Título:

Dimensionality reduction and feature selection applied to modeling of sepsis patients

Orientador

José Borges

Enquadramento (Indicar adicionalmente Ramo/Área de Especialidade caso aplicável):

Health care decision-makers have begun to look toward engineering systems concepts and approaches for solutions to the challenging problems or improving quality and reducing costs. Several recent reports from the U.S. National Academies of Science, Engineering and the Institute of Medicine have highlighted the large number of preventable deaths from medical errors that occur annually, and have suggested that they result from "systems problems" [1,2]. Similar conclusions can be found in Europe, including Portugal. To date, there are very few examples where engineering systems principles have been applied to the healthcare domain in order to improve quality, safety or clinical effectiveness. The exact degree to which the structure or design of a system influences outcome probably varies across clinical settings and conditions, but is likely to be significant in high-acuity, complex environments such as the intensive care unit (ICU) or operating room. ICUs provide a particular opportunity in the hospital to examine the benefits from implementing engineering systems approaches. Patients here are among the sickest patients in the hospital, and decisions that are made can literally mean the difference between life and death. The design and implementation of good interventions has been hampered by our lack of data-driven knowledge about what processes are "broken" and what changes would be most effective.

Objectivos:

The objectives of this dissertation are:

- To develop black-box models (Hybrid Subspace Models, Support Vector Machines and Artificial Neural Networks) from the sepsis dataset in order to predict, within suitable confidence limits, which patients in ICU will experience sepsis.
- To apply feature selection algorithms (Hierarchical Cluster Analysis and Decision Trees) in order to unveil which are the most important features in the sepsis dataset and how do they related with each other.
- To apply dimensionality reduction methods (Principal Component Analysis, Linear Discriminant Analysis and Kernel Principal Component Analysis) to the existing sepsis dataset and, therefore, reduce the amount of data processed by both feature selection and modeling algorithms.

Descrição:

Data collected in ICUs for both problems is aggregated in large-scale datasets consisting of relevant patient data, which is collected over long time periods. The handling of such datasets affects the efficient use of computational methods. In order to improve performance, while alleviating the effect from the curse of dimensionality, two approaches will be followed in this

work: dimensionality reduction and feature selection. The objective is to enhance the model generalization capability, improve the learning process and as well the model interpretability.

An initial data pre-processing step results from applying dimensionality reduction techniques to the complete dataset and, therefore, reduce the amount of data processed by both feature selection and modeling algorithms. Two approaches will be used in this work and their results compared: the application of two linear methods, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), and the application of the nonlinear method Kernel PCA (kPCA) [3].

Feature selection methods also contribute to acquire a better understanding about data, since they unveil which are the important features and how do they related with each other. In this work it is proposed to use the unsupervised Hierarchical Cluster Analysis method and Decision Trees towards this objective.

In our project, models will be derived to perform both classification and prediction and will address the medical condition of sepsis. The approaches proposed in this work to provide prediction models are Hybrid Subspace Identification [4], Support Vector Machines [5] and Artificial Neural Networks [6]. These model structures have proven to be suitable candidates to model non-linear, multivariable systems in an effective way, due to their favorable function approximation properties.

[1] Institute of Medicine, “Crossing the Quality Chasm: A New Health System for the 21st Century”. National Academy Press, pp. 1- 22, Executive Summary, (2001).

[2] National Academy of Engineering and Institute of Medicine, “Building a Better Delivery Sytem: A New Engineering/Health Care Partnership”. National Academies Press, pp. 1-26, (2005).

[3] J.A. Lee and M. Verleysen. Nonlinear dimensionality reduction. Springer, New York, NY, USA, 2007

[4] José Borges (2007). State-Space System Identification: New Developments and Applications. Ph. D. thesis, Instituto Superior Técnico/UTL.

[5] Huang T.-M., Kecman V., Kopriva I. (2006), Kernel Based Algorithms for Mining Huge Data Sets, Supervised, Semi-supervised, and Unsupervised Learning, Springer-Verlag, Berlin

[6] S. Haykin. Neural Networks and Learning Machines (Third Edition), Prentice Hall Pub. November 2008.

Requisitos (e.g. média,disciplinas concluídas):

Sistemas Inteligentes.

Resultado esperado:

- Data pre-processing: the data needed to derive classifiers and predictors for the problem considered is prepared to be used in feature selection and modeling. Dimensionality reduction methods are applied to reduce complexity.

- Implementation of algorithms for feature selection: the implementation of feature selection techniques is expected.
- Modeling: both classification and prediction models will be derived, using hybrid subspace, neural networks and SVM modeling techniques.
- Software tool: Development of a software tool for modeling data mining problems in ICUs performing the following tasks: preprocessing of data; dimensionality reduction; feature selection; selection of the structure of the model; Development of neural models; simplification of the models; simulation and validation.
- Interpretation and evaluation of results: Interpretation and evaluation of the models developed for sepsis.

URL da descrição detalhada da dissertação:

Observações:

Este trabalho está inserido no projecto da Fundação para a Ciência e para a Tecnologia "Systems redesign to improve the survival of critically ill patients using data based modeling", PTDC/SEN-ENR/100063/2008.

Localização da realização da dissertação:

IST - Centro de Sistemas Inteligentes/IDMEC