BAYESIAN ANALYSIS OF RELEVANCE IN PRESENCE OF MISSING AND ERRONEOUS DATA

Gábor HULLÁM
Advisors: Péter ANTAL, György STRAUSZ

I. Introduction

As recent technology and newly developed methods enable to address larger, more complex tasks in several different domains from telecommunication to biomedicine, the problem of focusing on the most relevant information has become increasingly important. This task has received significant attention in AI research and became known as the feature subset selection problem (FSS). In most of the cases, identifying a relevant subset of features (input variables) poses a statistical and computational challenge due to the enormous number of potential variables and to the limitations of parametric statistical methods for detecting cause-effect relationships. Motivated by this problem, we proposed a new structural model feature (property) called *Markov Blanket Graph* (MBG) in the Bayesian network framework and we formulated the *most probable MBG* (MP-MBG) problem. We demonstrated its applicability in a biomedical case study. Our current work aims to extend our present framework to handle missing and erroneous data in an integrated way.

II. The issue of relevance

The concept of relevance plays an important role in most scientific studies. Furthermore, the selection of relevant features is a central issue in several diverse fields such as information retrieval or pattern recognition. Throughout the paper we investigate this issue in the field of biomedicine.

In a typical biomedical study hundreds of clinical or genomic factors are measured in each patient sample, thus producing a dataset with a vast number of variables while the number of samples is usually moderate in comparison. In addition, the number of variables grows even further when environmental factors are added. The goal of such studies is to select the relevant features (clinical, genomic or environmental factors) w.r.t. a certain disease or condition. According to John et al.[5], given a set of features $X_i \in \boldsymbol{S}$, a target feature Y, feature relevance can be defined in the following way:

Definition 1: A feature X_i is *strongly relevant*, if there exists some x_i, y and $\boldsymbol{s_i} = x_1, . . . , x_{i-1}, x_{i+1}, . . . , x_n$ for which $p(x_i, \boldsymbol{s_i}) > 0$ such that $p(y \mid x_i, \boldsymbol{s_i}) \neq p(y|\boldsymbol{s_i})$. A feature X_i is *weakly relevant*, if it is not strongly relevant, and there exists a subset of features $\boldsymbol{S_i}' \in \boldsymbol{S_i}$ for which there exists some x_i, y and $\boldsymbol{s_i}'$ for which $p(x_i, \boldsymbol{s_i}') > 0$ such that $p(y \mid x_i, \boldsymbol{s_i}') \neq p(y| \boldsymbol{s_i}')$. A feature is *relevant*, if it is either weakly or strongly relevant; otherwise it is irrelevant.

Current solutions for this well known FSS problem include linear and logistic regression models, SVMs, PCA, likelihood based score tests, structured association methods and even ANOVA. For a detailed overview on the applicable statistical methods see [2]. A common theme in these methods is that they are based upon some form of conditional modeling. There is however another approach that can provide a solution for the FSS problem: the application of Bayesian methods.

III. Bayesian relevance analysis

The Bayesian approach has two main advantages over the others. First, a priori knowledge can be incorporated to aid the feature selection. In case that previous experiments or studies produced valuable information e.g. on clinical or genomic variable correlations, then it would be a waste to ignore these results, since they have the potential to improve the process of selecting relevant features.

Second, the Bayesian approach enables the handling of missing values in an integrated way, while the other methods based on conditional modeling require an additional (imputation) model to resolve that problem. Since the ratio of missing and erroneous values tends to be significant, especially in case of biomedical data, appropriate handling is a central issue. One of the frequent reasons for missing values is that not all factors are analyzed in each sample due to insufficient resources or to improper study protocol. Another typical reason is that the specific value of a certain factor can not be identified unambiguously. In that case the result is either marked as erroneous with a special failed measurement symbol in the result dataset or it is completely missing.

In order to find the relevant features we used our formerly devised Bayesian relevance analysis method [1] based on Markov Blanket Graphs.

Definition 2: A subgraph of Bayesian network structure G is called the *Markov Blanket* (sub)*Graph* MBG(X_i,G) of variable X_i if it includes the nodes of MB(X_i, G) and the incoming edges into X_i and into its children, where MB(X_i, G) denotes the Markov blanket of X_i , i.e. the set of parents, children and the children's other parents for X_i under the implicit assumption that the joint probability distribution p is Markov compatible with G and stable [8].

The advantage of identifying the MBG(Y, G) of a central variable Y is twofold. First, according to definition 1, MBG(Y, G) contains all the variables X_i that are relevant w.r.t. Y. Second, it also contains the dependency relationships between these relevant variables. In terms of biomedical data analysis, when the variable Y denotes the presence or a state of the complex, multifactorial disease under study, the MBG(Y, G) contains a set of relevant factors and their interdependencies that influence that disease mechanism.

However, in case of real-world biomedical applications, the amount of data available for analysis is typically not sufficient to select a single best feature subset, i.e. a single best MBG. The *Most Probable Features* (MPFs) method described in [7] aims to resolve this matter by selecting a predefined K number of feature values with high posteriors, which minimize a given loss function. The *MP-MBG* problem is a specific case of MPFs, with the goal of finding the K most probable MBGs. By selecting the K most probable MBGs the K most probable relevant feature subsets and their interdependencies are identified. In order to facilitate this task we devised an *ordering-based estimation and search method* for Markov Blanket sets and Markov Blanket subGraphs using Markov Chain Monte Carlo sampling and the concept of MBG space as core elements [1].

Since the MBG based relevance analysis requires a completed dataset, the missing and erroneous cases must be handled in some way. The following sections describe possible approaches and methods to solve that problem.

IV. Missing value handling methods

Missing values can be classified into the following three classes [3]:

- **Missing Completely at Random (MCAR):** the probability of a missing value for a variable of interest Y is the same for all units in the population. This means that the probability of missing does not depend on either auxiliary variables X or the variable of interest Y
- **Missing at Random (MAR):** the probability of a missing value for a variable of interest Y is related to auxiliary variable(s) X
- **Not Missing at Random (NMAR):** the probability of a missing value for a variable of interest Y is related to Y or to other variables that were not observed

The correct assessment of missing value types in a dataset is crucial, because the possible ways of handling of missing values depend on it. In the following sections we discuss the most popular methods found in the literature.

A. *Complete cases method*

This is the simplest of all methods, where only the complete records are used for analysis. All other samples with at least one value missing are eliminated. If missing is unrelated to all the variables of interest (MCAR), then this method is free of bias. Even then, by rejecting incomplete samples, valuable information contained within the non-missing values is not utilized. In addition, the elimination of incomplete records may result in an unacceptably low sample size.

B. *Ad-hoc methods*

The methods in this group manipulate the variables in some way. One of the basic approaches is to transform the variable with missing values by recoding. The down side of this simple method is the considerable bias it produces. Another possible action is the elimination of the variable from the analysis. The main risk of this method is obvious, leaving a variable out means losing information, furthermore it may also have a sever impact on the dependency relationship pattern of the variables.

C. *Weighting*

Weighting methods can only be used in such a rare case, when a complete model describing the probabilities of all missing values is available. Then based on the probabilities a weight can be assigned to each of the observed values, which compensate for the missing ones.

D. *Imputation*

This is the preferred way of handling MAR type missing values. The core of imputation is to create an artificial value according to some method and replace the missing value with it. This process eventually leads to a complete data file.

Currently, we are planning to embed different imputation schemes in our Bayesian relevance analysis method and to make a comparison study. In the following section we investigate the different classes of such imputation methods.

V. Imputation schemes

A. *Heuristic filtering and imputation*

These methods either use a random number or an artificial value that is independent from the domain, as imputation source. The most general forms of these methods are *random hot deck*, *mean imputation*, *regression imputation* and *k nearest neighbour imputation*. Another possible approach is to input values according to the most probable value or configuration based on the observations.

B. *Heuristic imputation with background knowledge*

If some domain knowledge is available a priori, then it can be used to aid the imputation of missing values, e.g. in the form of predefined rules.

C. *Multiple imputation*

This class consists of methods that generate plausible values for missing observations that are imputed in every possible way, thus creating multiple completed datasets which are analyzed using complete-data methods. Finally, the results are combined, which allows the uncertainty regarding the imputation to be taken into account. Note, that multiple imputation is originated from Bayesian network methods. Although the basic principle is the same, these approaches are handled separately in the literature.

D. *Bayesian approach – likelihood based methods*

Likelihood based methods have a close connection with Bayesian methods since their implementation is most likely based on Bayesian networks. Typically, the primary goal of these approaches given a target variable Y, predictor variables $\boldsymbol{X}$ and their parameters $\boldsymbol{\theta}$ is to assess the conditional distribution: $P(Y|\boldsymbol{X}, \boldsymbol{\theta})$. When some of the predictors have missing values, then these have to be estimated in a way so that the joint probability distribution is maximized. This is typically achieved through the use of the EM (expectation-maximization) algorithm. This approach has several variants such as the structural EM and the Bayesian structural EM [3].

VI. Asthma case study

In cooperation with SOTE DGCI we participated in a genome-wide association study (GWAS) on the asthma disease, which has a complex pathological mechanism, and possibly several genes are in connection with its symptoms. The goal of the GWAS was to study the genetic variation across the human genome in order to identify genetic associations with observable traits, or the presence or absence of the disease. More specifically we were interested in identifying relevant *single nucleotide polymorphisms* (SNPs) and *haplotypes* [4] that are connected to asthma.

The analyzed dataset consisted of 760 samples (349 cases, 411 controls) with 144 genotyped SNP variables and 20 clinical factors. Regarding only SNPs, 24.46% of the dataset was missing (7.81% completely missing, 16.65% erroneous due to measurement failure). This relatively high measurement failure rate shows the limits of the currently used genotyping technology. It allows the genotyping of only certain SNPs fulfilling strict criteria. In some cases however, even when the criteria are met, the genotype measurement can be unsuccessful, typically due to the poor quality of the sample. In this study, the erroneous values were marked with a special symbol and were processed as missing data.

In order to create a completed dataset for the Bayesian relevance analysis we applied a series of missing value handling methods. As a first step, we examined each SNP variable and removed those which had a missing value rate over 50%. This improved the overall missing value rate by 10.54%. The next step was to filter the samples in a similar way, resulting in a missing rate of 12%. After these preliminary steps, on the reminder of the dataset (685 samples, 122 SNP variables) we proceeded with the most probable value imputation calculated separately for each SNP. Finally, the Bayesian relevance analysis was applied to the completed dataset, revealing multiple relevant SNPs in connection with asthma.

The applied missing value handling methods provided a completed dataset of acceptable quality. In our future works we plan to investigate and integrate further methods into our present framework.

Acknowledgement

I am grateful to the members of SOTE DGCI SNP Lab for the chance to participate in the asthma project and to Peter Antal for many valuable discussions related to this article.

References

[1] P. Antal, G. Hullám, A. Gézsi and A. Millinghoffer, ”Learning complex bayesian network features for classification”, *In Proc. of third European Workshop on Probabilistic Graphical Models*, pp. 9–16, 2006.

[2] D. J. Balding, “A tutorial on statistical methods for population association studies”, *Nature Reviews*, *Genetics*, vol.7, pp. 781-791, Nature Publishing Group, 2006.

[3] N. Friedman, ”The Bayesian structural EM algorithm”, G. F. Cooper, S. Moral, eds., *In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pp. 129-138, Morgan Kaufmann, 1998.

[4] International HapMap Consortium, “The international HapMap Project”, *Nature*, vol. 426, pp. 789-796, NPG, 2003.

[5] G. H. John, R. Kohavi, K. Pfleger, “Irrelevant features and the subset selection problem”, In Proceedings of the 11th International Conference on Machine Learning, pp.121-129. 1994.

[6] N. J. Horton, K. P. Kleinman, “Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models”, *The American Statistician*, Vol. 61, No. 1, pp. 79-90, ASA, 2007.

[7] A. Millinghoffer, G. Hullám and P. Antal, “On inferring the most probable sentences in bayesian logic”, *In Workshop notes on Intelligent Data Analysis in bioMedicine And Pharmacology (IDAMAP-2007)*, pp. 13–18, 2007.

[8] J. Pearl. *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2000.