Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Diellor Hoxhaj		
Název práce	Techniques Applicable to the Analysis of Educational Data		
Rok odevzdání	2024		
Studijní program	Informatika	**Studijní obor**	Umělá inteligence

Autor posudku	Jan Hric	**Role**	Oponent
Pracoviště	Dept. of Theoretical Computer Science and Mathematical Logic		

Text posudku:

Study programme: Computer science
Study branch: Artificial Intelligence
Reviewer's review of the diploma thesis

The presented work is about use of data mining techniques to four datasets using three approaches. It is presented in English.
The first chapter describes a theory used for futher analysis. The second chapter describes datasets and their analysis. The first dataset is about graduation of students on US universities. It is used for clustering, and for analysis of differences among ethnic groups. The second and third dataset is about connections in UK and Kosovo parliament, respectively. They are analysed as social networks using graph techniques. The fourth dataset is public OULAD dataset from Open University, UK. A main goal is to create a predictive model of a student success rate based on their behaviour in a virtual learning environment.

The amount of work is big and used techniques are heterogenous. First three datasets were created by the author and especially the first dataset is big (3975 data points) and scrapped using python scripts which are described in the thesis as well. So in principle described methods can be used for scrapping data about other study programs. Although it is mentioned in the thesis that the dataset can be used for futher analyses (and so it is important result of the thesis), I did not find any description or a link.
The fourth dataset was public and the author used four method for constructing predictive models. As the dataset is public it was used by other authors and a quality of models is comparable. The best method was random trees. The dataset needs an intensive cleaning and preprocessing so these methods were used as well.
A theory described in first chapter is mostly usable and relevant. Some explanations on the beginning of the thesis in ch. 1.1-1.3 seem to me as to oscillate between a too high level description and technical details. English language is mostly good, but at some places needs a final proofreading (h,i,j,n). Mathematical language improves and some particular problems are mentioned later (a,e,g). The bibliography has 49 items but about 1/3 are web links concerning data sources.

This thesis shows that the author is able to do a technical analysis of various datasets using different methods and tools. A domain analysis and an interpretation (in a context) usually need an interaction with domain expert.

In the first dataset, found clusters were described mostly technically (using median and range of values) (k). Only 3 of 6 clusters have somehow clear meaning, other seems fuzzy and arbitrary. Analysis of changes during three years is mostly uninteresting as we have data for short interval. Including house prices adds some information but is still too coarse. The second and third dataset about parliaments are to coarse as only one type of interaction is used. A graph constructed from the third dataset is unconnected, so this is not typical example of social network. Moreover, data are incomplete and during analysis a part of data is thrown away, so result can be biased (t). Some methods are not directly suitable for unconnected graphs.
The last dataset is publicly available so result were compared with other work. Suggested questions for analysis are relevant. A necessity of massive preprocessing was a bit surprising but the author was able to clean the data. A domain interpretation is uneasy (q,u).

Comments on particular issues:
a) p.11, formula 1.7: from which values is computed sigma? All d_ij or d_ij for fixed winning i?
b) ch 1.3, p.11, 2nd para: explanation of an interconnection graph of links (edges?) is in the context of social networks unclear.
c) I expect an explanation if we have in an analysed social network only one type of edges or multiple types (for multiple types of interactions)
d) formula 1.17 and thus the centrality measure is applicable only to connected graphs
e) A constraint before formula 1.20 is unclear. It should be probably the same as in 1.20
f) cit. [20] at p.19 is unusual: definition of common decision trees refers to Handbook of AI techniques in some specific area.
g) formula 1.22 uses C and p_i, but the description is about data points x_i and their classes y_i
h) p.26, comparisons, text: probably atribute *two* has a max. gain
i) p.35 typos: inlcude
j) p.56, from cluster 3: a description of prices should be uniform (all in dollars or in thousand of dollars), probably *k* (for kilo-/thousands) is not used consistently and the dot and comma characters are significant in english numbers.
k) p.52+ : medians for a cluster description give only too coarse information
l) ch.2.2.3: which methods are usable for unconnected graphs?
m) for parliament analysis, common university or domain of study are too coarse criteria
n) p.69, item 3: unclear text
o) p.36, para 2: incomplete sentences. If beggining of a paragraph is something like paragraph headline, use other (font) style.
p) transformations from 2.3.4: A reason of these transformations is interesting for a datamining community (but not for final users). E.g. transformations can be necessary due some technical reasons of used methods and/or libraries, recommended after preliminary experiments or simply are usual approach (or other reason, e.g. interpretability of results).
q) Algorithms found that atributes ‘FFF course’ (in 2.3.5) and ‘STEM domain’ (in 2.3.6) are important features. But a domain interpretation of these findings is not clear: are there differences in courses (e.g. number of quizzes or an overall difficulty) or in a student behaviour?
r) ch.2.4.1: it seems that clusters were renamed after first paragraph
s) ch.2.5.1: there are possibly other public features like ‘how long is a member in a parliament’.
t) ch.2.5: not all members are included in analysis and consequently in clusters, so domain interpretation can be biased

u) last sentence: A domain interpretation of higher diversity can be subtle. It can be result of some unknown or hidden factors and supporting diversity per se can be irrelevant and ineffective.

I recommend the thesis for defence (to pass).

Práci doporučuji **k obhajobě.**

Práci nenavrhuji **na zvláštní ocenění.**

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 5.9.2024 **Podpis**