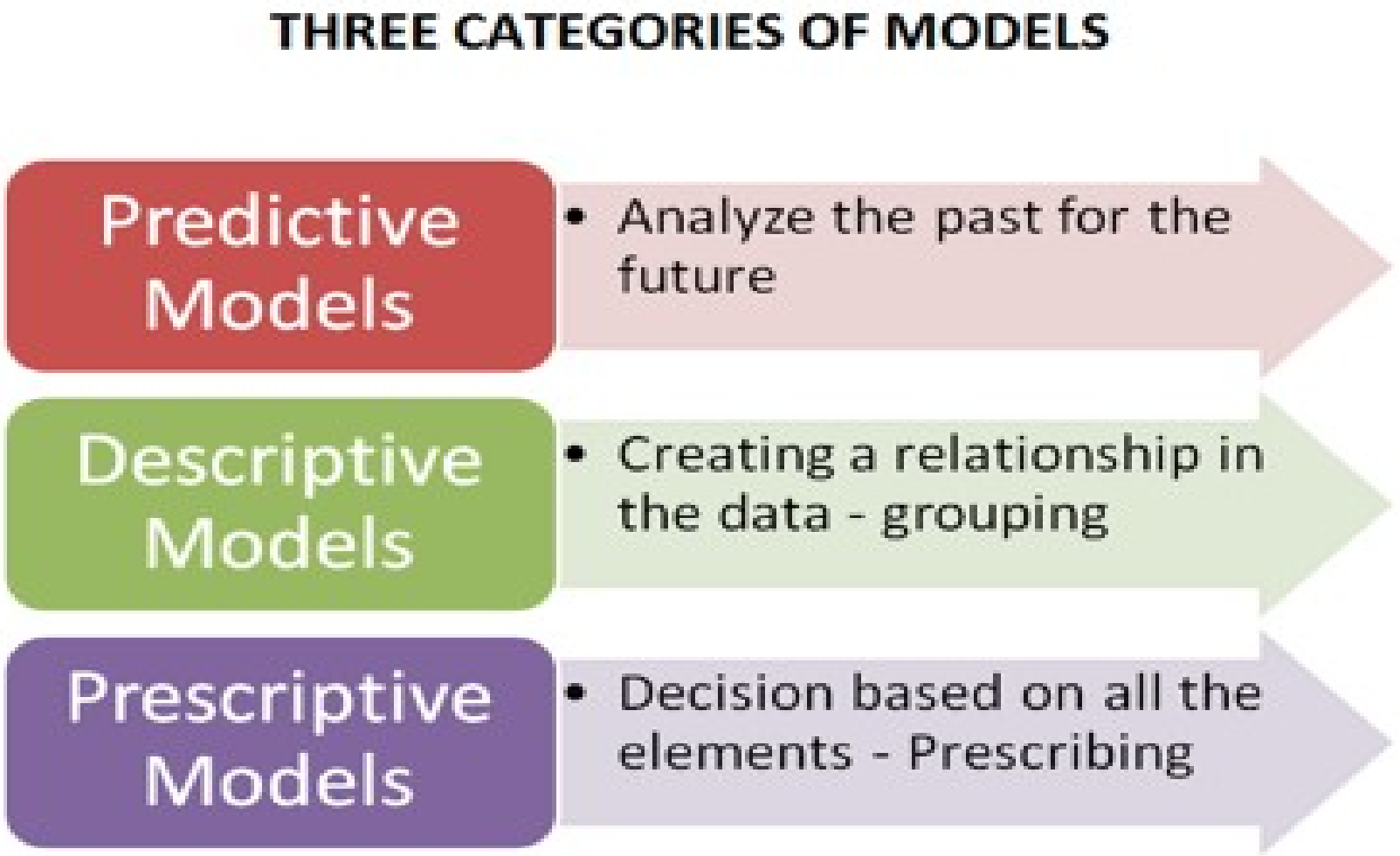
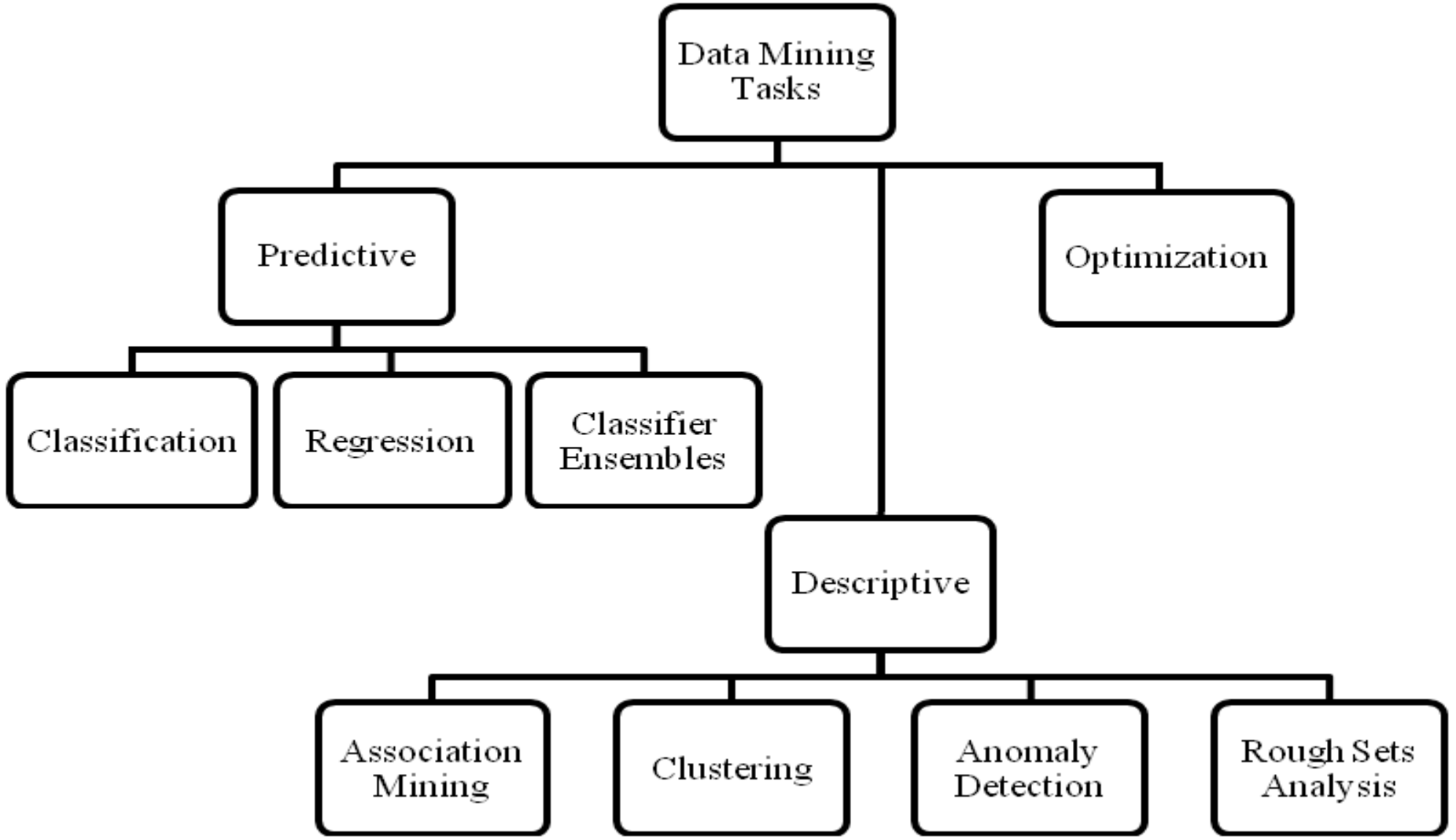
Introduction

- Models are very simple representation of a real system. For the prediction of earth climate, model can also be used. For this sake, there is a need to represent how oceans and atmosphere act, how winds blow, how temperature patterns changes and so on. Like any mathematical model about natural systems, climate model is also a simplification of real system. One important thing to remember is the choosing of complexity of a model. Therefore, complexity of a selected model sets limitations to the application of climate model.

Models


THREE CATEGORIES OF MODELS
Predictive Models
• Analyze the past for the future
Descriptive Models
• Creating a relationship in the data - grouping
Prescriptive Models
• Decision based on all the elements - Prescribing

Data Mining Tasks



Predictive Modeling

- The predictive task uses specific variables or values in the data set to predict unknown or future values of other variables of interest [33]. Several approaches have been proposed for prediction as follows:
- Classification
- The data mining task identifies the class to which a new observation belongs. Given a training data set that has several attributes, where a model is identified as a function of the other attributes' values. This requires a training set of correctly identified observations.
- The classification is applied to automatically assign records to pre defined classes, ex: to classify credit card transactions as legitimate or fraudulent, or to classify news stories as finance, entertainment,sports, etc.

Predictive Modeling

- Many techniques have emerged for classification. However, the most common approaches that have been
- used in solving real world problems are
- decision tree-based methods [10], neural networks [11], and
- support vector machines (SVM), naive bayes classifier, and k-nearest neighbor (KNN) [11].
- Decision tree-based methods deduce meaningful rules for predictive information in order to be used for data classification.
- One of the most popular algorithms is CART (Classification and Regression Tree), ID3 (Iterative Dichotomiser 3), and C4.5 [10].

Predictive Modeling

- Neural networks, which are also used in classification because of their ability to extract meaningful information from complex data, they are applied to detect patterns that are considered to be too complicated to be performed by humans.
-

Descriptive Modeling

- Descriptive models analyze past events in the data for insight on how to approach future events. These models can understand past performance by mining historical data to look for the reasons behind past success or failure. This can be used to quantify relationships in data in a way to classify, for example, customers into assemblies.
- Thus, it differs from the other predictive models that concentrate on evaluating the behavior of a single customer [28], [34].
- Descriptive mining is complimentary to predictive mining. but it is closer to decision support that decision making.
- Several approaches have been deduced from descriptive models as follows:

Descriptive Modeling(cont'd)

- Association Rules Mining
- It is an approach for exploring the relationships of interest between variables in huge databases [13].
- Considering groups of transactions, it discovers rules that forecast the existence of an item depending on the existences of other items in the transaction. It is applied to guide positioning products inside stores in such a way to increase sales, to investigate web server logs in order to deduce information about visitors to websites, or to study biological data to discover new correlations.
- Examples for association rules mining techniques are: Frequent Pattern (FP) Growth and Apriori. Apriori explores rules satisfying support and confidence values that are greater than a predefined minimum threshold value [34].

Descriptive Modeling

- Clustering
- Cluster Analysis is one of the unsupervised learning techniques, which collects similar objects together that are far different from the rest of objects in other groups [56]. Examples include grouping of related documents in emails, or proteins and genes having similar functionalities.
- Many types of clustering techniques have been introduced like the non-exclusive clustering, where the data may belong to multiple clusters. Whereas fuzzy clustering considers a data item to be a member to all clusters with different weights ranging from 0 to 1.
- Hierarchical (agglomerative) clustering, on theother hand, creates a group of nested clusters that are arranged in the form of a hierarchical tree.
- K-means is the most famous clustering algorithm, where it uses a partitioned approach to separate the data items into a pre-determined number of clusters having a centroid; data items that are in one cluster are closer to its centroid. K-medoids algorithm is a clustering algorithm related to K-means algorithm, which chooses data points as centers

Clustering

- Cluster :- is a collection of data objects
- Similar to one another within the same cluster
- Dissimilar to the objects in the other clusters
- So Clustering is Unsupervised way of grouping a set of data objects into clusters.
- It is to Determine object groupings such that objects within the same cluster are similar to each other, while objects in different groups are not
- Typically objects are represented by data points in a multidimensional space with each dimension corresponding to one or more attributes.
- Clustering problem in this case reduces to the following: Given a set of data points, each having a set of attributes, and a similarity measure, find cluster such that Data points in one cluster are more similar to one another Data points in separate clusters are less similar to one another

Clustering

- Why do Clustering?
- As a standalone tool to get insight into data distribution
- As a pre-processing step for other algorithms.
-

Clustering Applications

- Marketing : Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use : Identification of areas of similar land use in an earth observation database
- Insurance : Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning : Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies : Observed earth quake epicenters should be clustered along continent faults

Clustering Algorithms

- Partitioning approach:
- Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
- e.g: k-means, k-medoids, CLARANS
- Hierarchical approach:
- Create a hierarchical decomposition of the set of data (or objects) using some criterion
- e.g: Diana, Agnes, BIRCH, ROCK, CAMELEON

Clustering Algorithms

- Density-based approach:
- Based on connectivity and density functions
- e.g: DBSACN, OPTICS, DenClue
- Grid-based approach :
- based on a multiple-level granularity structure
- e.g: STING, WaveCluster, CLIQUE
- Model-based:
- A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other
- e.g : EM(Expectation Maximization), SOM, COBWEB

Clustering Algorithms

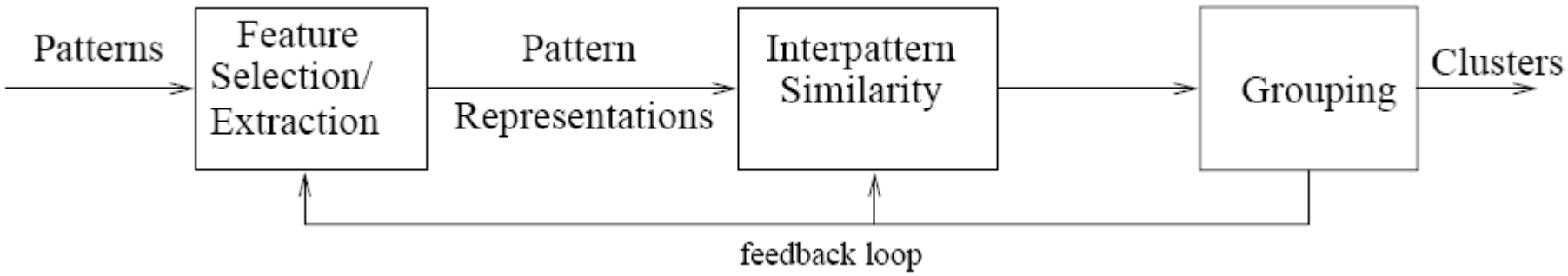
- Frequent pattern-based:
- Based on the analysis of frequent patterns
- e.g : pCluster
- User-guided or constraint-based:
- Clustering by considering user-specified or application-specific constraints
- e.g: COD (obstacles), constrained clustering

Pros and Cons of Descriptive Modeling

- Pros
- Abundance of algorithms with various grouping techniques.
- serve as a useful data-preprocessing step to identify homogeneous groups on which to build predictive models
- Help uncover natural groupings (clusters) in the data. The model can then be used to assign groupings labels (cluster IDs) to data points.
- Outliers(i.e objects that do not belong to any cluster) can easily be spotted which inturn could be used in anomaly detection applications(e.g IDS)

- Cons
- the outcome of the process is not guided by a known result, that is, there is no target attribute.
- Some algorithms need a predefined configuration parameters(K in K-means for example) which is sometimes hard to come by.
- The quality of a clustering is very hard to evaluate
- Because We do not know the correct clusters/classes
- In most Unsupervised learning applications, expert judgments are still the key for evaluation.

Architecture



- Feature Selection
- identifying the most effective subset of the original features to use in clustering
- Feature Extraction
- transformations of the input features to produce new salient features.
- Inter-pattern Similarity
- measured by a distance function defined on pairs of patterns.
- Grouping
- methods to group similar patterns in the same cluster