Science Foundation Ireland

LERO Centre for Science, Engineering Technology

Work Package:
D1
Survey of data mining algorithms

Project Reference:	10/CE/I1855
Date:	May 6, 2013
IBM SoW Title:	Survey of data mining algorithms
Total Number of Pages:	28

Executive Summary

The deliverable corresponds to a talk given by Sandra Buda (PhD student in PEL) in november 2011 on data mining technics: data classification, regression, data clustering, summarisation, etc.

A Brief Introduction to Data Mining

Teodora Sandra Buda

http://pel.ucd.ie

To be discussed:

- Following articles:
  - From Data Mining to Knowledge Discovery in Databases, by U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, published in *AI Magazine* 1996 – Cited by 4021;
  - Statistical Modeling: The two cultures, by L. Breiman published in *Statistical Science* 2001 – Cited by 586;
  - Data Mining: An overview from a Database Perspective, by MS Chen, J. Han, P.S. Yu, published in *IEEE Transactions on Knowledge and Data Engineering* 1996 - Cited by 1851;

Data Mining

• **Data mining** means a process of nontrivial **extraction** of implicit, previously unknown and potentially useful **information** (such as knowledge rules, constraints, regularities) from observed data.

•The data-mining component of the KDD process often **involves**:
- repeated iterative application of particular data-mining methods;
- fitting models to observed data;
- determining patterns from observed data;

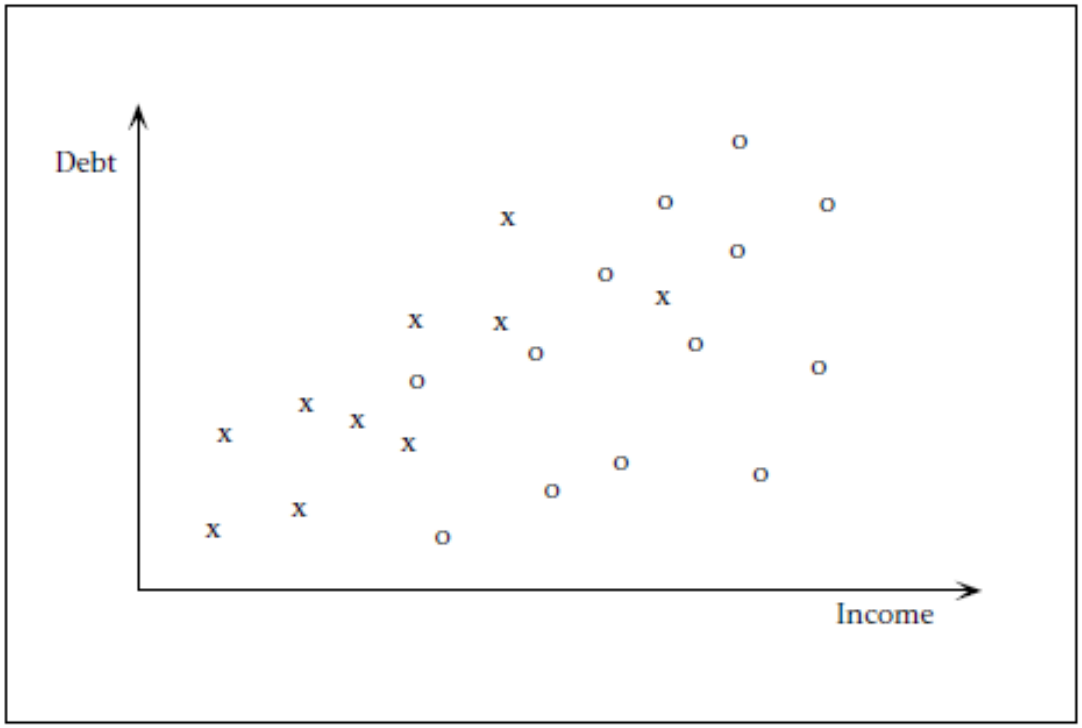
• **Goals**:
- **Prediction**. To be able to predict what the responses are going to be to future input variables;
- **Information**. To extract some information about how nature is associating the response variables to the input variables.

Data Mining Methods

- Data classification
- Regression
- Data clustering
- Summarization
- Dependency modeling
- Change and deviation detection
- Decision Trees and Rules
- Nonlinear Regression and Classification Methods
- Example-Based Methods

- Tools associated with database system products:
  - Data generalization and summarization tools, also referred as on-line analytical processing(OLAP), multiple-dimensional databases, data cubes, data abstraction, generalization, summarization, characterization, etc.

A Simple Data Set with 2 Classes



Data Classification

• Classification is learning a function that maps (classifies) a data item into one of several predefined classes.

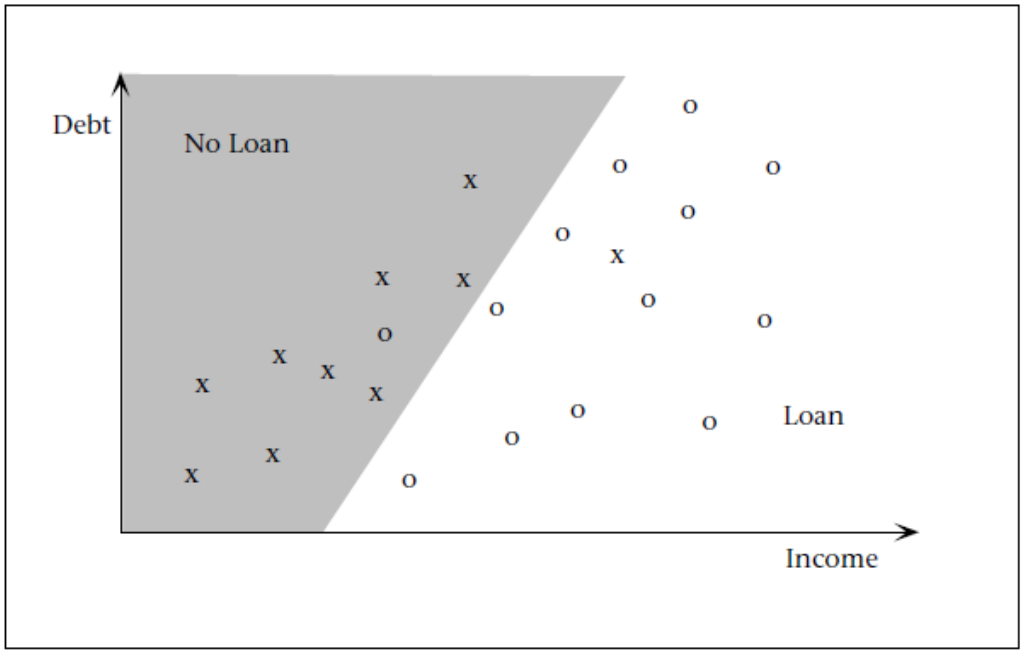


Fig. 2. A simple linear classification boundary for the loan data set

Regression

• Regression is learning a function that maps data item to a real-valued prediction variable.

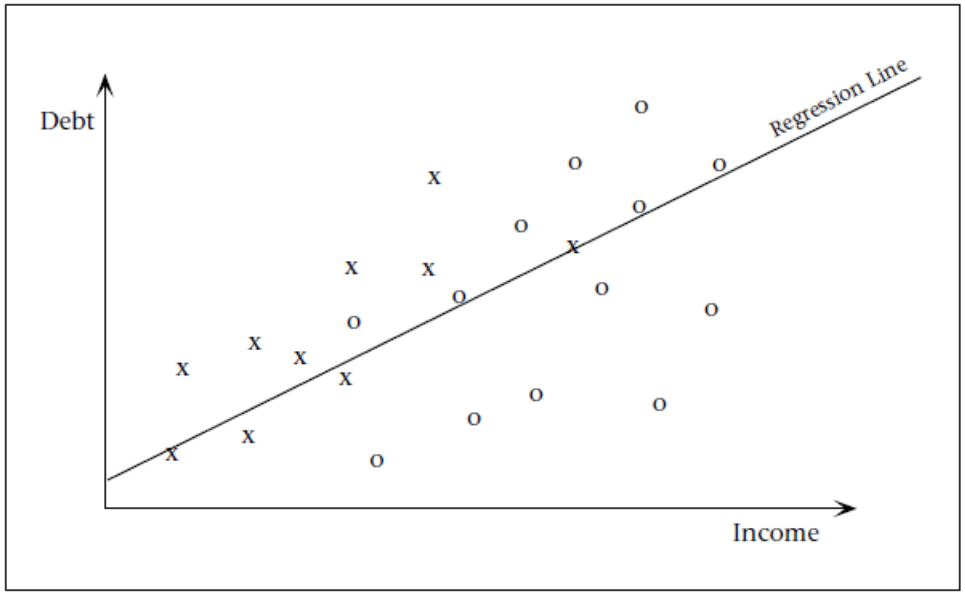


Fig. 3. A Simple Linear Regression for the Loan Data Set

Clustering

• Clustering is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data

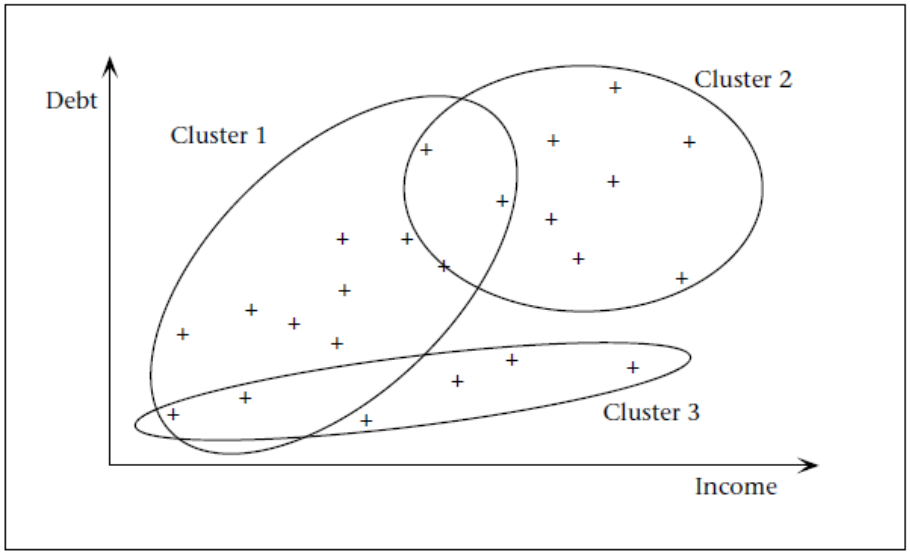


Fig. 4. A Simple Clustering of the Loan Data Set into Three Clusters. Note that original labels are replaced by a +.

Decision Trees and Rules

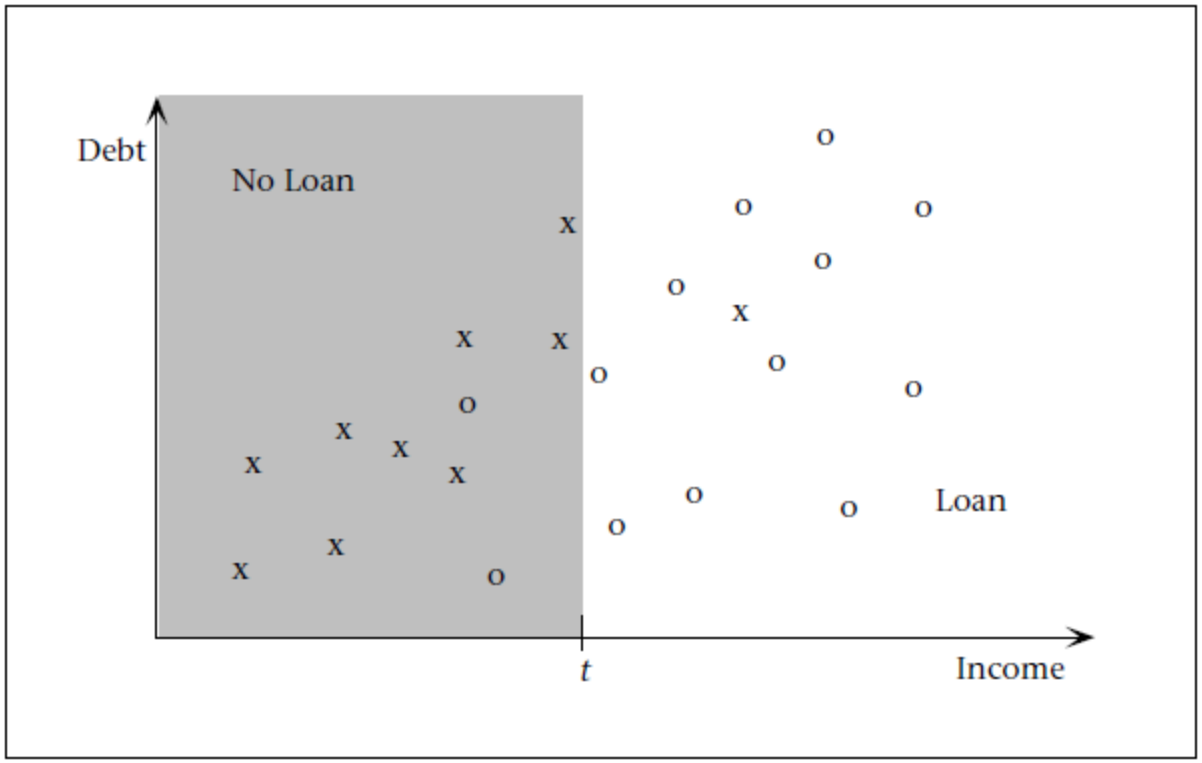


Fig. 5. Using a Single Threshold on the Income Variable to try to Classify the Loan Data Set.

Nonlinear Regression and Classification Methods

• These methods consist of a family of techniques for prediction that fit linear and nonlinear combinations of basis functions (sigmoids, splines, polynomials) to combinations of the input variables;

• Examples include: feedforward, neural networks, adaptive spline methods, and projection pursuit regression;

• Drawback: difficult to interpret;

Nonlinear Regression and Classification Methods

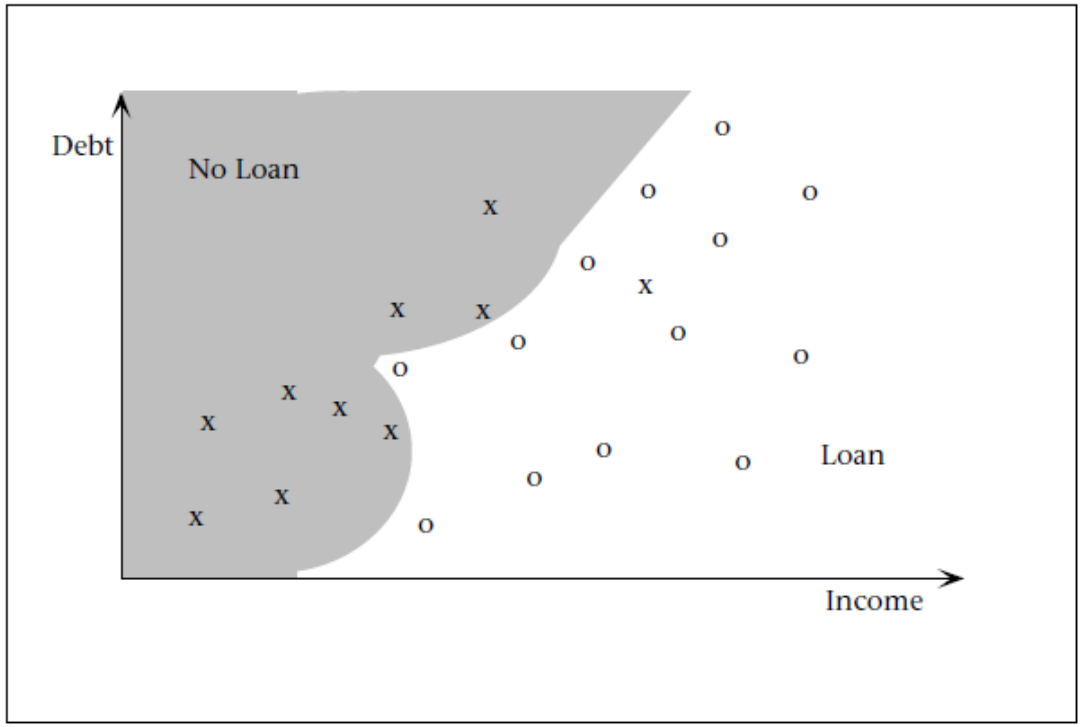


Fig. 6. An Example of Classification Boundaries Learned by a Nonlinear Classifier (Such as a Neural Network) for the Loan Data Set.

Example-Based Methods

•Use representative examples from the database to approximate a model; (predictions on new examples are derived from the properties of similar examples in the model whose prediction is known.)

•Examples: nearest neighbor classification and regression algorithms and case-based reasoning systems;

• Drawback: a well-defined distance metric for evaluating the distance between data points is required;

Example-Based Methods

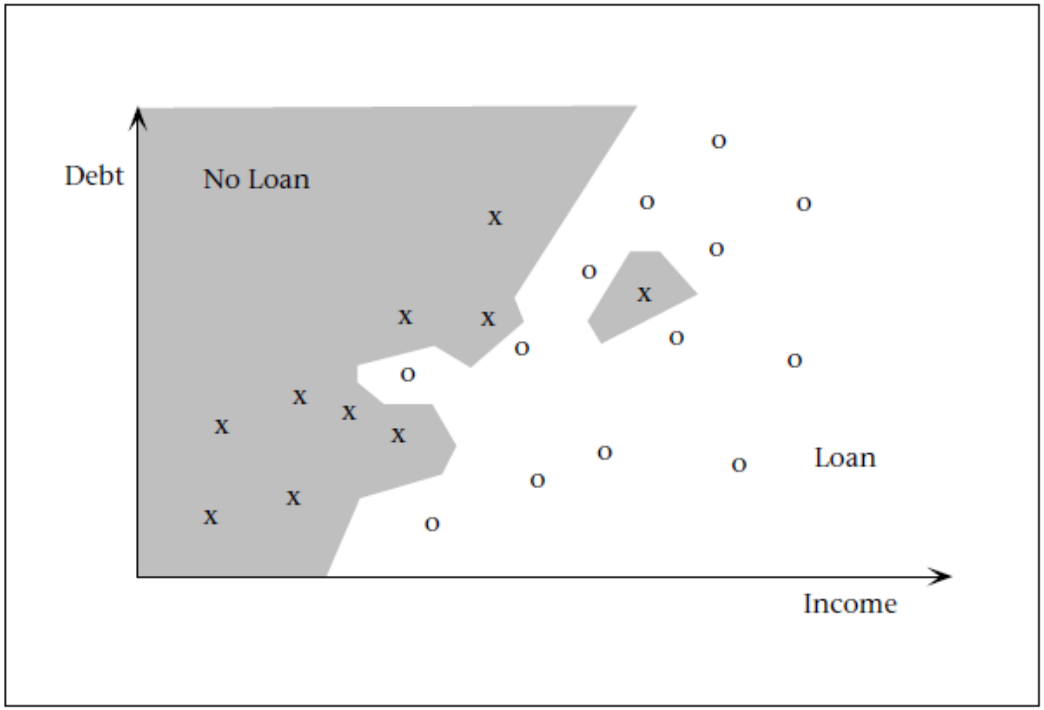
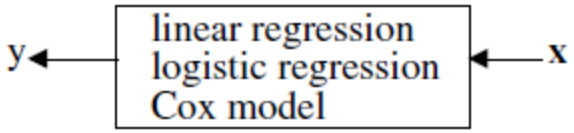


Fig. 7. Classification Boundaries for a Nearest-Neighbor Classifier for the Loan Data Set.

Data Modelling vs. Algortihmic Modelling

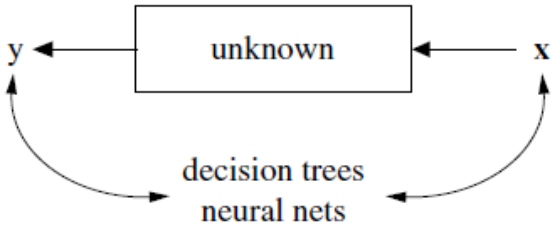
The Data Modeling Culture



- The analysis in this culture starts with assuming a stochastic data model for the inside of the black box.
- Model validation. Yes–no using goodness-of-fit tests and residual examination. (98% statisticians)

The Algorithmic Modeling Culture

- The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function f(x)—an algorithm that operates on x to predict the responses y.
- Model validation. Measured by predictive accuracy. (2% by statisticians)



Limitations of Data Modelling

- "If all a man has is a hammer, then every problem looks like a nail."

- More complicated data models are appearing in current published applications(Bayesian methods combined with Markov Chain Monte Carlo used for modeling the data)

- Drawback:
  - more complex data models;
  - losing the advantage of presenting a simple and clear picture of nature's mechanism.
  - restricts the ability to deal with a wide range of statistical problems;

Interesting studies of Algorithmic Modelling

•Rashomon: the multiplicity of good models; (a multitude of different descriptions [equations f(x)] in a class of functions giving about the same minimum error rate.)

•Occam: the conflict between simplicity and accuracy;

•Bellman: dimensionality—curse or blessing?

Decision Trees

• A decision-tree-based classification method(influential in machine learning studies) involves:

  • Choosing a subset of the training examples (a window) to form a decision tree in which each leaf carries a class name, and each interior node specifies an attribute with a branch corresponding to each possible value of the attribute; (a selection of the exceptions)

• The quality (function) of a tree depends on both the classification accuracy and the size of the tree

Growing Forests for Prediction. Comparison.

• Compare the performance of single trees (CART) to random forests on a number of small and large data sets,

• For the five smaller data sets above the line, the test set error was estimated by leaving out a random 10% of the data, then running CART and the forest on the other 90%. The left-out 10% was run down the tree and the forest and the error on this 10% computed for both. This was repeated 100 times and the errors averaged.

•The larger data sets below the line came with a separate test set.

Growing Forests for Prediction. Comparison.

Data set	Forest	Single tree
Breast cancer	2.9	5.9
Ionosphere	5.5	11.2
Diabetes	24.2	25.3
Glass	22.0	30.4
Soybean	5.7	8.6
Letters	3.4	12.4
Satellite	8.6	14.8
Shuttle $\times 10^3$	7.0	62.0
DNA	3.9	6.2
Digit	6.2	17.1

Fig. 8. Test set misclassification error (%)

Mining Association Rules. Algorithm Apriori

• Given a database of sales transactions, discover the important associations among items such that the presence of some items in a transaction will imply the presence of other items in the same transaction.

Database D

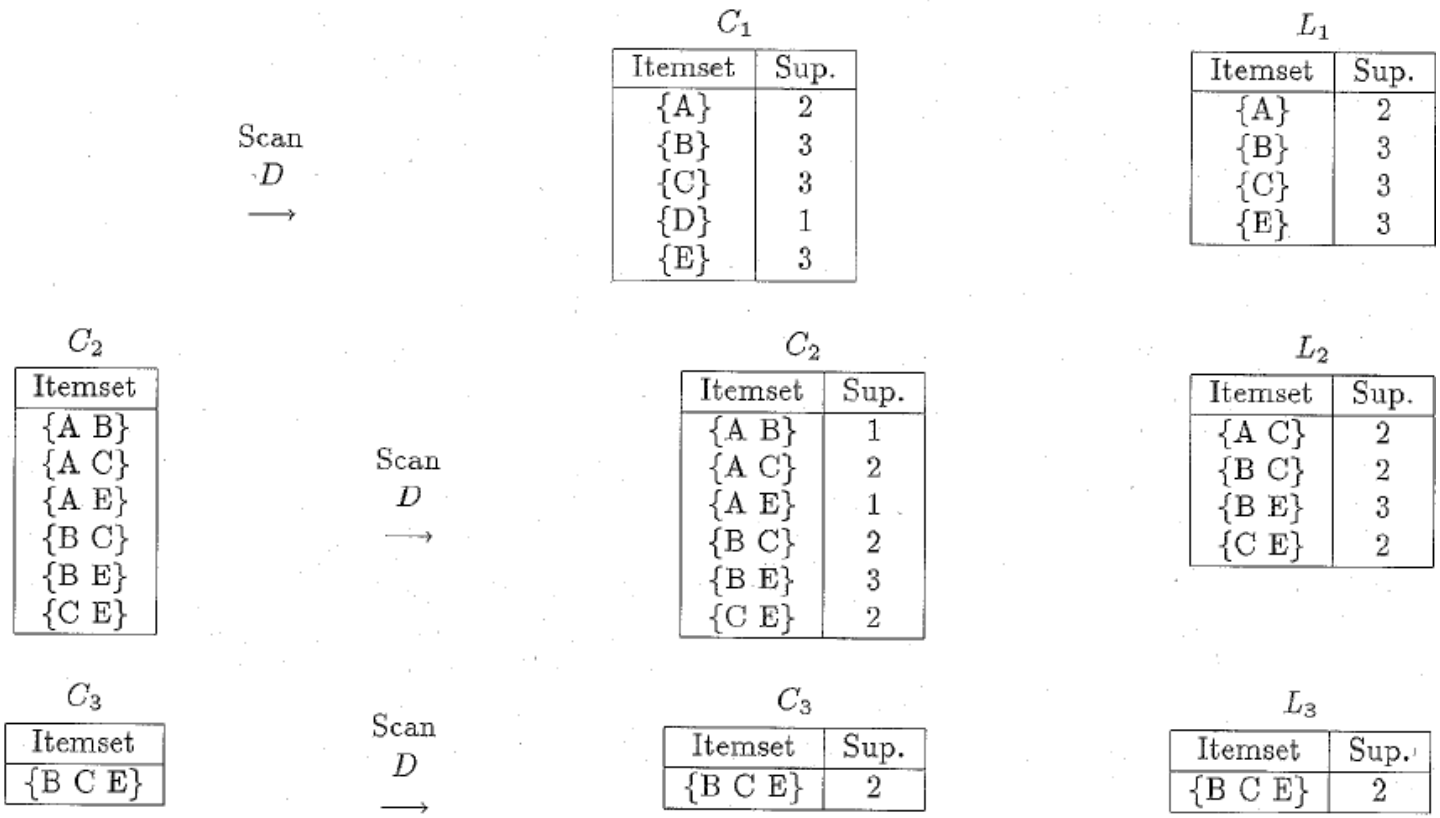
TID	Items
100	A C D
200	B C E
300	A B C E
400	B E

Fig. 9. An example transaction database for data mining.

•*k-itemsets*

• Apriori simply scans all the transactions to count the number of occurrences for each item.

Algortihm Apriori



Interestingness of Discovered Association Rules

•Survey in a school of 5,000 students.
•The data show that
- 60% of students (i.e., 3,000 students) play basketball,
- 75% of students (i.e., 3,750 students) eat cereal
- 40% of them (i.e., 2,000 students) both play basketball and eat cereal.

Discovering association with the minimal student support is 2,000 and the minimal confidence is 60%: “(play basketball) => (eat cereal)"
since this rule contains the minimal student support and the corresponding confidence(2000/3000 = 0:66).
⇒Misleading since the overall percentage of students eating cereal is 75% even larger than 66%;
=>Playing basketball and eating cereals are in fact **negatively associated**. Being involved in one **decreases** the likelihood of being involved in the other;

Data Cube Approach(Multilevel Data Generalization, Summarization, and Characterization

- Also called as “multidimensional databases“, “materialized views“ and “OLAP (On-Line Analytical Processing)“;

- Idea:
  - **materialize certain expensive computations** that are frequently inquire (egg. aggregate functions, such as count, sum, average, max, etc.)
  - store the materialized views in a multi-dimensional database (called a “data cube") for decision support, knowledge discovery, etc.

- Operations: “roll-up” or “drill-down”;

Data Cube Approach(Multilevel Data Generalization, Summarization, and Characterization

- Example:
  - The schema “sales(part; supplier; customer; sale price)" can be materialized into a set of eight views where psc indicates a view consisting of aggregate function values (such as total sales) computed by grouping three attributes part, supplier, and customer

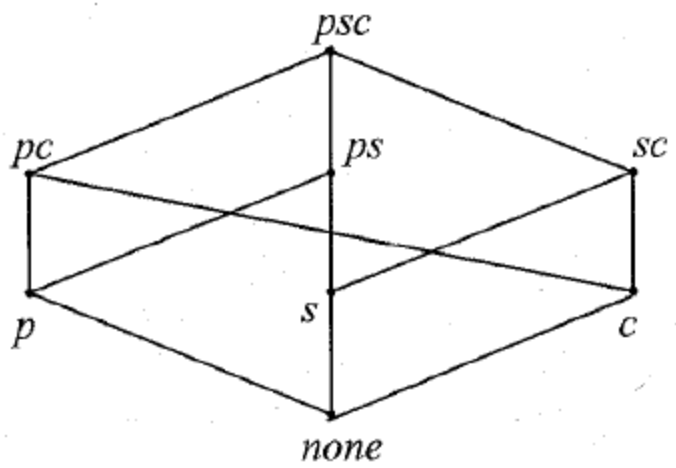


Fig. 10. Eight views of data cubes for sales information

Further readings

- Data Mining and Machine Oriented Modeling: A Granular Computing Approach, by Tsau Young Lin. published in *Applied Intelligence*, 2000;

- Model uncertainty, Data Mining and Statistical Inference, by Chris Chatfield, published in *J.R. Statistical Society,* 1995;

- A Survey of Data Mining and Knowledge Discovery Software Tools, by M. Goebel, L. Gruenwald, published in *ACM SIGKDD,* 1999;

Thank you!

Questions?