

Dissertation Critique: Exploring Machine Learning Techniques Using Patient Interactions In Online Health Forums to Classify Drug Safety

Christopher Jeschke

Engineering for Professionals

Johns Hopkins University

Elkridge, MD 20175, USA

CJESCHKE@GMAIL.COM

Editor: n/a

Abstract

Patient generated health data represents an area of active research interest for its potential applications in monitoring the public health. The study of Pharmacovigilance is one such area, focused on monitoring drugs once they have been released to market. Dr. Brant Chee's 2011 dissertation specifically explores the application of machine learning techniques to messages from online forums discussing patient interactions with drugs in an attempt to assemble a set of machine learning techniques and processes to detect drugs bound for the United States Food and Drug Administration's watch list. Watch list drugs which are those drugs deemed to pose a significant safety concern for consumers. The dissertation discusses a progression of techniques

Keywords: Drug Safety, Pharmacovigilance, NLP

1. Introduction

Probabilistic inference has become a core technology in AI, largely due to developments in graph-theoretic methods for the representation and manipulation of complex probability distributions (?). Whether in their guise as directed graphs (Bayesian networks) or as undirected graphs (Markov random fields), *probabilistic graphical models* have a number of virtues as representations of uncertainty and as inference engines. Graphical models allow a separation between qualitative, structural aspects of uncertain knowledge and the quantitative, parametric aspects of uncertainty...

2. Paper Criteria

The critique should include a summary of the research reported, a discussion of the major contributions claimed, and an assessment of the significance of those contributions and of the research itself. The critique should also include a brief literature review of the topic related to the thesis, discussion of relevant algorithms, and application areas for the research reported. Where appropriate, the critique should include a comparison with other issues discussed in class. Students are encouraged to select a dissertation that is related to their course projects. The evaluation criteria for the critique are as follows: Overview of

the research reported (20 Review of the related literature (15 Major contributions of the thesis (20 Understanding of techniques and algorithms (20 Application areas (15 Proper construction and readability of paper (10

3. Overview

4. Related Literature

5. Major Contributions

6. Techniques and Algorithms

7. Application Areas

Remainder omitted in this sample. See <http://www.jmlr.org/papers/> for full paper.

Acknowledgments

We would like to acknowledge support for this project from the National Science Foundation (NSF grant IIS-9988642) and the Multidisciplinary Research Program of the Department of Defense (MURI N00014-00-1-0637).

Appendix A.

In this appendix we prove the following theorem from Section 6.2:

Theorem *Let u, v, w be discrete variables such that v, w do not co-occur with u (i.e., $u \neq 0 \Rightarrow v = w = 0$ in a given dataset \mathcal{D}). Let N_{v0}, N_{w0} be the number of data points for which $v = 0, w = 0$ respectively, and let I_{uv}, I_{uw} be the respective empirical mutual information values based on the sample \mathcal{D} . Then*

$$N_{v0} > N_{w0} \Rightarrow I_{uv} \leq I_{uw}$$

with equality only if u is identically 0. ■

Proof. We use the notation:

$$P_v(i) = \frac{N_v^i}{N}, \quad i \neq 0; \quad P_{v0} \equiv P_v(0) = 1 - \sum_{i \neq 0} P_v(i).$$

These values represent the (empirical) probabilities of v taking value $i \neq 0$ and 0 respectively. Entropies will be denoted by H . We aim to show that $\frac{\partial I_{uv}}{\partial P_{v0}} < 0 \dots$

Remainder omitted in this sample. See <http://www.jmlr.org/papers/> for full paper.

References