Retrospective Publication Analysis - Validation and Improvements of the results

Exposé
for Bachelor Thesis

Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät II
Institut für Informatik

submitted by ... Dennis Wagner
born on 15.02.1994
in .....Berlin

supervised by: Prof. Dr.-Ing. Ulf Leser
Prof. Dr. Marius Kloft

submitted on: .....

1 Introduction

Developing new drugs is a very time and money consuming process. It can take around 10 to 15 years until a new drug gets the chance to be approved by the FDA (Food and Drug Administration). For a large number of therapeutica, the first step in this process is the identification of a drugable target, where a potential new drug can take effect. Much research is necessary to understand a disease on a sufficient level of detail to identify molecules like proteins, genes, etc involved and among them identify a drugable target. Once a suitable target has been found and validated, research can begin on finding a new drug, that, if accumulated in the malfunctioning cells, can interact with the target and change the course of the disease. The new drug has to be subjected to different stages of testing. First tests are employed to give an early estimation on its safety. Before tests on actual human subjects can be approved, tests on living cell cultures and animals in the next stage, the preclinical trials, determine if the drug is safe to test on humans. Good results in this stage make way for the clinical trials. Initially tests are done with healthy volunteers and later in larger groups of patients. After success in the clinical trials the results are submitted to the FDA, where they are reviewed and approval is decided.

Especially the clinical trials are very money and time consuming. Therefor predictions on future success of a new drug before entering clinical trials is desired. In collaboration with a pharma company, the chair of Knowledge Management in Bioinformatics has realized a project to make predictions for the success of a given drug in the clinical trials, by utilizing machine learning techniques on features extracted from medical publications on a given drug-disease pair. A variety of different features are extracted from papers published on pubMed, for example the count of papers mentioning related content or features extracted from qualifiers of mesh terms. On these features different classifiers have been tested and the results have been compared.

2 Goal

The aim of this work will be to validate and improve the results of the previous work. This will be accomplished by first reproducing the results, checking the scripts for errors and understanding the existing work. Then new features and classifiers can be tested and compared in order to improve the results.

3 Approach

Different features are used for classification, but they are not always able to represent the unique properties of underlying objects sufficiently enough to guarantee best possible classification results.

3.1 Time Series Representation

In the previous project different features like the number of articles published on the topic per year or the commitment of an author per year have been considered. A property of this application, that has not been exploited yet, is the fact that features are extracted over a long period of time and so far the correlation of values of features over time have not been considered for classification.

The data can be represented in a multivariate time series (MTS) $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ where $\boldsymbol{x}_i \in \mathbb{R}^m$ for all $i \in [n]$. If $m = 1$ $\boldsymbol{X}$ is called univariate time series (UTS). Currently 114 features (m) are observed per year over a period of 20 years (n). This representation allows to consider the changes of the features over time for classification. Let $\underline{\boldsymbol{X}} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_k)^T$ be the inputs and $\underline{\boldsymbol{Y}} = (y_1, ..., y_k)^T \in \{-1, 1\}^k$

the corresponding labels. In the previous project $k = 167$ manually annotated pairs have been considered.

3.2 Feature reduction

The MTS items have a high number of features in comparison to time samples so it is desired to reduce the number of variables. A too large number of dimensions can corrupt the results of the classification, while a well chosen subset of features can improve the results. Feature subset selection (FFS) is a technique for preprocessing data that is concerned with choosing a good subset of variables [3].

There are many ways to approach FFS. The theoretically best way is Exhaustive Search Selection (ESS), where every possible subset of features is tested. With 114 features this approach will take to much time to actually test with all classifiers. Fisher Criterion (FC) and Recursive Feature Elimination (RFE) can be done on a Autoregressive Model (AR) representation of the MTS items [5]. But these methods compute the rating score only per variable, ignoring possible correlation between variables. A widely used method for FSS is Principal Component Analysis (PCA). In PCA the Singular Value Decomposition (SVD) of the covariance or correlation matrix is considered for ranking the features. A PCA-based FSS method has been adapted for use with MTS data sets by [4]. First the Principle Components (PC) are computed for every MTS item using PCA, then the common PCs among all MTS items are obtained by bisecting the corresponding PCs of the MTS items. At last the variables have to be selected. Simple ranking risks the selection of redundant features, therefor the features are clustered, before the least redundant features from each cluster are chosen.

3.3 Time series classification

As a baseline classification will first be realized with all elements of the MTS in a single feature vector without considering the time correlation.

With the baseline in place the next step is to employ more advanced machine learning techniques on the MTS in order to improve the results. A Support Vector Machine (SVM) is a simple but flexible tool for classification. An additional advantage is the possibility to design kernels that can affect the resulting classifier, but

the kernels have to be adjusted for use with MTS data.

The simplest kernel is the linear kernel. This kernel is easy to understand and easy to apply on MTS data of the same length [2]. Another kernel that yields good results in practice is the RBF kernel. This kernel can also be applied directly on the MTS data. For this application the Euclidean distance should suffice, because the MTS are of the same length. Still the Euclidean distance is very sensitive to outliers, therefor later test can be done with different distance measures like discrete time warping.

3.3.1 Other kernel functions

The elements of some UTS in the MTS can take on any real value. So kernel functions based on subsequence analysis, expecting discrete values, can not directly be applied to the MTS data. Also the assumption of an underlying Markov Process does not apply to all UTS. So Pair Hidden Markov Models kernels can not be used directly on the MTS data either [1].

But during the work on this project based on the results of tests with different kernel functions other kernel functions may be explored as well. And when working with MKL some kernels that are not suited for use with whole MTS might be used with single UTS.

3.3.2 Multiple kernels

Kernel functions are closed under addition and multiplication, meaning the sum or product of multiple kernel functions is a kernel function itself. Therefor different kernel functions for UTS can be combined to create a new MTS kernel function. This gives a model where no interaction between single time series is assumed. Tests will show if this is the better approach for this application.

3.3.3 Feature based classification

The Problem of classifying a new time series can be approached a different way, where the time series is transformed into a feature vector. Therefor different features have to be extracted from the time series and then classification can be approached with different feature based methods.

Since in this application local as well as global trends are of importance for discriminating objects, Discrete Haar Transform (DHT) coefficients can be used as features. Additionally time domain features can be used as well. The usefulness of these features can be evaluated with non-parametric tests.

3.3.4 Evaluation and improvement of classifiers

The number of available data is relatively small, therefor the data will not be divided in training set and test set for evaluation. Classifiers will be evaluated based on k-fold cross validation. Based on the results of the previously introduced classification methods other other approaches can be explored to improve the results further.

3.4 Implementation

Because of lacking experience with R this project will be implemented in python. For the comparison of the results R is not necessary and classification will be done on MTS, so either way corresponding kernel methods have to be implemented. The existing methods will be used for reading the data and the results are exported in csv format. This can then be read in python and transformed in the previously introduced representation of MTS. With this representation in place the kernel methods can be implemented as described above.

4 Related work

Classification of sequential data is a problem found in many applications. Time series are ordered sequences and thus are a specialization of this problem. When analyzing DNA or protein sequences classification allows to order the data in different categories depending on structure or function. This can give valuable insight in their interactions and functions. Sequence classification has also medical applications. In Electrocardiography (ECG) the electrical activity of a heartbeat is measured with electrodes over a period of time. Analog brain activity can be measured in

Electroencephalography (EEG). Classification is used to determine if observations of a new patient are normal. Other applications are speech recognition, anomaly detection or automated text classification.

4.1 Sequence classification methods

For sequence classification there are three general approaches [6] [7] [8].

- Feature based classification
- Distance based classification
- Model based classification

The feature based classification approach uses feature selection to transform the MTS data into feature vectors. On these feature vectors classification can be handled with classic machine learning methods. This leaves the problem of selecting good features, which in practice is often the hardest part.

The distance based classification approach utilizes classic machine learning methods that make use of distances, like k-NN or SVMs. By defining a distance measure for sequences these methods can be applied directly on the sequential data. Widely used distance measures for sequential data are the Euclidean distance on sequences of same length and dynamic time warping on sequences of different length. For SVMs this can be done by defining a new kernel function [2]. Different kernel functions and their utility for this project have been discussed in the previous chapter. K-NN is very simple and can give good results in practice, but other methods yield generally better results.

The model based classification approach is based on the assumption, that the sequences are generated from a process. Examples for model based classification methods are Naive Bayes and Markov Model.

Bibliography

[1] S. Rüping, *"SVM kernels for time series analysis"*, 2001

[2] H. Shimodaira, K. Noma, M. Nakai and S. Sagayama, *"Dynamic Time-Alignment Kernel in Support Vector Machine"*, Advances in neural information processing systems, 2002

[3] I. Guyon and A. Elisseeff, *"An Introduction to Variable and Feature Selection"*, Journal of Machine Learning Research 3 1157-1182, 2003

[4] H. Yoon, K. Yang, and C. Shahabi, *"Feature Subset Selection and Feature Ranking for Multivariate Time Series"*, IEEE transactions on knowledge and data engineering vol. 17 no. 9, 2005

[5] T. N. Lal, M. Schröder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer and B. Scholkopf, *"Support Vector Channel Selection in BCI"*, IEEE Trans. Biomed. Eng. vol. 51 no. 6, 2004

[6] Z. Xing, J. Pei, E Keogh, *"A brief survey on sequence classification"*, ACM SIGKDD Explorations Newsletter, 2010

[7] T. Amr, *"Survey on Time-Series Data Classification"*, 2012

[8] M. Deshpande and G. Karypis, *"Evaluation of techniques for classifying biological sequences"*, In PAKDD '02: Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, 417-431, 2002