Objavovanie znalostí, prvá etapa

Martin Kalužník, Michal Kren

FIIT STU Bratislava

**Abstract.** Pozorovania UFO sú najmä v Amerike častým javom. V našej práci aplikujeme metódy analýzy dát, konkrétne zhlukovanie, aby sme v datasete 80 000 pozorovaní UFO z celého sveta našli medzi týmito pozorovaniami určité súvislosti. Z prieskumnej analýzy nám vyplynulo, že hlavné črty, podľa ktorých budeme záznamy zhlukovať, sú tvar a miesto úkazu.

**Keywords:** analýza dát, zhlukovanie, dbscan, space-time clustering

1. Opis problému, motivácia

Pozorovania neidentifikovateľných lietajúcich objektov (ďalej len UFO) sú napriek konšpiračnej a kontroverznej povahe relatívne časté. V našej práci sme sa snažili zistiť, či sa tieto pozorovania nedajú podľa ich parametrov zaradiť do niekoľko skupín podľa spoločných vlastností. Vďaka zaradeniu viacerých pozorovaní do rovnakej skupiny budeme schopný identifikovať viacnásobné pozorovania jedného UFO aj na základe netriviálnych parametrov. Taktiež budeme vedieť lepšie stanoviť jav alebo objekt, ktorý ľudia skutočne videli - alebo naopak, zistíme že niektorá skupina bude príliš početná na jej zanedbanie a bude sa javiť ako príliš “skutočná”.

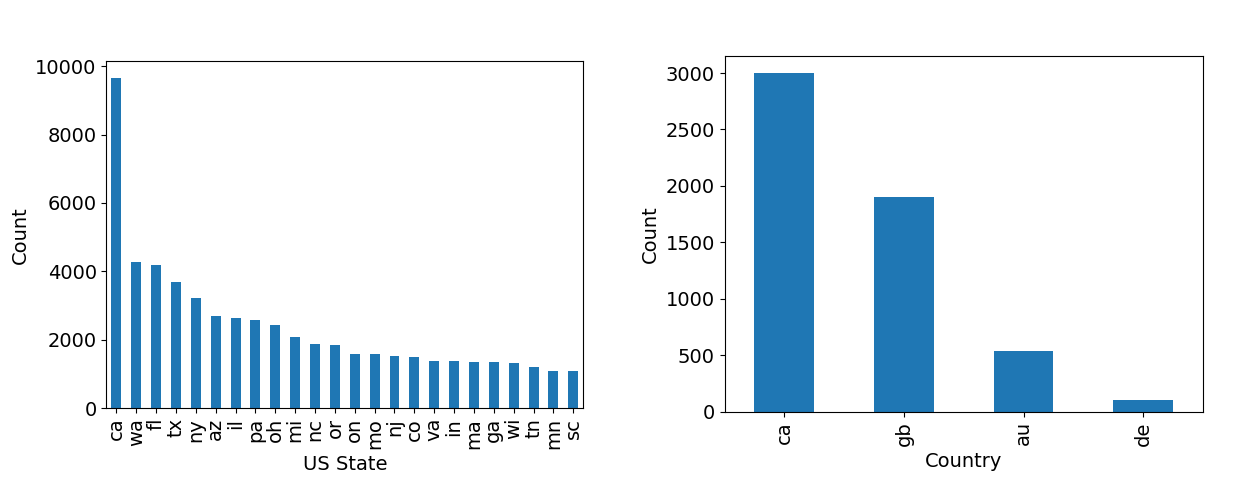
* 1. Opis dát spolu s charakteristikami dát

Dataset pozorovaní UFO sme získali z portálu Kaggle[[1]](#footnote-1) . Dáta obsahujú 11 čŕt, resp. stĺpcov a vyše 80,000 záznamov. Dataset obsahuje záznamy z celého sveta od 70tych rokov minulého storočia po súčasnosť. Dostupné boli dve verzie dát – pôvodné zozbierané dáta a vyčistená verzia zbavená nepoužiteľných či nekompletných záznamov. Dáta hovoria najmä o mieste a čase úkazu a taktiež obsahujú opis pozorovaného objektu a krátky slovný opis svedka. Základné charakteristiky týchto čŕt opisuje tabuľka 1.

**Table 1.** Všeobecný prehľad atribútov

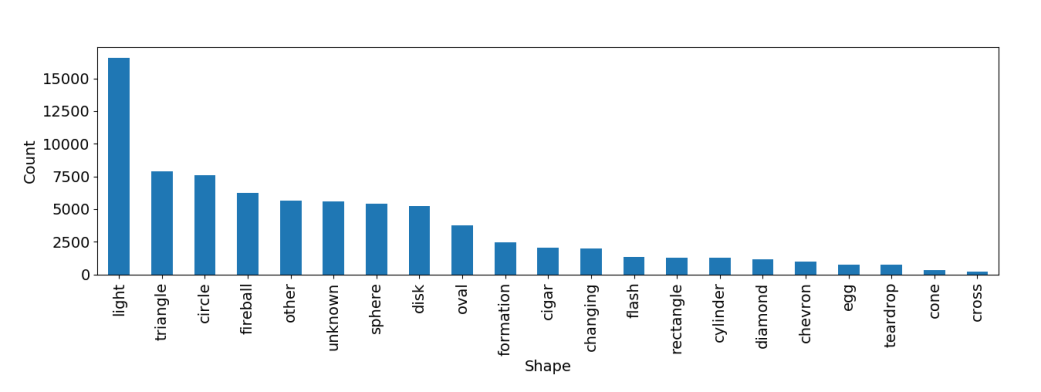
|  |  |  |
| --- | --- | --- |
| Názov | Typ | Popis |
| datetime | DateTime | dátum pozorovania |
| city | String | najbližšie mesto k pozorovaniu |
| state | String | štát pozorovania (len pre USA) |
| country | String | krajina pozorovania |
| shape | String | tvar pozorovaného UFO |
| duration (seconds) | Numeric | trvanie pozorovania |
| duration (hours/min) | String | slovný opis trvania pozorovania |
| comments | String | slovný opis pozorovania |
| date posted | DateTime | dátum nahlásenia pozorovania |
| latitude | Numeric | zemepisná šírka pozorovania |
| longitude | Numeric | zemepisná dĺžka pozorovania |

Dáta vznikali pravdepodobne ručným vypĺňaním veľkým počtom respondentov, čiže očakávali sme isté nečistoty v dátach, ako napr. hodnoty v nesprávnych stĺpcoch alebo formátoch. Aby sme odhadli, akú veľkú časť dát nepoužijeme, a taktiež aby sme si spravili prvotnú predstavu o samotných dátach, vykonali sme prieskumnú analýzu. Najzaujímavejšie pre nás vyšli stĺpce *country*, *state* a *shape*. Vytvorili sme pre tieto črty histogramy počtu pozorovaní (obr. 1 – 2).



**Fig. 1.** Počet pozorovaní UFO v jednotlivých krajinách a štátoch USA

Z obrázka 1 vyplýva, že UFO bolo najčastejšie pozorované v pobrežných štátoch USA. Obrázok 2 hovorí o tvare, ktorým svedkovia opísali údajne UFO. Už v tejto skorej fáze je vidno, že dáta si budú vyžadovať predspracovanie, keďže napr. “*Sphere*”, “*Circle*” a “*Disk*” sú podľa nás veľmi ťažko od seba rozpoznateľné a teda bolo by vhodné ich zlúčiť.



**Fig. 2.** Počet pozorovaní podľa tvaru UFO

Ostatne atribúty, ktoré majú pre nás význam sú :

* čas a trvanie pozorovania
* v ktorej časti dňa sa pozorovanie udialo - deň, noc, šero...
* opis udalosti, z ktorej vieme vyextrahovať ďalšie črty
* presné súradnice pozorovania
  1. Opis prác iných autorov

Zhlukovanie pozorovaní UFO bolo predmetov skúmania už dávnejšie. (Malozemo 2012) skúmali aplikovanie viacerých metód strojového učenia, okrem iného aj zhlukovanie. Využili algoritmus K-Means pri k=8. Použili takmer rovnaké atribúty ako sme využili my. Z datasetu uvažovali len pozorovania v intervale január 2000 – november 2012 (dátum písania práce). Pozorovania UFO sa snažili klasifikovať ako planéty, satelity, ISS apod. Z ich experimentu vyplynulo, že trénovanie bez učiteľa, t.j. zhlukovanie, neprinieslo požadované výsledky. Oveľa úspešnejšia bola klasifikácia, pri ktorej využili dodatočné informácie o pozorovaniach z NASA. Tieto informácie predstavovali vysvetlenie pozorovania, t.j. vopred zmienené planéty, satelity, ISS. Tieto informácie však boli poskytnuté iba k 4.4% záznamov z datasetu. Na tejto podmnožine dosiahla ich klasifikácia úspešnosť 43%.

Problém, ktorý riešime, sa vo všeobecnosti označuje ako časo-priestorové zhlukovanie (ang. Space-time clustering). Analýza pozorovaní UFO je iba jednou z mála aplikácií tejto metódy. Oveľa častejšie využitie nachádza v oblasti medicíny, pri analýze šírenia rôznych chorôb. (Root et al. 2012) skúmali rozšírenie vírusu RSV na Filipínach v rokoch v 2000 až 2004. Ľudí pozitívnych na tento vírus rozdelili pomocou zhlukovania na dve pod-skupiny RSV-A a RSV-B. Využitím metódy *spatial scan statistic,* Bernoulliho a Multinomiálneho modelu zistili, že vírus sa rozširuje vo v jednoznačných lokálnych oblastiach a v rôznych bodoch v čase. Tento výskum má výrazne pomôcť pri prevencii rozšírenia vírusu v budúcnosti.

Tento typ zhlukovania sa používa tiež na analýzu prírodných javov, napr. hurikánov. (Paliwal et al. 2011) analyzovali správanie hurikánov v oblasti Indického oceánu. Včasná predikcia trajektórie, intenzity a trvania hurikánu môže výrazne prispieť pri prevencií, príprave a znižovaní škôd spôsobených hurikánmi. Pomocou algoritmu K-Means zhlukovali hurikány podľa 5 atribútov – x,y súradnice a 3 atribúty odpovedajúce kovariančnej matici. Poradilo sa im hurikány rozdeliť do 4 zhlukov.

1. Definovanie úlohy objavovania znalostí

V predošlej kapitole sme opísali črty nášho datasetu a ich potenciál pri určovaní podobností. Objavovať súvislosti medzi pozorovaniami sme sa rozhodli pomocou metódy zhlukovania. Hlavným problémom tejto práce predspracovanie dát, identifikácia vzťahov medzi atribútmi a voľba relevantných vstupných atribútov pre zhlukovanie (tzv. feature engineering).

* 1. Predspracovanie a výber atribútov

V rámci predspracovania sme sa chceli zaoberať atribútmi *city,* *state* a *country* a pokúšať sa overiť ich správnosť pomocou atribútov *latitude* a *longitude*. Zistili sme ale, že samotné mesto, štát a krajina by boli kategorické atribúty a nemali by dostatočnú výpovednú hodnotu. Navyše geografická šírka a dĺžka sa dá priamočiaro previesť do intervalu požadovaného DM metódami a oveľa lepšie vypovedá o mieste pozorovania. Preto sme sa rozhodli pracovať len s atribútmi *latitude* a *longitude* a atribúty *city,* *state* a *country* zanedbať.

Z atribútu *datetime* sme sa rozhodli vytvoriť kategorické atribúty, ktoré budú zovšeobecňovať čas pozorovania. Jeden z týchto atribútov hovorí o časti dňa počas ktorej sa pozorovanie odohralo (doobedie, podvečer, večer, hlboká noc...) – *time\_of\_day*. Iný o ročnom období počas ktorého pozorovanie nastalo – *season*.

Kategorický atribút *shape* bolo potrebné pretransformovať na numerický atribút. Vytvorili sme binárne stĺpce pre každú unikátnu hodnotu v tomto stĺpci, čím vzniklo 26 nových atribútov. Chýbajúce hodnoty sme sa pokúsili odvodiť zo stĺpca *comments*, avšak hľadanie hodnoty stĺpca *shape* pomocou metódou *bag-of-words* neprinieslo žiadne výsledky. Sémantická analýza komentárov presahuje rozsah tejto práce.

Atribút *duration* predstavuje dĺžku úkazu, resp. jeho pozorovania. V dátovej sade je reprezentovaných v dvoch formách - v textovej a číselnej (počet sekúnd) podobe. Pre naše potreby sa samozrejme vhodnejšia číselná forma.

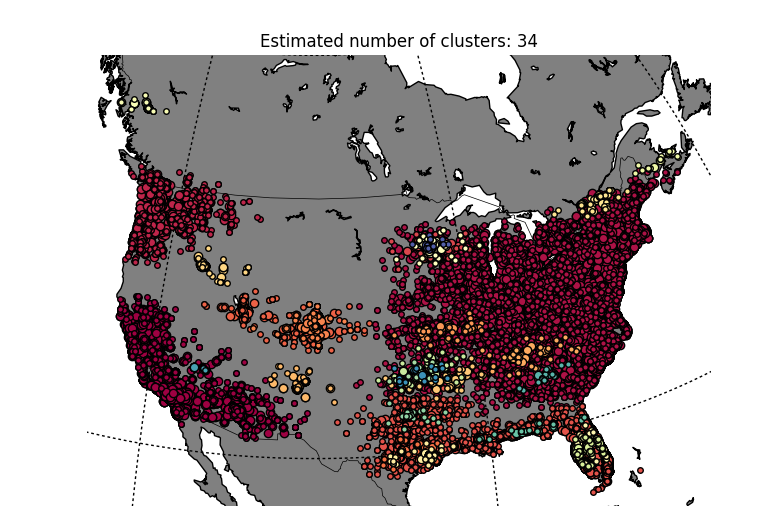
Atribút *date posted* sme sa rozhodli porovnať s atribútom *datetime* a zanedbať záznamy, kde ich rozdiel dosahoval hodnôt vyše pár mesiacov, pretože očakávame, že takéto záznamy nebudú dostatočne presné.

* 1. Hľadanie súvislostí v dátach

Plánujeme sa zamerať na zhlukovanie podľa dvoch hlavných atribútov – tvar pozorovaného objektu a predspracovanej polohy. Postupne budeme pridávať ostatné atribúty a pozorovať ich vplyv na výsledky. Našou hlavnou úlohou je tiež odhaliť korelácie, ktoré pri určovaní správnych čŕt pre zhlukovanie nie sú žiaduce.

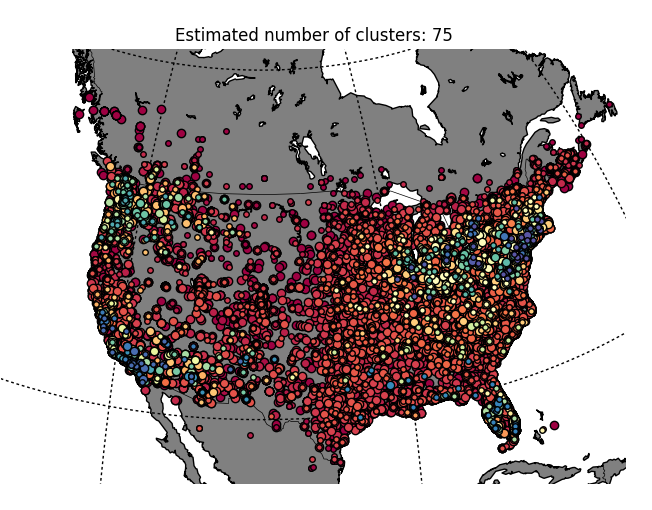
Na hľadanie súvislostí medzi pozorovaniami sme využili zhlukovanie. Rozhodli sme sa pre algoritmus DBSCAN, keďže v našom prípade je vhodnejšie ak počet zhlukov nie je pevne daný, ale algoritmus si sám určí ich optimálny počet. Toto napríklad algoritmus K-Means nedokáže a preto je pre naše použitie nevhodný.

* 1. Prvotné experimenty

Podarilo sa nám pomocou metódy DBSCAN rozdeliť pozorovania v Amerike a v Európe do rôznych zhlukov. Začali sme s využitím iba parametrov *latitude*, *longitude* a *timestamp*. Výstup zhlukovania v tejto konfigurácií vidno na obrázku Fig. 3.

**Fig. 3.** Zhluky s využitím súradníc a časovej stopy

K pôvodným trom parametrom sme pridali parameter *shape*, upravený do binárnej formy, čo predstavuje 26 tzv. *dummy* stĺpcov. Obrázok Fig. 4 predstavuje výstup tejto konfigurácie.

**Fig. 4.** Zhluky pri pridaní tvaru pozorovaného objektu

* 1. Vyhodnocovanie

Úspešnosť našej metódy overíme pomocou niekoľkých metrík. Keďže ide o metódu učenie bez učiteľa, nevieme overiť presnosť podľa žiadnej testovacej sady. Na overenie použijeme štandardnú metriku hodnotenia úspešnosti zhlukovania – meraním variability prvkov v rámci jedného zhluku a medzi zhlukmi. Zhlukovanie je úspešnejšie, čím viac sú si prvky v rámci zhluku podobné a čím viac sa odlišujú od prvkov iných zhlukov. Hlavným problémom je zadefinovať si kritérium úspešnosti, predbežne sme si zvolili kosínusovú vzdialenosť.

Okrem kvantitatívneho vyhodnocovania vykonáme aj kvalitatívne, manuálnym skúmaním podobností jednotlivých prvkov v zhluku. Pre každý zhluk vytvoríme histogram najviac opakujúcich sa hodnôt jednotlivých atribútov. Z týchto hodnôt budeme vedieť určiť, či zhlukovanie nie je príliš ovplyvnené jedným atribútom alebo, naopak, či hodnoty jednotlivých atribútov v rámci zhluku sú si dostatočne podobné.

* 1. Plán na dokončenie projektu

Vykonali sme prvotné experimenty, ktoré už v tejto skorej fáze považujeme za úspešné. Ďalším krokom bude predspracovať ďalšie atribúty a experimentovať s ich zapojením do zhlukovania, konkrétne pôjde o časové atribúty *time\_of\_day* a *season*.. Plánujeme tiež vyskúšať redukciu dimenzionality pomocou algoritmu PCA, keďže prevedenie kategorických atribútov do binárnej podoby rapídne zvýšilo počet atribútov vstupujúcich do zhlukovania. Taktiež je ešte potrebné implementovať samotné metódy vyhodnocovania, kvantitatívne aj kvalitatívne.

1. Zdroje

Malozemo, A.J., 2012. CMSC 726 Final Project Report : Classifying UFO Sightings. , pp.1–14. Available at: https://www.cs.umd.edu/~amaloz/ufo/report.pdf.

Paliwal, M., Patwardhan, A. & Sarda, N.L., 2011. Analyzing tropical cyclone tracks of North Indian Ocean. *Proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications - COM.Geo ’11*, pp.1–6. Available at: http://dl.acm.org/citation.cfm?id=1999320.1999338.

Root, E.D. et al., 2012. Evolution of respiratory syncytial virus (RSV) over space and time in rural Filipino children. *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, (c), pp.57–62.

1. https://www.kaggle.com/NUFORC/ufo-sightings [↑](#footnote-ref-1)