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”.

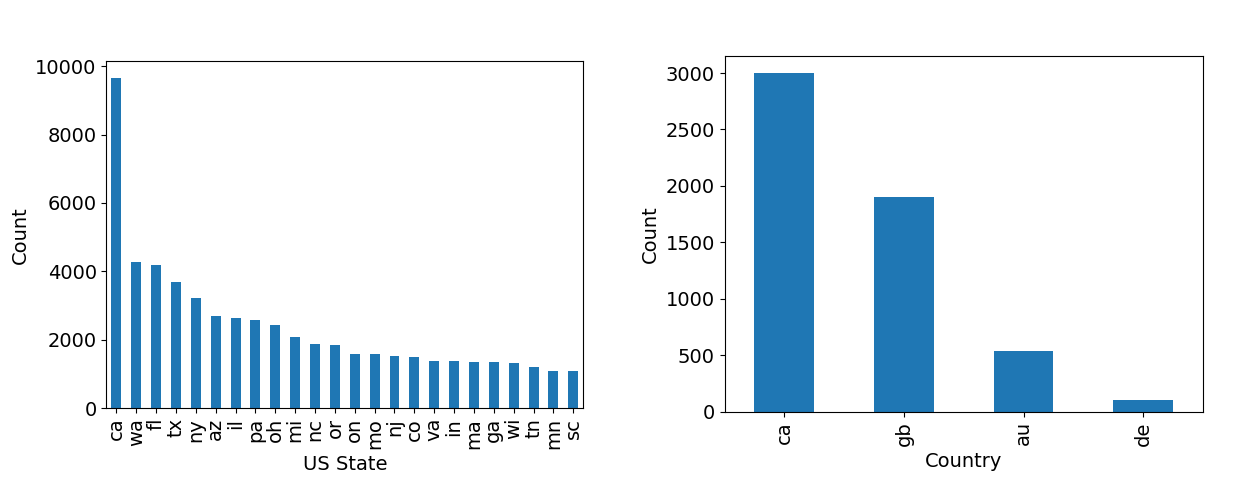
* 1. Opis dát spolu s charakteristikami dát

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

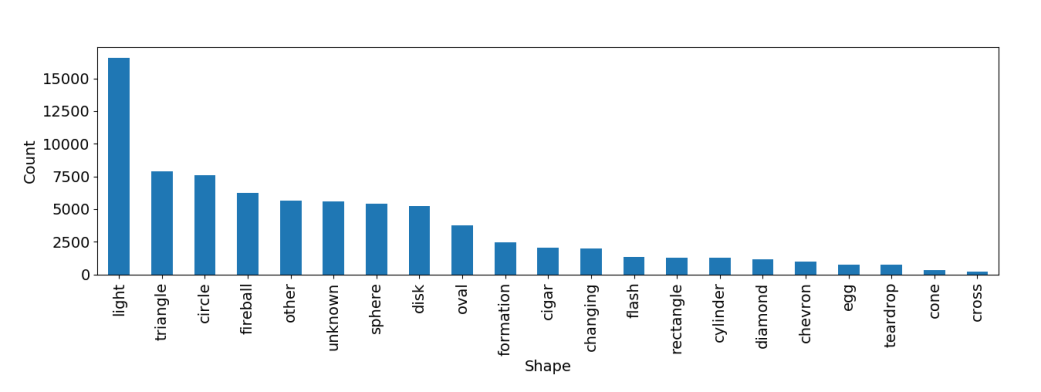
|  |  |  |
| --- | --- | --- |
| Názov | Typ | Popis |
| datetime | DateTime | dátum pozorovania |
| city | String | najbližšie mesto k pozorovaniu |
| state | String | štát pozorovania (len pre USA) |
| country | String | krajina pozorovania |
| shape | String | tvar pozorovaného UFO |
| duration (seconds) | 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. (Alex J Malozemoff) vykonali 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 poskytnutú iba k 4.4% záznamom 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. FIXME – doplniť podrobnosti

Taktiež sa takéto zhlukovanie používa na analýzu prírodných javov, napr. hurikánov. FIXME – doplniť podrobnosti

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 je správne určenie dôležitosti jednotlivých čŕt (tzv. feature engineering).

* 1. Predspracovanie 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, ...). Iný o ročnom období počas ktorého pozorovanie nastalo.

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. Pokúsili sme sa chýbajúce hodnoty odvodiť zo stĺpca *comments*, avšak hľadanie hodnoty stĺpca *shape* pomocou metódou *bag-of-words* neprinieslo žiadne výsledky. Sémantická analýza komentárov presahuje rozsah tejto práce.

Duration text vs netext – jeden zanedbáme lebo sa nám to nechce overovať

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. Výber atribútov

Vyberieme všetky čo sme v predošlom odseku nezanedbali, pretože ich nemáme až tak veľa.

* 1. ~~Predpokladaný scenár riešenia~~

~~Prvým krokom riešenia je predspracovanie datasetu. Okrem vyčistenia dát od nepoužiteľných záznamov tiež potrebujeme napr. vyextrahovať kľúčové slová z popisu pozorovania, na základe ktorých vieme určiť podobnosť jednotlivých pozorovaní. Zaujíma nás aj vzťah tohto popisu a tvaru UFO, keďže predpokladáme, že tieto dve črty budú spolu korelovať~~

* 1. ~~Predspracovanie a čistenie dát~~

~~Názvy miest pozorovania sú zadávané používateľom a často obsahujú aj iné slová a znaky, plánujeme preto využiť zemepisnú šírku a dĺžku na kontrolu alebo aj nahradenie parametrov~~ *~~city~~*~~,~~ *~~state~~* ~~a~~ *~~country~~*~~.~~

~~Dataset obsahuje dva atribúty týkajúce sa dĺžky pozorovania -~~ *~~duration (seconds)~~* ~~a~~ *~~duration (hours/min)~~*~~, pričom jeden z nich bol zadávaný používateľom ako voľný text, preto chceme skontrolovať konzistenciu týchto údajov a vybrať vhodnejší na ďalšie spracovanie.~~

~~Atribút~~ *~~shape~~* ~~obsahuje síce limitovaný počet hodnôt – 29 rôznych tvarov, z ktorých však viaceré sú synonymá. Chceme preto ich počet znížiť a manuálne vybrať niekoľko najvhodnejších ‘tvarov’. Zároveň, pre pozorovania, pre ktoré nebol nahlásený tvar, sa pokúsime chýbajúcu informáciu získať z atribútu~~ *~~comment~~*~~, ktorý môže obsahovať zmienku o tvare - jedno z nami určených slov.~~

~~Rozdiel medzi atribútmi~~ *~~datetime~~* ~~a~~ *~~date posted~~* ~~značí časový rozdiel medzi pozorovaním a jeho nahlásením. V prípadoch, keď je príliš veľký, nemusí byť pozorovanie nahlásené v dostatočnej kvalite a teda pre nás nie je významné.~~

* 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.

* 1. DM metódy

Na hľadanie súvislostí medzi pozorovaniami sme využili zhlukovanie. Rozhodli sme sa pre algoritmus DBSCAN. Keďže nepoznáme vopred počet zhlukov, nemohli sme použiť algoritmus K-Means.

* 1. Prvotné experimenty

Podarilo sa nám pomocou metódy DBSCAN a súradníc dostať pozorovania v Amerike a v Európe do rôznych zhlukov. Začali sme s využitím iba parametrov *latitude*, *longitude* a *timestamp*. Výstup zhlukovania v takejto konfigurácií vidno na obrázku N – FIXME – pridať obrázok.

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

Keďže z jedného kategorického stĺpca vzniklo 26 binárnych, pokúsili sme sa zredukovať dimenzionalitu pomocou PCA. Počet týchto stĺpcov sme znížili na FIXME – zistiť ideálnu hodnotu. Obrázok N predsatvuje výsledok zhlukovania v takejto konfigurácii. FIXME - obrázok

* 1. Vyhodnocovanie

FIXME Čo budete vyhodnocovať, aké miery, grafy, kritériá použijete.

Vieme kresliť na mapu sveta

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