ICDM ’06 Panel on “Top 10 Algorithms in Data Mining”

1. The 3-step identification process
2. The 18 identified candidates
3. Algorithm presentations
4. Top 10 algorithms: summary
5. Open discussions

The 3-Step Identification Process

1. **Nominations.** ACM KDD Innovation Award and IEEE ICDM Research Contributions Award winners were invited in September 2006 to each nominate up to 10 best-known algorithms.
    - All except one in this distinguished set of award winners responded.
    - Each nomination was asked to come with the following information: (a) the algorithm name, (b) a brief justification, and (c) a representative publication reference.
    - Each nominated algorithm should have been widely cited and used by other researchers in the field, and the nominations from each nominator as a group should have a reasonable representation of the different areas in data mining.

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 were invited to each vote for up to 10 well-known algorithms.
   - 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. The 18 identified candidates
3. Algorithm presentations
4. Top 10 algorithms: summary
5. Open discussions

The 18 Identified Candidates

- Classification
  - #1. C4.5: Quinlan, J. R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc.
  - #2. CART: L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, Belmont, CA, 1984.
  - #3. K Nearest Neighbours (kNN): Hastie, T. and Tibshirani, R. 1996. Discriminant Adaptive Nearest Neighbor Classification. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI). 18, 6 (Jun. 1996), 607-616.
  - #4. Naive Bayes Hand, D.J., Yu, K., 2001. Idiot's Bayes: Not So Stupid After All? Internat. Statist. Rev. 69, 385-398.
- Statistical Learning
  - #5. SVM: Vapnik, V. N. 1995. The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc.
  - #6. EM: McLachlan, G. and Peel, D. (2000). Finite Mixture Models. J. Wiley, New York.
- Association Analysis
  - #7. Apriori: Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In VLDB '94.
  - #8. FP-Tree: Han, J., Pei, J., and Yin, Y. 2000. Mining frequent patterns without candidate generation. In SIGMOD '00.
- Link Mining
  - #9. PageRank: Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In WWW-7, 1998.
  - #10. HITS: Kleinberg, J. M. 1998. Authoritative sources in a hyperlinked environment. In Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, 1998.

18 Candidates (2)

- Clustering
  - #11. K-Means: MacQueen, J. B., Some methods for classification and analysis of multivariate observations, in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, 1967.
  - #12. BIRCH Zhang, T., Ramakrishnan, R., and Livny, M. 1996. BIRCH: an efficient data clustering method for very large databases. In SIGMOD '96.
- Bagging and Boosting
  - #13. AdaBoost: Freund, Y. and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 1 (Aug. 1997), 119-139.
- Sequential Patterns
  - #14. GSP: Srikant, R. and Agrawal, R. 1996. Mining Sequential Patterns: Generalizations and Performance Improvements. In Proceedings of the 5th International Conference on Extending Database Technology, 1996.
  - #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
  - #17. Finding reduct: Zdzislaw Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Norwell, MA, 1992
- Graph Mining
  - #18. gSpan: Yan, X. and Han, J. 2002. gSpan: Graph-Based Substructure Pattern Mining. In ICDM '02.

Agenda

1. The 3-step identification process
2. The 18 identified candidates
3. Algorithm presentations
4. Top 10 algorithms: summary
5. Open discussions

Algorithm Presentations

- Each algorithm presentation provides
  a) a description of the algorithm,
  b) the impact of the algorithm, and
  c) current and further research on the algorithm
- Each presenter will introduce himself
  – Is an experienced researcher with the algorithm
  – Uses the original authors' slides if available, with possible modifications
  – Provides his own insights on the identified algorithm

Agenda

1. The 3-step identification process
2. The 18 identified candidates
3. Algorithm presentations
4. Top 10 algorithms: summary
5. Open discussions

Top 10 Algorithms: Summary

- **#1: C4.5** (61 votes), presented by Hiroshi Motoda
- **#2: K-Means** (60 votes), presented by Joydeep Ghosh
- **#3: SVM** (58 votes), presented by Qiang Yang
- **#4: Apriori** (52 votes), presented by Christos Faloutsos
- **#5: EM** (48 votes), presented by Joydeep Ghosh
- **#6: PageRank** (46 votes), presented by Christos Faloutsos
- **#7: AdaBoost** (45 votes), presented by Zhi-Hua Zhou
- **#7: kNN** (45 votes), presented by Vipin Kumar
- **#7: Naive Bayes** (45 votes), presented by Qiang Yang
- **#10: CART** (34 votes), presented by Dan Steinberg

Agenda

1. **The 3-step identification process**
2. **The 18 identified candidates**
3. **Algorithm presentations**
4. **Top 10 algorithms: summary**
5. **Open discussions**

Open Discussions

- A survey paper is being generated by the original authors and presenters.
- How to make a good use of these top 10 algorithms?
- Is there a need for generating a book out of them?
- Any particular questions on any of these 10 algorithms
- …

Open Votes for Top Algorithms

- Top 3 Algorithms:
  - C4.5: 52 votes
  - SVM: 50 votes
  - Apriori: 33 votes

- 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.