Introduction and Sharing: My Research in CSE

Hu Ding

Department of Computer Science & Engineering
State University of New York at Buffalo

October 14, 2014

cse@buffalo

Self Intruduction

- In 2009, I got my BS in *mathematics* from *Zhongshan University* in China, and start my PhD life in the same year.
- My advisor is Prof. Jinhui Xu.
- My research area: algorithms for machine learning and pattern recognition.

Research in Our Group

Quite broad, and hard to summarize

- Currently 6 PhD students in our group, and more than 10 graduated. Each one works on one or more individual topics.
- Generally, algorithmic aspect for any real world problem.

Research in Our Group

Quite broad, and hard to summarize

- Currently 6 PhD students in our group, and more than 10 graduated. Each one works on one or more individual topics.
- Generally, algorithmic aspect for any real world problem.
  1. **Data analytics**: clustering, regression, classification *et al.*
  2. **Computer vision**: image matching, pattern recognition, reconstruction *et al.*
  3. **Medical imaging**: Computed Tomagraphy (CT) reconstruction, segmentation *et al.*
  4. **Computational biology**: normal/cancer cells discrimination, pattern discovery for chromosome territories *et al.*
  5. **Fundamental topics in algorithms** (from 531): matching, max flow, shortest path, minimum cut *et al.*

Research in Our Group

Quite broad, and hard to summarize

- Currently 6 PhD students in our group, and more than 10 graduated. Each one works on one or more individual topics.
- Generally, algorithmic aspect for any real world problem.
  1. **Data analytics**: clustering, regression, classification *et al.*
  2. **Computer vision**: image matching, pattern recognition, reconstruction *et al.*
  3. **Medical imaging**: Computed Tomagraphy (CT) reconstruction, segmentation *et al.*
  4. **Computational biology**: normal/cancer cells discrimination, pattern discovery for chromosome territories *et al.*
  5. **Fundamental topics in algorithms** (from 531): matching, max flow, shortest path, minimum cut *et al.*
- We care both theory and practice:
  1. Time and space complexities, quality guarantee *et al.*
  2. Performance on real data.

Overview of My Research

Geometric Algorithms for Machine Learning and Pattern Recognition:

1. Constrained clustering in high dimensional space
2. Robust algorithms for classification and regression
3. Pattern matching, recognition, and retrieval

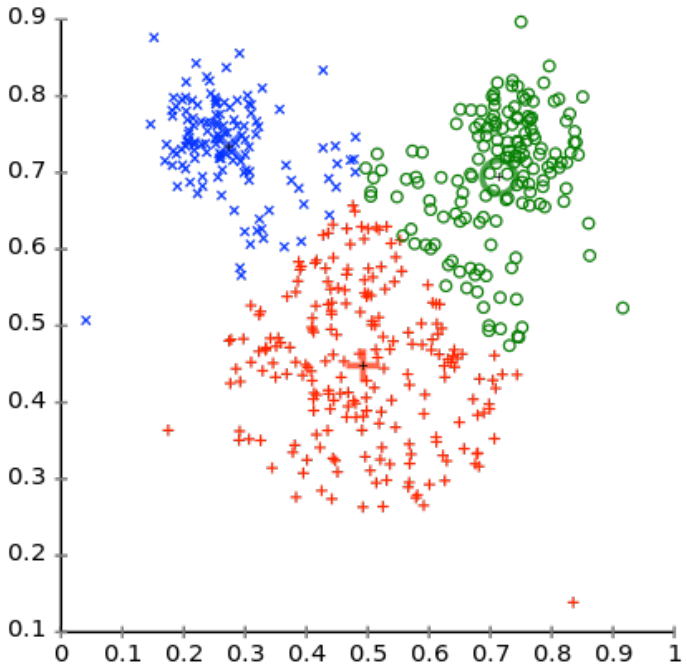
Overview of My Research

Geometric Algorithms for Machine Learning and Pattern Recognition:

1. ▷**Constrained clustering in high dimensional space**
2. Robust algorithms for classification and regression
3. Pattern matching, recognition, and retrieval

Constrained Clustering in High Dimensional Space

- Ordinary clustering assumes that all data items are independent from each other, and clustering is based only on distance or cost, *e.g.*, *k-means*, *k-medians*



Constrained Clustering in High Dimensional Space (cont.)

- Data items in real world applications are often correlated. Thus, clustering needs to consider both distance and some additional constraints, such as coloring, cluster size, etc.
  - l-Diversity clustering.
  - Chromatic clustering.
  - r-Gather clustering.
  - Capacitated clustering.
  - Semi-supervised clustering.
  - Uncertain data clustering.
- The additional constraints could complicate the problems considerably. No problem above has been solved satisfactorily.

Constrained Clustering in High Dimensional Space (cont.)

- Data items in real world applications are often correlated. Thus, clustering needs to consider both distance and some additional constraints, such as coloring, cluster size, etc.
  - l-Diversity clustering.
  - Chromatic clustering.
  - r-Gather clustering.
  - Capacitated clustering.
  - Semi-supervised clustering.
  - Uncertain data clustering.
- The additional constraints could complicate the problems considerably. No problem above has been solved satisfactorily.
- **Our result:** A unified framework (based on new geometric techniques) yielding good quality guarantees for all above constrained clustering problems in any dimensional space (accepted to **SODA'15**, one of the best conferences in algorithms).

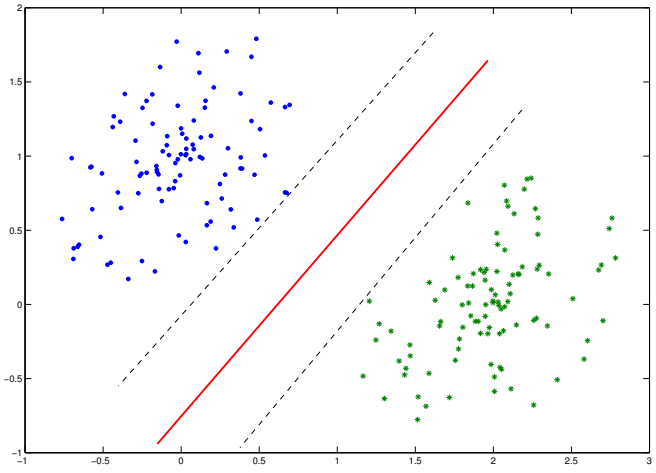
Overview of My Research

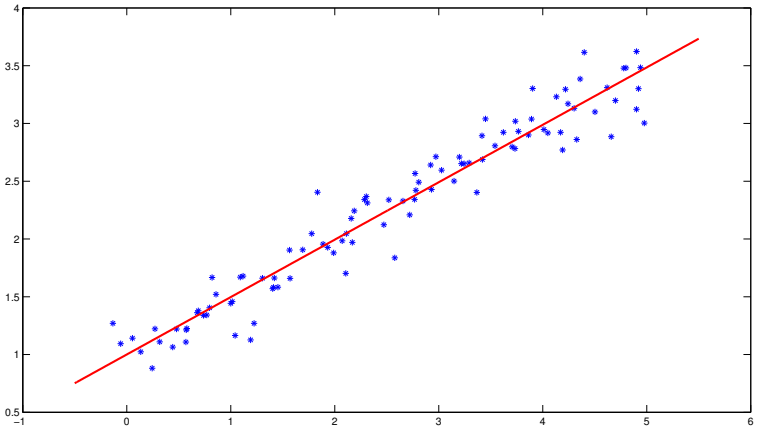
Geometric Algorithms for Machine Learning and Pattern Recognition:

1. Constrained clustering in high dimension
2. ▷**Robust algorithms for classification and regression**
3. Pattern matching, recognition, and retrieval

Robust Algorithms for Classification and Regression

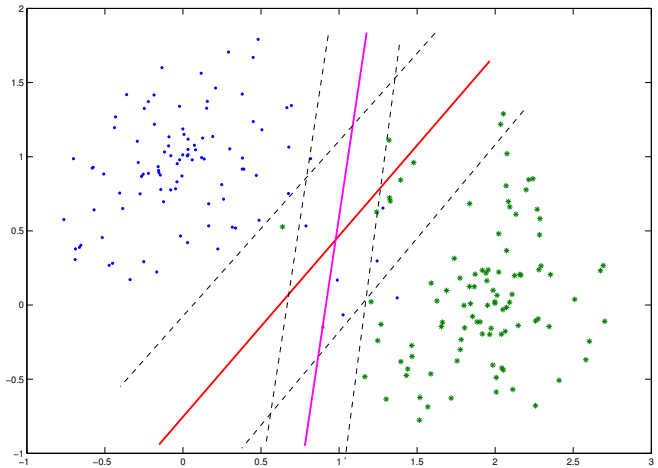
- **Support Vector Machine (SVM)** and **Linear Regression**:

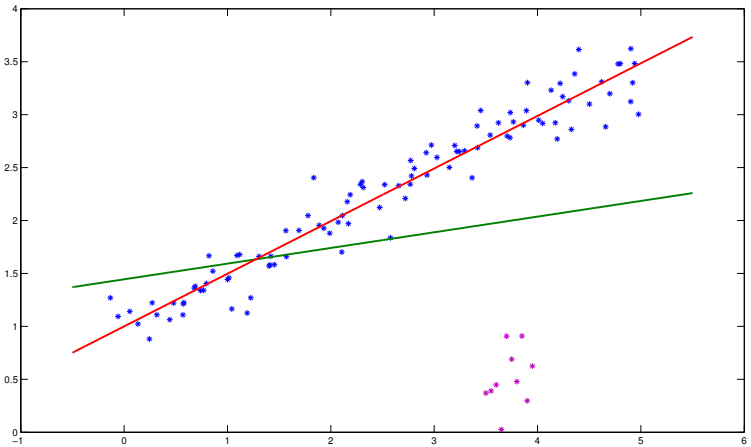




Robust Algorithms for Classification and Regression

- **Support Vector Machine (SVM)** and **Linear Regression**:
- Outliers could significantly deteriorate the solution:





Robust Algorithms for Classification and Regression

- **Support Vector Machine (SVM)** and **Linear Regression**:
- Outliers could significantly deteriorate the solution:
- Soft margin method and/or additional penalty term in the objective function do not work, if there is a considerable number of outliers.

Robust Algorithms for Classification and Regression

- **Support Vector Machine (SVM)** and **Linear Regression**:
- Outliers could significantly deteriorate the solution:
- Soft margin method and/or additional penalty term in the objective function do not work, if there is a considerable number of outliers.
- **Our results:** new combinatorial approaches for explicit outliers detection.
  - Theoretical guarantee on quality of solution.
  - Better performance in practice.

Overview of My Research

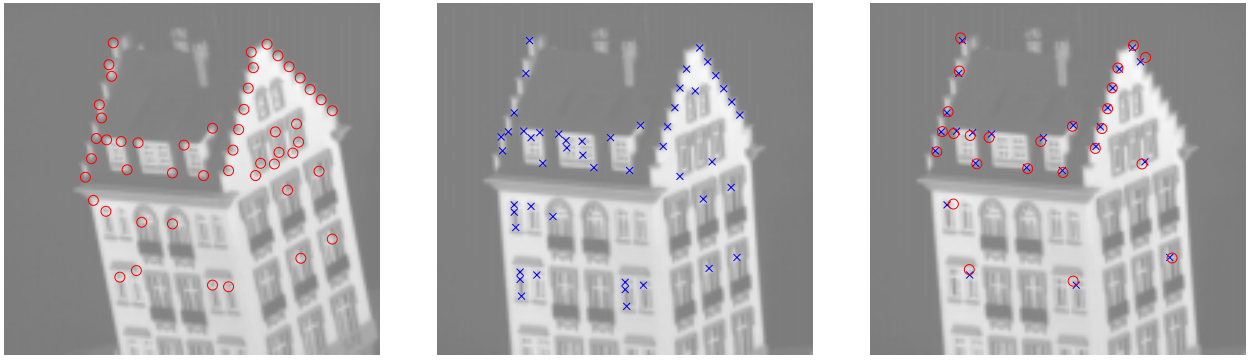
Geometric Algorithms for Machine Learning and Pattern Recognition:

1. Constrained clustering in high dimension
2. Robust algorithms for classification and regression
3. ▷**Pattern matching, recognition, and retrieval**

- **Point-sets matching** under *Earth Mover's Distance (EMD)*, which can be viewed as a min-cost max flow problem in the Euclidean space.

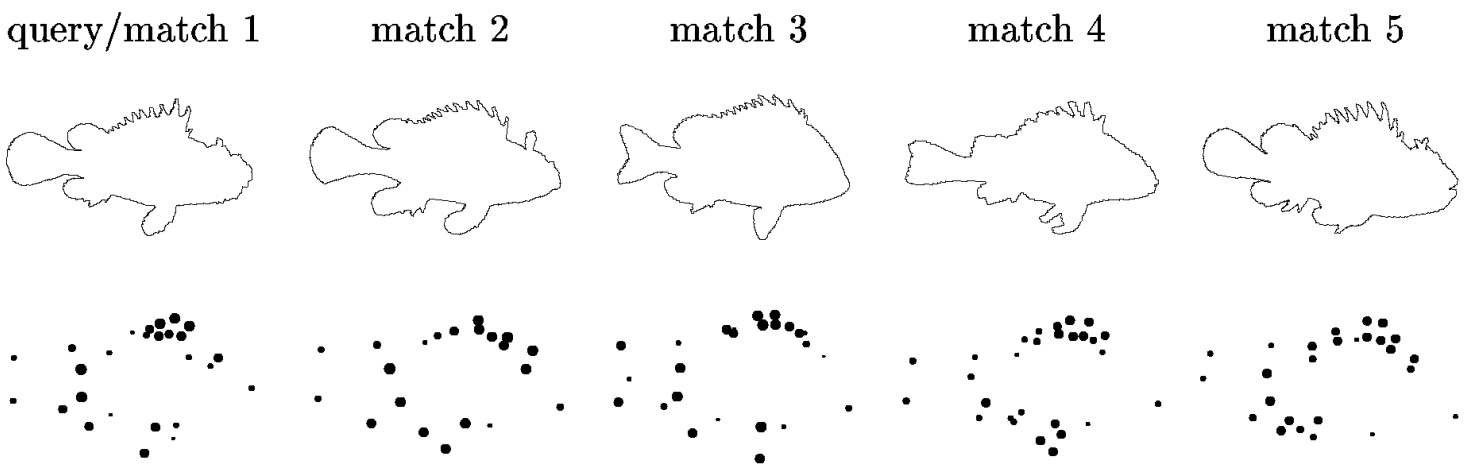
- **Point-sets matching** under *Earth Mover's Distance (EMD)*, which can be viewed as a min-cost max flow problem in the Euclidean space.
- EMD has been extensively studied in computer vision:

- **Point-sets matching** under *Earth Mover's Distance (EMD)*, which can be viewed as a min-cost max flow problem in the Euclidean space.
- EMD has been extensively studied in computer vision:
  - Registration [Cohen and Guibas, ICCV99]



Pattern Matching, Recognition, and Retrieval

- **Point-sets matching** under *Earth Mover's Distance (EMD)*, which can be viewed as a min-cost max flow problem in the Euclidean space.
- EMD has been extensively studied in computer vision:
  - Registration [Cohen and Guibas, ICCV99]
  - Pattern classification [Giannopoulos and Veltkamp, ECCV02]



Pattern Matching, Recognition, and Retrieval

- **Point-sets matching** under *Earth Mover's Distance (EMD)*, which can be viewed as a min-cost max flow problem in the Euclidean space.
- EMD has been extensively studied in computer vision:
  - Registration [Cohen and Guibas, ICCV99]
  - Pattern classification [Giannopoulos and Veltkamp, ECCV02]
  - Image retrieval [Rubner *et al.* IJCV00]

Pattern Matching, Recognition, and Retrieval (cont.)

Our results:

- The first FPTAS for minimizing EMD under certain transformation in any fixed dimensional space.

Pattern Matching, Recognition, and Retrieval (cont.)

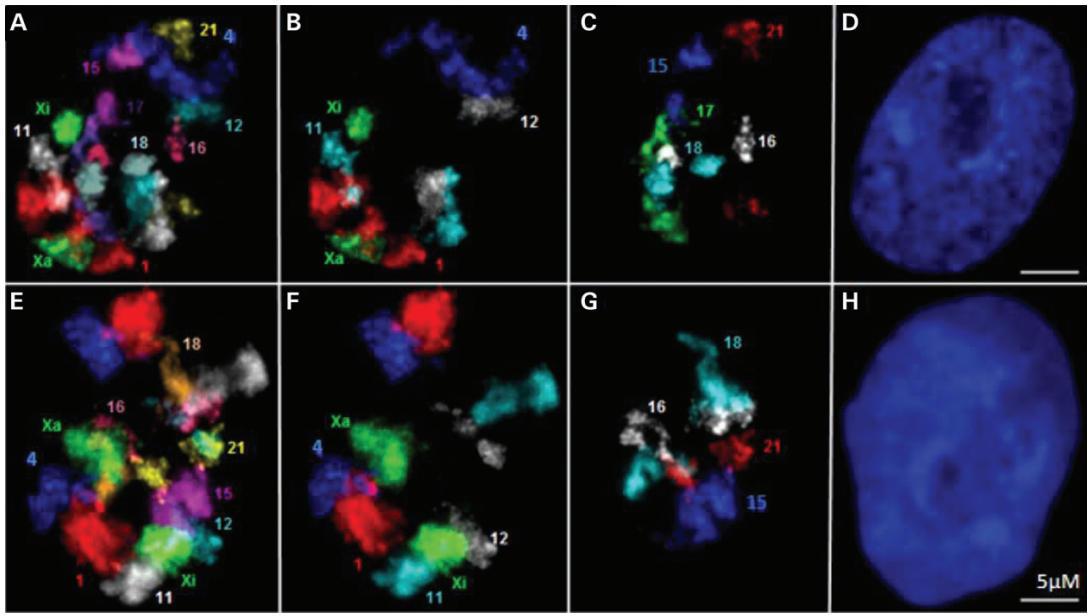
Our results:

- The first FPTAS for minimizing EMD under certain transformation in any fixed dimensional space.
- **Prototype learning** algorithms for association graphs, rigid structures, and affine deformable structures
  - Based on geometric techniques and EMD
  - Avoiding encoding and decoding between geometry and graph domains.
  - More robust and efficient for pattern recognition and retrieval.

Pattern Matching, Recognition, and Retrieval (cont.)

Our results (cont.):

- Applications of prototype learning algorithms: Extracting inter-chromosomal association and chromosome topological patterns from a population of cells:
  - Determine the difference in normal and cancer cells
  - Reveal the dynamics of the association pattern during cancer progression.
- Published in **CVPR'13** and **Plos Computational Biology**.



My Feelings and Advices

- **What I really like:** solving a real world problem
- **What's the challenging:** finding a really good research topic.
- **My suggestions:**
  - Open mind.
  - Intuition is much more important than mathematics.

Thank you!

Question?