

Řešení – Tenisové zápasy

Vadym Hladyuk – xhlady01@stud.fit.vut.cz

Jakub Svoboda – xsvobo0z@stud.fit.vut.cz

1. Úloha 1 – predikce výsledku pátého setu

1.1 Předzpracování

Námi zvolená data se nacházejí v celkem osmi souborech ve formátu `.csv`. Těchto osm souborů obsahuje data zaznamenaná na čtyřech turnajích (French Open, Australian Open, US Open a Wimbledon), jeden soubor je vždy pro mužskou část turnaje, jeden pro ženskou.

Jelikož je účelem této úlohy predikce výsledku pátého setu zápasu z dat z předchozích setů, byla použita pro predikci pouze mužská část turnaje. Ženské zápasy jsou hrány na tři sety a jejich predikce by tak vyžadovala jiný model.

Po naimportování datové sady byly převedeny jednotlivé sloupce do správných datových typů. Především u hodnot ST2.1 až ST5.1 byl nesprávně určen datový typ (chybějící položky byly v datasetu zapsány jako "NA", což RapidMiner považoval za textový typ) z *polynomial* na *integer*. RapidMiner pak již korektně rozpoznával chybějící hodnoty.

Po naimportování datové sady bylo potřeba ještě přejmenovat sloupce značící počet vyhraných setů. Ten byl v souborech označen nekonzistentně, buď jako "FNL1" nebo "FNL.1". Po přejmenování již bylo možné soubory spojit do jedné datové sady. Pro trénování pak byly údaje sloupce *Result* označeny za *label*.

Odtud byly následně vyfiltrovány položky datasetu z her, které skončili třetím nebo čtvrtým setem. Po této úpravě zbylo v datasetu celkem 86 položek. Celý proces je znázorněn na obrázku 1.

Ne všechny položky byly obsaženy v každém turnaji, některé chybějící bylo potřeba doplnit. Před samotným vstupem do modelu byly tyto chybějící hodnoty doplněny jako průměrná hodnota daného atributu.

Dále byla pro zlepšení predikcí do datasetu přidána informace o rychlosti povrchu kurtu¹. Před samotným vstupem do modelu jsou data normalizována.

1.2 Modelování

Samotným modelům nebyly předány všechny atributy datasetu, konkrétně sloupce obsahující počet vyhraných setů, jména hráčů, kolo v turnaji a počty vyhraných games v pátém setu (FNL1/2, Player1/2, Result, Round, ST5.1/2). Tyto položky byly vyřazeny, protože buď nemají návaznost na výsledek pátého setu nebo se jedná pouze o jiné vyjádření řešení úlohy.

Data byla následně rozdělena v poměru 3:7 pro validaci a trénování. Pro predikci byly vyzkoušeny dva modely: neuronová síť a SVM. Neuronová síť byla trénována s parametrem kroku učení 0,1 v průběhu 200 trénovacích cyklů. Na validačním setu pak dosáhla přesnosti 65,38 %. Algoritmus SVM po trénování s původními parametry schopen predikovat přesností 69,23 % a byl tedy přesnější, než neuronová síť. Toto může být způsobeno několika faktory, především velmi omezenou velikostí datasetu. Podrobné výsledky jednotlivých modelů jsou znázorněny v tabulkách 1 a 2.

1.3 Závěr

V této úloze bylo úkolem predikovat výsledky pátého setu tenisových zápasů podle dat z předchozích setů. Data byla nejprve očištěna a převedena do správného datového typu. Pro zpřesnění predikcí byl dataset doplněn o informace o povrchu kurtu. Pro samotné predikce byly vyzkoušeny dva modely, kde přesnější z nich – SVM dosahoval přesnosti 69,23 %.

Pro lepší přesnost je zřejmě nejvíce limitující velikost datasetu, která je hlavně pro neuronové sítě ne-

¹Data převzata z roku 2017 z <http://www.tennisabstract.com/blog/2019/11/27/the-speed-of-every-surface-2019-edition/>

SVM	True false	True true	Class Prediction
Pred. false	8	5	61,54 %
Pred. true	3	10	76,92 %
Class recall	72,73 %	66,67 %	
Accuracy: 69,23 %			

Tabulka 1. Znázornění výsledků predikce pro SVM.

Neural net	True false	True true	Class Prediction
Pred. false	6	4	60,00 %
Pred. true	5	11	68,75 %
Class recall	54,55 %	73,33 %	
Accuracy: 65,38 %			

Tabulka 2. Znázornění výsledků predikce pro neuronovou síť.

dostatečná. Kromě většího datasetu by pro přesnější modelování bylo vhodné obstarat všechna data odděleně pro každý set zápasu. V současném datasetu jsou totiž některé statistiky (například procentuální podíl prvních podání) počítány pro celý zápas a mohou tak ovlivnit predikce, jelikož model má nepřímo přístup k datům, která byla z části vytvořena během pátého setu.

2. 2. Úloha 2 - shlukování typů podání

2.1 Předzpracování

Použitá datová sada je stejná jako v první úloze viz **Předzpracování**. Dále jsme vybrali hodnoty, které se týkají podání – FSP, FSW, SSP, SSW, ACE a DBF, tedy procentuální úspěšnosti prvního a druhého podání a podíl vyhraných podání a také počet es a dvojchyb. Jelikož se jedná o různorodá data (například procentuální úspěšnost prvních podání a počet es) bylo nutné tato data normalizovat, aby nedocházelo ke zkreslení. Všechny tyto hodnoty máme ke každému zápasu k obou hráčům. Rozhodli jsme se hodnotit každý zápas a podání v něm odděleně abychom měli co nejvíce vzorků podání. Chybějící data byla nahrazeny průměrem.

2.2 Shlukování

Pro shlukování jsme vybrali metodu k-means. Metoda k-means je popsána v článku ². Jelikož u metody k-means zadáváme počet shluků my, vybrali jsme 2 a 4 shluky. Dále jsem z experimentálních důvodů ještě vybral metodu k-means fast, která je rychlejší než klasická metoda k-means. Níže na grafech zhodnotíme výsledky shluků.

Na snímcích **Přehled 2 shluků metodou k-mean fast**. a **Znázornění grafu 2 shluků metodou k-mean fast**. vidíme, při rozdělení na 2 shluky je početnější první shluk, který má vysoké procento prvních podání, nízké

procento vyhraných prvních podání, vysoké procento druhých podání a nižší počet es a dvojchyb. Naopak druhý shluk má naopak nízké procento prvních podání ovšem vysoké procento vyhraných prvních podání. Tudíž když se těmto hráčům povede první podání, je značně razantní, že už to jim zajistí výhru bodu. Z důvodu riskantních prvních podání mají vysoké procento druhých podání, které ovšem také vyhrávají. Z důvodu, že jsou podání v druhém shluku riskantnější je zde více es ale i dvojchyb.

Dále zde máme přehled shlukování metodou k-means na 4 shluky na snímcích **Přehled 4 shluků metodou k-mean fast**. a **Znázornění grafu 4 shluků metodou k-mean fast**.. Při rozdělení nám vzniknou shluky, které bych charakterizoval následovně. První shluk jsou nejhorsí podání. Mají nízkou úspěšnost prvních podání a také výhernost prvního podání je nízká. Dále mají vysoké procento druhých podání, které ovšem také nevyhrávají. Nízký počet es a vyšší počet dvojchyb. Dále zde máme shluk 2, které bych charakterizoval jako riskantní typy podání. Mají nízkou úspěšnost prvních podání, ovšem když se jim povede většinou ho vyhrávají. Dále jsou i při druhém podání agresivnější, protože riskují a vyhrávají druhé podání nejčastěji ovšem na to se váže nejvyšší počet dvojchyb.

Třetí cluster jsou opatrná podání. Mají největší úspěšnost prvních podání ovšem výhernost prvních podání je nižší. Jelikož jsou úspěšní během prvního podání, nedostanou se ke druhému, takže logicky mají nižší počet druhých podání. Eso ani dvojchyby většinou nedělají, mají jich nejméně.

Shluk číslo 4 bych charakterizoval jako nejlepší podání. Mají vysoký podíl prvních podání a nejvyšší počet úspěšných prvních podání. Dále mají mají vysokou výhernost i druhých podání. Počet es mají nejvyšší, což také svědčí o kvalitě podání z tohoto shluku a naopak mají nízký počet dvojchyb, co kvalitu těchto podání potvrzuje.

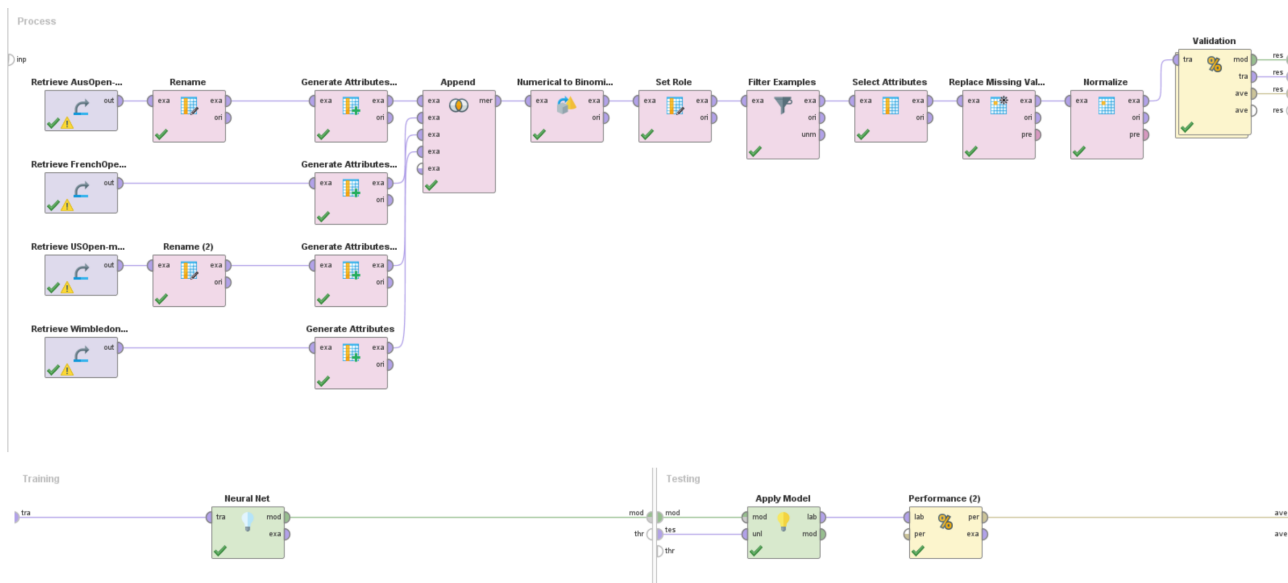
Dále jsme provedl experimentální shlukování zrychlenou metodou k-means fast. Toto byl pouze experiment a jeho výsledky jsou zobrazeny níže. Při dvou shlucích jsou výsledky ještě použitelné, ovšem při čtyřech shlucích z výsledků nejde vyčíst tak jasné archetypy podání, tudíž bych tuto metodu na tyto data nedoporučoval. Metoda k-means se osvědčila více.

2.3 Závěr

Výsledky, které jsme získali byly pomocné pro určení archetypů tenisových podání. Zřetelně jsem mohl poznat například typy výkonů, kteří sází na svoje podání nebo typy výkonů, které přistupují k podání opatrněji nebo také typy výkonů, kteří měli svoje podání téměř perfektní, většinu podání vyhrávají a nedělají chyby.

²<https://projecteuclid.org/euclid.bsm/1200512992>

Výsledky měření se dají použít pro přípravu na oponenty, kdy hráči budou vědět jaký typ podání mohou v zápase očekávat. Spíše se toto dá použít na nižší příčky žebříčku ATP, protože na grand-slamových turnajích, je tak úzký výběr hráčů, že se znají navzájem a zkoumat jejich podání pro ně, je zbytečné.



Obrázek 1. Znáznornění prediktivního modelu.

Number of Clusters: 2
Distance Measure: Squared Euclidean Distance
Average Cluster Distance: 4.315
Davies-Bouldin Index: 1.437

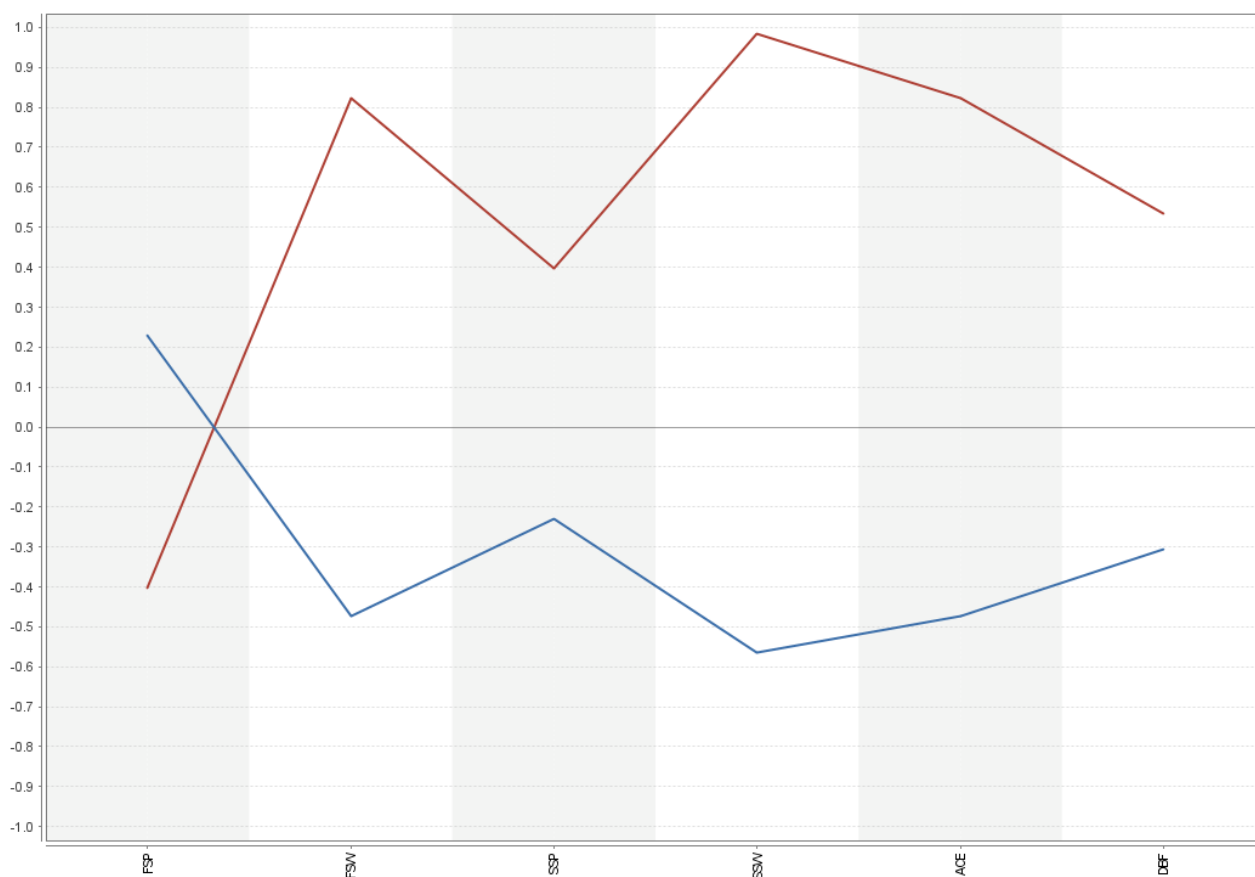
Cluster 0 1,198 Average Distance: 3.673

ACE is on average **47.69%** smaller, **SSW** is on average **30.99%** smaller, **DBF** is on average **22.39%** smaller

Cluster 1 688 Average Distance: 5.434

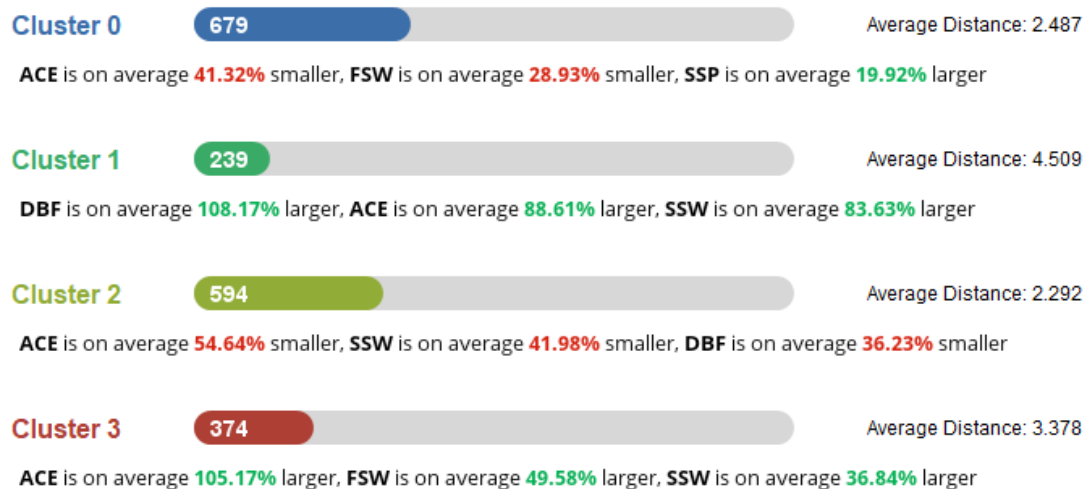
ACE is on average **83.04%** larger, **SSW** is on average **53.96%** larger, **DBF** is on average **38.99%** larger

Obrázek 2. Přehled 2 shluků metodou k-mean.

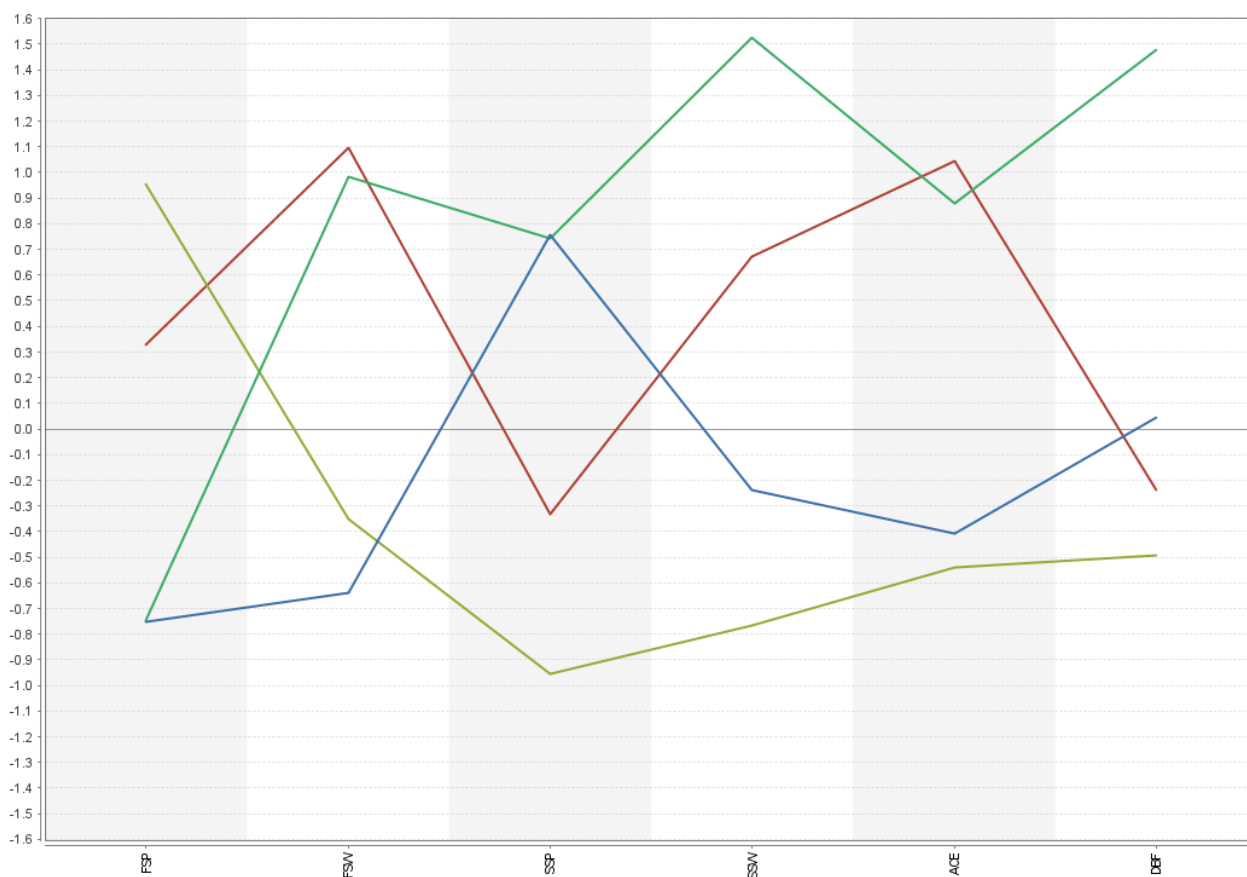


Obrázek 3. Znázornění grafu 2 shluků metodou k-mean.

Number of Clusters: 4
Distance Measure: Squared Euclidean Distance
Average Cluster Distance: 2.859
Davies-Bouldin Index: 1.311



Obrázek 4. Přehled 4 shluků metodou k-mean.



Obrázek 5. Znázornění grafu 4 shluků metodou k-mean.

Number of Clusters: 2
Distance Measure: Euclidean Distance
Average Cluster Distance: 4.314
Davies-Bouldin Index: 1.447

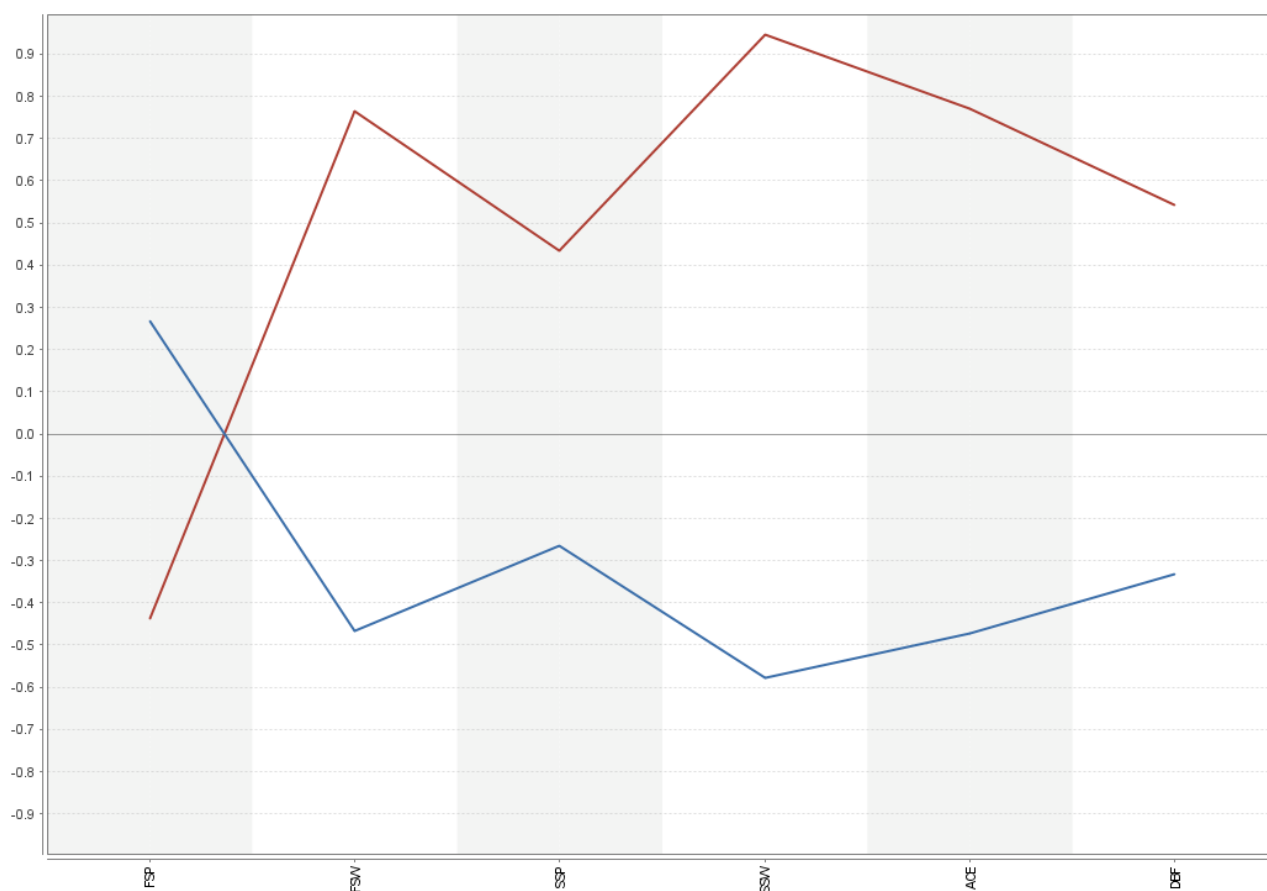
Cluster 0 1,171 Average Distance: 3.572

ACE is on average **47.56%** smaller, **SSW** is on average **31.68%** smaller, **DBF** is on average **24.29%** smaller

Cluster 1 715 Average Distance: 5.531

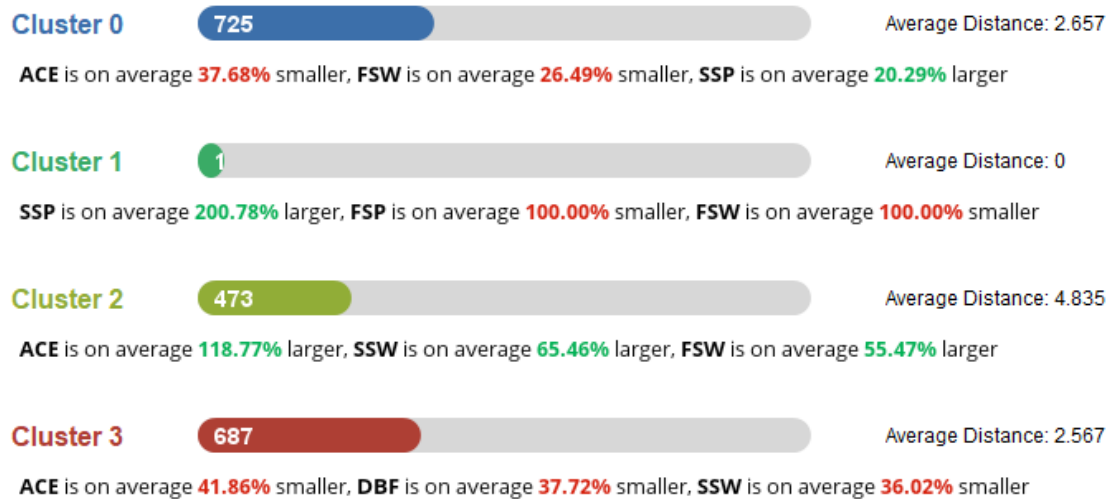
ACE is on average **77.90%** larger, **SSW** is on average **51.89%** larger, **DBF** is on average **39.78%** larger

Obrázek 6. Přehled 2 shluků metodou k-mean fast.

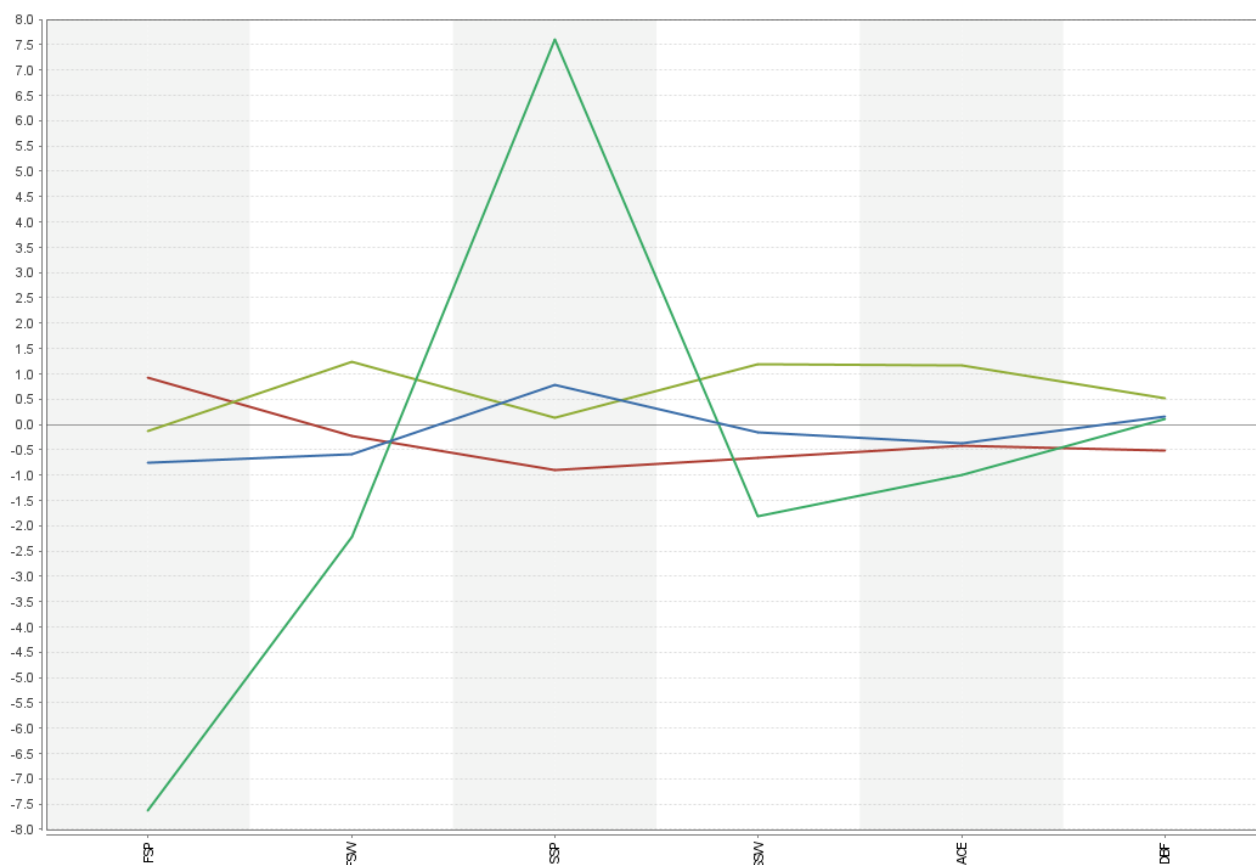


Obrázek 7. Znáznornění grafu 2 shluků metodou k-mean fast.

Number of Clusters: 4
Distance Measure: Euclidean Distance
Average Cluster Distance: 3.169
Davies-Bouldin Index: 0.951



Obrázek 8. Přehled 4 shluků metodou k-mean fast.



Obrázek 9. Znáznornění grafu 4 shluků metodou k-mean fast.