Data mining with Evolutionary Computation

Prabhas Chongstitvatana
Department of Computer Engineering
Chulalongkorn University

prabhas@chula.ac.th
www.cp.eng.chula.ac.th/faculty/pjw

Data mining

Identify valid, novel, potentially useful and ultimately understandable patterns in data.
(Frowley, Piatetsky-Shapiro & Matheus, 1991)

Data : Fact F
Pattern : Expression E describing subset F_E of F.

E is called pattern if it is simpler than the enumeration of all facts in F_E.

Data mining

a process that produces a particular enumeration of patter E_j over F. Space of patterns is often infinite and the enumeration of patterns involves some form of search in this space.

Two high level goals of data mining prediction and description (finding human interpretable pattern describing data)

using :

- Classification : maps a data item into predefined classes.
- Regression : maps a data item into a real-valued prediction variable.
- Clustering : identify a finite set of categories or clusters to describe the data.
- Summarization : finding compact description for a subset of data.
- Dependency modeling : finding a model which describes significant dependencies between variables.

- Change and deviation detection : discovering the most significant changes in the data from previously measured.

Applications

- Transforming rules and trees into comprehensible knowledge structure

- Finding patterns in time series using Dynamic Time Warping fig. DJIA 1989-1993 pp. 244,5,6,7

- Discovering association rules "98% of customers that purchase Discrete Math text book also get Introduction to Algorithm text book"

- Modeling subjective uncertainty in image annotation

Principle of supervised learning :
The system is trained using training data where the signal or category of interest is known. The trained model is subsequently used on data where the target variable is unknown and must be predicted.

fig. find Venus volcanoes p 521, 523

There are estimated 10^6 small (< 15 km diameter) visible volcanoes scatter throughout 30,000 images.

- Prediction equity return from securities data

The domain can be model by classification rules induced from available historical data for the purpose of making gainful predictions for equity investments.

S&P 500, 774 securities, 78 months, 40 variables each month for each security
fig p 552,3,4,5
rule 1 strongly under-performing
rule 481 strongly performing
fig 22.3
S&P 500 passive, investing in those companies in proportion of their capitalization.

(Advances in knowledge discovery and data mining, Fayyad and others (eds). MIT press, 1996)

- Finding association in collections of text (223)

association -- pattern of cooccurence -- amongst the keywords labeling items

- Learning patterns in images :detection of blasting caps in X-ray image of luggages (249) fig p 252, 254

- Learning to classify biomedical signals (409) : sleep classification, EEG patterns ; move the left hand, move a foot . . .

(EEG electroencephalogram : brain , EMG electromyogram : muscle)

(Machine Learning and Data Mining, Michalski, Bratko and Kubat, John Wiley & Son, 1998)

Evolutionary Computation for Data Mining

- Using GA to extract rules from trained NN (793, GECCO 99)

- GP for data mining of medical knowledge (254, GP 98)

Children with fractures 6500 records with 7 attributes

- Humerus fracture is the most common fracture of children between 2-5.
- Radius fracture is the most common fracture for boys between 11-13.
- Radius and ulna fracture are usually treated with plaster.
- Operation is usually not need for tibia fracture.
- Open reductions are more common for elder children with age > 11.

- Evolution of decision trees (350, GP 98)

- Rule acquisition with GA (778, GECCO 99)

- Discovering comprehensible classification rules using GP : medical domain (953, GECCO99)

If *low-voltage-of-the-electrocardiogram*
and *abnormality-found-in-chest-x-ray*
and *evidence-of-tampode-in-the-echocardiogram*
then *cardiac-tamponade*

If *air-in-pleural-space-observed-in-the-chest-x-ray*
then *pneumothorax*

- GP for classification and generalized rule induction (96, GP97)

- Discovering commonalities in collection of objects using GP (200, GP96)

Finding queries to compute the set of objects that have common characteristics. "*What are the common characteristics that the objects that belong to this set share*".

- GP for data mining : biochemistry of protein interactions (375, GP96)

GP determines functional relationships among the features.

The water displacement problem address predicting whether two molecules "bind" together.

Database contains 1700 water molecules bound to 13 different protein structures : thermo-liability, number of hydrogen bonds, atomic density, hydrophilicity.

75% accuracy. (other approaches best 72%)