Indicative Support Vector Clustering

Application on Anomaly Detection

Huang Xiao

Chair of IT Security (I20)
Department of Informatics
Technische Universität München

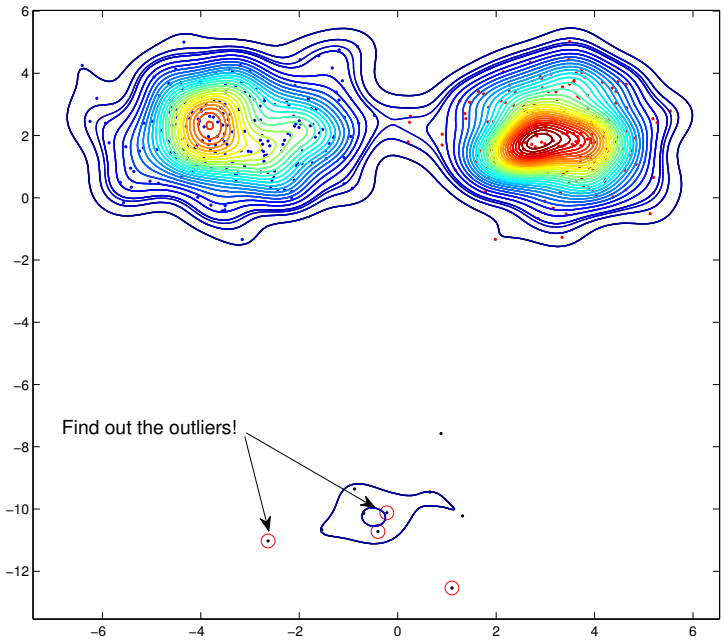
December 25, 2015

Overview

Fast KDE

To find the outliers...

Data points are scattered in ”some” space, e.g., Euclidean space, Hilbert space. Outliers are those points with lower density, which can be measured by Kernel Density Estimation (*KDE*).



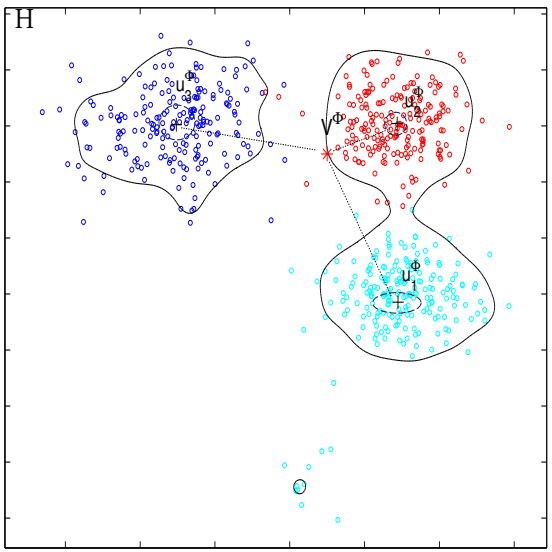
The kernel density estimate of x^* is

$$f(x^*) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x^* - x_i}{h})$$

Testing density is too much overhead!

Fast KDE: use clustering

Clustering the data points first



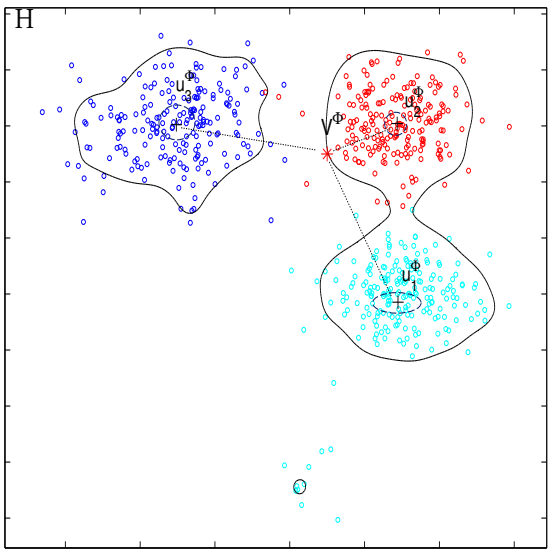
Averaging on m cluster centers:

$$\hat{V}^{\Phi} = \frac{1}{m} \sum_{i=1}^{m} w_i \Phi_i(x_s)$$

which is more efficient.

Fast KDE: use clustering

Clustering the data points first



Averaging on m cluster centers:

$$\hat{V}^{\Phi} = \frac{1}{m} \sum_{i=1}^{m} w_i \Phi_i(x_s)$$

which is more efficient.

BUT, lose too much precision, when m is small

Problems

- Density approximation only rely on cluster centers.
- Require to define how many clusters we have, i.e., $m =$?
- Outliers are averaged equally, not robust.
- We also need to integrate user feedback.

Another clustering?

Solution

Instead of looking for the cluster center, we look for the cluster bounds using support vectors.

Overview

How it works

In a mysterious space...

Data points can be structured into a hypersphere where low density points are fading away.

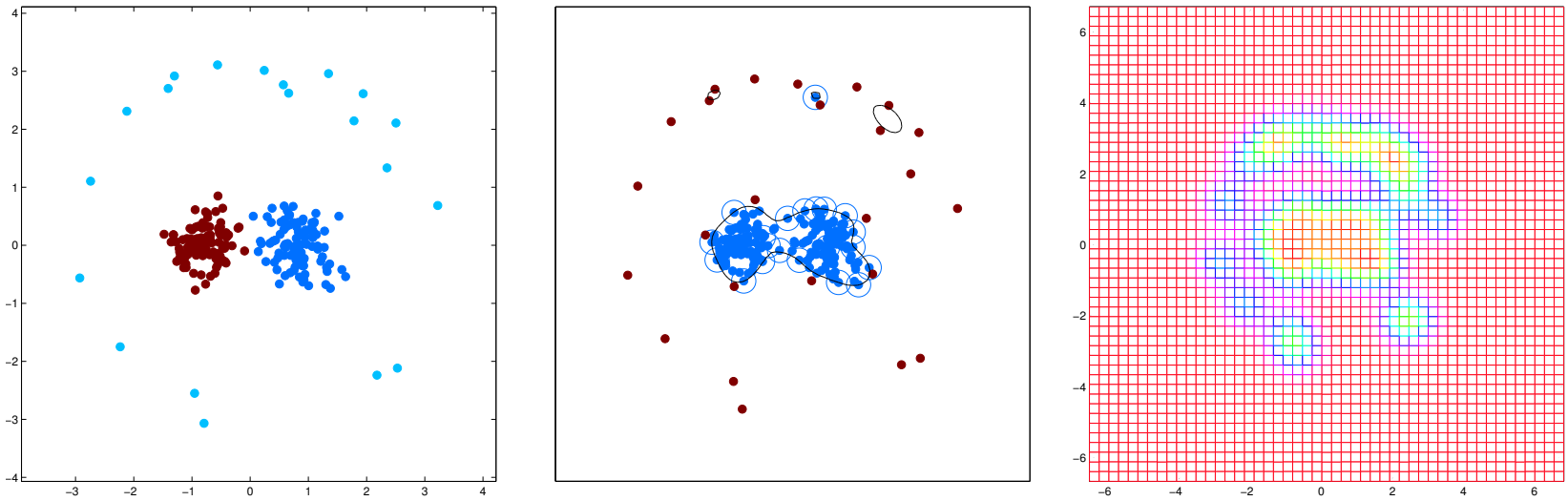
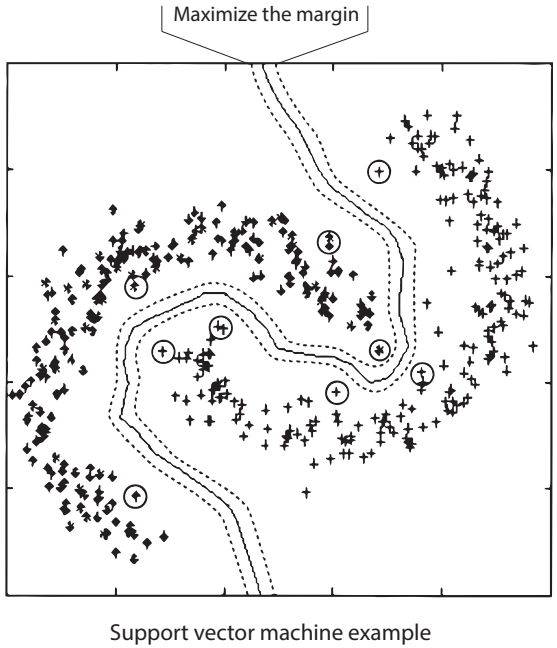


Figure: Support vector clustering example

Cutting the low density area

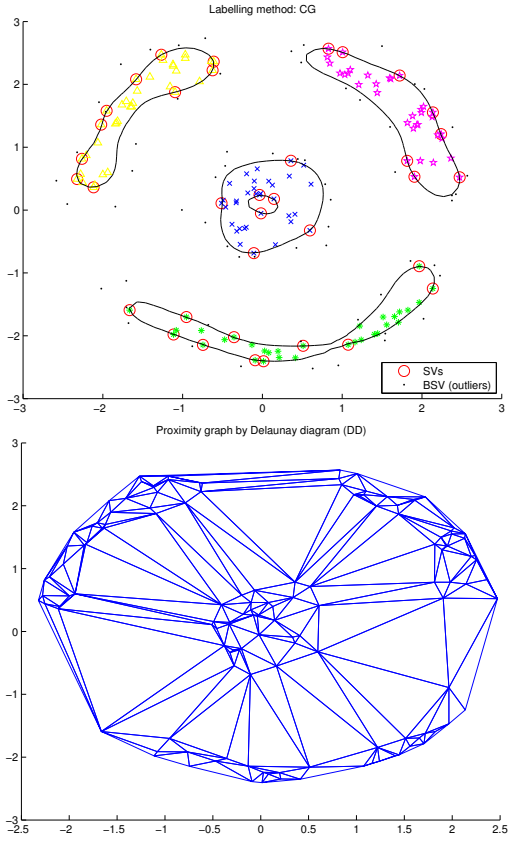
Support Vector Machine



Support vector machine example

Maximize the margin!

Support Vector Clustering



Minimize the hypersphere

Optimization problem

Looking for a smallest sphere $\langle R, g \rangle$ enclosing all data points:

$$\begin{aligned} & \min \left(R^2 + C \sum \xi_i \right) \\ \text{subject to} \quad & \|\Phi(x_j) - g\|^2 \leq R^2 + \xi_j, \ \forall j \\ & \xi_j \geq 0 \end{aligned}$$

R : the radius of the hypersphere

g : the center of the hypersphere

$\|\Phi(x_j) - g\|^2$: the distance of point x_j to center g

ξ_i : slackness variables

C : punishments on slackness

Optimization problem

Looking for a smallest sphere $\langle R, g \rangle$ enclosing all data points:

$$\begin{aligned} & \min \left(R^2 + C \sum \xi_i \right) \\ \text{subject to} \quad & \|\Phi(x_j) - g\|^2 \leq R^2 + \xi_j, \ \forall j \\ & \xi_j \geq 0 \end{aligned}$$

R : the radius of the hypersphere

g : the center of the hypersphere

$\|\Phi(x_j) - g\|^2$: the distance of point x_j to center g

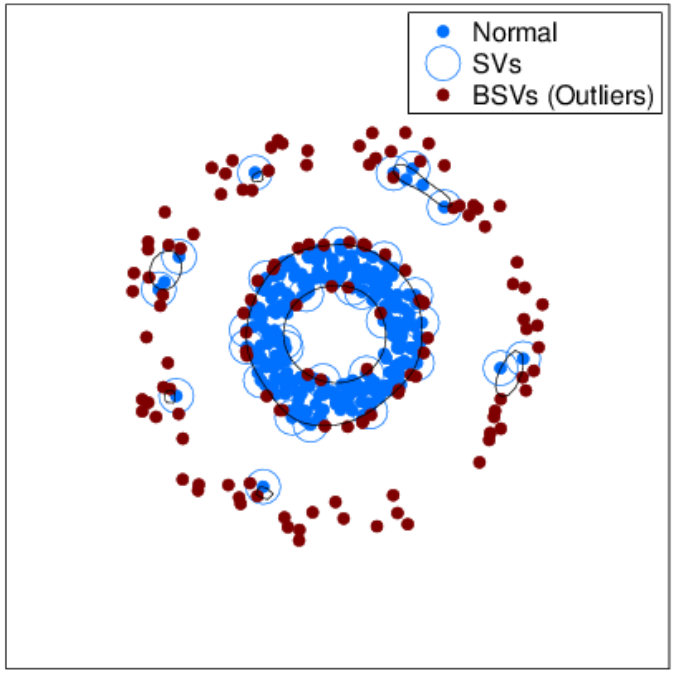
ξ_i : slackness variables

C : punishments on slackness

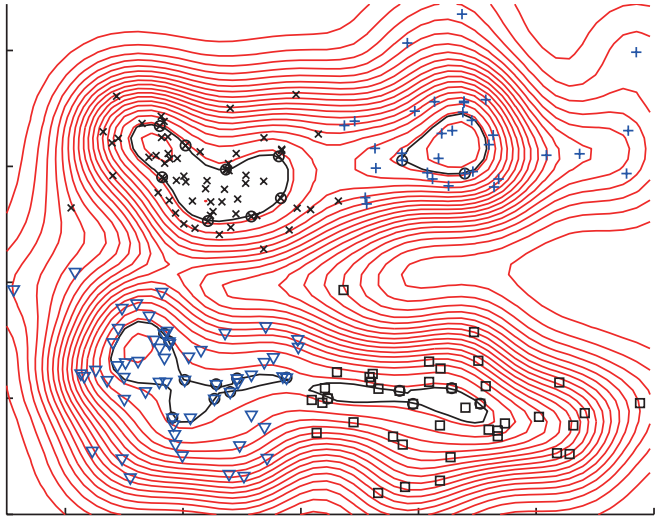
Minimizing the ball allowing part of points fade way (slackness).

Properties

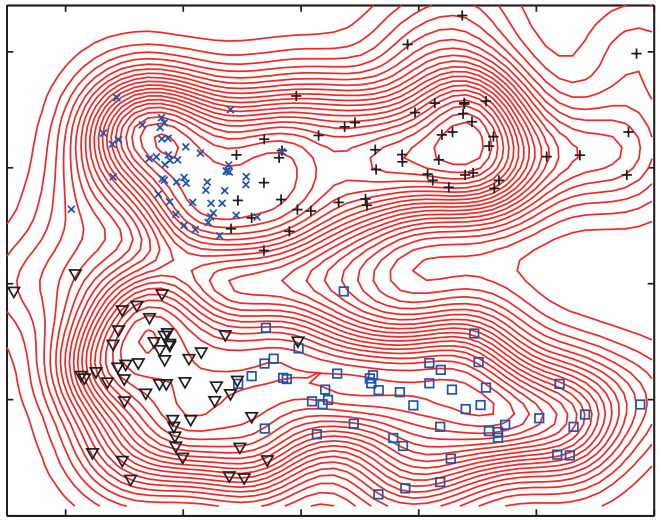
- The minimal ball with R is cutting the data points from low density area.
- Normal data points are inside the ball.
- Support vectors X_{SVs} are directly on the sphere.
- Outliers with $\xi_i > 0$ is outside the ball.
- The clustering bound is a natural threshold finding out the outliers.
- No need to define the m (#clusters)



From the cluster bound to centers



(a) SVC topographic map



(b) KDE topographic map

- Shrink the cluster bounds to centers.
- Allow all the points to be repelled outside the ball.

Advantages over KDE

- Computational easier
- Define a core region rather than a peak
- Topographically more reliable
- Global optimal solution rather than many local maxima

Optimization results

Finally we got...

$$g = \sum \beta_{\mathrm{sv}} \Phi(x_{\mathrm{sv}}) + \sum \beta_{\mathrm{bsv}} \Phi(x_{\mathrm{bsv}}) + \sum \beta_{\mathrm{in}} \Phi(x_{\mathrm{in}}),$$

$$0 < \beta_{\mathrm{sv}} < C,\ \beta_{\mathrm{bsv}} = C,\ \beta_{\mathrm{in}} = 0$$

Optimization results

Finally we got...

$$g = \sum \beta_{\mathrm{sv}} \Phi(x_{\mathrm{sv}}) + \sum \beta_{\mathrm{bsv}} \Phi(x_{\mathrm{bsv}}) + \sum \beta_{\mathrm{in}} \Phi(x_{\mathrm{in}}),$$

$$0 < \beta_{\mathrm{sv}} < C,\ \beta_{\mathrm{bsv}} = C,\ \beta_{\mathrm{in}} = 0$$

Note: the ball center g is still heavily biased by outliers (β_{bsv})

Problems

- Only deal with low density outliers
- Learning is not robust (g is biased)
- We need to integrate user feedbacks

Overview

Punishments are not fair

Recall the penalty term...

$$\min \left(R^2 + C \sum \xi_i \right)$$

The higher the punishment value C is...

- the lower the summation $\sum \xi_i$ is
- the less likely $\xi_i > 0$
- the less likely data points $\{x_i\}$ run out of sphere
- the less outliers are allowed!

Punishments are not fair

Recall the penalty term...

$$\min \left(R^2 + C \sum \xi_i \right)$$

The higher the punishment value C is...

- the lower the summation $\sum \xi_i$ is
- the less likely $\xi_i > 0$
- the less likely data points $\{x_i\}$ run out of sphere
- the less outliers are allowed!

We see that, for all the outliers, penalty is the same C.

Reweight the penalties

Now we treat data points individually instead of equivalently penalizing all the input data with a constant C...

$$\min \left(R^2 + \sum c_i \xi_i \right)$$

The larger the penalty c_i is, the less possible the point x_i would be driven away from the hypersphere, and vice versa.

Reweight the penalties

Now we treat data points individually instead of equivalently penalizing all the input data with a constant C...

$$\min \left(R^2 + \sum c_i \xi_i \right)$$

The larger the penalty c_i is, the less possible the point x_i would be driven away from the hypersphere, and vice versa.

Full control over each point!

Assumption

Anomalies are the patterns found to behave distinctly from the normal patterns, and similarly behaving instances are more likely hosted in the same cluster.

User given labels

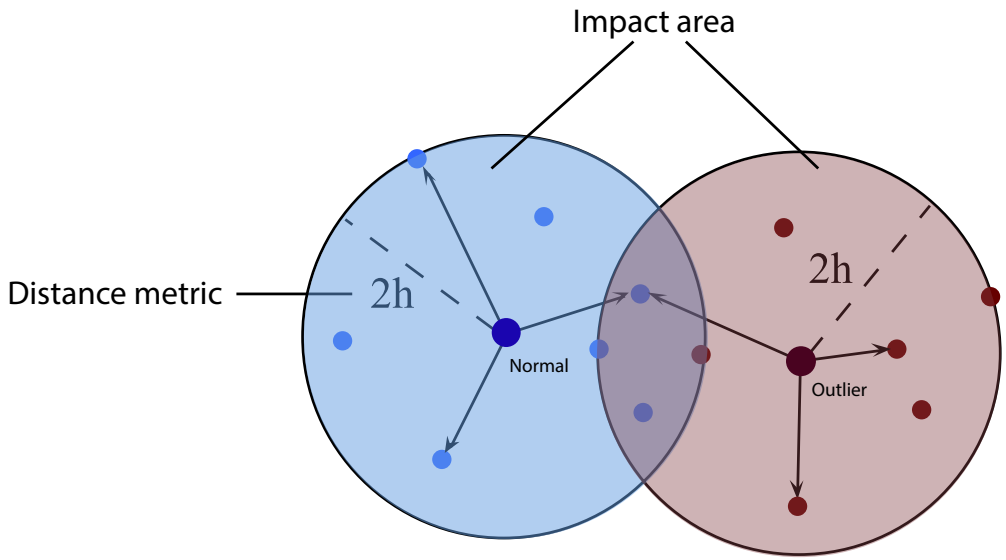
Suppose part of user feedback is available,

$$\mathcal{X}^+ = \{(x_l, y_l) \mid x_l \in \mathcal{X}, y_l = 1)\}$$
$$\mathcal{X}^- = \{(x_r, y_r) \mid x_r \in \mathcal{X}, y_r = -1)\}$$
$$l, r \in \{1, \ldots, N\}$$

$\mathcal{X}^+$ is set of outliers indicated by users, and $\mathcal{X}$ a set of normal samples.

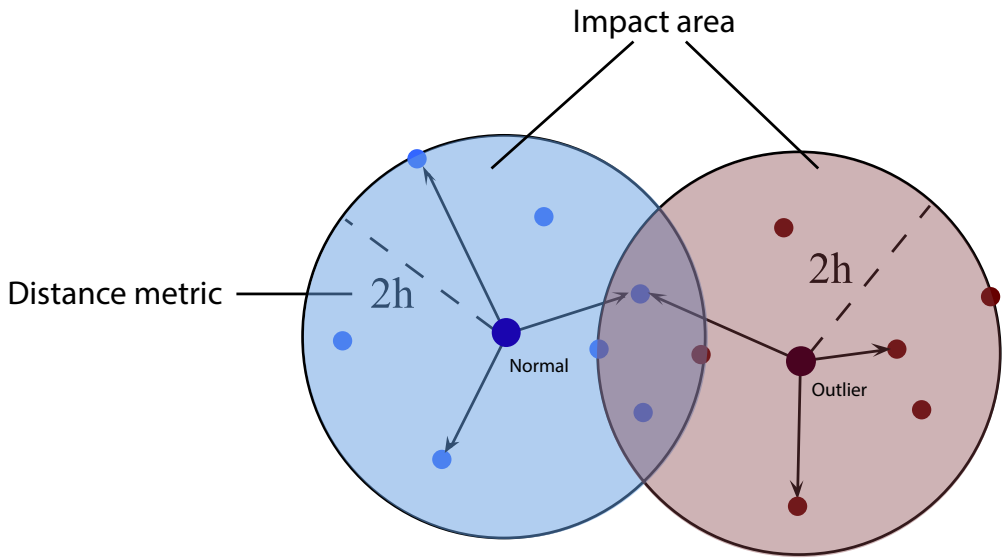
Impact function

The closer to the indicated data, the more likely they belong to the same set (normal/outlier)



Impact function

The closer to the indicated data, the more likely they belong to the same set (normal/outlier)



So, the penalties can be reweighed by distances to the given labeled data points.

We observed

Points near the given outliers $\mathcal{X}^+$

- Penalties c_j are small
- More likely to be repelled
- Impact on cluster center g is small

We observed

Points near the given outliers $\mathcal{X}^+$

- Penalties c_j are small
- More likely to be repelled
- Impact on cluster center g is small

Points near the given normals $\mathcal{X}^-$

- Penalties c_j are big
- More likely to be enclosed into the ball
- Impact on cluster center g is larger

We observed

Points near the given outliers $\mathcal{X}^+$

- Penalties c_j are small
- More likely to be repelled
- Impact on cluster center g is small

Points near the given normals $\mathcal{X}^-$

- Penalties c_j are big
- More likely to be enclosed into the ball
- Impact on cluster center g is larger

Leads to...

- Robustness
- Reliable outlier detection
- Better clustering results
- More compact bound

Overview

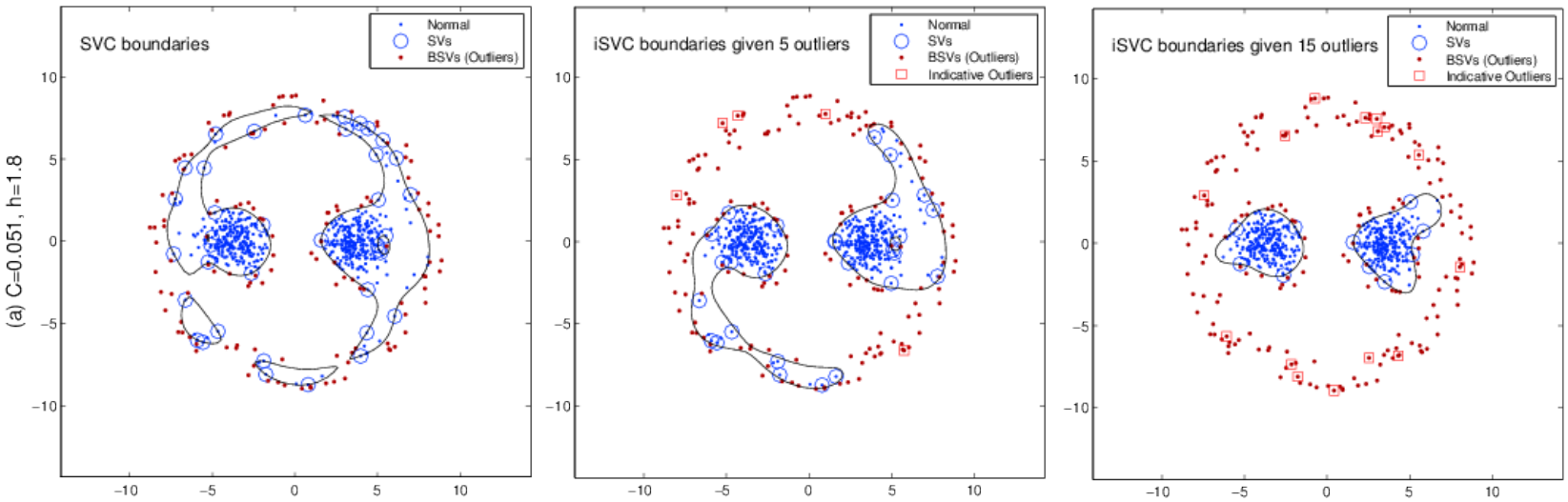
On synthetic data set

Setup

- Two normal classes data: 250 samples of each
- A ring-like Outlier data set: 150 samples
- Given different sets of data labels

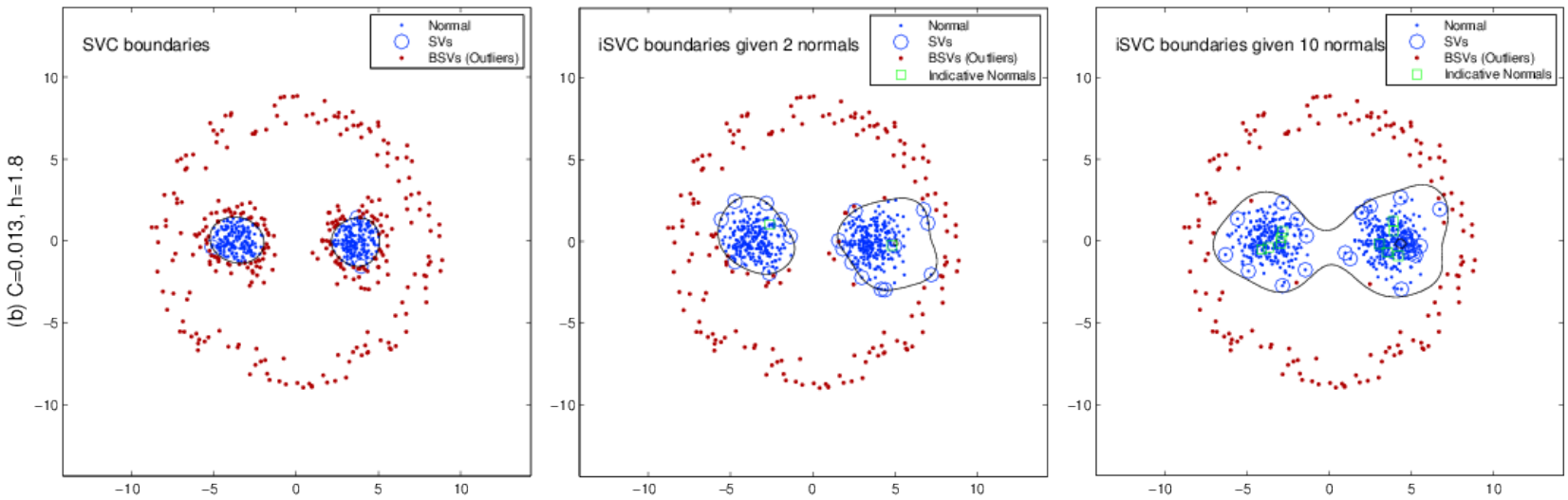
Results I

Given 5 outliers / 15 outliers...



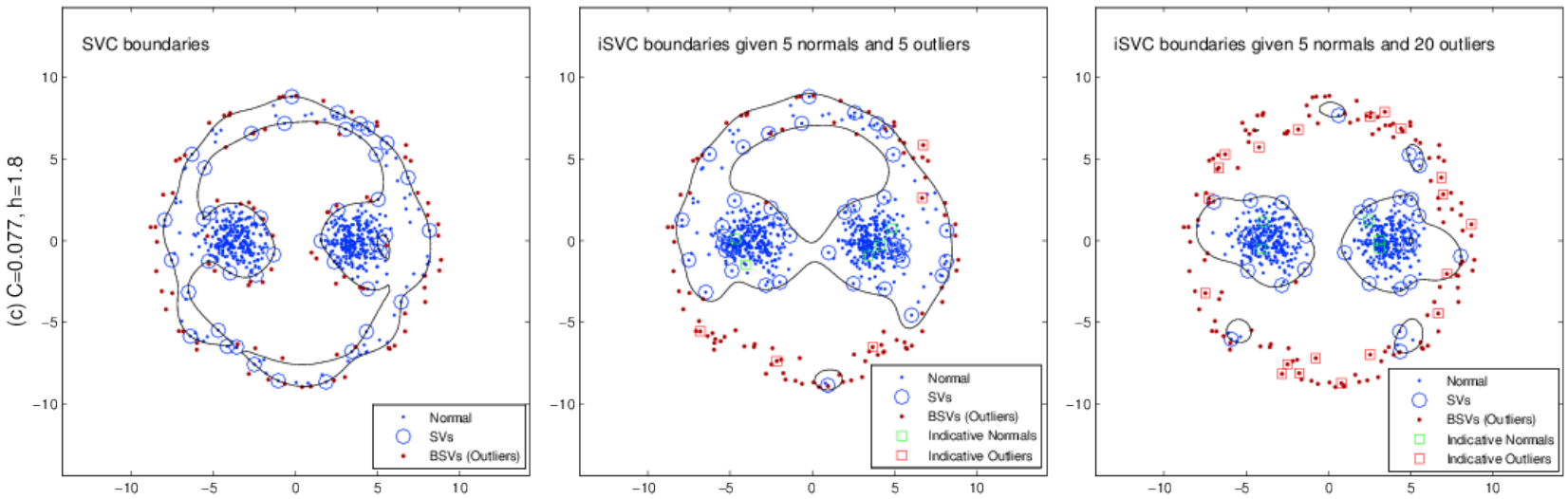
Results II

Given 2 Normals / 10 Normals...



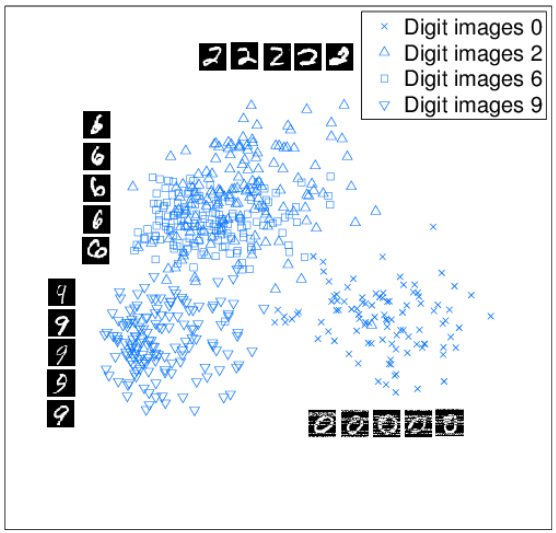
Results III

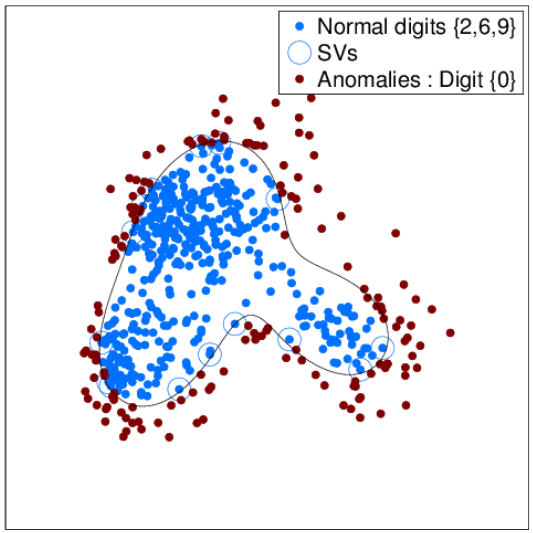
Given 5 Normals and 5 Outliers / 5 Normals and 20 outliers...

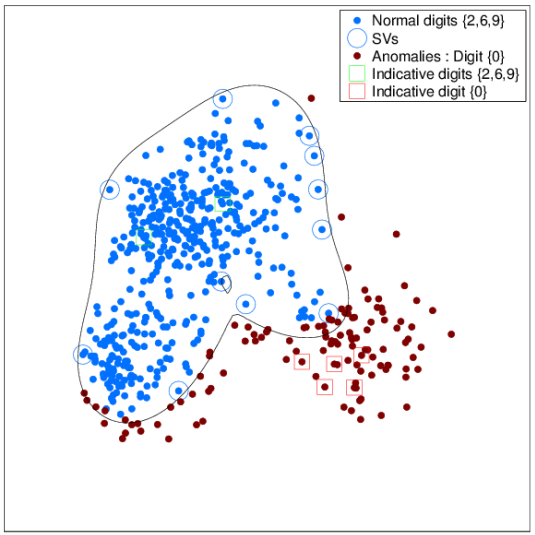


On digits recognition data set

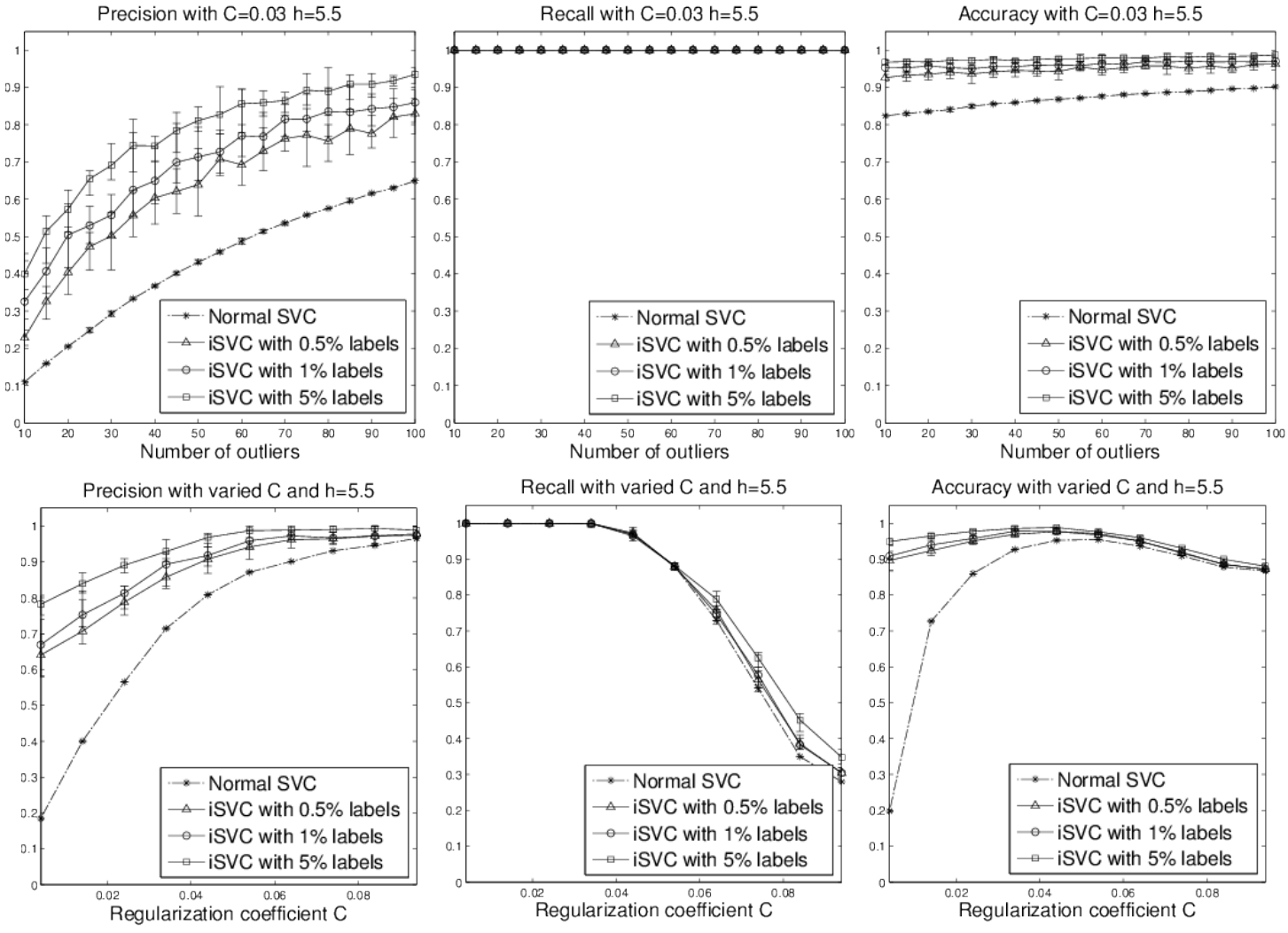
- MINST digit images data set
- Digits 2,6,9, 150 images of each
- 100 digit 0 tampered manually as outliers







Numerical evaluation



On the WDBC data set

- Wisconsin Diagnostic Breast Cancer (WDBC) data set
- 569 samples, out of which 212 are malignant and 357 are benign
- Test it against semi-supervised fast linear SVM given various labels

Results

	WDBC data set with h $= 0.9,$ C $= 0.001$			
	Accuracy	F_1-measure	FPR	FNR
SVC	63.4%	50.5%	28.6%	50%
	5% *positive* and 0% *negative* labels			
iSVC	**90.5**%	**87.0**%	**6.4**%	14.6%
S3VM	39.2%	55.0%	96.9%	**0**%
	0% *positive* and 5% *negative* labels			
iSVC	**89.8**%	**87.7**%	14.8%	2.4%
S3VM	64.9%	10.7%	**0**%	94.3%
	5% *positive* and 5% *negative* labels			
iSVC	**92.3**%	**89.8**%	**7.0**%	8.9%
S3VM	88.4%	86.3%	17.4%	**1.9**%
	10% *positive* and 10% *negative* labels			
iSVC	**93.9**%	**91.9**%	**6.4**%	5.6%
S3VM	92.4%	90.6%	10.4%	**2.8**%

Overview

Conclusion

We have

- More robust clustering method
- Outliers are more reliable
- User's feedback can be integrated

Future work

- More sophisticated impact function
- Online mode of iSVC
- Automation of model selection

Overview

References


Huang Xiao, Claudia Eckert. Indicative support vector clustering and its application on anomaly detection. *In IEEE 12th International Conference on Machine Learning and Applications* , Dec. 2013.

Appendix

Given a data set of N points $\{x_i\}_{i=1}^{N} \subseteq \mathcal{X}$, defining a non-linear feature mapping Φ from $\mathcal{X}$ to a higher dimensional Hilbert space $\mathcal{H}$,

$$\begin{aligned} & \min R^2 + \sum c_i \xi_i \\ \text{subject to} \quad & \|\Phi(x_j) - g\|^2 \leq R^2 + \xi_j, \ \forall j \\ & \xi_j \geq 0 \end{aligned} \tag{1}$$

Introducing the Lagrangian

$$\begin{aligned} L = R^2 - \sum_j \left(R^2 + \xi_j - \|\Phi(x_j) - g\|^2\right) \beta_j \\ - \sum \xi_j \mu_j + \sum c_i \xi_j, \end{aligned} \tag{2}$$

$\beta_j \geq 0$ and $\mu_j \geq 0$ are Lagrange multipliers, c_i is the regularization constant.

Two supervised data sets are indicated by users,

$$\begin{aligned}\mathcal{X}^{+} &= \{(x_l, y_l) \mid x_l \in \mathcal{X}, y_l = 1)\} \\ \mathcal{X}^{-} &= \{(x_r, y_r) \mid x_r \in \mathcal{X}, y_r = -1)\} \\ l, r &\in \{1, \ldots, N\}\end{aligned}$$

where $\mathcal{X}^{+}$ is the outlier set, and $\mathcal{X}^{-}$ is normal set.

An impact function f is defined in the input space $\mathcal{X}$ given $\mathcal{X}^+$ and $\mathcal{X}^-$.

$$f(x_i) = \frac{\sum_{x_l \in \mathcal{X}^+} K(x_l, x_i)}{\sum_l \mathbf{1}^+(x_l, x_i)} + \frac{\sum_{x_r \in \mathcal{X}^-} K(x_r, x_i)}{\sum_l \mathbf{1}^-(x_r, x_i)} \tag{3}$$

where the kernel function is taken as a similarity measurement, note that

$$K(x, x_i) = \exp\left(\frac{\|x - x_i\|^2}{-2h^2}\right)$$

And $\mathbf{1}^+$ and $\mathbf{1}^-$ are both indicator functions defined as

$$\mathbf{1}^+(x_l, x_i) = \begin{cases} 1 & \|x_l - x_i\| \leq 2h \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{1}^-(x_r, x_i) = \begin{cases} -1 & \|x_r - x_i\| \leq 2h \\ 0 & \text{otherwise} \end{cases}$$

The regularization weights c_j can be computed

$$c_j = \begin{cases} c_0 \cdot \frac{1-f\left(x_j\right)}{1-\exp(-2)} + \frac{1}{N} \cdot \frac{f\left(x_j\right)-\exp(-2)}{1-\exp(-2)} & \text{if} \quad f\left(x_j\right) > 0 \\ c_0 \cdot \frac{1-\left|f\left(x_j\right)\right|}{1-\exp(-2)} + \frac{\left|f\left(x_j\right)\right|-\exp(-2)}{1-\exp(-2)} & \text{if} \quad f\left(x_j\right) < 0 \end{cases} \tag{4}$$

where c_0 is the initial value of c_j and N is the sample size.

Now setting the derivative of the Lagrange L to zero, we have:

$$\sum \beta_j = 1 \tag{5}$$

$$g = \sum \beta_j \Phi(x_j) \tag{6}$$

$$\beta_j = c_j - \mu_j \tag{7}$$

The KKT conditions again require that

$$\xi_j \mu_j = 0 \tag{8}$$

$$\left(R^2 + \xi_j - \|\Phi(x_j) - g\|^2\right) \beta_j = 0 \tag{9}$$

Eliminating the variables R, g, and u_j, the Wolf dual form is obtained, where β_j are the only variables.

$$W = \sum_j \Phi(x_j)^2 \beta_j - \sum_{i,j} \beta_i \beta_j \Phi(x_i) \cdot \Phi(x_j) \tag{10}$$

$$\text{subject to} \quad 0 \leq \beta_j \leq c_j, \; j = 1, \ldots, N. \tag{11}$$

Solve the dual problem, we have found $\{\beta_j\}$.