methods

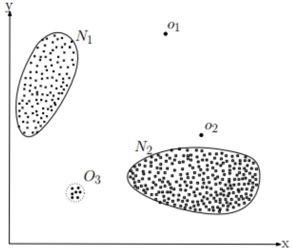


Figure : 2D-anomalies[a].

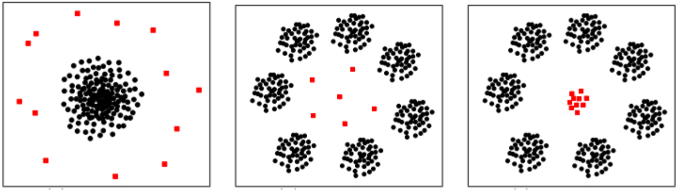
a)Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. ACM Comput. Surv.

Problem

Given a data set of the form

$$\mathcal{T} = \{s_i\}_{i=1}^{N}, \tag{1}$$

where $s_i = \{\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \ldots, \mathbf{x}_{L_i}^{(i)}\} \sim \mathbb{P}_i$, and $\mathbb{P}_i$ defined on $(\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D))$.
Try to detect anomalies or *group anomalies* from $\mathcal{T}$



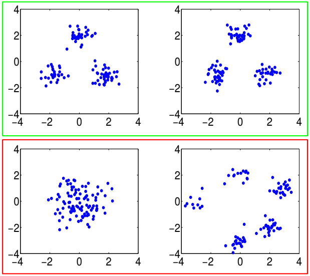
Problem

Given a data set of the form

$$\mathcal{T} = \{s_i\}_{i=1}^{N}, \tag{2}$$

where $s_i = \{\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \ldots, \mathbf{x}_{L_i}^{(i)}\} \sim \mathbb{P}_i$, and $\mathbb{P}_i$ defined on $(\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D))$.
Try to detect anomalies or *group anomalies* from $\mathcal{T}$

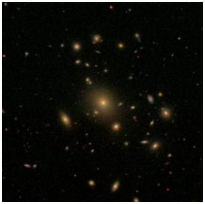
A more complex scenario



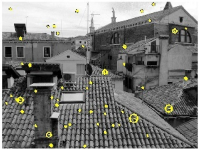
Introduction

Examples of datasets of this form

$$\mathcal{T} = \{s_i\}_{i=1}^{N}, \tag{3}$$



(a) Cluster of galaxies



(b) SIFT.

(c) USPS.

a) http://www.sdss3.org/. b) http://www.vlfeat.org. c) *Chang-Dong et al. "Multi-Exemplar Affinity Propagation",*

Group anomaly

Group of points with unexpected behavior wrt a dataset of group of points.

- **Point-based group anomalies** aggregation of anomalous points

- **Distributed-based group anomalies** anomalous aggregation of non-anomalous points.

Previous work

- Feature engineering approach
  - Feature extraction from each group[a,b]
  - Clustering point anomalies

  However, they ignore distributed-based group anomalies.
- Generative approach
  - Flexible genre models[c]
  - Hierarchical probabilistic models[d]

  However, procedures rely on parametric assumptions
- Discriminative approach
  - Support measure machines[e]
  - Our work

  nonparametric, performance depends on the kernel choice.

[a] Chan et al. "Modeling multiple time series for anomaly detection," in Data Mining, IEEE.
[b] Keogh et al. "HOT SAX: efficiently finding the most unusual time series subsequence," in Data Mining , IEEE.
[c] L Xion et al. "Group Anomaly Detection using Flexible Genre Models", NIPS.

[d] L Xion et al. "Hierarchical Probabilistic Models for Group Anomaly Detection", AISTATS. [e] Muandet et a.

"One-Class Support Measure Machines for Group Anomaly Detection", UAI.

Outline

RKHS, where the magic happens

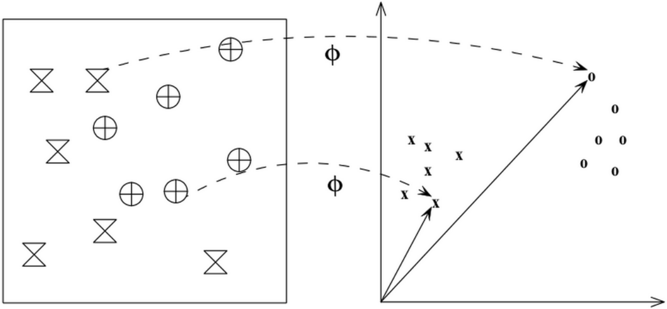
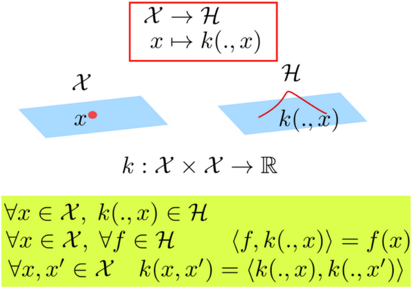


Figure : Kernel mapping[a]

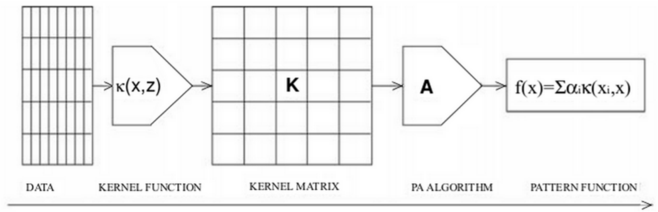
Figure from Shawe-Taylor et al. "Kernel Methods for Pattern Analysis". Cambridge University Press.

Main ingredient

- A real-valued symmetric positive definite kernel k.
  $\sum_{i=1,j=1}^{N} c_i c_j k(x_i, x_j) \geq 0$

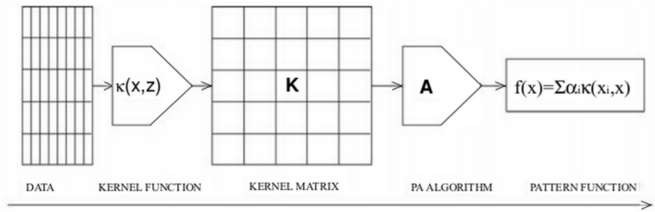


Kernel methods



- SVM
- kernel PCA
- Gaussian process
- SVDD
- MKL
- SMDD
- Kernel regression
- Kernel two-sample test
- kernel spectral clustering

Kernel methods



- SVM
- kernel PCA
- Gaussian process
- SVDD
- MKL
- SMDD
- Kernel regression
- Kernel two-sample test
- kernel spectral clustering

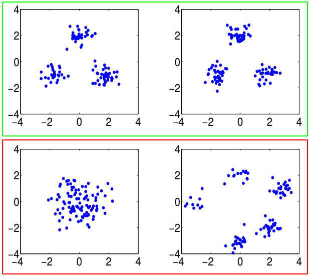
Representer Theorem

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}} Cost((x_1, y_1, f(x_1)), \ldots, (x_N, y_N, f(x_N))) + \Omega(\|f\|) \tag{4}$$

$$f^*(.) = \sum_{i=1}^{N} \alpha_i k(., x_i) \tag{5}$$

Remembering the problem

Detecting group anomalies from the set $\mathcal{T} = \{s_i\}_{i=1}^{N}$.



Hilbert space embedding for distributions

Framework for embedding probability measures into a RKHS $\mathcal{H}$.

Definition

The embedding of probability measures $\mathbb{P} \in \mathcal{P}$ into $\mathcal{H}$ is given by the mapping

$$\begin{aligned} \mu : \mathcal{P} &\rightarrow \mathcal{H} \\ \mathbb{P} &\mapsto \mu_{\mathbb{P}} = \mathbb{E}_{\mathbb{P}}[k(X, .)] = \int_{\mathbf{x} \in \mathbb{R}^D} k(\mathbf{x}, .) d\mathbb{P}(\mathbf{x}). \end{aligned}$$

Hilbert space embedding for distributions

Properties

- if $\mu_{\mathbb{P}}(X) = \mathbb{E}_{\mathbb{P}}[k(X, X)] < \infty$, with k being measurable then $\mu_{\mathbb{P}} \in \mathcal{H}$

Hilbert space embedding for distributions

Properties

- if $\mu_{\mathbb{P}}(X) = \mathbb{E}_{\mathbb{P}}[k(X, X)] < \infty$, with k being measurable then $\mu_{\mathbb{P}} \in \mathcal{H}$
- Reproducing property $\langle f, \mu_{\mathbb{P}} \rangle = \langle f, \mathbb{E}_{\mathbb{P}}[k(X, .)] \rangle = \mathbb{E}_{\mathbb{P}}[f(X)]$ holds for all $f \in \mathcal{H}$

Hilbert space embedding for distributions

Properties

- if $\mu_{\mathbb{P}}(X) = \mathbb{E}_{\mathbb{P}}[k(X,X)] < \infty$, with k being measurable then $\mu_{\mathbb{P}} \in \mathcal{H}$
- Reproducing property $\langle f, \mu_{\mathbb{P}} \rangle = \langle f, \mathbb{E}_{\mathbb{P}}[k(X,.)] \rangle = \mathbb{E}_{\mathbb{P}}[f(X)]$ holds for all $f \in \mathcal{H}$
- $\mu_{\mathbb{P}}$ is the *representative function* of $\mathbb{P}$

Hilbert space embedding for distributions

Properties

- if $\mu_{\mathbb{P}}(X) = \mathbb{E}_{\mathbb{P}}[k(X, X)] < \infty$, with k being measurable then $\mu_{\mathbb{P}} \in \mathcal{H}$
- Reproducing property $\langle f, \mu_{\mathbb{P}} \rangle = \langle f, \mathbb{E}_{\mathbb{P}}[k(X, .)] \rangle = \mathbb{E}_{\mathbb{P}}[f(X)]$ holds for all $f \in \mathcal{H}$
- $\mu_{\mathbb{P}}$ is the *representative function* of $\mathbb{P}$
- if k is *characteristic* then $\mu : \mathcal{P} \to \mathcal{H}$ is injective

Hilbert space embedding for distributions

Properties

- if $\mu_{\mathbb{P}}(X) = \mathbb{E}_{\mathbb{P}}[k(X,X)] < \infty$, with k being measurable then $\mu_{\mathbb{P}} \in \mathcal{H}$
- Reproducing property $\langle f, \mu_{\mathbb{P}} \rangle = \langle f, \mathbb{E}_{\mathbb{P}}[k(X,.)] \rangle = \mathbb{E}_{\mathbb{P}}[f(X)]$ holds for all $f \in \mathcal{H}$
- $\mu_{\mathbb{P}}$ is the *representative function* of $\mathbb{P}$
- if k is *characteristic* then $\mu : \mathcal{P} \rightarrow \mathcal{H}$ is injective
- The term $\|\mu_{\mathbb{P}} - \mu_{emp}\|$, is bounded, where μ_{emp} is a empirical estimator of μ_{emp}

Kernel on probability measures

The mapping

$$\begin{array}{rcl} \mathcal{P} \times \mathcal{P} & \to & \mathbb{R} \\ (\mathbb{P}, \mathbb{Q}) & \mapsto & \langle \mathbb{P}, \mathbb{Q} \rangle_{\mathcal{P}} = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \end{array}$$

defines an inner product on $\mathcal{P}$. the real-valued kernel on $\mathcal{P} \times \mathcal{P}$, defined by

$$\begin{aligned} \tilde{k}(\mathbb{P}, \mathbb{Q}) =& \langle \mathbb{P}, \mathbb{Q} \rangle_{\mathcal{P}} = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\ =& \int_{\mathbf{x} \in \mathbb{R}^D} \int_{\mathbf{x}' \in \mathbb{R}^D} k(\mathbf{x}, \mathbf{x}') d\mathbb{P}(\mathbf{x}) d\mathbb{Q}(\mathbf{x}') \end{aligned} \tag{6}$$

is positive definite.

Outline

- a MV-set is a set satisfying some optimization criteria.

Definition (MV-set for probability measures)

Let $(\mathcal{P}, \mathcal{A}, \mathcal{E})$ be a probability space, where $\mathcal{P}$ is the space of all probability measures $\mathbb{P}$ on $(\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D))$, $\mathcal{A}$ is some suitable σ-algebra of $\mathcal{P}$ and $\mathcal{E}$ is a probability measure on $(\mathcal{P}, \mathcal{A})$. The MV-set is the set

$$G^*_\alpha = \underset{G \in \mathcal{A}}{\operatorname{argmin}} \{\rho(G) | \mathcal{E}(G) \geq \alpha\}, \tag{7}$$

where ρ is a reference measure on $\mathcal{A}$ and $\alpha \in [0, 1]$. The MV-set G^*_α, describes a fraction α of the mass concentration of $\mathcal{E}$.

Minimum volume sets (MV-set)

- Assuming that $\{\mathbb{P}_i\}_{i=1}^{N}$ is an i.i.d. sample distributed according to $\mathcal{E}$
- Assuming that each $\mathbb{P}_i$ is unknown.

Minimum volume sets (MV-set)

- Assuming that $\{\mathbb{P}_i\}_{i=1}^{N}$ is an i.i.d. sample distributed according to $\mathcal{E}$
- Assuming that each $\mathbb{P}_i$ is unknown.

Examples of three different classes of volume-sets

$$\hat{G}_1(R, c) = \{\mathbb{P}_i \in \mathcal{P} \mid \|\mu_{\mathbb{P}_i} - c\|_{\mathcal{H}}^2 \leq R^2\}, \tag{8}$$

Minimum volume sets (MV-set)

- Assuming that $\{\mathbb{P}_i\}_{i=1}^{N}$ is an i.i.d. sample distributed according to $\mathcal{E}$
- Assuming that each $\mathbb{P}_i$ is unknown.

Examples of three different classes of volume-sets

$$\hat{G}_1(R,c) = \{\mathbb{P}_i \in \mathcal{P} \mid \|\mu_{\mathbb{P}_i} - c\|^2_{\mathcal{H}} \leq R^2\}, \tag{8}$$

$$\hat{G}_2(R,c) = \{\mathbb{P}_i \in \mathcal{P} \mid \|\mu_{\mathbb{P}_i} - c\|^2_{\mathcal{H}} \leq R^2,\ \|\mu_{\mathbb{P}}\|^2_{\mathcal{H}} = 1\}. \tag{9}$$

Minimum volume sets (MV-set)

- Assuming that $\{\mathbb{P}_i\}_{i=1}^{N}$ is an i.i.d. sample distributed according to $\mathcal{E}$
- Assuming that each $\mathbb{P}_i$ is unknown.

Examples of three different classes of volume-sets

$$\hat{G}_1(R,c) = \{\mathbb{P}_i \in \mathcal{P} \mid \|\mu_{\mathbb{P}_i} - c\|_{\mathcal{H}}^2 \leq R^2\}, \tag{8}$$

$$\hat{G}_2(R,c) = \{\mathbb{P}_i \in \mathcal{P} \mid \|\mu_{\mathbb{P}_i} - c\|_{\mathcal{H}}^2 \leq R^2,\ \|\mu_{\mathbb{P}}\|_{\mathcal{H}}^2 = 1\}. \tag{9}$$

$$\hat{G}_3(\mathcal{K}) = \{\mathbb{P}_i \in \mathcal{P} \mid \mathbb{P}_i(\|k(X_i,.) - c\|_{\mathcal{H}}^2 \leq R^2) \geq 1 - \kappa_i\}. \tag{10}$$

First SMDD model

$$\hat{G}_1(R, c) = \{\mathbb{P}_i \in \mathcal{P} \mid \|\mu_{\mathbb{P}_i} - c\|^2_{\mathcal{H}} \leq R^2\}$$

Given the mean functions $\{\mu_{\mathbb{P}_i}\}_{i=1}^N$ of $\{\mathbb{P}_i\}_{i=1}^N$, the SMDD is:

Problem

$$\begin{array}{ll} \underset{c\in\mathcal{H}, R\in\mathbb{R}^+, \boldsymbol{\xi}\in\mathbb{R}^N}{min} & R^2 + \lambda \sum_{i=1}^{N} \xi_i \\ subject\ to & \|\mu_{\mathbb{P}_i} - c\|^2_{\mathcal{H}} \leq R^2 + \xi_i, i = 1, \ldots, N \\ & \xi_i \geq 0,\ i = 1, \ldots, N. \end{array}$$

Proposition (Dual form)

The dual form of the previously problem is given by:

Problem

$$\begin{aligned} \max_{\boldsymbol{\alpha}\in\mathbb{R}^N} \quad & \sum_{i=1}^{N} \alpha_i \tilde{k}(\mathbb{P}_i, \mathbb{P}_i) - \sum_{i,j=1}^{N} \alpha_i \alpha_j \tilde{k}(\mathbb{P}_i, \mathbb{P}_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \lambda, \; i = 1, \ldots, N \\ & \sum_{i=1}^{N} \alpha_i = 1 \end{aligned}$$

where $\tilde{k}(\mathbb{P}_i, \mathbb{P}_j) = \langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j} \rangle_{\mathcal{H}}$ *and* $\boldsymbol{\alpha}$ *is a Lagrange multiplier vector with non negative components* α_i.

Proposition (Representer theorem)

$$c(.) = \sum_i \alpha_i \mu_{\mathbb{P}_i}, \quad i \in \{i \in \mathcal{I} \mid 0 < \alpha_i \leq \lambda\},$$

where $\mathcal{I} = \{1, 2, \ldots, N\}$. *Furthermore,*

- *all* $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid \alpha_i = 0\}$ *are inside the MV-set* $\hat{G}^*_\alpha$.
- *All* $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid \alpha_i = \lambda\}$ *are the training errors.*
- *All* $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid 0 < \alpha_i < \lambda\}$ *are the support measures.*

Proposition (Representer theorem)

$$c(.) = \sum_i \alpha_i \mu_{\mathbb{P}_i}, \quad i \in \{i \in \mathcal{I} \mid 0 < \alpha_i \leq \lambda\},$$

where $\mathcal{I} = \{1, 2, \ldots, N\}$. *Furthermore,*

- *all* $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid \alpha_i = 0\}$ *are inside the MV-set* $\hat{G}^*_\alpha$.
- *All* $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid \alpha_i = \lambda\}$ *are the training errors.*
- *All* $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid 0 < \alpha_i < \lambda\}$ *are the support measures.*

Theorem

Let η *be the Lagrange multiplier of the constraint* $\sum_{i=1}^N \alpha_i = 1$, *then* $R^2 = -\eta + \|c\|^2_{\mathcal{H}}$.

Proposition (Representer theorem)

$$c(.) = \sum_i \alpha_i \mu_{\mathbb{P}_i}, \quad i \in \{i \in \mathcal{I} \mid 0 < \alpha_i \leq \lambda\},$$

where $\mathcal{I} = \{1, 2, \ldots, N\}$. *Furthermore,*

- *all* $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid \alpha_i = 0\}$ *are inside the MV-set* $\hat{G}^*_\alpha$.
- *All* $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid \alpha_i = \lambda\}$ *are the training errors.*
- *All* $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid 0 < \alpha_i < \lambda\}$ *are the support measures.*

Theorem

Let η *be the Lagrange multiplier of the constraint* $\sum_{i=1}^N \alpha_i = 1$, *then* $R^2 = -\eta + \|c\|^2_{\mathcal{H}}$.

If $\mathbb{P}_t$ is described by this SMDD model, then this must be true:

$$\|\mu_{\mathbb{P}_t} - c\|^2_{\mathcal{H}} = \tilde{k}(\mathbb{P}_t, \mathbb{P}_t) - 2\sum_i \alpha_i \tilde{k}(\mathbb{P}_i, \mathbb{P}_t) + \sum_{i,j} \alpha_i \alpha_j \tilde{k}(\mathbb{P}_i, \mathbb{P}_j) \leq R, \quad (11)$$

Second SMDD model

- Mean maps with stationary kernels do not have constant norm $\|\mu_{\mathbb{P}}\|_{\mathcal{H}} = \|\mathbb{E}_{\mathbb{P}}[k_I(X.,)]\|_{\mathcal{H}} \leq \mathbb{E}_{\mathbb{P}}[\|k_I(X.,)\|_{\mathcal{H}}] = \sqrt{|\epsilon|}$
- normalize mean maps to lie on a surface of some hypersphere

$$\tilde{\tilde{k}}(\mathbb{P}_i, \mathbb{P}_j) = \frac{\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}}{\sqrt{\langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}}}, \tag{12}$$

- the injectivity of $\mu : \mathcal{P} \to \mathcal{H}$ is preserved.

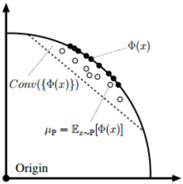


Figure : Figure from [a]

Muandet et al "One-class support measure machines for group anomaly detection."

This could be modeled by the class of volume sets

$$\hat{G}_2(R, c) = \{\mathbb{P}_i \in \mathcal{P} \mid \|\mu_{\mathbb{P}_i} - c\|_{\mathcal{H}}^2 \leq R^2,\ \|\mu_{\mathbb{P}}\|_{\mathcal{H}}^2 = 1\}. \tag{13}$$

and formulated by the following optimization problem:

Problem (M2)

$$\begin{aligned} \max_{\boldsymbol{\alpha} \in \mathbb{R}^N} \quad & -\sum_{i,j=1}^{N} \alpha_i \alpha_j \tilde{\tilde{k}}(\mathbb{P}_i, \mathbb{P}_j) \\ \textit{subject to} \quad & 0 \leq \alpha_i \leq \lambda,\ i = 1, \ldots, N \\ & \sum_{i=1}^{N} \alpha_i = 1. \end{aligned}$$

Third SMDD model

Estimate the MV-set over the class of the volume-sets given by

$$\hat{G}_3(\mathcal{K}) = \{\mathbb{P}_i \in \mathcal{P} \mid \mathbb{P}_i(\|k(X_i, .) - c\|^2_{\mathcal{H}} \leq R^2) \geq 1 - \kappa_i\}. \qquad (14)$$

Given the mean functions $\{\mu_{\mathbb{P}_i}\}_{i=1}^N$ of $\{\mathbb{P}_i\}_{i=1}^N$, and $\{\kappa_i\}_{i=1}^N$, $\kappa_i \in [0, 1]$, the SMDD model is the following chance constrained problem:

Problem

$$\begin{aligned}
\underset{c\in\mathcal{H}, R\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^N}{\mathit{min}} \quad & R^2 + \lambda \sum_{i=1}^{N} \xi_i \\
\mathit{subject\ to} \quad & \mathbb{P}_i(\|k(X_i, .) - c(.)\|^2_{\mathcal{H}} \leq R^2 + \xi_i) \geq 1 - \kappa_i, \\
& \xi_i \geq 0, \\
& \mathit{for\ all}\ \ i = 1, \ldots, N.
\end{aligned}$$

- *Markov's inequality*.

$$\mathbb{P}_i(\|k(X_i,.) - c(.)\|^2_{\mathcal{H}} \geq R^2 + \xi_i) \leq \frac{\mathbb{E}_{\mathbb{P}}[\|k(X_i,.) - c(.)\|^2_{\mathcal{H}}]}{R^2 + \xi_i}, \quad (15)$$

holds, for all $i = 1, 2, \ldots, N$.

- Trace of the covariance operator: $\Sigma^{\mathcal{H}} : \mathcal{H} \to \mathcal{H}$,

$$tr(\Sigma^{\mathcal{H}}) = \mathbb{E}_{\mathbb{P}}[k(X,X)] - \tilde{k}(\mathbb{P},\mathbb{P}), \quad (16)$$

- This allows to use the following result

Lemma

$$\mathbb{E}_{\mathbb{P}}[\|k(X,.) - c(.)\|^2_{\mathcal{H}}] = tr(\Sigma^{\mathcal{H}}) + \|\mu_{\mathbb{P}} - c(.)\|^2_{\mathcal{H}}.$$

Deterministic form

Given the mean functions $\{\mu_{\mathbb{P}_i}\}_{i=1}^N$ of $\{\mathbb{P}_i\}_{i=1}^N$ and $\{\kappa_i\}_{i=1}^N$, $\kappa_i \in (0,1]$, the SMDD model is:

Problem

$$\begin{array}{ll} \underset{c\in\mathcal{H},R\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^N}{min} & R^2 + \lambda \sum_{i=1}^{N} \xi_i \\ subject\ to & \|\mu_{\mathbb{P}_i} - c(.)\|_{\mathcal{H}}^2 \leq (R^2+\xi_i)\kappa_i - tr(\Sigma_i^{\mathcal{H}}), \\ & \xi_i \geq 0, \end{array}$$

for all $i = 1, \ldots, N$

Proposition (Dual form)

The dual form is given by the following fractional programming problem

Problem (M1)

$$
\begin{aligned}
\max_{\boldsymbol{\alpha}\in\mathbb{R}^N} \quad & \sum_{i=1}^{N} \alpha_i \langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_i} \rangle_{\mathcal{H}} - \frac{\sum_{i,j=1}^{N} \alpha_i \alpha_j \langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j} \rangle_{\mathcal{H}}}{\sum_{i=1}^{N} \alpha_i} \\
& + \sum_{i=1}^{N} \alpha_i tr(\Sigma_i^{\mathcal{H}}) \\
\text{subject to} \quad & 0 \le \alpha_i \kappa_i \le \lambda, \; i = 1, \dots, N \\
& \sum_{i=1}^{N} \alpha_i \kappa_i = 1,
\end{aligned}
$$

where $\langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j} \rangle_{\mathcal{H}}$ *is computed by* $\tilde{k}(\mathbb{P}_i, \mathbb{P}_j)$

Proposition (Representer theorem)

$$c(.) = \frac{\sum_i \alpha_i \mu_{\mathbb{P}_i}}{\sum_i \alpha_i}, \ i \in \{i \in \mathcal{I} \mid 0 < \alpha_i \kappa_i \leq \lambda\}, \tag{17}$$

where $\mathcal{I} = \{1, 2, \dots, N\}$. *Furthermore,*

- *all* $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid \alpha_i = 0\}$ *are inside the MV-set* $\hat{G}^*_\alpha$.
- *All* $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid \alpha_i \kappa_i = \lambda\}$ *are the training errors.*
- *All* $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid 0 < \alpha_i \kappa_i < \lambda\}$ *are the support measures.*

Theorem

Let η *be the Lagrange multiplier of the constraint* $\sum_{i=1}^N \alpha_i \kappa_i = 1$ *then* $R^2 = -\eta$.

Group anomalies could be detected if the term $\|\mu_{\mathbb{P}_t} - c\|^2_{\mathcal{H}} + tr(\Sigma_t^{\mathcal{H}}) \geq R$ is true, where $\|\mu_{\mathbb{P}_t} - c\|$ is given by

$$\tilde{k}(\mathbb{P}_t, \mathbb{P}_t) - 2\sum_i \alpha_i \tilde{k}(\mathbb{P}_i, \mathbb{P}_t) + \sum_{i,j} \alpha_i \alpha_j \tilde{k}(\mathbb{P}_i, \mathbb{P}_j) + tr(\Sigma_t^{\mathcal{H}}) \qquad (18)$$

Outline

Group anomaly detection in astronomical data

The Star KIC 8462852 phenomena

SCIENCE

The Most Mysterious Star in Our Galaxy

Astronomers have spotted a strange mess of objects whirling around a distant star. Scientists who search for extraterrestrial civilizations are scrambling to get a closer look.

173k 10k

ROSS ANDERSEN | OCT 13, 2015

In the Northern hemisphere's sky, hovering above the Milky Way, there are two constellations—Cygnus the swan, her wings outstretched in full flight, and Lyra, the harp that accompanied poetry in ancient Greece, from which we take our word "lyric."

Between these constellations sits an unusual star, invisible to the naked eye, but visible to the Kepler Space Telescope, which stared at it for more than four years, beginning in 2009.

The Star KIC 8462852 phenomena

SCIENCE

The Most Mysterious Star in Our Galaxy

Astronomers have spotted a strange mess of objects whirling around a distant star. Scientists who search for extraterrestrial civilizations are scrambling to get a closer look.

173k 10k

ROSS ANDERSEN | OCT 13, 2015

In the Northern hemisphere's sky, hovering above the Milky Way, there are two constellations—Cygnus the swan, her wings outstretched in full flight, and Lyra, the harp that accompanied poetry in ancient Greece, from which we take our word "lyric."

Between these constellations sits an unusual star, invisible to the naked eye, but visible to the Kepler Space Telescope, which stared at it for more than four years, beginning in 2009.

Planet Hunters X.
KIC 8462852 – Where's the flux? *†

T. S. Boyajian[1], D. M. LaCourse[2], S. A. Rappaport[3], D. Fabrycky[4], D. A. Fischer[1], D. Gandolfi[5,6], G. M. Kennedy[7], M. C. Liu[8], A. Moor[9], K. Olah[9], K. Vida[9], M. C. Wyatt[7], W. M. J. Best[8], F. Ciesla[10], B. Csák[11], T. J. Dupuy[12], G. Handler[13], K. Heng[14], H. Korhonen[15,16], J. Kovács[11], T. Kozakis[17], L. Kriskovics[9], J. R. Schmitt[1], Gy. Szabo[9,11,18], R. Szabo[9], J. Wang[1,19], S. Goodman[2], A. Hoekstra[2], K. J. Jek[2]

[1] *Department of Astronomy, Yale University, New Haven, CT 06511, USA*
[2] *Amateur Astronomer*
[3] *Department of Physics, and Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[4] *Department of Astronomy and Astrophysics, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, USA*
[5] *Landessternwarte Königstuhl, Zentrum für Astronomie der Universität Heidelberg, Königstuhl 12, D-69117 Heidelberg, Germany*
[6] *Dipartimento di Fisica, Universitá di Torino, via P. Giuria 1, I-10125, Torino, Italy*
[7] *Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*
[8] *Institute for Astronomy, University of Hawaii, 2680 Woodlawn Drive, Honolulu HI 96822, USA*
[9] *Konkoly Observatory, Research Centre of Astronomy and Earth Sciences, Hungarian Academy of Sciences, H-1121 Budapest, Konkoly Th. M. út 15 –17, Hungary*
[10] *Department of the Geophysical Sciences, The University of Chicago, 5734 South Ellis Avenue, Chicago, IL 60637*
[11] *ELTE Gothard Astrophysical Observatory, H-9704 Szombathely, Szent Imre herceg ut 112, Hungary*
[12] *The University of Texas at Austin, Department of Astronomy, 2515 Speedway C1400, Austin, TX 78712, USA*
[13] *Copernicus Astronomical Center, Bartycka 18, 00-716 Warsaw, Poland*
[14] *University of Bern, Center for Space and Habitability, Sidlerstrasse 5, CH-3012, Bern, Switzerland*
[15] *Finnish Centre for Astronomy with ESO (FINCA), University of Turku, Väisäläntie 20, FI-21500 Piikkiö, Finland*
[16] *Centre for Star and Planet Formation, Niels Bohr Institute, University of Copenhagen, Øster Voldgade 5-7, DK-1350, København K, Denmark*
[17] *Carl Sagan Institute, Cornell University, Ithaca, NY 14853, USA*
[18] *Gothard-Lendület Research Team, H-9704 Szombathely, Szent Imre herceg út 112, Hungary*
[19] *California Institute of Technology, Pasadena, CA 91109, USA*

17 October 2015

ABSTRACT
Over the duration of the *Kepler* mission, KIC 8462852 was observed to undergo irregularly shaped, aperiodic dips in flux down to below the 20% level. The dipping activity can last for between 5 and 80 days. We characterize the object with high-resolution spectroscopy, spectral energy distribution fitting, and Fourier analyses of the *Kepler* light curve. We determine that KIC 8462852 is a main-sequence F3 V/IV star, with a rotation period ~ 0.88 d, that exhibits no significant IR excess. In this paper, we describe various scenarios to explain the mysterious events in the *Kepler* light curve, most of which have problems explaining the data in hand. By considering the observational constraints on dust clumps orbiting a normal main-sequence star, we conclude that the scenario most consistent with the data in hand is the passage of a family of exocomet fragments, all of which are associated with a single previous breakup event. We discuss the necessity of future observations to help interpret the system.

Group anomaly detection in astronomical data

The Star KIC 8462852 phenomena

SCIENCE

The Most Mysterious Star in Our Galaxy

Astronomers have spotted a strange mess of objects whirling around a distant star. Scientists who search for extraterrestrial civilizations are scrambling to get a closer look.

173k 10k

ROSS ANDERSEN | OCT 13, 2015

In the Northern hemisphere's sky, hovering above the Milky Way, there are two constellations—Cygnus the swan, her wings outstretched in full flight, and Lyra, the harp that accompanied poetry in ancient Greece, from which we take our word "lyric."

Between these constellations sits an unusual star, invisible to the naked eye, but visible to the Kepler Space Telescope, which stared at it for more than four years, beginning in 2009.

Planet Hunters X.
KIC 8462852 – Where's the flux? *†

T. S. Boyajian[1], D. M. LaCourse[2], S. A. Rappaport[3], D. Fabrycky[4], D. A. Fischer[1], D. Gandolfi[5,6], G. M. Kennedy[7], M. C. Liu[8], A. Moor[9], K. Olah[9], K. Vida[9], M. C. Wyatt[7], W. M. J. Best[8], F. Ciesla[10], B. Csák[11], T. J. Dupuy[12], G. Handler[13], K. Heng[14], H. Korhonen[15,16], J. Kovács[11], T. Kozakis[17], L. Kriskovics[9], J. R. Schmitt[1], Gy. Szabo[9,11,18], R. Szabo[9], J. Wang[1,19], S. Goodman[2], A. Hoekstra[2], K. J. Jek[2]

[1] *Department of Astronomy, Yale University, New Haven, CT 06511, USA*
[2] *Amateur Astronomer*
[3] *Department of Physics, and Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[4] *Department of Astronomy and Astrophysics, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, USA*
[5] *Landessternwarte Königstuhl, Zentrum für Astronomie der Universität Heidelberg, Königstuhl 12, D-69117 Heidelberg, Germany*
[6] *Dipartimento di Fisica, Universitá di Torino, via P. Giuria 1, I-10125, Torino, Italy*
[7] *Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*
[8] *Institute for Astronomy, University of Hawaii, 2680 Woodlawn Drive, Honolulu HI 96822, USA*
[9] *Konkoly Observatory, Research Centre of Astronomy and Earth Sciences, Hungarian Academy of Sciences, H-1121 Budapest, Konkoly Th. M. út 15 –17, Hungary*
[10] *Department of the Geophysical Sciences, The University of Chicago, 5734 South Ellis Avenue, Chicago, IL 60637*
[11] *ELTE Gothard Astrophysical Observatory, H-9704 Szombathely, Szent Imre herceg ut 112, Hungary*
[12] *The University of Texas at Austin, Department of Astronomy, 2515 Speedway C1400, Austin, TX 78712, USA*
[13] *Copernicus Astronomical Center, Bartycka 18, 00-716 Warsaw, Poland*
[14] *University of Bern, Center for Space and Habitability, Sidlerstrasse 5, CH-3012, Bern, Switzerland*
[15] *Finnish Centre for Astronomy with ESO (FINCA), University of Turku, Väisäläntie 20, FI-21500 Piikkiö, Finland*
[16] *Centre for Star and Planet Formation, Niels Bohr Institute, University of Copenhagen, Øster Voldgade 5-7, DK-1350, København K, Denmark*
[17] *Carl Sagan Institute, Cornell University, Ithaca, NY 14853, USA*
[18] *Gothard-Lendület Research Team, H-9704 Szombathely, Szent Imre herceg út 112, Hungary*
[19] *California Institute of Technology, Pasadena, CA 91109, USA*

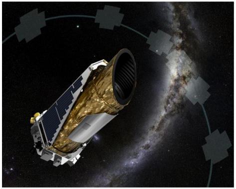
17 October 2015

ABSTRACT
Over the duration of the *Kepler* mission, KIC 8462852 was observed to undergo irregularly shaped, aperiodic dips in flux down to below the 20% level. The dipping activity can last for between 5 and 80 days. We characterize the object with high-resolution spectroscopy, spectral energy distribution fitting, and Fourier analyses of the *Kepler* light curve. We determine that KIC 8462852 is a main-sequence F3 V/IV star, with a rotation period ~ 0.88 d, that exhibits no significant IR excess. In this paper, we describe various scenarios to explain the mysterious events in the *Kepler* light curve, most of which have problems explaining the data in hand. By considering the observational constraints on dust clumps orbiting a normal main-sequence star, we conclude that the scenario most consistent with the data in hand is the passage of a family of exocomet fragments, all of which are associated with a single previous breakup event. We discuss the necessity of future observations to help interpret the system.

Has the Kepler Space Telescope Discovered an Alien Megastructure?

By Ian O'Neill, Discovery News | October 15, 2015 12:13pm ET

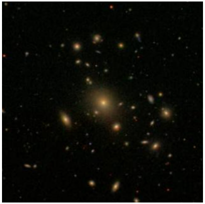
2893 Share
202 Tweet
17 Submit
374 Reddit
More



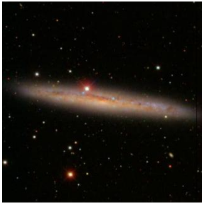
The artist's concept shows NASA's planet-hunting Kepler spacecraft operating in a new mission profile called K2.
Credit: NASA Ames/JPL-Caltech/T Pyle
View full size image

Pin it

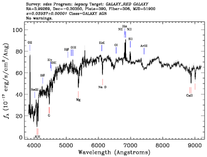
Group anomaly detection in astronomical data



(a) Cluster of galaxies

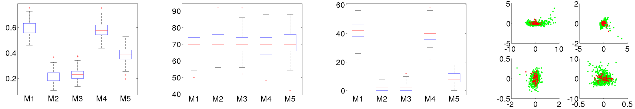


(b) A galaxy.

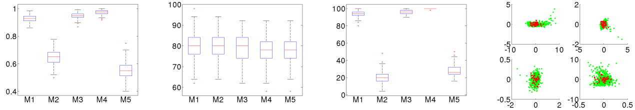


(c) Spectrum of a galaxy.

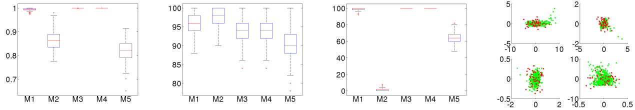
Pictures from http://www.sdss3.org/.



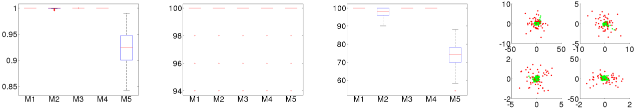
(a) AUC. (b) ACC Non A. (c) ACC Anom. (d) Means.



(e) AUC. (f) ACC Non A. (g) ACC Anom. (h) Means.



(i) AUC. (j) ACC Non A. (k) ACC Anom. (l) Means.



Outline

Conclusions and further research

- SMDD is a non parametric kernel method, with promising results in group anomaly detection task

Conclusions and further research

- SMDD is a non parametric kernel method, with promising results in group anomaly detection task
- Performance depends on the kernel choice, and regularization parameter

Conclusions and further research

- SMDD is a non parametric kernel method, with promising results in group anomaly detection task
- Performance depends on the kernel choice, and regularization parameter
- More models can be easily induced based on the idea of MV-sets

Conclusions and further research

- SMDD is a non parametric kernel method, with promising results in group anomaly detection task
- Performance depends on the kernel choice, and regularization parameter
- More models can be easily induced based on the idea of MV-sets

Further research

- Create models for classification, regression and other ML tasks

Conclusions and further research

- SMDD is a non parametric kernel method, with promising results in group anomaly detection task
- Performance depends on the kernel choice, and regularization parameter
- More models can be easily induced based on the idea of MV-sets

Further research

- Create models for classification, regression and other ML tasks
- exploit structural information of groups

Conclusions and further research

- SMDD is a non parametric kernel method, with promising results in group anomaly detection task
- Performance depends on the kernel choice, and regularization parameter
- More models can be easily induced based on the idea of MV-sets

Further research

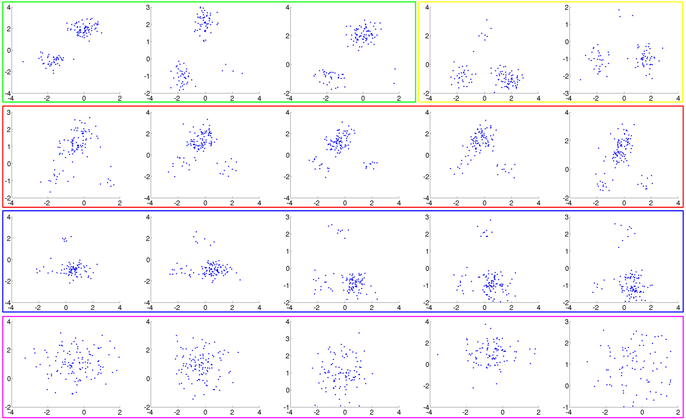
- Create models for classification, regression and other ML tasks
- exploit structural information of groups

References

- Guevara, Jorge and Canu, Stephane and Hirata Jr, Roberto Support Measure Data Description for Group Anomaly Detection 2015, SIGACM KDD, ODDx3'15.
- Guevara, Jorge and Canu, Stephane and Hirata Jr, Roberto Support Measure Data Description, 2014, Techical Report IME-USP.

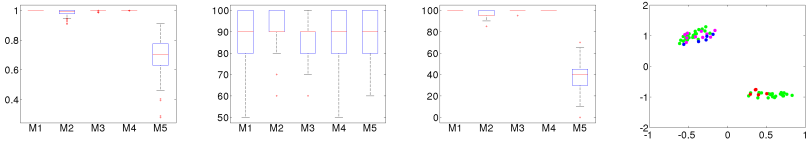
Any questions?

Point-based group anomaly detection over a Gaussian mixture distribution dataset



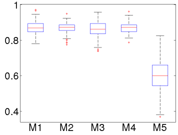
Red, blue and magenta boxes: group anomalies. Green and yellow boxes: non-anomalous groups.

- M4: OCSMM
- M5: SVDD
- kernel on probability measures: gaussian kernel, γ given by the median heuristic.
- 200 runs,training set=50 groups, test set = 30 groups (20 group anomalies).
- métricas AUC e ACC

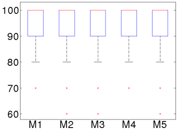


(a) AUC. (b) ACC Non A. (c) ACC Anom. (d) Médias grup.

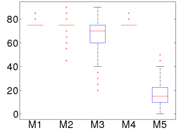
Distribution-based group anomaly detection over a Gaussian mixture distribution dataset



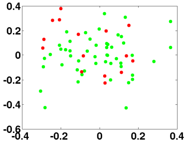
(a) AUC.



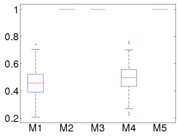
(b) ACC Non A.



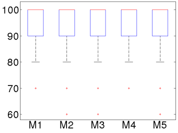
(c) ACC Anom.



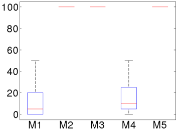
(d) Médias grup.



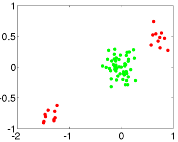
(e) AUC.



(f) ACC Non A.



(g) ACC Anom.



(h) Médias grup.