Automatic Discovery of the Statistical Types of Variables in a Dataset

Isabel Valera
Max Planck Institute for
Software Systems
ivalera@tmpi-sws.org

Zoubin Ghahramani
Department of Engineering
University of Cambridge
zoubin@eng.cam.ac.uk

Data analysis problems often involve pre-processing *raw* data, which is a tedious and time-demanding task due to several reasons: i) raw data is often unstructured and large-scale; ii) it contains errors and missing values; and iii) documentation may be incomplete or inexistent. As a consequence, as the availability of data increases, so does the interest of the data science community to automate this process. In particular, there is a growing number of works which focus on automating the different stages of data pre-processing, including data cleaning [2], data wrangling [4] and data integration and fusion [1].

The outcome of the data pre-processing is commonly a structured dataset, in which each of the objects is described by a set of attributes. However, before being able to proceed with the predictive analytics step of the data analysis process, the data scientist also needs to identify which kind of variables (*i.e.*, real-values, categorical, ordinal, etc.) these attributes represent. This labeling of the data is necessary to select the appropriate machine learning approach to explore, find patterns or make predictions on the data. As a example, a prediction task is solved differently depending on the kind of data to be predicted – while the prediction of categorical variables is usually formulated as a classification task, in the case of ordinal variables it is formulated as an ordinal regression problem. Similarly, a probabilistic modeling approach requires the definition of a likelihood model, which should be selected to fit the observed data – different probability distributions are appropriate for continuous and discrete variable, *e.g.*, while a Gaussian distribution is often use to model real-valued variables, it is not appropriate for categorical data. However, there is a lack of tools to perform such labeling automatically, and therefore, it is still done manually by the data scientist.

In order to complete the framework of techniques for automatic data analysis, it is necessary to derive approaches to automatically identify the different types of variables in a dataset. In this context, while it is possible to distinguish between continuous and discrete variables by simply applying concise logical rules (*e.g.*, by counting the number of unique values), it is difficult to come up with simple rules that, by taking a look to the data, are able to differentiate between certain types of discrete variables (*e.g.*, between categorical and ordinal data) or continuous variables (*e.g.*, between interval and circular variables). Previous work [3] has proposed to distinguish between categorical and ordinal data by comparing the model evidence and the predictive test log-likelihood of ordinal and categorical models. However, this approach can be only used to distinguish between ordinal and categorical data, and it does so by assuming that it has access to a real-valued variable that contains information about the presence of an ordering in the observed discrete (ordinal or categorical) variable. As a consequence, it cannot be easily generalizable to label the data type of all the variables (or attributes) in a dataset.

Our approach. In this work, our aim is to solve this gap by developing a general and scalable tool to automatically discover the statistical types of the variables in the data. In particular, we aim to automatically distinguish among the following types of data:

- *Continuous variables*:
  1. Real-valued data, i.e., $x \in \Re$.
  2. Positive real-valued data, i.e., $x \in \Re^+$.
  3. Interval data, *i.e.*, $x \in [a, b]$ with $a, b \in \Re$.

NIPS 2016 Workshop on Artificial Intelligence for Data Science (AI4DataSci), Barcelona, Spain.

   4. Circular or periodic data [5], *e.g.*, $x \in [0, 2\pi)$ or $x \in [0, 24)$.

- *Discrete variables*:
   1. Categorical data, which takes values in a finite unordered set, *e.g.*., $x \in \{$'blue', 'red', 'black'$\}$.
   2. Ordinal data, which takes values in a finite ordered set, *e.g.*, $x \in \{$'never', 'sometimes', 'often', 'usually', 'always'$\}$.
   3. Count data, *i.e.*, $x \in \{0, 1, \dots, \infty\}$.

To this end, we propose a Bayesian method that exploits the latent structure in the data to discover the statistical types of the data. More in detail, the proposed method is based on the following two key ideas:

- There exist some dependencies among the attributes describing each object in a dataset, and these dependencies can be capture by a low rank representation of the $N \times D$ observed dataset $\mathbf{X}$, such that $\mathbf{X} \approx f(\mathbf{ZB})$, where $f(\cdot)$ represent a transformation from the real numbers to the attribute space [6]. Under this low-rank representation, the likelihood function factorizes as

$$p(\mathbf{X}|\mathbf{Z}, \mathbf{B}) = \prod_{d=1}^{D} p(\mathbf{x}^d|\mathbf{Z}, \boldsymbol{b}^d),$$

  where $\mathbf{x}^d$ and $\boldsymbol{b}^d$ denote the $d-$th column of, respectively, $\mathbf{B}$ and $\mathbf{X}$.
- The likelihood function of each attribute describing the data can be written as mixture of likelihood models, one likelihood model per each type of data, as

$$p(\mathbf{x}^d|\mathbf{Z}, \{\boldsymbol{b}_j^d\}_{j=1}^J) = \sum_{j=1}^{J} w_j^d p_j(\mathbf{x}^d|\mathbf{Z}, \boldsymbol{b}_j^d),$$

  where w_j^d denotes the weigh of the likelihood model $p_j(\mathbf{x}^d|\mathbf{Z}, \boldsymbol{b}_j^d)$ in the d-th attribute in the data. Note that, the expression above corresponds to a valid likelihood model as long as $\sum_{j=1}^J w_j^d = 1$ and the likelihood functions $p_j(\mathbf{x}^d|\mathbf{Z}, \boldsymbol{b}_j^d)$ are normalized probability distributions. Ideally, each attribute corresponds to a unique type of data and, therefore, the weight vector $\boldsymbol{w}^d = (w_j^d)_{j=1}^J$ is a sparse vector containing a single one associated to the corresponding type of data. However, we can exploit this representation to infer the type of data (*i.e.*, the likelihood model) that better fits each attribute in the dataset.

We derive an efficient MCMC inference algorithm to infer both the low-rank representation and the weight of each likelihood model for each attribute of the observed data. This algorithm scales linearly with the number of observations N and likelihood functions J, and can be parallelized in the number of attributes D. Our preliminary results show that the proposed method can accurately discover the statistical types of the variables automatically in a large number of datasets.

References

[1] X. Luna Dong and D. Srivastava. Big data integration. In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, pages 1245–1248. IEEE, 2013.

[2] J. M. Hellerstein. Quantitative data cleaning for large databases, 2008.

[3] D. Hernandez-Lobato J. M. Hernandez-Lobato, J. R. Lloyd and Z. Ghahramani. Learning the semantics of discrete random variables: Ordinal or categorical? In *NIPS Workshop on Learning Semantics*, 2014.

[4] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4):271–288, 2011.

[5] A. Navarro, R. E. Turner, and J. Frellsen. The multivariate generalised von mises: Inference and applications. *arXiv preprint arXiv:1602.05003*, 2016.

[6] I. Valera and Z. Ghahramani. General table completion using a bayesian nonparametric model. In *Advances in Neural Information Processing Systems 27*, 2014.