Babeș-Bolyai University
Faculty of Mathematics and Computer Science

CLASSIFICATION MODELS
FOR MEDICAL AND PSYCHOLOGICAL PROBLEMS

PhD Thesis Summary – English version

Author:
Adriana Mihaela COROIU

Scientific Supervisor:
Phd. Prof. Horia F. POP

Cluj-Napoca
2018

1 Introduction

The present paper is being done at a time of technology presence for most of the standard activities in everyday life.

Artificial Intelligence (*IA*) has evolved rapidly, and we have constantly examples of this issue.

In addition, innovation in this area is becoming more and more important and represents a real help for people who choose to benefit from these research area offers.

Using the advantages offered by *IA* in medical science is becoming more and more necessary and gives better results, impossible to achieve otherwise, without the help of techniques and tools from computer science area. Thus, the latest news resulting from the combination of the two areas are:

- *creating micro-computers inside living cells* to study life system of the body and even reprogram it in some contexts coskun2014computational;
- *creating a "blind robot"* (without a camera or a visual sensors) by MIT researchers, a robot able to detect the optimal path, overcoming all obstacles encountered mrva2015tactile;
- *treating the corneal blindness phenomenon* by making a natural eye tissue similar to the eye using 3D printers [**?**].

The above mentioned facts show once again the importance of using *IA* in the medical sciences and also support the relevance of the present paper.

The paper studies the applicability of artificial intelligence field called *Machine Learning* (*ML*) in two areas important for humanity: medicine and psychology.

The objective of this PhD thesis is to identify and present *ML* (and *IA*) implications in establishing a diagnosis, namely in improving the decision for a particular diagnosis, and also in discovering templates in analyzed data sets and also the dependencies or methods used in order to achieve accurate results for new instances of data sets.

Each chapter of the paper focuses on different aspects, from the perspective of used techniques and from medical (or psychological) perspective. Different analysis and prediction methods are compared to check the behavior of these and which one could give better results for a particular problem addressed.

The results achieved from our experiments are promising and from our point of view, it could be used as an assistant in making decision process.

We consider useful this thing because the results of the analyzes are based on data set with many patients and capacity of technical tools used is higher when exits enough information and also subjectivity disappears.

As future research directions, these areas allow the development of new systems to facilitate essential processes necessary for life and society. To have a better understanding of the importance of *IA* in our lives, we mention an article recently published by the Harvard Business newspaper, where one of the most influential scientists in the field of Artificial Intelligence, Andrew Ng, states that if there is momentary an activity carried out by a person in about a second of time, most likely, that activity will be automated by IA either now or in near future.

2 Structure of the paper

The paper is structured on three chapters, each chapter has a presentation of the defining elements and the proposed approaches, and, as well the results of the experiments.

The chapter **Theoretical Prerequisites** describes the most relevant historical approaches to machine learning in different areas with emphasis on medical and psychological fields. Also, is presented the interdisciplinary work from these fields until nowadays.

The chapter **Classification of numerical and multicategorical data. Applications in medicine** consists of three different parts:

- first part of the chapter addresses an unsupervised classification problem based on the most relevant *linkage* methods and how they yield results (grouping data in clusters) and also their evaluation using internal and external evaluation methods;
- second part intend to determine the most relevant attributes selected to be included in a model so the results determined by the application of the classification methods to be as accurate as possible;
- last part of this chapter aims to improve the classification methods by selecting the optimal values of the classification methods parameters.

Chapter **Classification of binary data. Applications in psychology** is divided into two parts, in terms of data sets analyzed:

- first, the communication style questionnaire (SCq). Answers to this questionnaire create a set of data with 220 instances (the number of patients) and 63 attributes (60 the answers to the questions of the questionnaire, and 3 other attributes quantify other information about the participants in the questionnaire: age of the person, sex of the person and self-perceived stress). For this dataset we analyzed the predicted probability of the determined class (target variable) after applying the classification methods.
- second, the cognitive-emotional coping questionnaire (tCERQ). Patient responses to this questionnaire led to the creation of a set with 240 instances and 36 binary attributes. For this data set we applied detection methods to discover outliers.

  In addition, we wanted to check if exists some correlations between our achieved results, therefore for both data sets analyzed in this chapter we also considered techniques for identifying the correlations existing in collected data sets.

Chapter **Opinion Mining with Text Analysis** addresses issues related to the *Opinion Mining* (OM) with applicability to a data set collected by the author consisting of students opinion related to their feedback for a seminar.

The purpose of this part was to determine whether students have a positive or negative opinion (feedback) related to seminar class. The data set contains

180 instances (representing the number of students who have expressed their opinion) and a single text attribute representing the answer to the question "How was the seminar throughout the semester, from your perspective?"

Finally, the chapter **Conclusions and Future Work Directions** outlines the ideas and results that led to this work, as well as proposing other working directions, either by improving the issues already presented or by proposing other innovative ideas and methods.

3 Original contributions

From the point of view of relevance of this work for scientific community, the results obtained are useful and have applicability in the addressed fields.

We summarize these results on chapters and we will highlight the utility and advantages of the experiments.

- Results achieved in the chapter **Classification of numerical and multicategorical data. Applications in medicine** are:

    1. An experimental analysis of linkage methods in agglomerative clustering in order to show which of the methods compute the optimal results. The evaluation of clustering methods in our analysis was done with internal and external validation measures. The practical utility of this analysis is the extension of its application (with linkage methods and evaluation metrics discovered by us as being optimal) for new data sets with the same attributes types.
    2. a particular approach based on meta-heuristic Genetic Algorithms to determine a relevant subset of attributes from the entire existing set. The proposed method based on Genetic Algorithms, was compared with other methods, the differences being evaluated applying the classification methods for the achieved subsets with each method and entire data set. The classification methods validation was performed by specific metrics in order to quantify performance of these methods. The utility of these results is this new approach to determine a relevant subset of attributes.
    3. a method for selecting the optimal values of the classification algorithm parameters based on Genetic Algorithms. To validate the efficiency of our method, we performed the experiments with two other methods in order to determine the values of the classification methods parameters. The results highlighted the utility of the proposed method by improving the accuracy metric value as the measure of the performance of the classification methods.

- The achieved results in chapter **Classification of binary data. Applications in psychology** are:

1. collecting and creating a mixed data set based on responses to the communication style questionnaire (220 instances with 63 mixed attributes);
2. determine an optimal classification method for our particular data set and compute the correlations between the participants communication styles and the level of self-perceived stress.
3. collecting and creating a data set based on responses to the adapted cognitive-emotional coping questionnaire (240 instances with 36 binary attributes);
4. selection of instances with abnormal behavior from data set and evaluation of the performance for classification methods after we deleted these data;
5. investigations the effectiveness of methods to detect outliers from a data set: Elliptic Envelope, Isolation Forest and Local Outlier Factor, and perform an analysis of these detected outliers.
6. analysis of correlations between existing coping strategies in the data set.

The relevance of these results comes from the perspective of the originality of collected data sets, the combination of applied methods and corroboration of these results with the theoretical aspects from psychology scientific literature.

- The achieved results in chapter **Opinion Mining with Text Analysis** are:

1. collecting a set of data with 180 instances and an attribute (text type) corresponding to the answers of the students participating in the Logical and Functional Programming Seminar. These answers summarize students views related to seminar;
2. classification of these opinions in an objective way *(Opinion Mining)* and the ability to make predictions for new instances;
3. the recommendation of a meta-classification method in which each classification method gets a degree of confidence based on previously achieved results.

   Similar approaches were not found in the literature, when this paper was wrote, therefore we consider our approach relevant from the point of view of the originality of the data set collected and analyzed and from the perspective of approaching a text from *Opinion Mining* against to *Sentiment Analysis.*
4. documents (instances) clustering based on *topic modeling.*

We think this approach was and is necessary, especially in the context of the constant use of the mechanisms offered by the technical field to other fields of research (or study), either scientifically or clinically. The two areas addressed in this PhD thesis are representative areas for each of us because physical and mental health is vital. In this context, any mechanisms to improve them are more than welcome.

4 CUVINTE CHEIE

- **classification**
- **mixed data sets**
- **tunning parameters**
- **genetic algorithms**
- **outliers detection**
- **opinion mining**

5 PHD Thesis Content

Cuprins