Statistics 215a – 12/1/03 – D. R. Brillinger

Data mining – some details.

Problem. Suppose that a power station stores data about power consumption levels by time and by region, and power usage information per customer in each region.

a) Find similar power consumption curve fragments for a given region on Fridays.

b) Every time a power consumption curve rises sharply, what may happen within 20 minutes?

c) How can we find the most influential features that distinguish a stable power consumption region from an unstable one?

Han & Kamber (2001)

Time series case.

time series analysis / sequence analysis

Large archives of time series data sets, e.g. stock market, sensor data from space

shuttle missions (10 Gbytes/mission), human genome, …

Bioinformatics – science of storing, extracting, organizing, analyzing, interpreting and utilizing information from biological sequences and molecules

Wish investment strategies, decisions, strategies, knowledge, …

DM techniques used in the analysis and discovery of sequence, structure and functional patterns, models

Retrieval. Find the subsequence best matching a given query sequence.

E.g. finding customers whose spending patterns over time are similar to a given spending profile,

searching for similar past examples of unusual current sensor signals for real-time monitoring and fault diagnosis of complex systems (such as aircraft)

noisy matching of substrings in protein sequences.

Problems:

Identify components (trends, seasonals, cycles, irregular, …)

Similarity search – identification of a pattern sequence that is similar to a given pattern

Sequential pattern mining – identify sequences that occur frequently, e.g. weather X, then Y 5 days later

Periodicity analysis – identify patterns that repeat

Pattern – a local feature of the data, departure from general run, e.g. transient waveform

Time series, {x(1),…,x(T)},

define *queries*, e.g.

$$Q = \{q(t),\ldots,q(t+m)\}$$

Similarity measures

e.g. *Euclidian*

$$[x(t)-q(t)]^2 + \ldots + [x(t+m)-q(t+m)]^2$$

Cosine, r

$$(c,d)/(c,c)^{1/2}(d,d)^{1/2}$$

City block

$$|c_1 - d_1| + \ldots + |c_r - d_r|$$

(Data might be binary, nominal/categorical, ordinal, mixed type, …)

Approach - slide along / moving window

(*Prediction method)* – nearest neighbor

General strategy.

1. Determine a set of features to describe objects of interest (e.g. Fourier or wavelet coefficients – use “largest”)

2. Convert objects into vector representation

3. Perform matching / distance calculations

Steps of knowledge discovery (KD) ≡ DM:

1. Data cleaning (database)

2. Data integration (data warehouse)

3. Data selection

4. Data transformation

5. Data mining (extract patterns/knowledge)

6. Pattern evaluation (interesting?)

7. Knowledge presentation

(with feedback)

Questions.

How to implement computationally efficient search algorithm?

How to incorporate user feedback and interaction in the retrieval process?

How to evaluate the performance of a specific retrieval algorithm?