PKDD'99 Discovery Challenge

Guide to the Medical Data Set

Domain

Databases are collected at the University hospital. Each patient came to the outpatient clinic on collagen diseases, who is introduced by some home doctors or general physicians in the local hospitals.

Collagen diseases are auto-immune diseases whose patients generate antibodies attacking to their bodies. For example, if a patient generates antibodies to lung, he/she will lose the respiratory function in a chronic course and finally lose their lives. The disease mechanisms are only partially known and their classification are still fuzzy. Some patients may generate many kinds of antibodies and their manifestations may include all the characteristics of collagen diseases.

In these diseases, thrombosis is one of the most important and severe complications, one of the major cause of death in collagen diseases. Recently, this complication is closely related with anti-cardiolipin antibodies, which was discovered by the medical physicians, one of whom donated the datasets for discovery challenge.

Thrombosis is emergency and it is important to detect and predict the possibilities of thrombosis. However, such database analysis has not been made by any experts on immunology. Domain experts are very much interested in discovering some regularities behind patients' observations, which may be a really new discovery in the world.

Goal

1. Search for good patterns which detect and predict thrombosis.
2. Search for temporal patterns specific/sensitive to thrombosis. (Examination date is very close to the date on thrombosis. If we can find specific/sensitive patterns before/after the thrombosis, they are very useful.)
3. Search for good features which classifies collagen diseases correctly.
4. Search for temporal patterns specific/sensitive to each collagen diseases.

Domain experts told us that if useful patterns are discovered then they are acceptable in major journals on rheumatology(collagen diseases.)

Evaluation Scheme

One of the domain experts, who is well known for rheumatology, will attend PKDD'99 conference and evaluate all the results. The results will be also evaluated in the clinical environment in the future.

Databases

Databases consists of three tables. (tsumoto_a.csv, tsumoto_b.csv, tsumoto_c.csv). The patients in these three tables are connected by ID number.

tsumoto_a.csv

Basic Information about Patients (Input by Experts). This dataset includes all patients.

item	meaning	remark
ID	identification of the patient	
Sex		
Birthday		YYYY/M/D
Description date	the date when a patient was input	
First date	the date when a patient came to the hospital	
Admission	patient was admitted to the hospital or followed at the outpatient clinic	
Diagnosis	some patients may suffer from several diseases	multivalued attribute

tsumoto_b.csv

Special Laboratory Examinations (Input by Experts) (Measured by the Laboratory on Collagen Diseases)This dataset does not include all the patients, but includes the patients with these special tests.

item	meaning	remark
ID	identification of the patient	
Examination Date	date of the test	YYYY/MM/DD
aCL IgG	anti-Cardiolipin antibody (IgG)	
aCL IgM	anti-Cardiolipin antibody (IgM)	

ANA	anti-nucleus antibody	
ANA Pattern	pattern observed in the sheet of ANA examination	
aCL IgA	anti-Cardiolipin antibody (IgA)	
Diagnosis		multivalued attribute
KCT	meassure of degree of coagulation	
RVVT	meassure of degree of coagulation	
LAC	meassure of degree of coagulation	
Symptoms	other symptoms observed in each patient	
Thrombosis	degree of thrombosis (the Target)	0: negative 1: positive 2: positive and very severe

Examination date is very close to the date on thrombosis. In negative examples, these tests are examined when thrombosis is suspected.

tsumoto_c.csv

Laboratory Examinations stored in Hospital Information Systems (Stored from 1980 to 1999.3) All the data includes ordinary laboratory examinations and have temporal stamps.

item	meaning	normal range
ID	identification of the patient	
Date	Date of the laboratory tests (YYMMDD)	
GOT	AST glutamic oxaloacetic transaminase	N < 60
GPT	ALT glutamic pylvic transaminase	N < 60
LDH	lactate dehydrogenase	N < 500
ALP	alkaliphophatase	N < 300
TP	total protein	6.0 < N < 8.5
ALB	albumin	3.5 < N < 5.5
UA	uric acid	N > 8.0 (Male) N > 6.5 (Female)
UN	urea nitrogen	N > 30
CRE	creatinine	N > 1.5
T-BIL	total bilirubin	N < 2.0
T-CHO	total cholesterol	N < 250

TG	triglyceride	N < 200
CPK	creatinine phosphokinase	N < 250
GLU	blood glucose	N < 180
WBC	White blood cell	3500 < N < 9000
RBC	Red blood cell	350 < N < 600
HGB	Hemoglobin	10 < N < 17
HCT	Hematoclit	29 < N < 52
PLT	platelet	100 < N < 400
PT	prothrombin time	N < 14
APTT	activated partial prothrombin time	N < 45
FG	fibrinogen	150 < N < 450
PIC	plasmininhibitor-plasmin complex	N < 0.8
TAT	thrombin-antithrombin III complex	N < 3.0
U-PRO	proteinuria	0 < N 30
IGG	Ig G	900 < N < 2000
IGA	Ig A	80 < N < 500
IGM	Ig M	40 < N < 400
CRP	C-reactive protein	N= -, +-, or N < 1.0
RA	Rhuematoid Factor	N= -, +-
RF	RAHA	N < 20
C3	complement 3	N > 35
C4	complement 4	N > 10
RNP	anti-ribonuclear protein	N= -, +-
SM	anti-SM	N= -, +-
SCl70	anti-scl70	N= -, +-
SSA	anti-SSA	N= -, +-
SSB	anti-SSB	N= -, +-
CENTROMEA	anti-centromere	N= -, +-
DNA	anti-DNA	N < 8
DNA-II	anti-DNA	N < 8

This database was donated by dr. Katsuhiko Takabayashi and prepared by prof. Shusaku Tsumoto

For more details on the data description, please use following contact addresses:

Katsuhiko Takabayashi
Email: takaba@ho.chiba-u.ac.jp
Department of Internal Medicine II
Chiba University Hospital

Shusaku Tsumoto, M.D., Ph.D(Comp Sci)
Email: tsumoto@computer.org
Department of Medical Informatics
Shimane Medical University, School of Medicine
89-1 Enya-cho, Izumo-city
Shimane 693-8501 Japan
TEL +81-853-20-2172
FAX +81-853-20-2170