Accuracy of Distance Metric Learning Algorithms

Aurélien Sérandour

École Polytechnique
DMMT'09

June 28th, 2009

Table of contents

Motivation & problem

- recommendation problem in feature space

Motivation & problem

- recommendation problem in feature space
- use a simple Mahanalobis distance

Motivation & problem

- recommendation problem in feature space
- use a simple Mahanalobis distance
- real data are not perfect: try with imperfect ones

Chosen algorithms

Many algorithms $\longrightarrow$ only 4 chosen

Chosen algorithms

Many algorithms $\longrightarrow$ only 4 chosen

Euclidean	Xing	Coding Similarity	ITML
non iterative	iterative	non iterative	iterative
do nothing	minimize distance between similar pairs	shared information	respect thresholds
no data used	both	similar	both
	infinite loops		infinite loops

Experiments

- using 6 UCI datasets: Iris, Ionosphere, Wine, WDBC, Soybean-small, Balance-scale

Experiments

- using 6 UCI datasets: Iris, Ionosphere, Wine, WDBC, Soybean-small, Balance-scale
- datasets have properties: size, dimensionality, # of classes.

Experiments

- using 6 UCI datasets: Iris, Ionosphere, Wine, WDBC, Soybean-small, Balance-scale
- datasets have properties: size, dimensionality, # of classes.
- create similarity sets

Experiments

- using 6 UCI datasets: Iris, Ionosphere, Wine, WDBC, Soybean-small, Balance-scale
- datasets have properties: size, dimensionality, # of classes.
- create similarity sets
- lots of dissimilar data

Experiments

- using 6 UCI datasets: Iris, Ionosphere, Wine, WDBC, Soybean-small, Balance-scale
- datasets have properties: size, dimensionality, # of classes.
- create similarity sets
- lots of dissimilar data
- inject some errors because no perfect-real datasets $\rightarrow$ flipping similarities (noise)

Experiments

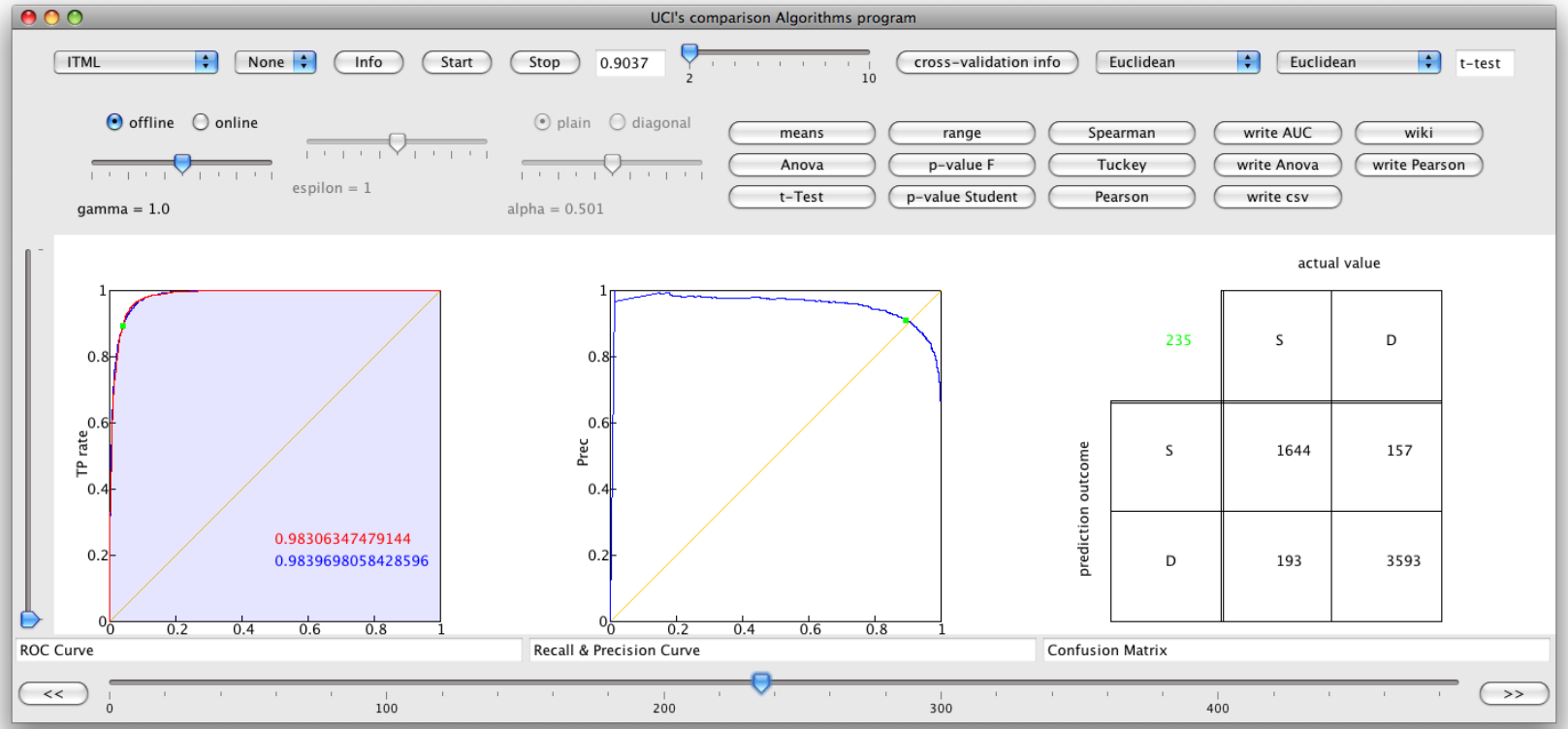

Figure: Java application

Results

noise	Iris	Ionosphere	Wine	WDBC	Soybean	Balance
0	ITML	Xing	CS	ITML	CS	ITML
5	ITML	Xing	CS	Xing	ITML	CS
10	ITML	CS	CS	Euclid.	ITML	CS
20	Euclid.	CS	Xing	Euclid.	ITML	CS
30	Euclid.	CS	Xing	Euclid.	Euclid.	CS

Table: Best algorithms

Conclusion

- there is no unique solution
- difficult to predict the behavior
- data drives the result
- use other kind of distances?

Thank you!