



I. IDENTIFIKACE

Název práce:	Shlukovací algoritmy pro nečíselná data
Jméno autora:	Jan Jakeš
Typ práce:	diplomová
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	Katedra počítačové grafiky a interakce
Oponent práce:	Zdeněk Buk
Pracoviště oponenta práce:	Katedra počítačů

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání	průměrně náročné
Zadání je průměrně náročné, převážně implementačního charakteru.	
Splnění zadání	splněno s menšími výhradami
Název práce i zadání explicitně zmiňuje "nečíselná data", ale v práci je tomuto kritériu věnován velmi malý prostor (prakticky pouze sekce 2.3). Předpokládám, že je to dáno tím, že v případě vhodné zvolené metriky pro nominální data se algoritmy pro shlukování číselných a nečíselných dat neliší. Uvítal bych ale, kdyby takový závěr v práci byl, případně, aby práce obahovala více metrik pro nominální data.	
Zvolený postup řešení	C - dobře
Student postupuje od analýzy a popisu jednotlivých metod přes implementaci po testování. Největší výhrady mám k fázi testování - student porovnává několik implementací - vlastní implementaci, Matlab a SYSTAT. Není však jasné, zda se shlukování provádí pouze jednou a následně se testují pouze implementované operátory (Dunn Index, Silhouette, Rand Index a Density), nebo zda je prováděno i opakované shlukování v jednotlivých nástrojích. Ocenil bych také lepší výkonové porovnání jednotlivých implementací a otestování škálovatelnosti zvolených operátorů	
Odborná úroveň	A - výborně
Student musel detailně pochopit několik metod pro vyhodnocování výsledků shlukování, které následně implementoval. V práci student čerpal informace především z anglicky psaných zdrojů. Vytvoření vlastních modulů pro Rapidminer vyžadovalo nastudování funkcionality již existujícího netriviálního softwarového díla. Odborná úroveň je odpovídající diplomové práci.	
Formální a jazyková úroveň, rozsah práce	B - velmi dobře
Rozsah textu považuji za odpovídající diplomové práci. Z hlediska jazykové úrovně mi vadí časté používání "počeštěných"anglických výrazů. Autor na toto sice sám v úvodu upozorňuje a - cituji - "recipientům této práce"se omlouvá, nicméně některé termíny bylo možné zvolit lépe.	



Výběr zdrojů, korektnost citací	A - výborně
Bez výhrad	

Další komentáře a hodnocení
Předložená diplomová práce je jednoznačně přínosem, student kromě toho, že nastudoval a přehledně popsal několik shlukovacích algoritmů, včetně metod hodnocení kvality shlukování, rozšířil také o tyto metody existující program Rapidminer. Rozšiřující moduly jsou k dispozici veřejně na serveru github.org. Jako největší problém práce vidím v testování - jak je uvedeno výše, není zcela jasný postup testování a chybí výkonostní srovnání různých implementací.

III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE

Předloženou závěrečnou práci hodnotím klasifikačním stupněm

B - velmi dobře

Považuji za vhodné, aby se v rámci obhajoby student vyjádřil k následujícímu:

- Srovnajte rychlost vaší implementace operátorů s jejich alternativami v programech Matlab a SYSTAT.
- Jaké jsou asymptotické složitosti vámi implementovaných operátorů? Uvádíte např. že výpočet siluety pro 20 000 instancí s 13-ti atributy trvá 25 minut. Je vůbec reálné aplikovat takový operátor pro řádově větší datové soubory?

Datum: 27. května 2015

Podpis: