ν-Anomica: A Fast Support Vector based Novelty Detection Technique

IEEE International Conference on Data Mining series, Miami, FL,USA, on December 6-9, 2009.

Santanu Das[1], Kanishka Bhaduri[2], Nikunj Oza[3]
and Ashok Srivastava[3]

[1]UARC UC Santa Cruz, [2]MCT Inc.,
and [3]NASA Ames Research Center

Outline

- Motivation
- Anomaly detection algorithm
  - Description of benchmark algorithm
  - Proposed method
    - Approach
    - ν-criterion
    - Key steps to implementation
- Results on data sets with variety of conditions.
- Conclusions and future work

This talk is about..

While developing algorithms to detect, classify, and predict events in large data streams for scientific and engineering systems one of the important issues of today is..

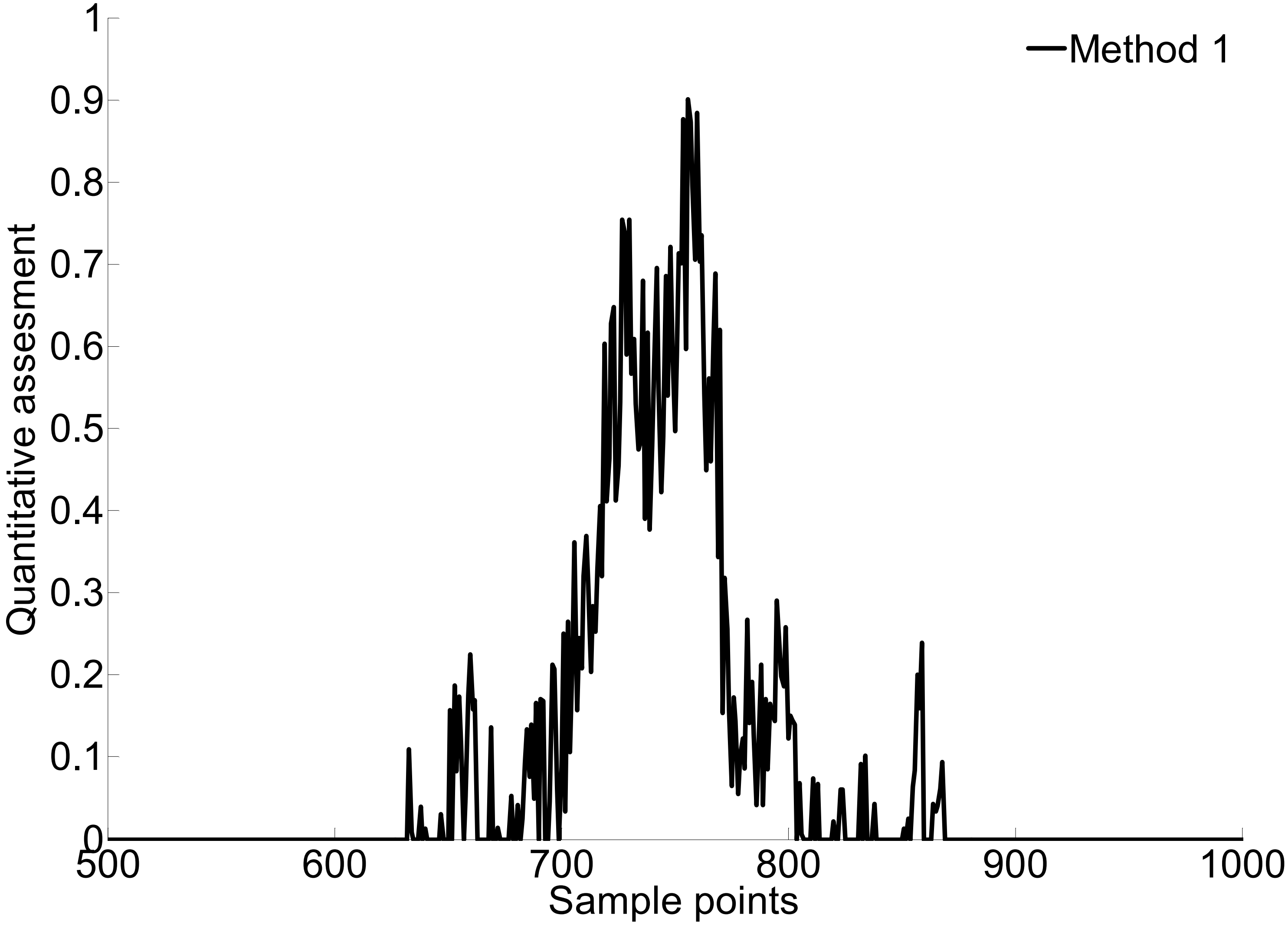
Scalable analysis of data

In summary..

Computationally efficient implementation of anomaly detection algorithm

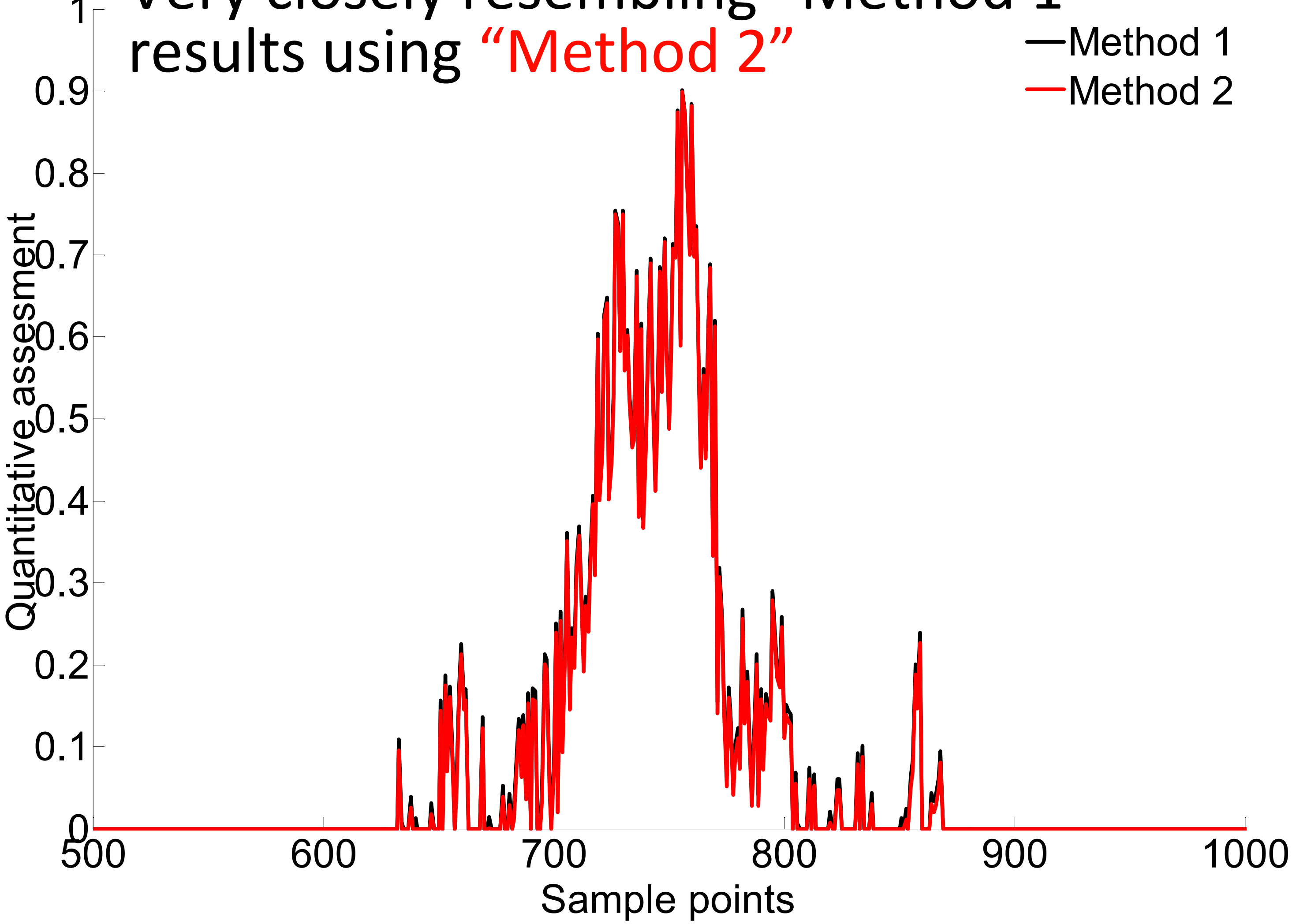
Some results..

Detecting outliers using “Method 1”



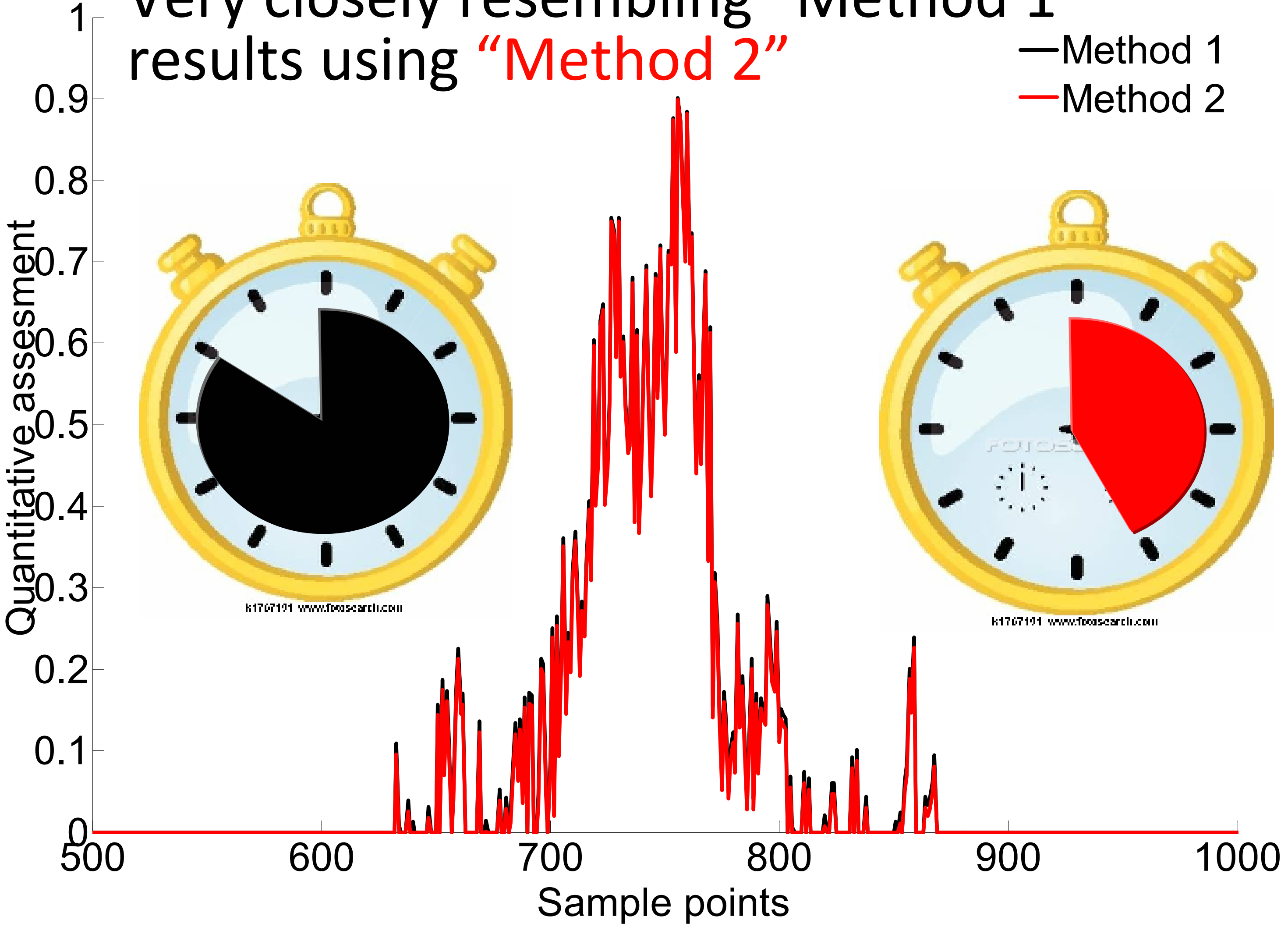
Reproducing similar results..

Very closely resembling “Method 1” results using “Method 2”



..but in reduced time

Very closely resembling “Method 1” results using “Method 2”



From research standpoint

Develop an anomaly detection algorithm "Method 2" for continuous data sources, such that "Method 2" retains same accuracy with lower running time compared to "Method 1" which is a benchmark algorithm.

Experimentally demonstrate "Method 2" with wide verity of data sets of varying sizes and dimensionalities.

Method 1: One-class SVMs

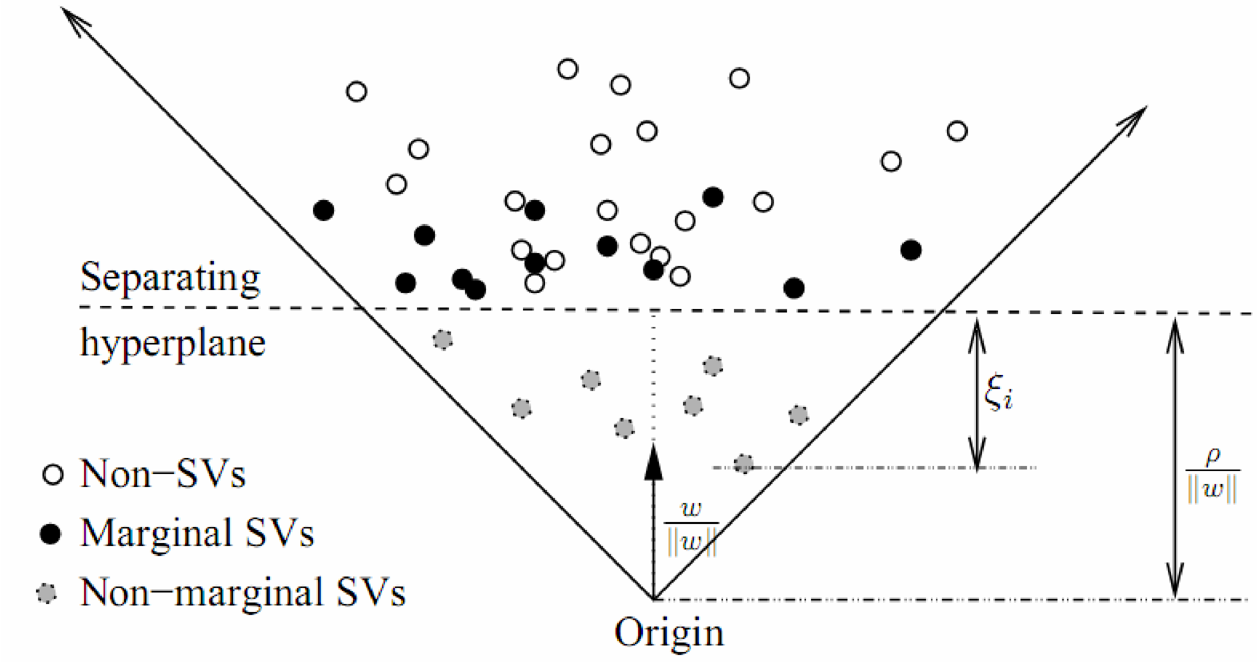
One-class SVMs (Schölkopf et al.) is a kernel based approach that solves an optimization problem which results in a model that can be used to perform anomaly detection.

$$\min_{\alpha} \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j)$$

Subject to

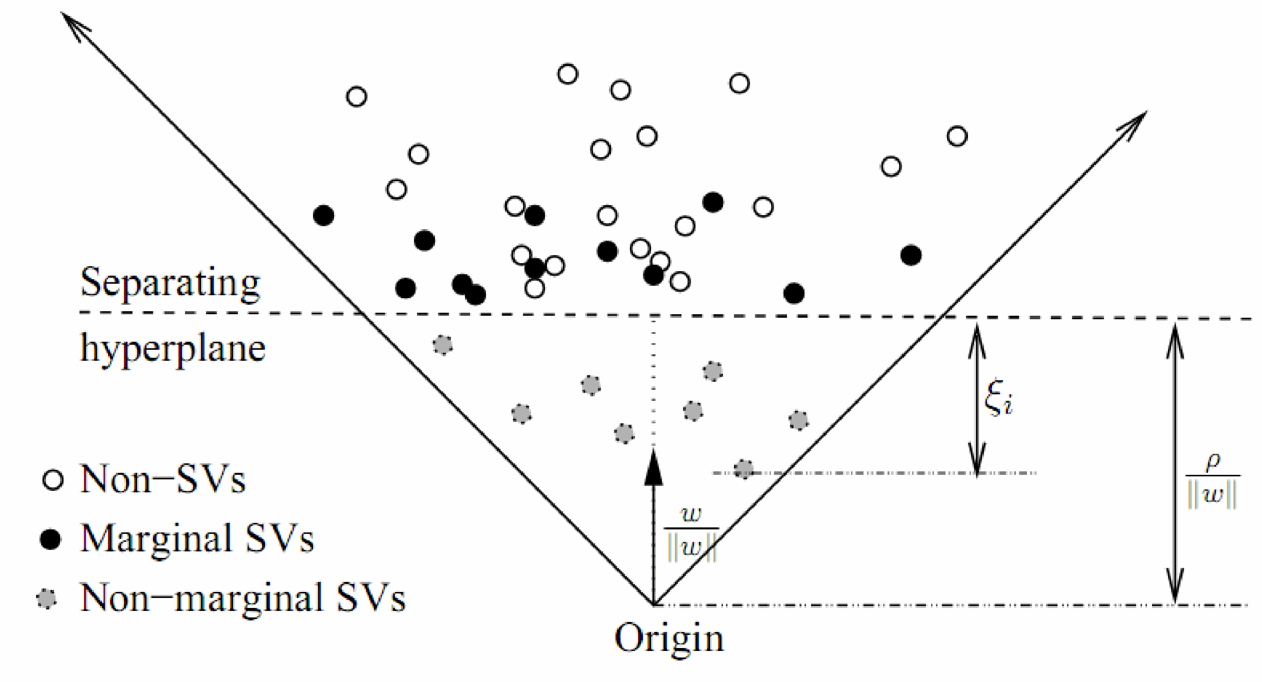
$$0 \leq \alpha_i \leq \frac{1}{lv}, \sum \alpha_i = 1 \qquad \rho = \sum_i \alpha_i K(x_i, x_j)$$

One-class SVMs : cont.



Finding a small fraction of the training data that can be linearly separated from the remainder. The resulting model can be used to classify new examples.

ν-criterion



This means the one-class model, once built, should be able to correctly classify 1−ν fraction of the entire training set as normal examples.

Approach: “Method 2”

“Method 2” is a variant of one-class SVMs that closely approximates the exact decision plane.

“Method 2” does this by,

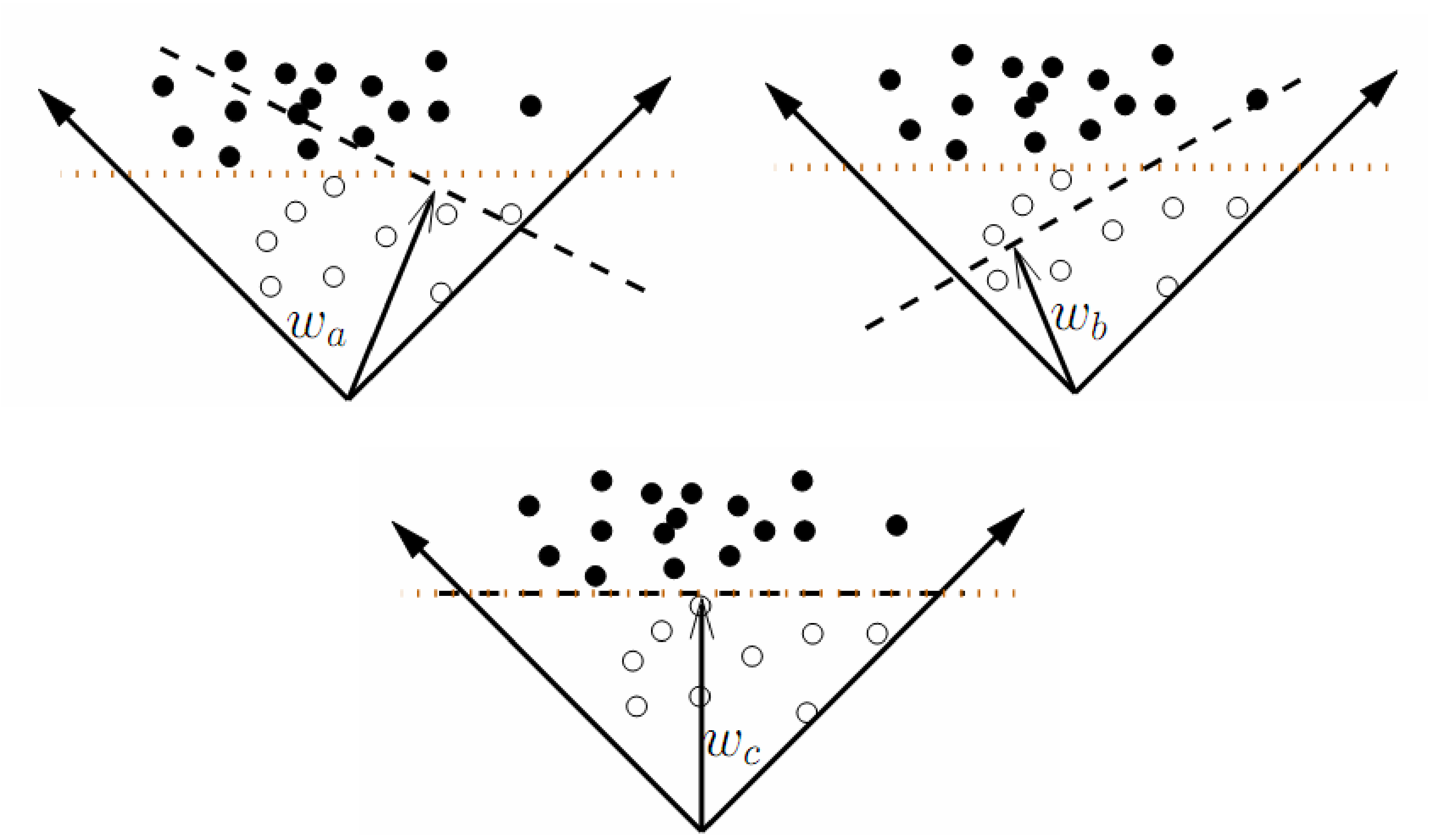
- Dividing the entire data set into a training set and a hold out set for validation purpose
- Building a model on “active training set” which is a reduced subset of the entire training set

If this newly developed model is a close approximation of exact solution, it must meet the “ν-criterion”.

Approach: “Method 2” (cont.)

- Evaluating the current model on the hold out set
- Checking “v-criterion”
- Iteratively updating the “active training set” by adding enough examples from the remaining training examples to satisfy the “v-criterion”.

Update rules



Nomenclature

Since “ν” is an important parameter that guides the “target solution” in “Method 2” and the overall goal is to build the model on large training set with the intention to find anomalous or unusual events, we have named the algorithm as “ν−Anomica”.

Anomie: Social disorder (Wikipedia)

Here onwards ‘~~Method 2~~’………….

Experimental conditions

- Tested on a dual core Pentium4 computer running Windows XP with 4 GByte of memory.
- Based on the OSU SVM Classifier Toolbox (ver. 3.00, Matlab)
- Nonlinear RBF kernel used

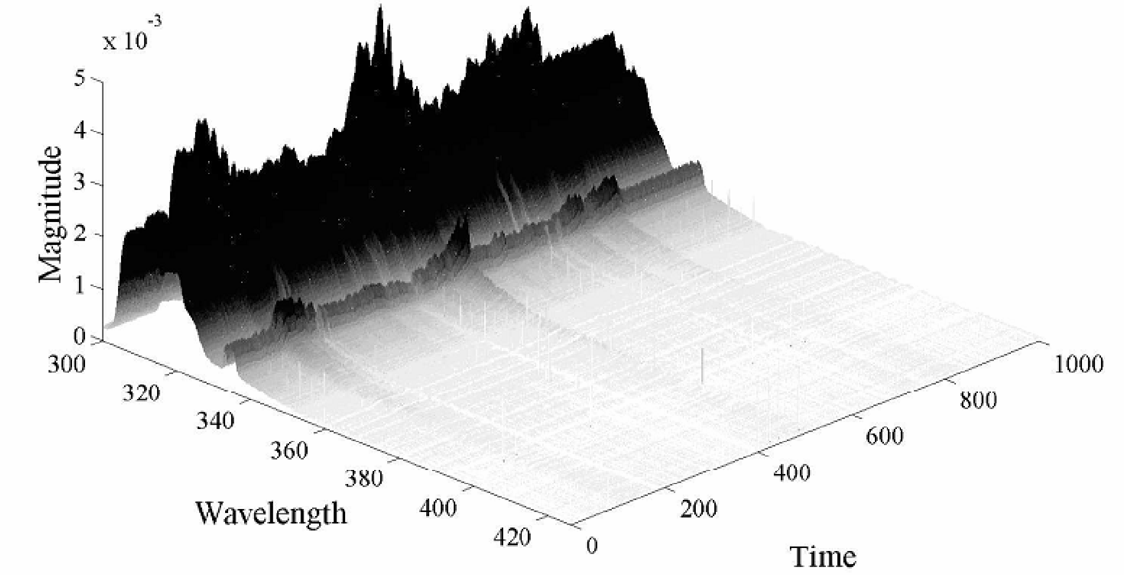
$$K\left(\vec{x}, \vec{x}_i\right) = \exp\left(-\frac{1}{2\sigma^2}\left\|\vec{x} - \vec{x}_i\right\|^2\right)$$

- Initial subset: 15% of the entire training set

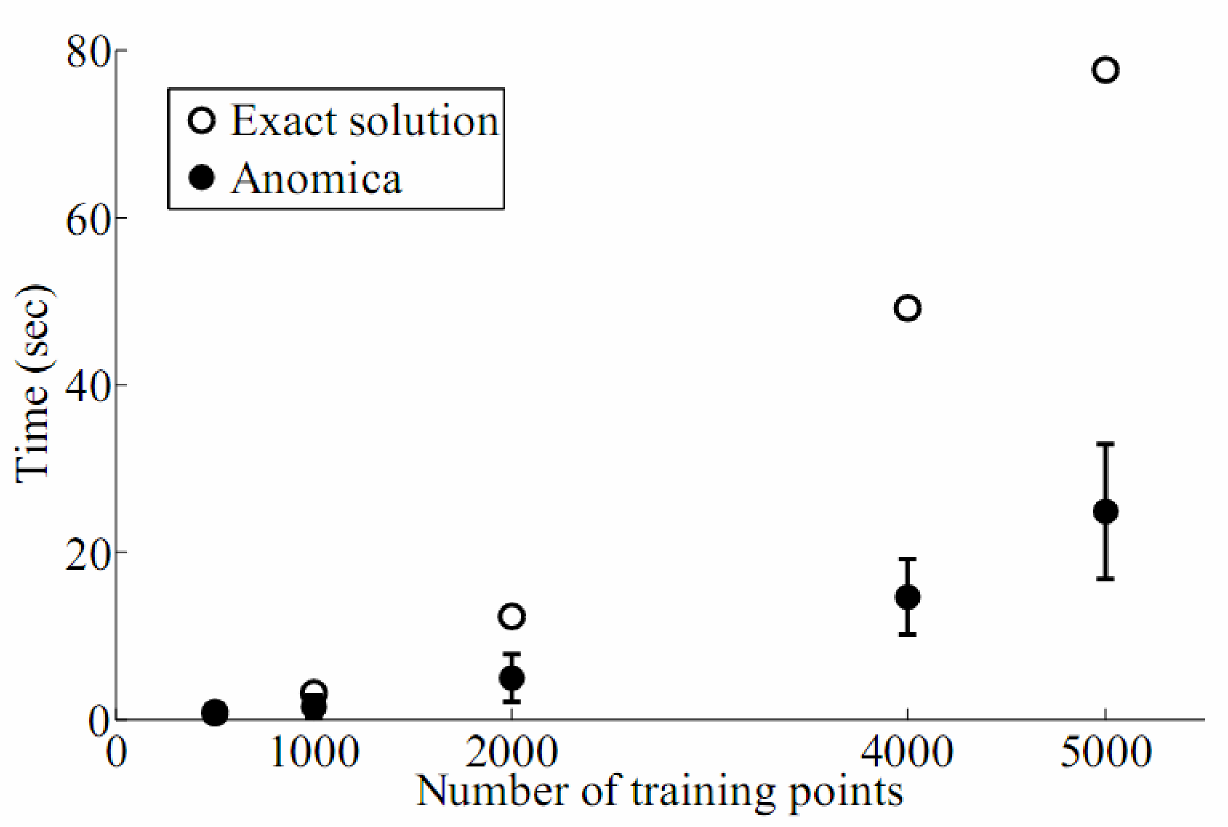
Data set-I

Emulated OPAD (Optical Plume Anomaly Detection) data

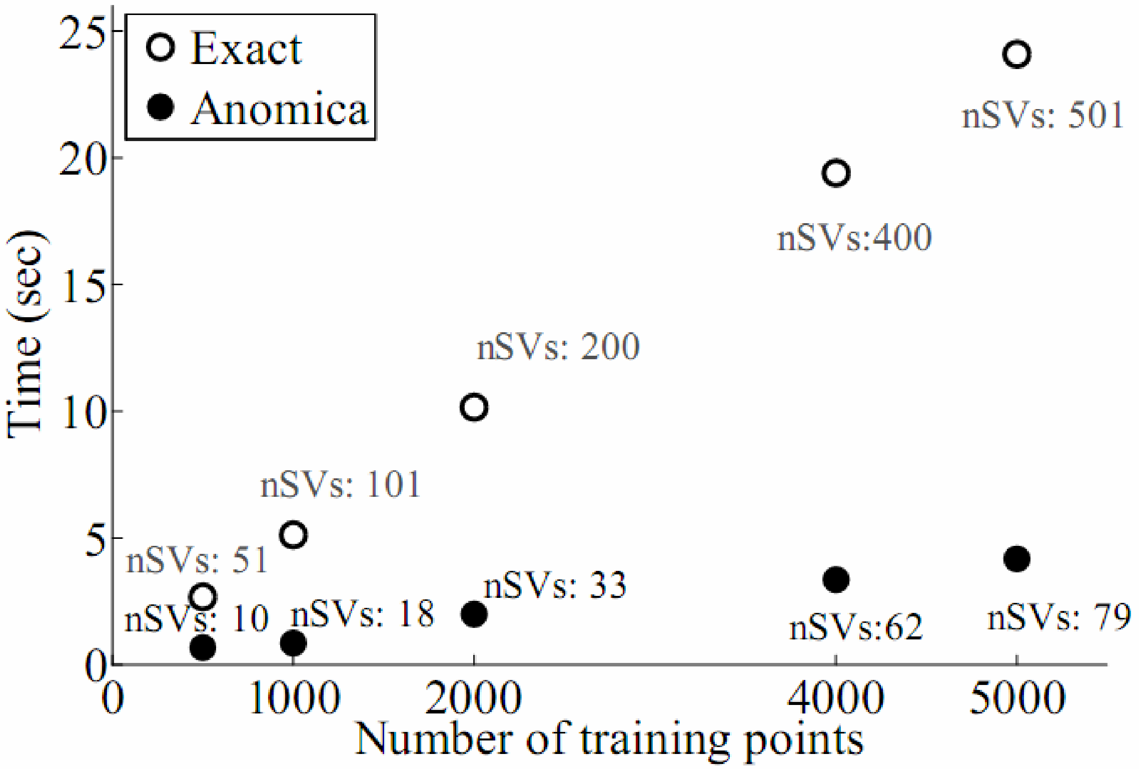
- Entire training set: 5k x 1024
- Entire hold out set: 5k x 1024
- Entire testing set: 2k x 1024



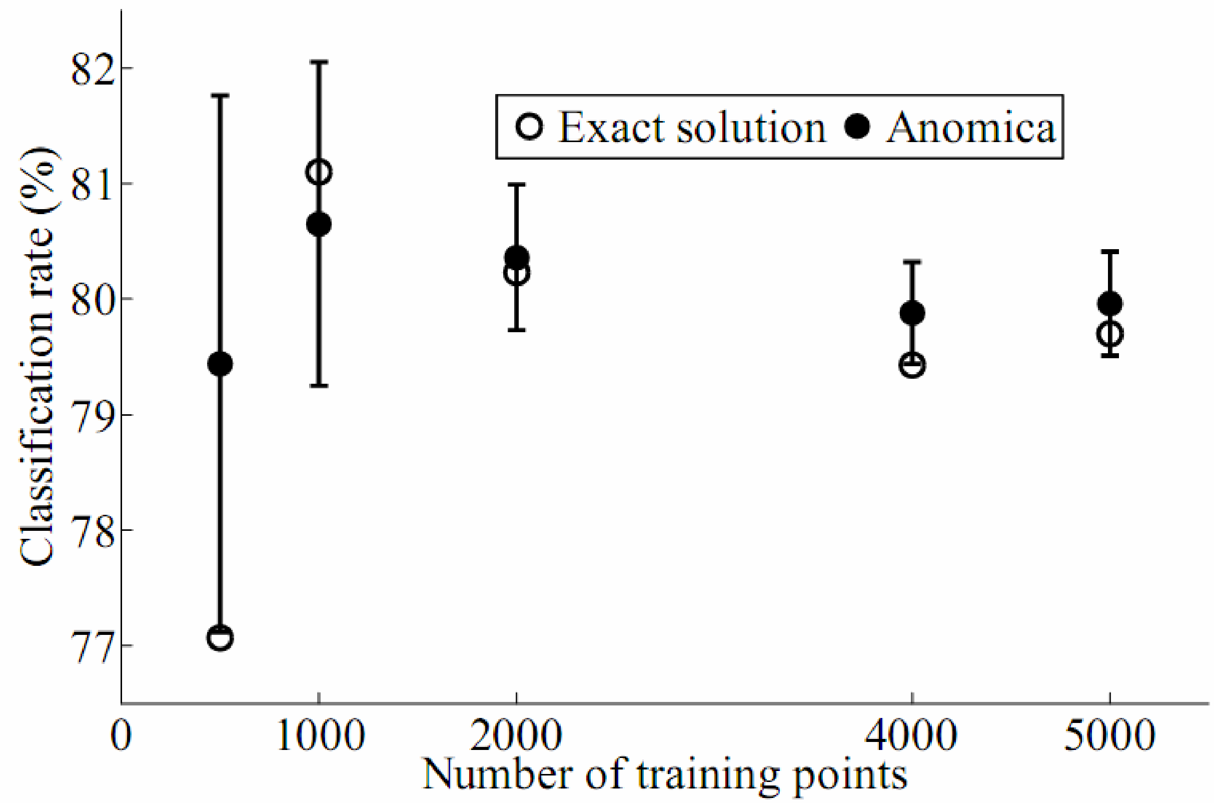
Run-time: training



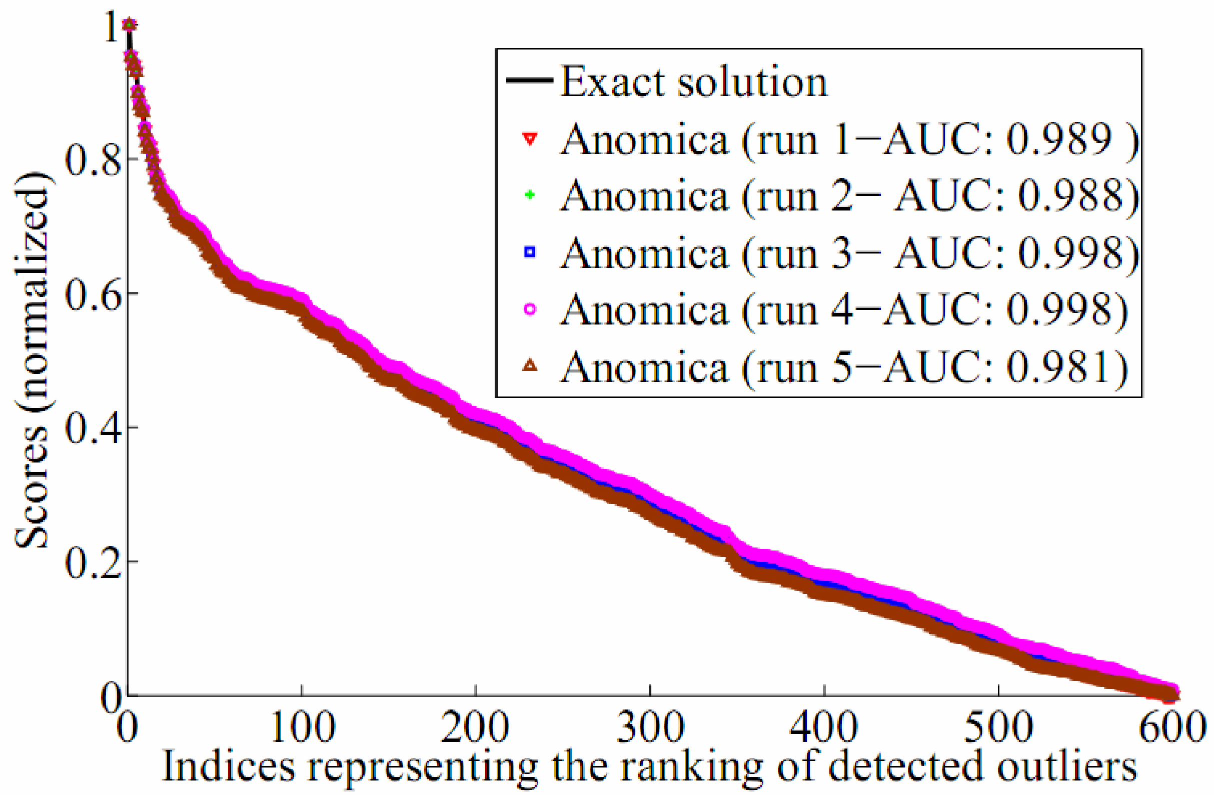
Run-time: testing



Test accuracy



Detection: top n outliers



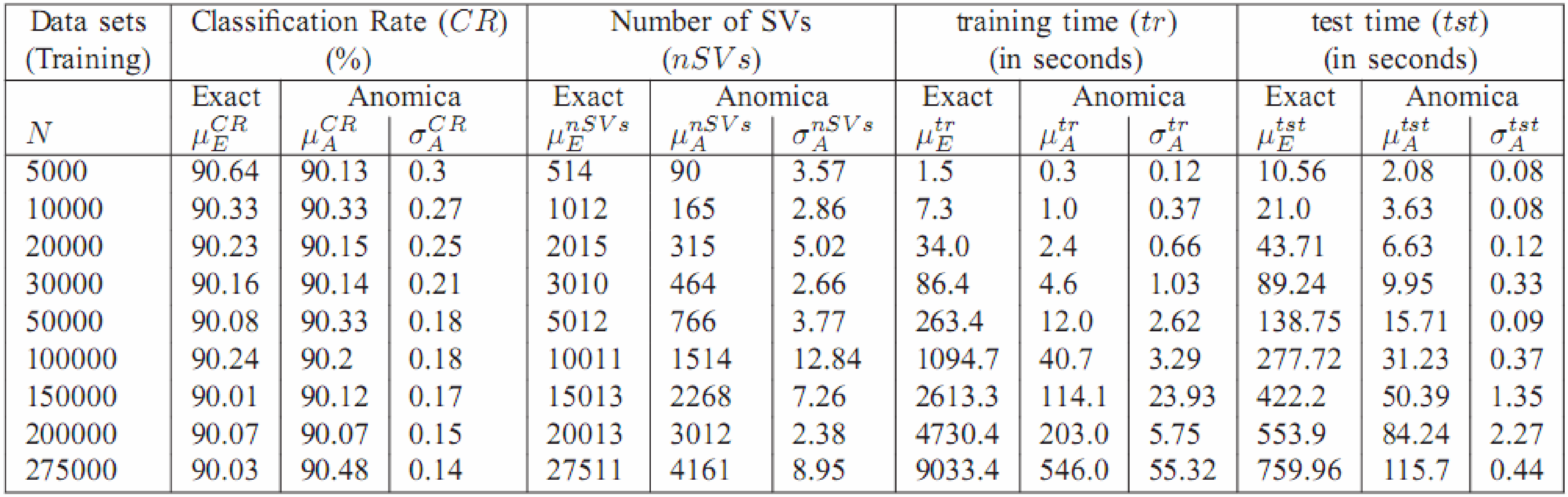
Data set-II

SDSS* (Sloan Digital Sky Survey) data

- Entire training set: 275k x 12
- Entire hold out set: 10k x 12
- Entire testing set: 130k x 12

* https://dashlink.arc.nasa.gov/data/sdss-dr6-data-for-photometric-redshift-calculations/

SDSS galaxy data set: results



Data sets (Training)	Classification Rate (CR) (%)			Number of SVs ($nSVs$)			training time (tr) (in seconds)			test time (tst) (in seconds)		
	Exact	Anomica		Exact	Anomica		Exact	Anomica		Exact	Anomica	
N	μ_E^{CR}	μ_A^{CR}	σ_A^{CR}	μ_E^{nSVs}	μ_A^{nSVs}	σ_A^{nSVs}	μ_E^{tr}	μ_A^{tr}	σ_A^{tr}	μ_E^{tst}	μ_A^{tst}	σ_A^{tst}
5000	90.64	90.13	0.3	514	90	3.57	1.5	0.3	0.12	10.56	2.08	0.08
10000	90.33	90.33	0.27	1012	165	2.86	7.3	1.0	0.37	21.0	3.63	0.08
20000	90.23	90.15	0.25	2015	315	5.02	34.0	2.4	0.66	43.71	6.63	0.12
30000	90.16	90.14	0.21	3010	464	2.66	86.4	4.6	1.03	89.24	9.95	0.33
50000	90.08	90.33	0.18	5012	766	3.77	263.4	12.0	2.62	138.75	15.71	0.09
100000	90.24	90.2	0.18	10011	1514	12.84	1094.7	40.7	3.29	277.72	31.23	0.37
150000	90.01	90.12	0.17	15013	2268	7.26	2613.3	114.1	23.93	422.2	50.39	1.35
200000	90.07	90.07	0.15	20013	3012	2.38	4730.4	203.0	5.75	553.9	84.24	2.27
275000	90.03	90.48	0.14	27511	4161	8.95	9033.4	546.0	55.32	759.96	115.7	0.44

With 275k training and 130k test instances, ν —Anomica is on an average ~15 times faster than classical SVMs.

Data set-III

Simulated CMAPSS* (Commercial Modular Aero-Propulsion System Simulation) data

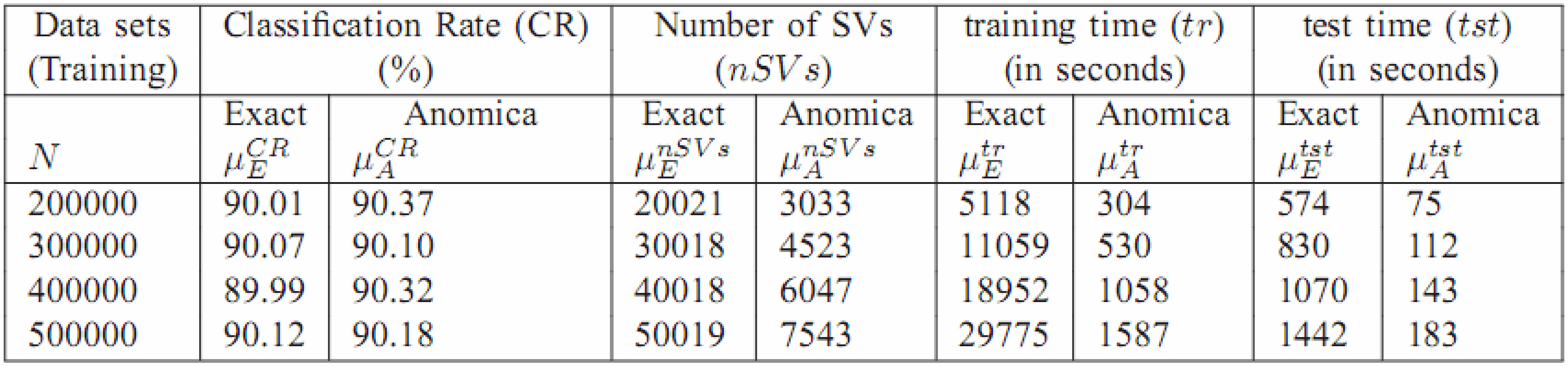
Entire training set: 500k x 29

Entire hold out set: 20k x 29

Entire testing set: 100k x 29

* https://dashlink.arc.nasa.gov/data/c-mapss-aircraft-engine-simulator-data/

CMAPSS data set: results



Data sets (Training)	Classification Rate (CR) (%)		Number of SVs ($nSVs$)		training time (tr) (in seconds)		test time (tst) (in seconds)	
N	Exact μ_E^{CR}	Anomica μ_A^{CR}	Exact μ_E^{nSVs}	Anomica μ_A^{nSVs}	Exact μ_E^{tr}	Anomica μ_A^{tr}	Exact μ_E^{tst}	Anomica μ_A^{tst}
200000	90.01	90.37	20021	3033	5118	304	574	75
300000	90.07	90.10	30018	4523	11059	530	830	112
400000	89.99	90.32	40018	6047	18952	1058	1070	143
500000	90.12	90.18	50019	7543	29775	1587	1442	183

With 500k training and 100k test instances, ν —Anomica is on average ~18 times faster than classical SVMs.

Highlights

- ν−Anomica runs in much reduced time in training and testing
  - while retaining accuracy
  - while maintaining similar ranking of anomalies
- With increasing training and test instances, the gain factor increases.
- Can be use for near real time detection
- Theoretical upper bound on the number of support vectors