Data Cleaning

Barbara Calabrese, University "Magna Graecia", Catanzaro, Italy

Introduction

Raw experimental data is highly susceptible to noise (i.e., contain errors or outliers), missing values (i.e., data lack attribute values, lack certain attributes of interest, or contain only aggregate data) and inconsistency (i.e., data contain discrepancies in codes or names). The representation and quality of the data is crucial because the quality of data affects the data analysis pipeline and thus results. Irrelevant and redundant information in the data, or noisy and unreliable data, could prevent the correct analysis (Han *et al.*, 2012).

Data pre-processing is a fundamental step in the data analysis process to limit such problems. Data pre-processing methods are divided into following categories: (i) data cleaning, that aims to fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies; (ii) data integration, i.e., the merging of data from multiple data stores; (iii) data transformation, that includes normalization and aggregation methodologies, and (iv) data reduction, that attempts to reduce the volume, but produces the same or similar analytical results. In the following paragraphs, the main data cleaning methodologies are presented and discussed.

Data Cleaning

Data cleaning includes all methodologies whose aims are "detecting and removing errors and inconsistencies from data in order to improve the quality of data" (Lenzerini, 2002). Specifically, data quality methods "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers and resolving inconsistencies (Van den Broeck *et al.*, 2005). Instance-level cleaning refers to errors within the data itself; schema level usually concerns integrating the databases into a new schema. Data cleaning is the major part of ETL (extraction, translation and loading) process in data warehouse.

Data Cleaning Methods for Missing Values in Bioinformatics

The presence of missing data may depend on (Herbert and Wang, 2007):

- Malfunctions of data collection systems;
- Inconsistency with values of other data set attributes;
- Not entered data, due to "misunderstanding";
- Some data may not be considered important at the time of insertion;
- Failure to register changes in data.

There are several approaches to solve missing the data problems. A first approach is "do not do anything". Some analysis algorithms are quite robust and insensitive to missing data. Another approach is to eliminate the records that contain missing data: this could introduce a distortion in the data. Otherwise, if only few columns have these characteristics it is better to ignore, but this is not very effective. An alternative approach is to predict new values by using different methods. It is possible to use mean or median value or to pad unknown values with zeros. These methods could cause some big imputation errors. To reduce these errors, more sophisticated methods, such as classification/regression techniques, could be used to estimate missing values.

In Liew *et al.* (2011), a comprehensive survey about missing value estimation techniques for microarray gene expression data is reported. A first approach for missing value estimation exploits the correlation structure between entries in the data matrix. In fact, in gene expression data matrix, rows are correlated because the genes involved in similar cellular processes usually have similar expression profiles. Moreover, there is correlation between columns because the set of genes is expected to behave similarly under similar conditions. Thus, missing values estimation can be based on subset of related genes or subset of related conditions. Another approach exploits the domain knowledge about the data or the process that generated the data. A recent work (Wei *et al.*, 2018) present a review and comparison of eight missing values imputation techniques (zero, half minimum, mean, median, random forest, singular value decomposition, KNN and QRILC quantile regression imputation of left-censored data) for different types of missing values using metabolomics datasets. Generally, missing values imputation methodologies could be grouped in the following classes:

- Global approach-based algorithms;
- Local approach-based algorithms;
- Hybrid approach-based algorithms;
- Knowledge assisted approach-based algorithms.

Global Approach-Based Algorithms

Global approach-based algorithms perform missing value estimation analysing the global correlation information of the entire data matrix. The most used algorithms are the Single Value Decomposition (*SVDimpute*) method and the Bayesian Principal Component Analysis (*BPCA*) method (Oba *et al.*, 2003).

SVD method employs singular value decomposition to obtain a set of mutually orthogonal expression patterns that can be linearly combined to approximate the expression of all genes in the data set (Troyauskaya *et al.*, 2001). These patterns, which in this case are identical to the principle components of the gene expression matrix, are referred to as eigengenes. SVDimpute first regresses the gene against the k most significant eigengenes and then use the coefficients of the regression to reconstruct the missing values from a linear combination of the k eigengenes.

In BPCA, the N-dimensional gene expression vectors γ is expressed as a linear combination of k principal axis vectors v_l. The factor scores w_l and the residual error ϵ are regarded as normally distributed random variables in the probabilistic PCA model (Eq. (1)).

$$\gamma = \sum_{k} w_l v_l + \varepsilon \tag{1}$$

An EM-like algorithm is then used to estimate the posterior distributions of the model parameter and the missing values simultaneously (Liew *et al.*, 2011).

Local Approach-Based Algorithms

The second class of algorithms examine only local similarity structure in the dataset in order to estimate missing value. The most used algorithm is KNN (k-Nearest Neighbors). The KNN-based method (*KNNimpute*) selects genes with expression profiles similar to the gene of interest to impute missing values. If we consider gene A that has one missing value in experiment 1, this method would find K other genes, which have a value present in experiment 1, with expression most similar to A in experiments 2–N (where N is the total number of experiments). A weighted average of values in experiment 1 from the K closest genes is then used as an estimate for the missing value in gene A (Troyauskaya *et al.*, 2001). In the weighted average, the contribution of each gene is weighted by similarity of its expression to that of gene A. Euclidean distance is the most accurate metrics for gene similarity. In Troyanuskaya (2001), KNN- and SVD-based methods were compared for DNA microarray missing value estimation. Both methods provide fast and accurate ways of estimating missing values for microarray data and far surpass the traditional accepted solutions, such as filling missing values with zeros or row average, by taking advantage of the correlation structure of the data to estimate missing expression values. Nevertheless, the authors recommend KNN-based method for imputation of missing values, because it shows less deterioration in performance with increasing percent of missing entries. In addition, the KNNimpute method is more robust than SVD to the type of data for which estimation is performed, performing better on non-time series or noisy data. KNNimpute is also less sensitive to the exact parameters used (number of nearest neighbors), whereas the SVD-based method shows sharp deterioration in performance when a non-optimal fraction of missing values is used.

A number of local imputation algorithms that use the concept of least square regression to estimate the missing values, have been proposed. In least square imputation (*LSimpute*) (Bø *et al.*, 2004), the target gene y and the reference gene x are assumed to be related by a linear regression model. LSimpute first select the K most correlated genes based on absolute Pearson correlation values. Then a least square estimate of the missing value is obtained from each of the K selected genes using single regression. Finally, the K estimates are linearly combined to form the final estimate.

Unlike LSimpute, local least square imputation (*LLSimpute*) (Kim *et al.*, 2005) uses a multiple regression model to impute the missing values from all K reference genes simultaneously. Despite its simplicity, LLSimpute has been shown to be highly competitive compared to KNNimpute and BPCA (Brock *et al.*, 2008).

Sequential LLSimpute (*SLLSimpute*) is an extension of LLSimpute (Zhang *et al.*, 2008). The imputation is performed sequentially starting from the gene with the least missing rate, and the imputed genes are then used for later imputation of other genes. However, only genes with missing rate below a certain threshold are reused since genes with many imputed missing values are less reliable. SLLSimpute has been shown to exhibit better performance than LLSimpute due to the reuse of genes with missing values.

In iterated LLSimpute (*ILLSimpute*) (Cai *et al.*, 2006), different target genes are allowed to have different number of reference genes. The number of reference genes is chosen based on a distance threshold which is proportional to the average distance of all other genes to the target gene. ILLSimpute iteratively refines the imputation by using the imputed results from previous iteration to re-select the set of coherent genes to re-estimate the missing values until a preset number of iterations is reached. ILLSimpute has been shown to outperform the basic LLSimpute, KNNimpute and BPCA due to these modifications.

In Gaussian mixture clustering imputation (*GMCimpute*) (Ouyang *et al.*, 2004), the data is clustered into S components Gaussian mixtures using the EM algorithm. Then the S estimates of the missing value, one from each component, are averaged to obtain the final estimate of the missing value. The clustering and estimation steps are iterated until the cluster memberships of two consecutive iterations are identical. GMCimpute uses the local correlation information in the data through the mixture components.

In Dorri *et al.* (2012) an algorithm that is based on conjugate gradient (**CG**) method is proposed to estimate missing values. k-nearest neighbors of the missed entry are selected based on absolute values of their Pearson correlation coefficient. Then a subset of genes among the k-nearest neighbors is labeled as the best similar ones. CG algorithm with this subset as its input is then used to estimate the missing values.

MINMA (Missing data Imputation incorporating Network and adduct ion information in Metabolomics Analysis) (Jin *et al.*, 2017) implements a missing value imputation algorithm for liquid chromatography-mass spectrometry (LC MS) metabolomics. MINMA is an R package whose algorithm combines the afore-mentioned information and traditional approaches by applying the support vector regression (SVR) algorithm to a predictor network newly constructed among the features. The software provides a function to match feature m/z values to about 30 positive adduct ions, or over 10 negative adduct ions.

Hybrid Approach-Based Algorithms

The correlation structure in the data affects the performance of imputation algorithms. If the data set is heterogeneous, local correlation between genes are dominant and localized imputation algorithms such as KNNimpute or LLSimpute perform better than global imputation methods such as BPCA or SVDimpute. On the other hand, if the data set is more homogenous, a global approach such as BPCA or SVDimpute would better capture the global correlation information in the data (Liew *et al.*, 2011). Jornsten *et al.* (2005) proposes a hybrid approach called *LinCmb* that captures both global and local correlation information in the data.
In LinCmb, the missing values are estimated by a convex combination of the estimates of five different imputation methods: row average, KNNimpute and GMCimpute, that use local correlation information in the estimation of missing values, and SVDimpute and BPCA that are global-based methods for missing values imputation.

To obtain the optimal set of weights that combine the five estimates, LinCmb generates fake missing entries at positions where the true values are known and uses the constituent methods to estimate the fake missing entries. The weights are then calculated by performing a least square regression on the estimated fake missing entries. The final weights for LinCmb are found by averaging the weights obtained in 30 iterations (Liew *et al.*, 2011).

Knowledge Assisted Approach-Based Algorithms

The algorithms that belong to this category exploit the integration of domain knowledge or external information into the missing values estimation process. The use of domain knowledge has the potential to significantly improve the estimation accuracy, especially for data sets with small number of samples, noisy, or with high missing rate (Liew *et al.*, 2011). Knowledge assisted approach-based algorithms can make use of, for example, knowledge about the biological process in the microarray experiment (Gan *et al.*, 2006), knowledge about the underlying biomolecular process as annotated in Gene Ontology (GO) (Tuikkala *et al.*, 2006), knowledge about the regulatory mechanism (Xiang *et al.*, 2008), information about spot quality in the microarray experiment (Johansson and Hakkinen, 2006), and information from multiple external data sets (Sehgal *et al.*, 2008).

Data Cleaning Methods for Noisy Data in Bioinformatics

Noise is a random error or variance in a measured variable. In biological experiments, such as genomics, proteomics or metabolomics, noise could arise from human errors and/or variability of the system itself. An accurate experimental design and technological advances help to reduce noise, but some uncertainty remains always. Specifically, noisy data could include:

- Input errors due to operators;
- Incorrect data related to transcription operations;
- Input errors due to programs;
- Incorrect data related to programming errors;
- Errors due at a glance of insertion;
- Data stored in different formats for the same attribute.

In the following some methods to manage and pre-process noisy data are described. Binning methods smooth a sorted data value by examining the value of its "neighbourhood". In smoothing by *bin means*, each value in a bin is replaced by the mean value of the bin. Similarly, smoothing by *bin medians* can be employed, in which each bin value is replaced by the bin median. In smoothing by *bin boundaries*, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value. In general, the larger the width, the greater the effect of the smoothing (Han *et al.*, 2012).

Data smoothing can also be performed by regression. Linear regression involves finding the "best" line to fit two attributes (or variables) so that one attribute can be used to predict the other. Multiple linear regression is an extension of linear regression, where more than two attributes are involved and the data are fitted to a multidimensional surface.

Clustering techniques are also applied to noise detection tasks. Clustering algorithms, used for biological data reduction are particularly sensitive to the noise. In Sloutsky *et al.* (2012), a detailed study relative to the sensitivity of clustering algorithms to noise has been presented, by analysing two different case studies (gene expression and protein phosphorylation).

Another approach employs Machine Learning (ML) classification algorithms, which are used to detect and remove noisy examples.

The work presented in Libralon *et al.* (2009) proposes the use of distance-based techniques for noise detection. These techniques are named distance-based because they use the distance between an example and its nearest neighbors. The most popular distance-based technique is the k-nearest neighbor (k-NN) algorithm. Distance-based techniques use similarity measures to calculate the distance between instances from a data set and use this information to identify possible noisy data. Distance-based techniques are simple to implement and do not make assumptions about the data distribution. However, they require a large amount of memory space and computational time, resulting in a complexity directly proportional to data dimensionality and number of examples., which is the simplest algorithm belonging to the class.

Libralon *et al.* stated that for high dimensional data sets, the commonly used Euclidian metric is not adequate, since data is commonly sparse. Thus, they use the HVDM (Heterogeneous Value Difference Metric) metric to deal with high dimensional data. This metric is based on the distribution of the attributes in a data set, regarding their output values, and not only on punctual values, as is observed in the Euclidian distance and other similar distance metrics.

Outlier Detection

Outliers detection is a fundamental step in the preprocessing stage aiming to prevent wrong results. To detect anomalous measurements and/or observations from normal ones, data mining techniques are widely used (Oh and Gao, 2009). Generally, statistical methods often view objects that are located relatively far from the center of the data distribution as outlier. Several distance measures were implemented, such as Mahalanobis distance. Distance-based algorithms are advantageous since model learning is not required.

Outliers may be detected by clustering, for example, where similar values are organized into groups, or clusters. Values that fall outside of the set of clusters may be considered outliers. In Wang (2008), the authors proposed an effective cluster validity measure with outlier detection and cluster merging strategies for support vector clustering. In Oh and Gao (2009), the authors present an outlier detection method based on KL divergence. Angiulli *et al.* (2006) proposed a distance-based outlier detection method which finds the top outliers and provides a subset of the dataset called outlier detection solving set, that can be used to predict if new unseen objects are outliers.

Concluding Remarks

Data cleaning is a crucial step in data analysis pipeline because errors and noise can affect the quality of collected data, preventing an accurate subsequent analysis. Several methods could be used for data cleaning according to the specific required task.

References

Angiulli, F., Basta, S., Pizzuti, C., 2006. Distance-based detection and prediction of outliers. IEEE Transactions on Knowledge and Data Engineering 18, 145–160.

Bø, T.H., Dysvik, B., Jonassen, I., 2004. LSimpute: Accurate estimation of missing values in microarray data with least squares methods. Nucleic Acids Research 32, e34.

Brock, G.N., Shaffer, J.R., Blakesley, R.E., *et al.*, 2008. Which missing value imputation method to use in expression profiles: A comparative study and two selection schemes. BMC Bioinformatics 9, 12.

Cai, Z., Heydari, M., Lin, G., 2006. Iterated local least squares microarray missing value imputation. Journal of Bioinformatics and Computational Biology 45, 935–957.

Dorri, F., Azmi, P., Dorri, F., 2012. Missing value imputation in DNA microarrays based on conjugate gradient method. Computers in Biology and Medicine 42, 222–227.

Gan, X., Liew, A.W., Yan, H., 2006. Microarray missing data imputation based on a set theoretic framework and biological knowledge. Nucleic Acids Research 34, 1608–1619.

Han, J., Kamber, M., Pei, J., 2012. Data Mining: Concepts and Techniques. Elsevier.

Herbert, K., Wang, J.T., 2007. Biological data cleaning: A case study. International Journal of Information Quality 1, 60–82.

Jin, Z., Kang, J., Yu, T., 2017. Missing value imputation for LC-MS metabolomics data by incorporating metabolic network and adduct ion relations. Bioinformatics. 1–7.

Johansson, P., Hakkinen, J., 2006. Improving missing value imputation of microarray data by using spot quality weights. BMC Bioinformatics 7, 306.

Jornsten, R., Wang, H.Y., Welsh, W.J., *et al.*, 2005. DNA microarray data imputation and significance analysis of differential expression. Bioinformatics 21, 4155–4161.

Kim, H., Golub, G.H., Park, H., 2005. Missing Value Estimation for DNA microarray gene expression data: Local least squares imputation. Bioinformatics 21, 187–198.

Lenzerini, M., 2002. Data integration: A theoretical perspective. In: Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 233–246. Madison, Wisconsin.

Libralon, G.L., Ferreira de Carvalho, A.C., Lorena, A.C., 2009. Pre-processing for noise detection in gene expression classification data. Journal of the Brazilian Computer Society 15, 3–11.

Liew, A., Law, N., Yan, H., 2011. Missing value imputation for gene expression data: Computational techniques to recover missing data from available information. Briefings in Bioinformatics 12, 498–513.

Oba, S., *et al.*, 2003. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics 19, 2088–2096.

Oh, J.H., Gao, J., 2009. A kernel-based approach for detecting outliers of high-dimensional biological data. BMC Bioinformatics 10 (Suppl. 4), S7.

Ouyang, M., Welsh, W.J., Georgopoulos, P., 2004. Gaussian mixture clustering and imputation of microarray data. Bioinformatics 20, 917–923.

Sehgal, M.S., Gondal, I., Dooley, L.S., *et al.*, 2008. Ameliorative missing value imputation for robust biological knowledge inference. Journal of Biomedical Informatics 41, 499–514.

Sloutsky, R., *et al.*, 2012. Accounting for noise when clustering biological data. Briefings in Bioinformatics 14, 423–436.

Troyauskaya, O., *et al.*, 2001. Missing value estimation methods for DNA microarray. Bioinformatics 17, 520–525.

Tuikkala, J., Elo, L., Nevalainen, O., *et al.*, 2006. Improving missing value estimation in microarray data with gene ontology. Bioinformatics 22, 566–572.

Van den Broeck, J., Argeseanu Cunningham, S., Eeckels, R., Herbst, K., 2005. Data cleaning: Detecting, diagnosing, and editing data abnormalities. PLOS Medicine 2, e267.

Xiang, Q., Dai, X., Deng, Y., *et al.*, 2008. Missing value imputation for microarray gene expression data using histone acetylation information. BMC Bioinformatics 9, 252.

Wei, R., *et al.*, 2018. Missing value imputation approach for mass spectrometry based metabolomics data. Scientific Reports 8.

Zhang, X., Song, X., Wang, H., *et al.*, 2008. Sequential local least squares imputation estimating missing value of microarray data. Computers in Biology and Medicine 38, 1112–1120.