Oponentský posudok diplomovej práce Jakuba Kratochvíla „Dimension Reduction Techniques in Morphometrics“

Pri spracovávaní mnohorozmerných dát sa často používajú postupy, ktoré redukujú počet dimenzií dát. Na takto transformovaných dátach s malým počtom dimenzií fungujú mnohé postupy dobývania znalostí ako napr. klastrovanie a klasifikácia jednoduchšie, spoľahlivejšie a ich výsledky sa ľahšie interpretujú. Cieľom práce bolo vybrať metódy redukcie dimenzií dát vhodné predovšetkým pre spracovanie morfometrických dát, porovnať ich a prípadne navrhnúť vlastné modifikácie známych metód.

Autor vybral z literatúry najpoužívanejšie techniky redukcie dimenzií dát – PCA (Principal Component Analysis), LLE (Locally Linear Embedding) a MDS (MultiDimensional Scaling). Tieto metódy kombinoval s vhodnými normalizáciami GPA (General Procrustes Analysis) a EDMA (Euclidean Distance Matrix Analysis) a niekoľkými metódami zhlukovej analýzy založených na princípe k-means clustering.

Diplomant v práci stručne popísal vyššie uvedené postupy až na 4 metódy zhlukovania, ktoré sú neskôr v práci požívané. Na str. 59 je uvedený iba ich holý zoznam s odkazmi na články. Hlavná časť práce porovnáva popísané metódy na konkrétnej úlohe, kde z trojrozmerných súradníc 13 vybraných bodov (tzv. landmarkov) na tvári osoby mal algoritmus určiť pohlavie osoby. Testovacie dáta obsahovali 24 záznamov o 17 osobách mužského pohlavia a 7 osobách ženského pohlavia. Autor používal postup skladajúci sa z normalizácie, redukcie dimenzií a konečného klastrovania do dvoch zhlukov. Cieľom bolo, aby každý z výsledných dvoch zhlukov obsahoval osoby rovnakého pohlavia.

Okrem vyššie uvedeného postupu autor skúšal taktiež iterované znižovanie dimenzie dát, kde namiesto redukcie počtu dimenzií na cieľovú hodnotu v jednom priechode, danú metódu aplikuje niekoľkokrát pričom počet dimenzií dát klesá geometrickou postupnosťou. Z výsledkov vykonaných experimentov autor odvodil, že iterované znižovanie počtu dimenzií je lepší postup, než jednopriechodová redukcia a tiež určil najvhodnejšie parametre použitých metód.

Všetky postupy, ktoré autor skúšal, implementoval v rámci programu Morphome3cs a priložený program a dáta umožňujú jeho experimenty jednoducho zopakovať.

Výsledky a odporučenia, ktoré autor odvodil však v posudku neuvádzam, pretože ich nepovažujem za príliš relevantné. Predovšetkým autor nezvážil vhodnosť úlohy, na ktorej metódy testoval. Prečo by výsledkom klastrovania mali byť práve skupiny zodpovedajúce pohlaviu? Čo ak sú „prirodzenejšie“ zhluky podľa vzdialenosti očí, alebo šírky nosa? Čo ak je vhodnejšie klastrovanie, kde jednému pohlaviu zodpovedá viacej než jeden zhluk?

Autor v práci vôbec nepopísal, ako vyhodnocuje kvalitu výsledného klastrovania vzhľadom k úlohe klasifikácie pohlavia. Usudzujem, že pre obidve možné priradenia rôznych pohlaví rôznym zhlukom spočítal pomer počtu správne klasifikovaných vzorov k celkovému počtu vzorov. Z dvoch hodnôt, ktoré dostal, napokon vybral tú lepšiu.

Autor nedostatočne analyzoval výsledky experimentov. Nekontroloval, ako vyzerajú výsledné zhluky. Triviálny algoritmus, ktorý označí všetky vzorky za mužov, má na dátach, ktoré autor použil, úspešnosť 71% a nulový rozptyl. Aj keď budeme vyžadovať, aby výsledkom boli dva neprázdne zhluky, tak pri náhodnom rozdelení na jeden zhluk s jediným vzorom a druhý s 23 vzormi s pravdepodobnosťou 7/24 = 0,29 bude úspešnosť klastrovania 18/24*100% = 75% a s pravdepodobnosťou 17/24 = 0,71 bude úspešnosť klastrovania 16/24*100% = 67%, čo v priemere dáva úspešnosť 69%. Táto hodnota sa blíži výsledkom dosiahnutým najlepšou metódou – viacpriechodovou LLE s GPA. Skutočne pre niektoré metódy ako napr. MDS najlepšie dosiahnuté výsledky väčšinou nastávali, keď menší zhluk mal 2 alebo 4 prvky. Takéto patologické javy môžu nastať preto, že množina vzoriek je silne nevyvážená.

Autor spúšťal experimenty pre jednu sadu parametrov 5-7 krát, ale nezamyslel sa, či je to dostatočný počet. Keď som dvakrát zopakoval vyššie uvedený viacprechodový algoritmus LLE s GPA, tak pri veľkosti okolia 8 som dostal priemerné úspešnosti líšiace sa až o 5%. To značne znehodnocuje numerické porovnávanie výsledkov experimentov. Na to, aby závery odvodené z experimentov boli podložené, by bolo nutné jednak pracovať s väčším počtom vzoriek a výsledky porovnávania rôznych metód robiť formou štatistického testovania hypotéz.

Text práce je napísaný angličtinou, ktorá vyžaduje drobné korekcie na mnohých miestach. Z terminologického hľadiska najviac vadí systematické používanie nesprávneho termínu pre rozptyl (autor používa „variability“, správne má byť „variance“). V niekoľkých vzorcoch sú zjavné preklepy (napr. druhý vzorec zdola na str. 16, vzorec zo str. 55).

Celkovo autor urobil rozumný výber metód a navrhol postupy (ako viacprechodovú redukciu dimenzií), ktoré by mohli viesť na zlepšenie výsledkov analýz mnohorozmerných dát používaných v morfometrii. Implementácia týchto metód v programe Morphome3cs umožňuje ich jednoduchú aplikáciu. Avšak triviálny algoritmus, ktorý zaradí všetky vzorky do jediného zhluku, dokáže riešiť autorom nevhodne zvolenú úlohu s úspešnosťou prevyšujúcou väčšinu výsledkov dosiahnutých v práci. Napriek nekvalitnej analýze výsledkov aplikácie týchto metód na reálne dáta, doporučujem, aby práca Jakuba Kratochvíla bola uznaná ako diplomová práca.

Praha, 27. 8. 2011,

RNDr. František Mráz, CSc.

KSVI MFF UK