

CLINICAL DATA MINING

Basic Information

This is a course, which contributes to MSc award in Biology

Course period	From September 1st till February 1st, 1 semester
Study credits	3 ECTS credits
Duration	108 hours
Language of instruction	English
Academic requirements	<ul style="list-style-type: none">- BSc degree in Biology, Chemistry, Physics or Environmental Sciences or equivalent (transcript of records),- Good command of English (certificate or other official document)

Course Description

«Clinical Data Mining» is an extensive course, which is designed to advance a student's ability to use the known repositories with medical data, collect the data from the clinical devices and the physical examination.

The students are expected to be able to download files of various formats to the working machine, rearrange and integrate the datasets into one massive of data for the further analysis.

The most important part of this course is the mined data reprocessing, which is dedicated to the primary analysis of the mined data. It includes the machine-learning techniques, which is thought to be useful to find the hidden structures in the mined data. The Representation of the Data part of the course includes the basic techniques to represent the data structure in the clear and condensed manner, which will significantly impact the efficiency of further analysis.

The course is designed to cover a wide range of possible data structures and formats of medicine.

Course Aims

- To help students study new discipline or to ameliorate the existing knowledge in data mining.
- To assist students to manipulate the data to get rid of the missing data and to make it informative.
- To give students the understanding of how to integrate the data from various sources.
- To introduce the basic principles of rough sets processing.
- To give the introduction in machine learning and classification methods and its application to medical data mining.

Course Objectives

The course has been designed to:

- ensure that students are familiar with a basic data mining routines.
- ensure that students know the main repositories of the medical data and medical data formats.
- give students the optimal methods for the data integration.
- give students an appreciation of the missing data processing.
- provide students with the basic routines of mined data representation.
- provide students with the basic routines of medical image processing.
- make students know and understand the principles of the data mining in the particular cases of the interest.

Learning Outcomes of the Course

By the end of the course, the students will be able to:

- collect the various medical data from any source in clinic or from the repositories,
- perform the missing data processing in the given dataset,
- detect the hidden structure of the given dataset using the machine learning approach,
- clearly represent the structure of the mine data and build the massive, convenient for the further analysis.

Course (module) Structure

Learning Activities	Hours
Lectures	16
Practice sessions / Seminars,	16
Self-study Assignments	40
Final Exam (including preparation)	36
Total study hours	108

Course Outline

Week	Lectures	Practice session / Assignments	Hours
1-5	Introduction to Clinical Data Mining	<ul style="list-style-type: none"> Types of Variable Operating the Large Volume Datasets Loading and Exploring Data Home assignment No 1 	(5/5/10)
6-10	Clinical Data Mining Process	<ul style="list-style-type: none"> Repositories of the Clinical Datasets Dealing with Missing Data Rough Sets Frequent Patterns Home assignment No 2 	(5/5/14)
11-16	Mined Data Reprocessing	<ul style="list-style-type: none"> Representation of Data Mining Results Documenting Data Mining Data Integration Machine Learning in Data Mining Home assignment No 3 	(6/6/16)
	Final Exam		36

Course Instructors and Tutors, Contact Information

Instructor	Contact Information
Andrey SHUVAEV Ph.D. in Biophysics, Associate professor, School of Biology and Biotechnology Siberian Federal University, Krasnoyarsk	Svobodny, 79 Room 32-13 Tel +7 391 206 2165 Email: AShuvaev@sfu-kras.ru, andrey.n.shuvaev@gmail.com

Assessment

The final exam is an individual problem in the form of a simulated experimental dataset. The student must demonstrate the abilities in:

- finding and downloading the appropriate database for the given problem (20 points maximum),
- getting rid of the missing values, dealing with the rough sets of the given data, integrating the data from the different sources in one table (40 points maximum),
- preprocessing the mined data: machine learning classification with the following clear representation of the data structure using the chosen programming language (40 point maximum).

Grade policy:

A (excellent work)	91-100 %
B (above average work)	81-90 %
C (average work)	71-80 %
D (below average work)	50-70 %
F (failed work)	<50 %

Core reading

1. Bramer, M. (2013) *Principles of Data Mining*. Springer, London.
2. Olson, D. and Delen, D. (2008) *Advanced Data Mining Techniques*. Springer, Berlin, Heidelberg.
3. Sullivan, R. (2012) *Introduction to Data Mining for the Life Sciences*. Humana Press.
4. Mamitsuka, H. (2018) *Data Mining for Systems Biology*. Humana Press, New York, NY.
5. Porcu, V. (2018) *Python for Data Mining Quick Syntax Reference*. Apress, Berkeley, CA.
6. Williams, G. (2011) *Data Mining with Rattle and R*. Springer, New York, NY.