Incremental Generalized Eigenvalue Classification on Data Streams

Mario R. Guarracino, Salvatore Cuciniello, Davide Feminiano

High Performance and Networking Institute
National Research Council, Italy

Overview

- ✓What is supervised learning?
- ✓Classification on data streams
- ✓GEPSVM: Generalized Eigenvalue Problem Support Vector Machine
- ✓Problems
- ✓Incremental classification
- ✓SI-ReGEC: Stream Incremental Reguralized Eigenvalue Classifier
- ✓Numerical results
- ✓Conclusions

What is supervised learning?

- ✓ Supervised learning refers to the capability of a system to learn from examples
- ✓ The trained system is able to provide an answer for each new question
- ✓ Supervised means the desired output for the training set is provided by an external teacher
- ✓ Binary classification is among the most successful methods for supervised learning

Classification on data streams

Many applications:

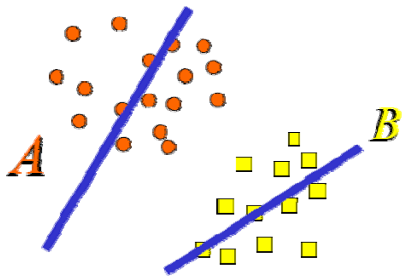
- ✓ Computer network traffic (spam, intrusion detection,…)
- ✓ Bank transactions (fraud, credit cards,…)
- ✓ Web search (link rating)
- ✓ Video/Audio sensors (video surveillance, face identification,…)

European Workshop on Data Stream Analysis, March 2007
Belvedere di San Leucio

GEPSVM:Generalized Eigenvalue Problem Support Vector Machine

Binary classification problem can be formulated as a generalized eigenvalue problem (GEPSVM).

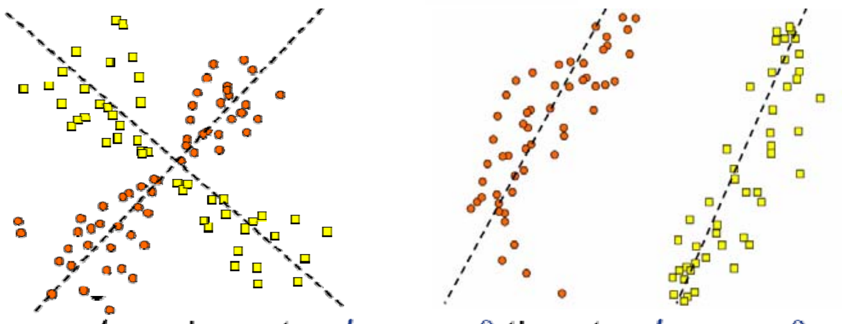
Find $x'w_1=\gamma_1$ the closest to A and the farthest from B



$$\min_{w,\gamma\neq 0} \frac{\|Aw-e\gamma\|^2}{\|Bw-e\gamma\|^2}$$

GEPSVM:Generalized Eigenvalue Problem Support Vector Machine

Let $[w_1\ \gamma_1]$ and $[w_m\ \gamma_m]$ be eigenvectors associated to min and max eigenvalues of $Gx=\lambda Hx$:



$a \in A \Leftrightarrow$ closer to $x'w_1 - \gamma_1 = 0$ than to $x'w_m - \gamma_m = 0$,

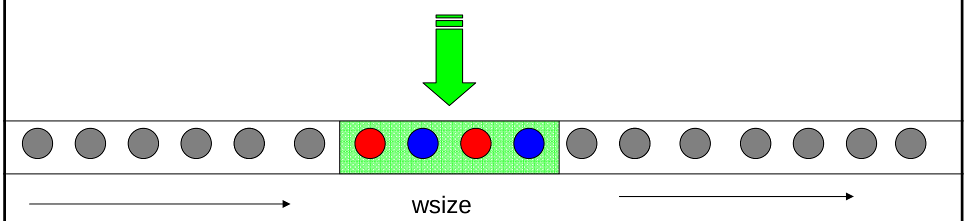
$b \in B \Leftrightarrow$ closer to $x'w_m - \gamma_m = 0$ than to $x'w_1 - \gamma_1 = 0$.

Problems

- ✓ Standard classification methods rely on the persistence of a complete training set
- ✓ Data not well represented by a persistent collection of items
- ✓ Data may be accessed but not completely loaded in main memory or stored

Incremental classification

A new approach consists in finding a small and robust subset of the training set while accessing data available in the window



When the window is full, all points within are processed by the classifier

European Workshop on Data Stream Analysis, March 2007
Belvedere di San Leucio

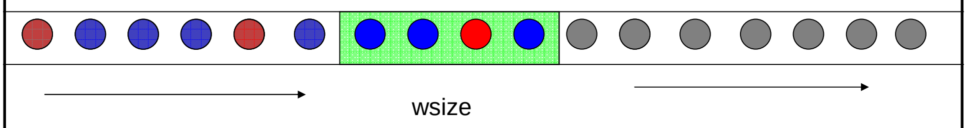
Incremental classification

It is possible to incrementaly train an algorithm using one point at time, analyzing its contribution of information.

Improvements:

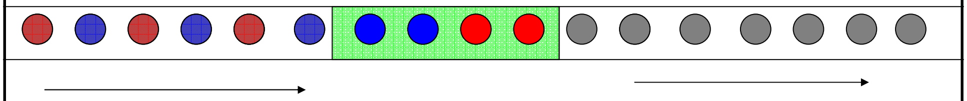
- ✓A smaller set of points reduces the probabilty of overfitting the problem
- ✓It is computationally more efficient in predicting new points
- ✓As new points become available in the window, their influence is evaluated with respect to the existing classifier

SI-ReGEC:Stream Incremental Regularized Eigenvalue Classifier



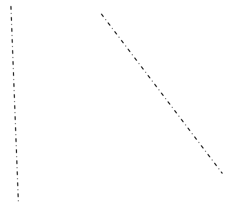
European Workshop on Data Stream Analysis, March 2007
Belvedere di San Leucio

SI-ReGEC:Stream Incremental Regularized Eigenvalue Classifier



New data

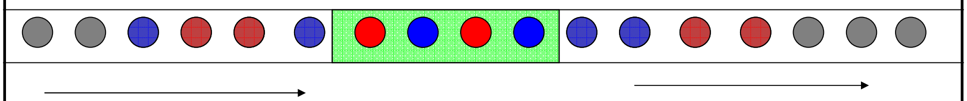
Old data



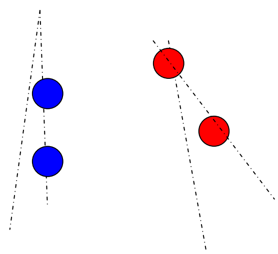
At each step, data in window are processed with the incremental learning classifier…

And hyperplanes are built

SI-ReGEC:Stream Incremental Regularized Eigenvalue Classifier

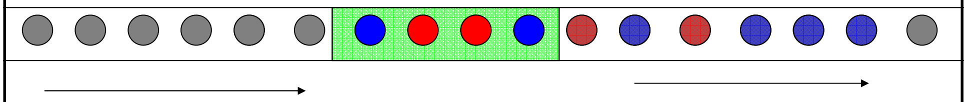


Step by step new points are processed



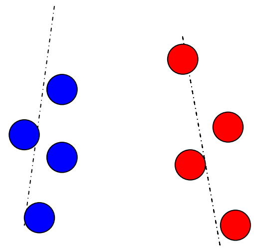
…and SI-ReGEC updates hyperplanes configuration

SI-ReGEC:Stream Incremental Regularized Eigenvalue Classifier

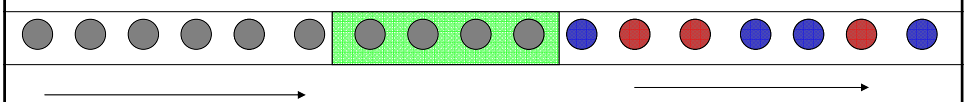


But not all points are considered…

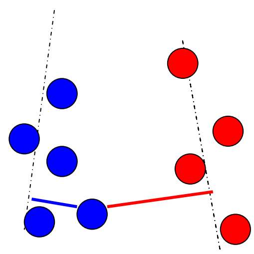
Some of them are discarted if their information contribution is useless



SI-ReGEC:Stream Incremental Regularized Eigenvalue Classifier



New unknown incoming points are classified by their distance from the hyperplanes



European Workshop on Data Stream Analysis, March 2007
Belvedere di San Leucio

SI-ReGEC:Stream Incremental Regularized Eigenvalue Classifier

An incremental learning technique based on GEPSVM that determines classification models based on a very small sample of data from the stream

$C = load(stream, wsize)$
$C_0 = kmean(C, km)$
$\Gamma_0 = C \setminus C_0$
$\{R_0, M_0\} = Classify(C, C_0)$
repeat
  $k = 1$
  while $|\Gamma_k| > 0$ **do**
    $x_k = x : \max_{x \in \{M_k \cap \Gamma_{k-1}\}} \{dist(x, P_{class(x)})\}$
    $\{R_k, M_k\} = Classify(C, \{C_{k-1} \cup \{x_k\}\})$
    if $R_k > R_{k-1}$ **then**
      $C_k = C_{k-1} \cup \{x_k\}$
    end if
    $\Gamma_k = \Gamma_{k-1} \setminus \{x_k\}$
    $k = k + 1$
  end while
  $C_0 = C_0 \cup C_k$
  $\Gamma_0 = load(stream, wsize)$
until $|\Gamma_0| = 0$

Numerical Results

Large-noisy-crossed-norm
Data set

200.000 points with 20 features
equal divided in 2 classes

100.000 train points

100.000 test points

Each class is drawn from a
multivariate normal distribution

Numerical Results

Parameter	B	ED	FP	EM	EM+E	SI-ReGEC
Error (%)	3.2	9.1	3.2	4.5	6.7	2.88
subset train	8321	4172	8452	1455	5308	413
window size	100000	500	500	500	500	500

SI-ReGEC has the lowest error and uses the smallest incremental set

B: Batch SVM
ED: Error-driven KNN
FP: Fixed partition SVM
EM: Exceeding-margin SVM
EM+E: Fixed margin + errors

Numerical Results

Larger windows lead to smaller train subset and execution time increases with window growth

Wsize	Acc.	Growth Rate	Avg. Time	Time
500	96.13%	15.75	4.25s	8.5e-4s
1000	96.92%	4.31	15.16s	1.5e-3s
2000	96.50%	2.63	61.79s	3.1e-3s
4000	97.45%	1.81	232.49s	7.0e-3s

Data is processed at 123.5 Gb/day on standard hardware

Conclusions

✓ SI-ReGEC:
1. achieves a classification accuracy well comparable with other methods
2. produces smaller incremental training sets

Future Work:
Investigate how to dinamically adapt window size to stream rate and nonstationary data streams

Thanks!