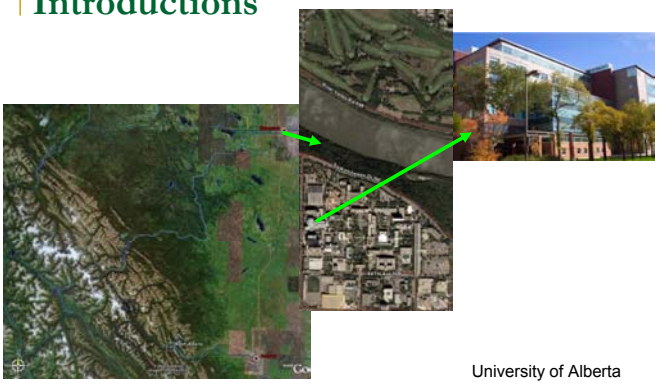
Discovering Structure in Data

Lukasz Kurgan
University of Alberta

Introductions



University of Alberta
Department of Electrical and Computer Engineering
data mining with applications to computational biology

Discovering Structure in **Data**

- Tabular (relational) multi-attribute and multi-sample
  - e.g. clinical patient records, microarray data, protein sequence data banks…
  - Numerical and nominal values
- Highly dimensional
  - # data samples (few thousands to few millions, or more…)
  - # attributes (few to several hundred, or more…)
- Analysis of such data is possible only using automated computational methods

Discovering Structure in Data

- Data Mining
  - defined as extraction of valid, useful, easily understandable knowledge from large collections of data, for high level decision making
  - research interests
    - data preprocessing (discretization, missing data imputation)
    - automated generation of data models
      - production and association rules
        (IF *a* and *b* and *c* THEN *Z*)
    - classification
      - discrete target concept
    - prediction
      - continuous target concept

Discovering Structure in Data

- automated generation of data models

  - does not require restrictive statistical assumptions such as independence, linear relationships, multi-colinearity, normality, etc.
  - finds rules for which a set of (independent) variables are correlated with a result, which simply means that given the 'IF' condition, the 'THEN' result occurs a given percentage of time.

DECISION	CONDITION1	CONDITION2	INDEX
A	low	normal	2
A	low	normal	3
A	normal	normal	3
A	normal	low	2
A	normal	low	1
B	low	high	4
B	low	low	4
B	high	normal	4
A	normal	low	2
A	normal	normal	2
A	normal	normal	2
A	normal	normal	3
A	low	normal	1
B	normal	high	4
B	high	low	4
B	high	normal	4
A	normal	low	4
A	low	normal	2
A	low	normal	2
A	normal	normal	4
A	normal	normal	2
B	low	high	4
B	high	low	4
B	high	normal	4

Discovering Structure in Data

- automated generation of data models

  - does not require restrictive statistical assumptions such as independence, linear relationships, multi-colinearity, normality, etc.
  - finds rules for which a set of (independent) variables are correlated with a result, which simply means that given the 'IF' condition, the 'THEN' result occurs a given percentage of time.

DECISION	CONDITION1	CONDITION2	INDEX
A	low	normal	2
A	low	normal	3
A	normal	normal	3
A	normal	low	2
A	normal	low	1
B	low	high	4
B	low	low	4
B	high	normal	4
A	normal	low	2
A	normal	normal	2
A	normal	normal	2
A	normal	normal	3
A	low	normal	1
B	normal	high	4
B	high	low	4
B	high	normal	4
A	normal	low	4
A	low	normal	2
A	low	normal	2
A	normal	normal	4
A	normal	normal	2
B	low	high	4
B	high	low	4
B	high	normal	4

RULES for DECISION: A (4 rules)
1. IF CONDITION1 = normal AND CONDITION2 = normal THEN DECISION = A
2. IF INDEX = 2 THEN DECISION = A
3. IF CONDITION1 = normal AND CONDITION2 = low THEN DECISION = A
4. IF CONDITION1 = low AND CONDITION2 = normal THEN DECISION = A

RULES for DECISION: B (3 rules)
1. IF CONDITION1 = high AND INDEX = 4 THEN DECISION = B
2. IF CONDITION1 = low AND INDEX = 4 THEN DECISION = B
3. IF CONDITION1 = normal AND CONDITION2 = high AND INDEX = 4 THEN DECISION = B

Discovering Structure in Data

- automated generation of data models
  - does not require restrictive statistical assumptions such as independence, linear relationships, multi-colinearity, normality, etc.
  - finds rules for which a set of (independent) variables are correlated with a result, which simply means that given the 'IF' condition, the 'THEN' result occurs a given percentage of time.

DECISION	CONDITION1	CONDITION2	INDEX
A	low	normal	2
A	low	normal	3
A	normal	normal	3
A	normal	low	2
A	normal	low	1
B	low	high	4
B	low	low	4
B	high	normal	4
A	normal	low	2
A	normal	normal	2
A	normal	normal	2
A	normal	normal	3
A	low	normal	1
B	normal	high	4
B	high	low	4
B	high	normal	4
A	normal	low	4
A	low	normal	2
A	low	normal	2
A	normal	normal	4
A	normal	normal	2
B	low	high	4
B	high	low	4
B	high	normal	4

ASSOCIATIONS (no "decision" target)
1. DECISION = B AND INDEX = 4
2. CONDITION1 = normal AND CONDITION2 = normal AND INDEX = 2
3. CONDITION2 = normal AND INDEX = 3
4. DECISION = A AND INDEX = 2
etc.

Discovering Structure in Data

- data models (rules, and others)
  - can be generated very fast
    - log-linear time with respect to number of data points
  - associations and rules allow to find hidden relations
  - rules (and other models) allow to perform classification and prediction
    - both associations (a special type called association classification) and rules can be used
    - other models include: decision trees, bayesian, regression, support vector machines, instance-based,… (may require more computations)

Discovering **Structure** in Data

- relevance
  - biology is a source of large and often unexplored databases
  - many biological problems can be translated into model generation and analysis, prediction and/or classification tasks
    - the goal is to find structure in the data
  - my recent interest is in protein structure analysis and prediction
    - analysis of both individual proteins and large protein clusters based on data stored in protein data banks
    - structure of other molecules and well as other biological data can also be analyzed and predicted…

Discovering **Structure** in Data

- protein structure
  - analysis of relation between structure and certain sequence and residue properties (physical, chemical, structural, etc.)
  - prediction of
    - secondary structure
    - secondary structure content
    - structural class

    based on protein sequence
  - analysis of prediction of tertiary protein structure

Discovering **Structure** in Data

- Prediction and Analysis of Secondary Protein Structure

NARBONIN (1NAR) protein

PKPIFREYIGVKPNSTTLHDFPTEIINTETLEFHYILGFAIESYYESGKGTGTFEESWDVELFGPEKVKNLKRRHPEVKVVISIGGRGVNTPFDPAEENVWSNAKESLKLIIQKYSIDDSGNLIDGIDIHYEHIRSDEPFATLMGQLITE
LKKDDDLNINVVSIAPSENNSSHYQKLYNAKKDYINWVDYQFSNQQKPVSTDDAFVEIFKSLEKDYHPHKVLPGFSTDPLDTKHNKTTRDIFGGCTRLVQTFSLPGVFFWNANDSVIPKRDGDKPFIVELTLQQLLAAR

Molecular Weight	RESIDUE COMPOSITION	RESIDUE COMPOSITION moment	RESIDUE PROPERTIES (electric charge, chemical group, etc.)	HYDROPHOBICITY based residue properties
33071.5	0.0345 0.0034 0.0759 0.0724 0.0586 0.0586 0.031 0.0828 0.0793 0.0759 0.0034	0.0211 0.0029 0.0449 0.0249 0.0307 0.0257 0.0147 0.0379 0.0403 0.0435 0.0017	0.3207 0.1103 0.2793 0.1414 0.1483 0.2793 0.2379 0.2276 0.4931 0.2172 0.2414 0.2241 0.169 0.1897 0.1724	0.2544 0.1154 0.2198 0.2893 0.1343 -0.02268 138.8 0.4788 139.6 0.05426 -0.05693 0.01952 0.03342 ....

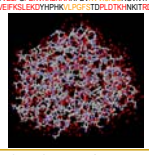
secondary structure content prediction

helix 35.9%; strand 21.7%; coil 42.4%

structural class classification

α-β proteins

secondary structure classification (black coil, yellow strand, red helix)

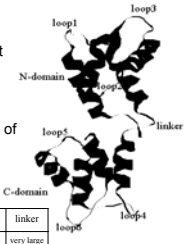


Kurgan L. and Homaeian L., Prediction of Secondary Protein Structure Content from Primary Sequence Alone - a Feature Selection Based Approach, *Proceedings of the International Conference on Machine Learning and Data Mining* (MLDM 2005), pp. 334-345, Leipzig, Germany, 2005

Kurgan L. and Kedarisetti K., Classification of Protein Secondary Structural Fragments without Alignment: Performance, Findings and Applications, *Computational Biology and Chemistry*, in review, 2005

Discovering **Structure** in Data

- Prediction and Analysis of Tertiary Calmodulin (CaM) Structure
  - known CaM structures were analyzed and four factors that influence the structure were found
  - The degree of influence of specific factors on different structural regions was investigated
  - a novel method for prediction of the CaM structure in complex with novel segments, given that the surroundings of the complex, was developed



Structure		loop1	loop2	loop3	loop4	loop5	loop6	linker
Flexibility		large	little	large	large	little	large	very large
Structure affecting factors	binding of Ca^{2+}	essential	no impact	essential	essential	no impact	essential	no impact
	binding segment and type	very little	very little	very little	very little	very little	very little	essential
	X-ray crystal surrounding	little	very little	little	little	little	little	essential
	mutation at linker	very little	very little	very little	very little	very little	very little	essential

Chen K., Ruan J. and Kurgan L., Prediction of Three Dimensional Structure of Calmodulin, *The Protein Journal*, in print, 2005

Discovering **Structure** in Data

- Intelligent Analysis of Cystic Fibrosis (CF) data
  - CF is a deadly genetic disease; it affects respiratory system, digestive system, endocrine system, and reproductive system
  - Project involved analysis of clinical CF data
    - in collaboration between the University of Colorado and the Denver's Children Hospital
    - (temporal) data on 856 patients collected starting in 1982
  - Goals
    - discovery of important factors that influence the pace of development of CF
      - several categories were defined based on an attribute that quantifies the progress of the disease in terms of the respiratory functions
    - discovery of important factors that are related to particular kinds of CF
      - CF is caused by at least 500 different genetic mutations but approximately 70% of the mutations are found to be "delta F508" gene (the most common CF mutation)
      - three kinds of CF were defined and analyzed: 1) both, Genotype 1 and Genotype 2 are F508, 2) either Genotypes 1 or Genotype 2 is F508, and the other is any other genotype, 3) both Genotype 1 and Genotype 2 are not F508

Discovering **Structure** in Data

- Intelligent Analysis of Cystic Fibrosis (CF) data
  - sample results for goal 1
    - significant and previously unknown finding was a relation between high value of sweatelectr1 (potassium levels) and the improvement of the disease

ATRIBUTE	VALUE	MARK	FASTDEGRAD					IMPROV					NOCHANGE					SLOWDEGRAD				
			TI1	TI2	TI3	TI4	TI5	TI1	TI2	TI3	TI4	TI5	TI1	TI2	TI3	TI4	TI5	TI1	TI2	TI3	TI4	TI5
CFtypes (cf)	Type4	2+																				
race (dem)	Black	3+																				
group (dem)	C	3+/4+																				
group (dem)	NBS	3+/4+																				
group (dem)	MI	3+/4+																				
group (dem)	FN	3+/4+																				
motage (dem)	(22.50,48.50)	3+																				
motage (dem)	(19.50,22.50)	3+																				
mecil (dem)	TreatedSurgically	2+																				
sweatelectr1 (dia)	**(24.50,46.99)**	**4+**																				
sweatelectr2 (dia)	(-11.00,95.50)	3+																				
tobaresistent7 (sul)	Suscept.	2+																				
na (mic)	(129.00,143.00)	2+																				
prot (mic)	(-1.95,7.95)	2+																				
vita (mic)	(0.44,748.00)	2+																				
wbc (hem)	(4.05,18.00)	2+																				
hct (hem)	(27.40,45.50)	2+																				
mch (hem)	(24.90,91.40)	2+																				
mchc (hem)	(30.40,35.80)	2+																				
rdw (hem)	(-0.85,15.40)	2+																				
HAZ (per)	(-2.91,-1.87)	3+																				
age@diagnosis (dia)	0258i92	3+																				

Kurgan L., Cios K., Sontag M., and Accurso F., Mining the Cystic Fibrosis Data, In: Zurada J. and Kantardzic M., (Eds.), *Next Generation of Data-Mining Applications*, pp. 415-444, IEEE Press - Wiley (ISBN 0-471-65605-4), 2005

THANK YOU