

Promoting DM Research or Technology?

Jeffrey Xu Yu
Chinese University of Hong Kong
yu@se.cuhk.edu.hk

Promote DM: Enough?

- ❑ Books.
- ❑ Conferences, workshops, symposiums, and journals.
- ❑ Commercial and free data mining software.
- ❑ KDDCup competition.
- ❑ Datasets.
- ❑ Data Mining Markup Language.

Promote DM: Books

- ❑ Searching www.amazon.com using "data mining" returns **12,386** results.
- ❑ The top three books on the list are
 - Data Mining: Practical Machine Learning Tools and Techniques (2nd Edition) by Ian H. Witten and Eibe, 2005.
 - Data Mining - Introduction to Data Mining, by Pang-Ning Tan, Michael Steinbach, and *Vipin Kumar*, 2005.
 - Data Mining: Concepts and Techniques (2nd Edition), by Jiawei Han and Micheline Kamber, 2005.

Promote DM: More Books

- ❑ Advances in Knowledge Discovery and Data Mining by Usama M. Fayyad, *Gregory Piatetsky-Shapiro*, Padhraic Smyth, and Ramasamy Uthurusamy, 1996.
- ❑ Intelligent Technologies for Information Analysis by *Ning Zhong* and *Jiming Liu*, 2004.
- ❑ Introduction to Business Data Mining by David L Olson and *Yong Shi*, 2005.
- ❑ Next Generation of Data Mining by *Hillol Kargupta*, Jiawei Han, *Philip S. Yu*, and Rajeev Motwani, 2008.
- ❑ Mobility, Data Mining and Privacy: Geographic Knowledge Discovery by *Fosca Giannotti* and *Dino Pedreschi*, 2008.
- ❑ The Top Ten Algorithms in Data Mining by *Xindong Wu* and *Vipin Kumar*, 2009.

What applications can DM support?

- ❑ I taught a data mining course in an Executive Master of Science in Logistics and Supply Chain Management (CUHK&Tsinghua).
 - Show me a case, and show me a Harvard case study.
 - Can I do it for L&SCM?
- ❑ Can DM support any applications?
 - The answer is Yes!
 - Is there something concrete to show managers about it?
 - There is ER model in DB.

More DM techniques Better?

- ❑ Need more DM techniques to solve different problems.
- ❑ But, more techniques may sometime confuse users, and make 'us' difficult to convince people.
- ❑ Let's recall the discussions we had in the forum.
 - Starting from association rules mining (min support and min confidence)
 - Charu Aggarwal & Philip Yu criticizes support and confidence (PODS'98).

Criticism to Support and Confidence

□ Example:

- Among 5000 students
 - 3000 play basketball
 - 3750 eat cereal
 - 2000 both play basket ball and eat cereal
- *play basketball* \Rightarrow *eat cereal* [40%, 66.7%] is misleading because the overall percentage of students eating cereal is 75% which is higher than 66.7%.
- *play basketball* \Rightarrow *not eat cereal* [20%, 33.3%] is far more accurate, although with lower support and confidence

	basketball	not basketball	sum(row)
cereal	2000	1750	3750
not cereal	1000	250	1250
sum(col.)	3000	2000	5000

New measures: Lift?

Drawback of Lift

- In Vipin Kumar's book, it says to be careful.

	y	\bar{y}	
X	10	0	10
\bar{X}	0	90	90
	10	90	100

$$Lift = \frac{0.1}{(0.1)(0.1)} = 10$$

	y	\bar{y}	
X	90	0	90
\bar{X}	0	10	10
	90	10	100

$$Lift = \frac{0.9}{(0.9)(0.9)} = 1.11$$

Statistical independence: If $P(X,Y)=P(X)P(Y) \Rightarrow Lift = 1$

More to come on Association Rules

- There are issues on
 - Spurious patterns (Geoff Webb)
 - Too many rules (Andrew Wong)
- *Rao made a comment to use support/confident like technique after Ee-Ping's talk.*

What should we tell people?

□ Typical Answers:

- There are many cases. There is no single one which is the best for all.
- Tell me your problem, and I will try to find a solution for you.

What is DM?

- ❑ Is DM art or science? (Hillol Kargupta)
- ❑ Production Intelligence (Andrew Wong)
- ❑ Autonomic Computing (Michele Sebag)
 - Which algorithm/system is best suited to MY problem. Just take the best one! No free lunch theorem.
 - Filter out bad parameter settings.
 - Meta learning.
- ❑ Can we do DM like google?
- ❑ Can we do DM without telling people what DM is about?

News Sensitive Investment

- Recently, Reuters, Thomson Financial and Dow Jones provides news mining technology to investment bank and hedge fund.
- The investment banks, such as Paribas Investment Partners, Citigroup, uses this technology for automated trading.

Final Question

- What are the best ways to convince people to use DM?