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 da
e.g. the fraction of supermarket baskets with diapers t
also contain beer.
• *Phenomenal* data mining concerns relations between
data and the phenomena underlying the data, e.g. yo
married couples keeping old friends buy diapers and b
• Example: The sales receipts of a supermarket usually
not identify the customers. Grouping baskets by custor
is possible and useful but requires new techniques.

OBSERVATIONS versus PHENOMENA

Events occur in the world.

The events sometimes cause some observations in an
server.

Two cars collide and a blind person hears the noise.

A person buys some groceries and a database entr
generated.

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

OBSERVATIONS and PHENOMENA

- Databases of purchases are observations of custo
behavior.

- Programs going beyond observations need knowle
of the world.

- A supermarket program needs facts about rates
consumption, items that go together, needs of vari
kinds of customers.

- The data-mining program infers that 30 baskets w
purchased by the same female customer with at le
three children whose husband goes on long trips.

THE MAIN TOOLS

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

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

- Minimize the total anomaly of the extended datab

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, customer info is useful.

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

- We call this identifying customers even though it do give us the customers' names.

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

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

EXAMPLES OF FACTS

- Rates of consumption vary less than rates of purcha
- Children consume milk at steady rates.
- A family that buys diapers will soon buy baby f
  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 p
terns among the 10,000 items to constitute an id
tifying signature.

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

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

Signatures are only one of many tools for identifying c
tomers.

ASSIGNMENTS AND THEIR ANOMALIES

- An *assignment* assigns each basket to a putative c tomer.

- A *partial assignment* assigns some baskets to c tomers.

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

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

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

PER CUSTOMER ANOMALIES

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

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

- Badness of demographic ascription.

SIGNATURES CHANGE

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

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

- A fact about the world:

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

- A corresponding fact about the data mining:

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

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

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 b increasing, they are only among preeteen girls, they buy just one.

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

REMARKS

- An experiment to identify customers from superm ket data is worth making. The experiment would best if customer identification were available but c used to verify identifications. Enough facts are rea obtained.

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

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

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

APPLICATION TO CATCHING TERRORISTS

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

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

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

FORMULAS

Separate credit cards for terrorist expenses (dubious)

$$
\begin{array}{l}
Has(person, creditcard1) \wedge Has(person, creditcard2) \\
\wedge Approximately\text{-}included(Purchases(creditcard1), Terr \\
\wedge Approximately\text{-}disjoint(Purchases(creditcard1), Terr \\
\qquad \rightarrow TwoCards \in Suspicions(person)
\end{array}
$$

TERRORIST FORMULAS 2

Signatures:

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

$$
\begin{aligned}
&(\exists signature)((\forall person \in group)(adheres(signature, pers\\
&\wedge\neg(\exists employer)(Members(group) \subset Employees(employe\\
&\qquad\qquad \rightarrow suspicious(group)
\end{aligned}
$$

TERRORIST FORMULAS 3

Identifying a group as common postponers of trip:

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

MORE REMARKS

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

- Classifying customers into discrete types may not g
as good results as a more complex model that t
into account the age of the customer as a continu
variable.

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

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

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

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

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

HARANGUE about BAD PHILOSOPHY and INADEQUATE COMPUTER SCIENCE

Extreme positivism held that science consisted of r
tions among sense data.

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

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

The phenomena are not immediately apparent in the
servations and are not just relations among observatio

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