Advanced School in Artificial Intelligence

Data Mining and Association Rules

Elena Bellodi
elena.bellodi@unife.it

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Regione Emilia-Romagna

FERRARIAE UNIVERSITAS · 13 91 · EX LABORE FRUCTUS

Università degli Studi di Ferrara

Outline

- Introduction to Data Mining
- KDD Process
- Association rules
  - Metrics
  - APRIORI

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Introduction to Data Mining

- Data mining looks for «patterns» in databases
  - a database is a set of interrelated data stored on some electronic storage device

- In data mining, the search for patterns is automated by computers

- **Large volumes of data** are processed to extract **useful information** that highlights **hidden patterns in the data**

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Introduction to Data Mining

Example

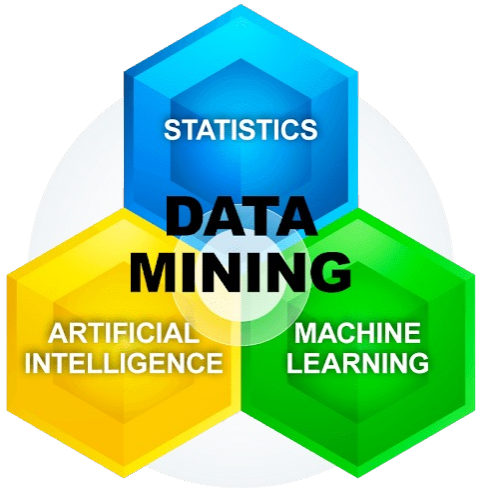
- The problem is fickle customer loyalty in a highly competitive marketplace.
- A database of customer choices, along with customer profiles, is available
- **Patterns of behavior** of former customers can be analyzed to identify distinguishing characteristics of those likely to switch products and those likely to remain loyal
- Once such characteristics are found, they can be put to work to identify present customers who are likely to jump ship

- GOAL: This group can be targeted for special treatment, treatment too costly to apply to the customer base as a whole

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Introduction to Data Mining

Data mining is designed to extract rules from large quantities of data (born in the 30s)



Machine learning trains a system to perform complex tasks and uses harvested data and experience to become smarter
(born in the 50s)

Data Mining is a step of the *Knowledge Discovery in Databases* (KDD) process

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Knowledge Discovery in Databases

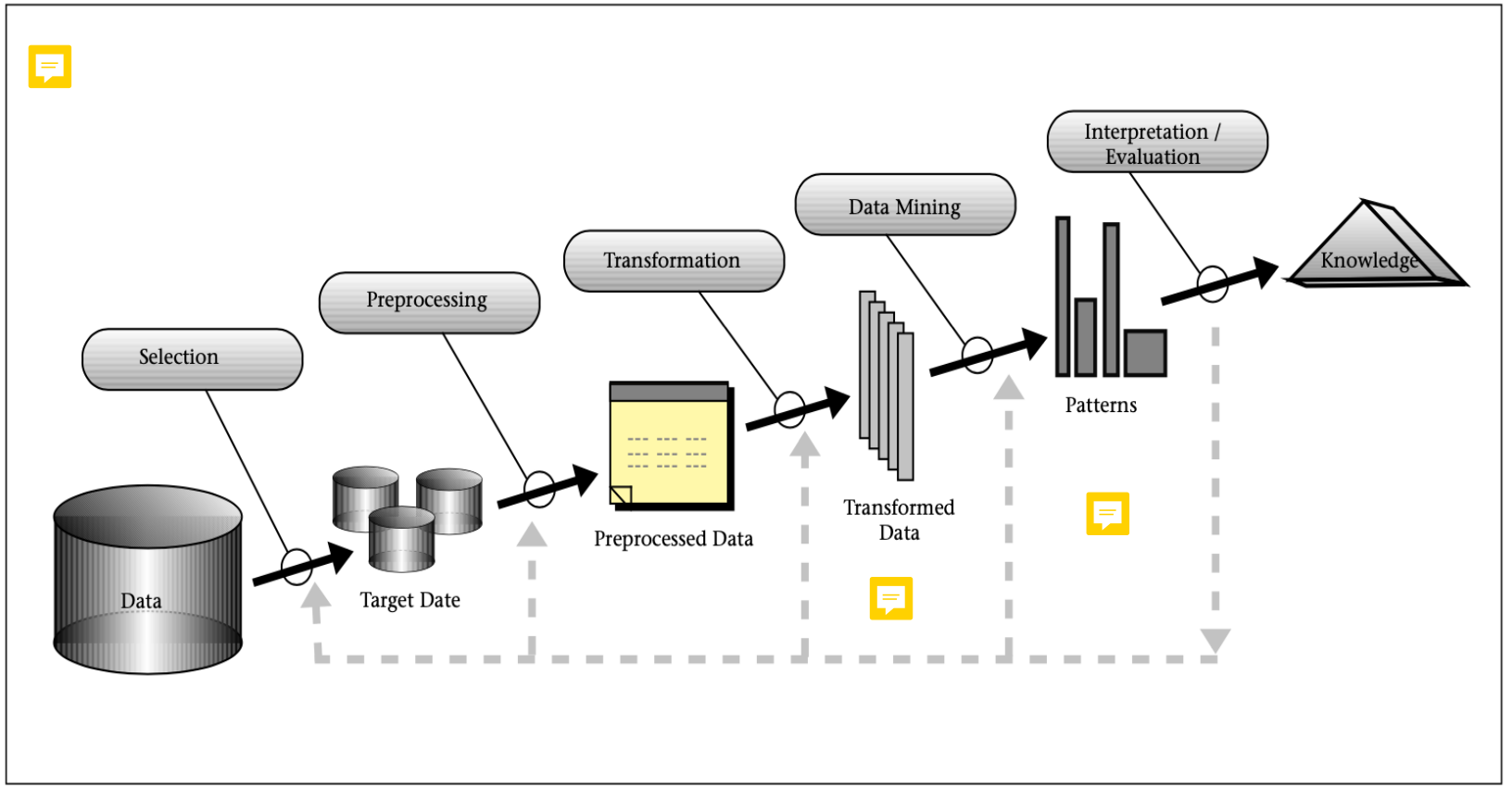


Figure 1. An Overview of the Steps That Compose the KDD Process.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Knowledge Discovery in Databases

- "KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37]

- **Data**: a set of records F in a database
- **Pattern**: an expression E in a language L describing facts in a subset F_E of F
- E is called a pattern if it is simpler than the enumeration of the facts of F_E

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Knowledge Discovery in Databases

- Process: composed of several steps. It must be non trivial, i.e. it must involve a kind of search or inference, it cannot be the calculation of a predefined quantity
- Valid: the discovered patterns must be applicable to new data
- New: the discovered patterns must not be previously known
- Potentially useful: they must potentially lead to useful actions
- Understandable: patterns must be understandable to a human being in order to facilitate the comprehension of the underlying data

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

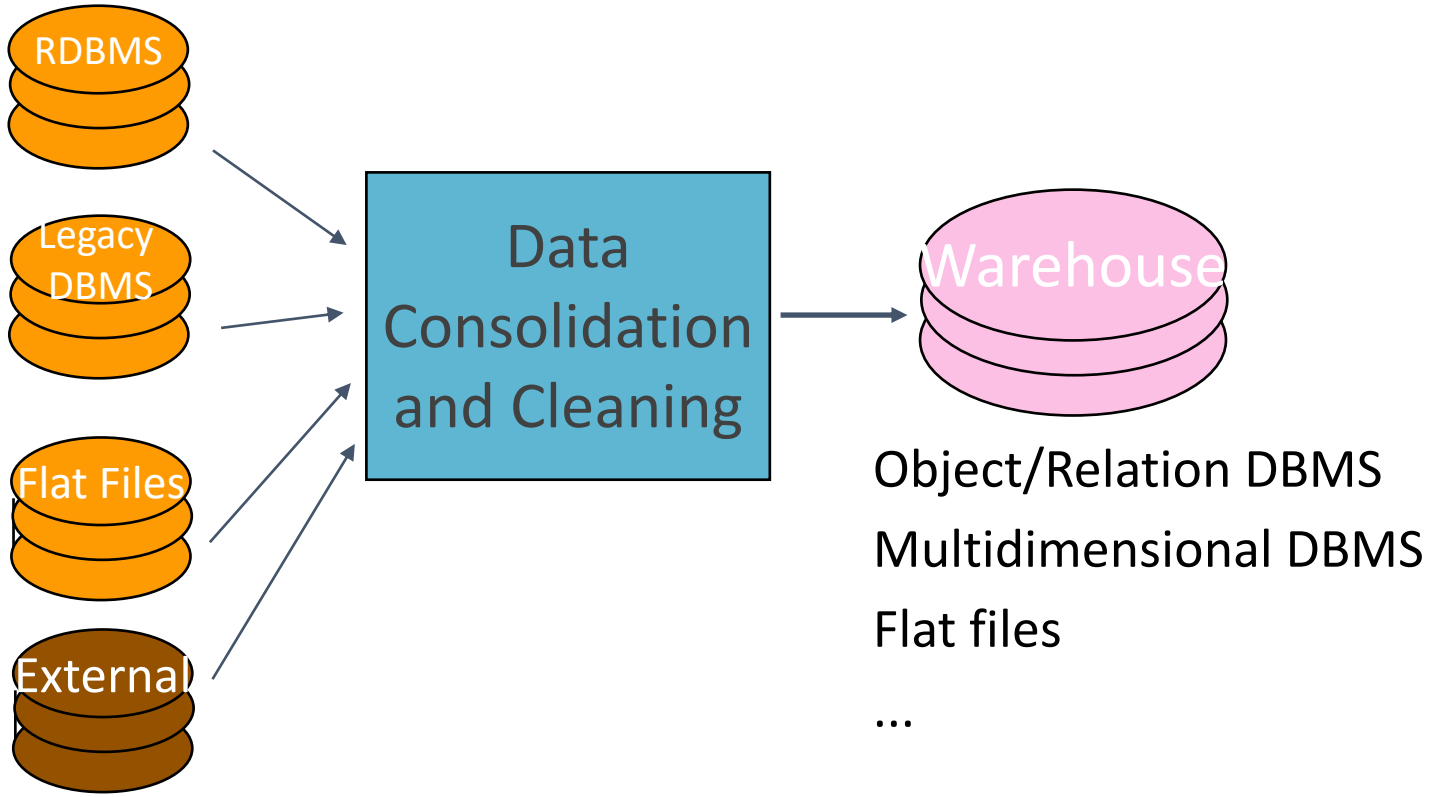
Knowledge Discovery in Databases

- The KDD process
  - is iterative as the steps of the algorithm can be repeated:
  - is interactive because it requires the intervention of the analyst for many decisions and for the choice of constraints to be imposed on algorithms
- 50%-70% of the effort on the KDD process is spent on selecting and preprocessing the data
- The quality of the results is directly related to the quality of the data

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

KDD: consolidation

- From heterogeneous sources to consolidated data repositories



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

KDD: Selection & Preprocessing

- Selection of a sample of records in case it is impossible to use the entire database
- Reduction of the attribute dimensionality
  - Removal of redundant and/or related attributes
  - Combination of attributes (sum, multiplication, difference)
- Reduction of the attribute domains
  - Grouping of values for discrete attributes
  - Quantization of continuous attributes

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

KDD: Transformation

- Data transformation: normalization of values, e.g. in neural networks, the input must be included within a domain of values between 0 and 1 or between -1 and 1

- Data encoding
  - The representation must be suitable for the data mining tool that will be used

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

KDD: Data Mining goals

Goal is description

- **Clustering**: discovery of subgroups of data such that the data within the same subgroup are similar to each other and are dissimilar from those in the other groups
- **Discovery of associative rules**: discovery of rules that describe regularities in data

- The discovered patterns may have different «formats» (language L)
  - We will study **association rules**

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Data Mining applications

- Marketing (Market Basket Analysis (MBA), Customer Relationship – Management(CRM))
- Investment / banking (predict customer profitability)
- Risk analysis (fraud detection)
- Manufacturing
- Telecommunications
- Decision support/decision making
- Clinical databases
- Text Mining
- Web Mining

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Association Rules

- They describe correlations of events (attribute-value) and can be seen as *probabilistic* rules

- Two events are related when they are frequently observed together
- An associative rule is an implication of the form $X \Rightarrow Y$, where X and Y are sets of disjoint events ($X \cap Y = \varnothing$)

- Example: sales transaction database in a supermarket. The associative rules describe what objects they are frequently bought together

  `Panettone ⇒ sparkling wine`

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Association Rules: example

each row is a transaction or record

1) Database D

No	Outlook	Temp (° F)	Humid (%)	Windy	Play
D1	sunny	75	70	T	Y
D2	sunny	80	90	T	N
D3	sunny	85	85	F	N
D4	sunny	72	85	F	N
D5	sunny	69	70	F	Y
D6	overcast	72	90	T	Y
D7	overcast	83	78	F	Y
D8	overcast	64	65	T	Y
D9	overcast	81	75	F	Y
D10	rain	71	80	T	N
D11	rain	65	70	T	N
D12	rain	75	80	F	Y
D13	rain	68	80	F	Y
D14	rain	70	96	F	Y

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Association Rules: example

2) Suppose that the database is already the result of the Selection, Preprocessing and Transformation steps

3) A specific Data Mining algorithm is applied to extract association rules

4) Final result:

```
Outlook = overcast ⇒ Play = Y

Windy = F and Play = N ⇒ Outlook = sunny
                              and Humidity = 85
…
```

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Association Rules Metrics

- *$X \Rightarrow Y$ has **support** s in database D iff a fraction s of the transactions in D contain $X \cup Y$:*

$$s(\text{Outlook=overcast} \Rightarrow \text{Play=Y}) = 4/14$$

  - A high value indicates that the rule affects a large part of the database

- *$X \Rightarrow Y$ has **confidence** c in database D iff, among all transactions containing X, there is a fraction c which contains also Y:*
  - *$c(X \Rightarrow Y) = s(XUY)/s(X)$*

$$c(\text{Outlook=overcast} \Rightarrow \text{Play=Y}) = 4/4 = 1$$

  - is an estimate of the conditional probability $P(Y|X)$

- Confidence and support can also be indicated as a percentage

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Example

Transaction ID	Purchased Items
T1	{1, 2, 3}
T2	{1, 4}
T3	{1, 3}
T4	{2, 5, 6}

For minimum support = 50% and minimum confidence = 50% we have the following rules:

1 => 3 with 50% support and 66% confidence
3 => 1 with 50% support and 100% confidence

- We will see next how to mine them from this database

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Terminology

- item/transaction: attribute=value, ex. *outlook=rain*
- itemset: a set of items, a transaction
  - {*Outlook=rain*, Temp =65, Humid =70, Windy=T, Play=N}
- or a subset of it (ex. {*Outlook=rain, Play=N*})
- k-itemset: set of k items
- database: a set of transactions tipically stored as a table

Transaction ID	Item

- frequent itemset: if it meets a minimum threshold for support and confidence

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Association Rules in practice

- Given a database D, the task of discovering associative rules is:
  - discovering all the association rules with a minimum support (called minsup) and a minimum confidence (called minconf), where minsup and minconf are values specified by the user

- The task can be decomposed into two **subproblems**:
  1. Find all itemsets that have support above the minimum (frequent itemsets). This subproblem is solved by the APRIORI algorithm
  2. Generate all the association rules with at least the minimum confidence from the set of large itemsets

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Step 1: Apriori

- The first algorithm for discovering association rules
  - Rakesh Agrawal and Ramakrishnan Srikant, *Fast algorithms for mining association rules*. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994

- Iterative algorithm to discover the most frequent itemsets
- **Idea**: given an item I, if
  - s(I) < minimum support threshold (given by the user), I is not frequent in the database
  - s(I+A) < minimum support threshold, then I + A is not frequent, where A also belongs to the itemset

that is, if an itemset has support less than minimum support, then all of its supersets will fall below the threshold, and therefore can be ignored.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Step 1: Apriori

1) At the first iteration of the algorithm (k = 1), each item is considered as a candidate of the 1-itemset. The algorithm will count the occurrences of each item.
2) let *min_sup* be the minimum support (for instance 2). The set of 1-itemsets whose occurrence satisfies min_sup is determined. Only candidates with support > = *min_sup* are carried over to the next iteration.
3) At the second iteration **k=2**, the 2-itemsets are generated by combining the items in pairs (*join*)
4) Candidate 2-itemsets are removed (***pruning***) or held using the min-sup threshold value.

…and so on

It stops at step k when the set of candidate k-itemsets (C_k) is empty. The most frequent itemset(s) are those obtained in step k-1 (L_{k-1} is the set of frequent itemsets).

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Apriori: join step

1. The items are kept **sorted** inside the k-itemsets- if the items are numerical values, they are kept sorted in *ascending order* with respect to the first item, then with respect to the second, etc (ex: {1,2},{1,3},{2,3},{2,4})

2. At the k-th join step, an itemset of k elements is generated by joining two frequent itemsets of k-1 elements that **have the first k-2 elements equal** (thanks to the ordering of the items, they are found in sequence)
   - the new itemset will have k elements: k-2 shared by the two k-1-itemsets, the other two by two different items

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Step 1: Apriori. Example

min_sup=50%=2 transactions
L_k: set of frequent k-itemsets (with min support)
C_k: new candidate k-itemsets

Database D

TID	Items
1	{1, 2, 3}
2	{1, 4}
3	{1, 3}
4	{2, 5, 6}

Scan D →

L_1

Itemset	Supp.
{1}	3
{2}	2
{3}	2

C_2

Itemset
{1,2}
{1,3}
{2,3}

Scan D →

C_2

Itemset	Supp.
{1,2}	1
{1,3}	2
{2,3}	1

L_2

Itemset	Supp.
{1,3}	2

C_3

Itemset	Supp.

→

L_3

Itemset	Supp.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Step 2: Generation of association rules

- **We need to use confidence and minconf set by the user**

1. For each frequent itemset A, find all non-empty subsets
2. For each subset X of A, generate the rule **X ⇒ (A-X)** if and only if the ratio between the support of A and the support of X (i.e., the confidence) is at least *minconf*

- So, if minimum confidence is minconf=50%:
- Itemset {1,3} //from L_2
- Rules:
- {1} ⇒ {3} // confidence = s(1,3)/s(1) = 2/3 = 0.66
- {3} ⇒ {1} // confidence = s(1,3)/s(3) = 2/2 = 1
- Both can be considered association rules wrt minconf

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Apriori: advantages and disadvantages

(+) Easy to understand

(+) The Join and Pruning steps are easy to implement on large itemsets in large databases

(-) Requires high computation if itemsets are very large and minimum support is kept very low

(-) The entire database must be scanned

- Applied in the fields of education, medicine, within organizations, ...
- Apriori is available in Weka open-source software

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Other database formats

Association rules can also be learned from databases of the form:

$Attribute_1$	**$Attrbute_2$**	….	**$Attribute_n$**
$Value_{1,1}$	$Value_{1,2}$	….	$Value_{1,n}$
….	….	….	….
$Value_{m,1}$	$Value_{m,2}$	….	$Value_{m,n}$

In practice, every record is considered as a transaction and every possible equality **Attribute = Value an item**

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Other database formats

In this case an associative rule has the form

$$A_1 = v_{A1},\ldots, A_j = v_{Aj} \quad \Rightarrow \quad B_1 = v_{B1},\ldots, B_k = v_{Bk}$$

where

- $A_1,\ldots, A_j, B_1,\ldots,B_k$ are attribute names
- $v_{A1},\ldots,v_{Aj}, v_{B1},\ldots,v_{Bk}$ are values such that v_{Ai} (v_{Bh}) belongs to the domain of the A_i (B_h) attribute

The databases of the first form can be translated into this second form by considering a Boolean attribute for each possible item.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021