Overview of Data Mining

1. Introduction

The fast developing computer science and engineering techniques has made the information easy to capture process and store in databases. Information/data are considered as elementary variable facts. Knowledge is considered as a set of instructions, which describes how these facts can be interpreted and use. Data describes the actual state of world; however knowledge describes the structure of the world and consists of principal and laws. How to gather, store and retrieve data is considered in database theory. In the knowledge engineering, in the center of interest there is knowledge and methods of its formalization and gaining are studied. Data mining can be viewed as a result of the natural evolution of information technology. An evolutionary path has been witnessed in the database industry in the development of the following functionalities (Image. 1): data collection and database creation, data management (including data storage and retrieval, and database transaction processing), and data analysis and understanding (involving data warehousing and data mining).

Since the 1960's, database and information technology has been evolving systematically from primitive file processing systems to sophisticated and powerful databases systems. The research and development in database systems since the 1970's has led to the development of relational database systems, data modeling tools, and data organization techniques. In addition, users gained convenient data access through query languages, query processing, and user interfaces. Efficient methods for on-line transaction processing (OLTP), where a query is viewed as a read-only transaction, have contributed substantially to the evolution and wide acceptance of relational technology as a major tool for efficient storage, retrieval, and management of large amounts of data. Figure 1 presents the evolution phase of databases to mining these databases for extracting knowledge from the same.

Database technology since the mid-1980s has been characterized by the popular adoption of relational technology and an upsurge of research and development activities on new and powerful database systems. These employ advanced data models such as extended-relational, object-oriented, object-relational, and deductive models Application-oriented database systems, including spatial, temporal, multimedia, active, and scientific databases, knowledge bases, and office information bases, have flourished. Issues related to the distribution, diversification, and sharing of data have been studied extensively.

Heterogeneous database systems and Internet-based global information systems such as the World-Wide Web (WWW) also emerged and play a vital role in the information industry.

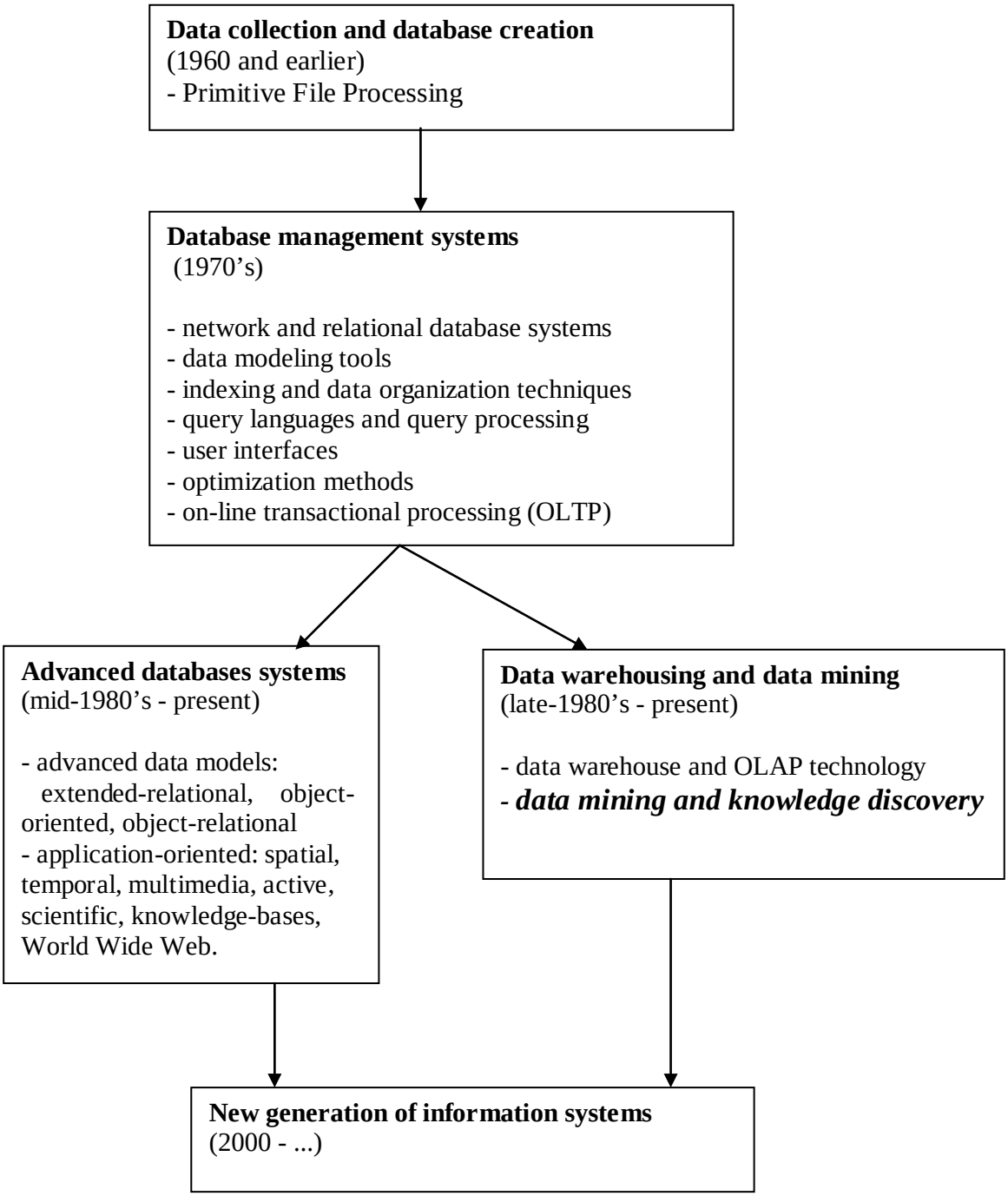

Image 1: The evolution of database technology

2. Knowledge Discovery Process

There is huge gap from the stored data to the knowledge that could be constructed from the data, that's where data mining comes into picture. Knowledge Discovery in Databases (KDD) refers to the overall process of discovering useful patterns from the data. Data mining is major step in KDD process and at times synonym to KDD (Image. 2).

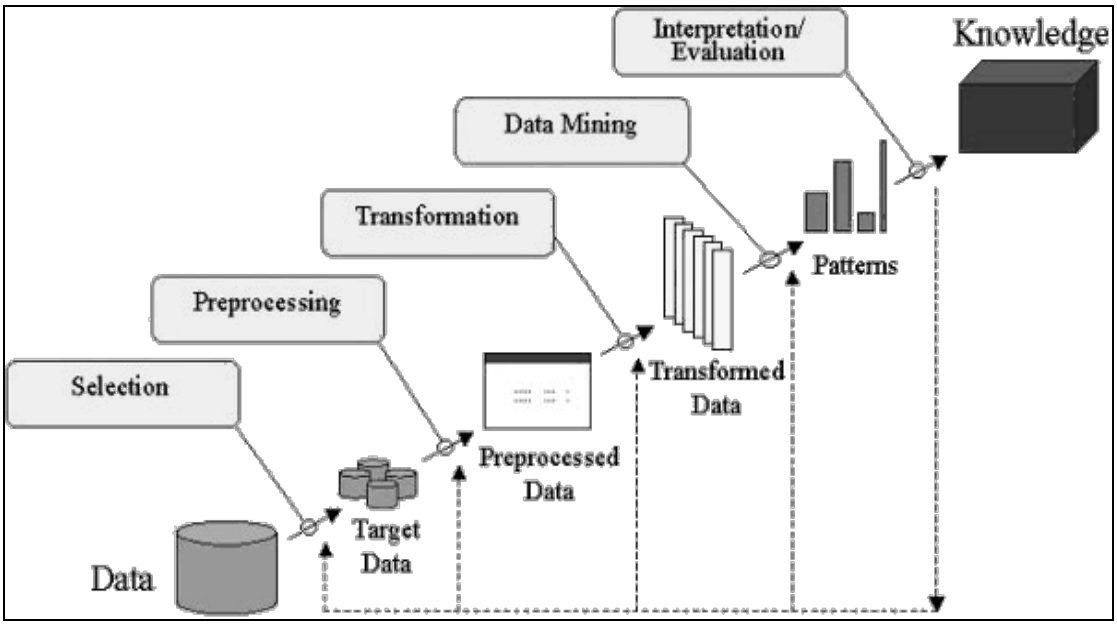

Image 2: KDD/ Data Mining Architecture

2.1. Knowledge Discovery Process Steps

The process of knowledge discovery using data mining can be divided into defined steps presented in Image 2.

i. **Selection**: This step involves identification or extraction of relevant data for analysis.

ii. **Preprocessing**: This involves preparing/ cleaning the data set by resolving problems like missing data, skewed data, irrelevant fields, removal of outlying points, format conversion etc. This step might consists of following operations that need to be performed before a data mining technique is applied:

- **Data cleaning** – It consist of some basic operations like normalization, noise removal and handling of missing or inconsistent data. Data from real world sources are often erroneous, incomplete and inconsistent, may be due to operational errors or implementation flaws.
- **Data integration** – This includes integrating multiple, heterogeneous datasets generated from different sources.
- **Transformation** – consolidation of data into the form appropriate for mining e.g. performing aggregation or summary of data
- **Reduction** – This includes finding useful features to represent the data and using dimensionality reduction, feature discretization, and feature extraction/ transformation methods.

iii. **Data Mining**: This step involves application of knowledge discovery algorithms to the cleaned, transformed data in order to extract meaningful patterns from the data.

iv. **Pattern evaluation**: This step involves evaluation of pattern for interestingness. One can evaluate the mined patterns automatically or semi automatically to identify the truly interesting or useful patterns for the user.

v. **Knowledge presentation and Interpretation:** This involves representation of discovered knowledge in proper format.

3. Data Mining

Data mining is defined as a non trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. The term *process* implies that data mining consists of many steps, *non trivial* means process is not straight forward and some search or inference is involved. The term *Pattern* is an expression in some language describing a subset of data, finding structures from data, or, in general making any high level description of a set of data. Pattern should be *novel* and *potentially useful*, that is, it should lead to some benefits to the user or task. Ultimately pattern should be *understandable*, if not immediately then at a later stage after some post processing.

Data mining is a highly inter disciplinary area spanning a range of disciplines; statistics, machine learning, databases, pattern recognition and other areas (Image 3).

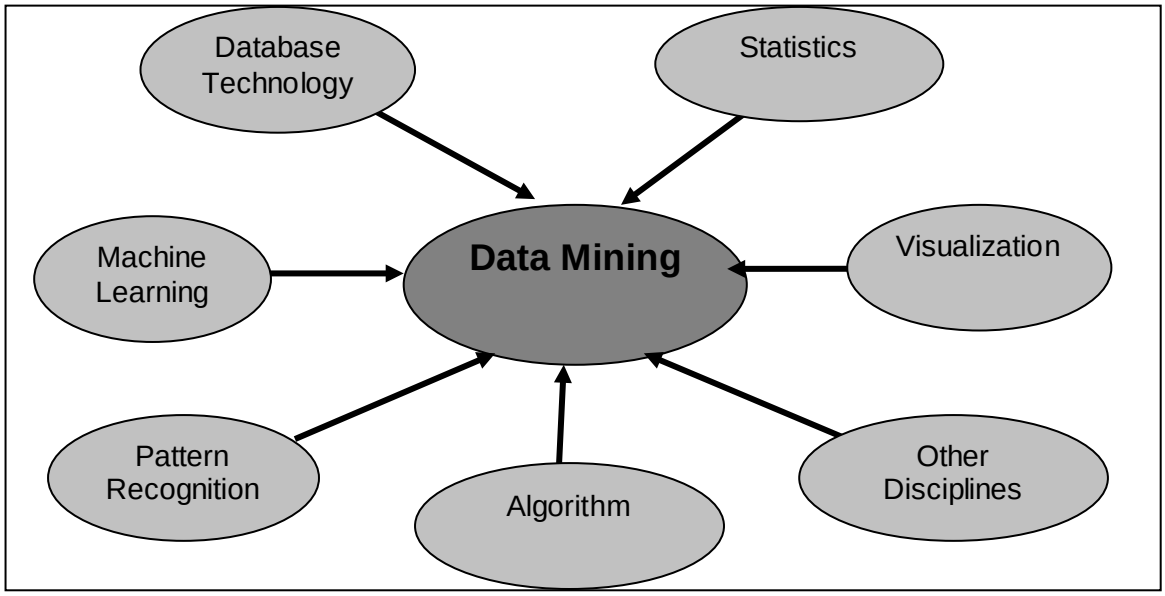

Image 3: Data mining as a confluence of multiple disciplines

All of these fields are concerned with certain aspects of data analysis, so they have much in common but each also has its own distinct flavor. Thus methods from these disciplines are welcome in data mining in their capacity to do the job. However the focus is different in various disciplines. In machine learning and statistics the stress is on the consistency of the algorithm, however in data mining it is the consistency of pattern that matters the most.

4. DATA MINING METHODS

In general, data mining methods can be classified into two categories: predictive and descriptive. Predictive data mining methods predicts the values of data, using some already known results that have been found using a different set of data. Predictive data mining tasks include: Classification, Prediction. Descriptive mining tasks characterize the general properties of the data in database. This is done by identifying the patterns and relationships in the data. These models are not based on any underlying theory or mechanism through which the data arose rather they are simply a description of the observed data. Descriptive data mining tasks include: Clustering, Association analysis and Summarization. These data mining methods are briefly discussed here.

4.1. Classification and Prediction

Classification is the process of finding a set of models that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known). There are many methods to present the derived model such as classification (IF-Then) rules, decision trees, mathematical formulae or neural networks. A decision tree is a flowchart likes tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can be easily converted to classification rules. A neural network is a collection of linear threshold units that can be trained to distinguish objects of different classes.

Classification can be used for predicting the class label of data objects. However, in many applications, one may like to predict some missing or unavailable data values rather than class labels. This is usually the case when the predicted values are numerical data, and is often specifically referred to as prediction. Although prediction may refer to both data value prediction and class label prediction, it is usually confined to data value prediction and thus is distinct from classification. Prediction also encompasses the identification of distribution trends based on the available data.

4.2. Association Rule Mining

Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used in market basket or transaction data analysis. This process analyzes customer buying habits by finding associations between the different items that customer purchases on a given tour to store. The discovery of such associations can help in planning marketing/ advertising strategies and catalog design. For example, if customers are buying milk, how likely are they to also buy bread on the same trip to the market? For example by placing milk and bread together may further encourage the sale of these items together within single visit to store. Formally association rules are of the form $X \Rightarrow Y$, means objects satisfy X are also likely to satisfy Y. For example $computer \Rightarrow printer$ [support=2%, confidence=60%]. Support of 2% means that 2% of all the transaction under analysis show that computer and printer are purchased together. 60% confidence means that 60% of the customer who purchased a computer also bought the printer.

Algorithm Apriori and its variant are used to find association in the large databases.

4.3. Clustering

Clustering maps the data items into clusters, where clusters are natural grouping of data items based on similarity or probability density methods. Unlike classification and prediction which analyzes class-label data objects, clustering analyzes data objects without class-labels and tries to generate such labels. In some cases, however, cluster analysis is only a useful starting point for other purposes, such as data summarization. Clustering algorithms are mainly classified into partitional and hierarchical methods.

The hierarchical clustering approach builds a tree of clusters. The root of this tree can be a cluster containing all the data. Then, branch by branch, the initial big cluster is split in sub-clusters, until a partition having the desired number of clusters is reached. In this case, the hierarchical clustering is referred to as divisive. Moreover, the root of the tree can also consist of a set of clusters, in which each cluster contains one and only one sample. Then, branch by branch, these clusters are merged together to form bigger clusters, until the desired number of clusters is obtained. In this case, the hierarchical clustering is referred to as agglomerative.

The partition technique recursively divides the data into non overlapping partitions. K-means is the popular clustering algorithm in this category. The k value refers to the number of clusters in which the data are partitioned. Clusters are represented by their centers. The basic idea is that each sample should be closer to the center of its own cluster. If this is not verified, then the partition is modified, until each sample is closer to the center of the cluster it belongs to. The distance function between samples plays an important role, since a sample can migrate from a cluster to another one based on the values provided by the distance function.

4.4. Outlier Analysis

A database may contain data objects that do not comply with the general behavior or model of the data. These objects are called outliers. The analysis of outlier data is referred to as outlier mining. Most data mining methods discard outliers as noise or exception. However, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring ones. In addition it is useful in customized marketing for identifying the spending behavior of customers with exceptional high and low income, or in medical analysis for finding unusual responses to various medical treatments.

5. DATA MINING USES/ APPLICATIONS

Data mining techniques are now being applied to all kinds of domain, which are rich in data. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. For example, the insurance and banking industries use data mining applications to detect fraud and assist in risk assessment (e.g., credit scoring). Using customer data collected over several years, companies can develop models that predict whether a customer is a good credit risk, or whether an accident claim may be fraudulent and should be investigated more closely. The medical community sometimes uses data mining to help predict the effectiveness of a procedure or medicine. Pharmaceutical firms use data mining of chemical compounds and genetic material to help guide research on new treatments for diseases. Retailers can use information collected through affinity programs (e.g., shoppers' club cards, frequent flyer

points, contests) to assess the effectiveness of product selection and placement decisions, coupon offers, and which products are often purchased together. Companies such as telephone service providers and music clubs can use data mining to create a "chum analysis," to assess which customers are likely to remain as subscribers and which ones are likely to switch to a competitor. It has been reported that data mining has helped the federal government recover millions of dollars in fraudulent Medicare payments. The Justice Department has been able to use data mining to assess crime patterns and adjust resource allotments accordingly. Similarly, the Department of Veterans Affairs has used data mining to help predict demographic changes in the constituency it serves so that it can better estimate its budgetary needs. Another example is the Federal Aviation Administration, which uses data mining to review plane crash data to recognize common defects and recommend precautionary measures. Recently, data mining has been increasingly cited as an important tool for homeland security efforts. Some observers suggest that data mining should be used as a means to identify terrorist activities, such as money transfers and communications, and to identify and track individual terrorists themselves, such as through travel and immigration records.

Data mining is widely applied to agricultural problems. For instance, the prediction of wine fermentation problems can be performed by using a k-means approach. Knowing in advance that the wine fermentation process could get stuck or be slow can help the enologist to correct it and ensure a good fermentation process. Weather forecasts can be improved using a k-nearest neighbor approach, where it is assumed that the climate during a certain year is similar to the one recorded in the past. The same data mining technique can also be used for estimating soil water parameters. Apples and other fruits are widely analyzed in agriculture before marketing. Apples running on conveyors can be checked by humans and the bad apples (the ones presenting defects) can be removed. The same task can be efficiently performed by the data mining tasks. This task uses X-ray images for checking the apple watercore. It is based on an artificial neural network which learns from a training set how to classify the X-ray images. Neural networks are also used for classifying sounds from animals such as pigs for checking the presence of diseases. Support vector machines can be used for recognizing animal sounds as well, such as sounds from birds. Other applications of data mining techniques include the detection of meat and bone meal in feedstuffs destined to farm animals, analysis of the effects of energy use in agriculture, it can be used to derive plant and animal taxonomies, characterization of diseases and varieties, in bioinformatics- categorization of genes with similar functionally and prediction of crop yield etc.

6. MAJOR ISSUES IN DATA MINING

- **Mining different kinds of knowledge in databases.**

  Since different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery tasks, including classification, prediction, cluster and association analysis.

- **Interactive mining of knowledge at multiple levels of abstraction.**

  Since it is difficult to know exactly what can be discovered within a database, the data mining process should be interactive. Interactive mining allows users to focus the search for patterns, providing and refining data mining requests based on returned results.

- **Presentation and visualization of data mining results.**

Discovered knowledge should be expressed in high-level languages, visual representations, or other expressive forms so that the knowledge can be easily understood and directly usable by humans.

- **Handling outlier or incomplete data.**

  The data stored in a database may have outliers | noise, exceptional cases, or incomplete data objects. System should be able to deal with these.

- **Pattern evaluation: the interestingness problem.**

  A data mining system can uncover thousands of patterns. Many of the patterns discovered may be uninteresting to the given user, representing common knowledge or lacking novelty. Several challenges remain regarding the development of techniques to assess the interestingness of discovered patterns, particularly with regard to subjective measures.

- **Efficiency and scalability of data mining algorithms.**

  To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable.

- **Mining information from heterogeneous databases and global information systems.**

  Local and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases. The discovery of knowledge from different sources of structured, semi-structured, or unstructured data with diverse data semantics poses great challenges to data mining. Data mining may help disclose high-level data regularities in multiple heterogeneous databases that are unlikely to be discovered by simple query systems and may improve information exchange and interoperability in heterogeneous databases.

7. DATA MINING RESOURCES/ SOFTWARES

Many data mining books contain introductions to various kinds of data mining systems and products. KDnuggets maintains an up-to-date list of data mining products at www.kdnuggets.com/companies/products.html and the related software at www.kdnuggets.com/software/index.html, respectively. Detailed information regarding specific data mining systems and products can be found by consulting the Web pages of the companies.

7.1. Software

- IBM Intelligent Miner: www.ibm.com/software/data/iminer
- Microsoft SQL Server : www.microsoft.com/sql
- Oracle Data Mining (ODM): www.oracle.com
- SPSS (Clementine): www.spss.com/clementine
- SAS Miner: www.sas.com/technologies/datamining/miner
- Enterprise Miner: www.insightful.com/products/iminer
- R environment for statistical computing and graphics: www.R-project.org
- CART and See5/C5.0 : www.salfordsystems.com and www.rulequest.com
- Weka: www.cs.waikato.ac.nz/ml/weka.

7.2. Books

- B.Mirkin, Clustering for Data Mining: Data Recovery Approach, Chapman & Hall/CRC, 2005.
- D. Hand, H. Mannila, P. Smyth, Principles of Data Mining, Prentice Hall of India, 2001.
- J. Han, M. Kamber, Data Mining: Concepts and Techniques, 2nd ed., Morgan Kaufmann Publisher, 2006, ISBN 1-55860-901-6.
- S. Mitra, T. Acharya, Data Mining: Multimedia, Soft Computing, and Bioinformatics, John Wiley & Sons, 2004, ISBN 9812-53-063-0.
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann, 2001.

References:

A. Arulselvan, G. Baourakis, V. Boginski, E. Korchina, P.M. Pardalos, Analysis of Food Industry Market using Network Approaches, *British Food Journal*, **110**(9), 916–928, 2008.

B. Rajagopalan, U. Lall, *A*, k Nearest Neighbor Simulator for Daily Precipitation and Other Weather Variables, *Water Resources Research*, **35** (10), 3089–3101, 1999.

R. Xu, D.Wunsch II, Survey of Clustering Algorithms, *IEEE Transactions on Neural Networks*, **16** (3), 645–678, 2005.

R. Chinchuluun, W.S. Lee, J. Bhorania, P.M. Pardalos, Clustering and Classification Algorithms in Food and Agricultural Applications: A Survey, in Advances in Modeling Agricultural Systems, Springer Optimization and Its Applications Series, P.J. Papajorgji, P.M. Pardalos (Eds.), Springer, 433–454, 2008.

J.-M. Aerts, P. Jans, D. Halloy, P. Gustin, D. Berckmans, Labeling of Cough Data from Pigs for On-Line Disease Monitoring by Sound Analysis, *American Society of Agricultural and Biological Engineers*, **48** (1), 351–354, 2004.

M.Aznar, R. Lopez, J. Cacho, andV. Ferreira, Prediction of Aged Red Wine Aroma Properties from Aroma Chemical Composition, Partial Least Squares Regression Models, *Journal of Agriculture and Food Chemistry*, **51**, 2700–2707, 2003.

G.A. Baigorria, J.W. Jones, J.J. O'Brien, Potential Predictability of Crop Yield using an Ensemble Climate Forecast by a Regional Circulation Model, *Agricultural and Forest Meteorology*, **148**, 1353–1361, 2008.

P. Berkhin, Survey Of Clustering Data Mining Techniques, *Tech. Report*, Accrue Software, San Jose, CA, 2002.

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, *From Data Mining to Knowledge Discovery in Databases*, Artificial Intelligence Magazine, **17**, 37–54, 1996.

S. Mitra, T. Acharya, *Data Mining: Multimedia, Soft Computing, and Bioinformatics*, John Wiley & Sons, 2004, ISBN 9812-53-063-0.

J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, 2nd ed., Morgan Kaufmann Publisher, 2006, ISBN 1-55860-901-6.