PHENOMENAL DATA MINING: FROM OBSERVATIONS TO PHENOMENA

www-formal.stanford.edu/jmc/data-mining.html

John McCarthy
Stanford University
jmc@cs.stanford.edu

• Conventional data mining infers relations among data, e.g. the fraction of supermarket baskets with diapers that also contain beer.

• *Phenomenal* data mining concerns relations between the data and the phenomena underlying the data, e.g. young married couples keeping old friends buy diapers and beer.

• Example: The sales receipts of a supermarket usually do not identify the customers. Grouping baskets by customer is possible and useful but requires new techniques.

OBSERVATIONS versus PHENOMENA

Events occur in the world.

The events sometimes cause some observations in an observer.

Two cars collide and a blind person hears the noise.

A person buys some groceries and a database entry is generated.

The observer infers from the sound of collision and subsequent shouts that someone was injured. He further infers that it was someone he knows.

OBSERVATIONS and PHENOMENA

- Databases of purchases are observations of customer behavior.
- Programs going beyond observations need knowledge of the world.
- A supermarket program needs facts about rates of consumption, items that go together, needs of various kinds of customers.
- The data-mining program infers that 30 baskets were purchased by the same female customer with at least three children whose husband goes on long trips.

THE MAIN TOOLS

- Extend the relational database to include entities like customers not present in the original database.

- Knowledge base of facts represented as sentences in a first order logical language.

- Minimize the total anomaly of the extended database

SUPERMARKET PROBLEM WITH MADE-UP NUMBERS

- Chain has 1,000 supermarkets.
- Supermarket stocks 10,000 items.
- Supermarket has 10,000 customers.
- 1,000 purchase "baskets" per day.
- 20 items per "basket".

Group baskets purchased by the same customer.

GROUPING BASKETS BY CUSTOMER

- Data records purchases but not always customers, but customer info is useful.

- Can a suitable data miner group baskets by customers well enough to be useful?

- We call this identifying customers even though it doesn't give us the customers' names.

- Grouping by customer is not a *clustering* problem, although there are some resemblances. Why?

- Use any available information about people's consumption and buying habits.

EXAMPLES OF FACTS

- Rates of consumption vary less than rates of purchase.
- Children consume milk at steady rates.
- A family that buys diapers will soon buy baby food and six months later junior food.
- Variety in detergents is not a consumer goal.
- Variety in soft drinks is often wanted.

- Italians buy much olive oil.

Which of these facts can a program use—and how?

THE SIGNATURE HYPOTHESIS

- Most customers have enough unique purchase patterns among the 10,000 items to constitute an identifying signature.

- Signature based on items for which variety is not especially desired by customer, e.g. brand of dishwasher detergent.

- Problem: Customers don't buy much of their signatures each time they go to the store.

Signatures are only one of many tools for identifying customers.

ASSIGNMENTS AND THEIR ANOMALIES

- An *assignment* assigns each basket to a putative customer.

- A *partial assignment* assigns some baskets to customers.

- If α is an assignment, $anomaly(\alpha)$ measures how bad the assignment is. (Partial assignments too.)

- $anomaly(\alpha)$ is a sum with terms associated with the putative customers and terms associated with the assignment as a whole.

- The data miner hill climbs in the space of (partial) assignments minimizing (total) anomaly.

PER CUSTOMER ANOMALIES

- Badness of best signature. The signature ascription gives probabilities of purchase.

- Badness of consumption continuity. It is unlikely, though not impossible, that a family of three will buy ten pounds of sugar on each of two successive days.

- Badness of demographic ascription.

SIGNATURES CHANGE

- Customers change their buying habits and hence their signatures. If the changes aren't too great, they can be tracked.

- Some changes don't count as raising an anomaly, e.g. change from buying baby food to buying junior food.

- A fact about the world:

$$Buys(Babyfood, customer, s) \rightarrow (\exists s')(s < s' \\ Buys(Juniorfood, customer, s'). \qquad (1)$$

- A corresponding fact about the data mining:

$$x \in Purchases(Basket1) \wedge x \in Babyfood$$
$$\wedge Time(Basket1) < Time(Basket2) \wedge y \in Basket2$$
$$\wedge y \in Juniorfood \wedge Ascribed(Basket1, customer)$$
$$\rightarrow Anomaly(y, Basket2, customer) = 0. \tag{2}$$

What does it take to derive (2) from (1)? What information must be in the knowledge base for this?

BADNESS OF ASSIGNMENT AS A WHOLE

- Number of distinct customers
- Wrong demographics
- Violates beliefs of marketing experts

WHAT USE IS PHENOMENAL DATA MINING?

- Stop buying hula hoops. Although sales have been increasing, they are only among preeteen girls, and they buy just one.

- Decide that product A will sell well in stores whose customers have been identified by phenomenal data mining as having a certain distribution of age, sex, ethnic, social class and taste characteristics. It is a waste of shelf space and of capital to sell it in other stores.

REMARKS

- An experiment to identify customers from supermarket data is worth making. The experiment would be best if customer identification were available but only used to verify identifications. Enough facts are readily obtained.

- How far away does the customer live? Don't be sure this can't be inferred.

- There are other applications and experiments. NASA wants data mining on data returned from spacecraft. Phenomenal data mining is what they need.

- Donal Lyons and Gregory Tseytain did PDM with Dublin Transport data.

APPLICATION TO CATCHING TERRORISTS

- The members of a terrorist group may use facilities in a common way that yields a signature. Thus one component of the Sept 11 terrorist signature would be using Travelocity.

- Groups with signatures can be inferred without any individual having been previously suspected.

- The FBI does a lot of what is essentially phenomenal data mining by hand, but some methods of finding groups are computationally intensive.

FORMULAS

Separate credit cards for terrorist expenses (dubious):

$$Has(person, creditcard1) \wedge Has(person, creditcard2)$$
$$\wedge Approximately\text{-}included(Purchases(creditcard1), TerroristItem$$
$$\wedge Approximately\text{-}disjoint(Purchases(creditcard1), TerroristItem$$
$$\rightarrow TwoCards \in Suspicions(person) \quad (3)$$

TERRORIST FORMULAS 2

Signatures:

Terrorists, like other groups of people, undoubtedly use the facilities of our society in special ways, some of which show up in databases of air travel, car rentals, telephone calls, credit card use, etc. They need to be distinguished from other groups, e.g. employees of some company or researchers in AI.

$$(\exists signature)((\forall person \in group)(adheres(signature, person)))$$
$$\wedge\neg(\exists employer)(Members(group) \subset Employees(employer))$$
$$\rightarrow suspicious(group) \tag{4}$$

TERRORIST FORMULAS 3

Identifying a group as common postponers of trip:

$$Occurs(Postponement(meeting), s) \rightarrow (\forall x)(Attendee(x, meeting)$$
$$\rightarrow Holds(Must(x, Postpone(Trip\text{-}Meeting(x))), Next(s))). \quad (5)$$

MORE REMARKS

- Suppose a customer of type i has a probability P_{ij} of including item j in a basket. We can infer an approximate number of types by looking at the approximate rank of the matrix P_{ij}.

- Classifying customers into discrete types may not give as good results as a more complex model that take into account the age of the customer as a continuous variable.

- A linear relation between phenomena and observations is the simplest case, and such relations can probably discovered by methods akin to factor analysis.

- We could infer that there were two subpopulations if we didn't already know about sex.

- We might infer from data from our stores in India, that there was a substantial part of the population that didn't purchase meat products. We can tell this from a situation in which everyone buys meat but less, because certain other purchase patterns are associated with not buying meat.

- If a customer buys a certain product but doesn't buy a necessary complementary product, we can infer that he buys the complementary product from someone else.

- Some brain storming is appropriate in thinking of customer patterns, because the more we can think of, the better the chances of identification.

HARANGUE about BAD PHILOSOPHY and INADEQUATE COMPUTER SCIENCE

Extreme positivism held that science consisted of relations among sense data.

Much learning research and even logical AI research involves making inferences about existing data expressed directly in terms of this data.

Science does better. We and our environment are complex structures built up from atoms.

The phenomena are not immediately apparent in the observations and are not just relations among observations.

Like science, phenomenal data mining uses whatever domain dependent information about the phenomena may be available and useful.