Bachelor Thesis Review

Faculty of Mathematics and Physics, Charles University in Prague

Thesis author	Aakash Ravi
Thesis title	Machine learning-based identification of separating features in molecular fragments
Year submitted	2017
Study program	Computer Science
Study branch	General Computer Science
Review author	David Hoksza — Advisor
Department	Department of software engineering

Overall

	good	OK	poor	insufficient
Assignment difficulty	X			
Assignment fulfilled		X		
Total size *... text and code, overall workload*	X			

The thesis required to become acquainted with the problem of virtual screening and with possibilities of application of machine learning for identification of bioactive molecules properties. After analyzing possible approaches, the student chose subspace clustering as the method to use for this task. The idea was to identify subspaces in a space of molecular fragments properties which would comprise primarily of properties present in active molecules. The subspace clustering proved later as not the best possible choice. The reason was the sensitivity of the approach in terms of the sensitivity of parameters which significantly impacts the ability of the method to find clusters in high dimensional spaces. This was further complicated by the fact that in this concrete application, subspace clustering needed to be used in a nontrivial way when points in the high dimensional space are of different types based on which molecule they come from. Necessity of such distinction led to a design which further increased the number of parameters of the system. In the experimental part, the student analyzed capabilities of subspace clustering to reveal subspace clusters in an optimal case where the characteristics of generated data matched the expected structure in real data. This experiment revealed that subspace clustering is not able to find clusters all the time but also indicated values of parameters which should be used with real data. Finally, the system was used with the learned parameters against real data. This last phase of the project should have been given more time. With well chosen evaluation method on distributed architecture, one could get a wider range or results which could be used for deeper analysis of the, rather bellow the average, results of the method. Although the evaluation on a distributed infrastructure (Metacentrum) was implemented, there was not enough time to obtain sufficient amount of results for proper analysis of the results on real data. On the other hand, the choice of using subspace clustering lead to a very complex task difficulty of which was close or comparable to a diploma thesis. Within the scope of the thesis the student: (i) developed a framework which shows how of subspace clustering can be utilized for the problem of identification of fragments' properties of bioactive molecules, (ii) identified weak spots of the approach for given types of data and (iii) evaluated the problem on both simulated and real data.

Thesis Text good OK poor insufficient

Form *... language, typography, references*		X		
Structure *... context, goals, analysis, design, evaluation, level of detail*		X		
Problem analysis		X		
Developer documentation		X		
User documentation		X		
The thesis is well structured and includes both verbal description of the problems and solutions, and formal definitions and pseudocodes of the proposed algorithms.				

Thesis Code

	good	OK	poor	insufficient
Design *... architecture, algorithms, data structures, used technologies*		X		
Implementation *... naming conventions, formatting, comments, testing*		X		
Stability		X		

Overall grade Velmi dobře

Award level thesis Ne

20. 1. 2017 Signature