Variable Interaction Networks in Medical Data

Stephan M. Winkler, Heuristic and Evolutionary Algorithms Laboratory, Bioinformatics Research Group, University of Applied Sciences Upper Austria, Hagenberg, Austria

Gabriel Kronberger, Heuristic and Evolutionary Algorithms Laboratory, Bioinformatics Research Group, University of Applied Sciences Upper Austria, Hagenberg, Austria

Michael Affenzeller, Heuristic and Evolutionary Algorithms Laboratory, Bioinformatics Research Group, University of Applied Sciences Upper Austria, Hagenberg, Austria

Herbert Stekel, Central Laboratory, General Hospital Linz, Linz, Austria

ABSTRACT

In this paper the authors describe the identification of variable interaction networks based on the analysis of medical data. The main goal is to generate mathematical models for medical parameters using other available parameters in this data set. For each variable the authors identify those features that are most relevant for modeling it; the relevance of a variable can in this context be defined via the frequency of its occurrence in models identified by evolutionary machine learning methods or via the decrease in modeling quality after removing it from the data set. Several data based modeling approaches implemented in HeuristicLab have been applied for identifying estimators for selected continuous as well as discrete medical variables and cancer diagnoses: Genetic programming, linear regression, k-nearest-neighbor regression, support vector machines (optimized using evolutionary algorithms), and random forests. In the empirical section of this paper the authors describe interaction networks identified for a medical data base storing data of more than 600 patients. The authors see that whatever modeling approach is used, it is possible to identify the most important influence factors and display those in interaction networks which can be interpreted without domain knowledge in machine learning or informatics in general.

Keywords: Data Based Modeling, Data Mining, Evolutionary Algorithms, Medical Data Analysis, Variable Interaction Networks

1. INTRODUCTION, RESEARCH GOALS

In this paper we present research results achieved within the Josef Ressel Centre for Heuristic Optimization Heureka! (http://heureka.heuristiclab.com/): Data of thousands of patients of the General Hospital (AKH) Linz, Austria, have been analyzed in order to identify mathematical models for cancer diagnoses. We have used a medical database compiled at the central laboratory of AKH in the years 2005 – 2011: A series of routinely measured blood values of thousands of patients are available. The results of empirical research work done on the data based identification of estimation

DOI: 10.4018/ijphim.2013070101
models for standard blood parameters are presented in this paper: The main goal is to generate mathematical models for standard blood parameters using other available parameters in this data set. For each variable we identify those variables that are most relevant for modeling it; relevance of a variable can in this context be defined via the frequency of its occurrence in models identified by evolutionary machine learning methods or via the decrease in modeling quality after removing it from the data set prior to the learning stage.

2. MACHINE LEARNING METHODS APPLIED

In this section we describe the modeling methods applied for identifying estimation models for medical variables and cancer diagnoses: On the one hand we apply hybrid modeling using machine learning algorithms and evolutionary algorithms for parameter optimization and feature selection (as described in Section 2.1), on the other hand we use genetic programming (as described in Section 2.2). In Winkler et al. (2011), for example, these methods have also been described in detail.


Feature selection is often considered an essential step in data based modeling; it is used to reduce the dimensionality of the datasets and often conducts to better analyses. Given a set of \( n \) features, \( F = \{f_1, f_2, \ldots, f_n\} \), our goal here is to find a subset of \( F \), \( F' \), that is on the one hand as small as possible and on the other hand allows modeling methods to identify models that estimate given target values as well as possible. Additionally, each data based modeling method (except plain linear regression) has several parameters that have to be set before starting the modeling process.

The fitness of feature selection \( F' \) and training parameters with respect to the chosen modeling method is calculated in the following way: We use a machine learning algorithm \( m \) (with parameters \( p \)) for estimating predicted target values \( \text{est}(F', m, p) \) and compare those to the original target values \( \text{orig} \); the coefficient of determination \( R^2 \) function is used for calculating the quality of the estimated values. Additionally, we also calculate the ratio of selected features \( |F'|/|F| \). Finally, using a weighting factor \( \alpha \), we calculate the fitness of the set of features \( F' \) using \( m \) and \( p \) by comparing the estimated values to the originally given target values:

\[
\text{fitness}(F', m, p) = \alpha \cdot \frac{|F'|}{|F|} + (1 - \alpha) \cdot \left(1 - R^2(\text{est}(F', m, p), \text{orig})\right) \in [0,1]
\]

In (Alba et al. 2007), for example, the use of evolutionary algorithms for feature selection optimization is discussed in detail in the context of gene selection in cancer classification. We have now used evolutionary algorithms for finding optimal feature sets as well as optimal modeling parameters for models for tumor diagnosis; this approach is schematically shown in Figure 1: A solution candidate is here represented as \( [s_1, \ldots, s_n, p_1, \ldots, p_q] \) where \( s_i \) is a bit denoting whether feature \( F_i \) is selected or not and \( p_j \) is the value for parameter \( j \) of the chosen modeling method \( m \). This rather simple definition of solution candidates enables the use of standard concepts for genetic operators for crossover and mutation of bit vectors and real valued vectors: We use uniform, single point, and 2-point crossover operators for binary vectors and bit flip mutation that flips each of the given bits with a given probability. Explanations of these operators can for example be found in Holland (1975) and Eiben (2003).

We have used strict offspring selection (Affenzeller & Wagner, 2005; Affenzeller et al., 2009): Individuals are accepted to become members of the next generation if they are evaluated better than both parents.
14 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage:

www.igi-global.com/article/variable-interaction-networks-in-medical-data/102626?camid=4v1


www.igi-global.com/e-resources/library-recommendation/?id=2

Related Content

An Evaluation of E-Learning in Healthcare

www.igi-global.com/chapter/evaluation-learning-healthcare/46531?camid=4v1a

Managing Operating Room Cost in Cardiac Surgery with Three Process Interventions

www.igi-global.com/chapter/managing-operating-room-cost-cardiac/13018?camid=4v1a

TEDx Conference Story

www.igi-global.com/article/tedx-conference-story/75187?camid=4v1a
Ambulance Dispatching System with Integrated Information and Communication Technologies on Cloud Environment