Data mining fundamentals

Data Mining DM Defined Is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner Process of analyzing data from different perspectives and summarizing it into useful information A class of database applications that look for hidden patterns in a group of data that can be used to predict future behavior. DM Defined The relationships and summaries derived are referred to as models or patterns. Examples include linear equations, rules, clusters, graphs, tree structures and recurrent patterns in time series.

Utilizes observational data as opposed to experimental data. Data that have already been colleted for some purpose other than data mining analysis. The relationship and structures sort, should be novel. Its of little point regurgitating unless the ‘ confirmatory hypothesis’ is used. “ Concepts” •Definition: A “ concept” is a set of objects, symbols or events grouped together because they share certain characteristics. Concept set, class, group, cluster, roughly • Classical View: Concept Set with well defined deterministic inclusion rules. E. g. A home owner is a good credit risk. Probabilistic View: A set with probabilistic inclusion rules. E. g. A home owner has an 80% chance of being a good credit risk. • Exemplar View: this states that a given instance is determined to be an example of a particulalr concept if the instance is “ similar enough” to a set of “ one or more known examples” of the concept. Eg. Mr. Smith owns his own home and is a good credit risk. Example: An Investment Dataset Possible Business Questions “ Supervised” Leaning In last two questions, we distinguish ONE of the attributes that we would like to be able to determine from the values of the

others. What characteristics distinguish between Online and Broker investors? (DISCRIMINATION). (Transaction method (categorical)) is the target variable . • Can I develop a model which will predict the average trades/month for a new investor? (PREDICTION). (Trades/month (real)) is the target variable. The Target variable is called the “ Output variable”. The other variables are called “ Input variables”. Clearly, which attributes are the output and input variables depends on your question. For these questions, and output variables, we KNOW the values of the output variables for the cases in the dataset.

In such cases we say that we do “ SUPERVISED” learning since the learning is controlled by the known values of the output variable in the dataset. “ Unsupervised” Learning For the question: “ Can I develop a general characterisation/profile of different investor types? (CLASSIFICATION)”, NO particular attribute is singled out as an OUTPUT variable. •The question is open-ended. •We do not know if there are any different investor types at all. •If there are different investor types, we do not know how many types there are. •If there are different investor types then we do not know what the various investor type (or classes, or concepts) mean.

We have to determine the meaning of the concepts, and appropriate names, after we have determined that they exist. •The method of induction based learning used is said to be UNSUPERVISED in such a situation, because the there are no known output classes to control the learning process. Another Example Dataset Two Concept Learning Paradigms •Supervised Learning – builds a learner model, or concept definitions, using data instances of known origin. – and uses the model to determine the outcome new instances of

unknown origin. •Unsupervised Learning – A data mining method that builds models from data without predefined classes. Usually for classification/clustering. Return to supervised and unsupervised learning later Elements in Data Mining •Extract, transform, and load transaction data onto the data warehouse system. •Store and manage the data in a multidimensional database system. •Provide data access to business analysts and information technology professionals. •Analyze the data by application software. •Present the data in a useful format, such as a graph or table. Seeking relationships The process should aim at accurate, convenient and useful summaries. The steps are as follows: •Determine the nature and structure of the representation to be used Decide how to quantify and compare how well different representations fit the data (choosing the score function) •Choose an algorithmic process to optimize the score function •Decide what principles of data management are required to implement the algorithm efficiently. Seeking relationships: example In simple regression, one can build a predictive model to relate the predictor variable, X to a response variable Y through a relationship Y = aX + b. e. g. we might build a model which would allow us to predict a persona’s annual credit card spending given their annual income. Not perfect but good for a rough characterization.

Step wise: •The representation is a model in which the response variable, spending, is linearly related to the predictor variable, income. •The score function: the sum of squared discrepancies between predicted spending and observed spending in group of people described by the data •The optimization algorithm is quite simple: a and b are expressed as explicit

functions of the observed values of spending and income •Unless the data set is very large, simple sum, sums of squares and sums of products of X and Y can compute the estimates of a and b. Meaning a simple run through the data will give results. Relationships sought •Classes:

Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials. •Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities. •Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining. •Sequential patterns: Data is mined to anticipate behavior patterns and trends.

For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes. Mining Tasks •Data Mining is interdisciplinary! Statistics, Database technology, Machine learning, Pattern recognition, Artificial intelligence , Visualization –Data Mining Tasks: •Exploratory Data Analysis (EDA) –No particular ideas of what is being sort –Interactive and visual •Descriptive Modeling –Describe all data (generates data!) –Density estimation, partitioning, cluster analysis, segmentation, dependency modeling •Predictive Modeling –Classification and Regression Predict an unknown variable based on known values of variables •Discovering patterns and Rules –Pattern detection –Abnormally detection (outlier) •Retrieval by content –Find pattern of interest (similar to one already known) –Text and

image mining Levels of Analysis •Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure. •Genetic algorithms: Optimization techniques that use processes such as mutation, and natural selection in a design based on the concepts of natural evolution. •Decision trees: Tree-shaped structures that represent sets of decisions.

These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). •Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset. Sometimes called the k-nearest neighbor technique. •Rule induction: The extraction of useful if-then rules from data based on statistical significance. •Data visualization: The visual interpretation of complex relationships in multidimensional data.

Graphics tools are used to illustrate data relationships. Is Data Mining Appropriate for My Problem? Will Data Mining help me? •Can we clearly define the problem •Do potentially meaningful data exist? •Do the data contain hidden knowledge or the data is useful for reporting purposes only? •Will the cost of processing the data be less than the likely increase in profit seen by applying any potential knowledge gained from the data mining? Supervised Learning: A Decision Tree Example Back to supervised and unsupervised learning Notes on this Decision Tree: •The “ tree” is upside down. •The Decision Tree fits the data perfectly.

There are no errors. Accuracy = 100%. •The Decision Tree discards the unneccessary attributes •A computer algorithm to construct Decision Trees would be farly easy to programme, and would do the job much quicker than we humans can. Use of the Decision Tree for Prediction Production Rules We may summarize the Decision Tree by listing the decisions along each path from the starting node to each terminal node. 1. IF Swollen Glands = Yes THEN Diagnosis = Strep Throat 2. IF Swollen Glands = No & Fever = Yes THEN Diagnosis = Cold 3. IF Swollen Glands = No & Fever = No THEN Diagnosis = Allergy Unsupervised Clustering

A data mining method that builds models from data without predefined output classes. Data Mining Applications •Data mining is a young discipline with wide and diverse applications –There is still a nontrivial gap between general principles of data mining and domain-specific, effective data mining tools for particular applications •Some application domains –Biomedical and DNA data analysis –Financial data analysis –Retail industry – Telecommunication industry Biomedical Data Mining and DNA Analysis •DNA sequences: 4 basic building blocks (nucleotides): adenine (A), cytosine (C), guanine (G), and thymine (T). Gene: a sequence of hundreds of individual nucleotides arranged in a particular order •Humans have around 100, 000 genes •Tremendous number of ways that the nucleotides can be ordered and sequenced to form distinct genes •Semantic integration of heterogeneous, distributed genome databases –Current: highly distributed, uncontrolled generation and use of a wide variety of DNA data –Data cleaning and data integration methods developed in data mining will help DNA Analysis: Examples •Similarity search and comparison among DNA

sequences –Compare the frequently occurring patterns of each class (e. g. diseased and healthy) –Identify gene sequence patterns that play roles in various diseases •Association analysis: identification of co-occurring gene sequences –Most diseases are not triggered by a single gene but by a combination of genes acting together –Association analysis may help determine the kinds of genes that are likely to co-occur together in target samples •Path analysis: linking genes to different disease development stages –Different genes may become active at different stages of the disease –Develop pharmaceutical interventions that target the different stages separately •Visualization tools and genetic data analysis

Data Mining for Financial Data Analysis •Financial data collected in banks and financial institutions are often relatively complete, reliable, and of high quality •Design and construction of data warehouses for multidimensional data analysis and data mining –View the debt and revenue changes by month, by region, by sector, and by other factors –Access statistical information such as max, min, total, average, trend, etc. •Loan payment prediction/consumer credit policy analysis –feature selection and attribute relevance ranking –Loan payment performance –Consumer credit rating

Financial Data Mining •Classification and clustering of customers for targeted marketing –multidimensional segmentation by nearest-neighbor, classification, decision trees, etc. to identify customer groups or associate a new customer to an appropriate customer group •Detection of money laundering and other financial crimes –integration of from multiple DBs (e. g. , bank transactions, federal/state crime history DBs) –Tools: data

visualization, linkage analysis, classification, clustering tools, outlier analysis, and sequential pattern analysis tools (find unusual access sequences)

Data Mining for Retail Industry •Retail industry: huge amounts of data on sales, customer shopping history, etc. •Applications of retail data mining – Identify customer buying behaviors –Discover customer shopping patterns and trends –Improve the quality of customer service –Achieve better customer retention and satisfaction –Enhance goods consumption ratios – Design more effective goods transportation and distribution policies Data Mining in Retail Industry: Examples Design and construction of data warehouses based on the benefits of data mining –Multidimensional analysis of sales, customers, products, time, and region •Analysis of the effectiveness of sales campaigns •Customer retention: Analysis of customer loyalty –Use customer loyalty card information to register sequences of purchases of particular customers –Use sequential pattern mining to investigate changes in customer consumption or loyalty –Suggest adjustments on the pricing and variety of goods Purchase recommendation and cross-reference of items Data Mining for Telecomm. Industry (1) •A rapidly expanding and highly competitive industry and a great demand for data mining –Understand the business involved –Identify telecommunication patterns –Catch fraudulent activities –Make better use of resources –Improve the quality of service •Multidimensional analysis of telecommunication data –Intrinsically multidimensional: calling-time, duration, location of caller, location of callee, type of call, etc. Data Mining for Telecomm. Industry (2) Fraudulent pattern analysis and the identification of unusual patterns –Identify potentially fraudulent users and their atypical usage patterns –Detect attempts to gain

fraudulent entry to customer accounts –Discover unusual patterns which may need special attention •Multidimensional association and sequential pattern analysis –Find usage patterns for a set of communication services by customer group, by month, etc. –Promote the sales of specific services – Improve the availability of particular services in a region •Use of visualization tools in telecommunication data analysis