Data Redundancy

Redundancy is one of the major problem in data integeration. While integerating different data sources, some attributes with different names and values may representing the same meaning.

Another import part of data redundancy is Tuple Duplication, which can be completed in WEKA

In this paper, we treat data as numeric data, we calculate the **"Correlation Coefficient"** and **"Covariance"** of the dataset.

Efficiency Evalution

Correlation Coefficient & Covariance

> The following table is the correlation coefficient

Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chrom
1.0	0.6449125043512702	0.6545890800019234	0.4863562436767019	0.5218162199598521	0.5892964878134132	0.5584281622
0.6449125043512702	1.0	0.9068819130525921	0.7055818115571123	0.7517991298771314	0.6845689245821212	0.7557209811
0.6545890800019234	0.9068819130525921	1.0	0.6830792002304751	0.7196684371703601	0.7045287881705936	0.735948454
0.48635624367670194	0.7055818115571123	0.6830792002304751	1.0	0.5995990684254976	0.6657232870666341	0.6667153262
0.5218162199598521	0.7517991298771314	0.7196684371703601	0.5995990684254976	1.0	0.5829042875913927	0.616101840
0.5892964878134132	0.6845689245821212	0.7045287881705937	0.6657232870666341	0.5829042875913927	1.0	0.671544601
0.5584281622853955	0.7557209811005724	0.7359484540232973	0.6667153262640527	0.6161018408718495	0.671544601603347	1.0
0.5358345492129787	0.7228648219063575	0.7194463169532832	0.6033524122167611	0.6288806855890924	0.5720544684445475	0.665877809
0.35003385648596486	0.45869314741651085	0.4389109289282088	0.4176327800568878	0.47910147703474826	0.3427947614621469	0.3441694962

> The following table is coveriance matrix

Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromati
7.928395456464618	5.541164004246757	5.477690191882793	3.910307807715482	3.2534689343351677	6.041049227098883	3.834056839283
5.541164004246757	9.311402699722493	8.224213059179904	6.147785825842078	5.079790613688814	7.6051994047985	5.62299396190
5.477690191882793	8.224213059179904	8.832265495939772	5.796567753360308	4.73592647703843	7.622907879041274	5.33312837414
3.910307807715482	6.147785825842078	5.796567753360308	8.153190599751598	3.7910645990383207	6.920594709593332	4.64197523273
3.2534689343351677	5.079790613688814	4.73592647703843	3.7910645990383207	4.903123988013978	4.699152698697676	3.32649999385
6.041049227098883	7.6051994047985	7.622907879041274	6.920594709593332	4.6991526986976 76	13.254756078064869	5.96155170505
3.834056839283298	5.622993961902175	5.333128374140713	4.641975232731163	3.326499993851224	5.961551705055526	5.94562022701
4.6072346495812715	6.735682575599183	6.529071411881906	5.260800324655356	4.252278121426019	6.359752573262675	4.95804075408
1.690388643621056	2.4005660972900325	2.2371562321941694	2.0452303946284367	1.8194821910957506	2.1404441875622537	1.43931158306

Duplicated Tuples

Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	class
466906	1	1	1	1	2	1	1	1	1	2
1321942	5	1	1	1	2	1	3	1	1	2
704097	1	1	1	1	1	1	2	1	1	2
320675	3	3	5	2	3	10	7	1	1	4
1198641	3	1	1	1	2	1	3	1	1	2
1100524	6	10	10	2	8	10	7	3	3	4
1218860	1	1	1	1	1	1	3	1	1	2
1116116	9	10	10	1	10	8	3	3	1	4

The data above represents the duplicated tuples removed by WEKA

Redundancy analysis over, Table 1. shows that the unified Cell Shape and Cell Smooth are highly related to other attributes while Mitoses is relevently lower.

And, there are 8 duplicated d Pata tuples in the original data.