ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

FAKULTA BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

Katedra biomedicínské techniky

Komparace metod k-means a k-means++ při použití na reálném EEG záznamu

Comparison of the k-means and k-means++ methods used on a real EEG record

Bakalářská práce

Studijní program: Biomedicínská a klinická technika

Studijní obor: Biomedicínský technik

Autor bakalářské práce Daniela Kolíková

Vedoucí bakalářské práce: Ing. Jan Štrobl

Kladno 2018

České vysoké učení technické v Praze, Fakulta biomedicínského inženýrství

Katedra biomedicínské techniky | Akademický rok: 2017/2018

Zadání bakalářské práce

Student:	**Daniela Kolíková**
Obor:	Biomedicínský technik
Téma:	**Komparace metod k-means a k-means++ při použití na reálném EEG záznamu**
Téma anglicky:	Comparison of the k-means and k-means++ methods used on a real EEG record

Zásady pro vypracování:

Navrhněte a implementujte metodu k-means++ pro zpracování EEG záznamu. K realizaci metody využijte programovací prostředí MATLAB. Na simulovaných datech vytvořených pro porovnání metod k-means a k-means++ statisticky vyhodnoťte výhody metody k-means++ oproti metodě k-means. Prostřednictvím programu Wave-Finder otestujte výhody metody k-means++ v porovnání s metodou k-means u reálných EEG dat. Pro testování použijte ROC analýzu.

Seznam odborné literatury:

[1] Krajča V., Mohylová J., Číslicové zpracování neurofyziologických signálů, ed. Fakulta biomedicínského inženýrství, ČVUT Praha, 2011, ISBN 978-80-01-04721-7
[2] David Arthur, Sergei Vassilvitskii, k-means++: The Advantages of Careful Seeding, Society for Industrial and Applied Mathematics, ročník 18, číslo 1, 2007
[3] Vinay K. Ingle, John G. Proakis, Digital signal processing using MATLAB, ed. Third edition, CENGAGE Learning, 2012, ISBN 978-1-111-42737-5

Zadání platné do:	20.09.2019
Vedoucí:	Ing. Jan Štrobl
Konzultant:	Ing. Marek Piorecký

..... vedoucí katedry / pracoviště

..... děkan

V Kladně dne 19.02.2018

PROHLÁŠENÍ

Prohlašuji, že jsem Bakalářskou práci s názvem Komparace metod k-means a k-means++ při použití na reálném EEG záznamu samostatně a použila k tomu úplný výčet citací použitých pramenů, které uvádím v seznamu přiloženém k bakalářské práci.

Nemám závažný důvod proti užití tohoto školního díla ve smyslu §60 Zákona č.121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon).

V Kladně dne …… …..……..……...

PODĚKOVÁNÍ

Mé poděkování patří Ing. Janu Štroblovi za odborné vedení, trpělivost a ochotu, kterou mi v průběhu zpracování bakalářské práce věnoval.

ABSTRAKT

Komparace metod k-means a k-means++ při použití na reálném EEG záznamu

Tato práce je zaměřena na oblast elektroencefalografie neboli EEG. Měření EEG spočívá ve snímání elektrické aktivity mozku z povrchu hlavy. Naměřený EEG signál má stochastickou povahu a jeho vyhodnocování je obtížné a jsou proto využívány metody automatické klasifikace. Jednou hojně využívanou metodou v automatické klasifikaci EEG signálu je metoda k-means. Tato metoda v případě složitějších úloh může chybovat. Chybovost může být dána například náhodnou inicializací center shluků dat, a proto byla vytvořena metoda k-means++, který upravuje inicializaci center shluků. Metoda k-means++ vychází z původní metody k-means, tedy rozšiřuje původní verzi metody.

Cílem bakalářské práce je aplikovat metodu k-means++ na reálný EEG záznam a porovnat metody k-means a k-means++. Pro ověření správnosti metod a provedení statistického zhodnocení výhod, případně nevýhod metod, byla sestavena simulovaná data. Simulovaná data i kód inicializace dat metody k-means++ byl vytvořen v programovacím prostředí MATLAB. Po statistickém vyhodnocení výstupů metod, metoda k-means++ potvrdila menší chybovost a výstupy klasifikace simulovaných dat byly příznivější, než u základní metody k-means. Po tomto zhodnocení byla metoda k-means++ aplikována na reálná data, aby se porovnalo využití metod k-means a k-means++ na automatickou klasifikaci EEG záznamů. Klasifikace reálných EEG dat byla zobrazena v programu Wave-Finder a vyhodnocena expertem. Reálné záznamy byly naměřeny v Nemocnici Na Bulovce na pěti pacientech. Měření bylo schváleno etickou komisí Nemocnice Na Bulovce v roce 2011 a záznamy byly dále anonymizovány.

Automatická klasifikace EEG záznamů byla u obou metod statisticky vyhodnocena a bylo provedeno zhodnocení metod. Dle statistické analýzy bylo zjištěno, že metoda k-means v EEG záznamu nejlépe detekuje EMG artefakty, dobře detekuje fyziologickou aktivitu a epileptickou aktivitu. Nicméně hůře detekuje pomalé oční artefakty. Metoda k-means++ v EEG záznamu detekuje nejlépe fyziologickou aktivitu. Epileptická aktivita je lépe detekována metodou k-means, nežli metodou k-means++. Metoda k-means++ epileptickou aktivitu detekovala téměř o 50 % hůře. Mnohem lépe však metoda k-means++ detekovala pomalé oční artefakty, kde je zlepšení oproti metodě k-means výrazné.

Klíčová slova

EEG, shluková analýza, k-means, k-means++

ABSTRACT

Comparison of the k-means and k-means++ methods used on a real EEG record

This work is focused on electroencephalography or EEG. EEG measurement consists of sensing electrical activity of the brain from the surface of the head. The measured EEG signal has a stochastic nature and its evaluation is difficult and therefore the methods of automatic classification are used. One abundantly used method in the automatic classification of the EEG signal is k-means. This method may be mistaken for more complex tasks. Errors can be caused, for example, by the random initialization of data clustering centres, and an algorithm for k-means++ has been created to modify the initialization of clustering centres. The method k means ++ is based on the original k-means method, which extends the original version of the method.

The aim of the bachelor thesis is to apply the k-means++ method to a real EEG record and compare the k-means and k-means++ methods. To verify the correctness of the algorithms and to perform a statistical evaluation of the advantages, or the disadvantages of the algorithms, simulated data was compiled. The simulated data and the k-means ++ algorithm were created in the MATLAB programming environment. After statistical evaluation of algorithm outputs, the k-mean ++ method confirmed a smaller error rate and the outputs of the simulated data classification were more favourable than the k-means method. After this evaluation, the method k-means++ was applied to real EEG data to compare the use of the k-means and k-means++ algorithms to automatically classify EEG records. Classification of real EEG data was displayed in the Wave-Finder program and evaluated by an expert. Real records were measured at the Hospital Na Bulovce on five patients. The measurement was approved by the Hospital Na Bulovce Ethics Commission in 2011 and the records were further anonymized.

The automatic classification of EEG records was statistically evaluated for both methods and methods were evaluated. According to statistical analysis, it was found that the method k-means in the EEG record best detects EMG artefacts, well detects physiological activity and epileptic activity. However, it is less likely to detect slow eye artefacts. The method k-means++ in the EEG record best detects physiological activity. Epileptic activity is better detected by k-means than k-means ++. The k-means++ method epileptic activity detected almost 50% worse. Much better the method k-means++ detected slow eye artefacts, where the improvement over the k-means method is significant.

Keywords

EEG, cluster analysis, k-means, k-means++

1. Obsah

Seznam symbolů a zkratek ... **9**

1 Úvod ... **10**

1.1. Cíle práce ... 10

2 Encefalografie a EEG signál ... **11**

2.1 Definice ... 11

2.1.1 Mozková aktivita ... 12

2.2 Problematika měření signálu EEG ... 13

2.2.1 Zpracování signálu ... 14

2.3 Adaptivní segmentace ... 15

2.4 Extrakce příznaků ... 15

2.5 Metody umělé inteligence ... 16

2.6 Shluková analýza ... 17

2.6.1 Metoda k-means ... 18

2.6.2 Metoda k-means++ ... 20

3 Metodika ... **21**

3.1 Programovací prostředí ... 21

3.2 Data ... 21

3.2.1 Simulovaná data ... 21

3.2.2 Reálná data ... 22

3.3 K-means ... 23

3.4 K-means++ ... 23

3.5 Analýza dat ... 24

3.6 Kvalitativní analýza ... 25

3.6.1 Specificita ... 26

3.6.2 Senzitivita ... 26

3.6.3 Pozitivní prediktivní hodnota ... 26

4 Výsledky ... **27**

4.1 Klasifikace simulovaných dat ... 27

4.2 Klasifikace reálných EEG dat ... 40

4.2.1 Klasifikace EEG záznamu pomocí metody k-means ... 40

4.2.2 Klasifikace EEG záznamu pomocí metody k-means++ 41
5 Diskuze 43
6 Závěr 47
Seznam použité literatury 48
A Seznam příloh na CD 51

Seznam symbolů a zkratek

Seznam zkratek

Zkratka	Význam
EEG	Elektroencefalografie
2D	Dvoudimenzionální
EKG	Elektrokardiografie
EOG	Elektrookulografie
TN	True Negative
TP	True Positive
FN	False Negative
FP	False Positive
PPV	Pozitivní Prediktivní Hodnota
SE	Senzitivita
SP	Specificita

1 Úvod

Tato práce je zaměřená na oblast elektroencefalografie, zkráceně EEG, konkrétně na automatickou klasifikaci EEG signálu. V první části práce budou popsány základní pojmy vyskytující se v oblasti EEG a problematika měření EEG záznamu. EEG se obvykle měří ambulantně, kdy se pacientovi provede základní vyšetření, které trvá přibližně půl hodiny. V praxi je také žádáno dlouhodobé měření EEG. Dlouhodobé měření signálu EEG se zpravidla měří u pacientů s epilepsií, nebo jinými poruchami mozkové aktivity. Dlouhodobé monitorování mozkové aktivity trvá od několika hodin až po týden, proto je nutno takto dlouhý signál klasifikovat algoritmy umělé inteligence. Algoritmy umělé inteligence využívají určitých příznaků, které expert označí za důležité. Tyto příznaky jsou v praxi analyzovány v programu Wave-Finder, který v některých oblastech mimo jiné využívá funkci k-means. Analýza příznaků je prováděna i jinými programy určených pro analýzu signálu EEG, v mé práci však bude použit program Wave-Finder.

Ve druhé části práce budou porovnávány metody k-means a k-means++ na simulovaných datech. Dalším bodem druhé části práce bude popsání sestrojení simulovaných dat a naprogramování kódu kmeans++ v prostředí MATLAB. Při aplikaci metod k-means a k-means++ na simulovaná data se provede statistické zhodnocení pomocí ROC analýzy. Díky statistickému zpracování a změření časové náročnosti se zjistí, zda je možné mnou naprogramovanou metodu při pozitivních výsledcích této analýzy aplikovat na reálný EEG záznam. Dále se metoda k-means++ aplikuje na EEG záznam, statisticky se vyhodnotí a porovná s výsledky metody k-means.

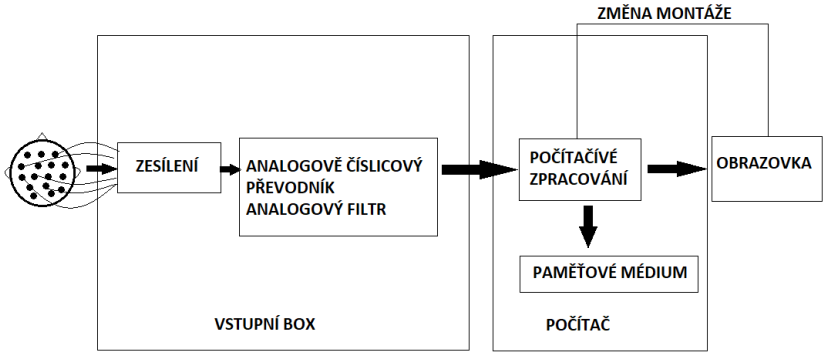
1.1. Cíle práce

Cílem práce je otestovat uplatnění metody k-means++ v klasifikaci EEG dat. Pro správné vyhodnocení výsledků je třeba porovnat klasifikaci dat do shluků pomocí metody k-means a metody k-means++. Jelikož je reálný EEG záznam obtížně klasifikovatelný, nejprve statisticky vyhodnotím výstupy metod pomocí sestavených simulovaných dat. Simulovaná data jsem sestavila tak, aby byla otestována správnost klasifikace metod k-means a k-means++. Po statistickém vyhodnocení výstupů metod simulovaných dat, aplikuji metodu k-means++ na reálný záznam EEG v programu Wave-Finder. Po klasifikaci reálných dat v programu Wave-Finder provedu statistické vyhodnocení metod k-means a k-means++. Na základě statistického vyhodnocení posoudím výhody, případně nevýhody metody k-means++ oproti meodě k-means.

2 Encefalografie a EEG signál

2.1 Definice

EEG měří elektrickou aktivitu mozku na různých místech hlavy, typicky pomocí elektrod umístěných na pokožce. Jeho hlavní výhodou, oproti ostatní záznamové technice, je vysoké časové rozlišení a skutečnost, že měření může být zaznamenáno neinvazivně (tj. bez nutnosti chirurgického zákroku). EEG nemá konstantní průběh v čase. Znamená to, že v určitém časovém úseku nenabývá konstantních hodnot jdoucích pravidelně po sobě. Jedná se o stochastický děj, což znamená, že EEG signál je sada dat, která jsou řazena náhodně po sobě v daném časovém úseku. K měření elektrické mozkové aktivity se využívá elektroencefalograf, skládající se ze snímacích elektrod, které zaznamenávají rozdíl elektrického potenciálu mezi oblastmi lebky, dále se skládá z tzv. head boxu (vstupní box), ve kterém je uložen zesilovač a A/D převodník (obrázek 2.1). Neodmyslitelnou součástí přístroje je displej, který slouží k zobrazení EEG signálu a počítač, kde probíhá další zpracování signálu. [1] [2] [3]



Obrázek 2.1: Blokové schéma přístroje pro měření EEG. Převzato z [2] a upraveno.

U měření EEG lze měřit bipolární, nebo unipolární záznam. Porovnáváním potenciálů dvou bodů na pokožce hlavy se získá bipolární záznam, jehož výhodou je eliminace amplitudových artefaktů [2]. Měřením rozdílu elektrického potenciálu mezi aktivním bodem mozkové tkáně (přímo pod aktivní elektrodou) proti bodu s nulovým potenciálem (referenční elektrody, které jsou typicky umístněný na ušních boltcích) se získá unipolární záznam. Elektrické potenciály naměřené elektrodou jsou řádově v desítkách mikrovoltů, tedy před dalším zpracováním je třeba signál zesílit. Zesílení je realizováno pomocí diferenčního zesilovače řádově sto tisíc až milion krát. Výstupní EEG signál je zobrazen na displeji monitoru, nebo vytištěn pomocí termotisku na papíře. [4]

Již v první zprávě o lidských EEG nahrávkách Hans Berger již poznamenal přítomnost různých mozkových laloků. Objevil zejména rytmickou aktivitu, která byla nejvýraznější v okcipitálních elektrodách, při zavřených očích. Aktivita těchto kmitů, které nazval alfa vlny, se prudce snížila, pokud měřená osoba otevřela oči. Tento průběh je dnes označen jako blokování alfa a používá se jako nejjednodušší demonstrace odrážení mozkových procesů pomocí EEG. [3]

Berger také popsal vlny beta, které kmitají na vyšších frekvencích než vlny alfa. Objevily se při otevřených očích měřené osoby a do jisté míry i při zavřených očích, za podmínky, že měřená osoba prováděla mentální výpočet. [3]

Na základě Bergerově práce, která byla průlomová v oblasti EEG, byly nalezeny další kmity v mozku – delta a theta. [3] [5]

2.1.1 Mozková aktivita

EEG signál nabývá obvykle amplitudy 1 až 100 μV a pohybuje se ve frekvencích od 0,5 až 30 Hz, někdy až 60 Hz. EEG signál se skládá ze čtyř základních rytmů, jak je popsáno výše. Jedná se o vlny alfa, beta, a později objevené vlny delta a theta.

Vlna, která byla historicky popsána jako první, alfa, nabývá frekvence od 8 do 13 Hz s přibližně sinusovým průběhem. Nabývá obvykle amplitudy do 50 μV. Tato vlna je považována za hlavního představitele mozkové aktivity pro dospělého člověka. Maximální aktivita alfa vlny je v bdělém stavu, při fyzické relaxaci a zavřených očích, utlumení aktivity se provede pomocí otevření očí (blokace alfa). Tato vlna je charakteristická pro stadium těsně před usnutím. Výskyt aktivity alfa je nejvýraznější v okcipitálních oblastech hlavy. [1] [3] [5]

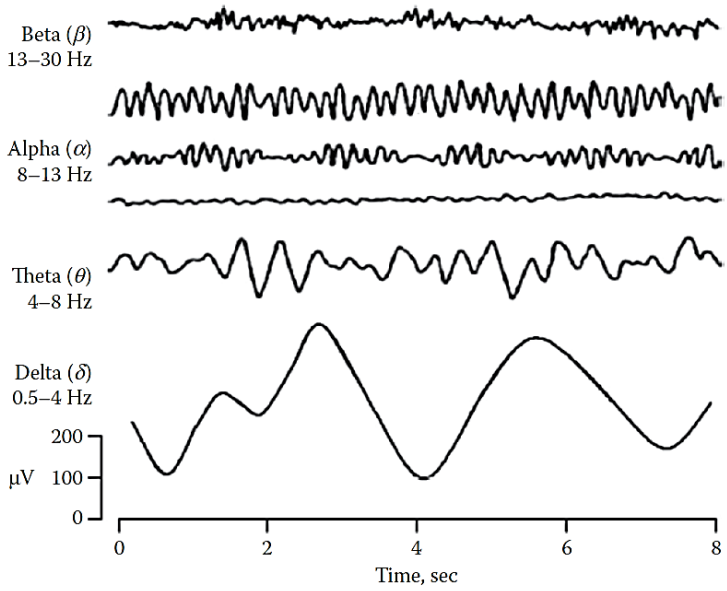
Vlna beta nabývá frekvence od 14 do 26 Hz a amplitudy, která je oproti vlně alfa nižší (do 30 μV). Spolu s vlnou alfa je představitelkou aktivity lidské mozkové tkáně. Při experimentálním měření se aktivita vlny beta zvýší při mentálních výpočtech. Nejlépe se aktivita vlny beta snímá z frontálních oblastí. [2] [3]

Vlna delta nabývá frekvence od 0,5 do 4 Hz a amplitudy do 100 μV. Tato vlna má zvýšenou aktivitu při hlubokém spánku nebo při závažných onemocněních mozku. Nejlépe aktivitu této vlny lze zaznamenat v parietálním laloku. Často je tato vlna zaměňována s pohybovými artefakty. [5]

Další vlna theta nabývá frekvence od 4 do 7 Hz a amplitudu do 100 μV. Aktivita této vlny je fyziologicky přítomna během spánku a stoupá ve stresových situacích. Její aktivita je nejlépe identifikována v temporální oblasti. V současné době se snímání této vlny využívá pro studii a analýzu stresu [5]. [6]

Gamma vlna má frekvenční rozsah nad 30 Hz. Vyznačuje se velmi nízkou amplitudu a vyskytuje se nepravidelně. Nález gamma vlny se vyskytuje u zvláštních případů

a onemocnění. Aktivita vlny gamma je nejlépe detekovatelná ve frontálně – středové oblasti mozku. [5]



Obrázek 2.2: Průběhy vln alfa (Alpha), beta (Beta), theta (Theta), delta (Delta) v čase (Time) v sekundách (sec). Převzato z [5]

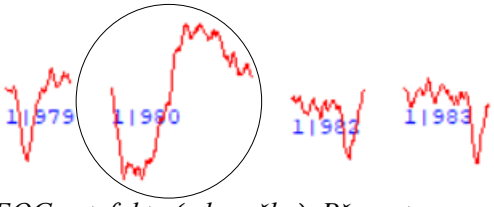
2.2 Problematika měření signálu EEG

Naměřené signály, které nemají svůj původ v mozku, se nazývají artefakty. Jelikož EEG signál se pohybuje v řádech mikrovoltů (nízká amplituda je způsobená především z důvodu přítomnosti lebky, tedy signál je utlumen kostní tkání), jsou artefakty přítomné téměř vždy [7]. Tyto artefakty je nutno ze signálu odstranit, případně je rozeznat, aby nedošlo k chybné diagnostické interpretaci. Artefakty lze dle povahy rozdělit na technické a biologické.

Mezi artefakty technické povahy se řadí elektrostatické potenciály, síťový brum, impulsní rušení a šum elektronických obvodů. Tyto artefakty pochází z vnějšího prostředí. Elektrostatické potenciály jsou způsobeny nízkou jakostí elektrod, nebo špatným kontaktem mezi elektrodou a kůží, případně vysycháním vodivého gelu. Síťovým brumem se rozumí rušení napětím síťového kmitočtu. Síťový brum se odstraňuje pomocí filtru typu pásmová zádrž (síťový brum je odstraňován ze všech diagnostických přístrojů, které snímají malé potenciály pomocí pásmové zádrže). Impulsní rušení způsobuje blízkost motorků, zapínání přístrojů ze stejné energetické sítě, nebo přepínání svodů. Šum elektronických obvodů vzniká na vstupních obvodech zesilovače. [1] [7]

Artefakty biologické povahy vznikají v těle měřeného subjektu. Patří sem například EOG a EKG artefakt. EOG artefakt je zapříčiněn rozdílem potenciálů mezi rohovkou

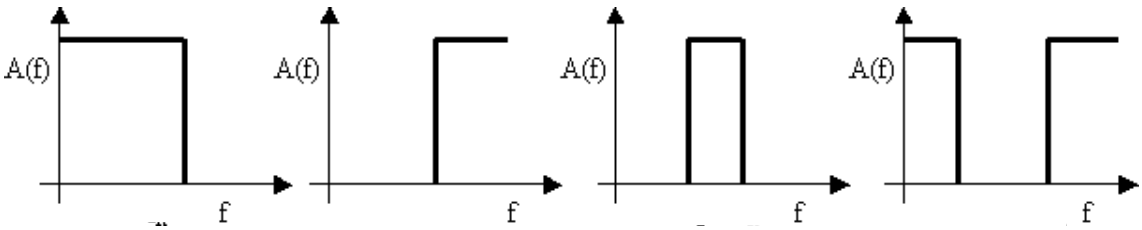
a oční bulvou. Tento rozdíl potenciálů je v porovnání s mozkovými potenciály velký. Mezi EOG artefakt se řadí i mrkání. EKG artefakt se vyskytuje v EEG signálu běžně a může být zaměněn za epileptický projev. Z tohoto důvodu jsou modernější EEG diagnostické přístroje rozšířeny o jeden kanál určený pro měření EKG křivky. [7] [1]



Obrázek 2.3: Příklad EOG artefaktu (v kroužku). Převzato z programu Wave-Finder.

2.2.1 Zpracování signálu

Artefakty jsou nežádoucími příznaky v EEG signálu. Typické EEG signály měřené elektrodami při klinickém nahrávání se pohybují frekvenčně do 30 Hz. Veškerá aktivita typických EEG vln (alfa, beta, delta, theta) patří právě do této oblasti (0-30 Hz). Ve výzkumu se však pracuje i mnohem vyššími frekvencemi, například s vlnou gamma (40-100 Hz) [9]. Frekvenční složka, která se vyskytuje nad tímto rozsahem (EMG artefakt způsobený svalovou aktivitou, jehož frekvence je vyšší než 30 Hz) je odstraněna pomocí nízkofrekvenčních filtrů, pouze v případě, kdy odborník chce měřit pouze vlny alfa, beta, delta a theta. Vlnu gamma tímto filtrem odstraní. Mezi základní digitální filtry patří konvenční filtry jako je dolní propust, horní propust, pásmová propust a pásmová zádrž. Toto rozdělení je dle impulzní charakteristiky. V případě, že je nutno odstranit součást signálu o nižší frekvenci (tato frekvence nabývá hodnoty do 0,1 Hz), jsou využívány vysokofrekvenční filtry. V případě odstranění síťového brumu (50 Hz) se využívá pásmová zádrž. [8] [2] [5]



Obrázek 2.4: Přenosová charakteristika ideálních typů filtrů, zleva – Dolní propust, horní propust, pásmová propust, pásmová zádrž. Na ose y A(f) značí zesílení v dB a osa x popisuje frekvence (f) v Hz. Převzato z [1].

Dle impulzní charakteristiky filtru rozlišujeme filtry FIR a IIR. Základními vlastnostmi FIR filtrů je konečná impulsní charakteristika a stabilita. Filtry FIR mají při správném návrhu lineární fázovou charakteristiku a jsou náročnější na čas zpracování signálu. IIR filtry mají nekonečnou dobu odezvy. Mají nelineární fázovou charakteristiku, což značí, že ovlivňují fázi a při nesprávném návrhu mohou být nestabilní. [2]

2.3 Adaptivní segmentace

Aby se mohl signál analyzovat pomocí dalších algoritmů, v případě dlouhého záznamu je nutno signál rozdělit na kratší časové úseky neboli provést segmentaci záznamu.

EEG záznam se časem mění, tedy nemá stacionární charakter. Frekvenční i amplitudové vlastnosti se v časovém úseku mění. V signálu se mohou vyskytovat artefakty, nebo vlny, které jsou charakteristické pro různé mozkové nemoci, například epilepsie. V případě, že by se dlouhodobé záznamy EEG rozdělily pro účely extrakce příznaků do konstantní délky, hranice nemají žádnou informaci o tom, jak mají vypadat. Nejsou závislé na povaze signálu, tedy je možné, že se hranice utvoří v polovině artefaktu. Mohou také vzniknout hybridní úseky, které obsahují směs vln různého tvaru a frekvencí.

Záznam EEG signálu je v mnoha případech velmi dlouhý a je potřeba správně identifikovat jednotlivé grafoelementy, je třeba jej rozdělit na části, které budou analyzovány. Pojmem grafoelementy rozumíme charakteristické aktivity vyskytující se v záznamu (např. epileptická aktivita, fyziologická aktivita, průběh EOG artefaktu, …). Tuto problematiku řeší metoda segmentce signálu, poprvé navržena Bedensteinem a Praetoriusem v roce 1977. Metoda využívá pořadí používající pro vyhledání nestacionarit dvou oddělených oken, klouzajících po signálu. Nevýhodou je, že metoda může být aplikována pouze v jednom kanálu jednoho průběhu algoritmu [2]. Metody pro adaptivní segmentaci signálu vyžadují detekci změny stacionarity a odhad přesného okamžiku změny. Předpokládají se skokové změny stacionarity signálu.

Segmentace se dělí na konstantní a adaptivní segmentaci. V případě konstantní segmentace je signál dělen na fixní úseky, kdy každý tento úsek má stejný počet vzorků. Nevýhodou této segmentace je, že mnohdy se rozdělí vlna nebo artefakt v polovině. Adaptivní segmentace produkuje různé délky segmentů, kdy je signál rozdělen na části, které mají průběh záznamu s podobným charakterem. U adaptivní segmentace nedochází k přerušení signálu, který patří do jedné části. [10] V adaptiní segmentaci se využívá psegmentace podle tzv. oken. Tato okna jsou buď spojená nebo oddělená a kontrolují průběh signálu, aby nedošlo k ukončení segmentu v polovině vlny, nebo hrotu. Princip a druhy adaptivní sementace jsou popsány docentem Krajčou v knižní publikaci [2].

2.4 Extrakce příznaků

Podstatou aplikace všech metod rozpoznávání příznaků je správné vyjmutí příznaků, které popisují vlastnosti objektů, které jsou tříděné. Kvalita každé metody automatické

klasifikace závisí na kvalitě použitých příznaků. Z tohoto důvodu je třeba věnovat pečlivost vyjmutí (extrakci) příznaků. [2]

Příznak charakterizuje určitou vlastnost záznamu v určité oblasti. Je vyjádřen číselnou hodnotou. Příznakem lze označit maximální amplitudu daného segmentu, nebo například různé spektrální charakteristiky. [9]

Výběr příznaků ze signálu je jedna z nejdůležitějších záležitostí automatické klasifikace. Pro automatickou klasifikaci EEG, která se má blížit vizuálnímu hodnocení experta, je lepší zvolit příznaky popisující nejen spektrální ale i grafické vlastnosti 0klasifikace nekorektní. Extrakce příznaků se v praxi provádí pomocí speciálních programů, jako je například Wave-Finder. [2] [9]

2.5 Metody umělé inteligence

Umělá inteligence v oblasti zpracování EEG, je metodika, která se zabývá vývojem programů a algoritmů, které řeší požadované problémy místo experta. Aby algoritmus, nebo program mohl být zařazen mezi metody umělé inteligence, musí dokázat uložit znalosti, pomocí těchto uložených znalostí řešit problém a během experimentu získat nové znalosti. Algoritmy umělé inteligence se podle učení dělí na algoritmy s učitelem a algoritmy bez učitele. [1] [10] [1]

U algoritmů učení s učitelem jsou známy požadované výsledky. Algoritmus tedy za podmínky známých výsledků přizpůsobuje váhy a prahy tak, aby se výstup co nejvíce blížil k požadovaným hodnotám. Pokud se vstupní a požadovaná výstupní data shodují, jedná se o autoasociativní učení. Jestliže jsou tato data odlišná, jedná se o heteroasociativní učení. Tyto algoritmy mohou probíhat offline i online. Při offline učení se algoritmus zpustí až tehdy, kdy jsou do sítě přivedeny všechny data. Počítá se aktuální velikost gradientu chyby pro celou množinu dat. Výhodou je téměř úplná konvergence. V případě online učení se přizpůsobují parametry při každém průchodu dat. Algoritmus tedy nečeká, na úplnou sadu dat, ale parametry upřesňuje při průchodu každého vzorku. Tento proces však konverguje k lokálnímu minimu, nikoli k absolutnímu minimu. Online učení konverguje rychleji než offline. [10] [2] [1]

Učení bez učitele nemá představu o požadovaných výstupních datech. Na základě vzoru stejných nebo blízkých vlastností třídí přicházející vektory do tříd. Výstup z algoritmů bez učitele není porovnáván s požadovaným výstupem. Mezi algoritmy bez učitele patří shluková analýza. Řazení do tříd probíhá na základě zjištění vzdálenosti (nebo jiné vlastnosti jako je například hustota) mezi daty a shluky. Díky vzdálenosti lze zjistit míru podobnosti dat. [10]

Podobnost, je hledání podobných vlastností vektorů. Vytváří se shluky všech vstupních vektorů, které mají společné nebo blízké vlastnosti.

Shlukovacích metod je využíváno mnoho. Při výběru správné shlukovací metody záleží především na typu dat. Na základě povahy dat, které je potřeba roztřídit do tříd se vybere konkrétní metoda. V této práci je využito shlukovacích metod založených na vzdálenosti, proto jsou zde uvedeny především tyto metody.

Pro zařazení do nějakého shluku je třeba znát vzdálenost mezi vzorem a shlukem. Na základě měření této vzdálenosti lze zjistit míru podobnosti. Mezi nejvyužívanější vzdálenosti pro rozřazování dat do shluků patří: Hammingova vzdálenost, Euklidova vzdálenost a Čtvercová vzdálenost. [2] [13]

Hammingova vzdálenost hledá rozdíly mezi jednotlivými daty, celková vzdálenost je součet absolutních hodnot těchto rozdílů. Euklidova vzdálenost je nejpoužívanější metrikou [2]. Euklidova metrika se orientuje v kartézském souřadném systému, ve kterých se vektory dat pohybují. Výpočet vzdálenosti probíhá dle rovnice:

$$E = \sqrt{\sum_{i=1}^{N}(A(i) - B(i))^2}, \tag{1}$$

kde E označuje Euklidovu vzdálenost A a B jsou vektory dat, N určuje dimenzi, i označuje konkrétní objekt vektoru.

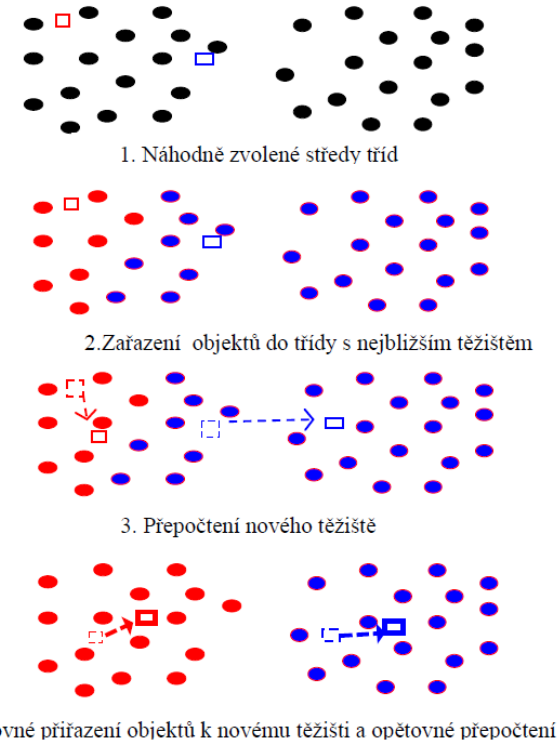
Čtvercová vzdálenost je zjednodušena Euklidova vzdálenost, pouze se dosadí za míru vzdálenosti největší rozdíl mezi jednotlivými elementy vektorů. [10] [8] [2]

2.6 Shluková analýza

Tato metoda je příznaková metoda učení bez učitele. K identifikace zkoumaných dat využívá podobnosti, resp. vzdálenosti mezi daty nebo hustotou dat. Data jsou popsána *n*-rozměrnými příznaky. Úkolem shlukové analýzy je rozdělení zkoumaných dat do stejnorodých tříd. Nevýhodou shlukové analýzy je, že neumožňuje online klasifikaci. Nedokáže objekty shlukovat, pokud není set dat kompletní. Metod, které se řadí do shlukové analýzy je celá řada a klasifikují se dle různých kritérií, například se klasifikují se matematického aparátu, kam se řadí metody deterministické, statistické nebo fuzzy. Další klasifikace je z hlediska zpracování dat, dle sdílení členství v různých shlucích a dle typu shlukového kritéria. Obecně shlukovací metody lze rozdělit na hierarchistické metody a nehierarchické metody. Hierarchické metody transformují matici vzdálenosti do posloupnosti hierarchicky seřazených rozdělení. Hierarchické algoritmy jsou náročné na paměť, je totiž nutné uchovávat v paměti matici vzdálenosti po dobu výpočtu. Nehierarchické metody hledají opakovaně optimální rozdělení dat, které minimalizují určitou kriteriální funkcí (funkce, dle které srovnáváme více dat). Mezi tyto metody patří metoda k-means. [1] [11]

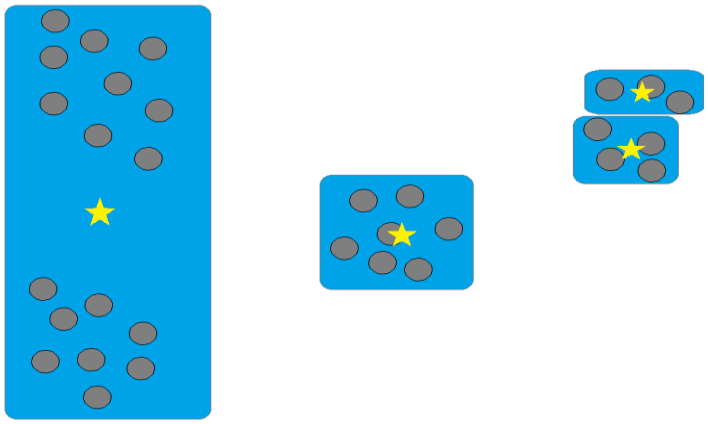
2.6.1 Metoda k-means

Principem metody k-means je počáteční rozdělení naměřených dat do shluků (počet shluků si uživatel navolí). Z těchto shluků se označí prvních *k* objektů jako těžiště konkrétních shluků s jedním členem, které se označí jako střed shluku. Dále se vypočítají vzdálenosti všech objektů od každého středu. V této metodě se zpravidla využívá Euklidovská vzdálenost. Objekt se přiřadí do toho shluku, k jehož středu má nejblíže. Následně se opět musí přepočítat střed nových shluků. Tento postup se opakuje do úplné konvergence. Základní princip metody k-means a rozřazování do shluků, je znázorněn na obrázku číslo 5. Tato metoda se používá v mnoha oblastech zpracování dat. Velmi často se využívá ke zpracování signálů ve zdravotnictví ([12], [13]) a mimo jiné byla využita při detekci leukémie z mikroskopického obrazu v roce 2016 v Indii [14]. [15] [16] [17]



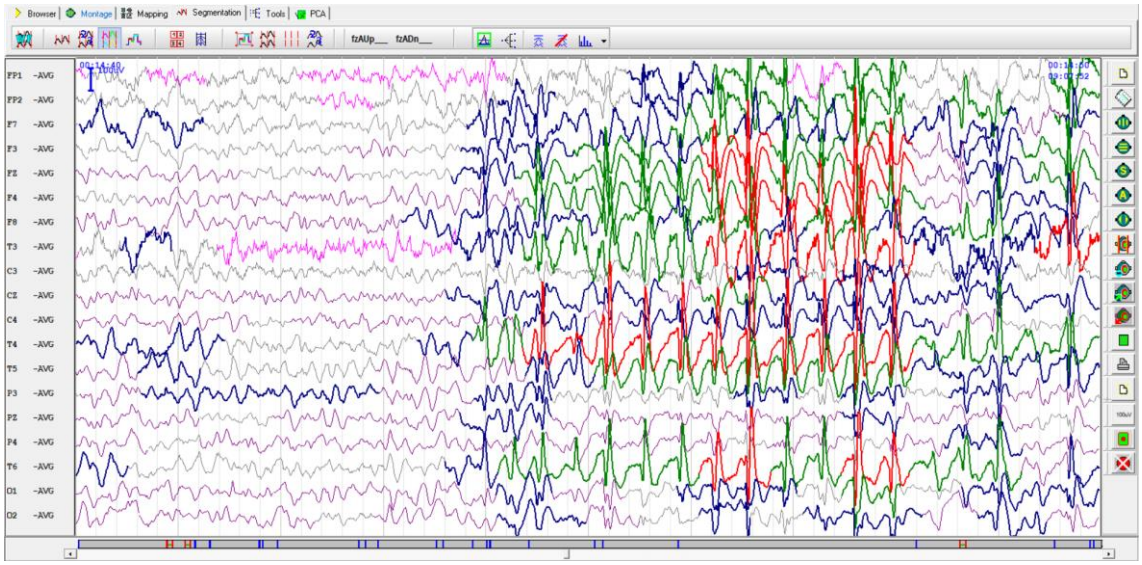
Obrázek 2.5: Popsaný princip metody k-means v jednotlivých krocích. V prvním kroku je naznačeno náhodné zvolení těžiště. Ve druhém kroku je znázorněno rozřazení dat ke středům (červená barva znázorňuje první shluk a modrá druhý shluk). Ve třetím kroku je znázorněn přepočet nového těžiště a jeho přesunutí do modrého shluku. V posledním kroku je znázorněno opětovné přiřazení objektů k novému těžišti a opětovné přepočtení těžišť. Převzato z [2].

Jednou z nevýhod této metody je náhodná iniciace dat, tedy důsledkem této iniciace je, že může určit dvě centra v jednom shluku, jako je naznačeno na obrázku číslo 2.6. V důsledku této skutečnosti, kdy metoda k-means špatně rozřadí zbytek naměřených dat (například z důvodu konvergence do lokálního minima), byla vyvinuta metoda k-means++. [16] [17]



Obrázek 2.6: Chybné určení centra shluku metodou k-means, kde hvězdička značí náhodně určený střed shluku, kolečka značí jednotlivé objekty a modrý obrazec značí označení shluku (klasifikace objektů). Objekty označeny v modrém obrazci přísluší konkrétnímu obrazci, tedy shluku (ty, které se nacházejí uvnitř obrazce). Převzato z [17]*a upraveno.*

V oblasti klasifikace EEG signálu se metoda k-means mimo jiné využívá v programu Wave-Finder. Po provedení segmentace jsou grafoelementy klasifikovány do shluků. Podle počtu sledovaných grafoelementů (zvoleno odborníkem) je určen i počet shluků. Využití metody k-means v praxi je znázorněno na obrázku 2.7. Konkrétně program Wave-Finder využívá rozdělení do sedmi shluků.

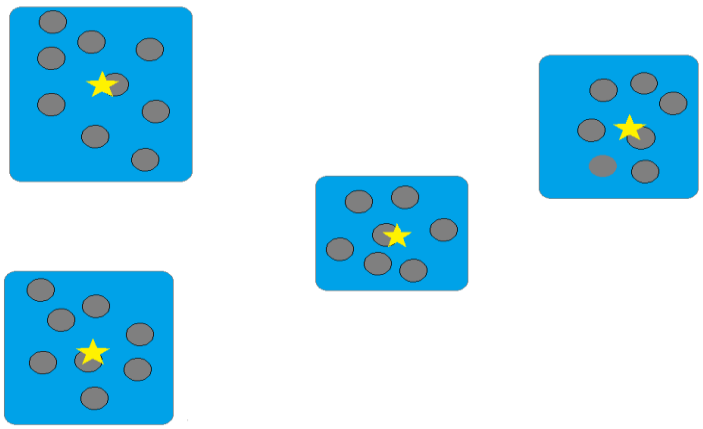


Obrázek 2.7: Výstup metody k-means využité v praxi, konkrétně v programu Wave-Finder. Barevně jsou označeny různé grafoelementy. Červeně je označena výrazná epilepsie, zeleně je označen epileptický grafoelement s pomalejším průběhem. Dalšími barvami jsou označeny artefakty (EMG a EOG) a fyziologická aktivita.

2.6.2 Metoda k-means++

Metoda k-means++ se liší od metody k-means v inicializaci dat. Rozšíření metody k-means o nenáhodnou inicializaci dat provedli v roce 2007 David Arthur a Sergei Vassilvitski. Toto rozšíření bylo provedeno z důvodu chybovosti metody k-means, které je znázorněno na obrázku 2.6 a popsáno v kapitole 2.5.3.

Metoda k-means++ dříve, než určí středy nových shluků, propočítá vzájemnou vzdálenost všech objektů. Na základě vzdáleností všech objektů vypočte pravděpodobnost pro všechny objekty. Vypočtená pravděpodobnost určuje, zda právě konkrétní objekt spadá k nejbližšímu středu, který je zvolen pomocí nejvzdálenějšího objektu. Díky této pravděpodobnosti má metoda k-means++ přehled o typu dat jako celku, tedy zvolení center už není čistě náhodné, jako je tomu u metody k-means. Nově zvolené středy shluků přesněji odpovídají povaze dat. Zvolení středů shluků dle metody k-means++ je znázorněno na obrázku 2.7. Po přepočtu vzdáleností a pravděpodobností, zbytek algoritmu má stejný průběh jako metoda k-means.[15] [17] [16]



Obrázek 2.7: Zvolení středů shluků pomocí metody k-means++, na základě počáteční inicializace dat (popsáno v kapitole 2.5.3). Hvězdička značí náhodně určený střed shluku, kolečka značí jednotlivé objekty a modrý obrazec značí označení shluku (klasifikace objektů). Objekty označeny v modrém obrazci přísluší konkrétnímu obrazci, tedy shluku (ty, které se nacházejí uvnitř obrazce). Převzato z [17] *a upraveno.*

3 Metodika

Metody využívané pro klasifikaci EEG signálu jsou založeny na různých přístupech, proto je potřeba porovnat metodu k-means++ s metodou založenou na podobném základu, tedy metodou k-means, která se v praxi využívá pro klasifikaci EEG záznamu jako metoda bez učitele.

Simulovaná data byla vytvořena tak, aby mohla odhalit rozdíly mezi těmito metodami.

3.1 Programovací prostředí

MATLAB

Matlab je programovací prostředí a skriptovací programovací jazyk vyvíjen společností Mathworks a umožňuje především počítání s maticemi, vykreslování 2D a 3D grafů funkcí, implementaci algoritmů, analýz a prezentaci dat i vytváření aplikací. V této práci jsem využila verzi MATLAB 2014b a Statistics and Machine Learning Toolbox pro funkci kmeans. [18]

Wave-Finder (WF)

Tento program je využíván v oblasti zpracování EEG záznamů, protože umožňuje vizuální i matematické hodnocení digitálních EEG záznamů. Uživatel si volí sám parametry ovlivňující charakter příznaků, které jsou využívané pro klasifikaci EEG záznamu. Autorem programu je pan doc. Ing. Vladimír Krajča, CSc. V programu Wave-Finder bylo provedeno předzpracování dat (segmentace a extrakce příznaků). Program Wave-Finder byl využit pro zobrazení klasifikace reálných záznamů EEG a následné zhodnocení výsledků klasifikace k-means++. [19]

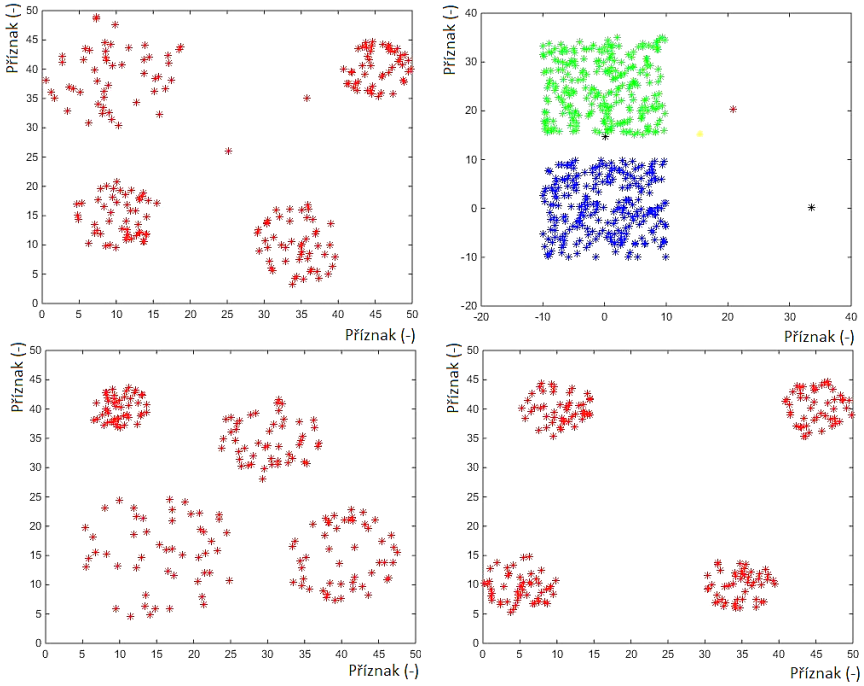
3.2 Data

Data reálného signálu EEG byla segmentována pomocí adaptivní segmentace konkrétně segmentace pomocí dvou spojených oken (princip viz kapitola 2.3). Výstup segmentace obsahuje i soubor, ve kterém jsou uloženy příznaky nacházející se ve všech segmentech signálu. Extrakce příznaků, kde jejich počet je 24 (dvacet čtyři), je popsána v diplomové práci Piorecký, 2016 [10].

3.2.1 Simulovaná data

Pomocí základních úkonů v programu MATLAB jsem si vytvořila simulovaná data ve 2D (vykreslená vzájemná závislost 2 příznaků). Tyto data tvoří shluky, které jsou od sebe vzdálené, nebo přiblížené s vyskytujícími se občasnými samostatnými objekty,

které na první pohled nepatří k žádnému shluku. Shluky jsou vytvořeny tak, aby bylo možné vizuálně stanovit rozdíl mezi výstupem metod k-means a k-means++.



Obrázek 3.7: Příklad čtyř různých simulovaných dat využívaných pro porovnání metod k-means a k-means++
V případě oválných shluků jsem nevolila počáteční barevné odlišení, jelikož data jsou vcelku přehledná. Pro shluky obdélníkového tvaru jsem zvolila barevné odlišení z důvodu lepší orientace mezi hranicí shluku a osamoceným objektem.

3.2.2 Reálná data

Testovanou metodu k-means++ jsem otestovala mimo simulovaných dat i na reálných záznamech EEG. Záznamy byly naměřeny v Nemocnici Na Bulovce přístrojem Brainquick. Data jsou měřena pomocí klinického vyšetření na subjektech, které mají podezření na epilepsii. Tato měření byla schválena etickou komisí Nemocnice Na Bulovce v roce 2011. Před měřením všechny subjekty podepsaly informovaný souhlas. Věk měřených subjektů je mezi 26 až 60 roky a jedná se o subjekty různého pohlaví.

Pro testování metody k-means++ jsem využívala záznam pěti subjektů. Tyto záznamy byly anonymizovány. Záznamy EEG byly naměřeny pomocí devatenácti kanálového EEG, kdy elektrody byly rozmístěny na hlavě pacienta v systému 10-20 (deset-dvacet) a využívaná montáž byla Average.

Tato data jsou již upravena pro aplikaci shlukové analýzy. Úprava dat byla provedena v programu Wave-Finder, kdy byla provedena segmentace dat pomocí dvou spojených

oken. Extrahované příznaky jsou popsány v diplomové práci Piorecký, 2016 [10]. Klasifikace předzpracovaných dat byla provedena pomocí metod automatické klasifikace, nejdříve metodou k-means a následně metodou k-means++.

3.3 K-means

Definice metody k-means [15] [3] [2]:

1. Algoritmus si náhodně vybere počáteční rozdělení s *K* shluky
2. Vygeneruje novou pozici objektu na základě přiřazení všech objektů k původnímu centru shluku
3. Vypočte nové středy shluků
4. Opakuje kroky 2 a 3 dokud přiřazení objektu k danému shluku nezůstane stabilní.

Princip metody k-means je podrobněji popsán v teoretické části a tuto definici popisuje obrázek 2.5.

Pro klasifikaci EEG záznamu však nemusí metoda k-means pracovat správně. V této práci jsem využila pro klasifikaci reálných EEG záznamů algoritmus k-means v programu Wave-Finder. Dle chybovosti popsané v teoretické části lze předpokládat, že některé příznaky vyhodnotí nesprávně. Proto metodu k-means porovnávám s metodou k-means++, kde by měla být odstraněna chyba způsobená náhodnou inicializací.

3.4 K-means++

Metoda k-means++ je popsána pomocí definice [15] [13] [20]:

1. Algoritmus si určí náhodně jeden libovolný střed c_1 (shluk $C=\{c_1\}$) z daného setu dat *X*.
2. Pro každou hodnotu objektu m z dat *X*, vypočítá vzdálenost k počátečnímu středu c_1. Označí vzdálenost mezi c_1 a pozorovaným objektem *m* jako $d(x_m, c_1)$.
3. Spočítá pravděpodobnost P pro každý objekt dle vzorce:

$$P = \frac{d^2(x_m, c_1)}{\sum_{j=1}^{n} d^2(x_j, c_1)}, \tag{4}$$

kde $d(x_m, c_1)$ značí vzdálenost mezi počátečním centrem c_1 a pozorovaným objektem *m,* ve jmenovateli se nachází součet všech vzdáleností všech objektů k počátečnímu centru.

4. Jako další centrum c_2 je označen objekt, který se nachází nejdále od centra c_1. Nejvzdálenější objekt je určen pomocí výše uvedené pravděpodobnosti.

5. Určení dalších středů probíhá obdobně jako v předchozím kroku, s rozdílem, že je brána v potaz vzdálenost konkrétních objektů pro určité centrum.
6. Opakuje se krok 4, dokud není vytvořeno *k* center příslušících počtu shluků navolených uživatelem.
7. Pro vytvořené iniciační středy shluků provede algoritmus k-means klasifikaci objektů do shluků.

Metodu k-means++ jsem implementovala na základě výše popsané definice. Stejně jako pro použití metody k-means, je třeba i pro metodu k-means++ vložit vstupní hodnoty. Jako vstup algoritmu je třeba vložit vstupní data (v mém případě simulovaná data a reálné EEG záznamy) a počet shluků. Počet shluků pro simulovaná data jsem volila podle konkrétních dat. V případě reálných EEG dat je zvoleno sedm shluků, jelikož tento počet využívá algoritmus k-means v programu Wave-Finder (empiricky zjištěno).

Metoda kmeans++ rozřadí všechny objekty do shluků. Nebere v úvahu šumové nebo okrajové objekty, na rozdíl od funkce fuzzy k-means nebo například hustotně založeného algoritmu DBSCAN. [21]

3.5 Analýza dat

U metod k-means a k-means++ jsem prováděla kvalitativní a kvantitativní zhodnocení pro simulovaná data. Kvantitativní porovnání jsem provedla pomocí měření časové náročnosti pro obě metody. Obě metody jsem spustila 20x na souhlasná data. Ze změřených časů jsem zvolila medián, který reprezentuje časovou náročnost metody pro patřičná data. Pomocí kvalitativní analýzy jsem srovnávala účinnost metod na testovaná data. Dohromady simulovaná data tvoří deset druhů, které jsou analyzovány pomoci metod k-means a k-means++.

Reálný EEG signál jsem analyzovala pro metody k-means a k-means++. Výstup obou metod byl vyhodnocen za pomoci a statisticky zpracován kvalitativní analýzou.

Klasifikováno bylo pět EEG záznamů. Expertem byla stanovena klasifikace do sedmi shluků. Vyhodnocení metod pro všech pět záznamů bylo sjednoceno expertem.

Sledovány byly čtyři grafoelementy: fyziologická aktivita, EMG artefakt, pomalé oční artefakty a epileptická aktivita. Počet grafoelementů se neshoduje s počtem shluků z důvodu různých průběhů těchto grafoelementů (např. fyziologická aktivita je rozřazena do více shluků podle amplitudy a frekvence segmentů fyziologické aktivity). Epileptická aktivita je také zařazena do více shluků dle průběhu. Počet shluků byl určen expertem podle empirických zkušeností s programem Wave-Finder.

Z důvodu automatické segmentace vznikly i segmenty, které nelze vyhodnotit. Tyto chybné segmenty byly z analýzy vyřazeny (obrázek 4.8).

Obrázek 4.8: Příklad chybného segmentu (v kroužku), který byl z analýzy vyřazen.

Požadovaný výstup klasifikace reálného EEG signálu je znázorněn na obrázku 2.7, který je uveden v kapitole 2.5.2.

3.6 Kvalitativní analýza

Pro vyhodnocení účinnosti metod k-means a k-means++ jsem využila ROC analýzu. ROC analýza se obvykle využívá pro binární případy tříd, jelikož její definice a interpretace je jednoduchá a přesná. V mém případě jsem porovnávala, zda objekty v souboru dat jsou správně rozřazeny do shluků, nebo nikoliv. Při analýze se soubor rozřadí na správně rozřazené objekty do shluků, nebo jsou objekty nesprávně rozřazené. Získáváme tak čtyři množiny pro každý shluk.

1. True positive (TP) – množina dat, které byly správně přiřazeny k příslušnému shluku
2. True negative (TN) – množina dat, které byly správně nepřiřazeny ke konkrétnímu shluku
3. False negative (FN) – data, která patří ke konkrétnímu shluku, avšak jsou přiřazeny nesprávně k jinému
4. False pozitive (FP) – data, která nepaří ke konkrétnímu shluku, avšak jsou k němu nesprávně přiřazeny

Z těchto hodnot lze definovat základní veličiny ROC analýzy, mezi které patří například senzitivita, specificita a pozitivní prediktivní hodnota testu. Mé vyhodnocení bude sestaveno ze specificity, selektivity a pozitivní prediktivní hodnoty. [22] [23]

3.6.1 Specificita

Specificitu (SP) lze také označit jako pravdivě negativní poměr. Tento poměr je definován (se vztažením na vyhodnocení dat využitých v této práci) jako podíl mezi správně nezařazeny do konkrétního shluku a všemi daty správně rozřazenými do ostatních shluků společně s daty nesprávně přiřazenými ke konkrétnímu shluku. [22] [23]

$$SP = \frac{TN}{TN+FP},$$

(6)

kde *SP* značí specificitu, *TN* objekty, které byly správně nepřiřazeny ke konkrétnímu shluku a *FP* značí objekty, které nepatří ke konkrétnímu shluku, avšak jsou k němu nesprávně přiřazeny.

3.6.2 Senzitivita

Senzitivita (SE) je označena jako pravdivě pozitivní poměr. Pravdivě pozitivní poměr je definován jako poměr mezi daty, které jsou správně přiřazeny ke konkrétnímu shluku a všemi daty zařazených do konkrétního shluku. [22] [23]

$$SE = \frac{TP}{TP + FN},$$

(7)

kde *SE* značí senzitivitu, *TP* objekty, které byly správně přiřazeny k příslušnému shluku a *FN* značí objekty, které patří ke konkrétnímu shluku, avšak jsou přiřazeny nesprávně k jinému.

3.6.3 Pozitivní prediktivní hodnota

Tato hodnota (PPV) určuje pravděpodobnost, že daný objekt je správně přiřazen ke konkrétnímu shluku, jestliže má k jeho centru nejblíže. Definice je dána jako poměr pravdivě pozitivních výsledků ke všem pozitivním výsledkům. [22] [23]

$$PPV = \frac{TP}{TP + FP},$$

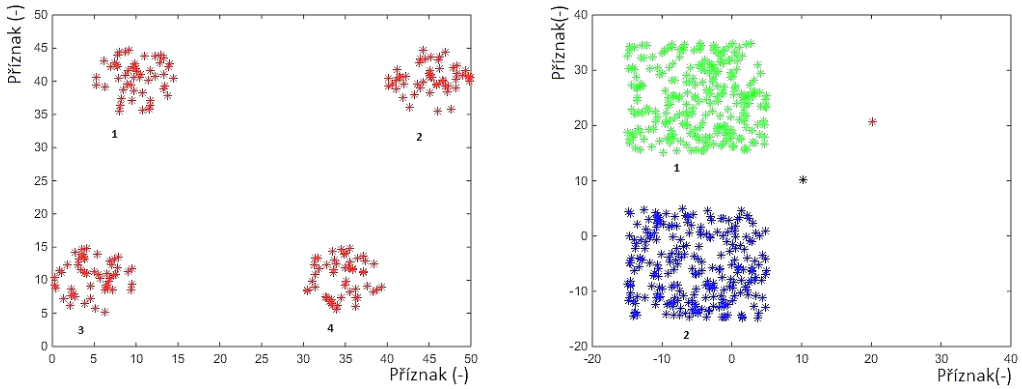
(8)

kde *PPV* je označení pro pozitivní prediktivní hodnotu, *TP* značí objekty, které byly správně přiřazeny k příslušnému shluku a *FP* značí objekty, které nepatří ke konkrétnímu shluku, avšak jsou k němu nesprávně přiřazeny.

4 Výsledky

4.1 Klasifikace simulovaných dat

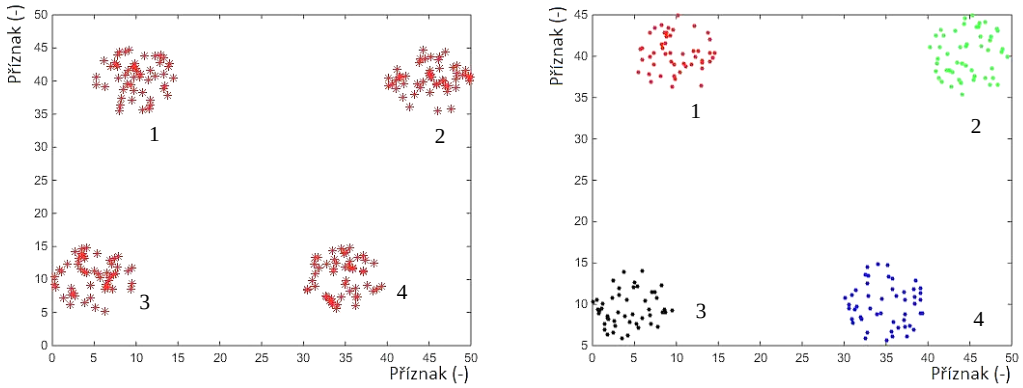
Klasifikováno bylo deset typů simulovaných dat metodami k-means a k-means++. U každého obrázku je uvedena tabulka s časovou náročností obou metod a druhá tabulka obsahuje výsledky kvalitativní analýzy, tedy hodnoty senzitivity (SE), specificity (SP) a PPV pro každý shluk konkrétních dat po aplikaci metod k-means a k-means++. V případě, že se chybná klasifikace objevila alespoň v deseti procentech případů pro danou metodu a typ dat, byl hodnocen výsledek této chybné klasifikace. Chybná klasifikace je znázorněna pod tabulkami. Pořadí shluků je označeno v následujícím obrázku a pořadí se nemění.



Obrázek 4.9: Příklad simulovaných dat a číselného označení shluků

Data 1

Data 1 je základní příklad dat v příznakovém prostoru. U tohoto typu nejsou předpokládány žádné odchylky nebo špatné roztřídění do shluků, jelikož hranice shluků jsou patrné. Data 1 jsou především kontrolní skupinou.



Obrázek 4.10: Simulovaná data 1 s čísly jednotlivých shluků (obrázek vlevo) a jeho správná klasifikace, kde jsou barevně odděleny jednotlivé shluky (obrázek vpravo)

Tabulka 1: Kvantitativní vyhodnocení

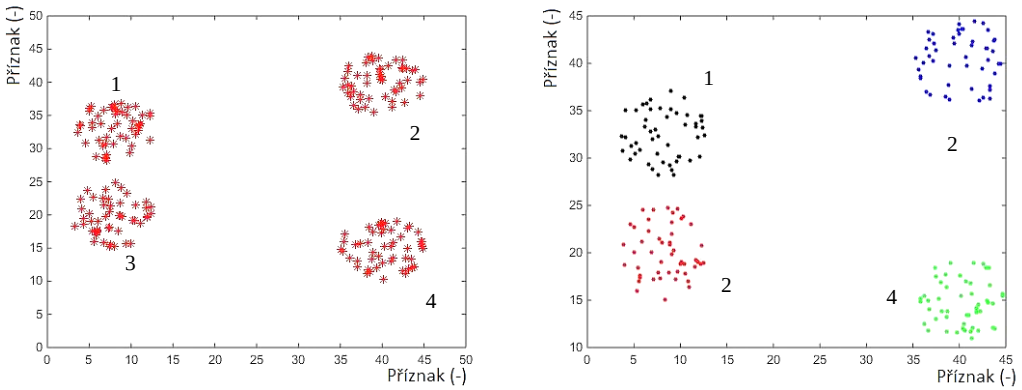
Metoda	**Časová náročnost [s]**
k-means	0,22
k-means++	0,22

Tabulka 2: Kvalitativní vyhodnocení simulovaných dat 1 (spočítaná senzitivita, specificita a PPV pro každý shluk a metody k-means a k-means++)

Shluk	**k-means**		**k-means++**	
1	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00
2	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00
3	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00
4	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00

Data 2

Simulovaná data 2 jsou velmi podobná typu simulovaných dat Data 1. Odlišují se v přiblížení dvou shluků. Poloměr a velikost shluků se mezi sebou nijak neliší, rozdíl od Dat 1 je pouze v poloze.



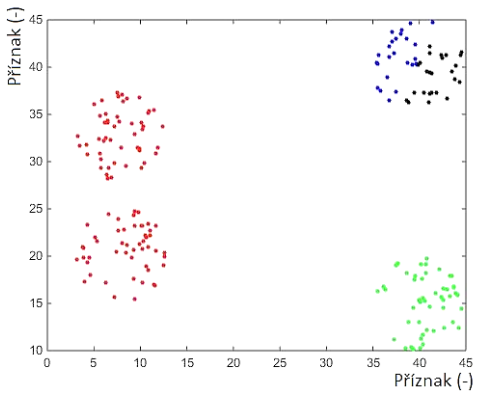
Obrázek 4.11: Simulovaná data 2 s čísly jednotlivých shluků (obrázek vlevo) a jeho správná klasifikace, kde jsou barevně odděleny jednotlivé shluky (obrázek vpravo)

Tabulka 3: Kvantitativní vyhodnocení simulovaných dat 2 (časová náročnost metod k-means a k-means++)

Metoda	**Časová náročnost [s]**
k-means	0,25
k-means++	0,20

Tabulka 4: Kvalitativní vyhodnocení simulovaných dat 2 (spočítaná senzitivita, specificita a PPV pro každý shluk a metody k-means a k-means++)

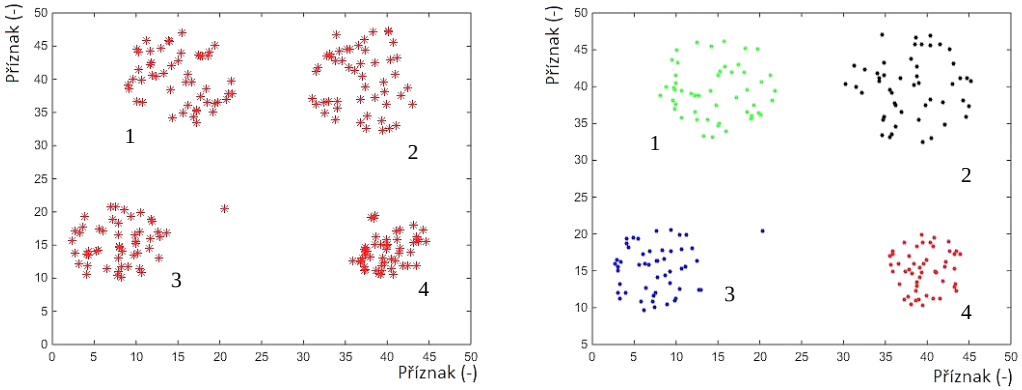
Shluk	**k-means**		**k-means++**	
1	Specificita	1,00	Specificita	1,00
	Senzitivita	0,00	Senzitivita	1,00
	PPV	0,00	PPV	1,00
2	Specificita	1,00	Specificita	1,00
	Senzitivita	0,50	Senzitivita	1,00
	PPV	1,00	PPV	1,00
3	Specificita	0,67	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	0,50	PPV	1,00
4	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00



Obrázek 4.12: Chybné vyhodnocení simulovaných dat 2 metodou k-means

Data 3

Simulovaná data 3 vychází ze základního typu. Navíc je přidaný objekt, který se nachází mezi dvěma spodními shluky. Polohu tohoto objektu lze určit i bez přepočtu. Je na první pohled jasné, že tento objekt patří ke shluku vlevo dole.



Obrázek 4.13: Simulovaná data 3 s čísly jednotlivých shluků (obrázek vlevo) a jeho správná klasifikace, kde jsou barevně odděleny jednotlivé shluky (obrázek vpravo)

Tabulka 5: Kvantitativní vyhodnocení simulovaných dat 3 (časová náročnost metod k-means a k-means++)

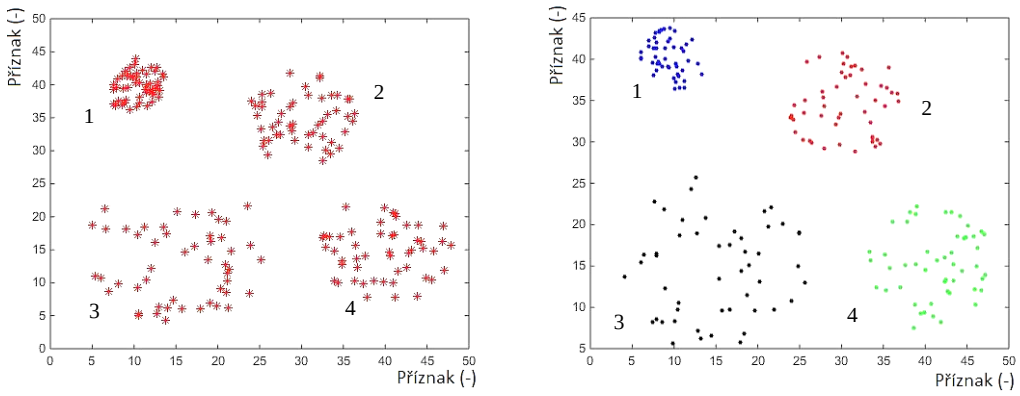
Metoda	Časová náročnost [s]
k-means	0,24
k-means++	0,23

Tabulka 6: Kvalitativní vyhodnocení simulovaných dat 3 (spočítaná senzitivita, specificita a PPV pro každý shluk a metody k-means a k-means++)

Shluk	k-means		k-means++	
1	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00
2	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00
3	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00
4	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00

Data 4

Simulovaná data 4 se liší od ostatních především s velikostí shluků. Shluk vlevo dole zaujímá největší část, kdežto shluk vlevo nahoře je nejmenší. Stěžejní je však umístnění těchto shluků celkově, spodní shluky jsou větší než horní. Horní shluk vpravo je více přiblížen spodním shlukům.



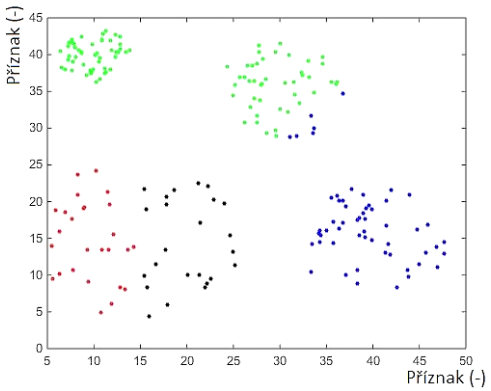
Obrázek 4.14: Simulovaná data 4 s čísly jednotlivých shluků (obrázek vlevo) a jeho správná klasifikace, kde jsou barevně odděleny jednotlivé shluky (obrázek vpravo)

Tabulka 7: Kvantitativní vyhodnocení simulovaných dat 4 (časová náročnost metod k-means a k-means++)

Metoda	Časová náročnost [s]
k-means	0,25
k-means++	0,24

Tabulka 8: Kvalitativní vyhodnocení simulovaných dat 4 (spočítaná senzitivita, specificita a PPV pro každý shluk a metody k-means a k-means++)

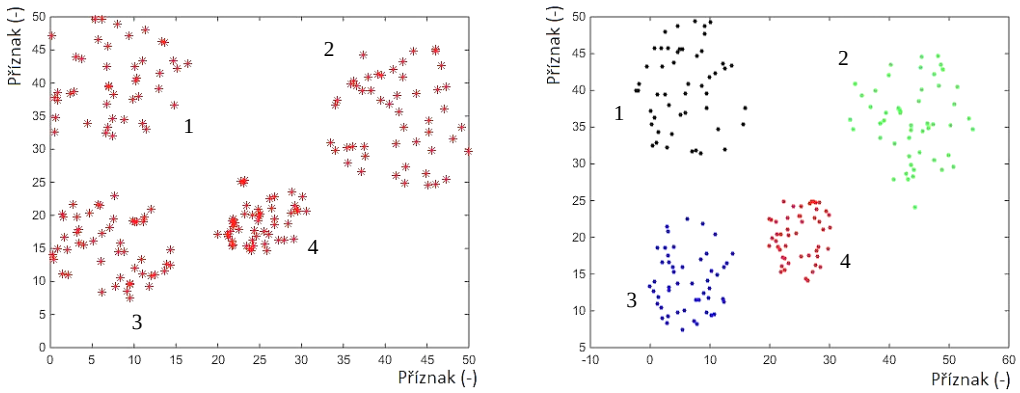
Shluk	k-means		k-means++	
1	Specificita	0,71	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	0,53	PPV	1,00
2	Specificita	0,83	Specificita	1,00
	Senzitivita	0,00	Senzitivita	1,00
	PPV	0,00	PPV	1,00
3	Specificita	1,00	Specificita	1,00
	Senzitivita	0,52	Senzitivita	1,00
	PPV	1,00	PPV	1,00
4	Specificita	0,96	Specificita	1,00
	Senzitivita	0,50	Senzitivita	1,00
	PPV	0,89	PPV	1,00



Obrázek 4.15: Chybné vyhodnocení simulovaných dat 4 metodou k-means

Data 5

Simulovaná data 5 se od ostatních dat liší především umístněním shluků. Nepočítaje shluk vlevo nahoře, shluky jsou umístěny po diagonále zobrazovacího okna. Velikost prostředního shluku na diagonále je nejmenší.



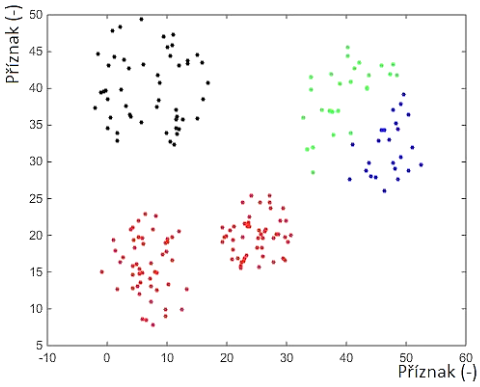
Obrázek 4.16: Simulovaná data 5 s čísly jednotlivých shluků (obrázek vlevo) a jeho správná klasifikace, kde jsou barevně odděleny jednotlivé shluky (obrázek vpravo)

Tabulka 9: Kvantitativní vyhodnocení simulovaných dat 5 (časová náročnost metod k-means a k-means++)

Metoda	**Časová náročnost [s]**
k-means	0,25
k-means++	0,24

Tabulka 10: Kvalitativní vyhodnocení simulovaných dat 5 (spočítaná senzitivita, specificita a PPV pro každý shluk a metody k-means a k-means++)

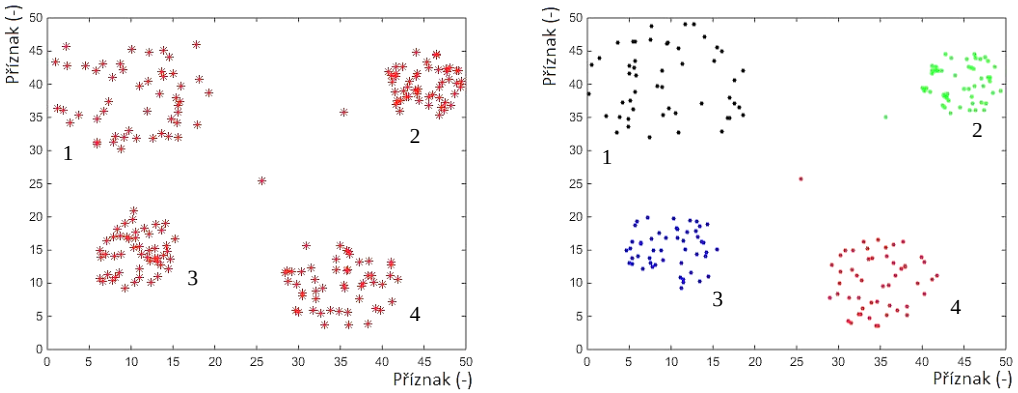
Shluk	**k-means**		**k-means++**	
1	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00
2	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	0,50
	PPV	1,00	PPV	1,00
3	Specificita	1,00	Specificita	0,67
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	0,50
4	Specificita	1,00	Specificita	0,71
	Senzitivita	1,00	Senzitivita	0,00
	PPV	1,00	PPV	0,00



Obrázek 4.17: Chybné vyhodnocení simulovaných dat 5 metodou k-means++

Data 6

Simulovaná data 6 vychází také ze základního souboru dat s oválnými shluky. V tomto souboru nehledě na velikost shluků jsou umístěny do prostoru 2 osamocené objekty, kdy nelze pouhým okem rozeznat ke kterému shluku patří.



Obrázek 4.18: Simulovaná data 6 s čísly jednotlivých shluků (obrázek vlevo) a jeho správná klasifikace, kde jsou barevně odděleny jednotlivé shluky (obrázek vpravo)

Tabulka 11: Kvantitativní vyhodnocení simulovaných dat 6 (časová náročnost metod k-means a k-means++)

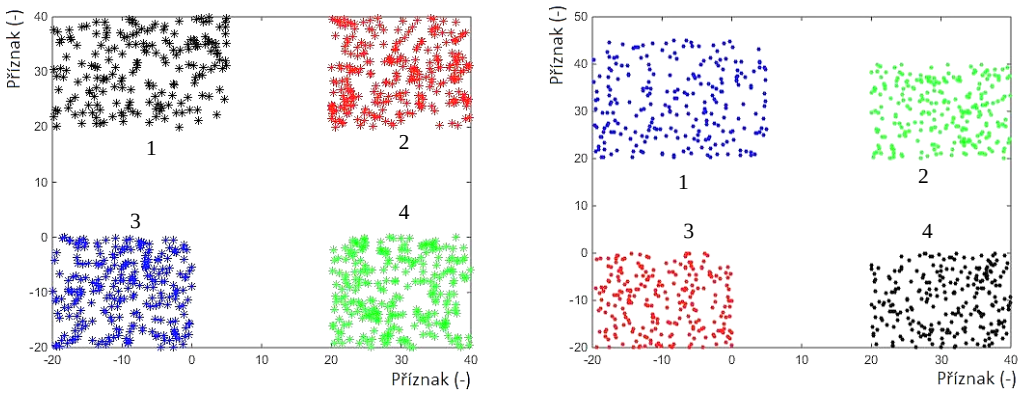
Metoda	Časová náročnost [s]
k-means	0,25
k-means++	0,24

Tabulka 12: Kvalitativní vyhodnocení simulovaných dat 6 (spočítaná senzitivita, specificita a PPV pro každý shluk a metody k-means a k-means++)

Shluk	k-means		k-means++	
1	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00
2	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00
3	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00
4	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00

Data 7

Shluky ze souboru dat 7 se liší především tvarem. Doposud byly shluky simulovaných dat oválného tvaru, nyní jsou shluky uspořádané do obdélníku. Data 7 slouží jako základní soubor dat se shluky obdélníkového tvaru.



Obrázek 4.19: Simulovaná data 7 s čísly jednotlivých shluků (obrázek vlevo) a jeho správná klasifikace, kde jsou barevně odděleny jednotlivé shluky (obrázek vpravo)

Tabulka 13: Kvantitativní vyhodnocení simulovaných dat 7 (časová náročnost metod k-means a k-means++)

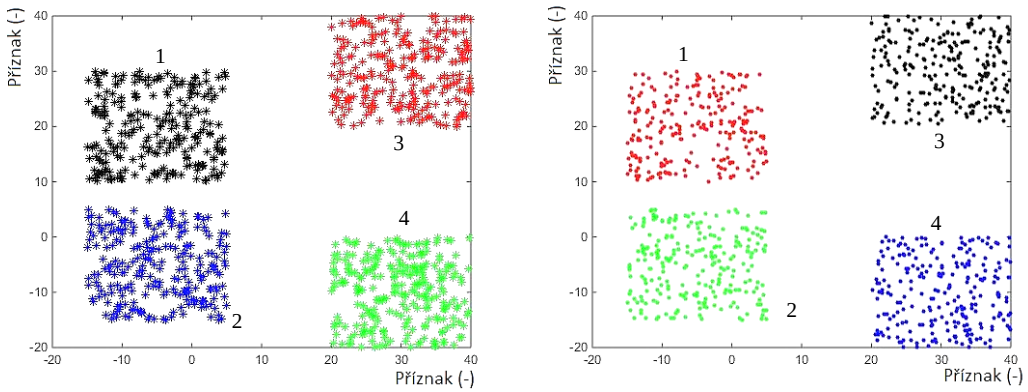
Metody	Časová náročnost [s]
k-means	0,25
k-means++	0,24

Tabulka 14: Kvalitativní vyhodnocení simulovaných dat 7 (spočítaná senzitivita, specificita a PPV pro každý shluk a metody k-means a k-means++)

Shluk	k-means		k-means++	
1	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00
2	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00
3	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00
4	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00

Data 8

Shluky ze souboru simulovaných dat 8 vycházejí ze základního souboru se shluky obdélníkového tvaru Data 7. Přiblížení dvou shluků o obdélníkovém tvaru je umístěno tak, aby se projevily schopnosti obou metod.



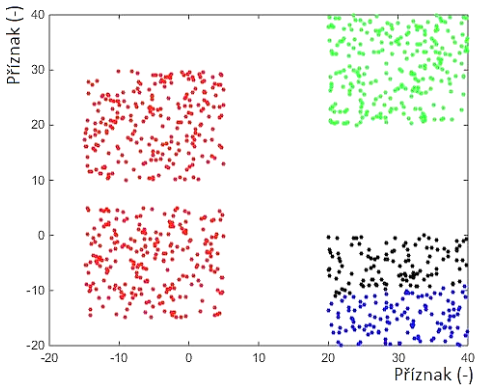
Obrázek 4.20: Simulovaná data 8 s čísly jednotlivých shluků (obrázek vlevo) a jeho správná klasifikace, kde jsou barevně odděleny jednotlivé shluky (obrázek vpravo)

Tabulka 15: Kvantitativní vyhodnocení simulovaných dat 8 (časová náročnost metod k-means a k-means++)

Metoda	Časová náročnost [s]
k-means	0,24
k-means++	0,24

Tabulka 16: Kvalitativní vyhodnocení simulovaných dat 8 (spočítaná senzitivita, specificita a PPV pro každý shluk a metody k-means a k-means++)

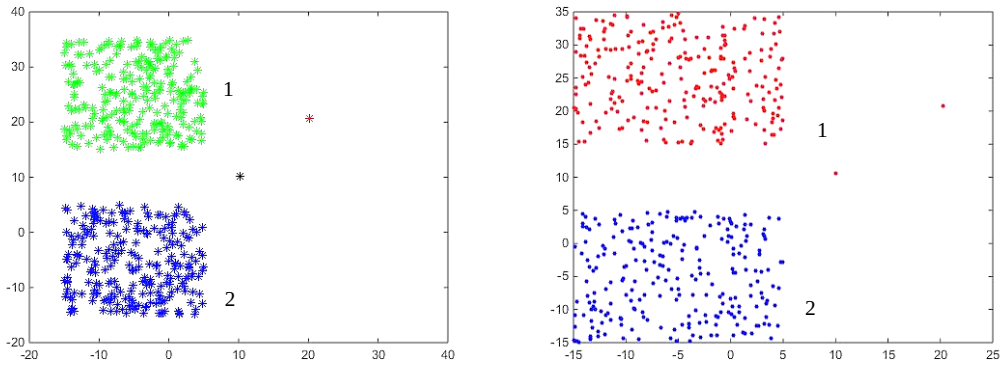
Shluk	k-means		k-means++	
1	Specificita	0,67	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	0,50	PPV	1,00
2	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00
3	Specificita	0,71	Specificita	1,00
	Senzitivita	0,00	Senzitivita	1,00
	PPV	0,00	PPV	1,00
4	Specificita	1,00	Specificita	1,00
	Senzitivita	0,50	Senzitivita	1,00
	PPV	1,00	PPV	1,00



Obrázek 4.21: Chybné vyhodnocení simulovaných dat 8 metodou k-means

Data 9

Tento typ simulovaných dat je odlišný od předchozích především počtem shluků. V tomto souboru dat jsou přítomné pouze dva osamocené objekty umístěné do prostoru.



Obrázek 4.22: Simulovaná data 9 s čísly jednotlivých shluků (obrázek vlevo) a jeho správná klasifikace, kde jsou barevně odděleny jednotlivé shluky (obrázek vpravo)

Tabulka 17: Kvantitativní vyhodnocení simulovaných dat 9 (časová náročnost metod k-means a k-means++)

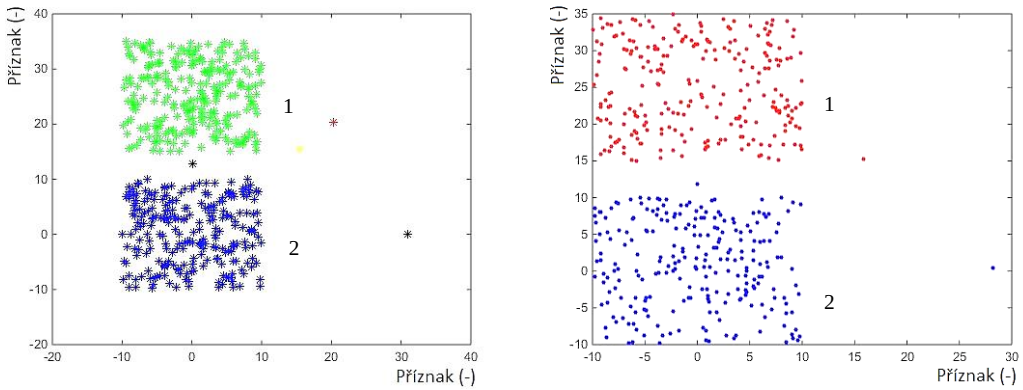
Metoda	**Časová náročnost [s]**
k-means	0,24
k-means++	0,23

Tabulka 18: Kvalitativní vyhodnocení simulovaných dat 9 (spočítaná senzitivita, specificita a PPV pro každý shluk a metody k-means a k-means++)

Shluk	**k-means**		**k-means++**	
1	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00
2	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00

Data 10

Poslední typ simulovaných dat se příliš neliší od Dat 9, pouze je umístěn třetí osamocený objekt.



Obrázek 4.23: Simulovaná data 10 s čísly jednotlivých shluků (obrázek vlevo) a jeho správná klasifikace, kde jsou barevně odděleny jednotlivé shluky (obrázek vpravo)

Tabulka 19: Kvantitativní vyhodnocení simulovaných dat 10 (časová náročnost metod k-means a k-means++)

Metoda	**Časová náročnost [s]**
k-means	0,25
k-means++	0,23

Tabulka 20: Kvalitativní vyhodnocení simulovaných dat 10 (spočítaná senzitivita, specificita a PPV pro každý shluk a metody k-means a k-means++)

Shluk	**k-means**		**k-means++**	
1	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00
2	Specificita	1,00	Specificita	1,00
	Senzitivita	1,00	Senzitivita	1,00
	PPV	1,00	PPV	1,00

4.2 Klasifikace reálných EEG dat

Reálná data byla klasifikována pomocí metod k-means a k-means++ v programu Wave-Finder. Klasifikováno bylo pět EEG záznamů. Expertem byla stanovena klasifikace do sedmi shluků. Expert sjednotil vyhodnocení záznamů, tedy hodnoty v tabulkách v kapitolách 4.2.1 a 4.2.2 platí pro segmenty ze všech signálů.

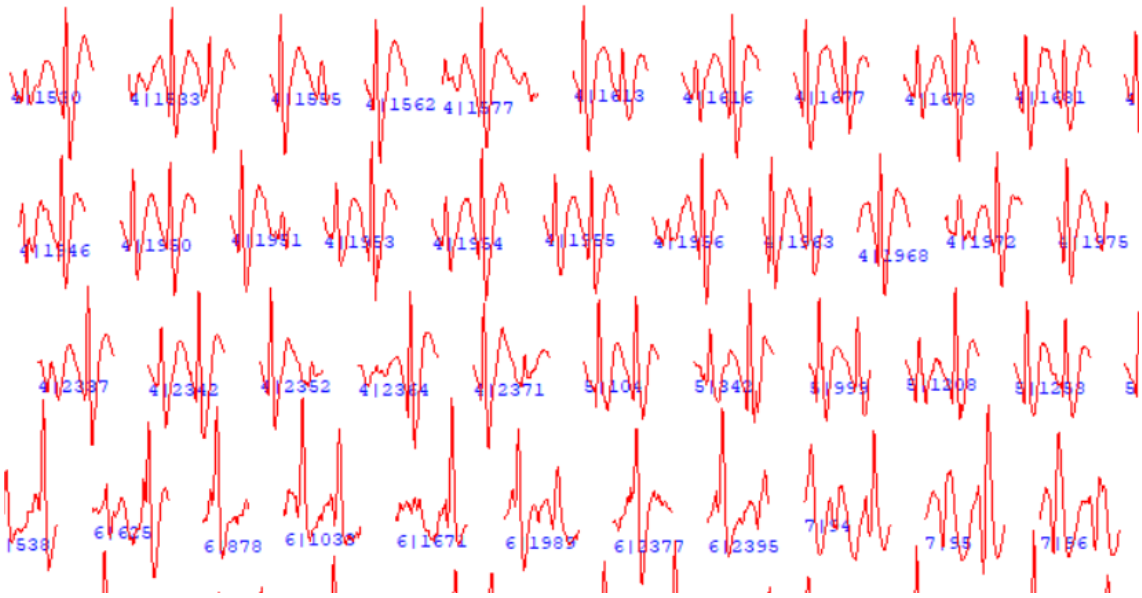
4.2.1 Klasifikace EEG záznamu pomocí metody k-means

V této kapitole je uvedena klasifikace dat pomocí metody k-means. Jednotlivé obrázky v této kapitole znázorňují příklad klasifikace reálných EEG dat v programu Wave-Finder pomocí metody k-means. V tabulce 21 je uvedeno statistické vyhodnocení klasifikace pomocí metodyu k-means.

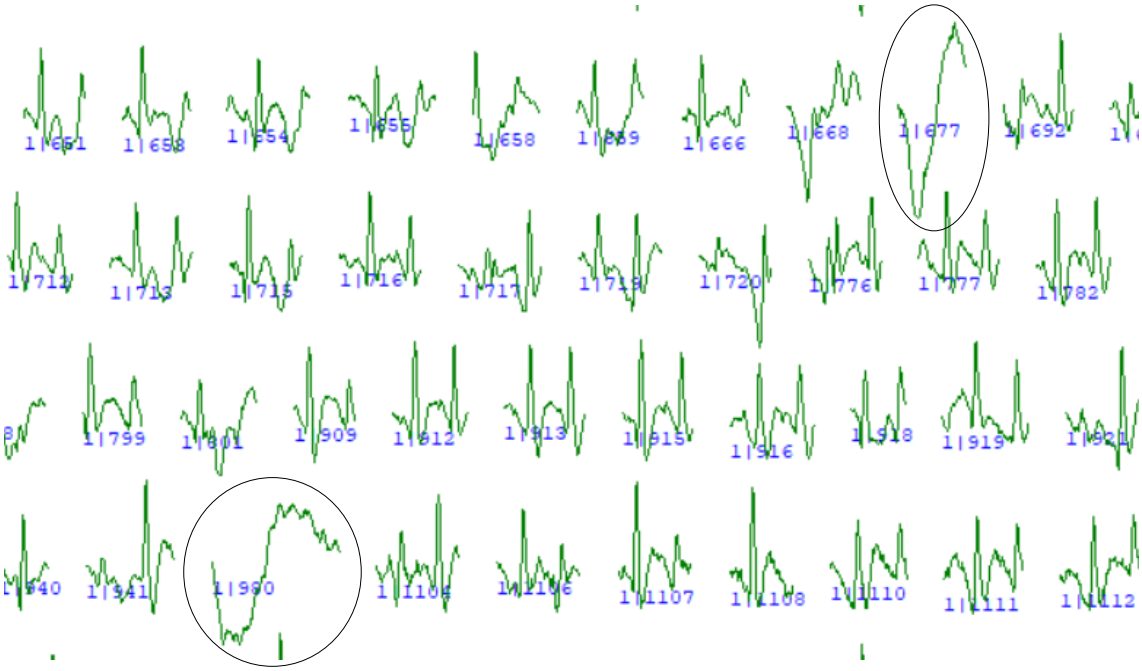
Tabulka 21: Kvalitativní vyhodnocení (senzitivita, specificita a PPV) reálných EEG záznamů metodou k-means. Hodnoceny byly shluky Fyziologické aktivity, EMG artefaktu, pomalých očních artefaktů a epileptické aktivity.

	Fyziologická aktivita	**EMG artefakt**	**Pomalé oční artefakty**	**Epileptická aktivita**
Senzitivita	0,49	0,99	0,70	0,71
Specificita	0,38	1,00	0,67	0,95
PPV	0,81	0,84	0,01	0,68

Klasifikace do shluků



Obrázek 4.24 Příklad segmentů rozřazených do shluku epileptické aktivity metodou k-means. Tento shluk je odlišen od ostatních grafoelementů červenou barvou. Obrázek převzat z programu Wave-Finder.



Obrázek 4.25: Příklad segmentů epileptických grafoelementů klasifikovaných metodou k-means a vyskytujících se mimo shluk epileptické aktivity. Ve shluku jsou obsaženy i segmenty pomalé oční aktivity (v kroužku). Obrázek převzat z programu Wave-Finder.

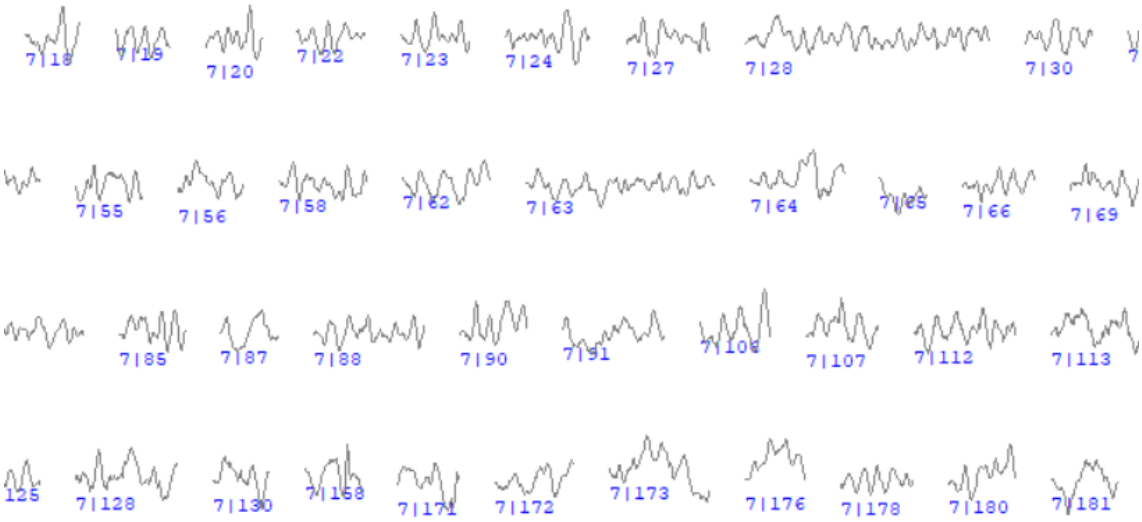
4.2.2 Klasifikace EEG záznamu pomocí metody k-means++

V této kapitole je uvedena klasifikace dat pomocí metody k-means++. Jednotlivé obrázky v této kapitole znázorňují příklad klasifikace reálných EEG dat v programu Wave-Finder. V tabulce 22 je uvedeno statistické vyhodnocení klasifikace pomocí metody k-means.

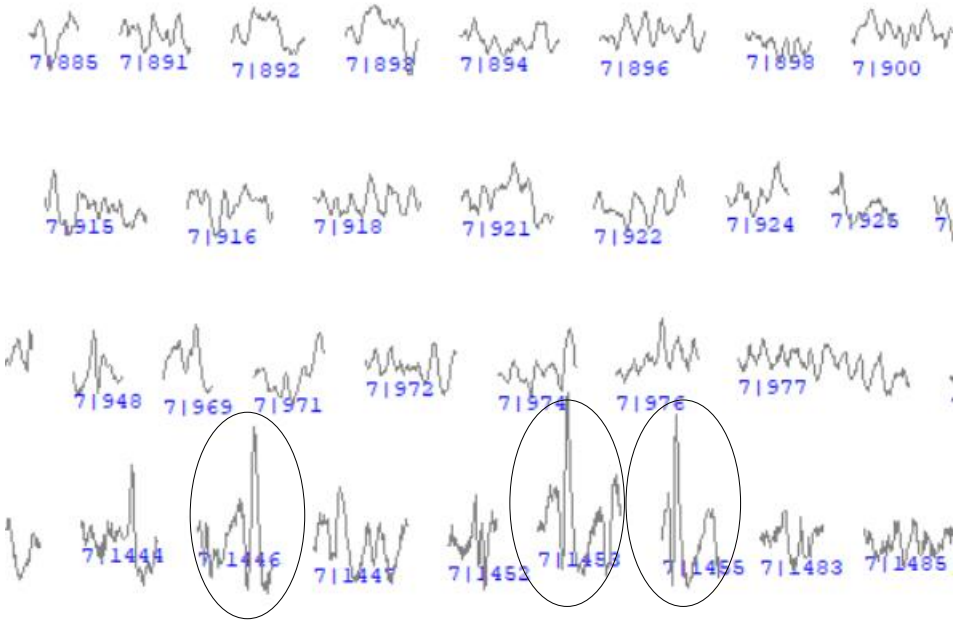
Tabulka 22: Kvalitativní vyhodnocení (senzitivita, specificita a PPV) reálných EEG záznamů metodou k-means++. Hodnoceny byly shluky Fyziologické aktivity, EMG artefaktu, pomalých očních artefaktů a epileptické aktivity

	Fyziologická aktivita	**EMG artefakt**	**Pomalé oční artefakty**	**Epileptická aktivita**
Senzitivita	0,41	0,54	0,83	0,31
Specificita	0,86	0,95	0,97	0,93
PPV	0,95	0,17	0,17	0,39

Klasifikace do shluků



Obrázek 4.26: Příklad shluku fyziologické aktivity klasifikovaného metodou k-means++. Tato část shluku znázorňuje správně klasifikované segmenty.



Obrázek 4.27: Příklad segmentů fyziologické aktivity klasifikovaných metodou k-means++. Ve shluku jsou obsaženy i segmenty epileptické aktivity (v kroužku). Obrázek převzat z programu Wave-Finder.

5 Diskuze

Cílem této práce bylo aplikovat metodu k-means++ na reálný EEG záznam. Nejdříve však bylo nutno provést porovnání metod k-means a k-means++ na simulovaných datech, aby se ověřila správnost mého návrhu metody k-means++. Pro porovnání byla vybrána metoda k-means, jelikož metoda k-means++ vychází z metody k-means. Po zhodnocení výstupů klasifikace simulovaných dat obou metod jsem aplikovala metodu k-means++ na reálný EEG záznam.

Jelikož metoda k-means++ se liší od metody k-means inicializací dat, tak při sestavování simulovaných dat byl brán ohled na povahu těchto metod. Simulovaná data tvoří shluky, které jsou již na první pohled téměř lehce klasifikovány pouhým okem. Avšak některá data jsou obohacena o osamocené objekty, které v některých případech nelze určit bez výpočtu vzdálenosti. Konkrétní data nikdy nebudou nabývat stejných hodnot, jelikož jsou tvořena náhodně. Pozice a rozsahu vytvořených shluků a osamocených dat je docíleno násobením náhodně vygenerovaného čísla určitou konstantou. V jednotlivých typech dat se pozice, přítomnost osamocených objektů, tvar a rozsah shluků mění tak, aby bylo možno vyhodnotit odlišnost metod.

Konkrétní simulovaná data byla klasifikována pomocí obou metod dvacetkrát. Při každém analyzování dat jednou z metod byla měřena časová náročnost pomocí funkce v prostředí MATLAB. Z těchto dvaceti měření časové náročnosti pro danou metodu a konkrétní data byl určen medián. Zvolila jsem medián z důvodu malého počtu měření, pro které nelze předpokládat normální rozdělení. Za chybnou klasifikaci dat danou metodou jsem označila soubor dat, kdy metoda chybně klasifikovala tato data více než dvakrát z celkového počtu dvacetkrát spuštěného kódu metody.

Metoda k-means při klasifikaci simulovaných dat vykazovala mnohem větší variabilitu pořadí shluků, tedy při klasifikaci konkrétních dat několikrát pozměnila barevné označení pořadí (vždy minimálně 10x pro data obsahující čtyři shluky a minimálně 3x pro data obsahující dva shluky). Dle výše uvedených nároků na označení klasifikace za chybnou, metoda k-means vyhodnotila chybně tři typy dat ze sestavených deseti. Jedná se o typy dat Data 2, Data 4 a Data 8 (viz obrázek 4.12, obrázek 4.15 a obrázek 4.21). Jedná se o chybu, která je popsána v teoretické části v kapitole 2.5.2. Tato nesprávná klasifikace vznikla z důvodu náhodného určení jednoho z původních center mezi dva shluky. Po přepočítávání vzdáleností a určování pozice nového centra byly pozice center k objektům tvořící tyto dva shluky konstantní, proto tyto dva shluky byly označeny jako jeden. Z důvodu požadavku zadaného na čtyři shluky, musel být jeden shluk rozdělen nesprávně na dva. Proto jeden ze shluků u nesprávně klasifikovaných dat je označen jako dva.

Klasifikace simulovaných dat pomocí metody k-means++ měla mnohem pozitivnější výsledky oproti metodě k-means. Variabilita pořadí shluků byla v případě typu dat

o čtyřech shlucích vždy pozměněna v průměru 7x a pro data se dvěma shluky byla pozměněna maximálně 4x. Chybovost metody k-means++ byla menší než chybovost metody k-means. Metoda k-means++ nesprávně klasifikovala pouze jeden typ dat z deseti. Jedná se o typ dat Data 5 (viz obrázek 4.17).

Po zhodnocení výstupů klasifikace simulovaných dat pomocí obou metod jsem aplikovala metodu k-means++ na reálný EEG záznam. V případě klasifikace reálných dat je metoda k-means již v praxi využívaný v programu Wave-Finder. Bylo provedeno statistické vyhodnocení pro reálná data (tabulka 21), kdy uvedené hodnoty senzitivity, specificity a PPV nabývají hodnot od 0 do 1.

Nejvyšší senzitivita byla zjištěna při klasifikaci dat metodou k-means pro grafoelement příslušící EMG artefaktu. Senzitivita dosáhla hodnoty 0,99, což znamená, že EMG grafoelementy byly správně označeny téměř všechny a velmi malá část segmentů znázorňující grafoelementy EMG byla přiřazena nesprávně k jinému shluku, proto hodnota senzitivity není rovna jedné. Specificita pro EMG grafoelementy dosáhla hodnoty 1,00, tedy lze tvrdit na základě této hodnoty, že metoda k-means odlišila správně všechny segmenty které neodpovídají průběhu EMG grafoelementu. Pozitivní prediktivní hodnota je 0,84, což lze považovat stále za relativně přesnou detekci. PPV s touto hodnotou určuje, že ve shluku, který přísluší EMG grafoelementům, se vyskytují navíc i segmenty, které nejsou EMG grafoelementy. V signálu jsou tedy odlišeny téměř všechny EMG grafoelementy spolu s nesprávně označenými segmenty, které odpovídají jinému typu grafoelementu , ale jsou jako EMG grafoelementy označeny. Fyziologická aktivita byla detekována s hodnotou senzitivity 0,49. Tato hodnota vypovídá o tom, že v polovině segmentů označených jako fyziologická aktivita, ve skutečnosti fyziologickou aktivitou není. Správně označenou fyziologickou aktivitu tvoří pouze polovina shluku. Hodnota specificity pro fyziologickou aktivitu je 0,38. Hodnota 0,38 je poměrně nízká, tedy do shluku označujícího fyziologickou aktivitu jsou přiřazeny i grafoelementy, které neodpovídají průběhu fyziologické aktivity. PPV dosahuje pro shluk fyziologické aktivity hodnoty 0,81. Hodnota 0,81 je naopak od hodnot senzitivity a specificity stanovených pro fyziologickou aktivitu vyšší. Pro tuto hodnotu shluku platí, že většina segmentů přiřazených do shluku je skutečně fyziologická aktivita, ale jsou označeny v malém množství i grafoelementy, které průběhu fyziologické aktivity neodpovídají. Shluk označující pomalé oční artefakty dosáhl celkem vysoké hodnoty senzitivity a specificity. Na základě těchto získaných hodnot lze tvrdit, že shluk určený pro pomalé oční artefakty obsahuje správně grafoelementy znázorňující pomalé oční artefakty, ale i malou část grafoelementů, které neodpovídají průběhu pomalému očnímu artefaktu. Malá část pomalých očních artefaktů vyskytujících se v signálu byla zařazena do jiného nesprávného shluku, nejsou tedy v signálu správně odlišeny všechny grafoelementy znázorňující pomalé oční artefakty. Pozitivní prediktivní hodnota pro pomalé oční artefakty nabyla hodnoty pouze 0,01. Tato velmi nízká hodnota stanovuje, že ve shluku je přítomno velké množství grafoelementů, které vůbec do shluku

s pomalými očními artefakty nepatří. V signálu jsou tedy nesprávně označeny jako grafoelementy pomalých očních artefaktů i segmenty, které reprezentují např. fyziologickou aktivitu, EMG artefakt atd. Posledním sledovaným grafoelementem je průběh epileptické aktivity. Hodnoty specifity a senzitivity jsou vysoké (SP=0,95 a SE=0,71), tedy lze tvrdit, že shluk označující epileptickou aktivitu mimo grafoelementů znázorňující průběh epilepsie obsahuje i malou část grafoelementů, která neodpovídá epileptickému průběhu. Hodnota senzitivity značí, že malá část segmentů značící epileptickou aktivitu je přiřazena k jinému shluku, tedy není správně odlišena od ostatních grafoelementů dle požadavků. PPV pro epileptickou aktivitu je 0,68. Hodnota PPV značí, že část signálu je nesprávně označena jako epileptická aktivita, přestože dané nesprávně zařazené grafoelementy do shluku značícího epileptickou aktivitu neodpovídají epileptickému grafoelementu. Shluk určený pro grafoelementy epileptické aktivity mimo většiny grafoelementů správně zařazených obsahuje i několik grafoelementů příslušících jiné aktivitě nežli epilepsii.

Metoda k-means++ klasifikoval příslušná data v programovacím prostředí MATLAB, kdy výstupem byla matice o dvou sloupcích a řádcích s hodnotami signálu. V prvním sloupci byla příslušná hodnota a ve druhém sloupci bylo k hodnotě přiřazeno číslo shluku. Tato matice byla také vyhodnocena expertem. Statistické vyhodnocení klasifikace reálného EEG záznamu je uvedeno v tabulce 22.

Při klasifikaci pomocí metody k-means++ nejlépe v signálu byla detekována fyziologická aktivita. Specificita dosáhla hodnoty 0,86. Takto vysoká hodnota poukazuje na to, že byly správně odlišeny v signálu grafoelementy fyziologické aktivity od ostatních grafoelementů, ale jako fyziologická aktivita byly označeny i grafoelementy příslušící jinému shluku, tedy značící jinou aktivitu než fyziologickou. Hodnota senzitivity je značně nižší nežli hodnota specificity (SE=0,41). Nižší hodnota senzitivity poukazuje na to, že velká část grafoelementů fyziologické aktivity jsou zařazeny k jinému shluku, tedy jsou metodou k-means++ označeny nesprávně jako jiná aktivita. PPV značící obsah shluku dosáhla hodnoty 0,95. Ve shluku označující fyziologickou aktivitu jsou přítomné správně zařazeny grafoelementy značící průběh fyziologické aktivity a je ve shluku přítomna malá část grafoelementů příslušících jinému shluku (jiná aktivita než fyziologická). Grafoelementy značící průběh EMG artefaktu dosáhli nejvyšší hodnoty pro specificitu (SP=0,95). Specificita pro shluk odlišující EMG artefakt značí správné odlišení EMG artefaktů od ostatních grafoelementů v signálu. Senzitivita osáhla hodnoty 0,54 a na základě této hodnoty lze tvrdit, že část grafoelementů značících EMG artefakty jsou nesprávně přiřazeny k jiném shluku, tedy jsou nesprávně označeny jako jiná aktivita. Hodnota PPV je 0,17. Tato nízká hodnota určuje, že ve shluku značící EMG artefakty jsou nesprávně přítomny i grafoelementy značící jinou aktivitu. Označení EMG artefaktů v signálu tedy není velmi přesné. Pomalé oční artefakty dosáhly vyšších hodnot senzitivity a specificity (SP=0,97 a SE=0,83). V signálu byla velká část grafoelementů značících průběh pomalého očního artefaktu správně odlišena, avšak velmi malá část

grafoelementů značících jinou aktivitu byla k tomuto shluku nesprávně přiřazena. Na základě vysoké hodnoty senzitivity lze tvrdit, že jen malá část grafoelementů značících pomalé oční artefakty byla přiřazena k jinému shluku, tedy označena nesprávně jako jiná aktivita. PPV pro grafoelementy pomalých očních artefaktů je oproti metodě k-means mnohem vyšší, ale ve srovnání s výsledky PPV hodnoty u jiných grafoelementů je hodnota 0,17 stále velmi nízká. V signálu jsou jako pomalé oční artefakty označeny i jiné grafoelementy. Pro epileptickou aktivitu senzitivita dosáhla poměrně nízké hodnoty (SE=0,31). Lze tedy tvrdit, že ne všechna epileptická aktivita je v signálu správně odlišena. Velká část epileptické aktivity je zaměněna za jinou aktivitu (je přiřazena nesprávně k jinému shluku). Specificita dosáhla hodnoty 0,93, což značí, že epileptická aktivita byla správně odlišena od ostatních a pouze jen malá část grafoelementů příslušících jiné aktivitě byla označena za epilepsii. PPV dosáhlo hodnoty 0,39, což poukazuje na velké množství grafoelementů příslušících jiné aktivitě nežli epileptické, nesprávně označených jako epilepsie.

Metoda k-means v klasifikaci reálných dat lépe detekoval epileptické průběhy, než metoda k-means++. Pomalé oční artefakty byly lépe detekovány metodou k-means++. V praxi je požadováno u pacientů s podezřením na epilepsii odlišit epileptickou aktivitu. Na základě statistického vyhodnocení klasifikace reálných dat metodou k-means (tabulka 21) se ověřila vysoká přesnost detekce epileptické aktivity. Pro ideálnější výstup klasifikace reálných dat by bylo vhodné využít výhod obou metod k-means a k-means++, což bude objektem dalšího zkoumání.

6 Závěr

V této práci jsem se zabývala porovnáním metod k-means a k-means++. Nejprve jsem v prostředí MATLAB sestavila kód pro inicializaci dat, která přísluší metodě k-means++. Kód pro metodu k-means jsem sestavovat nemusela, jelikož je v prostředí MATLAB již realizován v podobě funkce. Funkčnost kódu k-means++ a sledování rozdílů oproti metodě k-means jsem provedla na 2D simulovaných datech. Simulovaná data jsem sestrojila pomocí matic a tyto matice jsem upravovala podle povahy obou metod tak, abych mohla detekovat rozdíly mezi výstupy obou metod. Tyto výstupy jsem statisticky vyhodnotila. Provedla jsem kvantitativní analýzu, kdy jsem sledovala časovou náročnost metod u simulovaných dat a kvalitativní analýzu, kdy jsem hodnotila účinnost metod u simulovaných dat. Po statistickém hodnocení jsem aplikovala metodu k-means++ na reálný EEG záznam. Výstup byl zhodnocen odborníkem. Metoda k-means je již využívaný pro klasifikaci reálných EEG dat prostřednictvím programu Wave-Finder. Vypracovala jsem ROC analýzu pro výstupy metod k-means a k-means++ aplikovaných na reálné záznamy EEG. Tímto jsem splnila veškeré body zadání mé bakalářské práce.

Časová náročnost se mezi metodami k-means a k-means++ nijak nelišila i přes skutečnost, že kód metody k-means++ je rozšířen o počáteční přepočet vzdáleností. Po provedení kvalitativní analýzy pro výstupy metod spuštěných na simulovaná data lze tvrdit, že simulovaná data byla lépe klasifikována metodou k-means++. Reálná data byla lépe klasifikována metodou k-means, který je již využíván pro automatickou klasifikaci signálu v programu Wave-Finder. Metoda k-means lépe detekovala v signálu epileptickou aktivitu, pomalé oční artefakty a fyziologickou aktivitu detekovala hůře. Naopak metoda k-means++ hůře detekovala epileptickou aktivitu, ale podstatně lépe oproti metodě k-means detekovala pomalé oční artefakty. Tato práce zjistila, že samostatná metoda k-means++ není vhodná pro použití pro automatickou klasifikaci reálného EEG signálu. Pro zlepšení automatické klasifikace signálu pomocí shlukové analýzy založené na vzdálenosti by se mohly výhody metod k-means a k-means++ sloučit.

Seznam použité literatury

[1] KRAJČA, Vladimír a Jitka MOHYLOVÁ. *Zpracování biologických signálů*. 1. Ostrava: Ediční středisko VŠB – TUO, 2006. ISBN 978-80-248-1491-9.

[2] KRAJČA, Vladimír a Jitka MOHYLOVÁ. *Číslicové zpracování neurofyziologických signálů*. 1. Praha: nakladatelství ČVUT, 2011. ISBN 978-80-01-04721-7.

[3] FREEMAN, Walter a Rodrigo QUIROGA. *Imaging Brain Function With EEG: Advanced Temporal and Spatial Analysis of Electroencephalographic Signals*. 1. New York: Springer, 2013. ISBN 978-1-4614-4983-6.

[4] UHLIARIK, Michal. *Zpracování a analýza EEG signálu*. Brno, 2013. Bakalářská práce. Vysoké učení technické v Brně. Vedoucí práce Karolína Lankašová.

[5] ALI JATOI, Munsif a Nidal KAMEL. *Brain Source Localizatiion Using EEG Signal Analysis*. 1. Boca Raton: CRC Press, 2018. ISBN 978-1-4987-9934-8.

[6] RAMACHANDRAN, Vilayanur. *Encyclopedia of the Human Brain*. 1. USA: Academic Press, 2002. ISBN 9780080548036.

[7] WALCZYSKO, Martin. *ANALÝZA EEG SIGNÁLU POMOCÍ ANALÝZY HLAVNÍCH KOMPONENT (PCA)*. Brno, 2008. Bakalářská práce. VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ. Vedoucí práce Ing. Milan Rychtárik.

[8] ROZMAN, Jiří. *Elektronické přístroje v lékařství*. 1. Praha: Academia, 2006. ISBN 80-200-1308-3.

[9] SHENG BAO, Forrest, Ya-Liang LI, Jue-Ming GAO a Jin HU. Performance of Dynamic Features in Classifying Scalp Epileptic Interictal and Normal EEG. *Annual International Conference of the IEEE EMBS*. 2010, **2010**(1), 6308-6311. DOI: 10.1109/IEMBS.2010.5628091.

[10] TUČKOVÁ, Jana. Algoritmy učení. TUČKOVÁ, Jana. *Učební text ČVUT FEL* [online]. 1. Praha: České vysoké učení technické v Praze, 2009, s. 15 [cit. 2017-11-09]. ISBN 9788001042298. Dostupné z: http://amber.feld.cvut.cz/ssc/ssc-cv/bpg.pdf

[11] WEITKUNAT, Rolf, ed. *Digital biosignal processing*. Amsterdam: Elsevier Science Publishers, 1991. Techniques in the behavioral and neural sciences, vol. 5. ISBN 0-444-89144-7.

[12] SALONI, , R.K. SHARMA, GUPTA a K. ANIL. Classification of High Blood Pressure Persons Vs Normal Blood Pressure Persons Using Voice Analysis. *Modern Education and Computer Science Press*. Hong Kong, 2013, **2013**(1), 47-52. DOI: http://dx.doi.org.ezproxy.techlib.cz/10.5815/ijigsp.2014.01.07. ISSN 20749074.

[13] MACHADO, José. Improved Shape Parameter Estimation in K Clutter with Neural Networks and Deep Learning. *Nternational Journal of Interactive Multimedia and Artificial Intelligence*. 2016, **2016**(7), 96-103. DOI: 10.9781/ijimai.2016.3715.

[14] BAGWARI, Pragya, Bhavya SAXENA, Meenu BALODHI a Vishwanath BIJALWAN. A New Method for Image Segmentation Based on Integration Technique. *International Journal of Interactive Multimedia and Artificial Intelligence*. Španělsko, 2017, **2017**(5), 58-64. ISSN 1989-1660.

[15] JAIN, Anil. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*. 2010, **2010**(318), 651-666. ISSN 167-8655.

[16] LIKAS, Astridis, Nikos VLAASSIS a Jakob VERBEEK. The global k-means clustering algorithm. *Pattern Recognition*. 2003, **2002**(326), 451-461.

[17] BAHMANI, Bahman. *Scalable K-Means++* [online]. 2012 [cit. 2017-12-30]. Dostupné z: https://www.youtube.com/watch?v=cigXAxV3XcY. Přednáška. Stanford University.

[18] ZAPLATÍLEK, Karel a Bohuslav DOŇAR. *MATLAB pro začátečníky*. 2. vyd. Praha: BEN - technická literatura, 2005. ISBN 80-730-0175-6.

[19] PIORECKÝ, Marek. *Automaticka klasikace EEG segmentu metodou DBSCAN*. Kladno, 2016. Diplomová práce. ČVUT FBMI.

[20] Kmeans. *MathWorks* [online]. 1994 [cit. 2017-11-09]. Dostupné z: https://www.mathworks.com/help/stats/kmeans.html#bueftl4-1

[21] MICHELI-TZANAKOU, Evangelia. *Supervised and Unsupervised Pattern Recognition: Feature Extraction and Compuational Intelligence*. 1. Boca Raton: CRC Press, 1999. ISBN 0-8493-2278-2.

[22] VRÁNOVÁ, J., J. HORÁK, K. KRÁTKÁ, M. HENDRYCHOVÁ a K. KOVAŘÍKOVÁ. ROC analýza a využití analýzy nákladů a přínosů k určení

optimálního dělícího bodu. *ČASOPIS LÉKAŘŮ ČESKÝCH*. 2009, **2009**(148), 410-415.

[23] ZVÁROVÁ, Jana. *Základy statistiky pro Biomedicínské obory*. 1. Praha: Nakladatelství Karolinum, 2007. ISBN 978-80-7184-786-1.

[24] TONG, Shanbao a Nitish V. THAKOR. *Quantitative EEG Analysis Methods and Clinical Applications*. 1. Norwood: Artech House, 2009. ISBN 978-1-59693-204-3.

A Seznam příloh na CD

Tabulka 23: Seznam příloh, které jsou k této práci dodány na CD

Seznam příloh na CD
Kód inicializace dat pro metodu k-means++ vytvořen v prostředí MATLAB2014b
Kód pro vytvoření simulovaných dat