Introduction to Data Mining Techniques

Dr. Rajni Jain

1 Introduction

The last decade has experienced a revolution in information availability and exchange via the internet. In the same spirit, more and more businesses and organizations began to collect data related to their own operations. While the database technologists have been seeking efficient means of storing, retrieving and manipulating data, the machine learning community has focussed on developing techniques for learning and acquiring knowledge from the data. At times the data can be considered to be a gold mine for strategic planning for research and development in this area which is often referred to as *Data Mining* (DM) and *Knowledge Discovery in Databases* (KDD).

We are overwhelmed with data. People have been seeking patterns in data since human life began. Hunters seek patterns in animal migration behaviour, farmers seek patterns in crop growth, and politicians seek patterns in voter opinion. A scientist's job is to make sense out of data, to discover the underlying model that governs the functioning of the physical world and encapsulate the same in theories that can be used for predicting the future. As the background of all scientific discoveries especially theories has been same, what is new about Data Mining (DM)? The simple answer is that, in DM the volume of the stored data is in the digital form and the search is automated or augmented by a computer. In DM, it is important to understand the difference between a model and a pattern. Model is a global summary of the dataset and makes statements about any point in the full measurement space while pattern describes a structure, relationship to a relatively small part of the data or the space in which the data would occur [HMS01]. In 1960's, computers were increasingly applied to data analysis problems and it was noted that if one searches long enough, one can always find some model to fit in the dataset but complexity and size of the model were important considerations. Also the aim is to generalize beyond the available data. Figure 1 shows the history of databases systems and DM [HK01]. And Figure 2 presents the scope of data mining in KDD. Fayyad [Fay96] defined DM as a process of finding models, interesting trends or patterns in large datasets in order to guide decisions about future activities. It requires tools that can help in

explaining the data and which are also capable to make predictions out of that. The data takes the form of a set of examples and the output takes the form of predictions on the new examples.

2 Issues in Data mining

Data mining has evolved into an important and active area of research because of the theoretical challenges and practical applications associated with the problem of discovering interesting and previously unknown knowledge from real-world databases. The main challenges to the data mining and the corresponding considerations in designing the algorithms are as follows:

1. Massive datasets and high dimensionality.
2. Overfitting and assessing the statistical significance.
3. Understandability of patterns.
4. Non-standard incomplete data and data integration.
5. Mixed changing and redundant data.

3 Tasks of Data Mining

Data mining as a term used for the specific set of six activities or tasks as follows:

1. Classification
2. Estimation
3. Prediction
4. Affinity grouping or association rules
5. Clustering
6. Description and visualization

The first three tasks - classification, estimation and prediction are all examples of directed data mining or supervised learning. In directed data mining, the goal is to use the available data to build a model that describes one or more particular attribute(s) of interest (target attributes or class attributes) in terms of the rest of the available attributes. The next three tasks – association rules, clustering and description are examples of undirected data mining i.e. no attribute is singled out as the target; the goal is to establish some relationship among all the attributes.

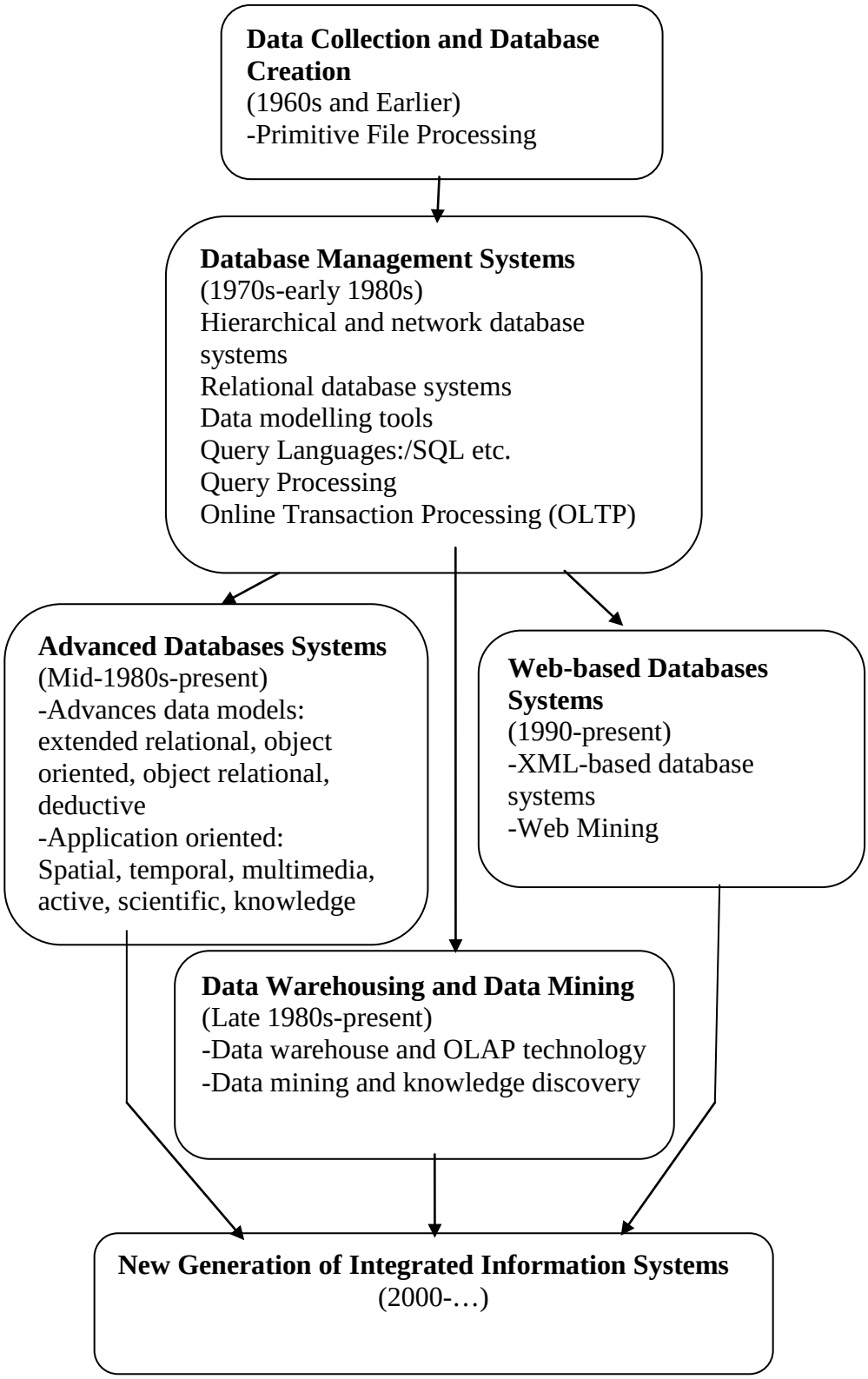

Figure 1: The history of databases systems and data mining

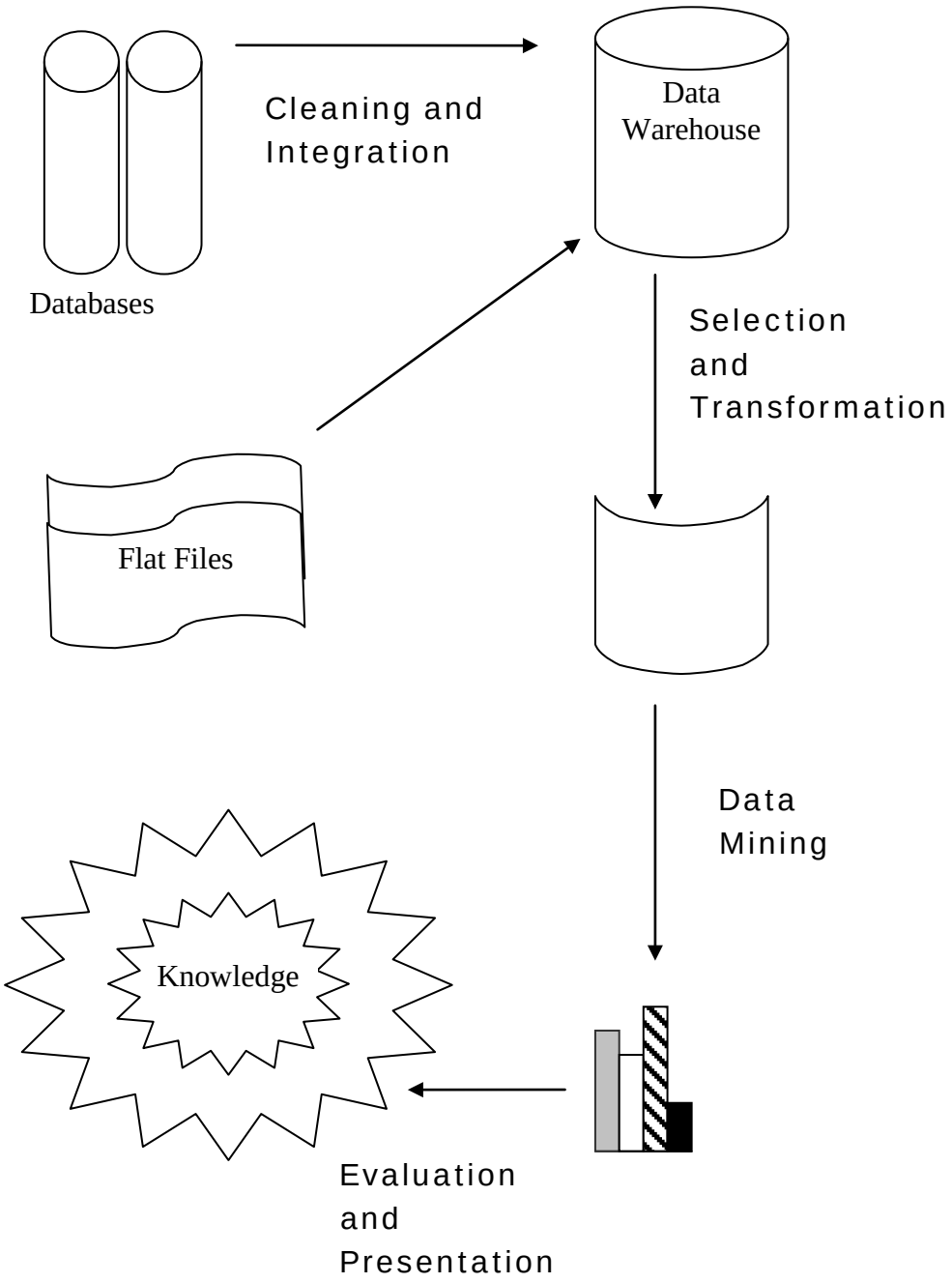

Figure 2: Data mining as a step in the process of knowledge discovery

3.1 Classification

Classification consists of examining the features of a newly presented object and assigning to it a predefined class. The classification task is characterized by the well-defined classes, and a training set consisting of preclassified examples. The task is to build a model that can be applied to unclassified data in order to classify it. Examples of classification tasks include:

- Classification of credit applicants as low, medium or high risk
- Classification of mushrooms as edible or poisonous
- Determination of which home telephone lines are used for internet access

3.2 Estimation

Estimation deals with continuously valued outcomes. Given some input data, we use estimation to come up with a value for some unknown continuous variables such as income, height or credit card balance. Some examples of estimation tasks include:

- Estimating the number of children in a family from the input data of mothers' education
- Estimating total household income of a family from the data of vehicles in the family
- Estimating the value of a piece of a real estate from the data on proximity of that land from a major business centre of the city.

3.3 Prediction

Any prediction can be thought of as classification or estimation. The difference is one of emphasis. When data mining is used to classify a phone line as primarily used for internet access or a credit card transaction as fraudulent, we do not expect to be able to go back later to see if the classification was correct. Our classification may be correct or incorrect, but the uncertainty is due to incomplete knowledge only: out in the real world, the relevant actions have already taken place. The phone is or is not used primarily to dial the local ISP. The credit card transaction is or is not fraudulent. With enough efforts, it is possible to check. Predictive tasks feel different because the records are classified according to some predicted future behaviour or estimated future value. With prediction, the only

way to check the accuracy of the classification is to wait and see. Examples of prediction tasks include:

- Predicting the size of the balance that will be transferred if a credit card prospect accepts a balance transfer offer
- Predicting which customers will leave within next six months
- Predicting which telephone subscribers will order a value–added service such as three-way calling or voice mail.

Any of the techniques used for classification and estimation can be adopted for use in prediction by using training examples where the value of the variable to be predicted is already known, along with historical data for those examples. The historical data is used to build a model that explains the current observed behaviour. When this model is applied to current inputs, the result is a prediction of future behaviour.

3.4 Association Rules

An association rule is a rule which implies certain association relationships among a set of objects (such as “occur together” or “one implies the other”) in a database. Given a set of transactions, where each transaction is a set of literals (called items), an association rule is an expression of the form X Y , where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y. An example of an association rule is: “30% of farmers that grow wheat also grow pulses; 2% of all farmers grow both of these items”. Here 30% is called the confidence of the rule, and 2% the support of the rule. The problem is to find all association rules that satisfy user-specified minimum support and minimum confidence constraints.

3.5 Clustering

Clustering is the task of segmenting a diverse group into a number of similar subgroups or clusters. What distinguishes clustering from classification is that clustering does not rely on predefined classes. In clustering, there are no predefined classes. The records are grouped together on the basis of self similarity. Clustering is often done as a prelude to some other form of data mining or modelling. For example, clustering might be the first step in a market

segmentation effort, instead of trying to come up with a one-size-fits-all rule for determining what kind of promotion works best for each cluster.

3.6 Description and Visualization

Data visualization is a powerful form of descriptive data mining. It is not always easy to come up with meaningful visualizations, but the right picture really can be worth a thousands association rules since the human beings are extremely practiced at extracting meaning from visual scenes.

Knowledge discovery goals are defined by the intended use of the system. There are two types of goals: (1) verification and (2) discovery. With verification, the system is limited to verifying the user's hypothesis. With discovery, the system autonomously finds new patterns. The discovery goal is further divided into prediction, where the system finds patterns for predicting the future behaviour of some entities and description, where the system finds patterns for presentation to a user in human understandable form.

4 DM Techniques

Each of the above problems is relevant to the derivation of useful knowledge from collection of data. Therefore methods for solving these problems, developed in the disciplines of statistics, machine learning, fuzzy sets, rough sets or hybrids are also directly relevant to DM, in particular, to conceptual data exploration. Various popular techniques of classification which have been in use pertaining to each disciplines are introduced in this section.

4.1 Statistics

The problem of abstracting knowledge from data has been tackled by statisticians, long before the first artificial intelligence papers were published. For example, correlation analysis applies statistical tools for analyzing the correlation between two or more variables. Cluster analysis offers methods for discovering clusters in large set of objects described by vector of values. Factor analysis tries to point the most important variables describing clusters. Some of the popular techniques that are used for supervised classification tasks are Linear Discriminants, Quadratic Discriminants, K-nearest Neighbour, Naïve Bayes, Logistic Regression and CART.

4.2 Machine Learning

Statistical methods have difficulty incorporating subjective, non quantifiable information in their models. They also have to assume various distributions of parameters and independence of attributes. Various studies have concluded that machine learning produces comparable (and often better) predictive accuracy. Its good performance as compared to statistical methods can be attributed to the fact that it is free from parametric and structural assumptions that underlie statistical methods. Another weakness of statistical approaches to data analysis is the problem of interpreting the results. Some of the machine learning techniques are mentioned below.

4.2.1 Neural Networks

Artificial neural networks are computational models composed of many non linear processing elements arranged in a pattern similar to biological neuron networks. A typical neural network has an activation value associated with each node and a weight value associated with each connection. An activation function governs the firing of nodes and the propagation of data through network connections in massive parallelism. The network can also be trained with examples through connection weight adjustments [TQT96].

4.2.2 Genetic Algorithms

Genetic algorithms are search algorithms based on mechanics of natural selection and natural genetics [Gol89]. They combine survival of the fittest among string structures with a structured yet randomized information exchange to form a search algorithm with some of the innovative flair of human search. In every generation, a new set of strings is created using bits and pieces of the fittest of the old; an occasional new part is tried for good measure. While randomized, genetic algorithms are no simple random walk. They efficiently exploit historical information to speculate on new search points with expected improved performance. A simple GA that yields good result, is composed of three operators namely reproduction, crossover and mutation. GAs differ from more normal optimization and search procedures in four ways:

1. GAs work with coding of parameter set, not the parameter themselves.
2. GAs search from a population of points, not a single point.

3. GAs use objective function information, not derivatives or other auxiliary knowledge.
4. GAs use probabilistic transitional rules, not deterministic rules.

4.2.3 Support Vector Machines

SVMs are the learning machines that can perform binary classification and regression estimation tasks. They are becoming increasingly popular as a new paradigm of classification and learning because of two important factors. First, unlike the other classification techniques, SVMs minimize the expected error rather than minimizing the classification error. Second, SVMs employ the duality theory of mathematical programming to get a dual problem that admits efficient computational methods.

4.2.4 Decision Tree Induction

A DT is a classification scheme which generates a tree and a set of rules, representing the model of different classes, from a given dataset. As per Hans and Kamber [HK01], DT is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test and leaf nodes represent the classes or class distributions. The top most node in a tree is the root node. Figure 3 refers to DT induced for dataset in Table 1. We have the following rules corresponding to the tree.

1. If Hair=blonde and Lotion=no then Sunburn=yes;
2. If Hair=blonde and Lotion=yes then Sunburn=no;
3. If Hair=red then Sunburn=yes;
4. If Hair=brown then Sunburn=no;

Table 1: Sunburn Dataset

ID	Hair	Height	Weight	Lotion	Sunburn
X1	blonde	average	light	no	yes
X2	blonde	tall	average	yes	no
X3	brown	short	average	yes	no
X4	blonde	short	average	no	yes

X5	red	average	heavy	no	yes
X6	brown	tall	heavy	no	no
X7	brown	average	heavy	no	no
X8	blonde	short	light	yes	no

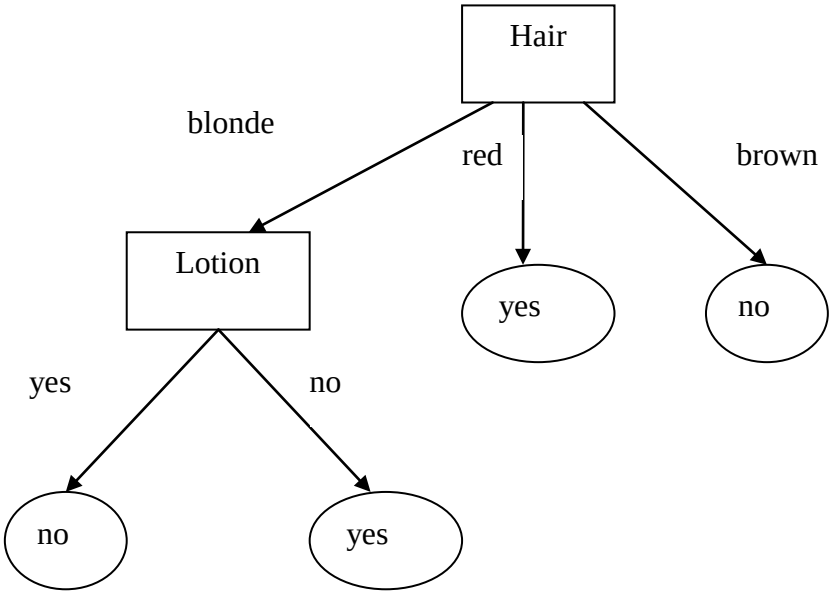

Figure 3: DT obtained by using ID3 algorithm on Sunburn data

4.3 Fuzzy Logic

Fuzzy logic, which may be viewed as an extension of classical logical systems, provides an effective conceptual framework for dealing with the problem of knowledge representation in an environment of uncertainty and imprecision [Zad89]. Some of the essential characteristics of fuzzy logic relate to the following:

1. In fuzzy logic, exact reasoning is viewed as a limiting case of approximate reasoning.
2. In fuzzy logic everything is a matter of degree.
3. Any logical system can be fuzzified.
4. In fuzzy logic, knowledge is interpreted as a collection of elastic or equivalently, fuzzy constraint on a collection of variables.

Summary of basic concepts and techniques underlying the application of fuzzy logic to knowledge representation and description of number of examples relating to its use as a computational system is provided in [Zad89]. Fuzzy logic in its pure form is not a technique for classification but it has been a very useful concept in many hybrid techniques for classification.

4.4 Rough Sets Techniques

RS theory deals with approximation of sets or concepts by means of binary relations constructed from empirical data based on the notion of indiscernibility and the inability to distinguish between objects. Such approximations can be said to form models of our target concepts, and hence in its typical use, falls in under the bottom up approach to model construction. Rough set applications to data mining generally proceed along the following directions:

1. Decision rule induction from attribute value table
2. Data filtration by template generation - This mainly involves extracting elementary blocks from data based on equivalence relation. Genetic algorithms are also sometimes used in this stage for searching.

5 Summary

Data mining involves extracting useful rules or interesting patterns from historical data. There are many data mining tasks each of them further has many techniques. No free lunch theorem exists i.e. a single technique is not suitable for all kinds of data for all types of domains. Sometimes, hybrid techniques have been observed to perform better as compared to the pure ones.

References

[HMS01] Hand, D., Mannila, H., Smyth, P., Principles of Data Mining, Prentice Hall of India, 2001

[HK01] Han, J., Kamber, M. Data Mining Concepts and Techniques, Morgan Kaufmann Publisher, 2001

[Fay96] Fayyad, U., Data Mining and Knowledge Discovery: Making Sense out of Data, IEEE Expert, Oct. 20-25, 1996

[TQT96] Tan, C. L., Quah, T. S. and Teh, H. H., An Artificial Neural Network that models Human Decision making, IEEE Computer, 64-70, 1996

[Gol89] Goldberg, D.E. Genetic Algorithms in Search Optimization and Machine Learning, Addison -Wesley, 1989

[Zad89] Zadeh, L. A., Knowledge Representation in Fuzzy Logic, IEEE TKDE, 1(1):89-99, 1989