Expert Systems With Applications 73 (2017) 69–81

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Approximate NORTA simulations for virtual sample generation

Guillaume Coqueret[a,b,*]

[a] Montpellier Business School, 2300 avenue des Moulins, 34080 Montpellier, France
[b] Fundvisory, 48 rue du Château Landon, 75010 Paris, France

ARTICLE INFO

Article history:
Received 31 August 2016
Revised 25 November 2016
Accepted 19 December 2016
Available online 21 December 2016

Keywords:
NORTA simulation
Multivariate sampling
Regression trees
Support Vector Machine

ABSTRACT

We introduce an approximate variant of the NORTA method which aims at generating structured data from a given prior sample. The technique accommodates for any combinations of marginals (especially continuous/discrete mixtures) and a wide range of correlation structures. We focus on the interesting case where the sample includes categorical data, both ordered and unordered. We provide an application in the financial industry through a test of our iterative Newton-like algorithm on a dataset comprising the results of a questionnaire. We show that the sampled data, similarly to the NORTA technique, matches both the marginal and correlation structures of the original dataset closely. Consequently, analyses such as decision tree modeling or Support Vector Machine classification and regression, can be carried out on the new, much larger, sample without altering the core properties of the original sample.

1. Introduction

With the *Big Data* hype, people tend to forget that very small companies (e.g. start-ups) are not overwhelmed by data, but, in contrast, strive for it. Given the promises of data science, they hope they can benefit from advanced machine learning and data mining tools[1] but they very often lack the essential ingredient: the data. The present article is meant to help them generate artificial data so that they can perform larger scale analyses at very early stages of their development (business plan construction, commercial pitch elaboration, etc.). This is typically useful for the purpose of testing classification algorithms because they are much more reliable when applied on large datasets.

The sampling of random variables is a well-known topic (see Gentle, 2006 and Kelton and Law, 2000 for overviews) which is quite mainstream in the Operational Research literature (e.g. Chen, 2005; Grimlund, 1992 and Ilich, 2009 to cite but a few). While it is quite straightforward to generate simple multivariate distributions (Gaussian, Student, Uniform, etc.), more exotic laws require ad-hoc simulation techniques.

The existing literature on data generation is vast and we review it briefly in Section 2. Our work relates closely to NORTA (NORmal To Anything) simulations which is one of the most flexible approaches even though it comes with a strong limitation. The idea behind NORTA is to combine a Gaussian generator with the cumulative distribution function inversion theorem so as to obtain any sought marginal distributions. Unfortunately, the correlations that can be reached are bounded by the choice of marginals, and this can be too restrictive. The present article aims at addressing this drawback.

The main application of our methodology is to generate large samples with customized marginals, both continuous and discrete (categorical). This is useful in order to augment small datasets once their statistical properties have been identified. Apart from the marginal distribution of the variables, this means to characterize their dependence structure. While the most general structure would require copula-based analyses, we choose, for tractability purposes, to simply work with Pearson correlations. This is of course imposed by the choice of the Gaussian distribution as intermediate sampling tool.

When the original dataset includes categorical data (as is often the case), analyses are not as mainstream as with numerical data.[2] We handle categorical variables in the two steps. First, we treat unordered (nominal) variables separately and within each

* Correspondence to: Montpellier Business School, 2300 avenue des Moulins, 34080 Montpellier, France.
E-mail address: g.coqueret@montpellier-bs.com

[1] Possible applications are e.g. logistics (Waller & Fawcett, 2013), recommendation (Adomavicius & Zhang, 2012), fraud detection (Kirkos, Spathis, & Manolopoulos, 2007; Ngai, Hu, Wong, Chen, & Sun, 2011), credit scoring (Huang, Chen, & Wang, 2007) and customer relationship management (Jahromi, Stakhovych, & Ewing, 2014; Malthouse, Haenlein, Skiera, Wege, & Zhang, 2013; Ngai, Xiu, & Chau, 2009; Shaw, Subramaniam, Tan, & Welge, 2001; Spiess, T'Joens, Dragnea, Spencer, & Philippart, 2014). Large scope reviews of applications can be found in Chen, Chiang, and Storey (2012), Power (2014) and Provost and Fawcett (2013).

[2] We refer to the monograph Agresti and Kateri (2011) for more details on the subject.

http://dx.doi.org/10.1016/j.eswa.2016.12.027
0957-4174/

such category, we transform the ordered categorical (ordinal) variables into numerical variables. We are then able to work with numerical variables only and the relationships between them is then simply modeled via their sample correlation.

The determination of the marginal distribution of numerical and ordinal variables can also be a genuine challenge when the sample frequencies are far from those of the actual population we try to model. It can hence be convenient to be able to impose a user-specified distribution in this case. However, the choice of a marginal distribution may enter in conflict with the sample covariance structure observed with the initial data and this is why the NORTA approach can fail in our framework.

Our main contribution is thus to propose a variant of the NORTA technique which can handle seemingly incompatible marginals and correlation structures. It is applied to each subset of nominal variable separately and ordinal variables are treated as (i.e. converted into) discrete numerical items. Our procedure is iterative (Newton-like) and the correlation structure of the simulated data converges to the sought correlation matrix. The modeling of marginals is left to the appreciation of the user. We test our method on a use-case and show with Monte-Carlo sampling and regression trees that the core properties of the original sample are not strongly altered by the simulation process. Finally, we test several Support Vector Machine (SVM) schemes on two samples generated via our algorithm.

The remainder of the paper is structured as follows. In Section 3, we formally detail the context, notations and algorithms at the core of our methodology. Section 4 is dedicated to the numerical analyses. Section 5 concludes. Finally, we prove in Appendix A a theoretical result required by our procedure.

2. Related work

In this section, we provide a brief overview of the similar methods already developed in the literature.

2.1. Duplication with noise

The most simple way to increase the size of a dataset without altering its properties is to duplicate it and add some uniform noise. Several studies have shown the link between this process and Tikhonov regularization (e.g. An, 1996; Bishop, 1995 and Lee, 2000). More precisely, resorting to averaging across noisy samples for regressions (or generalizations such as neural networks) is equivalent to introducing a quadratic penalization in the minimization of the error term. More recently, Yang, Yu, Xie, and Zhang (2011) discuss the application of duplication with Gaussian noise to improve classification algorithms based on small datasets.

Now, this approach can have two drawbacks for some applications. The first drawback is when the marginals of the sample does not faithfully reflect the distribution of the population. This can happen for instance if you consider the age of a population which can be fairly well approximated by a continuous unimodal distribution (in many cases). If the sample is very small and essentially multimodal, adding noise will not be sufficient to obtain a correct distribution (the introduction of a prior is a possible solution).

A second drawback is when dealing with discrete distributions with very limited support (a few points), which happens with categorical variables. In this case, the generation of white noise is at best complicated because shifting categories has a crucial impact on the relationships between the variables, especially when the categorical variables are unordered.

2.2. Data augmentation

Our research question is of course closely linked to the topic of data augmentation. Beyond the seminal works of Tanner and Wong (1987) and Wei and Tanner (1990), the interested reader can rely on the comprehensive survey of Van Dyk and Meng (2012). In contrast to these techniques, our approach does not require the intricate apparatus of bayesian statistics. Moreover, as is shown in the numerical applications of Van Dyk and Meng (2012), data augmentation is not straightforward, even for well-known parametric distributions.[3] Satisfying our goal requires customized distributions (both for marginals *and* dependence structures) and it is not clear how to proceed with data augmentation with very general multivariate distributions.

2.3. Parametric approaches

Beyond the classical methods which aim at simulating random variables (rejection–acceptance Von Neumann, 1951, cdf inversion technique – see also Chapter 4 of Gentle, 2006 for an exhaustive treatment), some techniques are focused on parametric families (Gaussian Box, Muller et al., 1958, Stable Chambers, Mallows, & Stuck, 1976). Overviews can be found in Johnson, Kotz, and Balakrishnan (1994) for univariate Cauchy, Gamma, Chi-square distributions and in Balakrishnan and Lai (2009) for univariate beta, Student and Weibull distributions and bivariate Cauchy, Pearson, beta, Liouville, extreme value distributions. Chapter 5 in Gentle (2006) and Chapter 14 in Balakrishnan and Lai (2009) provide useful summaries of these techniques. The introduction of Chen (2001) provides many additional examples of generation techniques in which the marginals belong to the same parametric families. These families are of course overly simplistic and incompatible with combinations of continuous and discrete marginals.

2.4. Recent data science driven techniques

The rise of data science needs and applications has led to the development of ad-hoc methods, depending on the application topic. To cite but a few: hypothesis testing (Li & Lin, 2013), neural networks algorithms (Li & Fang, 2009; Li & Wen, 2014), support vector approaches (Yang et al., 2011) or other classification tools (Li & Lin, 2014) and finally, prediction (Li, Chang, & Liu, 2012). The purpose of these methods is often the same as that of the present paper: to enhance and sharpen the results of quantitative tools on small datasets. Since the statistical machinery is very often more efficient when the scale is large, the generation of additional (relevant) data improves the performance of the algorithms and reduces errors and biases. For instance, in Li and Wen (2014), megatrend diffusion functions are combined to a genetic algorithm to generate virtual samples so as to increase the accuracy of backpropagation neural networks. Unfortunately, it is unclear how to adapt such technical tools to categorical data.

2.5. The NORTA generation method

Even though it could be included in some of the previous categories, we detail the NORTA approach separately. While the origins of NORTA go back to the 1970s at least, the core theoretical results were proven in Cario and Nelson (1997) and practical details are studied in Chen (2001). The idea is fairly simple: if one generates a multivariate Normal vector, then applying cdf and inverse

[3] As they put it, "*In general, however, constructing data augmentation schemes that result in both simple and fast algorithms is a matter of art in that successful strategies vary greatly with the (observed-data)models being considered.*"