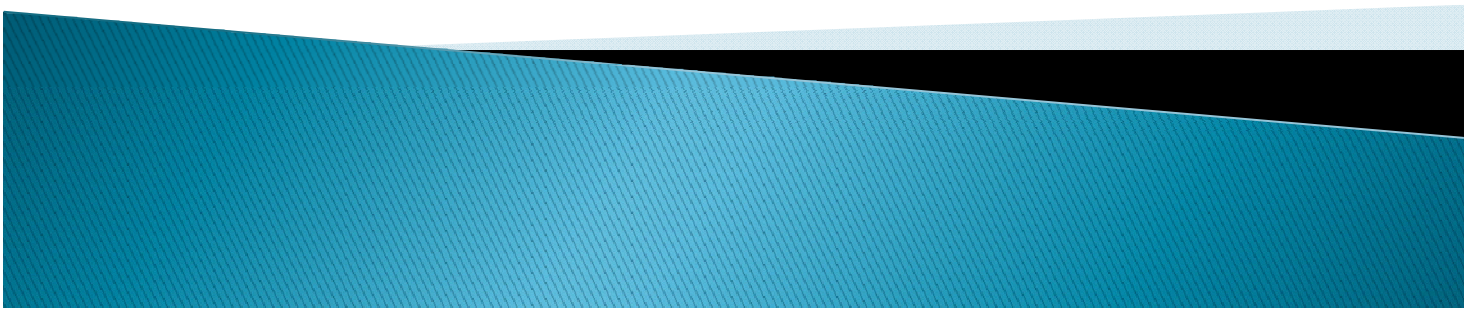
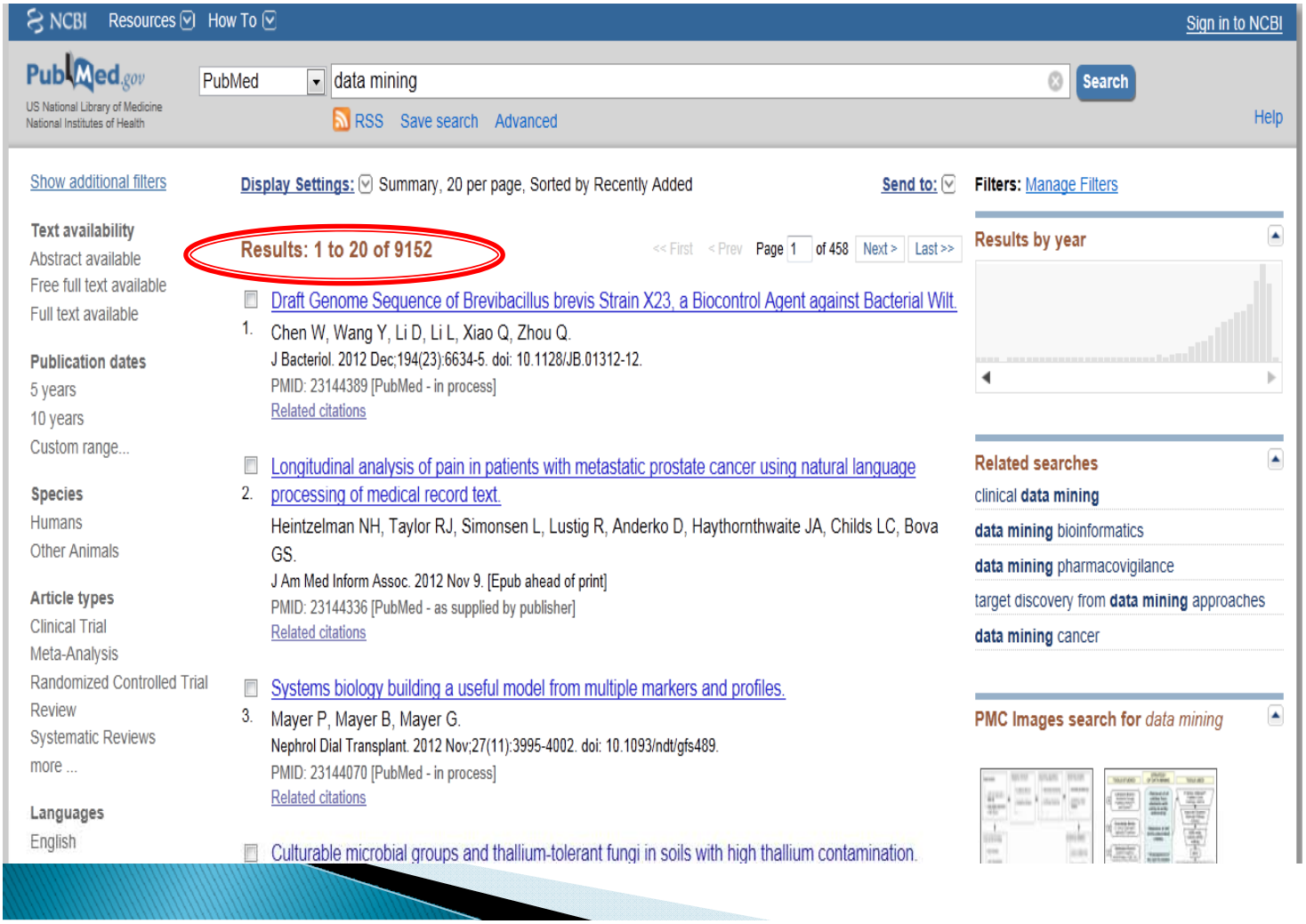
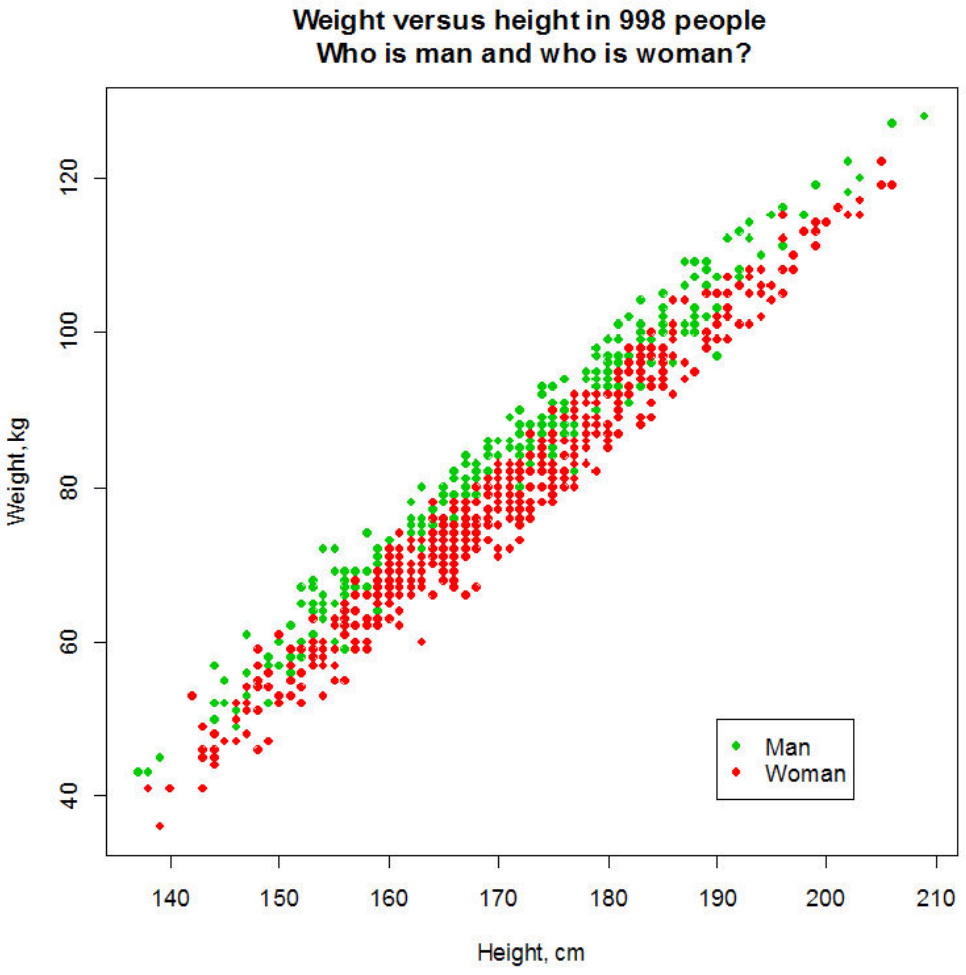
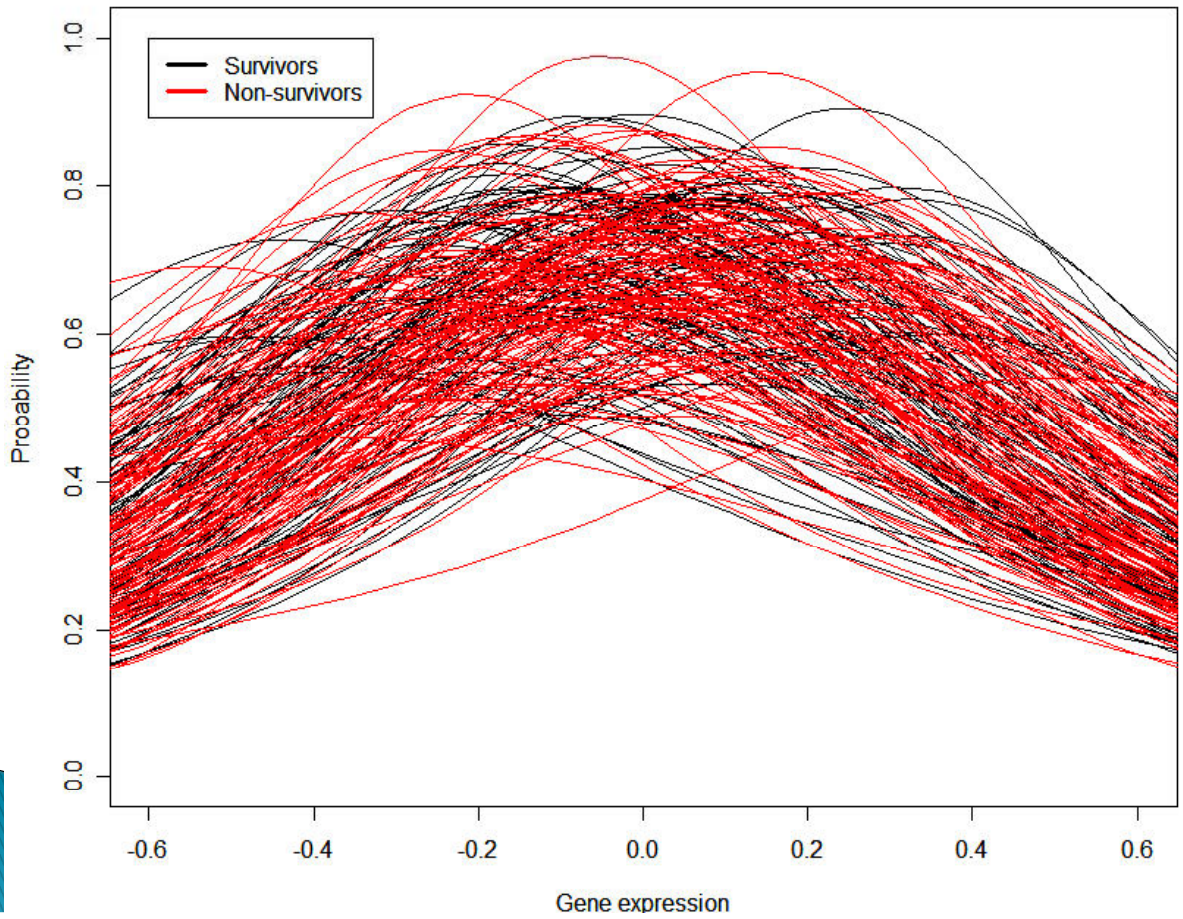
次級資料庫分析經驗(三)
DATA MINING簡介

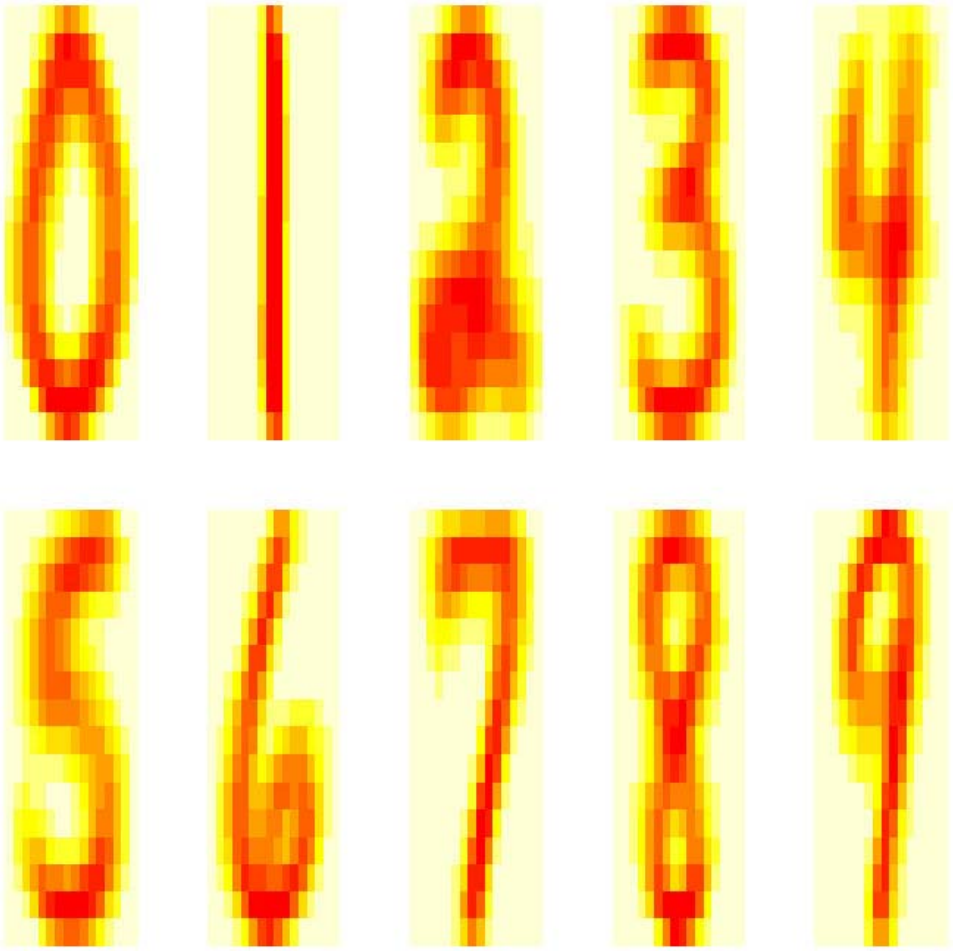
劉介宇
國立台北護理健康大學
護理助產研究所 / 通識教育中心副教授
兼教師發展中心教師評鑑組長
Nov 19, 2012

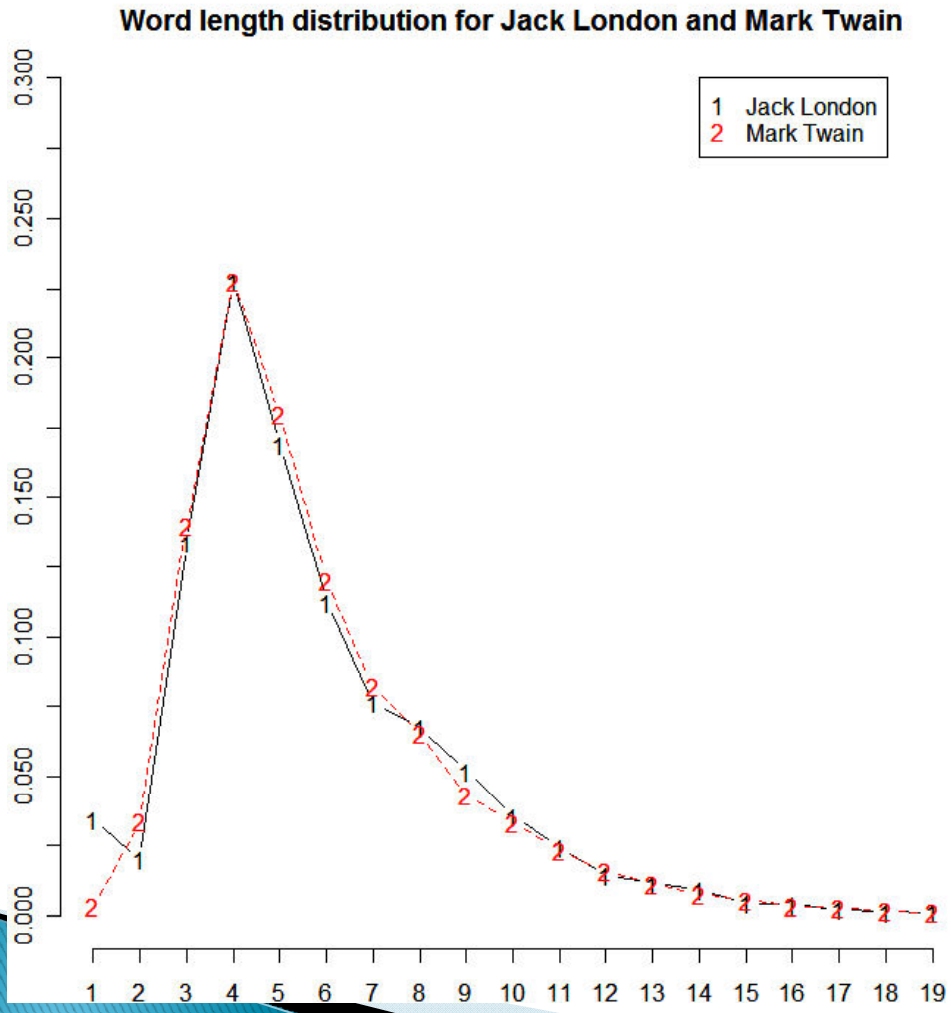



NCBI Resources How To
Sign in to NCBI
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
data mining
Search
RSS Save search Advanced
Help
Show additional filters
Display Settings: Summary, 20 per page, Sorted by Recently Added
Send to:
Filters: Manage Filters
Text availability
Abstract available
Free full text available
Full text available
Publication dates
5 years
10 years
Custom range...
Species
Humans
Other Animals
Article types
Clinical Trial
Meta-Analysis
Randomized Controlled Trial
Review
Systematic Reviews
more ...
Languages
English
Results: 1 to 20 of 9152
<< First < Prev Page 1 of 458 Next > Last >>
1. Draft Genome Sequence of Brevibacillus brevis Strain X23, a Biocontrol Agent against Bacterial Wilt.
Chen W, Wang Y, Li D, Li L, Xiao Q, Zhou Q.
J Bacteriol. 2012 Dec;194(23):6634-5. doi: 10.1128/JB.01312-12.
PMID: 23144389 [PubMed - in process]
Related citations
2. Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text.
Heintzelman NH, Taylor RJ, Simonsen L, Lustig R, Anderko D, Haythornthwaite JA, Childs LC, Bova GS.
J Am Med Inform Assoc. 2012 Nov 9. [Epub ahead of print]
PMID: 23144336 [PubMed - as supplied by publisher]
Related citations
3. Systems biology building a useful model from multiple markers and profiles.
Mayer P, Mayer B, Mayer G.
Nephrol Dial Transplant. 2012 Nov;27(11):3995-4002. doi: 10.1093/ndt/gfs489.
PMID: 23144070 [PubMed - in process]
Related citations
Culturable microbial groups and thallium-tolerant fungi in soils with high thallium contamination.
Results by year
Related searches
clinical data mining
data mining bioinformatics
data mining pharmacovigilance
target discovery from data mining approaches
data mining cancer
PMC Images search for data mining


Weight versus height in 998 people
Who is man and who is woman?
Weight, kg
40
60
80
100
120
140
150
160
170
180
190
200
210
Height, cm
Man
Woman


Gene expression Gaussion kernel densities for 240 patients
Survivors
Non-survivors
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
Gene expression





Outline

Overview of Data Mining (資料採礦)

- What is Data Mining?
- Steps in Data Mining
- Overview of Data Mining techniques
- Points to Remember

What is Data Mining?

- Data mining is an analytic process designed to explore large amounts of data in search of *consistent patterns* and/or *systematic relationships* between variables, and then to validate the findings by applying the detected patterns to new subsets of data.
- "Data Mining is a process of torturing the data until they confess"
- The typical goals of data mining projects are:
  - Identification *of groups, clusters, strata, or dimensions* in data that display no obvious structure,
  - The identification of factors that are related to a particular outcome of interest (**root-cause** analysis)
  - **Accurate prediction** of outcome variable(s) of interest (in the future, or in new customers, clients, applicants, etc.; this application is usually referred to as predictive data mining)

Steps in Data Mining

- Stage 1: Precise statement of the problem.
- Stage 2: Initial exploration.
- Stage 3: Model building and validation.
- Stage 4: Deployment.

Steps in Data Mining

Stage 1: Precise statement of the problem.

- Before opening a software package and running an analysis, the analyst must be clear as to what question he wants to answer. If you have not given a precise formulation of the problem you are trying to solve, then you are wasting time and money.

Stage 2: Initial exploration.

- This stage usually starts with data preparation that may involve the "cleaning" of the data (e.g., identification and removal of incorrectly coded data, etc.), data transformations/ normalization, selecting subsets of records, and, in the case of data sets with large numbers of variables, performing preliminary feature selection. Data description and visualization are key components of this stage (e.g. descriptive statistics, correlations, scatterplots, box plots, etc.).

Steps in Data Mining

Stage 3: Model building and validation.

- This stage involves considering various models and choosing the best one based on their predictive performance.

Stage 4: Deployment.

- When the goal of the data mining project is to predict or classify new cases (e.g., to predict the credit worthiness of individuals applying for loans), the third and final stage typically involves the application of the best model or models (determined in the previous stage) to generate predictions

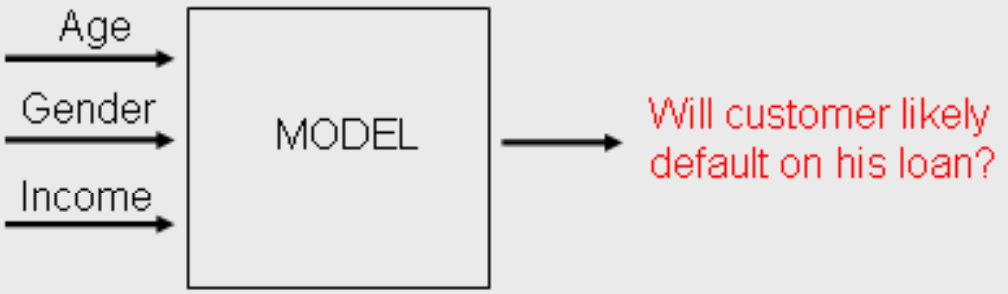
Initial exploration

- "Cleaning" of data,
  - Identification and removal of incorrectly coded data,

  e.g., Age=-90, 200, Height=60, Weight=160.
- Data transformations,
  - Data may be skewed (that is, outliers in one direction or another
  - may be present). Log transformation, Box-Cox transformation, etc.
- Data reduction, Selecting subsets of records, and, in the case of data sets with large numbers of variables ("fields"), performing preliminary feature selection.
- Data description and visualization are key components of this stage (e.g. descriptive statistics, correlations, scatterplots, box plots, brushing tools, etc.)
  - Data description allows you to get a snapshot of the important characteristics of the data (e.g. central tendency and dispersion).

Model building and validation.

- Data Mining involves creating models of reality
- A model takes one or more inputs and produces one or more outputs



- A model can be "transparent", for example, a series of if/then statements where structure is easily discerned, or a model can be seen as a black box, for example, neural network, where the structure or the rules that govern the predictions are impossible to fully comprehend.

Model building and validation.

- Validation of the model requires that you train the model on one set of data and evaluate on another independent set of data.
- There are two main methods of validation
  - Split data into train/test datasets (75-25 split)
  - If you do not have enough data to have a holdout sample, then use v-fold cross validation.

Data Mining Techniques

- Neural Networks
- Generalized EM And K-means Cluster Analysis
- General CART Models
- General CHAID Models
- Interactive Trees (C&RT and CHAID)
- Boosted Tree Classifiers and Regression
- Association Rules
- MARS(Multivariate Adaptive Regression Splines)
- Machine Learning(Bayesian, Support Vectors and Nearest neighbors)
- Random Forests for Regression and Classification
- Generalized Additive Models (GAM)
- Feature Selection and Variable Screening

Data Mining techniques

- Supervised Learning

Supervised learning is a machine learning technique for deducing a function from training data.

The training data consist of pairs of input variable and desired outputs. The task of the supervised learner is to predict the value of the function for any valid input object after having seen a number of training examples.

Classification and Regression are very popular techniques of supervised learning.

- Unsupervised Learning

In unsupervised learning training data set is not available in the form of input and output variable.

unsupervised learning is a class of problems in which researcher seeks to determine how the data are organized

Cluster analysis, and Principal component analysis are very popular techniques for unsupervised learning.

Points to Remember..

- Data mining is a tool, not a magic box.
- Data mining will not automatically discover solutions without guidance.
- To ensure meaningful results, it's vital that you understand your data.
- which leverages analytic tecUser-centric interactive processhnologies and computing power.
- Data mining central quest: Find true patterns and avoid overfitting (finding random patterns by searching too many possibilities)

Classification and Regression.

- Databases are rich with hidden information that can be used to make intelligent business decisions.
- Classification and Regression are two form of data analysis that can be used to extract models, describing important data classes or to predict future data trends.

- Classification is used to predict or classify categorical response variable, like to predict Iris type of flowers (Setosa,Verginica,Versocol).
- Regression is used to predict quantitative response variable, average income of household.

Steps of Classification and Regression models

Step 1: In the first step a model is built describing a predetermined set of data classes. (Supervised learning).

Step 2: In the second step the predictive accuracy of the model is estimated.

Step 3: If the accuracy of the model is considered acceptable, then the model can be used to classify future data for which the class label is unknown.

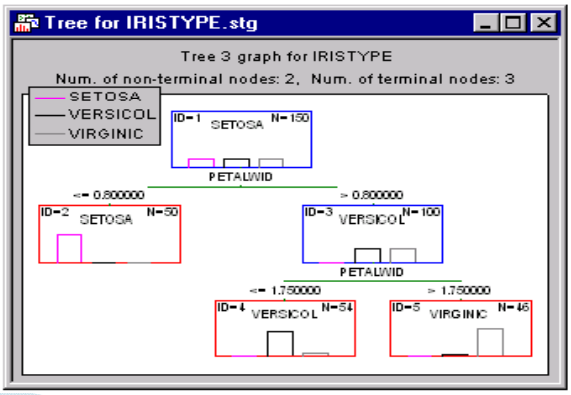
Techniques.以STATISTICA爲例

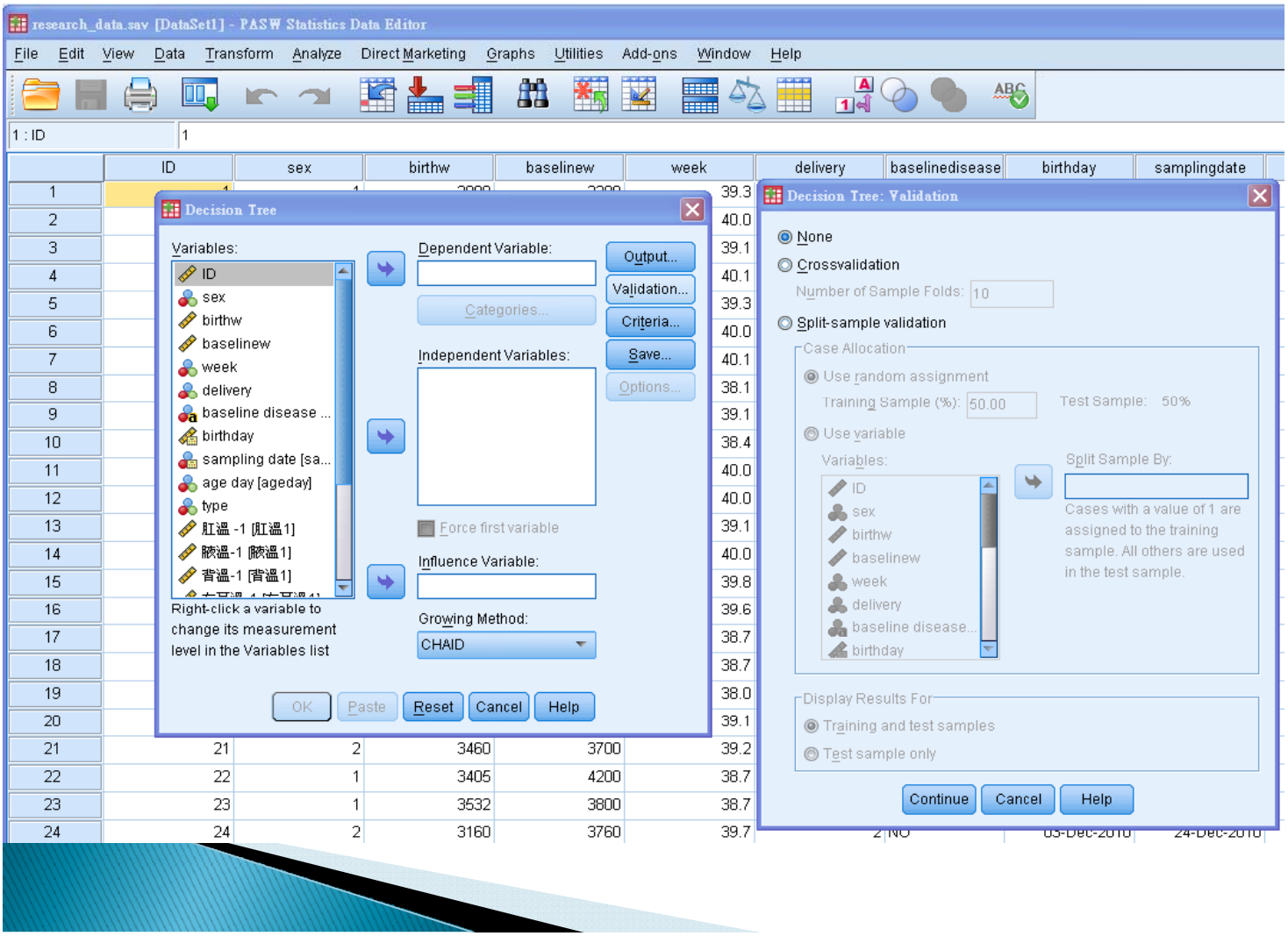
Different kind of Classification and Regression techniques are available in STATISTICA, including

1. Classification and Regression, through STATISTICA Automated Neural Network.
2. General Classification and Regression tree.
3. General CHAID model.
4. Boosted Tree Classification and Regression.
5. Random Forest for Classification and Regression, etc.

Decision Trees (also in SPSS)

- For example, consider the widely referenced Iris data classification problem introduced by Fisher (1936).
- The purpose of the analysis is to learn how one can discriminate between the three types of flowers, based on the four measures of width and length of petals and sepals.
- A classification tree will determine a set of logical if-then conditions (instead of linear equations) for predicting or classifying cases.





Advantages of tree methods.

Simplicity of results.

- In most cases, the interpretation of results summarized in a tree is very simple. This simplicity is useful not only for purposes of rapid classification of new observations .
- Often yield a much simpler "model" for explaining why observations are classified or predicted in a particular manner .

  e.g., when analyzing business problems, it is much easier to present a few simple if-then statements to management, than some elaborate equations.

Tree methods are nonparametric and nonlinear.

- The final results of using tree methods for classification or regression can be summarized in a series of logical if-then conditions .

  Therefore, there is no implicit assumption that the underlying relationships between the predictor variables and the dependent variable are linear, follow some specific non-linear link function , or that they are even monotonic in nature.

General Classification and Regression tree

- The *STATISTICA General Classification and Regression Trees* module *(GC&RT)* will build ***classification and regression trees*** for predicting continuous dependent variables (regression) and categorical predictor variables (classification).
- The program supports the classic C&RT algorithm and includes various methods for pruning and cross-validation, as well as the powerful ***v*-fold cross-validation** methods.
- **Classification and Regression Trees (C&RT)**

  In most general terms, the purpose of the analyses via tree-building algorithms is to determine a set of *if-then* logical (split) conditions that permit accurate prediction or classification of cases.

CHAID Model

- CHAID stands for CHi-squared Automatic Interaction Detector.
- CHAID, a technique whose original intent was to detect interaction between variables (i.e., find "combination" variables), recursively partitions a population into separate and distinct groups, which are defined by a set of independent (predictor) variables, such that the CHAID Objective is met – the variance of the dependent (target) variable is minimized within the groups, and maximized across the groups.
- Like other decision trees, its advantages are that its output is highly visual and easy to interpret.
- It uses multiway splits by default, it needs rather large sample sizes to work effectively.

CHAID and Exhaustive CHAID Algorithms:

- *Exhaustive CHAID*, a modification to the basic *CHAID* algorithm, performs a more thorough merging and testing of predictor variables, and hence requires more computing time.
- Specifically, the merging of categories continuous (without reference to any alpha-to-merge value) until only two categories remain for each predictor.
- The program then proceeds as described above in the *Selecting the split variable* step, and selects among the predictors the one that yields the most significant split.

For large data sets, and with many continuous predictor variables, this modification of the simpler *CHAID* algorithm may require significant computing time.

Machine Learning Algorithms.

These methods include

- *Support Vector Machines (SVM)*

(for regression and classification).

- *Naive Bayes* (for classification)

- *K-Nearest Neighbors (KNN)*

(for regression and classification.)

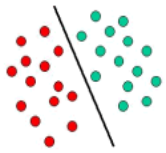
Support Vector Machines

Support Vector Machine (SVM) is primarily a classier method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels.

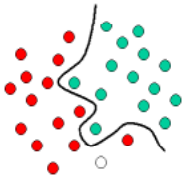
- To construct an optimal hyperplane, *SVM* employees an iterative training algorithm, which is used to minimize an error function. According to the form of the error function, SVM models can be classified into four distinct groups:

- Classification SVM Type 1 (also known as C-SVM classification).
- Classification SVM Type 2 (also known as nu-SVM classification).
- Regression SVM Type 1 (also known as epsilon-SVM regression).
- Regression SVM Type 2 (also known as nu-SVM regression).

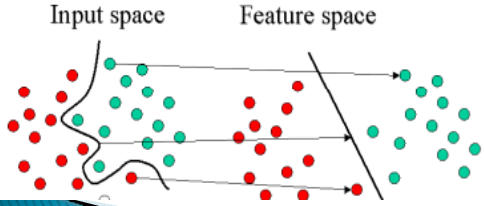
Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. A schematic example is shown in the illustration below. In this example, the objects belong either to class *GREEN* or *RED*. The separating line defines a boundary on the right side of which all objects are *GREEN* and to the left of which all objects are *RED*. Any new object (white circle) falling to the right is labeled, i.e., classified, as *GREEN* (or classified as *RED* should it fall to the left of the separating line).



The above is a classic example of a linear classifier, i.e., a classifier that separates a set of objects into their respective groups (*GREEN* and *RED* in this case) with a line. Most classification tasks, however, are not that simple, and often more complex structures are needed in order to make an optimal separation, i.e., correctly classify new objects (test cases) on the basis of the examples that are available (train cases). This situation is depicted in the illustration below. Compared to the previous schematic, it is clear that a full separation of the *GREEN* and *RED* objects would require a curve (which is more complex than a line). Classification tasks based on drawing separating lines to distinguish between objects of different class memberships are known as hyperplane classifiers. *Support Vector Machines* are particularly suited to handle such tasks.



The illustration below shows the basic idea behind *Support Vector Machines*. Here we see the original objects (left side of the schematic) mapped, i.e., rearranged, using a set of mathematical functions, known as kernels. The process of rearranging the objects is known as mapping (transformation). Note that in this new setting, the mapped objects (right side of the schematic) is linearly separable and, thus, instead of constructing the complex curve (left schematic), all we have to do is to find an optimal line that can separate the *GREEN* and the *RED* objects.



Classification SVM

Classification SVM Type 1

For this type of *SVM*, training involves the minimization of the error function:

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i$$

subject to the constraints:

$$y_i\left(w^T\phi(x_i)+b\right)\geq 1-\xi_i \text{ and } \xi_i \geq 0,\ \ i=1,\ldots,N$$

where C is the capacity constant, w is the vector of coefficients, b a constant and ξ_i are parameters for handling nonseparable data (inputs). The index i labels the N training cases. Note that $y \in \pm 1$ is the class labels and x_i is the independent variables. The kernel ϕ is used to transform data from the input (independent) to the feature space. It should be noted that the larger the C, the more the error is penalized. Thus, C should be chosen with care to avoid over fitting.

Classification SVM Type 2

In contrast to *Classification SVM Type 1*, the *Classification SVM Type 2* model minimizes the error function:

$$\frac{1}{2}w^T w - v\rho + \frac{1}{N}\sum_{i=1}^{N}\xi_i$$

subject to the constraints"

$$y_i\left(w^T\phi(x_i)+b\right)\geq \rho-\xi_i,\ \xi_i \geq 0,\ \ i=1,\ldots,N \text{ and } \rho \geq 0$$

Regression SVM

In a regression *SVM*, you have to estimate the functional dependence of the dependent variable y on a set of independent variables x. It assumes, like other regression problems, that the relationship between the independent and dependent variables is given by a deterministic function f plus the addition of some additive noise:

$y = f(x)$ + noise

The task is then to find a functional form for f that can correctly predict new cases that the *SVM* has not been presented with before. This can be achieved by training the *SVM* model on a sample set, i.e., training set, a process that involves, like classification (see above), the sequential optimization of an error function. Depending on the definition of this error function, two types of *SVM* models can be recognized:

Regression SVM Type 1

For this type of *SVM* the error function is:

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i + C\sum_{i=1}^{N}\xi^*_i$$

which we minimize subject to:

$$w^T\phi(x_i)+b-y_i \leq \varepsilon+\xi_i^*$$

$$y_i - w^T\phi(x_i)-b_i \leq \varepsilon+\xi_i$$

$$\xi_i,\xi_i^* \geq 0, i=1,\ldots,N$$

Regression SVM Type 2

For this *SVM* model, the error function is given by:

$$\frac{1}{2}w^T w - C\left(v\varepsilon+\frac{1}{N}\sum_{i=1}^{N}\left(\xi_i+\xi_i^*\right)\right)$$

which we minimize subject to:

$$\left(w^T\phi(x_i)+b\right)-y_i \leq \varepsilon+\xi_i$$

$$y_i-\left(w^T\phi(x_i)+b_i\right)\leq \varepsilon+\xi^*_i$$

$$\xi_i,\xi^*_i \geq 0, i=1,\ldots,N,\varepsilon\geq 0$$

Cross-Validation

- *K* can be regarded as one of the most important factors of the model that can strongly influence the quality of predictions.
- There should be an optimal value for *K* that achieves the right trade off between the bias and the variance of the model.
- *STATISTICA KNN* can provide an estimate of *K* using an algorithm known as "Cross-validation" .

Cross-Validation

- Cross-validation is a well established technique that can be used to obtain estimates of model parameters that are unknown. Here we discuss the applicability of this technique to estimating *K*.
- The general idea of this method is to divide the data sample into a number of *v* folds (randomly drawn, disjointed sub-samples or segments).
  - For a fixed value of *K*, we apply the *KNN* model to make predictions on the *v*th segment (i.e., use the *v-1* segments as the examples) and evaluate the error.
    The most common choice for this error for regression is sum-of-squared and for classification it is most conveniently defined as the accuracy (the percentage of correctly classified cases).
  - This process is then successively applied to all possible choices of *v*. At the end of the v folds (cycles), the computed errors are averaged to yield a measure of the stability of the model (how well the model predicts query points).
  - The above steps are then repeated for various *K* and the value achieving the lowest error (or the highest classification accuracy) is then selected as the optimal value for *K* (optimal in a cross-validation sense).

Note that cross-validation is computationally expensive and you should be prepared to let the algorithm run for some time especially when the size of the examples sample is large.

Association Rule.

- The goal of the ***Association rule*** is to detect relationships or associations among a large set of data items.
- It is an important data mining model studied extensively by the database and data mining community.
- Assume all data are categorical.
- Initially used for Market Basket Analysis to find how items purchased by customers are related.
- The discovery of such association rule can help people to develop marketing strategies by gaining insight into, which items are frequently purchased together by customer.

Cluster Analysis.

- The process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.
- Clustering is an example of unsupervised learning, where the learning do not rely on predefined classes and class labeled training examples.
- For the above reason , Clustering is the form of “Learning by observation” , rather than learning by “Example”.

Area of Application.

- Market Research.

  Clustering can help marketers discover distinct groups in their customer bases and characterize customer groups based on purchasing patterns.
- Biology.

  Biologist can use cluster to discover distinct groups of species depending on some useful parameters.

- ***k*-Means clustering.** The basic operation of this algorithm is relatively simple: Given a fixed number of (desired or hypothesized) k clusters, assign observations to those clusters so that the means across clusters (for all variables) are as different from each other as possible.
- **Extensions and generalizations.** The methods implemented in the *Generalized EM and k-Means Cluster Analysis* module of *STATISTICA* extend this basic approach to clustering in three important ways:
- Instead of assigning cases or observations to clusters so as to maximize the differences in means for continuous variables, the EM (expectation maximization) clustering algorithm rather computes probabilities of cluster memberships based on one or more probability distributions. The goal of the clustering algorithm is to maximize the overall probability or likelihood of the data, given the (final) clusters.
- Unlike the classic implementation of *k*-Means clustering in the *Cluster Analysis* module, the *k*-Means and EM algorithms in the *Generalized EM and k-Means Cluster Analysis* module then can be applied to both continuous and categorical variables.
- A major shortcoming of *k*-Means clustering has been that you need to specify the number of clusters before starting the analysis (i.e., the number of clusters must be known *a priori*); the *Generalized EM and k-Means Cluster Analysis* module uses a modified v-fold cross-validation scheme , to determine the best number of clusters from the data. This extension makes the *Generalized EM and k-Means Cluster Analysis* module an extremely useful data mining tool for unsupervised learning and pattern recognition.

Asia-Pacific Journal of
Public Health
Volume 22 Number 1
January 2010 51-62

10.1177/1010539509352024
http://aph.sagepub.com
hosted at
http://online.sagepub.com

Socioeconomic and Demographic Factors Associated With Health Care Choices in Taiwan

Chieh-Yu Liu, PhD, and Jih-Shin Liu, MSc

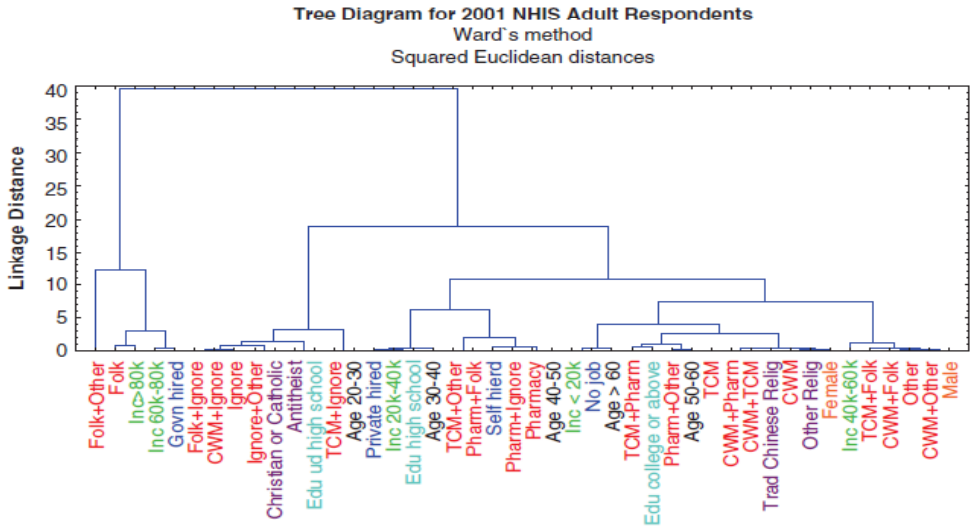


Figure 1. The dendrogram of hierarchical cluster analysis.

Expert Systems with Applications 38 (2011) 1400–1404

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Mining the optimal clustering of people's characteristics of health care choices

Chieh-Yu Liu [a,b,*], Jih-Shin Liu [c]

[a] Department of Nursing, National Taipei College of Nursing, 365, Min-der Rd., Bei-Tou District, Taipei City 112, Taiwan, ROC
[b] Department of Health Care Management, National Taipei College of Nursing, 365, Min-der Rd., Bei-Tou District, Taipei City 112, Taiwan, ROC
[c] Division of Biostatistics and Bioinformatics, National Health Research Institutes (NHRI), 35, Keyan Rd., Zhunan Town, Mioli County 350, Taiwan, ROC

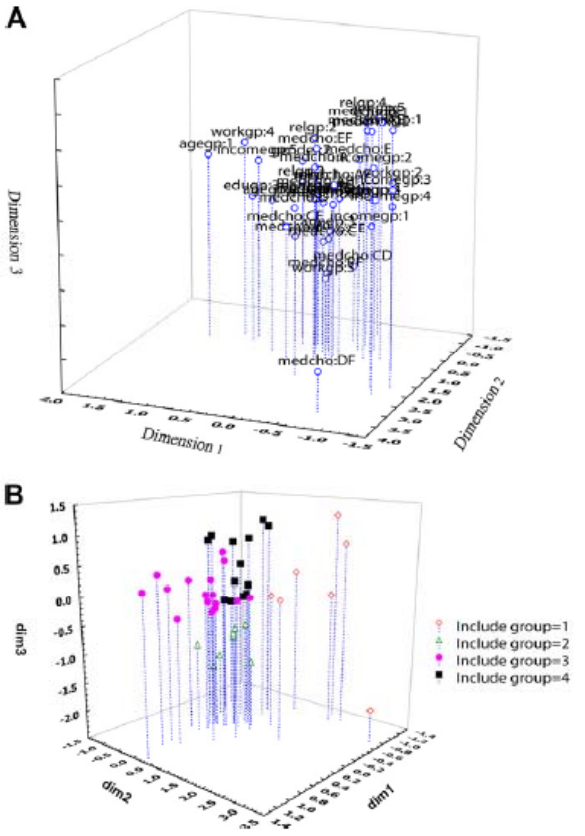
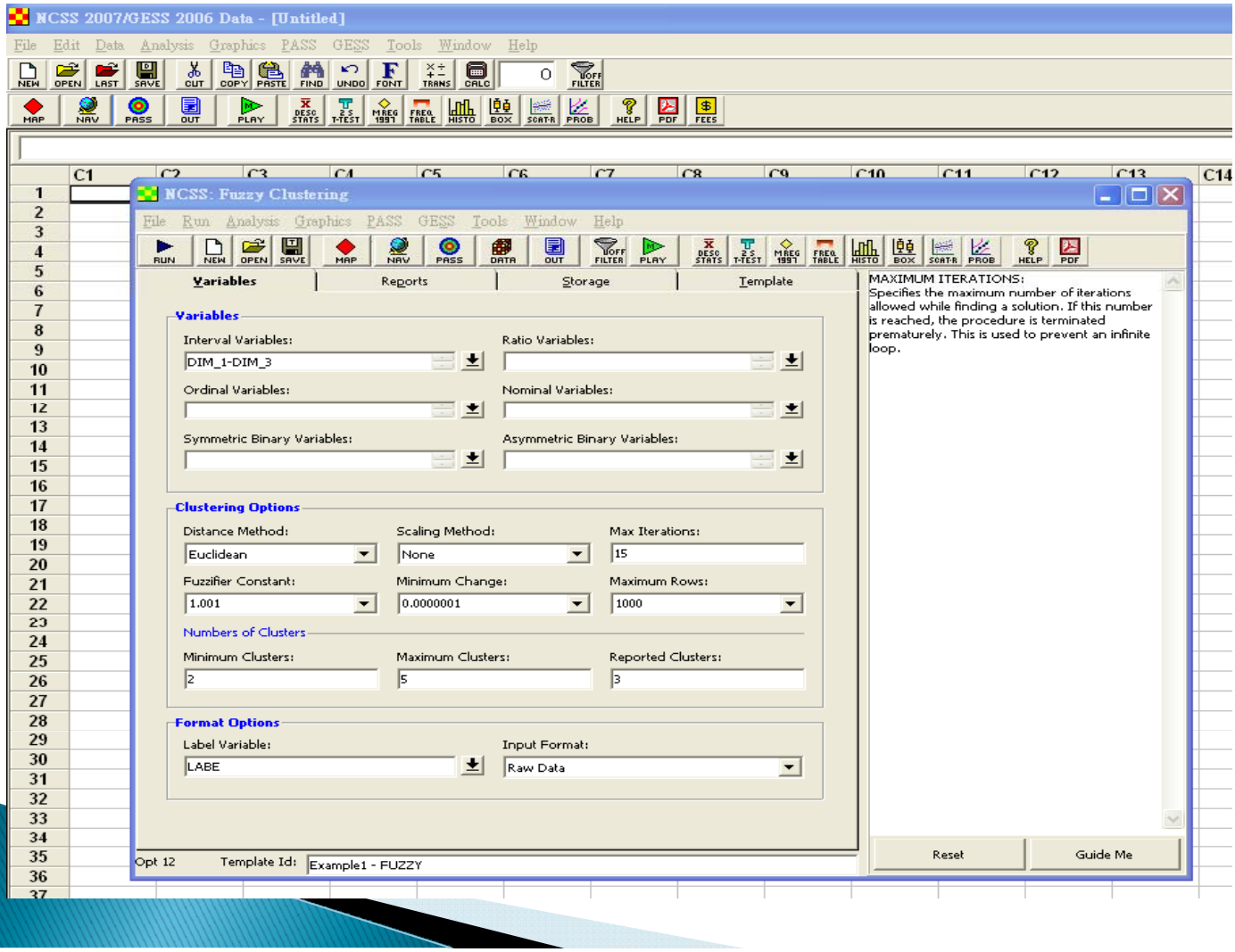
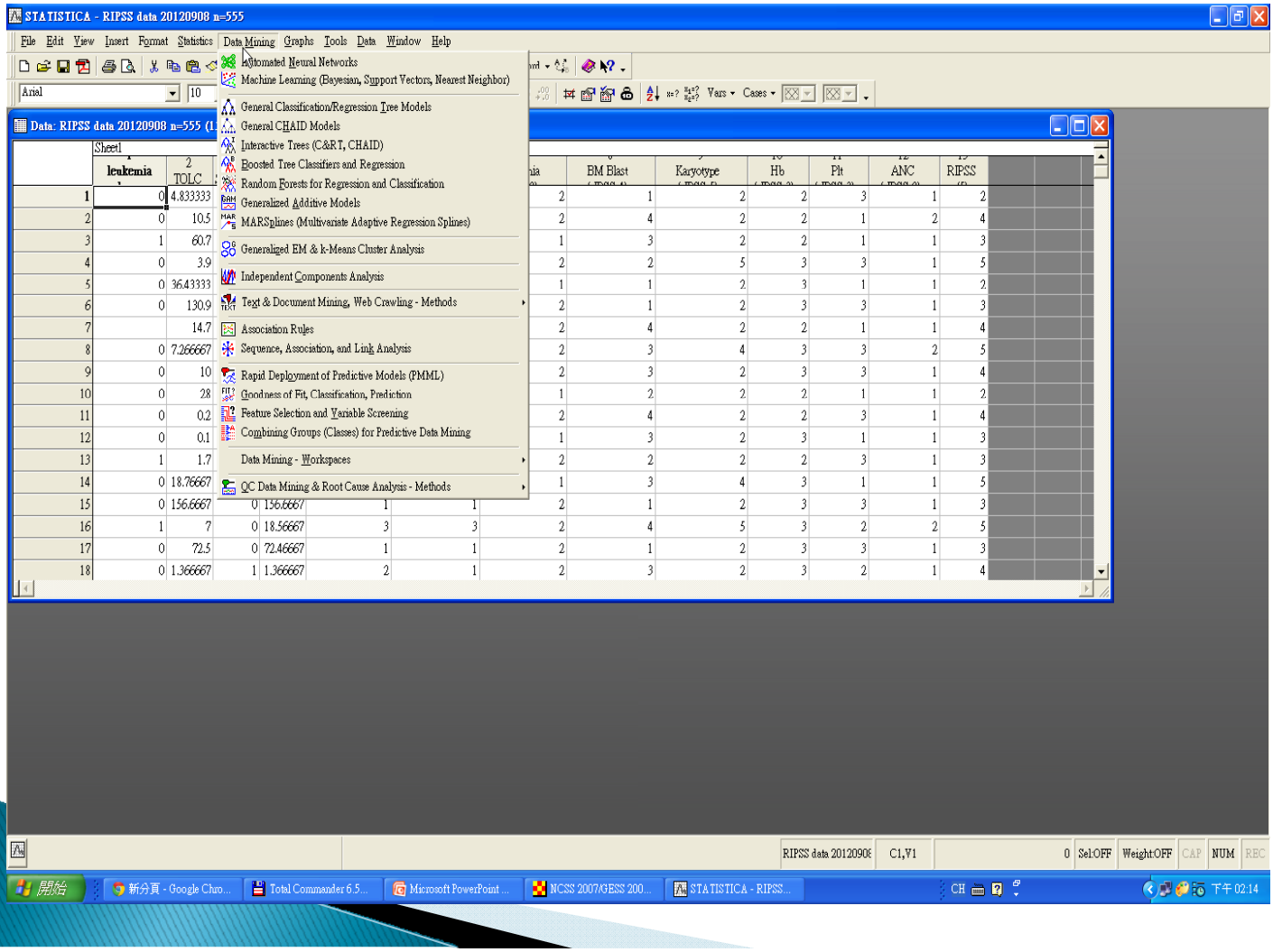


Fig. 2. (A) 3-Dimensional plot of coordinates of socioeconomic and demographic variables and people's health care choices. (B) The resulted clustering derived from the novel methodology proposed in this paper.

Table 3
The results of k-means cluster analysis with v-fold cross-validation.

Variable	Label	Group
workgp:1	Government-hired	1
incomegp:1	≧NTD$ 80k	1
incomegp:2	NTD$ 60–80k	1
incomegp:3	NTD$ 40–60k	1
medcho:BD	TCM + Folk med	1
medcho:D	Folk med	1
medcho:DF	Folk med + Others	1
agegp:3	Age 40–50	2
workgp:3	Self-hired	2
medcho:C	Pharmacy	2
medcho:CE	Pharmacy + Ignore	2
medcho:CF	Pharmacy + Others	2
medcho:CD	Pharmacy + Folk med	2
medcho:BC	TCM + Pharmacy	2
medcho:BF	TCM + Others	2
gender:2	Female	3
agegp:1	Age ≧ 60	3
agegp:2	Age 50–60	3
edugp:3	Under high school	3
relgp:1	Chinese traditional	3
relgp:2	Christian-related	3
relgp:3	Atheists	3
workgp:4	No job	3
incomegp:5	≦NTD$20, 000	3
medcho:A	Western med	3
medcho:F	Others	3
medcho:AC	Western + Pharmacy	3
medcho:AD	Western + Folk med	3
medcho:AB	Western + TCM	3
medcho:B	TCM	3
medcho:EF	Ignore + Others	3
gender:1	Male	4
agegp:4	Age 30–40	4
agegp:5	Age 20–30	4
edugp:1	College or above	4
edugp:2	High school	4
relgp:4	Other religions	4
workgp:2	Private-hired	4
incomegp:4	NTD$20–40 k	4
medcho:E	Ignore	4
medcho:AE	Western + Ignore	4
medcho:BE	TCM + Ignore	4
medcho:AF	Western + Others	4
medcho:DE	Folf med + Ignore	4


NCSS 2007/GESS 2006 Data - [Untitled]
File Edit Data Analysis Graphics PASS GESS Tools Window Help
NCSS: Fuzzy Clustering
File Run Analysis Graphics PASS GESS Tools Window Help
Variables
Reports
Storage
Template
Variables
Interval Variables:
DIM_1-DIM_3
Ratio Variables:
Ordinal Variables:
Nominal Variables:
Symmetric Binary Variables:
Asymmetric Binary Variables:
Clustering Options
Distance Method:
Euclidean
Scaling Method:
None
Max Iterations:
15
Fuzzifier Constant:
1.001
Minimum Change:
0.0000001
Maximum Rows:
1000
Numbers of Clusters
Minimum Clusters:
2
Maximum Clusters:
5
Reported Clusters:
3
Format Options
Label Variable:
LABE
Input Format:
Raw Data
MAXIMUM ITERATIONS:
Specifies the maximum number of iterations allowed while finding a solution. If this number is reached, the procedure is terminated prematurely. This is used to prevent an infinite loop.
Opt 12
Template Id:
Example1 - FUZZY
Reset
Guide Me


STATISTICA - RIPSS data 20120908 n=555
File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Help
Automated Neural Networks
Machine Learning (Bayesian, Support Vectors, Nearest Neighbor)
General Classification/Regression Tree Models
General CHAID Models
Interactive Trees (C&RT, CHAID)
Boosted Tree Classifiers and Regression
Random Forests for Regression and Classification
Generalized Additive Models
MARSplines (Multivariate Adaptive Regression Splines)
Generalized EM & k-Means Cluster Analysis
Independent Components Analysis
Text & Document Mining, Web Crawling - Methods
Association Rules
Sequence, Association, and Link Analysis
Rapid Deployment of Predictive Models (PMML)
Goodness of Fit, Classification, Prediction
Feature Selection and Variable Screening
Combining Groups (Classes) for Predictive Data Mining
Data Mining - Workspaces
QC Data Mining & Root Cause Analysis - Methods
Data: RIPSS data 20120908 n=555
Sheet1
leukemia
TOLC
BM Blast
Karyotype
Hb
Plt
ANC
RIPSS
RIPSS data 20120908
C1,V1
Sel:OFF
Weight:OFF
CAP
NUM
REC
開始
新分頁 - Google Chro...
Total Commander 6.5...
Microsoft PowerPoint ...
NCSS 2007/GESS 200...
STATISTICA - RIPSS...
CH
下午 02:14

Thank you for your attention!!