Metody analýzy dat II

ANALÝZA – MAD II

Analýza pro vybraný data set – Tamil Nadu Electricity Board

Katanik David
7.5.2017

Obsah

Tamil Nadu Electricity Board 2
Obecné informace o datech 2
Tamil Nadu 2
Atributy 2
Důvod výběru těchto dat 3
Ukázka dat 3
Analýza 4
Předpracování dat 4
Prováděné akce nad datovou sadou 4
Grafická reprezentace 4
Četnosti instancí podle typu subjektu 4
Spotřeba elektrické energie v kW pro jednotlivé subjekty 5
Četnosti pro jednotlivé subjekty v jiné formě 5
Koláčový graf zastoupení subjektů 5
Vizualizace dat v software Weka 6
Závislost spotřeby v kW na VA 6
Závislost spotřeby v kW na sektoru vzhledem ke spotřebě ve VA 7
Shlukování na základě typu a spotřeby v kW s vizualizací závislosti 8
Klasifikace pomocí Bayesova algoritmu 9
Shrnutí 9
Závěry analýzy 9
Zdroje 10

Tamil Nadu Electricity Board

Obecné informace o datech

Data byla pořízena 22. 12. 2013 K. Kalyanim. Poskytují informace o spotřebě elektrické energie rezidenčních, komerčních, industriálních a hospodářských subjektech ve svazovém státě Tamil Nadu v jihovýchodní Indii. Obsahuje 45781 instancí a 5 atributů.

Tamil Nadu

Tento svazový stát jihovýchodní Indie má rozlohu okolo 130 000 km^2. Počet obyvatel dosahuje 62,5 milionu obyvatel s hustotou zalidnění 478ob./km^2. Tamní náboženství je z 89 % hinduismus následuje s 6 % islám a 5 % křesťanství. [1]

Obrázek 1 - Tamil Nadu [2]

Atributy

Název atributu	Typ	Popis
ForkVA	Reálné číslo	Spotřebované Volt/Ampéry
ForkkW	Reálné číslo	Spotřebované k Watty
Type	Výčet hodnot[1]	Typ zákazníka
Sector	Hodnota 1	Není definováno, bude vynechán z analýzy
ServiceID	Výčet hodnot	Identifikační číslo služby

[1] - Bank, AutomobileIndustry, BpoIndustry, CementIndustry, Farmers1, Farmers2, HealthCareResources, TextileIndustry, PoultryIndustry, Residential(individual), Residential(Apartments), FoodIndustry, ChemicalIndustry, Handlooms, FertilizerIndustry, Hostel, Hospital, Supermarket, Theatre, University

Důvod výběru těchto dat

Data jsem si zvolil z důvodů jejich reálnosti. Jelikož se oblast skutečně nachází a data jsou z již uvedeného roku, tak mi přišlo zajímavé se pokusit o jejich analýzu.

Ukázka dat

```
 1 0.865935636652813,0.143762528699181,Bank,1,671004572
 2 0.12980418301167,0.088929797893535,Bank,1,671004572
 3 0.061801486824636,0.552047074067644,Bank,1,671004572
 4 0.099116455214686,0.848172019260837,Bank,1,671004572
 5 0.20570390723589,0.624722465632105,Bank,1,671004572
 6 0.164028974259382,0.038167963920294,Bank,1,671004572
 7 0.61983409749181,0.079610680955383,Bank,1,671004572
 8 0.011324217122891,0.71847239966481,Bank,1,671004572
 9 0.344461180776512,0.271178427937514,Bank,1,671004572
10 0.679169150233846,0.191223430741136,Bank,1,671004572
11 0.191592100144425,0.381792419765995,Bank,1,671004572
12 0.940152001775472,0.981296832913215,Bank,1,671004572
13 0.543696404804164,0.210494581806471,Bank,1,671004572
14 0.074784530965709,0.407647780821232,Bank,1,671004572
15 0.667057346464578,0.441407887791691,Bank,1,671004572
16 0.860207737177882,0.671407822707594,Bank,1,671004572
17 0.177475410998072,0.931109520974205,Bank,1,671004572
18 0.10762750700023,0.309155999893746,Bank,1,671004572
19 0.695464116086839,0.992785710573213,Bank,1,671004572
20 0.622045571545376,0.063885043959955,Bank,1,671004572
21 0.658741862850935,0.874707970961486,Bank,1,671004572
22 0.029141099470774,0.711535063683138,Bank,1,671004572
23 0.721738121835287,0.631547891941216,Bank,1,671004572
24 0.516696168915683,0.14756826504183,Bank,1,671004572
25 0.164995785940689,0.369273829793661,Bank,1,671004572
```

Analýza

Předzpracování dat

Datový soubor byl zkontrolován. Neobsahuje chybějící hodnoty, překlepy apod. Atribut Sector nabývá pouze jednu hodnotu (=1) a bude v analýze zanedbán.

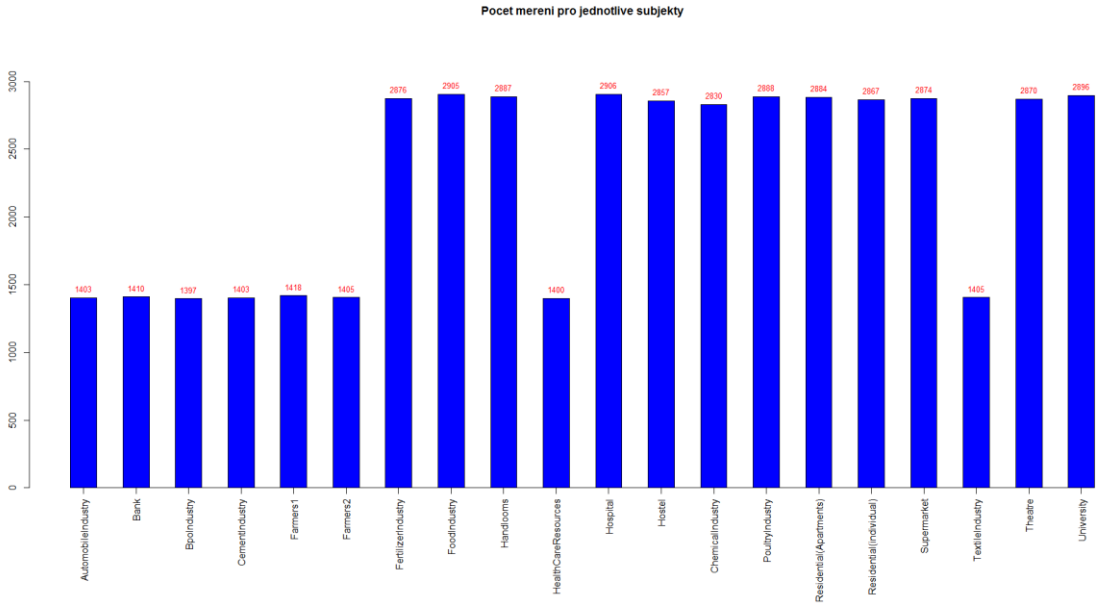
Prováděné akce nad datovou sadou

Cílem analýzy je zjistit, jakou minimální, maximální a průměrnou spotřebu elektrické energie mají jednotlivé sektory. Dále pokusit se definovat vztah mezi spotřebou ve VA a spotřebou v kW. Jelikož data neobsahují žádné speciální atributy, kterými se dá určit poloha jednotlivých subjektů, tak nelze určit spotřeba v určitém místě na světě (podle zdrojů totiž měl atribut service ID určovat i lokalizaci ve světě – toto se však nepotvrdilo při bližším prozkoumání).

Grafická reprezentace

Grafická reprezentace byla vytvořena pomocí více nástrojů. Konkrétně software R a Weka.

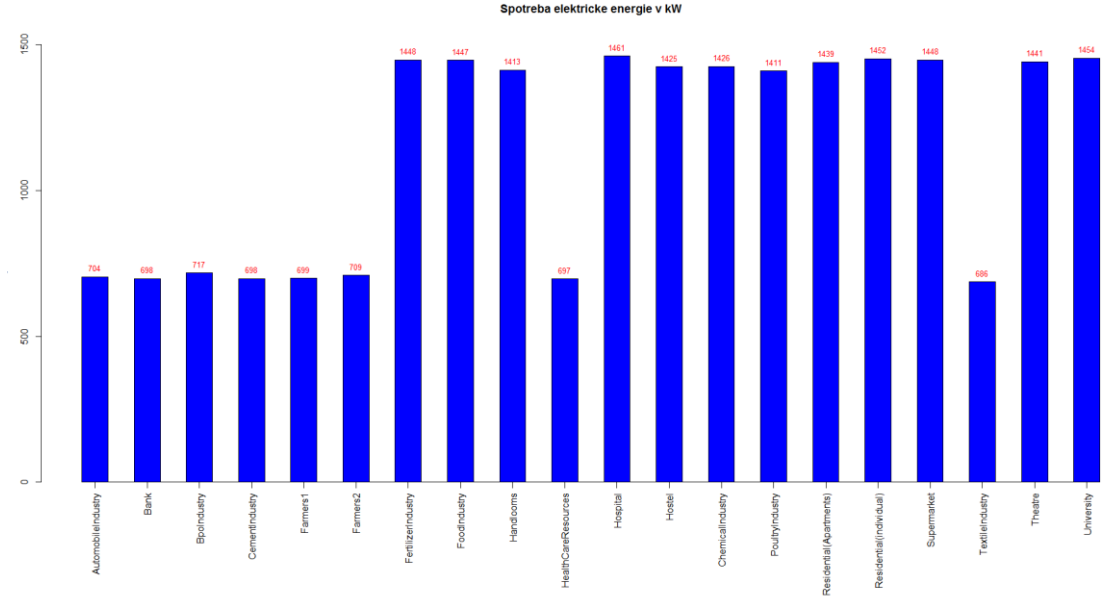
Četnosti instancí podle typu subjektu



Obrázek 2 Četnosti instancí podle typu

Z Obrázku 2 je patrné, že pro 12 subjektů máme dvojnásobný počet instancí. Z toho lze vyvozovat, že měření probíhá častěji u těchto 12ti subjektů. Další možným scénářem je, že bylo prováděno měření na více místech, tzn je větší počet subjektů daného typu.

Spotřeba elektrické energie v kW pro jednotlivé subjekty



Obrázek 3 - Spotřeba EE v kW pro subjekty

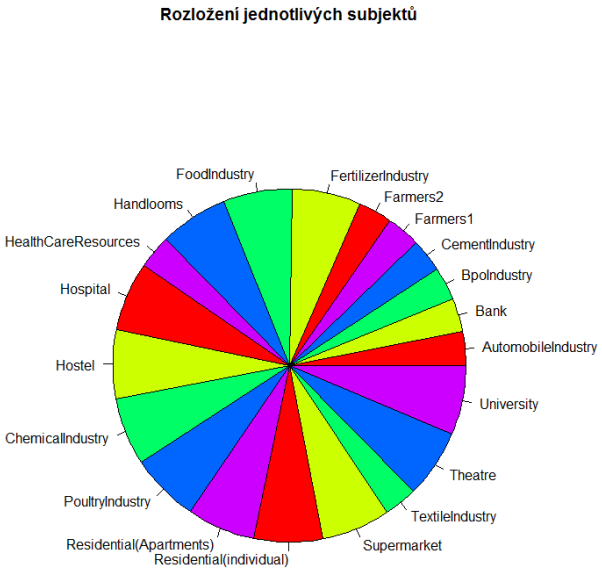
Na obrázku 3 lze zaznamenat celkovou spotřebu v kW pro jednotlivé typy subjektů. Z analýzy lze tedy vyvodit, že nejvíce spotřebovávají nemocnice. Naopak nejmenší odběr má textilní sektor.

Četnosti pro jednotlivé subjekty v jiné formě

AutomobileIndustry	Bank	BpoIndustry	CementIndustry	Farmers1
1403	1410	1397	1403	1418
Farmers2	FertilizerIndustry	FoodIndustry	Handlooms	HealthCareResources
1405	2876	2905	2887	1400
Hospital	Hostel	ChemicalIndustry	PoultryIndustry	Residential(Apartments)
2906	2857	2830	2888	2884
Residential(individual)	Supermarket	TextileIndustry	Theatre	University
2867	2874	1405	2870	2896

Jednotlivé absolutní četnosti instancí pro jednotlivé sektory v tabulce zobrazeny výše.

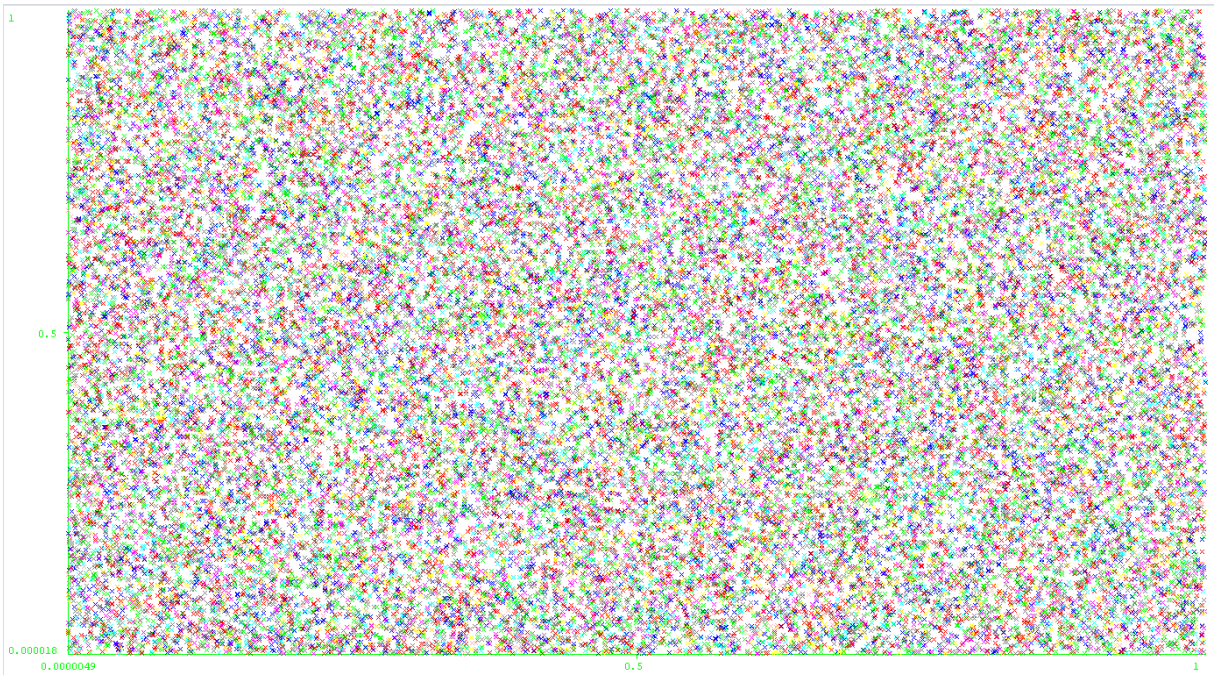
Koláčový graf zastoupení subjektů



Obrázek 4 - Zastoupení jednotlivých subjektů pomocí koláčového grafu

Vizualizace dat v software Weka

Vizualizace dat v software Weka proběhla následovně (na ose X je spotřeba ve VA a na ose Y je spotřeba v kW) a měla by ukázat závislost v převodu VA charakteristiky na reálnou spotřebu v kW:

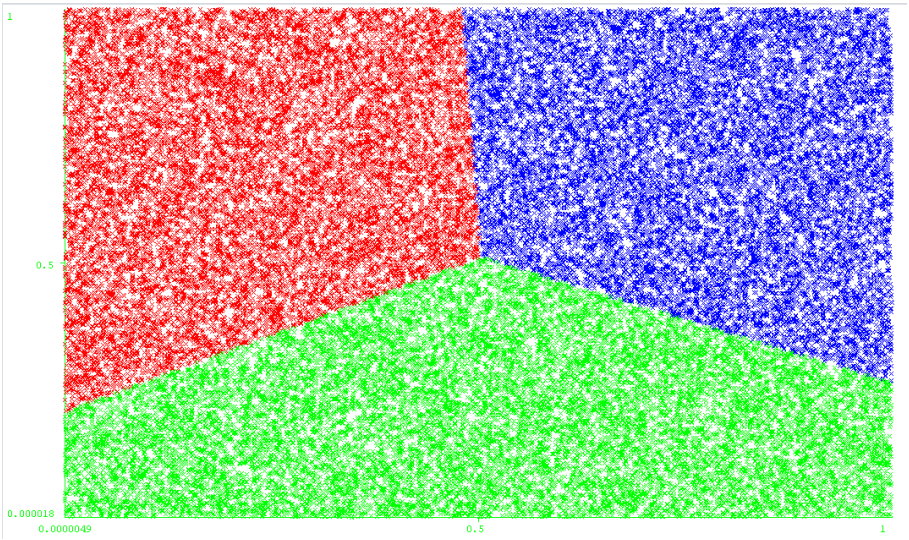


Obrázek 5 – Vizualizace kW a VA spotřeby

Z výsledků vizualizace můžeme usuzovat, že instancí je velmi mnoho a nejsou patrné žádné shluky. Nedá se tedy předpokládat, že by byl větší rozptyl mezi jednotlivými měřeními.

Závislost spotřeby v kW na VA

Následující vizualizace se snaží zobrazit závislost spotřeby v kW na VA. Z dostupných dat předpokládáme téměř rovnoměrné shlukování. Pomocí kMeans se třemi shluky.



Obrázek 6 - kMeans pro spotřebu

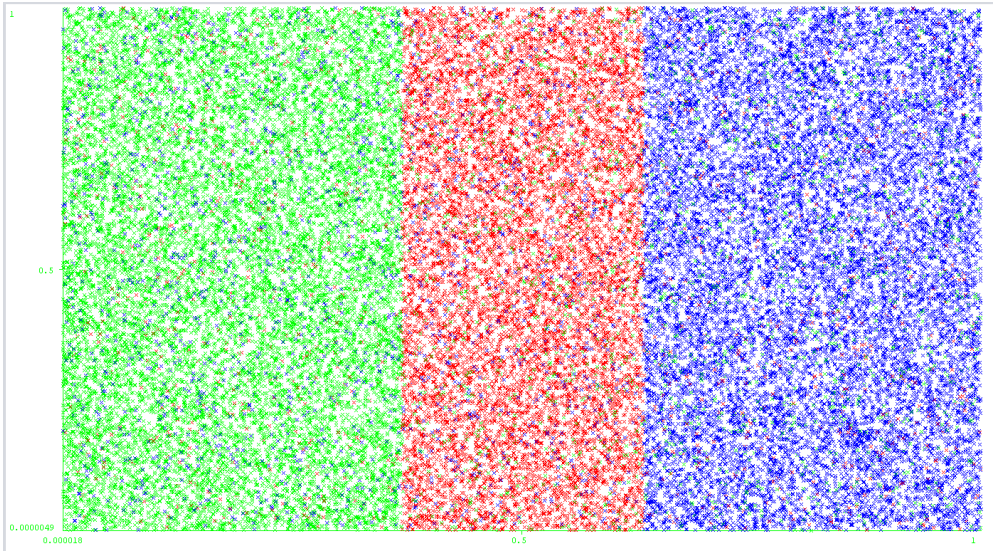
```
Clustered Instances

0       14192 ( 31%)
1       14503 ( 32%)
2       17086 ( 37%)
```

Z dostupného výsledku vidíme, že data jsou téměř rovnoměrně rozdělená a rozdíl mezi spotřebou v kW a VA je zanedbatelný. V další analýze by se teoreticky dalo pracovat pouze s jedním atributem.

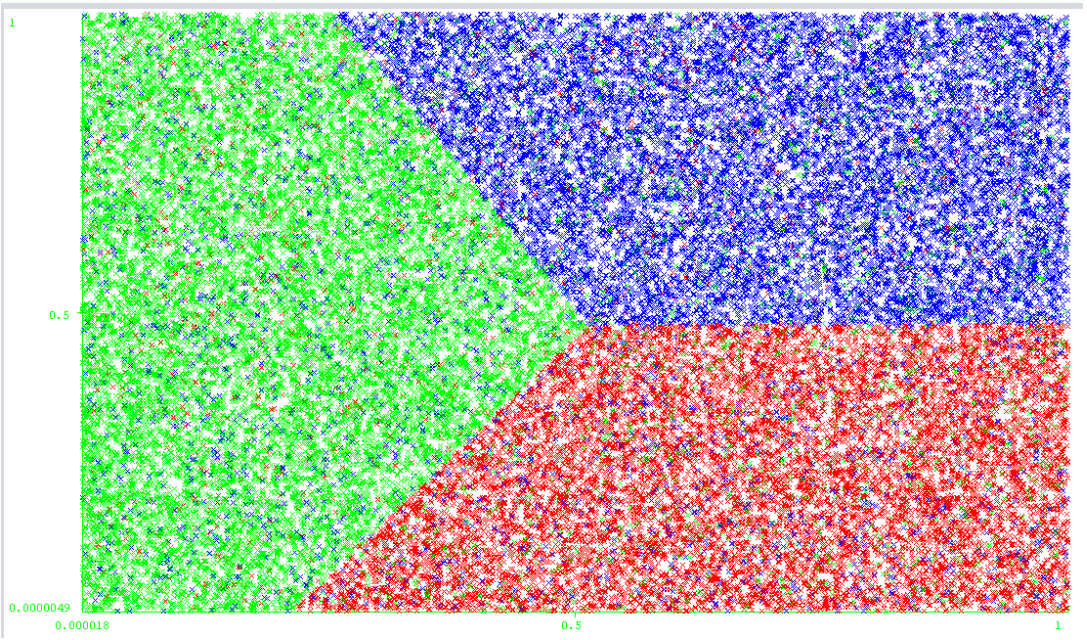
Závislost spotřeby v kW na sektoru vzhledem ke spotřebě ve VA

Dalším pokusem je zobrazení vztahu mezi spotřebou v kW a spotřebou VA při Kmeans shlukování pro tři shluky, kde shlukování bylo provedeno pouze s atributy kW a Typem. Zde předpokládáme (kvůli vztahu z obrázku 5), že instance by měly být zařazeny téměř stejně.



Obrázek 7 - kMeans s parametrem spotřeba v kW a typem

Z výsledků tedy můžeme potvrdit předpoklad, že většina bodů se neliší, a tedy použití spotřeby v kW nebo ve VA má minimální vliv na shlukování. Ovšem použití obou parametrů při shlukování by mělo výsledky upřesnit.



Obrázek 8 - Kmeans s parametry spotřeba v kW/VA a typem (vlajka)

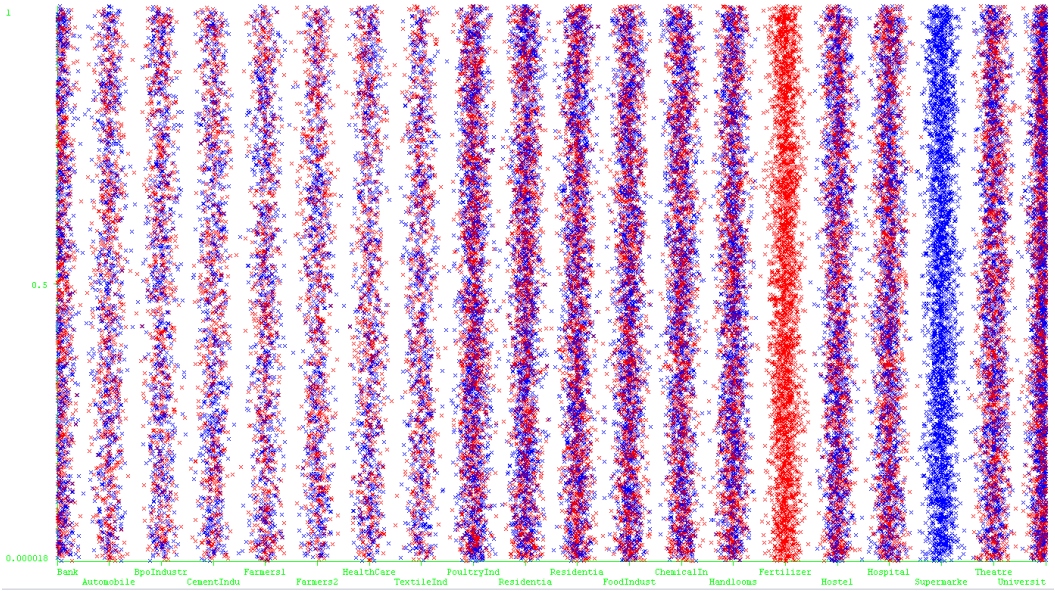
```
Clustered Instances

0       15278 ( 33%)
1       13144 ( 29%)
2       17359 ( 38%)
```

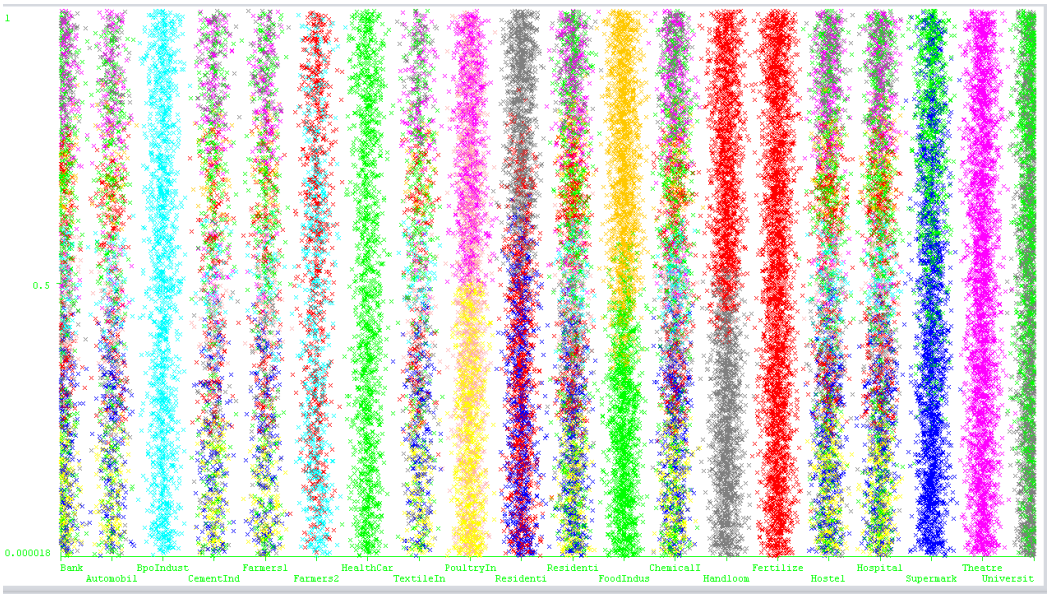
Z výsledků tedy potvrzujeme upřesnění rozložení instancí pro jednotlivé shluky. Oproti předchozím výsledkům se téměř vyrovnaly instance v jednotlivých shlucích. Navíc se vytvořil zajímavý obrazec připomínající vlajku.

Shlukování na základě typu a spotřeby v kW s vizualizací závislosti

Shlukování prostřednictvím atributu typ a spotřeby v kW lze znázornit takto:



Výsledek lze interpretovat tak, že instance některých subjektů drží „pospolu“, tzn. subjekt supermarket (modrý) je kompletně shlukován spolu s ostatními supermarkety. Což dokonale zobrazuje skutečnost, že všechny supermarkety mají velmi podobnou spotřebu. To samé se dá říci i o subjektech pracujících s hnojivy. Další subjekty již mají různě přeházené instance pro k rovno třem. Pokud položíme k rovno 20 (počtu subjektů), tak bychom mohli předpokládat, to samé, co popisuje tento text, ale u všech subjektů.



Předpoklad byl tedy špatný, jelikož některé instance se prolínají a není jednoznačně možné určit, že veškeré subjekty se shlukují společně. Toto značí malé odchylky v datech, je tedy velmi nelehké určit jasně daný shluk pro jakoukoliv instanci.

Klasifikace pomocí Bayesova algoritmu

Klasifikace pomocí Bayesova algoritmu by měla zařadit instanci do správné třídy. K tomuto účelu by teoreticky měl sloužit atribut serviceID, jelikož určuje identifikační číslo služby.

```
Correctly Classified Instances       29793               65.0772 %
Incorrectly Classified Instances     15988               34.9228 %
Kappa statistic                          0.6387
Mean absolute error                      0.023
Root mean squared error                  0.1057
Relative absolute error                 36.8046 %
Root relative squared error             59.8308 %
Total Number of Instances            45781
```

Z výsledků lze vyvodit, že serviceID není vhodný pro zařazování instancí do tříd. Avšak nástroj Weka není pro tento typ analýzy příliš přehledný.

Shrnutí

```
     forkVA             forkkW                                type         sector       serviceId
 Min.   :0.0000049   Min.   :0.0000175   AutomobileIndustry     :1403   Min.   :1   198346752 :1433
 1st Qu.:0.2513885   1st Qu.:0.2500633   Bank                   :1410   1st Qu.:1   256835671 :1401
 Median :0.5008203   Median :0.5003568   BpoIndustry            :1397   Median :1   286130985 :1454
 Mean   :0.5009826   Mean   :0.4996393   CementIndustry         :1403   Mean   :1   374897109 :1442
 3rd Qu.:0.7508845   3rd Qu.:0.7495545   Farmers1               :1418   3rd Qu.:1   389457902 :1429
 Max.   :0.9999962   Max.   :0.9999699   Farmers2               :1405   Max.   :1   450012212 :1435
                                         FertilizerIndustry     :2876               450023897 :1400
                                         FoodIndustry           :2905               455007891 :1418
                                         Handlooms              :2887               457008451 :1403
                                         HealthCareResources    :1400               486589321 :1442
                                         Hospital               :2906               498710889 :1434
                                         Hostel                 :2857               524100231 :1440
                                         ChemicalIndustry       :2830               548542561 :1430
                                         PoultryIndustry        :2888               562321452 :1405
                                         Residential(Apartments):2884               568730109 :1442
                                         Residential(individual):2867               581000256 :1397
                                         Supermarket            :2874               600124212 :1463
                                         TextileIndustry        :1405               609822556 :1446
                                         Theatre                :2870               652132542 :1439
                                         University             :2896               671004572 :1410
                                                                                    693421673 :1425
                                                                                    775001231 :1403
                                                                                    785200123 :1405
                                                                                    785643218 :1432
                                                                                    785643223 :2906
                                                                                    800145754 :1433
                                                                                    819034567 :1449
                                                                                    894536726 :1428
                                                                                    945678934 :1463
                                                                                    978045321 :1439
                                                                                    5783456902:1435
```

Závěry analýzy

Analýza v tomto dokumentu byla prováděna na datech z reálného prostředí. Konkrétně byla pořízena jako měření spotřeby elektrické energie pro svazový stát jihovýchodní Indie Tamil Nadu. Jednotlivá měření byla prováděna pro různé sektory (zdravotnictví, zemědělství, těžký průmysl apod.). Selským rozumem lze předpokládat, že některé subjekty budou mít vyšší spotřebu elektrické energie. Překvapivě to však nebyl průmysl ani potravinářství, ale zdravotnictví (nemocnice a sanitární zařízení). Podle různých informací z internetu, není zdravotnictví v oblasti na takové úrovni, ale i tak dosáhlo nejvyšší spotřeby. Je nutné podotknout, že potravinářství je téměř na stejné úrovni jako zdravotnictví. Naopak na druhé straně, tedy minimální spotřebu má taktéž překvapivě textilní sektor.

Dalšími tématy analýzy bylo prozkoumání relace mezi spotřebou ve Voltampérech a kilo Wattech. Z výsledku bylo jasně patrné, že využití obou těchto atributů pouze zpřesňuje nepatrně výsledek. Není tedy nutné využívat oba parametry a postačila by spotřeba v kW, která je pro běžného člověka vnímána reálněji.

Následně byl proveden pokus s algoritmem na klasifikaci. Pokus spočíval v ověření, že atribut serviceId není vhodný pro klasifikaci. Kdy velké množství dat (35 %) klasifikuje špatně.

Zdroje

[1] Tamil Nadu. In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2017-05-07]. Dostupné z: https://cs.wikipedia.org/wiki/Tamiln%C3%A1du

[2] Tamil Nadu [online]. [cit. 2017-05-07]. Dostupné z: https://www.google.cz/maps/place/Tamiln%C3%A1du,+Indie/@10.7789549,76.0439992,7z/data=!3m1!4b1!4m5!3m4!1s0x3b00c582b1189633:0x559475cc463361f0!8m2!3d11.1271225!4d78.6568942

Zdroje dat:

http://archive.ics.uci.edu/ml/datasets/Tamilnadu+Electricity+Board+Hourly+Readings