Minería de Datos

U2 Entendimiento de los datos

Héctor Maravillo

Section 1

Data Preparation

Data Preparation

Data preparation or **data preprocessing** are the set of techniques that initialize the data properly to serve as input for a certain Data Mining or Machine Learning algorithm.

Data preparation is normally a **mandatory step**. It converts prior useless data into new data that fits a DM process.

Data Preparation

What are the basic issues that must be resolved in data preparation?

- How do I clean up the data? — **Data Cleaning**.
- How do I provide accurate data? — **Data Transformation**.
- How do I incorporate and adjust data? — **Data Integration**.
- How do I unify and scale data? — **Data Normalization**.
- How do I handle missing data? — **Missing Data Imputation**.
- How do I detect and manage noise? — **Noise Identification**.

Data Preparation

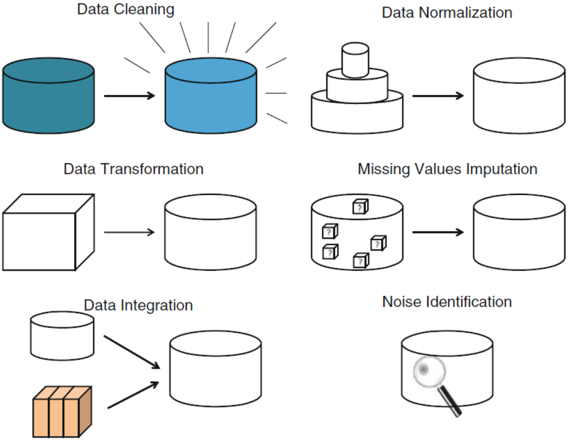


Figure: Data Preparation

Data Preparation

Three elements define data quality:

- **Accuracy**: The degree to which data correctly reflects the real-world scenario it represents.
- **Completeness**: The extent to which all required data is present.
- **Consistency**: The uniformity of data across different sources or systems.

Section 2

Data Integration

Data Integration

Data integration is the process of combining data from different sources into a single dataset.

It is not an easy task, for example:

- Different attribute names or table schemes will produce uneven examples that need to be consolidated.
- Attribute values may represent the same concept but with different names creating inconsistencies in the instances obtained.
- If some attributes are calculated from the others, the data sets will present a large size, but the information contained will not scale accordingly.

Finding Redundant Attributes

Redundancy is a problem that should be avoided as much as possible.

- It will usually cause an increased in the **data set size**, meaning that the **modeling time** of DM algorithms is incremented as well.
- It may also induce **overfitting** in the obtained model.

Redundancy

Redundancies in attributes can be detected using **correlation analysis**. Through this analysis we can **measure** how strong the **implication** of one attribute is on the other.

- When the data is **nominal** and the set of values is thus finite, the χ^2 **(chi-squared) test** is commonly applied.
- In **numeric attributes** the use of the **correlation coefficient** and the **covariance** is typical.

χ^2 Correlation test

Suppose that two **nominal attributes**, $A = \{x_1, \ldots, x_m\}$ and $B = \{y_1, \ldots, y_m\}$, contain c and r distinct values each, namely $a_1, \ldots, a_c$ and $b_1, \ldots, b_r$. We can check the correlation between them using **the χ^2 test**.

In order to do so, a **contingency table**, with the joint events (A_i, B_j) in which attribute A takes the value a_i and the attribute B takes the value b_j, is created.

Every possible joint event (A_i, B_j) has its own entry in the table.

χ^2 Correlation test

The χ^2 value (or **Pearson χ^2 statistic**) is computed as:

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where o_{ij} is the **observed frequency** of the joint event (A_i, B_j), and e_{ij} is the **expected frequency** of (A_i, B_j).

χ^2 Correlation test

The is the **expected frequency** e_{ij} is computed as:

$$\begin{aligned} e_{ij} &= \frac{count(A = a_i) \times count(B = b_j)}{m} \\ &= \frac{\sum_{i=1}^{m} 1_{a_i}(A) 1_{b_j}(B)}{m} \end{aligned}$$

Where m is the number of instances in the data set, $count(A = a_i)$ is the number of instances with the value a_i for attribute A and $count(B = b_j)$ is the number of instances having the value b_j for attribute B.

Let $1_x(X)$ be the indicator function of the random variable X.

χ^2 Correlation test

The χ^2 test checks the hypothesis that A and B **are independent**, with $(r-1)(c-1)$ **degrees of freedom.**

The χ^2 statistic obtained is compared against any χ^2 table using the suitable degrees of freedom or any available software that is able to provide this value.

- If the p-**value** is **below** the established significance level (or the computed **statistic value** computed is **above** the needed one in the table), we can saw that the **null hypothesis is rejected** and therefore, A and B are **statistically correlated**.

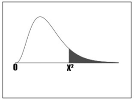
χ^2 Correlation test

The **assumptions** of the Chi-square include:

- Both variables must be nominal or categorical.
- The levels (or categories) of the variables are **mutually exclusive**. That is, a particular subject fits into one and only one level of each of the variables.
- The expected value should be 5 or more in at least 80% of the cells, and no cell should have an expected of less than one. This assumption is most likely to be met if the sample size equals at least the number of cells multiplied by 5.

χ^2 Correlation test

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Correlation Coefficient

When we have two numerical attributes, checking whether they are highly correlated or not is useful to determine if they are redundant.

The most well-known correlation coefficient is the **Pearson's product moment coefficient** or **Pearson's coefficient**, given by:

$$r_{A,B} = \frac{\sum_{i=1}^{m}(x_i - \bar{A})(y_i - \bar{B})}{m\sigma_A\sigma_B} = \frac{\sum_{i}^{m}(x_i y_i) - m\bar{A}\bar{B}}{m\sigma_A\sigma_B}$$

where m is the number of instances, a_i and b_i are the values of attributes A and B in the instances, $\bar{A}$ and $\bar{B}$ are the mean values of A and B respectively, and σ_A and σ_B are the standard deviations of A and B.

Note that

$$-1 \leq r_{A,B} \leq 1$$

Correlation Coefficient

- When $r_{A,B} \geq 0$ it means that the two attributes are **positively correlated**: when values of A are increased, then the value of B are incremented too. Having a high value of $r_{A,B}$ could also indicate that one of the two attributes can be removed.
- When $r_{A,B} = 0$, it implies that attributes A and B are independent and no correlation can be found between them.
- If $r_{A,B} \leq 0$, then attributes A and B are **negatively correlated** and when the values of one attribute are increased, the values of the other attribute are decreased.

Detecting Tuple Duplication

It is interesting to check, when the tuples have been obtained, that there are not any **duplicated** tuple.

Having duplicate tuples can be troublesome, not only wasting space and computing time for the DM algorithm, but they can also be a source of inconsistency.

Detecting Tuple Duplication

One of the most common sources of mismatches in the instances is the **nominal attributes**. Analyzing the similarity between nominal attributes is not trivial, as distance functions are not applied in a straightforward way and several alternatives do exist.

Several character-based distance measures for nominal values can be found in literature. These and can be helpful to determine whether two nominal values are similar (even with entry errors) or different.

Edit distance

The **edit distance** between two strings σ_1 and σ_2 is the minimum number of string operations (or edit operations) needed to convert one string in the other.

Three types of edit operations are usually considered:

- Inserting a character
- Replacing a character
- Deleting a character.

Using **dynamic programming** the number of operations can be established.

References

- García, S., Luengo, J. & Herrera, F. *Data Preprocessing in Data Mining*, Springer, 2015.