INSTITUTE OF COMPUTER SCIENCE
The Czech Academy of Sciences

Opponent's report for doctoral thesis of Issam Salman: *Bayesian Networks for Medical Data Analysis*

The dissertation thesis of Salman Issam deals with machine learning methods for the analysis of medical data, and it focuses in more detail on the problem of learning the Bayesian network structure from incomplete, imbalanced, and noisy data.

The dissertation is divided into four chapters and one appendix. Chapter 1 provides an introduction to the thesis including the motivation, statement of the problem, related work, and description of the goals and structure of the dissertation thesis. In this chapter, the Bayesian networks which are the focus of the thesis, are introduced as one of the machine learning methods.

The second chapter provides background to the Bayesian network models. Reasons for choosing Bayesian networks are presented, as well as their use in health care. The author then presents a representation of probability distribution by a Bayesian network and defines the D-separation and some assumptions for learning the causal structure. Finally, some algorithms for learning the Bayesian network structure are presented.

The third chapter contains the main results of the thesis. Section 3.1. which was published in *Turkish Journal of Electrical Engineering and Computer Science* presents a study of heart attack mortality prediction using different machine learning methods, including decision trees, logistic regression, Naive Bayes classifier, and Bayesian network classifiers. Section 3.2. then focuses on tree-augmented naive Bayes algorithms for learning network structure. The study focuses on algorithms for incomplete and imbalanced data. Section 3.3. and 3.4. which were partly published in the *Informatica* journal then propose a new approach to learning optimal Bayesian network structures from incomplete data, and the proposed approach is compared with other methods in a simulation study. Newly added Section 3.5. which has been submitted to *Applied Clinical Informatics* provides analysis of the cardiac Single Proton Emission Computed Tomography images.

The last fourth chapter provides a summary of the thesis.

The thesis contains 40 figures, 24 tables 7 algorithms, and a list of 75 references. The thesis overall reads well. Mathematical notation seems proper and neat. This version of the thesis includes a list specifying the mathematical notation that I believe is beneficial.
I believe, more of the equations could have been numbered so they would have easily been referred to. E.g., the measures of the prediction quality defined on page 28, if numbered, could then be explicitly referred to on subsequent pages to make Section 3.2.4 clearer.

A graphical display of the results seems nice and informative. Only in Figures 3.8.-3.13., I am not sure if the x-axis is properly labeled, "Dataset Size" seems inappropriate. Should this refer to the ratio of missing values? With many abbreviations being used in the thesis it would be helpful if the thesis included a list of abbreviations somewhere at the beginning.

For the replicability of the results, it is very helpful that the code for analyses is available to the readers in an online repository at https://github.com/issamsalman/PhD-thesis-algorithms. This is important not only to allow for replicability of the study, it may also support dissemination of the methods and any follow-up research.

Overall, the thesis offers a comparison of several machine-learning methods for the analysis of medical data. It also brings a new algorithm for learning optimal Bayesian networks, which is shown to be superior in many of the selected designs of the simulation study, and which is then practically implemented on real medical data. I have these questions and points for discussion:

1. Please describe how you selected the designs and true models for your simulation study. For example, what was the motivation behind the true models discussed in Section 3.2.4. and summarized in Table A.1?

The goals of the dissertation were certainly met and new useful theoretical and practical results have been achieved. I believe that the submitted work meets the requirements for a dissertation in the study field and, upon successful defense of the thesis, I recommend awarding the scientific degree Ph.D. to Issam Salman.

doc. RNDr. Patrícia Martinková, Ph.D.
Department of Statistical Modelling
Institute of Computer Science of the Czech Academy of Sciences

February 20, 2024, Prague