Computing expert's intelligence
Neocleous, Andreas

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2016

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 1
Introduction

Classification and data analysis techniques are nowadays applied in most of the fields of science. In medicine, a significant amount of work has been done for predicting or diagnosing diseases, using historical data from examinations of previous patients (Fang et al. 2016).

For applications and tasks such as medical diagnosis, weather forecast, forex prediction, pattern recognition in arts and culture, speech recognition, text identification and others, typically the information is digitally stored in databases and processed computationally. In computer science, these datasets are used to build systems that automatically perform actions and return results. In this thesis, I used two different types of datasets to apply machine learning and signal processing techniques for two tasks. In the first case which involves the early detection of chromosomal abnormalities, the data that I worked on contain medical observations from pre-natal examinations, including both fetal and maternal characteristics. In the second case, signal processing techniques and filter based methods for the identification of repeating patterns and ornamentations in ethnomusicological data are proposed. These are presented in the second part of this thesis.

In medicine, data storage and health record of patients has been a common practice for medical doctors since the age of Hippocrates. Medical information related to the evolution and the causes of a disease was stored. The modern health record is structured in a more detailed way, including information from examinations such as blood tests, heart rate, allergies, biophysical measurements, medicine taken and many more. The advance of digital technology gave from one hand the opportunity to develop specialized equipment for medical examinations such as ultrasound screenings, MRI, automated systems for medicine control of patients, tools for surgeries etc, and on the other hand gave the ability of storing digital information from many private doctors, networks of hospitals from several countries operating electronic health record systems.

The benefits of using an electronic health record system that ideally will be globally interoperable are many, both on a personal level, as well as on a social level. Some of the benefits from the patients’ point of view are the opportunities for a better health care, the safety of the personal data, the freedom of choosing more than one doctor or hospital for a diagnosis and treatment among others. On a social level,
the hospitals have digital libraries for the records of patients and there is substantially reduced need for manpower, papers and storage rooms. There is also better communication between the doctors of different hospitals. From the computer science and machine learning points of view, we take advantage of the multitude of data that are stored in such electronic health records in order to create systems that identify and predict several diseases and thus for better health management.

What is important in machine learning, is to analyze with big data and extract the knowledge and the usefulness through this analysis. In the last years, there has been an attempt to analyze unstructured data and build models that are able to identify patterns of interest. The Watson platform from IBM (Wolf et al. 2016) is an example of such systems. In medicine, such unstructured data can be for example the documents of the doctor.

The dataset I used for the identification of chromosomal abnormalities consists of normal cases, called euploid, and five abnormal cases: trisomies 21 (T21), 18 (T18) and 13 (T13), Turner syndrome and triploidy. More than 99% of the cases are euploid and thus it is a highly imbalanced set. It is common in medical datasets that the populations of the normal and the abnormal cases have an imbalanced relation due to the fact that from the entire population of cases that are examined, only a small percentage will be positive to a disease. Examples of such datasets among others, are described in (Klement et al. 2012) for computed tomography imaging of children, in (Boughorbel et al. 2016) for breast cancer prognosis, in (Sheikhi and Altınçay 2016) for type II diabetes detection and in (Dinov et al. 2016), for Parkinson’s disease prediction.

The fact that in medical datasets typically the positive cases consist the minority of the total population is encouraging for human nature and health, however it is many times an obstacle for classification methods in terms of accurate generalizations. Every popular classifier such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), k-Nearest Neighbours (k-NN) and others have some drawbacks in their methodology regarding the class imbalanced problem. In the literature, a lot of work has been done for creating balanced from imbalanced datasets for classification (Mazurowski et al. 2008, Wilk et al. 2016, Hoens et al. 2013, Stefanowski and Wilk 2008). This is typically done either by oversampling the minority class (Perez-Ortiz et al. 2015) or downsampling the majority class (Kubat et al. 1997).

In both approaches, it is important to make sure that the information that is stored in the data that are used for training is not lost when downsampling the majority class or distorted when oversampling the minority class. One significant contribution among others in this thesis, is the application of a method for downsampling the majority class using the k-means algorithm. It is based on the assumption that the euploid class can be represented with a number of subclasses. This assumption rises from the large amount of cases that consist the euploid class and thus the normality can be consisted with more than one pattern and it is difficult to be de-
fined. It is also possible that subclasses exist in the minority class of the aneuploidy cases, but it is difficult to safely drive such conclusion due to the limited number of cases. Nevertheless, if the above assumption is valid, then the supervised methods are forced to generalize a multi class feature space such as the euploid, into a single class. In Chapter 3 it is shown that ANNs perform better when the populations in both classes that are used for training are consisted with equal or similar number of cases.

The findings in this thesis contributes also to the determination of the relative significance that each feature has to the prediction or the diagnostic findings. This has been facilitated through heuristic and through formal mathematical approaches. In Chapter 3 I present the results of the ANNs that are built using different combinations of input markers. In (Neocleous et al. 2016b), we applied the Generalized Matrix Learning Vector Quantization (Biehl et al. 2006b) (GMLVQ) method to measure the relevances of the input markers. The results of the GMLVQ are in line with the ones presented in Chapters 2 and 3.

Concerning the data in ethnomusicology, these are consisted of one for monophonic or two for stereophonic dimensional feature spaces. These are typically the sound pressure over time for the two channels left and right, that are converted into electronic signal and stored in digital/audio format. Identifying specific features that model and distinguish the folk music of non-Western countries requires several adaptations of the existing algorithms in order to meet the particularities of each country’s musical culture. Typically, for problems such as instrument identification, segmentation of important musical parts in a song and music similarity, a set of low features, called descriptors are extracted directly from the audio signal (Neocleous et al. 2014b). Then, by applying dynamic programming, higher level features such as the scale of a song can be extracted and used for classification and similarity measures (Gómez 2006). The contribution of this thesis in the field of the Computational Ethnomusicology is presented in Chapters 5 and 6. An adaptation of the COSFIRE filters (Azzopardi and Petkov 2013, Azzopardi and Petkov 2014) to 1D signals is proposed for tasks such as repeating pattern identification that is found in melodies and for ornamentation detection. Further work was done for symbolic representation of musical signals, tonal similarity, segmentation into important musical parts of songs and published in several conferences and workshops (Neocleous et al. 2012, Neocleous et al. 2014a, Neocleous et al. 2014b). The COSFIRE filters outperform state-of-the-art methods in digital signal processing such as dynamic time warping (DTW), as shown in Chapter 5.
1. Introduction

1.1 Scope

In part I of this thesis, the problem of the early diagnosis of fetal chromosomal abnormalities during pregnancy is addressed. The goal of this study is to explore the potential use of machine learning techniques, such as ANNs, SVMs and k-NNs, in order to generate classification models for predicting fetal chromosomal abnormalities in the first trimester of pregnancy. The aim of the work done as presented in Chapters 2, 3 and 4 is to achieve 100% diagnostic rate of the T21 at the lowest false positive rate (FPR) possible.

Other equally important objective of this study include the optimization of a cut-off value (threshold) that is applied to the output of the ANN of an unknown case in order to binarize that value into a class between normal or abnormal. In most of the classification methods, there is a need of using a cut-off value that separates a similarity measure or a system output into a binary decision and that is in the range between 0 and 1. In the identification of chromosomal abnormalities, we represent the normal cases (euploid) with the number 0 and the abnormal cases with the number 1. A cut-off value that is closer to 0, returns high detection rates, at a cost of high FPR.

The question here is how important is to identify the entire population of the abnormal cases with respect to the number of the false positive classifications. These are cases that will have to perform additional unnecessary examinations, including invasive tests that may put a mother in risks for complications and the fetus in risks for miscarriage.

In the optimization stage of the cut-off value, we consider several social and financial aspects, such as the cost of a false positive classification with respect to the cost of a false negative classification. There are several methods to optimize the cut-off value as it has been discussed in (Neocleous et al. 2016). However, our criterion is to use a cut-off value that will return a 100% detection rate (DR) for T21.

In addition to the objectives stated above, several other research questions of equal importance are addressed that are related to practical aspects of the pre-natal examination procedure. We attempted to minimize the number of parameters that are needed as input to the system and at the same time to meet the above mentioned objective (100% DR for T21). In Chapter 3 we separate the available medical markers into two groups that come from two different examinations. From the findings of this study, we raised another question: is it possible to propose a two-stage screening for reducing the number of unnecessary examinations? From the findings in Chapter 4 it is shown that a two-stage approach can reduce dramatically the number of invasive tests that are currently performed. The research questions are summarized as follows:
1.1. Scope

- How can we build an effective system that will require the least possible prenatal examinations?

- Which machine learning techniques are appropriate for the modelling of medical data?

- How to deal with highly imbalanced datasets?

- How do ANNs contribute to the problem of the detection of chromosomal abnormalities?

- How to determine the best cut-off value that guarantees a 100% DR of T21?

In the second part of this thesis, I propose a novel algorithm for the identification of important patterns in 1D signals. This method is an adaptation of the COSFIRE filters that are found effective in 2D and 3D signals such as in image processing. The research questions in this part of the thesis are the following:

- Are the proposed COSFIRE filters effective and appropriate for applications using 1D signals?

- How does their performance compare with other state-of-the-art methods?

- What differentiates the COSFIRE filters from other existing methods?

- How can the COSFIRE filters be used in other 1D signals than the ethnomusicological data?

In Chapters 5 and 6 I first describe the proposed method and then I apply such filters to ethnomusicological data for the identification of repeating patterns and ornamentations. Particularly, in Chapter 5 I have applied the COSFIRE filters to a dataset of 38 songs that I created and published online. From the results, it is shown that COSFIRE filters outperform state-of-the-art methods. The COSFIRE filters are found to be effective in the identification of patterns in signals that have temporal variations, such as the singing voice. Similar to this idea is the well known method DTW (Müller 2007). However, what differentiate COSFIRE filters from DTW is the fact that the COSFIRE filters allow temporal tolerance that can be optimized with parameters. Additionally, COSFIRE filters are more computationally efficient than other methods. The COSFIRE filters proposed in this thesis can be applied in any type of 1D signal. However, additional experiments need to be done to test their performance on other tasks, such as the identification of k-complexes in EEG signals.
1. Introduction

1.2 Outline

The rest of the thesis is based on journal and conference papers that are either published or accepted for publication or submitted to academic journals of high impact factors. Some of the content in this thesis might be repeated in order to make the chapters self contained.

In Chapter 2, the problem of the identification of chromosomal abnormalities is introduced. In that Chapter, a two-stage approach for the detection of the T21 (stage 1) and the detection of the other chromosomal abnormalities (stage 2) is proposed. A statistical analysis of the dataset is also presented and discussed with respect to the discriminative power of every independent variable.

In Chapter 3, a more technical approach to the problem is introduced. In that Chapter, an attempt is made to address questions such as whether to use normalized or raw data for training the classifiers, and on whether a balanced training set improves the performance of the models considering that the dataset is highly imbalanced, like in many medical datasets. Finally, experiments have been done in a heuristic way to explore the contribution of the parameters to the system.

In Chapter 4, a two-stage procedure is proposed for the detection of the chromosomal abnormalities. This is an extension of the work that is presented in Chapter 5. It is based on the premise that pregnant women perform the standard screening in the first trimester of pregnancy which contains an ultrasound examination and a blood test. In the first stage of our trials, when using 4 parameters (maternal age, PAPP-A, \( \beta \)-hCG and the nuchal translucency), we achieve 100% DR of T21 at a relatively high FPR of about 20%. In the second stage, all the cases that are ranked as positive in stage 1, are further processed with another examination which includes a specific ultrasonograhic scan to get information about the flow in two arteries of the fetal cardio system. In stage 2, we achieve 100% DR of T21 and more than 80% of the other chromosomal abnormalities, at a FPR of less than 3%.

In Chapter 5, we apply the COSFIRE filters for the identification of repeating patterns in ethnomusicological data. We have adapted the COSFIRE filters that are found effective in 2D and 3D signals into 1D signals. It is shown in that Chapter that the COSFIRE filters are effective for such problems and they can be applied in signals of different scales.

In Chapter 6, I introduce another application for ornamentation detection in singing folk music that is approached with the use of COSFIRE filters. It is shown that the COSFIRE filters are able to capture such particular variations in 1D signals and to identify similar patterns.

Conclusions and future work of this thesis are reported in Chapter 7.
Part I

Computational Intelligence in Medicine