Data Mining notes

Marco Natali

INDICE

ELENCO DELLE FIGURE

1 INTRODUCTION TO DATA MINING

In the course of Data Mining are introduced and analyzed methods and models to use to analyze large amount of data, so we start giving the definition of Data Mining as

Def. Data Mining is the use of efficient techniques for the analysis of very large collections of data and the extraction of useful and possibly unexpected patterns in data (hidden knowledge)

Enormous data growth in both commercial and scientific databases, due to advances in data generation and collection technologies but also there is also a mantra that says to gather whatever data you can whenever and wherever possible.

Lots of data is being collected and warehoused, like for example Web data, where Yahoo has Peta Bytes of web data or also Facebook has billions of active users, also Amazon handles millions of visits each day and this also explains why know how to process and find useful information from huge amount of data will be very important.

Data are divided in two useful categories:

PRIMARY DATA: original data that has been collected for a specific purpose and they are not altered by humans.

SECONDARY DATA: data that has been already collected and made available for other purposes and may be obtained from many sources.

The process of Knowledge discovered in database (KDD) can be described by figure 1, where we have input data that will have a preprocessing phase and then we applied to a data mining modelation and we obtain the information gained.

We have so discovered that Data Mining is an phase of KDD process and we descrived all phases that can be viewed on 2, that will all analyzed during the course:

DATA INTEGRATION: involves the process of data understanding, data cleaning, merging data coming from multiple sources and transfoming them to load them into a Data Warehouse Databases.

DATA WAREHOUSE: is a database targeted to answer specific business questions.

DATA SELECTION: relevant data to analysis tasks are retrieved from data.

DATA TRANSFORMATION: transform data into appropriate form for mining (summary, aggregation, etc.)

Figura 1: Phases on a Data Mining approach

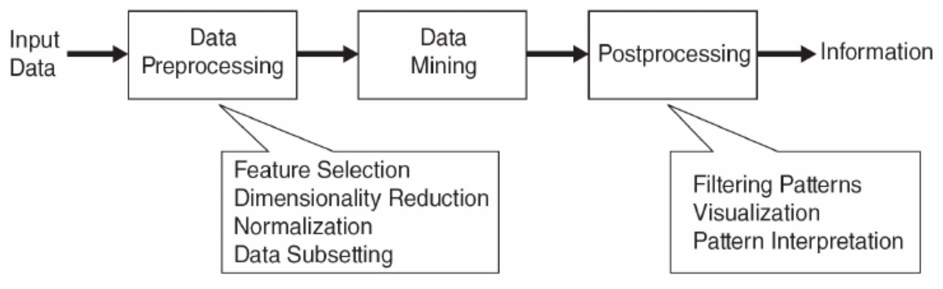
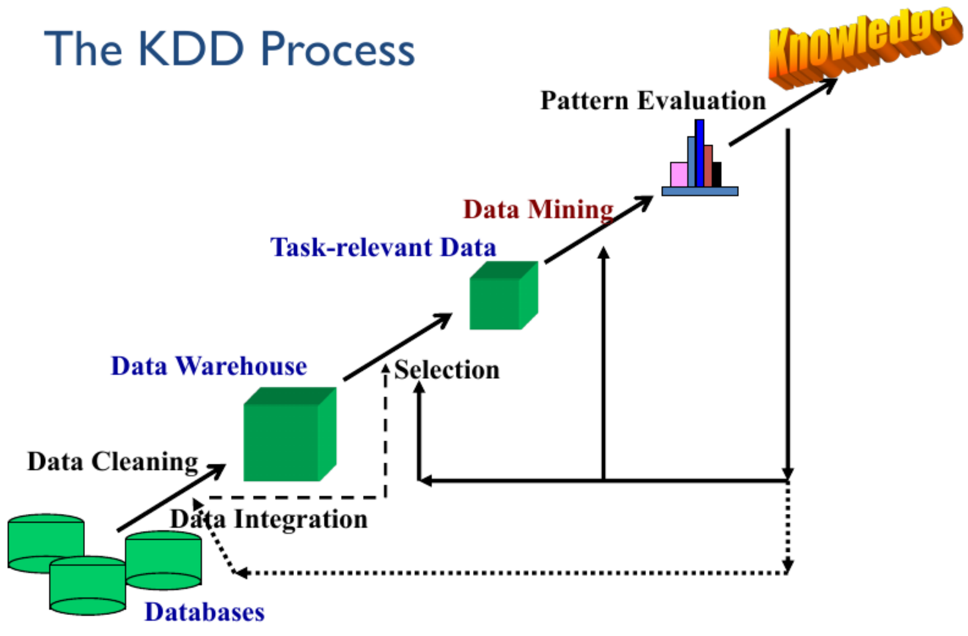


Figura 2: KDD process phases



PATTERN EVALUATION: Identify truly interesting patterns

KNOWLEDGE REPRESENTATION: use visualization and knowledge representation tools to present the mined data to the user

Data Mining approches can be divided in two different tasks:

PREDICTION METHODS: we use some variables to predict unknown or future values of other variables.

DESCRIPTION METHODS: the purpose is to find human-interpretable patterns that describe the data.

We now descrive the 4 modellation task that are developed and used on Data Mining:

CLASSIFICATION AND REGRESSION: refers to the task of building a model for the target variable as a function of explanatory variable and can be a *classification* (model for discrete class attribute) or a *regression* where we continuious data where we would like to predict as a function of the values of other attributes.

CLUSTERING: finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.

ASSOCIATION RULES: Given a set of records each of which contain some number of items from a given collection, produce dependency rules which will predict occurrence of an item based on occurrences of other items.
It is used in Market-basket analysis to optimize sales promotion and also in medical informatics to find combination of patient symptoms and test results associated with certain diseases.

ANOMALY DETECTION: detect significant deviations from normal behavior and it can be used in Credit Card fraud detection or in detect network intrusions.

Traditional techniques have often encountered practical difficulties in meeting the challenges posed by big data application in particular in scalability of peta/etabytes of data, in high dimension of attributes, in heterogeneous and complex data and also in ownership of data so this is why was introduced and also used data mining approach.

2 DATA VISUALIZATION

For preparing data for data mining task it is essential to have an overall picture of your data, so we have to gain insight in your data, with respect to your project goals and should be general to understand properties.
You should discover semantics of data and also to discover statistical charecteristics of your data in order to have a more understanding of data.

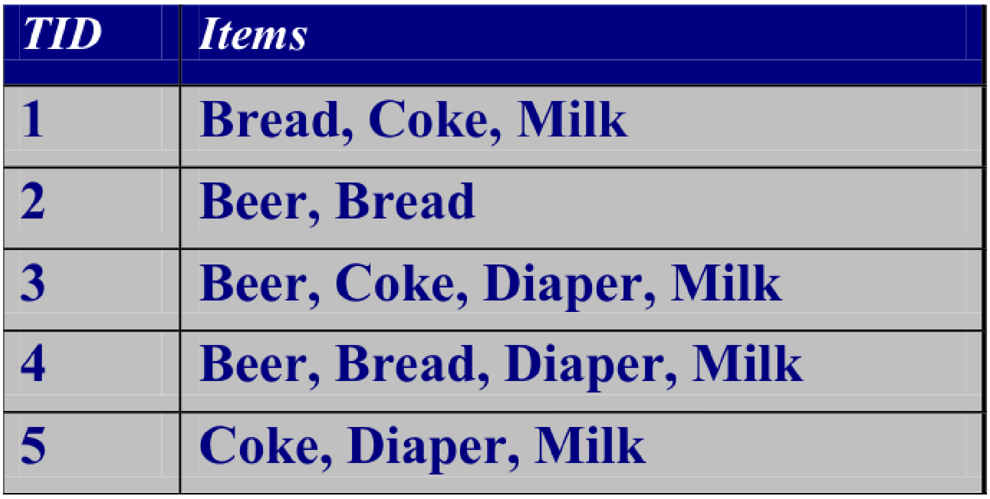
Data can be represent usually by 3 modes:

- Record: we have a matrix representation that will represent data and it is divided in:
  - Data Matrix: If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multidimensional space, where each dimension represents a distinct attribute.
    Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute.
  - Document Matrix: each document becomes a 'term' vector, where each term is a component (attribute) of the vector and the value of each component is the number of times the corresponding term occurs in the document.
  - Transiction Data: A special type of record data, where each record (transaction) involves a set of items, so for example consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items, as it possible to note in figure 3.
- Graph
- Order

Data is a collection of data objects and their attributes, where the last one is a property or characteristic of an object like eye color of a person and a data object is a collection of attributes that descrive an object.

There are different types of attributes:

Figura 3: Example of Transiction Data



TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

NOMINAL/CATEGORICAL: attribute values in a finite domain, categories, "name of things" like eye color, zip codes.

BINARY: nominal attribute with only 2 states (0 and 1) where we have *symmetric binary*, in which both outcomes are equally important, or also *asymmetric binary* where outcomes are not equally important, like medical test positive vs negative, and the convention is to assign 1 to most important outcome.

ORDINAL: finite domain with a meangniful ordering on the domain, like rankings, grades and height.

NUMERIC: quantity (integer or real-valued) measured on a scale of equal-sized units and which values have order, example of this data are temperatures in Celsius and calendar dates.

RATIO-SCALED: we can speak of values as being an order of magnitude larger than the unit of measurement and example are length, elapsed time and so on.

There is also a distinction about which data an attribute can have:

DISCRETE ATTRIBUTE: has only a finite or countably infinite set of values and often are represented as integer variables, where we can note that binary attributes are a special case of discrete attributes.

CONTINUOUS ATTRIBUTE: has real numbers as attribute values and practically real values can only be measured and represented using a finite number of digits. Examples are temperature, weight, or height and also continuous attributes are typically represented as floating-point variables.

The type of an attribute depends on which of the following properties/operations it possesses:

DISTINCTNESS: $= \neq$

ORDER: $<>$

DIFFERENCES: are $+-$

RATIOS: are $*/$

We have that Nominal attribute has only distinctness, ordinal attribute add the order property, interval attribute has also differences and in the end ratio attribute has all 4 properties.

Poor data quality negatively affects many data processing efforts infact we have that the most important point is that poor data quality is an unfolding disaster and also poor data quality costs to a typical company at least 10% of revenue, but 20% is probably a better estimate.

Some *Data quality* issues are the following:

SYNTACTIC ACCURACY: entry is not in the domain, like write "fmale" in gender, and that can be checked and solve quite easy.

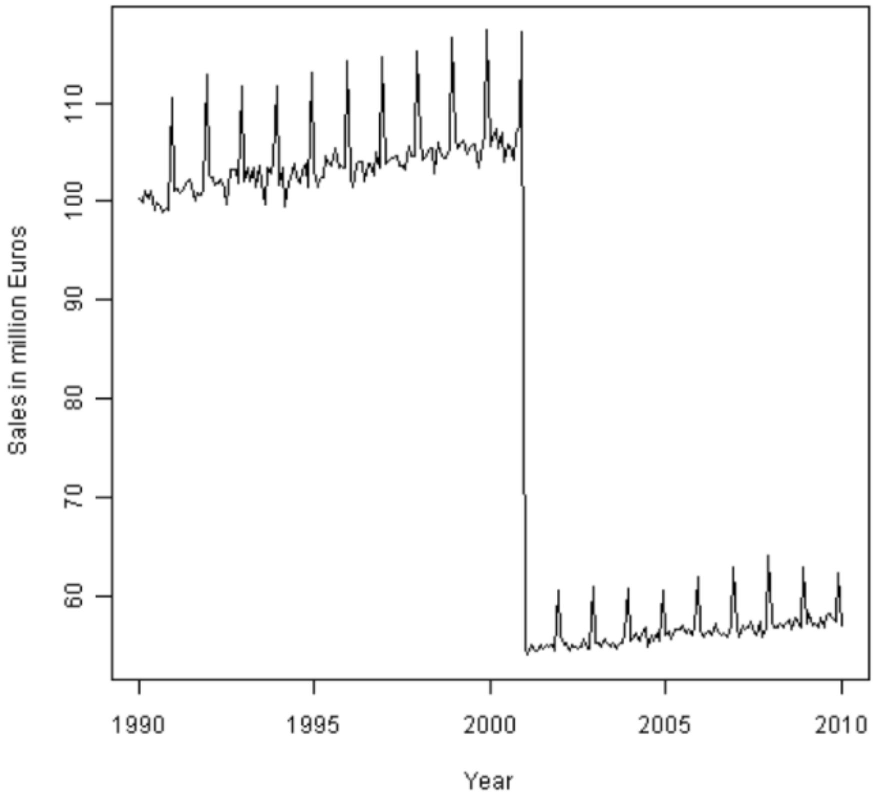
SEMANTIC ACCURACY: entry is in the domain but not correct, like "Bergamo is in France", and that type of error needs more information to be checked ("business rules").

COMPLETENESS: is violated if an entry is not correct although it belongs to the domain of the attribute, and that happens when complete records are missing so the data is biased.

UNBALANCED DATA: the data set might be biased extremely to one type of records.

TIMELINESS: is the available data up to date?

Figura 4: Change of scale in Revenue in Italy



Data set may include data objects that are duplicates, or almost duplicates of one another and that is a major issue when we are merging data from heterogeneous sources, an example can be a same person with multiple email addresses. *Data cleaning* is the process of dealing with duplicate data issues and consists to discover missing data and outlier with the purpose to remove all or at least a largest part.

In figure 4 we can see that in 2001 there was the change from Lira to Euro and that explain a dramatic reduction on a problem so we have to rescale values that makes consistent data.

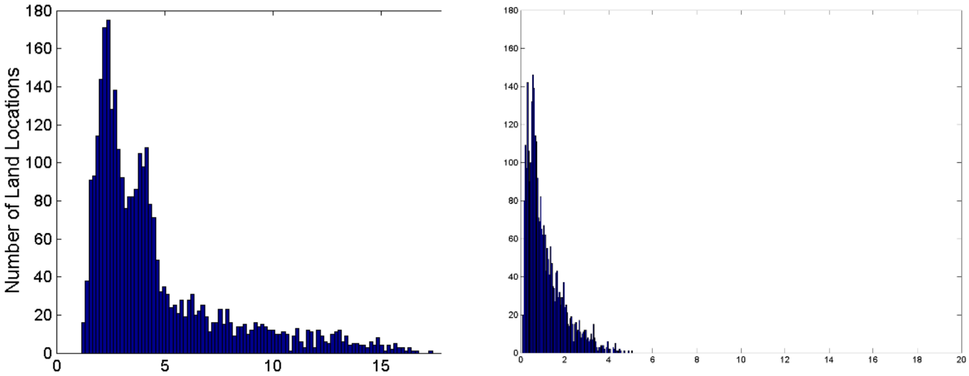
The distribution of data observation is important to recognize if we have symmetric data, skewed data, bimodal pattern (which usually shows subpopulation on attribute) and when visualisations reveal patterns or exceptions, then there is "something" in the data set; when visualisations do not indicate anything specific, there might still be patterns or structures in the data that cannot be revealed by the corresponding (simple) visualisation techniques.

We will define now what we intend when we talk about outlier:

Def. Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set.

There are two different cases: one when Outliers are noise that interferes with data analysis and another when Outliers are the goal of our analysis, like in credit card fraud and intrusion detection.
Outliers cause data quality problems and should be exceptional or an unusual data objects, so outliers coming from erroneous data should be excluded from the analysis and even if the outliers are correct (exceptional data), it is sometime useful to exclude them from the analysis.
For example, a single extremely large outlier can lead to completely misleading values for the mean value.

Figura 5: Example Aggregation on Australia's precipitation



To detect an outlier we can do the following actions:

SINGLE ATTRIBUTE: when we have an outlier in categorial attributes we can find it when we have a value that occurs with an extremely lower frequency than other; in case we want to discover an outlier in numerical attributes we use box plots.

MULTIDIMENSIONAL ATTRIBUTE: we can use Scatter plots for visually detect outliers for two attributes or also we can use PCA or MDS plots to discover outliers.

For some instances values of single attributes might be missing, due to a missing a collection of an information, broken sensors and also attributes may not be applicable to all cases, and also missing value might not necessarily be indicated as missing, we can use instead zero or default values.

2.1 DATA PREPARATION

Data preparation uses informations from data understanding to select attributes, reduce the data dimension, select records, treat missing values, treat outliers, integrate, unify and transform data and in the end we can improve data quality.

In Data Preparation can be execute the following operations:

AGGREGATION: combining two or more attributes (or objects) into a single attribute (or object) with the purpose of reduce the number of attributes/objects, have more stable data and also to change the scale, like aggregate cities into regions. In figure 5 it is possible that aggregation from number of precipitation from a month in Australia in number of precipitation in a year make more stable data.

DATA REDUCTION: we reduce the amount of data that can be done in two ways:

- Reduce the number of records, using Data Sampling and Clustering.
- Reduce the number of columns (attributes), where we select a subset of attributes and we generate a new (a smaller) set of attributes

Sampling is the main technique employed for data reduction and it is often used for both the preliminary investigation of the data and the final data analysis. Sampling is typically used in data mining because processing the entire set of data of interest is too expensive or time consuming.

The key principle for effective sampling is to using a sample that will work almost as well as using the entire dataset, if the sample is *representative*, that happens if the same has approximately the same properties as the original set of data.

The types of Samplings are the following:

SIMPLE RANDOM SAMPLING: there is an equal probability of selecting any particular item and we can have sampling without and with replacement.

STRATIFIED SAMPLING: split the data into several partitions and then draw random samples from each partition and we create an approximation of the percentage of each class; it is suitable for distribution with peaks, in which each peak is a layer.

DIMENSION REDUCTION: selection of a subset of attributes that is as small as possible and sufficient for the data analysis.
Consist in removing (more or less) irrelevant features, that contain no information useful for data mining (students ID is irrelevant in predicting student GPA), and also removing redundant features, that duplicate much or all of the information contained in one or more other attributes.
When dimensionality increases, data becomes increasingly sparse in the space that it occupies(curse of dimensionality) and also definitions of density and distance between points, which are critical for clustering and outlier detection, become less meaningful.

The reduction of dimension has the purpose to avoid curse of dimensionality, reduce amount of time and memory required by datamining algorithms, allow data to be more easily visualized and may help to eliminate irrelevant features or reduce noise.

It uses PCA(Principal Components Analysis), Singular Value Decomposition or other supervised and non-linear techniques.

For removing irrelevant features, a performance measure is needed that indicates how well a feature or subset of features performs and for removing redundant features, either a performance measure for subsets of features or a correlation measure is needed.

To reduce the dimension we usually use these 3 methods:

FILTER METHOD: selection after analyzing the significance and correlation with other attributes (preprocessing)

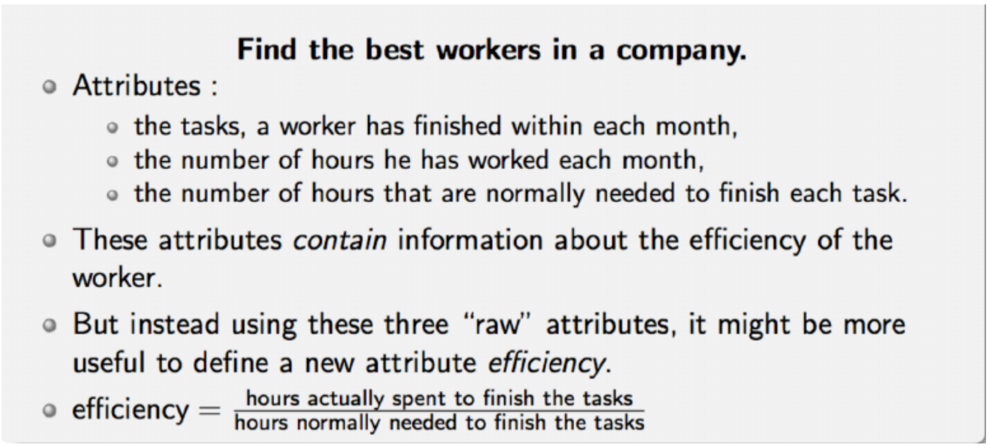
WRAPPER METHOD: selecting the top-ranked features using as reference a DM task and there is an incremental selection of the “best” attributes, with respect of information gain.

EMBEDDED METHOD: selection as part of the data mining algorithm, so during the operation of the DM algorithm, the algorithm itself decides which attributes to use and which to ignore, like Decision tree.

FEATURE CREATION: create new attributes that can capture the important information in a dataset much more efficiently than the original attributes and there are two general methodologies:

- Feature construction: in figure 6 can be seen an example on how to create a new feature.
- Feature projection: it transforms the data in the high-dimensional space to a space of fewer dimensions and this transformation may be linear, or nonlinear.
  It uses PCA (Principal Component Analysis), SVD (Singular Value Decomposition), Autoencoder and LDA (Linear Discriminant Analysis).

Figura 6: Example of Feature Creation



Find the best workers in a company.

- Attributes :
  - the tasks, a worker has finished within each month,
  - the number of hours he has worked each month,
  - the number of hours that are normally needed to finish each task.
- These attributes *contain* information about the efficiency of the worker.
- But instead using these three "raw" attributes, it might be more useful to define a new attribute *efficiency*.
- efficiency = $\frac{\text{hours actually spent to finish the tasks}}{\text{hours normally needed to finish the tasks}}$

Data Cleaning will deal on how to handle anomalous values, how to handle outliers and also on data transformations, so we will start to discuss on how we can manage missing values: a first approach can be to eliminate records that contain missing values or also it is possibile to substitute values, estimate using a probability distribution of existing values or also build a model (regression/classification) for computing missing values, or even using mean, median and mode.

Discretization is the process of converting a continuous attribute into an ordinal attribute, where potentially infinite number of values are mapped into a small number of categories and this technique is commonly used in classification, because many classification algorithms work best if both the independent and dependent variables have only a few values.
The advantages of Discretization is that original values can be continuous and sparse instead discretized data can be simple to be interpreted and also data distribution after discretization can have a normal shape.

There are two different approach to discretize data:

UNSUPERVISED DISCRETIZATION: we don't have label for the instances and the number of classes is unknown and there are 3 technique to bin data:

NATURAL BINNING: it is a simple approach, which subdivise sorted attributes values in k parts with the same size using a interval size defined as

$$\delta = \frac{x_{max} - x_{min}}{k}$$

so an element x_j belong to a class i if

$$x_j \in (x_min + i\delta, x_min + (i+1)\delta)$$

The problem of this approach is that can generate distribution very unbalanced.

EQUAL FREQUENCY BINNING: we sort, count the elements, and we define k intervals of f as

$$f = \frac{N}{k}$$

where N are the number of elements on the sample.
An element x_j belongs to a class j if we have

$$j \times f \leq i \leq (j+1) \times f$$

It is not always suitable for highlighting interesting correlations.

STATISICAL BINNING: it uses statistical information (Mean, variance, quartile) to create k interval

The optimal number of classes is function of N elements defined by [?] in 1929

$$C = 1 + \frac{10}{3} \log_{10} N$$

and the optimal width of the classes depends on the variance and the number of data as told in 1979 in [?]

$$h = \frac{3.5 * s}{\sqrt{N}}$$

SUPERVISED DISCRETIZATION: the discretization has a quantifiable goal and the number of classes is known and there are two techniques used:

- discretization based on percentiles
- discretization based on *Entropy*: Minimizes the entropy of a label, that maximizes the purity of the intervals, so in this starts by bisecting the initial values so that the resulting two intervals give minimum entropy, the splitting process is then with another interval, typically choosing the interval with the worst (highest) entropy and in the end we will stop when a user-specified number of intervals is reached, or a stopping criterion is satisfied.

Binarization maps a continuous or categorical attribute into one or more binary variables and it is typically used for association analysis.
We often convert a continuous attribute to a categorical attribute and then convert a categorical attribute to a set of binary attributes.

2.2 DATA TRANSFORMATION

In this section we will briefly talk about data transformation that can reduce in particular two issues: data with errors and imcomplete and data not adequately distributed (with many peaks and strong asymmetry in the data).

An *attribute transform* is a function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values.
It can be done by a simple function ($e^x, \log x$ for example) or by normalization and in statistics standardization refers to substracting off the means and dividing by the standard deviation.
We define a transformation T on the attribute X as $Y = T(X)$ such that Y preserve the relevant information of X, Y is more useful of X and in the end Y eliminates at least one of the problems of X.

The goals of data transformation are the following:

- stabilize the variances
- normalize the distributions
- make linear relationships among variables
- simplify the elaboration of data containing features you do not like
- represent data in a scale considered more suitable

To normalize data there are 3 techniques that are the following:

MIN-MAX NORMALIZATION: we transform data using a new max e min of our data as following

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Z-SCORE NORMALIZATION: we transform to a variable distributed as a Z with the following transformation

$$v' = \frac{(v - mean_A)}{stdev_A}$$

DECIMAL SCALING: we scale all values V by a decimal scaling defined as

$$v' = \frac{v}{10^j}$$

where j is the smallest number such that $max|v'| \leq 1$.

2.3 PCA: PRINCIPAL COMPONENT ANALYSIS

The goal of PCA is to find a new set of dimensions (attributes or features) that better captures the variability of the data and in PCA the first dimension is chosen to capture as much of the variability as possible; the second dimension is orthogonal to the first and, subject to that constraint, captures as much of the remaining variability as possible, and so on.
It is a linear transformation that chooses a new coordinate system for the data set and the steps done in PCA approach are the following:

1. Standardize the dataset
2. Calculate the covariance matrix for the features in the dataset.
3. Calculate the eigenvalues and eigenvectors for the covariance matrix.
4. Sort eigenvalues and their corresponding eigenvectors and pick k eigenvalues and form a matrix of eigenvectors.
5. Transform the original matrix.

PCA calculates the covariance matrix of all pairs of attributes given matrix D of data, where remove the mean of each column from the column vectors to get the centered matrix C (standardization) and then compute the matrix $V = C^T C$, the covariance matrix of the row vectors of C.

Exact value in Covariance matrixis not as important as it's sign, so a positive value of covariance indicates both dimensions increase or decrease together, a negative value indicates while one increases the other decreases, or vice-versa and in the end if covariance is zero the two dimensions are independent of each other.

We identify the principal components of data by computing the eigenvectors and eigenvalues from the covariance matrix, these components are new variables that are constructed as linear combinations of the initial variables, uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components, infact PCA tries to put maximum possible information in the first component, then maximum remaining information in the second and so on.

The eigenvectors of the Covariance matrix are actually the directions of the axes where there is the most variance(most information) and Eigenvalues are simply the coefficients attached to eigenvectors, which give the amount of variance carried in each PC.
By ranking your eigenvectors in order of their eigenvalues, highest to lowest, you get the principal components in order of significance.

3 DATA SIMILARITY

Similarity is a numerical measure of how alike two data objects are, it is higher when objects are more alike and often falls in the range $[0, 1]$.
We have another concept related is the *dissimilarity*, who is a numerical measure of how different are two data objects, it is lower when objects are more alike and minimum dissimilarity is often 0.

To compute the dissimilarity of data we will use some distance function, where the more common are the following:

EUCLIDEAN DISTANCE: it is a common distance function used in geometry to compute the distance between vectors and it is defined as

$$d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

MINKOWSKI DISTANCE: is a generalization of Euclidean distance and it is defined as

$$d(x,y) = (\sum_{k=1}^{n} (x_k - y_k)^r)^{1/r}$$

When $r = 1$ we have *Manhattan Distance*, with $r = 2$ we have Euclidean Distance and in the end with $r = \infty$ we have L_∞ norm.

Distances, such as the Euclidean distance, have some well-known properties:

1. $d(x,y) \geq 0 \quad \forall x,y$
2. $d(x,y) = 0 \quad x = y$
3. $d(x,y) = d(y,x) \quad \forall x,y$
4. $d(x,z) \leq d(x,y) + d(y,z) \quad \forall x,y,z$

A distance that satisfies these properties is a *metric* and also similarities have some well-known properties:

1. $s(x,y) = 1 \quad x = y$
2. $s(x,y) = s(y,x) \quad \forall x,y$

Common situation is that objects, p and q, have only binary attributes, so we have to compute similarities using the following two measure:

SMC (SIMPLE MATCHING): it is defined as the ratio between number of matches and number of attributes

JACCARD COEFFICIENT: it is defined as the ratio betwen number of 11 matches and number of not-both-zero attributes values

If d_1 and d_2 are two document vectors, we compute the similarity using the *cosine similarity*, that is defined as

$$\cos(d_1, d_2) = \frac{d_1 * d_2}{\|d_1\| \; \|d_2\|}$$

We can consider also weights in our similarities and distance, adding an weight elements that multiply each element in distance and similarity function.

Another important measure is the *Correlation*, who measures the linear relationship between objects (binary or continuous) and to compute correlation, we standardize data objects, p and q, and then we compute

$$corr(x,y) = \frac{s_{xy}}{s_x s_y}$$

Another important consideration to do is the amount of information of attributes and information relates to possible outcomes of an event, where the more certain an outcome, the less information that it contains and vice-versa.
To compute the quantity of information *Entropy* is the commonly used measure defined as

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

and this measure is between 0 and $logn$, with bits as measure unit.

We also consider *mutual information*, that captures information that one variable provide to another and it is computes as

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

where $H(X,Y)$ is defined as

$$H(X,Y) = -\sum_i \sum_j p_{ij} \log p_{ij}$$

4 CLUSTERING

Clustering analysis relate to finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups, as it is possible to note in figure 7

The notion of cluster can be ambiguous as can be note in figure 8 and a clustering is a set of clusters, where exist an important distinction between:

PARTITIONAL CLUSTERING: a division of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset as can be viewed in figure 9.

HIERARCHICAL CLUSTERING: a set of nested clusters organized as a hierarchical tree, as can be note in figure 10.

We have the following type of Clusters:

WELL-SEPARATED CLUSTERS: is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster

CENTER-BASED CLUSTERS: is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster; the center of a cluster is often a *centroid*, the average of all the points in the cluster, or a *medoid*, the most "representative" point of a cluster.

CONTIGUOUS CLUSTER (NEAREST NEIGHBOR OR TRANSITIVE): each point is closer to at least one point in its cluster than to any point in another cluster and this approach can have trouble when noise is present since a small bridge of points can merge two distinct clusters.

DENSITY-BASED: is a dense region of points, which is separated by low-density regions, from other regions of high density and used when the clusters are irregular or intertwined, and when noise and outliers are present.

Clusters are defined by an Objective Function, so we Finds clusters that minimize/maximize an objective function and enumerate all possible ways of dividing the

Figura 7: Cluster Example

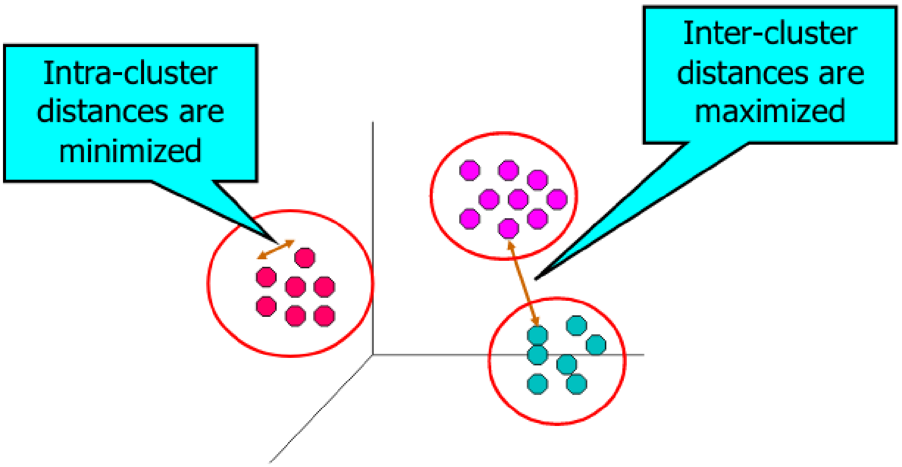

Intra-cluster distances are minimized
Inter-cluster distances are maximized

Figura 8: Ambiguity about number of Cluster

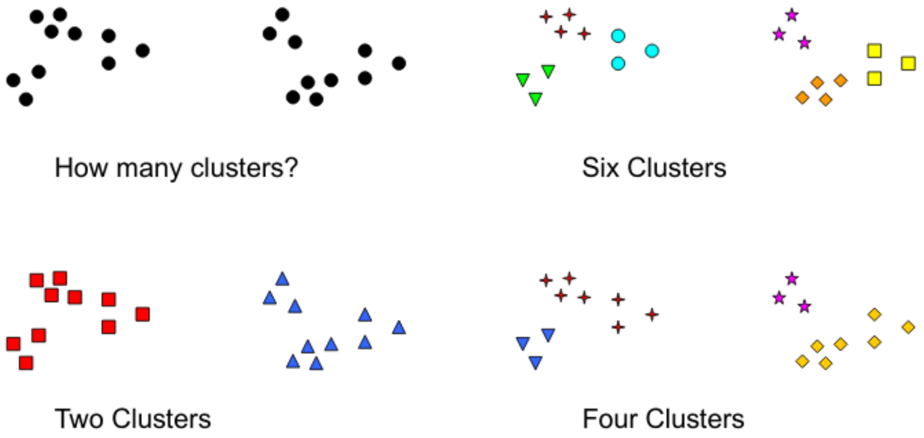


Figura 9: Example of Partitional Clustering

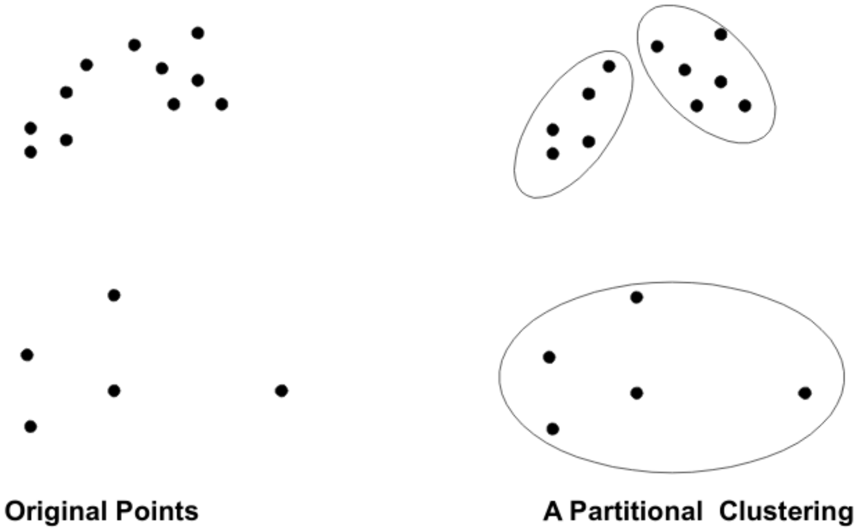
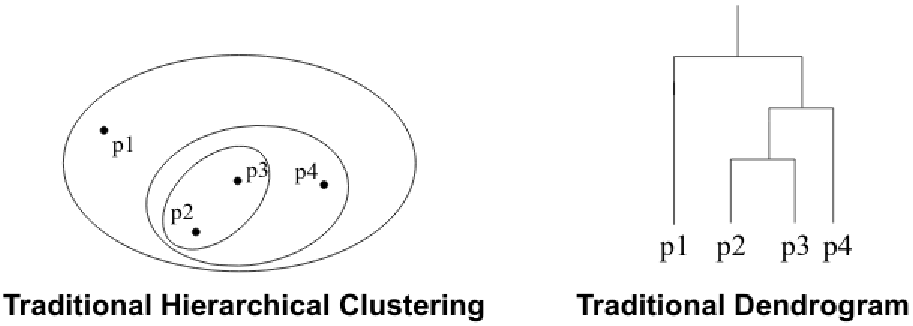


Figura 10: Example of Hierarchical Clustering



points into clusters and evaluate the ‘goodness’ of each potential set of clusters by using the given objective function is an NP Hard problem.
There can be global or local objectives: Hierarchical clustering algorithms typically have local objectives, instead Partitional algorithms typically have global objectives.