Top 10 Algorithms in Data Mining

Xindong Wu (吴信东)
Department of Computer Science
University of Vermont, USA;
合肥工业大学计算机与信息学院
“Top 10 Algorithms in Data Mining”
by the IEEE ICDM Conference

1. The 3-step identification process
2. 18 identified candidates in 10 data mining topics
3. The top 10 algorithms
4. Follow-up actions
1. **Nominations.** ACM KDD Innovation Award and IEEE ICDM Research Contributions Award winners were invited in September 2006 for nominations.

   Each nomination was asked to come with the following information:
   
   a) the algorithm name
   
   b) a brief justification
   
   c) a representative publication reference

   Up to 10 nominations from each nominator

   The nominations as a group should have a reasonable representation of the different areas in data mining.

   All except one in this distinguished set of award winners responded.
The 3-Step Identification Process (2)

2. **Verification.** Each nomination was verified for its citations on Google Scholar in late October 2006, and those nominations that did not have at least 50 citations were removed. 18 nominations survived and were then organized in 10 topics.

3. **Voting** by the wider community.
   - (a) Program Committee members of KDD-06, ICDM '06, and SDM '06 and
   - (b) ACM KDD Innovation Award and IEEE ICDM Research Contributions Award winners
   - The top 10 algorithms are ranked by their number of votes, and when there is a tie, the alphabetic order is used.
Agenda

1. The 3-step identification process
2. 18 identified candidates (in 10 data mining topics)
3. The top 10 algorithms
4. Follow-up actions
18 Identified Candidates

Classification

Statistical Learning

Association Analysis

Link Mining

Top 10 Algorithms in Data Mining: Xindong Wu and Vipin Kumar
18 Candidates (2)

Clustering

Bagging and Boosting

Sequential Patterns
- #15. **PrefixSpan**: J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M-C. Hsu. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In ICDE ’01.

Integrated Mining
- #16. **CBA**: Liu, B., Hsu, W. and Ma, Y. M. Integrating classification and association rule mining. KDD-98.

Rough Sets

Graph Mining
- #18. **gSpan**: Yan, X. and Han, J. 2002. gSpan: Graph-Based Substructure Pattern Mining. In ICDM ‘02. (Google Scholar Citations: October 2006: 155; 1/10/18: 2163)
Agenda

1. The 3-step identification process
2. 18 identified candidates
3. The top 10 algorithms
4. Follow-up actions
The Top 10 Algorithms

#1: C4.5, presented by Hiroshi Motoda
#2: K-Means, presented by Joydeep Ghosh
#3: SVM, presented by Qiang Yang
#4: Apriori, presented by Christos Faloutsos
#5: EM, presented by Joydeep Ghosh
#6: PageRank, presented by Christos Faloutsos
#7: AdaBoost, presented by Zhi-Hua Zhou
#7: kNN, presented by Vipin Kumar
#7: Naive Bayes, presented by Qiang Yang
#10: CART, presented by Dan Steinberg
Agenda

1. The 3-step identification process
2. 18 identified candidates
3. The top 10 algorithms
4. Follow-up actions
Top 3 Algorithms:
- C4.5
- SVM
- Apriori

Top 10 Algorithms
- The top 10 algorithms voted from the 18 candidates at the panel are the same as the voting results from the 3-step identification process.
Follow-Up Actions

A survey paper on Top 10 Algorithms in Data Mining (X. Wu, V. Kumar, J.R. Quinlan, et al., *Knowledge and Information Systems*, 14(1), 2008, 1~37)
- Written by the original authors and presenters
- Cited 3529 times on Google Scholar as of 1/10/2018

How to make a good use of these top 10 algorithms?
- Curriculum development
- 2nd Edition in 2018 ??

Various questions on these 10 algorithms?
- Why not this algorithm or that topic?
- New promising algorithms: Random Forests and Gradient Boosting

Will the votes change in the future?
- Sure, let’s work together to make positive changes!

Top 10 Algorithms in Data Mining: Xindong Wu and Vipin Kumar