

Zisťovanie príčiny požiaru pomocou 3D-PCA projekcie

Anotácia: Pri forenznom vyšetrowaní je často užitočné identifikovať ropný produkt, ktorý bol použitý pri založení požiaru. Ide o klasifikačný problém, ktorého cieľom je určiť, či bol objekt zapálený petrolejom, alebo naftou. Keďže obe látky majú veľmi podobné chemické zloženie, identifikácia pôvodu môže byť veľmi komplikovaná. Na reálnych dátach publikovaných v ¹ predstavujeme štatistickú 3D-PCA analýzu, ktorou možno klasifikovať jednotlivé vzorky do niektorej z dvoch tried.

Kľúčové slová: petrolej, nafta, metóda PCA, klasifikácia, zhuková analýza.

Úvod

Jednou z úloh forenzného vyšetrowania je aj analýza pôvodu požiaru. Keďže znalosť, či bol objekt zapálený petrolejom, alebo naftou, môže pomôcť pri vyšetrowaní, je dôležité určiť pôvodnú zápalnú látku čo najpresnejšie. V prípade, ak majú dve posudzované látky (ktoré mohli byť použité pri založení požiaru) odlišné fyzikálno-chemické vlastnosti, je možné pôvod určiť podrobnejšou analýzou ich štruktúry a zloženia. V prípade dvojice petrolej – nafta sú ale spomenuté fyzikálno-chemické vlastnosti látok veľmi podobné, čo výrazne sťažuje ich identifikáciu.

V ¹ preto analyzovali 22 rôznych vzoriek látky, 11 z triedy petrolej a 11 z triedy nafta (rozdelenie všetkých vzoriek do tried bolo teda vopred známe, aby mohla prebehnúť aj samotná validácia výsledkov). Chemicko-laboratórnou analýzou s využitím chromatografie získali pre každú z 22 vzoriek dáta, ktoré pozostávali z 9 premenných (relatívne zastúpenia n-alkánov). Cieľom práce bolo pritom zistiť, či možno uvedené 2 triedy pomocou získaných údajov nejako štatisticky rozlíšiť a navrhnúť čo najpresnejší spôsob klasifikácie. V ¹ pritom použili hierarchické zhukovanie a následne test pomerom vierohodností. V ² analyzovali rovnaké dáta, využili pritom model bayesovskej siete. V oboch článkoch pritom ukázali opodstatnenosť štatistického prístupu, keďže obe metódy správne klasifikovali väčšinu z 22 vzoriek. Problémom je skôr ťažšia implementácia navrhnutého postupu vyžadujúca pokročilejšie znalosti štatistiky a programovania (najmä prípad analýzy pomocou bayesovskej siete). V článku sa preto pokúšame analyzovať rovnaké dáta zrozumiteľnejším spôsobom, využívajúc pritom 3D-PCA vizualizáciu na jednoduchšiu používateľskú analýzu.

Predspracovanie dát

Vo všetkých nasledujúcich krokoch sme používali voľne stiahnuteľný štatistický softvér R ³ s balíkom *rgl* ⁴. Pred samotnou klasifikáciou sme dáta predspracovali rovnakým

¹ ZADORA, G., BORUSIEWICZ, R., ZIEGBA-PALUS J. Differentiation between weathered kerosene and diesel fuel using automatic thermal desorption GC-MS analysis and the likelihood ratio approach, In *Journal of Separation Sciences*, 2005, 28, s. 1467–1475. ISSN: 1615-9306

² ZADORA, G. Evaluation of the evidential value of physicochemical data by a Bayesian network approach. In *Wiley Journal of Chemometrics*, 2010, 24, s. 346-366. ISSN: 1099-128X

³ R DEVELOPMENT CORE TEAM (2007) R Development Core Team, R: a Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna.

⁴ ADLER, D. <dadler@uni-goettingen.de> and Duncan Murdoch <murdoch@stats.uwo.ca> (2011). *rgl*: 3D visualization device system (OpenGL). R package version 0.92.798. <http://CRAN.R-project.org/package=rgl>

spôsobom (na objektívne porovnanie výsledkov) ako v ⁵. Z 9 spomenutých premenných meraných na 22 vzorkách sme 4 vylúčili, keďže pri každej z nich nebol aspoň pri jednej vzorke detekovaný n-alkán. Dáta sme následne štandardizovali s cieľom eliminovať náhodné vplyvy nasledujúcim spôsobom:

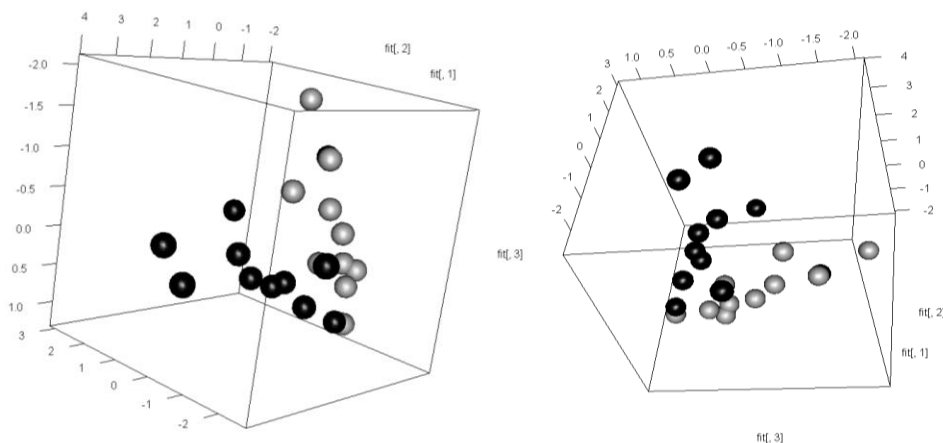
$$stdVar_i = \log_{10} \frac{Var_i}{Var_{12}} \quad i = 10, 11, 13, 14$$

Symbol Var_i pritom označuje i -tu premennú a $stdVar_i$ tú istú premennú po štandardizácii. Z 5 pôvodných premenných sme takto dostali 4 nové (štandardizované premenné), keďže sme pôvodné premenné delili premennou Var_{12} .

3D-PCA projekcia dát

Metóda hlavných komponentov⁶ sa používa na zníženie dimenzie dát najčastejšie do dvoch alebo troch rozmerov, aby boli dáta ľahšie vizualizovateľné. Pôvodné premenné (v našom prípade sú 4) sa snaží nahradiť novým počtom premenných (v našom prípade tromi), aby prišlo k čo najmenšej strate informácie.

Jednou z možností je vizualizovať dáta do roviny (teda dvoch rozmerov). V tomto prípade ale 2 nové premenné nemusia dostatočným spôsobom popisovať vzťahy medzi pôvodnými premennými. Použili sme preto projekciu vzoriek do 3 rozmerov, viď obrázok 1. Na tento účel sme použili spomínaný balík *rgl*, ktorý umožňuje interaktívne manipulovať s výsledným zobrazením. Pohybom myši sa teda dá objekt na obrázku 1 otáčať, približovať a vzdďaľovať. Výhodou je teda jednoduchá orientácia medzi vzorkami reprezentovaná čiernymi (petrolej) a sivými (nafta) guľčkami.



Obrázok 1: 3D-PCA projekcia dát (čierne guľčky predstavujú vzorky petroleja a sivé nafty), pohľad z dvoch rôznych uhlov.

Výsledky

⁵ ZADORA, G., BORUSIEWICZ, R., ZIEGBA-PALUS J. Differentiation between weathered kerosene and diesel fuel using automatic thermal desorption GC-MS analysis and the likelihood ratio approach, In *Journal of Separation Sciences*, 2005, 28, s. 1467–1475. ISSN: 1615-9306

⁶ PEARSON, K. On Lines and Planes of Closest Fit to Systems of Points in Space. In *Philosophical Magazine*, 1901 2 (6), s. 559–572. ISSN: 1478-6435

Úlohou výskumu v ⁷ a ⁸ bolo zistiť, či možno vzorky pochádzajúce z dvoch odlišných tried nejako štatisticky rozlíšiť, a tak identifikovať látku, ktorou bol zapálený objekt. Pri projekcii všetkých 22 vzoriek pomocou 3D-PCA metódy vidieť, že petrolejové vzorky (čierne guľičky na obrázku 1) metóda projektovala približne do tej istej časti priestoru. Podobne vzorky, pri ktorých bola ako zápalná látka použitá nafta, boli projektované do inej časti priestoru. Uvedený problém teda možno riešiť aj pomocou metódy 3D-PCA.

Uvedených 22 vzoriek predstavovalo len testovacie dáta, slúžiace na vytvorenie samotného klasifikačného modelu. V prípade obdržania novej vzorky, pri ktorej ale už nebudeme vedieť vopred príslušnosť ku konkrétnej triede, môžeme postupovať nasledovne:

1. Chemickou analýzou popísanou v ⁷ nameriame spomínaných 5 premenných (relatívne zastúpenia vybraných n-alkánov v látke).
2. Premenné štandardizujeme rovnakým spôsobom ako vyššie a dostaneme tak hodnoty 4 nových premenných pre analyzovanú vzorku.
3. Vzorku projektujeme do 3D priestoru použitím PCA metódy a podľa polohy novej vzorky posúdime, či patrí do prvej alebo druhej triedy. Treba pritom poznamenať, že podobnosť vzorky s nejakou triedou v štatistickom zmysle slova znamená vzdialenosť tejto vzorky od konkrétnej triedy. Pri klasifikácii teda jednoducho povedané sledujeme, či je naša vzorka bližšie k petrolejovým alebo naftovým vzorkám. V prípade, ak je problém určiť triedu vizuálne, môžeme vypočítať euklidovskú vzdialenosť ku všetkým 22 vzorkám a následne nájsť vzorku s najkratšou vzdialenosťou od novej vzorky. Ak bude táto vzorka z triedy petrolej, tak novú vzorku klasifikujeme do tejto triedy (v opačnom prípade do triedy nafta).

Záver

Navrhnutá metóda predstavuje len jeden z možných prístupov. Ako väčšina štatistických metód aj klasifikácia pomocou 3D-PCA projekcie pracuje s určitou chybou, čo treba zohľadniť pri vyvodzovaní možných dôsledkov. V budúcnosti plánujeme na uvedené dáta použiť ďalšie metódy s cieľom porovnania a validácie získaných výsledkov.

PodĎakovanie

Táto práca vznikla za podpory grantu VEGA 1/0103/10.

Literatúra

ADLER, D. <dadler@uni-goettingen.de> and Duncan Murdoch <murdoch@stats.uwo.ca> (2011). rgl: 3D visualization device system (OpenGL). R package version 0.92.798. <http://CRAN.R-project.org/package=rgl>

PEARSON, K. On Lines and Planes of Closest Fit to Systems of Points in Space. In *Philosophical Magazine*, 1901 2 (6), s. 559–572. ISSN: 1478-6435

R DEVELOPMENT CORE TEAM (2007) R Development Core Team, R: a Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna.

⁷ ZADORA, G. Evaluation of the evidential value of physicochemical data by a Bayesian network approach. In *Wiley Journal of Chemometrics*, 2010, 24, s. 346-366. ISSN: 1099-128X

⁸ ZADORA, G., BORUSIEWICZ, R., ZIEGBA-PALUS J. Differentiation between weathered kerosene and diesel fuel using automatic thermal desorption GC-MS analysis and the likelihood ratio approach, In *Journal of Separation Sciences*, 2005, 28, s. 1467–1475. ISSN: 1615-9306

ZADORA, G. Evaluation of the evidential value of physicochemical data by a Bayesian network approach. In *Wiley Journal of Chemometrics*, 2010, 24, s. 346-366. ISSN: 1099-128X

ZADORA, G., BORUSIEWICZ, R., ZIEGBA-PALUS J. Differentiation between weathered kerosene and diesel fuel using automatic thermal desorption GC-MS analysis and the likelihood ratio approach, In *Journal of Separation Sciences*, 2005, 28, s. 1467–1475. ISSN: 1615-9306

Key words: kerosene, diesel fuel, PCA method, classification, cluster analysis

Summary

In recent years, there have been many attempts to apply statistics in forensic sciences. Mathematical statistics methods are able to reveal hidden relations between objects, to identify similar objects or to classify samples to classes. Especially, the classification problem is the topic of this article. Authors propose different approach as proposed in ⁹. We are using principal component analysis on same data and we project all samples into 3D space. Membership to classes (kerosene or diesel fuel) is described using nearest neighbor approach (i.e. any new sample is classified as kerosene sample, if the nearest testing sample is also kerosene, analogously for diesel fuel sample). Visualization advantages of 3D- PCA projection are also presented, because they allow simple analysis also for specialists not so good in programming. We are using open-source statistical software R in this paper.

*Mgr. Miroslav Sabo
Katedra matematiky a deskriptívnej geometrie
Stavebná fakulta STU v Bratislave
Radlinského 11, 813 68, Bratislava
email: sabo@math.sk
Tel. č.: 0902 167 061*

*RNDr. Tatiana Hajdúková
Katedra manažmentu a informatiky
Akadémia Policajného zboru v Bratislave
Sklabinská 1, 835 17, Bratislava 35
email: tatiana.hajdukova@minv.sk
Tel. č.: 0961 057 297*

Recenzent: npor. Ing. Milan Marcinek

⁹ ZADORA, G., BORUSIEWICZ, R., ZIEGBA-PALUS J. Differentiation between weathered kerosene and diesel fuel using automatic thermal desorption GC-MS analysis and the likelihood ratio approach, In *Journal of Separation Sciences*, 2005, 28, s. 1467–1475. ISSN: 1615-9306