Imputing both 'missing' and 'censored' values in multivariate projection methods, a 'missing' link in pharmacological data mining?

Christophe Buyck
Tibotec BVBA, Gen.
De Wittelaan L11B 3 2800 Mechelen, Belgium

Multivariate projection methods like Principal Component Analysis (PCA), Partial Least Squares (PLS), Spectral Mapping Analysis (SMA) and many others aim at reducing the dimensionality of the data matrix with a minimal loss of information. Such well-established data reduction techniques are quite often used in data visualization and knowledge gathering in drug discovery research.

However, drug discovery data matrices are seldom complete. In a classical drug discovery data matrix, each row represents a different chemical compound and each column displays the readout from a different biological experiment. Typically, not every compound is tested in every biological assay of the matrix. Hence, the resulting data matrix will be a sparse matrix that could contain from 10 up to 90 percent missing values.

A second characteristic of pharmacological data is the use of censors. The classical example is the determination of the 50% inhibitory concentration (IC50) of a chemical compound on substrate conversion by an enzyme. No effect is seen at infinite compound dilution, while at a high compound concentration the enzyme is completely blocked. As the aim is to identify the more active compounds, most compounds are not screened at very high concentrations and compounds that do not inhibit the enzyme at the highest measured concentration will appear as censored values in the matrix.

The appearance of missing and censored data is not random (e.g. some assays are more difficult than others and are used as second line screens only). Therefore, distribution of holes in the matrix is likewise not random. A typical data matrix could look like a [100 (compounds) *100 (assays)] 50% sparse matrix with a non-random distribution of missing and censored data. Robust (partial) visualization and importance-based highlighting of such data is the Holy Grail in chemoinformatics data mining.

Disregarding missing values and/or censors, or simply setting them to zero is a very dangerous approach. Alternatively, missing and censored elements could be imputed from an ANOVA model or a similar approach. While better than simply setting the elements to zero, this approach is still not appropriate and ignores the (hidden) available information.

There are several algorithms described in literature for the estimation of missing values in expression profiles. Examples are based on K-nearest neighbor (KNNimpute) or on SVD (SVDimpute). In addition, probabilistic PCA (PPCA), Bayesian PCA (BPCA) and NIPALS are examples of such algorithms. All these algorithms are available in R, making them easily accessible. In addition, various algorithms are available for censored data imputation.

If censored data are turned into missing data, the above-mentioned algorithms can be evaluated on (simulated) pharmacological sparse matrices. These algorithms do however not deal with the extra information available in censored data.

A typical characteristic of pharmacological data that can be beneficial for imputing censored and missing values is the existence of potential Quantitative Structure-Activity Relationships (QSAR) and chemical structural descriptors. This characteristic enables the production of additional (temporary) columns that reduce the sparseness and might be quite beneficial in the imputation. Finding the relevant descriptors that help in solving the original problem (feature selection) is an open issue in this approach.

Key questions for this workshop include the availability, applicability and robustness of algorithms that cope with both missing and censored data. What are the potential pitfalls? Can possible solutions deal with large-scale data?