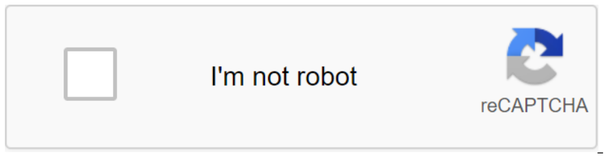

Continue

Data mining pdf notes

Wednesday, 5/17: H. Mannila, H. Toivonen, and A.I. Verkamo, opening frequent episodes in sequence. First International Conference on Knowledge Discovery and Data Analysis, page 210 - 215, AAAI Press, 1995. Postscript. Monday, 5/15: Christos Faluzos, M. Ranganatan and Yannis Manolopoulos, 'Fast Subsector Of Compliance in Time Row Databases,' SIGMOD, 1994, p. 419-429. Pdf. Wednesday, 5/10: S. Guha, R. Rastogi, and K. Shim, 'CURE: Effective Clustering Algorithm for Large Databases,' SIGMOD 1998. Pdf. Note: This PDF file requires a huge tempo space (over 200MB). Monday, 5/8: Venkatesh Santi, Ragu Ramakrishnan, Johannes Gehrke, Allison L. Powell and James K. French: Clustering large data sets in arbitrary metric spaces, ICDE, 502-511, 1999. Wednesday, 5/3: Christos Faloutsos and King-Ip (David) Lin, 'FastMap: Fast Algorithm for Indexing, Intelligent Data Analysis and Visualization of Traditional and Multimedia Data Sets,' ACM SIGMOD, May 1995, San Jose, Cay., page 163-174. Gzipped postscript. Wednesday, 4/26:. Bradley, U. Fayyad, and K. Reina, Scaling clustering algorithms for large databases,'1998 KDD. Postscript. Monday, 4/24: S. Breen, Extracting patterns and relationships from the world wide web. Postscript. Wednesday, 4/19: a) J. Kleinberg, Authoritative Sources in a Hyper-Connected Environment, J. ACM September, 1999, page 604-632. Pdf. b) S. Breen and L. Page, 'Dynamic Data Mining.' Postscript. Monday, 4/17: S. Breen and L. Page, Anatomy of a large-scale hypertextual search engine, WWW7/Computer Networks (1-7), 1998, page 107-117. Postscript. Wednesday, 4/12: D. Tsur et al., The Query Flocks: Generalization of the Mining Industry Rule Association, 1998 SIGMOD. Postscript. Monday, 4/10: E. Cohen et al., In Search of Interesting Associations Without Backing Pruning, ICDE 2000. Postscript. Wednesday, 4/5: a) M. Fang, N. Shivakumar, H. Garcia-Molina, R. Motwani, and J. Ullman, The Calculation of Iceberg Requests Effectively, 1998 VLDB. Postscript. b) H. Toivonen, Large Database for Association Rules, VLDB 1996, p. 134-145. Postscript. Monday, 4/3: J. S. Park, M.-S. Chen, and P. S. Yu, Effective hash algorithm for mountain association rules,' 1995 SIGMOD, page 175--186. PDF Wednesday, 3/29: a) R. Agrawal, T. Imielsky, A. Swami: Mountain Associations between sets of elements in massive databases, Prok. ACM SIGMOD Int'l Data Management Conference, Washington, D.C., May 1993, 207-216. Postscript. Pdf. b) R. Agrawal, R. Srikant: Fast Algorithms of the Rules of the Mountain Association, Prok. 20th Int'l Conference on Very Large Databases, Santiago, Chile, September 1994. Postscript. Pdf. LEC - Topics Data Sources 1 Data Review (PDF) Forecast and Classification with k-Nearest Neighbors Example 1: Riding Mower (PDF) Table 11.1 from page 584: Richard, and Dean Wichern. Applies Applies Statistical analysis. 5th. Prentice Hall, 2002. ISBN: 0-13-092553-5. 2 Classification and Bayes Rule, Naiv Bayes (PDF) 3 Tree Classification (PDF) Housing Database (Boston). Publicly available data from the University of California, the Irvine School of Information and Computer Science, machine learning databases. 4 Example of Disciplinary Analysis 2: Data from Fisher Iris (PDF) Iris Plant Database. Publicly available data from the University of California, the Irvine School of Information and Computer Science, machine learning databases. 5 Logistical Regression Case (PDF) Handlooms (PDF) 6 Neural Networks (PDF) 7 Homework Discussion - see Problem 1 in The Assignments section 8 Multiple Regression Review (PDF) 9 Multiple linear regressions in data extraction (PDF) 10 Tree Regressions, Case: IBM/GM Weekly Returns Comparison of Data Mining Methods (PDF) Homework Discussion - see Problem 2 in the Assignments section 11 k-Means Clustering, Hierarchical Clustering (PDF) 12 Case: Retail Merchandising 13 Medium-Term Examination 14 Key Components (PDF) Sample 1, Main Dimensions of Adult Sons: Rencher, Alvin. Multivariate analysis methods. 2nd o. Wylie-Internauka, 2002. Table 3.7, page 79. ISBN: 0-471-46172-5. Example 2, Charactersitics Wine: Wine Recognition Database. Publicly available data from the University of California, the Irvine School of Information and Computer Science, machine learning databases. 15 Guest Lecture by Dr. Ira Haimowitz: Data Mining and CRM at Pfizer 16 Association Rules (MARKET Basket Analysis) (PDF) Khan, Jiawei and Michelin Cumber. Data mining: concepts and methods. Morgan Kaufman Publishers, 2001. Example 6.1 (Figure 6.2). ISBN: 1-55860-489-8. 17 Recommendation Systems: Joint Filtering 18 Guest Lecture by Dr. John Elder IV, Senior Research: Practice Data Analysis Page 2 Introduction to Data Analysis Data Analysis Issues Pre-Processing Classification Data, Part 1 Classification, Part 2 Lecture Notes (MDL) Classification, Part 3 Classification, Part 4 Association Analysis, Part 1 Association Analysis, Part 2 Association Analysis, Part 2 Part 1 Clustering, Part 2 Clustering, Part 3 Clustering, Part 4 Clustering, Part 5 Visualization Introduction - SVM Document on SVM Detection of Cancer Detection Anomalies - WebMining App This lesson is a brief introduction in data analysis (which is also sometimes called It Adapted From Module 1: Introduction, Machine Learning and Data Analysis Course. 1.1 Data Flooding Data Current Trends Are Inexorably Lead to Data Flooding. telecommunications and other business operations. Additional data are generated from scientific experiments in astronomy, space exploration, biology, physics Additional data are created on the especially in text, image and other multimedia formats. For example, Europe's very long base interferometry (VLBI) has 16 telescopes, each producing 1 gigabit per second (yes, per second!) of astronomical data during a 25-day observation session. It really generates an astronomical amount of data. ATT processes so many calls a day that it can't store all the data - and data analysis needs to be done on the fly. Analysis by UC Berkeley professors. Peter Liman and Hal R. Varian calculated that 5 exabytes (5 million terabytes) of new data were created in 2002. Twice as much information was received in 2002 than in 1999 (30% growth rate). The U.S. produces 40% of the world's new stored data. According to the Winter Corp. Survey, according to winter Corp. Survey, according to the Winter Corp. Survey, France Telecom has the largest DB for decision-making, 30 TB (terabytes); ATT was in second place with 26 TB databases. Some of the largest databases on the Internet, as of 2003, include the Alexa (www.alexa.com) Internet Archive: 7 years of data, 500 TB Internet Archive (www.archive.org), 300 TB Google, over 4 billion pages, many, many TB data volumes are growing very quickly, and very few of them will ever be considered by man. Discovering knowledge is necessary to make sense and use data. 1.2 Examples of the application of mining data in the Region, which recently used data analysis, include: Scientific astronomy, bioinformatics, drug discovery, ... Business advertising, customer modeling and CRM (Customer Relations Management) e-commerce, detection of healthcare fraud, ... investment, production, sports/entertainment, telecommunications (phone and communication), target marketing, Internet: search engines, bots, ... Government counterterrorism efforts (we will discuss privacy disputes later) law enforcement, profiling tax scams One of the most important and widespread business applications for data collection is customer modeling, also called predictive analytics. This includes tasks such as predicting exhaustion or churn, i.e. finding which customers are likely to stop servicing targeted marketing: acquiring customers - finding what prospects can become cross-selling customers - for a given customer and product, finding which other product (s) they can buy credit risk - to determine the risk that that customer will not pay off credit or credit card fraud detection - is it a fraudulent transaction? The largest users of customer analytics are industries such as banking, telecommunications, retailers, where businesses with a large number of customers are widely these technologies. 1.2.1 Customer Service Level: Let's look at the example of a mobile phone company. The typical level of exhaustion (also called churn) rate on mobile phone customers is about 25-30% per year! The task is given customer information for the last N (N can vary from 2 to 18 months), predict who is likely to attrite in the next month or two. Also, rate the customer's customer and what a cost-effective offer to be made for that customer. Verizon Wireless is the largest wireless service provider in the United States with a customer base of 34.6 million subscribers by . Verizon has built a data repository of customers who have identified potential attriters developed several regional models Target customers with a high propensity to accept the offer of a reduction in the rate of exhaustion from more than 2%/month to less than 1.5%/month (huge impact of more than 34 million subscribers) 1.2.2 Credit Risk Assessment : Case Study Let's consider a situation where a person is applying for a loan. Should the bank approve the loan? Note: People who have the best credit don't need loans, and people with the worst credit are more likely not to repay. The bank's best customers are in the middle. Banks develop credit models using different machine learning methods. Mortgages and the proliferation of credit cards are the result of being able to successfully predict if a person is most likely to default on a loan. Credit risk assessment is widely used in the United States and is widely used in most developed countries. 1.2.3 Successful e-commerce - Case Study Amazon.com is the largest online retailer that has started with books and expanded in music, electronics and other products. Amazon.com has an active data analysis team that focuses on personalization. Why personalization? Consider the person who buys a book (product) in Amazon.com. The challenge: Recommend other books (and perhaps products) this person is more likely to buy Amazon's initial and fairly successful effort using clustering-based books bought. For example, customers who bought Advances in Knowledge Discovery and Data Mining from Fayada, Piatecki-Shapiro, Smith and Uturus also bought Data Mining: Practical Machine Learning Tools and Java Implementation Techniques by Witten and Abe. The recommendation project is quite successful, more advanced programs are being developed. 1.2.4 Unsuccessful e-commerce - Example (KDD Cup 2000) Of course, the application of data analysis is not a guarantee of success and during the Internet bubble of 1999-2000 we saw many examples. Consider the legwear and legcare e-tailer Gazelle.com whose link and purchase data were the subject of the KDD Cup 2000 (One of the questions was: The cost of visitors who spend more than $12 on an average order on the data website included a dataset of 3,465 purchases, 1831 customer very interesting and illuminating analysis was done by dozens of participants. The total time was thousands of hours, which would have been equivalent to millions of dollars in consulting fees. However, total sales Gazelle.com only a few thousand dollars, and no mining can't help them. No wonder Gazelle.com came out of August 2000. 1.2.5 Genomic Microarrays - Case Study DNA Microarrays is a revolutionary new technology that measures gene expression levels for many thousands of genes simultaneously (more on Microarrays later). Microarrays have recently become a popular area of application for data collection (see, for example, SIGKDD Research Special Issue on Microarray Data Mining , December 2003 (Vol. 5, Issue 2) www.acm.org/sigkdd/explorations/) One of the typical problems given microarrhe data for a number of patients (samples), can we accurately diagnose the disease? Predict the outcome of this treatment? Recommend the best treatment? Consider the Go99 leukemia dataset with 72 samples and about 7,000 genes. The samples belong to two classes of Acute Lymphoblastic (ALL) and Acute Myeloid (AML), which are similar under a microscope but have very different levels of genetic expression. The best diagnostic model PKR03 was studied on the training set (38 samples) and applied to the test set (the remaining 34 samples). The results were: 33 samples were diagnosed correctly (97% accuracy). Interestingly, one error was consistently misled by most algorithms and suspected of being mislabeled by the pathologist. So this may be one example where computer diagnostics is more accurate than a human expert. 1.2.6 Data detection, security and fraud detection are now numerous data collection applications for security and fraud detection. One of the most common is the detection of credit card fraud. Almost all credit card purchases are scanned by special algorithms that identify suspicious transactions for further action. I recently received such a call from my bank when I used a credit card to pay for a magazine published in England. It was an unusual deal for me (the first purchase in the UK on this card) and the software taged it. Other applications include money laundering detection - a notable system called FAIS, developed by Ted Senator for the U.S. Treasury. Se96. The National Association of Securities Dealers (NASD), which manages NASDA's, has developed a system called Sonar, which uses mining data to monitor insider trading and fraud by distorting (Many telecommunications companies, including ATT, Bell Atlantic, British Telecom/MCI have developed systems for catching phone fraud. In 2003, there was also a lot of evidence in the headlines that the U.S. government was making efforts to use data to detect terrorism as part of the now-closed General Awareness Program. (TIA). However, the problem of terrorism is unlikely to disappear any time soon, and the Government's efforts continue in other programmes, such as CAPPS II or MATRIX. Less controversial is the use of data to detect bioterro terrorism, as was done at the 2002 Salt Lake Olympics (the only thing that was there was a small outbreak of tropical diseases). The system used there has made a very interesting analysis of unusual events - we will return to this topic later in this course. 1.2.7 Problems suitable for mining data Previous case studies show some successful (and unsuccessful) data analysis applications. Areas in which data collection applications can be successful have these characteristics: knowledge-based solutions have changing environments, have a suboptimal current method, are available, sufficient, and relevant data provide a high buy-back for the right decisions, and if the problem is related to people, then due attention to privacy should be paid - otherwise, as the TIA example shows, the result will fail, regardless of technical problems. 1.3 Discovery of Knowledge We define Discovery of Knowledge in Data (KDD) as a non-trivial process of identifying a valid novel of potentially useful and ultimately understandable patterns in the data. of advances in knowledge discovery and data analysis, Fayyad, Piatetsky-Shapiro, Smith, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996 Discovery of Knowledge is an interdisciplinary area that is based on databases and statistics and applies machine learning and imaging techniques in order to find useful models. Other related areas also include information search, artificial intelligence, OLAP, etc. Some people say that data analysis is essentially a fancy name for statistics. It is true that data analysis has a lot in common with statistics and machine learning. However, there are differences. Statistics provide a solid theory for combating randoms and tools to test hypotheses. It does not study topics such as pre-processing data or visualization of results that are part of data analysis. Machine learning has a more guristic approach and is focused on improving the performance of the training agent. It also has other under fields such as real-time learning and robotics - which are not part of data analysis. The Data Mining and Discovery of Knowledge deposit integrates theory and guristics. The focus is on the entire knowledge discovery process, including data cleanup, training, integration, and visualization of results. 1.3.1 The process of discovering knowledge The key difference between focusing on the field of discovery of knowledge lies in the process. KDD is not a one-step solution to apply machine learning to a dataset, but a continuous process with a lot of loops and feedback. This process has been formalized by an industry group called CRISP-DM, which advocates for the CRoss Industrial Standard Process for Data Mining Key steps in this process include: 1. Business (or Problem) Understanding 2. Understanding 3. Preparing data (including all cleanup and pre-processing) 4. Simulation (use of machine learning and data analysis algorithms) 5. Ratings Ratings 6. Deployment While not officially part of CRISP, we should also consider the 7th step - Monitoring, which completes the circle. For more information www.crisp-dm.org crisp-DM, visit www.crisp-dm.org more. 1.3.2 Historic Note: Many of the names of the data Mining Data Mining and Knowledge Discovery field are called by many names. In the 1960s, statisticians used terms such as data fishing or data dredging to refer to what they considered to be poor data analysis practices without the a priori hypothesis. The term data mining appeared around 1990 in the database community. In short, there was the phrase mining data ™, but it was a trademark of the HNC (now part of the fair, Isaac), and the researchers turned to data mining. Other terms used include data archaeology, information collection, information discovery, knowledge extraction, etc. by Gregory Pyatitsky-Shapiro coined the term Discovery of Knowledge in Databases for the first seminar on the same topic (1989), and the term has become more popular in the AI and machine learning community. However, the term data mining has become more popular in the business community and in the press. As of January 2004, Google's search for data mining finds more than 2,000,000 pages, while the discovery search finds only 300,000 pages. In 2003, data mining gained a bad image because of its association with the U.S. government program TIA (General Information Awareness). Headlines such as the Senate Kills Data Mining Program, ComputerWorld, july 18, 2003, citing the U.S. Senate's decision to close the TIA, show how much data has become associated with TIA. Data Mining and Knowledge Discovery are currently interchangeable, and we also use these terms as synonyms. 1.4 Data analysis is associated with different types of patterns, and there are, accordingly, many types of data collection tasks. Some of the most popular are classification: prediction of cluster class elements: search of clusters in data associations: for example, often there is visualization: to facilitate human discovery Summary: description of the group Detection of deviations: Search for changes Score: predicting continuous analysis of link values: search of relationships ... Classification refers to the training method of predicting the class of instances from pre-marked (classified) instances. This is the most popular task, and there are dozens of approaches, including statistics (logistical regression), decision trees, neural networks, etc. 1.5 Technology trends leading to flood data analysis are needed to understand the data, data mining has many applications, successful rather than data mining and Knowledge Discovery Discovery Detection of the process of data processing Classification tasks, clustering, ... For more information on mining and discovery, including news, software publications, decision courses, meetings, education publications, websites, datasets companies, jobs visit www.KDnuggets.com. www.KDnuggets.com. data warehousing and data mining notes. data warehousing and data mining notes for it 7th sem. data preprocessing in data mining notes. data warehouse and data mining notes rejinpaul. data warehousing and data mining notes tutorialspoint. data warehousing and data mining notes aktu. data warehousing and data mining notes for mca. data mining notes for students pdf