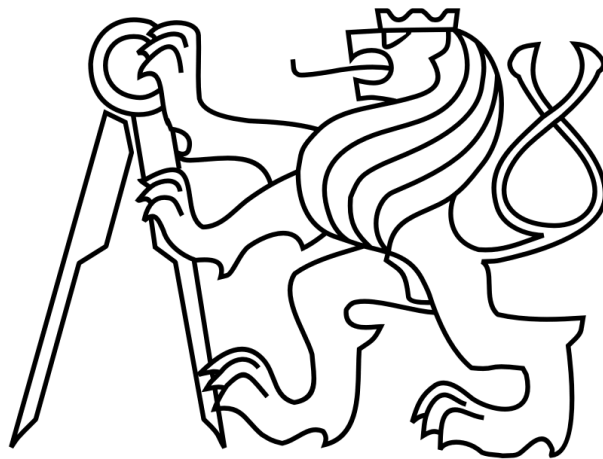


České vysoké učení technické v Praze

Fakulta elektrotechnická

Katedra kybernetiky



Diplomová práce

**Aplikace shlukovacích metod na časové řady se
zaměřením na záznamy FHR**

České vysoké učení technické v Praze
Fakulta elektrotechnická

Katedra kybernetiky

ZADÁNÍ DIPLOMOVÉ PRÁCE

Student: Bc. Tereza Janíčková
Studijní program: Biomedicínské inženýrství a informatika (magisterský)
Obor: Biomedicínské inženýrství
Název tématu: Aplikace shlukovacích metod na časové řady se zaměřením na záznamy FHR

Pokyny pro vypracování:

1. Seznamte se s problematikou hypoxie plodu a jejího vlivu na srdeční frekvenci.
2. Seznamte s metodami pro shlukování časových řad.
3. Navrhněte a implementujte vybrané metody.
4. Pomocí zvolených metod analyzujte simulované a reálné signály srdeční frekvence plodu.
5. Výsledky vyhodnoťte vzhledem k metabolické acidóze (měřeno pomocí pH) a klinickému ohodnocení.

Seznam odborné literatury:

- [1] Lin, J.; Li, Y.: Finding structurally different medical data, IEEE Computer Society, Washington DC, USA, 2009.
[2] Jain, M.K.: Numerical Methods For Scientific And Engineering Computation, New Age International, 2003.

Vedoucí diplomové práce: Ing. Václav Chudáček, Ph.D.

Platnost zadání: do konce letního semestru 2014/2015

L.S.

doc. Dr. Ing. Jan Kybic
vedoucí katedry

prof. Ing. Pavel Ripka, CSc.
děkan

V Praze dne 10. 1. 2014

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

V Praze dne 12. 05. 2014

.....

Podpis autora práce

Poděkování

Chtěla bych poděkovat Ing. Václavu Chudáčkovi, Ph.D. za vedení diplomové práce a za pomoc, rady a připomínky. Dále bych chtěla poděkovat své rodině a přátelům za podporu během studia.

Abstrakt

Tato práce se zabývá analýzou záznamů fetální tepové frekvence (FHR), které byly naměřeny během kardiokografického vyšetření. KTG je v současné době nejběžnější metodou určenou k diagnostice fetální hypoxie. Interpretace těchto záznamů je velmi složitá, díky čemuž se objevuje velká variabilita jejich hodnocení mezi lékaři. Proto se snažíme vytvořit systém, který by jejich klasifikaci zlepšil. FHR záznamy jsme v této práci klasifikovali pomocí shlukové analýzy. K dispozici máme data z databáze CTU-UHB, která obsahuje 552 reálných signálů. Pro získání příznaků vyjadřujících signál, které vstupují do shlukování, používáme metody založené na symbolické a lineární aproximaci. Hierarchickým shlukováním a algoritmem k-means signály klasifikujeme do skupin. Dva shluky odpovídají klasifikaci na normální a patologické signály. Klinicky se signály rozdělují do tří skupin, nejlepších výsledků ale dosáhneme při použití více shluků pro každou skupinu. Výsledek porovnáváme s klasifikací podle hodnoty pH. Při použití více než dvou shluků jsme dosáhli výsledku 63.4% senzitivity a 71% specificity. Úspěšnost klasifikátoru je tak srovnatelná s jinými metodami.

Klíčová slova: srdeční frekvence plodu, analýza, shlukování, časové řady

Abstract

This thesis deals with analysis of fetal heart rate recordings which were measured using cardiotocograph. CTG is currently the most common diagnostic method of fetal hypoxia. Interpretation of these recording is very complex resulting in a large variability of evaluation among doctors. We want to create a system that would help us improve CTG signals evaluation. In this thesis FHR signals are classified using cluster analysis. Analyzed signals are from CTU-UHB database which contains 552 CTG recordings. Methods of symbolic and piecewise linear approximation were used for feature extraction. We classify signals into two groups by k-means algorithm and hierarchical clustering. Two clusters correspond with classification to normal and pathological signal. Clinical evaluation uses three groups, but we get the best solution with more clusters. Results are compared with pH measurement. We get 63.4% sensitivity and 71% specificity with six clusters as the best result. The success of classification is comparable with other methods.

Key words: fetal heart rate, analysis, clustering, time-series

Obsah

Seznam zkratek	IX
Seznam symbolů	X
Seznam obrázků	XI
Seznam tabulek	XIII
1. Úvod.....	1
1.1. Cíle práce	2
2. Medicínský úvod.....	3
2.1. Fetální fyziologie	3
2.1.1. Fetální oběh.....	3
2.1.2. Metabolismus buňky.....	4
2.2. Diagnostika hypoxie.....	6
2.2.1. Kardiotokografie	6
2.2.2. Biofyzikální profil plodu	7
2.2.3. Fetální pulsní oximetrie	7
2.2.4. Fetální EKG	7
2.3. Hodnocení novorozence.....	8
2.3.1. Apgar skóre.....	8
2.3.2. ABR	8
3. Shlukování časových řad	9
3.1. Diskretizace časových řad.....	9
3.1.1. Fourierova transformace	10
3.1.2. PAA	10
3.1.3. PLA	11
3.1.4. SAX	12
3.1.5. Komprese algoritmem Lempel-Ziv	13
3.2. Míry podobnosti	14

3.2.1.	Euklidovská vzdálenost	14
3.2.2.	DTW	15
3.2.3.	Pearsonův korelační koeficient	15
3.3.	Metody shlukování	16
3.3.1.	Hierarchické shlukování	16
3.3.2.	K-means	17
3.3.3.	Modelově založené algoritmy	17
3.3.4.	Metody založené na hustotě objektů	18
3.4.	Hodnocení kvality shluků	18
4.	Metody analýzy srdeční frekvence	19
4.1.	Statistické metody	19
4.2.	Lineární a nelineární metody	20
4.3.	Waveletová transformace	20
4.4.	Gaussovské modelování	20
4.5.	Partial least squares discriminant analysis	20
4.6.	Phase-rectified signal averaging	21
5.	Databáze signálů	22
6.	Předzpracování dat	23
6.1.	Odstranění artefaktů a zkrácení signálu	23
6.2.	Rozdělení a seřazení signálů	24
7.	Diskretizace signálů	26
7.1.	Rozdělení signálu na okna	26
7.2.	Definice hranic intervalů	27
7.3.	Reprezentace	28
7.3.1.	PAA	28
7.3.2.	SAX	29
7.3.3.	PLA	30

7.3.4.	APCA.....	31
7.3.5.	APLA.....	32
7.4.	Eroze	33
7.5.	Kompresie algoritmem Lempel-Ziv	34
8.	Porovnávání podobnosti	36
8.1.	Euklidovská vzdálenost.....	36
8.2.	Frekvenční analýza.....	37
8.3.	Korelační analýza.....	40
9.	Shlukování	42
9.1.	Shlukování metodou k-means	42
9.1.1.	Shlukování aproximací signálů.....	43
9.1.2.	Shlukování signálů po kompresi LZ77	50
9.1.3.	Shlukování strukturních příznaků	51
9.2.	Hierarchické shlukování.....	53
9.2.1.	Shlukování aproximací	53
9.2.2.	Shlukování prvků slovníku LZ a strukturních příznaků	56
9.3.	Závislost výsledků na parametrech	56
9.3.1.	Závislost na délce segmentu	57
9.3.2.	Závislost na počtu intervalů	58
9.3.3.	Závislost na počtu shluků	60
10.	Závěr	61
	Použitá literatura	63
	Zdroje obrázků.....	69

Seznam zkratek

- ABR – Acidobazická rovnováha
- APCA – Adaptive piecewise constant approximation
- ASCII - American standard code for information interchange
- ATP - Adenosintrifosfát
- BE – Base excess
- CNS – Centrální nervová soustava
- DBSCAN - Density-based spatial clustering of applications with noise
- DFT – Diskrétní Fourierova transformace
- DTW – Dynamic time warping
- EEG – Elektroencefalografie
- EKG – Elektrokardiografie
- EM – Expectation-maximization
- FDA – Functional data analysis
- FHR – Fetal heart rate
- FKG – Fonokardiografie
- FPO – Fetální pulsní oximetrie
- HRV – Heart rate variability
- KTG – Kardiotografie
- LF, HF – Low fequency, high frequency
- LPA – Linear piecewise approximation
- LSF – PAA – Linear statistical feature based piecewise aggregate approximation
- LTI - Long-term irregularity index
- LZ77, LZ78, LZW – Lempel-Ziv 77, Lempel-Ziv 78, Lempel-Ziv-Welch
- PAA – Piecewise aggregate approximation
- PCA – Principal component analysis
- PLR – FP – Piecewise linear representation based on feature points
- PLS – DA - Partial least squares discriminant analysis

PRSA - Phase-rectified signal averaging

REM – Rapid eye movement

RNA – Ribonukleová kyselina

SAX – Symbolic approximation

SIFT – Scale invariant feature transform

TNR – True negative rate

TPR – True positive rate

Seznam symbolů

e – Eulerovo číslo

j – imaginární jednotka

ρ – korelační koeficient

ω – úhlová frekvence

Seznam obrázků

Obrázek 1 Fetální oběh [1]	4
Obrázek 2 Metabolismus buňky [2].....	5
Obrázek 3 Kardiotokografický záznam [2]	6
Obrázek 4 Shlukování [3]	9
Obrázek 5 Diskretizace časových řad [4]	12
Obrázek 6 SAX [4]	12
Obrázek 7 Porovnání Euklidovské vzdálenosti a DTW	15
Obrázek 8 Dendrogram [5]	16
Obrázek 9 K-means [6].....	17
Obrázek 10 Filtrace signálu, červeně vyznačeny interpolované segmenty	24
Obrázek 11 Obsah databáze.....	25
Obrázek 12 Přesnost aproximace PAA.....	26
Obrázek 13 Histogram dat a hranice intervalů	27
Obrázek 14 Hraniční úhly intervalů.....	28
Obrázek 15 Reprezentace pomocí PAA	29
Obrázek 16 Reprezentace pomocí SAX	30
Obrázek 17 Reprezentace pomocí PLA.....	31
Obrázek 18 Reprezentace pomocí APCA.....	32
Obrázek 19 Reprezentace pomocí APLA	33
Obrázek 20 Eroze signálu s vyznačením odstraněných úseků	34
Obrázek 21 Histogram četností znaků ze slovníku v datech	35
Obrázek 22 Vzdálenost signálů po aproximaci PAA	36
Obrázek 23 Vzdálenost signálů po aproximaci PLA.....	37
Obrázek 24 Četnost jednotlivých znaků – SAX	38
Obrázek 25 Četnosti dvojice znaků – SAX	38
Obrázek 26 Místa vyhodnocená jako decelerace.....	39
Obrázek 27 Korelační koeficienty PAA	40
Obrázek 28 Korelační koeficienty – PLA.....	41
Obrázek 29 Výsledek k-means pro metody PAA a SAX	44
Obrázek 30 Výsledek k-means pro metody PAA a SAX- korelace	45
Obrázek 31 Výsledek k-means pro metody PAA a SAX - 6 shluků	46
Obrázek 32 Výsledek k-means pro metodu PLA	47

Obrázek 33 Výsledek k-means pro metodu PLA - 6 shluků	48
Obrázek 34 Výsledek k-means pro metodu LZ - korelace	51
Obrázek 35 Výsledek k-means pro metodu Bag - of - Words.....	53
Obrázek 36 Diagram pro metodu PAA - euklidovská vzdálenost.....	54
Obrázek 37 Diagram pro metodu PAA – korelace	55
Obrázek 38 Diagram pro metodu APLA – korelace.....	56
Obrázek 39 Diagram pro metodu PLA – korelace.....	56
Obrázek 40 Diagram - Bag - of – Words.....	56
Obrázek 41 Diagram pro metodu LZ - korelace.....	56
Obrázek 42 Závislost úspěšnosti klasifikace na délce okna	57
Obrázek 43 Závislost úspěšnosti klasifikace na velikosti chyby	58
Obrázek 44 Závislost úspěšnosti klasifikace na počtu intervalů	59
Obrázek 45 Závislost úspěšnosti klasifikace na počtu intervalů	59
Obrázek 46 Závislost úspěšnosti klasifikace na počtu shluků	60

Seznam tabulek

Tabulka 1 Rozdělení podle hodnot pH	8
Tabulka 2 Vlastnosti metriky.....	14
Tabulka 3 Počty poklesů v úrovních pro jednotlivé metody	39
Tabulka 4 Kontingenční tabulka pro metody PAA a SAX.....	44
Tabulka 5 Kontingenční tabulka pro metody PAA a SAX - 6 shluků.....	46
Tabulka 6 Kontingenční tabulka pro metodu PLA.....	47
Tabulka 7 Kontingenční tabulka pro metodu PLA - 6 shluků.....	48
Tabulka 8 Výsledek klasifikace pro metodu APCA	49
Tabulka 9 Výsledek klasifikace pro metodu APLA	50
Tabulka 10 Kontingenční tabulka pro metodu LZ.....	50
Tabulka 11 Výsledek klasifikace pro metodu LZ.....	51
Tabulka 12 Kontingenční tabulka pro metodu Bag of Words	52
Tabulka 13 Kontingenční tabulka pro metodu PAA	54

1. Úvod

Porodnictví a péče o novorozence v České republice zaznamenaly v uplynulých letech výrazný pokrok a patří mezi nejlepší na světě, což dokazuje i nízká novorozenecká úmrtnost a snížený výskyt postižení novorozence v důsledku porodu (Kantor, 2013). Plod je během porodu vystaven stresu a nedostatku kyslíku, které je za normálních okolností schopen zvládnout. Zdravý plod má několik obranných mechanismů, pokud však dojde k jejich vyčerpání, plod už nedokáže kompenzovat snížený přísun kyslíku a dochází k hypoxickému poškození orgánů, především CNS. Trvalými následky jsou např. mentální retardace, epilepsie a dětská mozková obrna a v krajních případech může nastat i smrt (Ježová, 2014).

Z těchto důvodů je důležitá prevence spočívající hlavně v diagnostickém sledování plodu během těhotenství i porodu. Historicky první metodou monitorování srdeční frekvence plodu byla auskultace pomocí stetoskopu. Dnes je nejběžnějším typem vyšetření používanému k tomuto účelu kardiokografie (KTG). Poskytuje záznam změn tepové frekvence plodu v čase a dokáže tak odhalit hypoxii. Kardiokografie sice snížila incidenci záchvatů křečí novorozenců, ovšem kontinuální sledování plodu nepřispělo k snížení úmrtnosti a výskytu postižení oproti starší auskultační metodě (Janků, 2007).

Hodnocení KTG křivky provádí obvykle zkušený lékař, interpretace takto komplexního záznamu je velice složitá a značně subjektivní. U nejasných nálezů je možné použít objektivní analýzu hodnoty pH z krve plodu. Pro zlepšení vyhodnocení fetální srdeční frekvence byly zavedeny směrnice (FIGO, 1985). Spolu s tím se začaly rozvíjet metody automatického hodnocení KTG záznamů jako podpora hodnocení lékaře. Nejčastějšími způsoby jsou např. statistický popis nebo frekvenční a waveletová analýza (Chudáček a spol., 2010).

V této práci budeme KTG signály klasifikovat pomocí shlukování. Shluková analýza se již používá pro klasifikaci biologických signálů jako např. EKG (Pospíšil a spol., 2013), EEG (Horniak, 2010) nebo třeba FKG (Amit a spol., 2009). Největší výhodou shlukování je, že patří mezi algoritmy s učením bez učitele. Klasifikátor tedy

nemusí u vstupních dat znát výsledek klasifikace. Shluková analýza se aplikuje na prvky popsané skupinou příznaků. Pro výsledek shlukování je tedy klíčové správně extrahovat příznaky z dat. Podstatou shlukování je rozdělení daných signálů do několika skupin, v případě klinického hodnocení je to rozdělení na fyziologické, suspektní a patologické. Pro nás bude dostačující rozdělení do dvou skupin, na signály normální a patologické. Existuje několik druhů hodnocení novorozence, my pokládáme za rozhodující klasifikaci podle hodnoty pH.

1.1. Cíle práce

Cílem této práce je seznámit se s fyziologií plodu a problematikou hypoxie, dále zpracovat přehled o metodách používaných k analýze KTG signálů a zhodnotit je. Budeme se zabývat použitím shlukování časových řad ke klasifikaci signálů. Hlavním cílem je implementovat klasifikátor KTG signálů, který by zlepšil stávající interpretaci FHR záznamů. Tento klasifikátor použijeme k analýze reálných signálů z databáze. Výsledky z klasifikátoru budeme porovnávat s hodnotou pH, podle které se hodnotí výsledek porodu.

2. Medicínský úvod

Základy fetální fyziologie jsou nutné pro pochopení dějů probíhajících v organismu plodu při porodu a jejich vlivu na srdeční frekvenci. Plod je vystaven nadměrnému stresu, na který různým způsobem reaguje. Během kontrakcí dělohy dochází k omezení přísunu kyslíku, které zdravý plod kompenzuje pomocí obranných mechanismů (Sundström, 2000). Při vyčerpání těchto mechanismů dochází k rozvoji hypoxie, která při delším trvání přechází v asfyxii. Důsledkem je ischemické poškození CNS, které může vést až ke smrti. Asfyxie je jednou z nejčastějších příčin úmrtí nebo těžkého poškození novorozence. Je důležité průběžně sledovat, jak se plod se zhoršenými podmínkami vyrovnává. O stavu plodu nejlépe vypovídá průběh srdeční frekvence v závislosti na děložních kontrakcích získaný při KTG vyšetření. (Trojan, 2003)

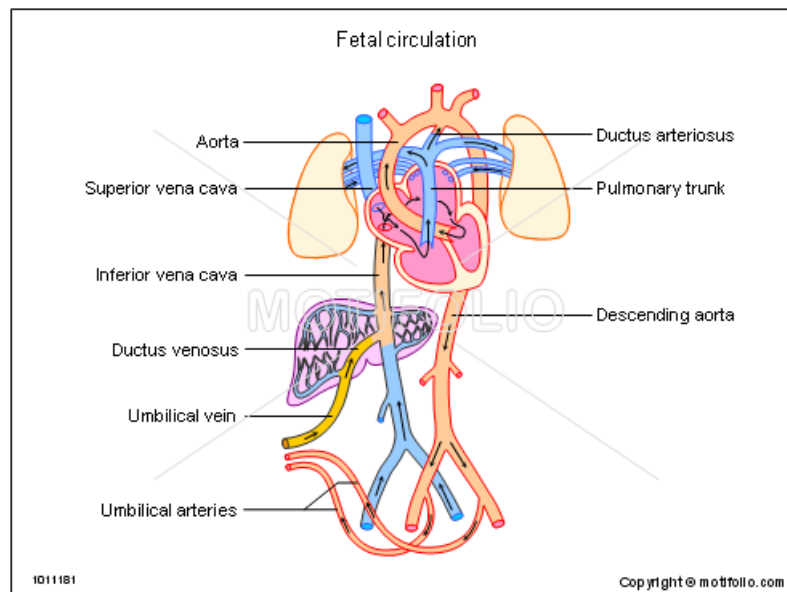
2.1. Fetální fyziologie

Těhotenství trvá průměrně 40 týdnů, během této doby dochází k růstu a vývoji plodu. Fetální období začíná 9. týdnem, následuje po období embryonálním. V tuto dobu už je vytvořena většina orgánů. Kardiovaskulární systém se začíná vyvíjet už ve 3. týdnu těhotenství a srdce začíná tepat ve 4. týdnu. Tepová frekvence je přibližně 155 tepů/min a před porodem klesá na 140 tepů/min. Oproti dospělému krevnímu oběhu má ten fetální několik specifíků, především přítomnost cévních zkratů. (Trojan, 2003)

2.1.1. Fetální oběh

Během těhotenství je plod zásoben mateřskou krví, důležitým orgánem pro krevní oběh plodu je placenta. Zajišťuje výměnu dýchacích plynů mezi plodem a matkou. Mateřská okysličená krev přichází z aorty přes iliakální arterie do uterinních arterií, které předávají krev placentě. Plod je s placentou spojen pupečnickem, který se dělí na jednu umbilikální věnu a dvě arterie. Krev přitéká vénou do dolní duté žíly plodu, kde se mísí s odkysličenou žilní krví. V srdci plodu jsou síně spojeny foramen ovale, krev z pravé síně se tedy přesune do levé a odtud přes komoru do aorty. Přednostně se zásobí mozek a srdce, naopak dolní část těla dostává krev s menším obsahem kyslíku. Ve fetálním oběhu existuje kromě foramen ovale několik dalších

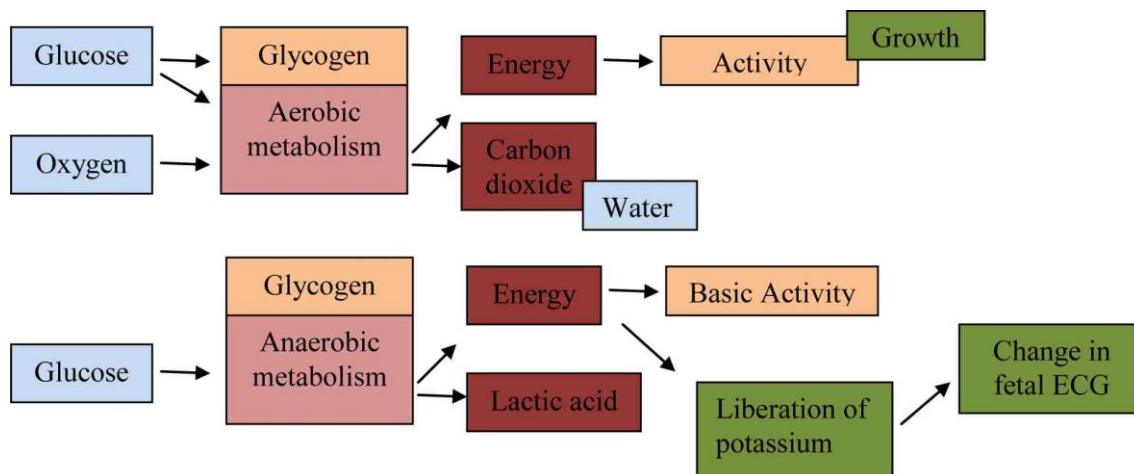
zkratů, díky kterým se obcházejí nefunkční plíce a částečně játra. Zpět do placenty se krev vrací umbilikálními arteriemi. (Trojan, 2003, Sundström, 2000)



Obrázek 1 Fetální oběh [1]

2.1.2. Metabolismus buňky

Buňky plodu získávají energii z přiváděné krve bohaté na živiny. Kyslík a glukóza vstupují v mitochondriích do reakce, při které se syntetizuje ATP sloužící jako zdroj energie. Vedlejším produktem je CO_2 a voda. Tento způsob vytváření energie se nazývá aerobní dýchání a probíhá za normálních okolností. Při dostatečném přísunu kyslíku mají buňky dostatek energie pro plnění své funkce. ATP může vznikat ze zásob glukózy i bez přísunu kyslíku, pak je odpadním produktem kyselina mléčná. Tomu se říká anaerobní dýchání a poskytuje buňkám energii pouze k udržení bazálního metabolismu. (Sundström, 2000)



Obrázek 2 Metabolismus buňky [2]

V případě nedostatečné dodávky kyslíku tkáním plodu, nastávají postupně následující tři úrovně deficitu. (Sundström, 2000)

Hypoxémie

Prvotní fáze kyslíkové nedostatečnosti je hypoxémie, projevuje se sníženou saturací krve. Jde o nezávažný stav, který nemá vliv na funkci orgánů. Nízká saturace je zaznamenána chemoreceptory umístěnými v cévách plodu. Reakcí na to je efektivnější využití kyslíku. Dále může dojít k snížení aktivity, což se projeví jako pokles pohybů plodu a omezení růstu. Tuto situaci může plod zvládat po dobu několika dní až týdnů. Rizikem je ale nedostatečná schopnost plodu vyrovnat se s případnou akutní hypoxií.

Hypoxie

Hypoxie následuje jako další fáze při ještě větším poklesu saturace. Obranné mechanismy aktivované při hypoxémii už nestačí pokrýt energetický deficit. Hypoxie ovlivňuje nejenom arteriální krev, ale i orgány plodu. Dochází k uvolnění stresových hormonů a k centralizaci krevního oběhu. To zajistí zásobení důležitých orgánů (mozek, srdce) na úkor periferie, kde začne probíhat anaerobní metabolismus. Hypoxii může plod kompenzovat po dobu až několika hodin.

Asfyxie

Posledním stadiem je asfyxie, při které dodávka kyslíku nestačí ani k zásobení centrálních orgánů a hrozí jejich selhání. Saturace krve je velmi nízká a dochází k anaerobnímu metabolismu i u důležitých orgánů. Uvolňují se další stresové hormony

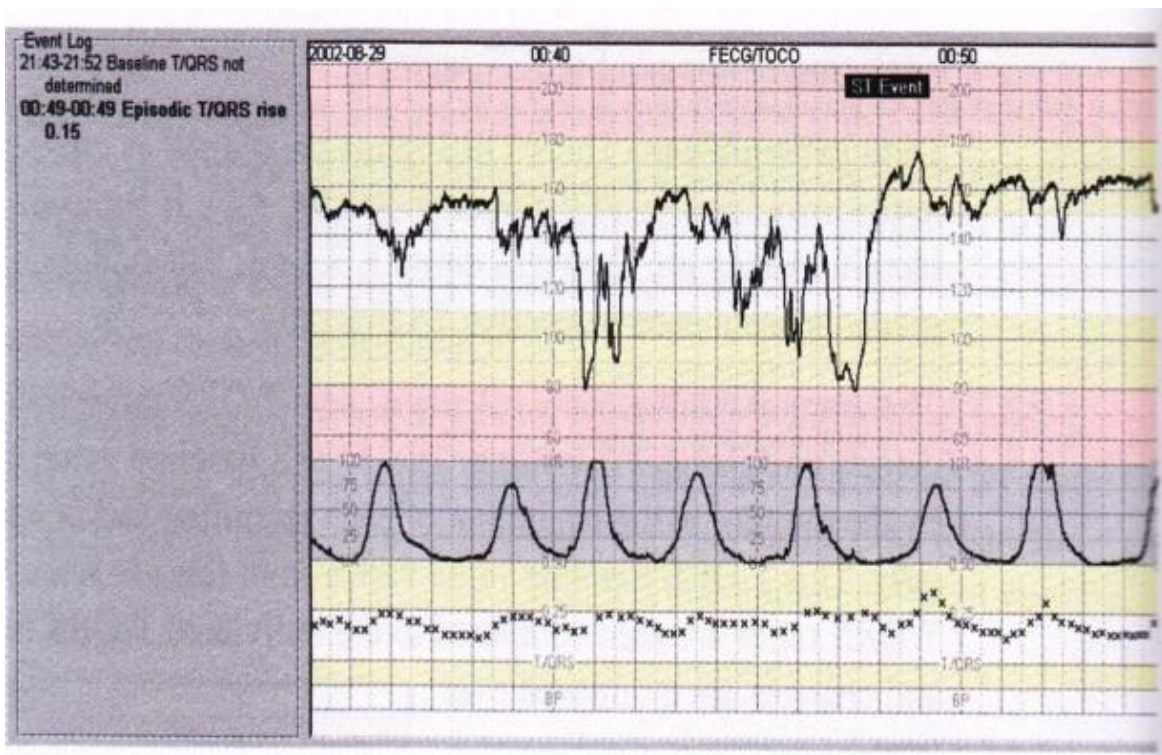
a aktivuje se sympatický nervový systém. Tento stav je plod schopen zvládnout pouze několik minut.

2.2. Diagnostika hypoxie

Odhalení hypoxie plodu je poměrně obtížné, existuje řada diagnostických metod prováděných před a během porodu. Každá metoda má své výhody a optimálních výsledků lze dosáhnout kombinací více vyšetření (Valuchová, 2012).

2.2.1. Kardiotokografie

Kardiotokografie je nejběžnějším typem vyšetření pro diagnostiku hypoxie plodu. Pomocí elektrod (zevní, vnitřní) se snímá srdeční frekvence plodu a zároveň nitroděložní tlak v závislosti na čase. Výsledkem jsou tedy dvě křivky. Metoda se vyznačuje vysokou senzitivitou a nízkou specifitou. To znamená, že počet falešně negativních výsledků je nízký, ale objevuje se vysoký počet falešně pozitivních nálezů (Novotná, 2001).



Obrázek 3 Kardiotokografický záznam [2]

V této práci se budeme zabývat pouze analýzou srdeční frekvence plodu. Na FHR záznamu můžeme hodnotit několik parametrů. (Papírníková, 2006)

Dlouhodobé frekvenční jevy

Nejdůležitější charakteristikou je základní tepová frekvence, která se sleduje po dobu 5-10 min. V normě je 110-150 tepů/min. Nižší tepová frekvence se označuje jako bradykardie a má různé stupně závažnosti. Stejně tak vysoká frekvence – tachykardie.

Střednědobé frekvenční jevy

Dále se hodnotí přechodné změny tepové frekvence trvající od 15s do 3 min. Akcelerace je definována jako přechodné zvýšení tepové frekvence o alespoň 15 tepů/min. Vzniká buď při pohybech plodu jako výraz dobrého stavu nebo v reakci na děložní kontrakce. Decelerace je naopak charakterizována jako dočasné snížení tepové frekvence. Fyziologicky vzniká při podráždění vagu, nežádoucí jsou decelerace závislé na kontrakcích, kdy dochází k sníženému průtoku krve pupečnickem.

Krátkodobé frekvenční jevy

Variabilita srdeční frekvence se projevuje jako oscilace kolem bazální frekvence. Normální variabilita má amplitudu 5-25 tepů/min. Při snížené variabilitě mluvíme o pásmu silentním, které je příznakem těžké hypoxie. Naopak oscilace vyšší než 25 tepů/min se označují jako pásmo saltatorní a jsou charakteristické pro kompenzační mechanismy plodu.

2.2.2. Biofyzikální profil plodu

Vyšetření se provádí kombinací KTG a ultrazvuku. Sleduje se několik parametrů chování plodu (aktivní pohyby, dýchací pohyby, tonus...) a každý parametr se bodově ohodnotí. (Novotná, 2001)

2.2.3. Fetální pulsní oximetrie

Poměrně přesnou metodou měření fetálního diskomfortu je FPO. Využívá rozdílné absorpce oxyhemoglobinu a deoxyhemoglobinu. Sleduje se tím hodnota saturace krve, která je u plodu nižší než u dospělého člověka (asi 60%). Za patologické se považuje saturace pod 30%. (Novotná, 2001)

2.2.4. Fetální EKG

Tato metoda patří mezi nejmodernější přístupy. EKG plodu lze snímat buď invazivně skalpovou elektrodou nebo neinvazivně přes břišní stěnu. Ke zjištění hypoxie se analyzuje hlavně ST úsek křivky. (Novotná, 2001, Janků, 2007)

2.3. Hodnocení novorozence

Výše uvedené metody slouží k diagnostice antepartální, po porodu lze hypoxii odhalit na základě určitých příznaků novorozence.

2.3.1. Apgar skóre

Apgar skóre je systém bodování novorozence v první, páté a desáté minutě života. Hodnotí se následující příznaky: puls, dech, svalové napětí, reakce na podráždění a barva kůže. Maximální hodnocení je 10-10-10 (Ahmadpour-Kacho, 2010).

2.3.2. ABR

Důležitým parametrem pro diagnostiku hypoxie je acidobazická rovnováha. Po porodu se provádí odběr krve z pupečnicku a z ní se určí hodnota pH. Analýza pH krve je spolehlivý ukazatel míry hypoxie a hodnota pH souvisí s hodnocením Apgar skóre. Acidóza může vzniknout respiračně, nebo metabolicky. U hodnocení novorozence je klíčová metabolická složka acidózy, kterou vyjadřuje hodnota přebytku bází (BE). (Ahmadpour-Kacho, 2010, Chudáček a spol., 2014)

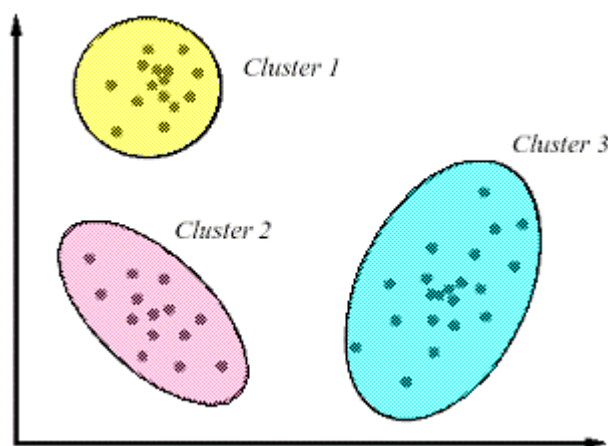
Tabulka 1 Rozdělení podle hodnot pH

Normální pH	>7.2
Abnormální pH	$7.05 - 7.2$
Patologické pH (acidóza)	< 7.05

3. Shlukování časových řad

Časová řada je soubor hodnot naměřených v průběhu času. Poskytuje informace o změnách veličiny v čase. Analýza časových řad je v současné době v popředí zájmu, na časové řady totiž můžeme narazit v nejrůznějších oborech, např. v medicíně, meteorologii, geologii.... Cílem analýzy je získat z dat co nejvíce informací, nalézt v časových řadách určitou vnitřní strukturu a případně předpovědět jejich budoucí chování. V této práci budeme za časovou řadu považovat KTG signál. (Meesrikamolkul a spol., 2012)

Shlukování se ukázalo být velice efektivní metoda pro získání informací. Cílem shlukování je rozdělit data do oddělených shluků tak, že data uvnitř shluku jsou si co nejvíce podobná a data z rozdílných shluků jsou co nejvíce odlišná. Shlukování bylo původně určeno pro statická data, lze ho ale použít i na časové řady. Umožňuje nalézt podobné signály a přiřadit je od stejné skupiny. Výhodou shlukování je, že jde o učení bez učitele, nemusíme tedy znát předem výsledek klasifikace dat. (Waren Liao, 2005)



Obrázek 4 Shlukování [3]

3.1. Diskretizace časových řad

Časové řady obsahují množství nadbytečných informací. Pro získávání znalostí z řad je důležité hledání podobnosti mezi zkoumanými řadami. Porovnávání řad

v originálním tvaru je časově náročné, data je tedy nutné zredukovat a použít jen některé příznaky. Cílem komprese je převést data s vysokou dimenzí do nízké dimenze. Tato redukce výrazně zefektivňuje práci s časovými řadami. Používá se celá řada kompresních metod (Zhu a spol., 2007).

3.1.1. Fourierova transformace

Jednou z nejzákladnějších metod ve zpracovávání signálů je Fourierova analýza. Převádí časovou oblast do frekvenční. Pro signály se používá především diskrétní Fourierova transformace (DFT). Časová řada je transformací rozložena na součet harmonických signálů.

$$X(\omega) = \int x(t)e^{-j\omega t} dt$$

Toho lze využít pro odstranění artefaktů ze signálu a pro kompresi. Redukce se provede použitím jen několika nejdůležitějších Fourierových koeficientů. Ztráta informace se zmenšuje s použitím více koeficientů. To ale znamená také méně efektivní kompresi. (Olver, 2014)

3.1.2. PAA

PAA (Piecewise Aggregate Approximation) je nejjednodušší metoda redukce dimenze. I přesto může konkurovat složitějším transformacím, jako např. DFT. PAA vyjadřuje časovou řadu jako sled stejně dlouhých segmentů. Signál se rozdělí na okna a v každém okně se vypočítá statistická charakteristika (průměr, medián...). Toto číslo pak reprezentuje celý segment. Délka segmentu určuje kvalitu aproximace, krátké okno lépe kopíruje signál, ale nedostatečně komprimuje. (Guo a spol., 2010)

Při tomto způsobu reprezentace dochází ke ztrátě informace, existuje několik vylepšených algoritmů. Jedním z nich je APCA, který patří mezi adaptivní metody reprezentace dat. Používá různě dlouhé segmenty podle charakteru časové řady, každý segment je pak vyjádřen dvěma čísly, průměrem segmentu a jeho délkou (Chakrabarti a spol., 2002). Dalším způsobem je LSF-PAA, který lineárně kombinuje dva statistické příznaky, průměr a varianci (Guo a spol., 2010). Tyto algoritmy jsou časově náročnější na výpočet než klasické PAA.

Aproximace pomocí PAA se navzdory své jednoduchosti objevuje v mnoha aplikacích. Angeles-Yreta a spol. (2004) pomocí PAA diskretizuje seismologické

signály a DTW algoritmem měří jejich podobnost. Ohsaki a spol. (2007) používá PAA s nepravidelným vzorkováním pro získání vzorů z klinických dat (rozvoj symptomů hepatitis). Gacek (2013) metodou PAA reprezentuje EKG signály a následně je shlukuje algoritmem fuzzy c-means.

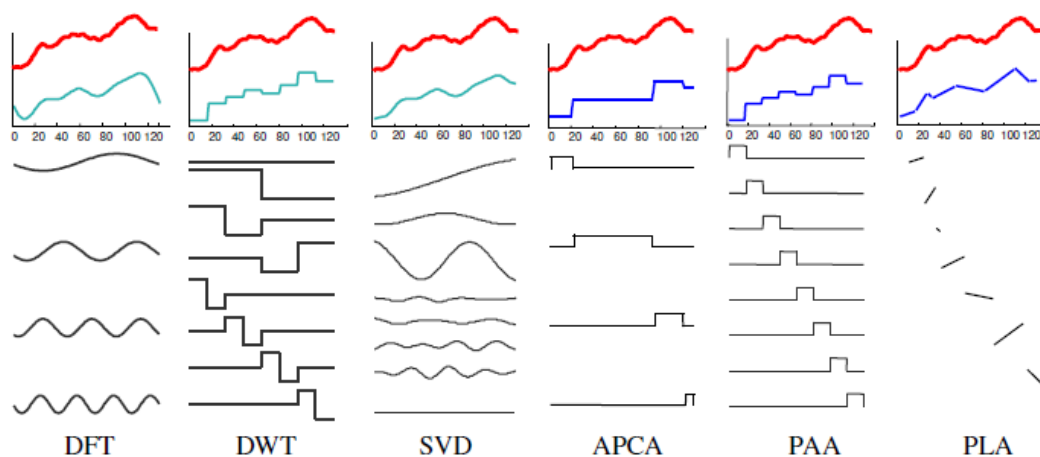
3.1.3. PLA

PLA (Piecewise Linear Approximation) je jedna z nejpoužívanějších reprezentací a poskytuje přesnější vyjádření časové řady než předchozí metoda. Stejně jako PAA rozděluje řadu na stejně dlouhé segmenty, ale charakteristikou segmentu je v tomto případě sklon přímky proložené daty. Na rozdíl od PAA tedy dokáže vystihnout trend v datech. Aproximace umožňuje dosáhnout jakékoli přesnosti zvýšením počtu segmentů. (Zhu a spol., 2007)

PLA může získat reprezentační přímky dvěma způsoby. Lineární interpolace vytvoří úsečku jednoduše procházející krajními body segmentu. Lineární regrese prokládá úsečku segmentem tak, aby byla minimalizována čtvercová vzdálenost úsečky a bodů v segmentu. (Keogh a spol., 2001)

Podle Keogha a spol. (2001) se adaptivní metody převodu časových řad do PLA nazývají segmentační algoritmy. Přestože existuje mnoho přístupů, lze je zařadit do některé ze tří skupin. Metoda posuvných oken postupně zvětšuje segment a počítá chybu aproximace oproti původním datům. Při dosažení předem dané maximální chyby se růst segmentu zastaví a postup se opakuje pro další segment. Algoritmy Top-down a Bottom-up postupně rozdělují nebo spojují časové řady, dokud není dosaženo ukončovacího kritéria. Tyto algoritmy tedy vytváří různě dlouhé segmenty, které lépe charakterizují časovou řadu.

Keogh a spol. (2001) porovnává efektivnost těchto algoritmů na datech z různých oblastí, např. záznamy hladiny vody, radiové vlny nebo EKG signály. Zhu a spol. (2007) představuje vylepšenou metodu reprezentace založenou na příznakových bodech. PLR-FP se aplikuje na periodické signály s velkou variabilitou vzoru během period. Tím se následně analyzují záznamy denního průtoku vody v řece. López-Avitia a spol. (2010) používá PLA pro zjištění délky QRS komplexu v EKG signálech a tím zjištění variability srdeční frekvence.

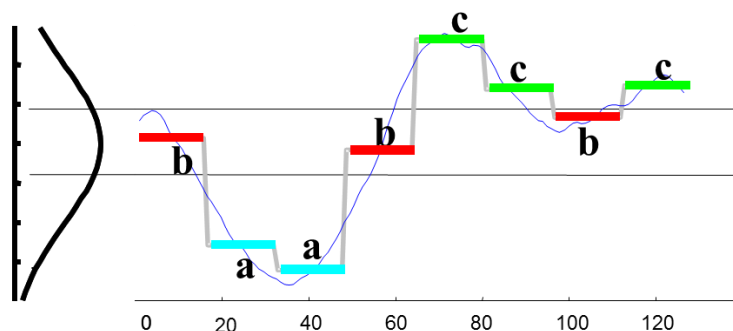


Obrázek 5 Diskretizace časových řad [4]

3.1.4. SAX

SAX (Symbolic Approximation) je hlavní metodou symbolického vyjádření časových řad a v současné době je často využíváný v analýze časových řad. Splňuje požadavky na dostatečnou redukci dimenze řad i na možnost výpočtu vzdálenosti mezi řadami. Reprezentace pomocí SAXu výrazně zrychluje proces získávání znalostí z dat, přitom zachovává přesnost výsledků. SAX vychází z PAA, ale místo čísla reprezentující každý segment, vyjadřuje signál řada symbolů. (Sun a spol., 2013, Lin a spol., 2003)

Pro získání symbolické reprezentace časové řady nejprve provedeme transformaci metodou PAA. Každý segment může být vyjádřen reálným číslem, je proto nutné tyto úrovně diskretizovat. Z histogramu časové řady získáme hranice intervalů. Hranice můžeme nalézt i ze statistických tabulek normálního rozdělení. Symbol přiřadíme segmentu podle toho, do jakého intervalu PAA reprezentace spadá. (Lin a spol., 2003)



Obrázek 6 SAX [4]

Jednou z nejdůležitějších vlastností SAXu je, že umožňuje dolní omezení měření vzdáleností. To znamená, že vzdálenost mezi aproximovanými řadami je vždy nižší než vzdálenost mezi řadami v originálním tvaru. Tato vlastnost je užitečná pro určení chyby a zrychlení výpočtu. (Sun a spol., 2013)

Nevýhodou SAXu je, že nerespektuje trend v datech. Sun a spol. proto vytvořili ESAX, který do řad vkládá další symboly, vyjadřující minimum a maximum v segmentu. To poskytuje vylepšení měření vzdáleností mezi řadami. Pham a spol. (2010) vytváří adaptivní verzi SAXu, která kombinuje SAX a k-means algoritmus a lépe tak vyjadřuje data. Lkhagva a spol. (2006) používá pro data z finančního odvětví verzi SAXu, která místo jednoho symbolu reprezentuje segment více symboly.

Hui a spol. (2011) pomocí SAXu analyzuje data z leteckého zapisovače s cílem zlepšit detekci poruch. Junejo a spol. (2012) algoritmem na základě SAXu rozpoznává lidské pohyby. Aplikací metody SAX na EKG signály se zabývá Kulahcioglu a spol. (2008). Cílem analýzy je detekovat anomálie v signálech.

Chyba aproximace vyjadřuje, jak přesně je řada reprezentována. Určí se jako suma čtvercových vzdáleností aproximované řady a původní.

3.1.5. Kompresí algoritmem Lempel-Ziv

LZ77 je univerzální kompresní algoritmus pro sekvenční data, který byl vytvořen Abrahamem Lempelem a Jacobem Zivem v roce 1977. Tento algoritmus položil základ několika dalším kompresním metodám. Mezi jeho výhody patří vysoký kompresní poměr a minimální ztráta informace. Postup lze aplikovat na řady vyjádřené konečnou abecedou.

Principem metody je vyjádření opakujících se segmentů odkazem na segment původní nekomprimované řady. Algoritmus prochází řadu metodou posuvného okna. Pokud je v řadě nalezena sekvence, která odpovídá již dříve zjištěné sekvenci, uloží se odkaz na polohu dřívější sekvence a informace o její délce. V případě, že je sekvence delší než odkaz, dochází ke kompresi. Komprimovaná data pak mohou být snadno zpět zrekonstruována procesem dekódování. (Ziv a spol, 1977)

Hned v následujícím roce byla představena inovace, algoritmus LZ78. Patří mezi slovníkové kompresní metody. V každé iteraci algoritmu se hledá nejdelší skupina znaků ve slovníku, ten se postupně rozšiřuje přidáním dalšího znaku z řady (Ziv a spol.,

1978). Zdokonalení této metody přinesl Terry Welch v roce 1984. (Welch, 1984) LZW zjednodušil kódování z dvojic indexu a znaku na pouze index.

Varianty algoritmu se používají v různých aplikacích, nejčastěji pro kompresi textových souborů (Arroyuelo, 2011) a obrázků (Glover, 1995). Fira a spol. (2008) komprimuje normální a patologické EKG signály algoritmem LZW a klasifikuje je pomocí neuronových sítí. Ferrario a spol. (2004) provádí LZ analýzu komplexity HRV signálů.

3.2. Míry podobnosti

Důležitou roli v shlukování hraje funkce, podle které budeme zjišťovat podobnost mezi řadami. Při porovnávání časových řad můžeme použít různá kritéria, nejčastěji jsou to různé formy vzdálenosti a korelace. Při výběru kritéria musíme zohlednit typ dat a metodu shlukování. Volba kritéria podobnosti přímo ovlivňuje výsledek shlukování. (Warren Liao, 2005)

3.2.1. Euklidovská vzdálenost

Měření vzdálenosti vyjadřuje odlišnost dvou objektů. Nejčastěji využívanou metrikou ve shlukování je euklidovská vzdálenost. Odpovídá geometrické vzdálenosti ve vícerozměrném prostoru. Jde o speciální případ obecné Minkowského vzdálenosti. (Karásek, 2012)

$$D(x, y) = \sqrt{\sum_{j=1}^J (x_j - y_j)^2}$$

Metriky mají několik základních vlastností: (Cilibrasi a spol., 2005)

Tabulka 2 Vlastnosti metriky

$D(x,y) \geq 0$	Nezápornost
$D(x,y) = D(y,x)$	Symetrie
$D(x,y) = 0 \leftrightarrow x = y$	Totožnost
$D(x,y) \leq D(x,z) + D(z,y)$	Trojúhelníková nerovnost

Euklidovská vzdálenost je jednoduchá, ale není příliš vhodná pro řady, ve kterých se objevuje časový posun. Ve shlukování se používá např. v algoritmu k-means, funguje totiž dobře u kulovitých tvarů shluků. (Meesrikamolkul a spol., 2012)

3.2.2. DTW

Častým problémem při porovnávání časových řad je jejich rozdílná délka. Nevýhodou Euklidovské vzdálenosti je, že ji můžeme určit pouze u stejně dlouhých signálů. Tento problém řeší DTW (Dynamic Time Warping), který umožňuje zarovnání dvou časově posunutých řad. Ve shlukování obvykle poskytuje lepší výsledky než Euklidovská vzdálenost. (Keogh a spol., 2004)



Obrázek 7 Porovnání Euklidovské vzdálenosti a DTW, (Keogh a spol., 2004)

Zarovnání dvou řad X , Y o délce m , n provedeme tak, že vytvoříme matici $(m \times n)$ vzdáleností mezi body z obou řad. V matici pak pomocí dynamického programování hledáme nejkratší cestu. (Meesrikamolkul, 2012)

$$DTW(x, y) = \sqrt{dist(x_n, y_m)}$$

$$dist(x_i, y_j) = (x_i - y_j)^2 + \min (dist(x_{i-1}, y_j), dist(x_i, y_{j-1}), dist(x_{i-1}, y_{j-1}))$$

DTW se často používá v různých oborech, zpracování medicínských dat pomocí DTW provádí např. Aach a spol. (2001), který zarovnáva časové řady s expresí RNA. Mezi časté aplikace patří rozpoznávání biometrických údajů, jako je dynamický podpis (Santosh, 2010) nebo otisky prstů (Kovacs-Vajna 2000).

3.2.3. Pearsonův korelační koeficient

Vztah mezi řadami lze charakterizovat korelačním koeficientem, který je maximální tehdy, pokud jsou řady lineárně závislé. Korelační koeficient nabývá hodnot od -1 do +1. Kladné hodnoty značí přímou závislost, hodnoty záporné naopak nepřímou. (Chae a spol., 2008)

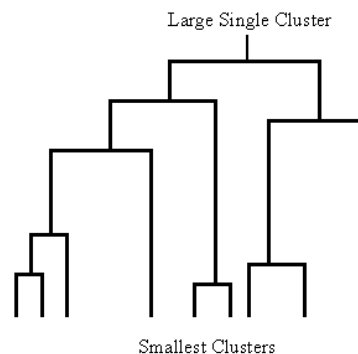
$$\rho_{XY} = \frac{\sum_{j=1}^J (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{j=1}^J (X_i - \bar{X})^2} \sqrt{\sum_{j=1}^J (Y_i - \bar{Y})^2}}$$

3.3. Metody shlukování

Existuje nepřeberné množství algoritmů pro vytváření shluků, důvodem je nepřesná definice toho, co je shluk. Algoritmy pro vytvoření shluků lze rozdělit na dvě skupiny, hierarchické shlukování a shlukování rozkladem (Hautamäki a spol., 2008). Pro shlukování časových řad se využívají 3 základní přístupy, lze shlukovat buď neupravená data, extrahované příznaky z dat, nebo parametry modelu získaného z dat (Warren Liao, 2005).

3.3.1. Hierarchické shlukování

Výsledkem této metody je tzv. dendrogram, který má podobu binárního stromu. Je na uživateli, aby dendrogram interpretoval. V algoritmu se postupuje směrem zdolana-horu nebo shora-dolů. Buď na začátku každý prvek představuje shluk a na základě podobnosti se postupně slučuje s jinými shluky (aglomerativní shlukování), nebo jsou všechna data v jednom shluku, který se rozděluje (divizivní shlukování). Postup se opakuje, dokud není dosaženo požadované struktury. (Štěpánková, 2013)

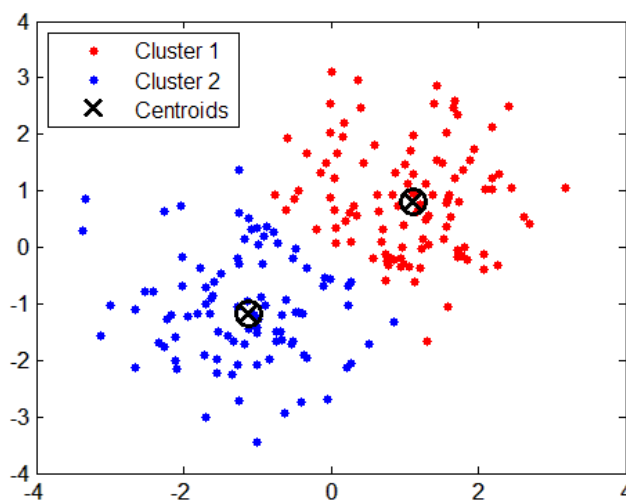


Obrázek 8 Dendrogram [5]

Barbu (2013) metodou SIFT a aglomerativním shlukováním rozpoznává obličej. Syed a spol. (2008) hierarchicky shlukuje EKG signály s cílem nalézt pacienty s vysokým rizikem infarktu myokardu. Chang a spol. (2009) používá shlukování pro analýzu HRV signálů pacientů se schizofrenií a porovnává je se zdravými lidmi. Costa Santos a spol. (2006) shlukuje FHR signály kompresí a následně hierarchickým shlukováním.

3.3.2. K-means

K-means je velice oblíbený iterativní algoritmus. Zařazujeme ho mezi shlukování rozkladem. Cílem k-means je nalézt shluky, kde by byla minimalizovaná chyba, určená jako součet vzdáleností prvků od středu svého shluku. Na začátku algoritmu se náhodně vyberou středy shluků, jejich počet je dán hodnotou K. Na základě nejmenší vzdálenosti se přiřadí prvky ke shlukům. V dalším kroku se přepočítá střed shluku. Postup se opakuje, dokud dochází ke změnám. Výhody tohoto algoritmu jsou jeho jednoduchost a efektivita, uživatel ale musí znát počet shluků a velmi záleží na počáteční inicializaci. (Meesrikamolkul a spol., 2012)



Obrázek 9 K-means [6]

Algoritmus k-means má uplatnění v mnoha oborech. Di Giuseppe a spol. (2014) jím analyzuje geologická data o seismické aktivitě. Donoso a spol. (2013) použil hierarchické shlukování a k-means na EKG záznamy pacientů s fibrilací síní.

3.3.3. Modelově založené algoritmy

Shlukování založené na směsi Gaussových rozdělání pracuje na podobném principu jako k-means. Předpokládá, že data byla vygenerována modelem, a snaží se nalézt parametry tohoto modelu. Vytváří se shluky, kde jsou data distribuována normálním rozdělením kolem středu. Postup se skládá z několika kroků. Zaprvé je to počáteční inicializace modelu, dále odhad parametrů modelu a nakonec stanovení počtu shluků. Pro určení parametrů modelu se většinou používá EM algoritmus, jehož cílem je vypočítat maximální věrohodnost klasifikace. Data se sdružují do shluků na základě nejvyšší věrohodnosti klasifikace všech možností. (Manning a spol., 2009, Fraley a spol., 2002)

Modelově založené shlukování pro analýzu časových řad používá Zhang a spol. (2004). Jeho postup lze aplikovat na řady s variabilní časovou osou získané nepravidelným vzorkováním. Klasifikaci HRV a echokardiografických signálů modelovým shlukováním provádí Granåsen a spol. (2013)

3.3.4. Metody založené na hustotě objektů

Shlukování založené na hustotě objektů se používá k vytvoření shluků libovolného tvaru. V tomto případě totiž selhávají k-means a EM algoritmy určené hlavně pro kulovité tvary shluků. Typickým zástupcem těchto metod je DBSCAN, který postupně spojuje objekty do shluků, dokud hustota prvků v okolí přesahuje danou mez. Výhodou DBSCANu je mimo jiné odolnost proti odlehlým hodnotám, naopak nevýhodou výpočetní náročnost a nutnost správné počáteční inicializace. (Shah, 2012)

Vylepšení tohoto algoritmu přinesl Shah (2012), který rozděljuje data podle hodnoty prahu před shlukováním. Tento postup funguje i na shluky uvnitř shluků a na objekty, jejichž hustota se mění. Kombinaci shlukování založeného na hustotě objektů a detekce odlehlých hodnot používá ve svém algoritmu Tao a spol. (2009)

3.4. Hodnocení kvality shluků

Správnost vytvořených shluků lze zhodnotit vnitřní a vnější kritériální funkcí. Jednou z vnitřních funkcí je součet odchylek uvnitř shluku. Vnější kritériem může být např. vyhodnocení úspěšnosti podle objektů se známou klasifikací. (Zhao, 2002)

4. Metody analýzy srdeční frekvence

Vyhodnocení záznamu z KTG monitorování lékařem je značně subjektivní a nejednotné. Objevuje se vliv inter- a intrapersonální variability při hodnocení, nedostatečné znalosti faktorů ovlivňujících FHR a individuální způsob hodnocení parametrů. Výsledky z KTG monitorování mají vysoký počet falešně pozitivních nálezů, což vedlo k častějším intervencím, především císařských řezů. Závěry z KTG vyšetření se proto někdy ověřují měřením hodnoty pH plodu. Z důvodů špatné interpretace vznikly snahy o vytvoření standardizovaného postupu klasifikace záznamů. (Novotná, 2001)

Sjednocení ve vyhodnocování záznamů srdeční frekvence plodu přineslo vydání směrnice (FIGO, 1985). Podle těchto pokynů určité vzory v záznamech úzce souvisí se stavem plodu. Při tomto postupu se klasifikace zaměřuje na morfologická kritéria jako bazální tepová frekvence, srdeční variabilita a přítomnost akcelerací a decelerací. Podle studií byla prokázána korelace mezi hodnocením algoritmu založeného na FIGO kritériích a výsledkem porodu. (Schiermeier a spol., 2008)

Metody automatického hodnocení KTG záznamu přispěly k zpřesnění klasifikace a k minimalizaci lidské chyby. Zatím však žádný z algoritmů nedosahuje vysoké úspěšnosti, proto je tento obor stále ve vývoji. V této kapitole popíšeme některé používané principy automatické klasifikace.

4.1. Statistické metody

Tyto přístupy používají ke klasifikaci FHR signálů a diagnostice hypoxie statistické charakteristiky. Hopkins a spol. (2006) používá základní statistiky (směrodatná odchylka, 95% percentil), odhady dimenze a entropii pro detekci snížené variability srdeční frekvence plodu. Hypotézy testuje t-testem. Santiago-Mozos a spol. (2013) diagnostikuje hypoxii plodu pomocí metod založených na teorii informace. Jako příznaky jsou použity zprvu statistické momenty a zadruhé vzdálenosti mezi časovými řadami. Klasifikaci FHR provádí k-NN klasifikátorem. Lepších výsledků je dosaženo křížovou validací. Cahill a spol. (2014) analyzuje vzory v FHR signálu metodou FDA. Snaží se identifikovat decelerace a porovnat je s klinickým výsledkem (pH, Apgar

skóre). Rozdíly mezi vzory decelerací byly měřeny metodou hlavních komponent. Souvislost decelerací a klinického výsledku byla odhadnuta lineární regresí.

4.2. Lineární a nelineární metody

Chudáček a spol. (2010) hodnotí FHR záznamy podle různých příznaků (statistické, frekvenční, nelineární) a rozděluje signály do skupin podle FIGO pravidel. Goncalves a spol. (2007) analyzuje FHR záznam v závislosti na stavu plodu (REM a non-REM spánek, klidné a aktivní bdění). K analýze byly použity segmenty dlouhé 10 min a určují se u nich parametry jako frekvenční spektrální indexy, entropie a LTI. Kikuchi a spol. (2006) používá nelineární analýzu HRV u normálních a růstově retardovaných plodů. Nelineární metody tu zahrnují rekonstrukci atraktoru, výpočet největšího Ljapunova exponentu a odhad korelace dimenze.

4.3. Waveletová transformace

German-Sallo (2013) se zaměřuje na spektrální metodu analýzy HRV s cílem zjistit index sympato-vagální rovnováhy dospělého člověka. Signál je dekomponován paketovou vlnkovou transformací. Tato metoda poskytuje více informací než Fourierova transformace. Deshpande a spol. (2012) se kromě vlnkové transformace zabývá i Wigner-Villeho transformací. HRV je analyzováno na základě LF/HF poměru (nízké a vysoké frekvence).

4.4. Gaussovské modelování

Daoud a spol. (2013) odhaduje spektrální parametry HRV na základě Gaussova modelu. Tato metoda modeluje výkonové spektrum jako směs dvou normálních rozdělání, což lépe vyjadřuje výkonové spektrum nízkých a vysokých frekvencí. Pro reprezentaci ve frekvenční oblasti používá Gaborovu transformaci. Cílem studie je analyzovat funkci autonomního nervového systému a rovnováhu mezi činností sympatického a vagálního nervu.

4.5. Partial least squares discriminant analysis

PLS-DA je statistická metoda, která se používá pro klasifikaci objektů. Tvoří ji PLS regrese, kde závislá proměnná je kategoriální. Rossini a spol. (2012) aplikuje PLS-DA na data ze smyslového hodnocení výrobků. Szymańska a spol. (2011) používá PLS-DA k hodnocení diagnostických modelů v metabolomických studiích.

4.6. Phase-rectified signal averaging

PRSA je metoda průměrování signálů, které byly nejprve zarovnány podle klíčových bodů. Metoda umožňuje analýzu zašuměných a nestacionárních signálů s různou fází. Kantelhardt a spol. (2007) tuto metodu používá pro detekci kvaziperiodických oscilací v srdeční frekvenci z EKG signálů. Cílem je určit decelerační kapacitu, podle které lze určit riziko smrti. Stejnou analýzu provádí Pan a spol. (2009), ale používá vylepšenou verzi algoritmu s přesnějším určením klíčových bodů. Georgieva a spol. (2014) tvrdí, že výsledek průměrování signálů indikuje stav plodu a lze z něj diagnostikovat acidémii.

5. Databáze signálů

Navzdory tomu, že kardiokografie vznikla už v 60. letech 20. století a stala se nejběžnějším vyšetřením pro diagnostiku fetální hypoxie, až dosud neexistovala žádná databáze KTG signálů. V této práci budeme analyzovat data z databáze CTU-UHB získané z porodnického oddělení ve Fakultní nemocnici Brno.

Databáze obsahuje 552 záznamů, v každém z nich je FHR signál (záznam průběhu srdeční frekvence plodu), tokografický signál (záznam průběhu tlaku v děloze) a klinická data. Všechny signály jsou maximálně 90 minut dlouhé a začínají maximálně 90 minut před koncem porodu. Klinická data zahrnují informace o porodu (typ porodu, délka první a druhé fáze), informace o plodu (pohlaví, váha, délka těhotenství), informace o matce a další. Nás bude zajímat především výsledek porodu vyjádřený hodnotou pH novorozence. Záznamy byly vybrány z celkového počtu 9164 porodů na základě medicínských a technických kritérií. Většina záznamů pochází z vaginálních porodů, pouze 46 z nich jsou císařské řezy. Analyzovat budeme pouze samotný FHR signál, bez ohledu na děložní činnost. (Chudáček a spol., 2014)

6. Předzpracování dat

Předzpracování signálu má velký vliv na další analýzu a na celou klasifikaci. Je nutné správně oddělit užitečnou informaci od rušení. Při odstranění artefaktů může dojít i ke ztrátě užitečné informace, ale ponechání chybných hodnot by vedlo k nesprávným výsledkům. Předzpracování dat je klíčová operace a někdy je nutné se k ní vracet a zdokonalit postup.

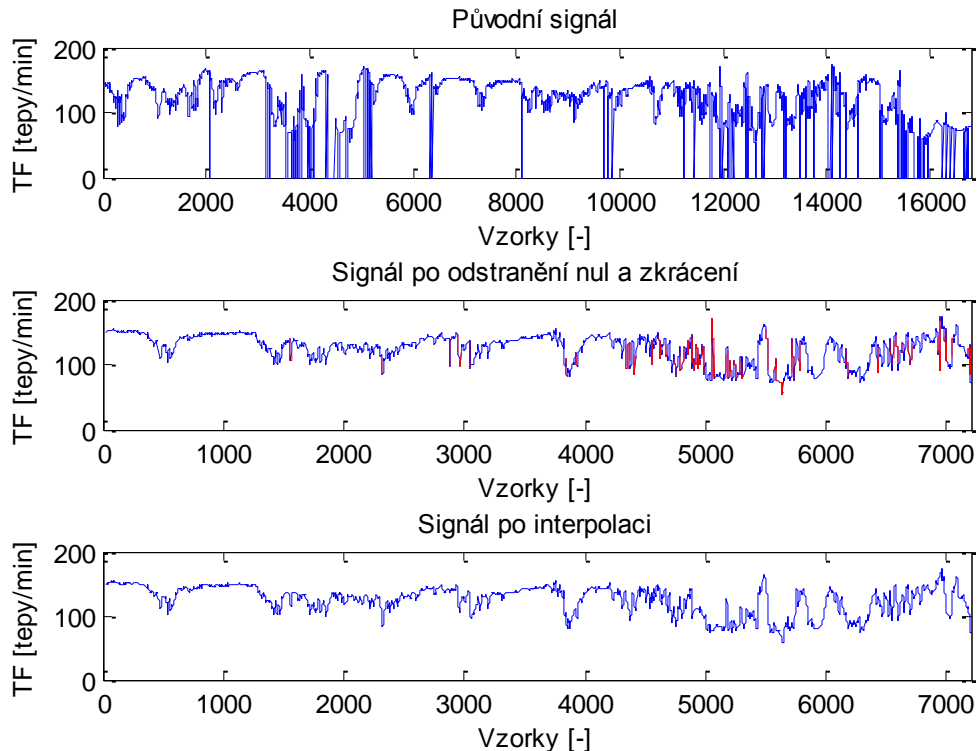
6.1. Odstranění artefaktů a zkrácení signálu

Signály z databáze jsou v původním tvaru, tak jak byly naměřeny. Obsahují tedy chybějící data, šum a artefakty. Chybějící data jsou vyjádřena nulovým signálem a jde o dobu, kdy neprobíhalo měření. Nejčastějším artefaktem je zachycení srdeční akce matky, která má nižší frekvenci než srdeční akce plodu. Vysokofrekvenční artefakty pak mohou být způsobené pohybem matky nebo plodu.

Ze signálů jsme nejprve odstranili nulové a chybné vzorky. Srdeční frekvence by neměla klesnout pod 30 tepů/min, ani být vyšší než 200 tepů/min. Tyto hodnoty jsou považovány za artefakty. Signály jsou různě dlouhé, pro další zpracování je nutné vybrat stejně dlouhé části. U vaginálních porodů máme k dispozici údaj o začátku druhé doby porodní. Vybrali jsme tedy část signálu o délce 30 min z konce první doby porodní. V případě císařských řezů budeme zpracovávat posledních 30 min signálu. Signály byly snímány s vzorkovací frekvencí 4 Hz, ke zpracování tedy použijeme úsek 7200 vzorků. Pokud nebudeme brát v úvahu údaj o začátku druhé doby porodní, budeme zpracovávat 8000 vzorků ze začátku signálu a 8000 vzorků z konce.

Rušení, které se v signálu projevuje jako impuls, jsme ze signálů odstranili metodou podle Bernardese a spol. (1991). Nejprve se určí stabilní segment, který je definován jako segment 5 následujících tepů, kde celková chyba je menší než 10 tepů/min. Pokud je rozdíl větší než 25 tepů/min, je provedena spline interpolace mezi posledním vzorkem předchozího stabilního segmentu a prvním vzorkem následujícího stabilního segmentu. Spline interpolaci jsme zvolili, protože v tomto případě lépe doplňuje data než lineární interpolace.

Tato metoda nám však neodstraňuje velmi častý artefakt, a to srdeční frekvenci matky. Jelikož není snadné ji odstranit beze ztráty užitečné informace, bude zanedbána až při pozdější aproximaci signálu.



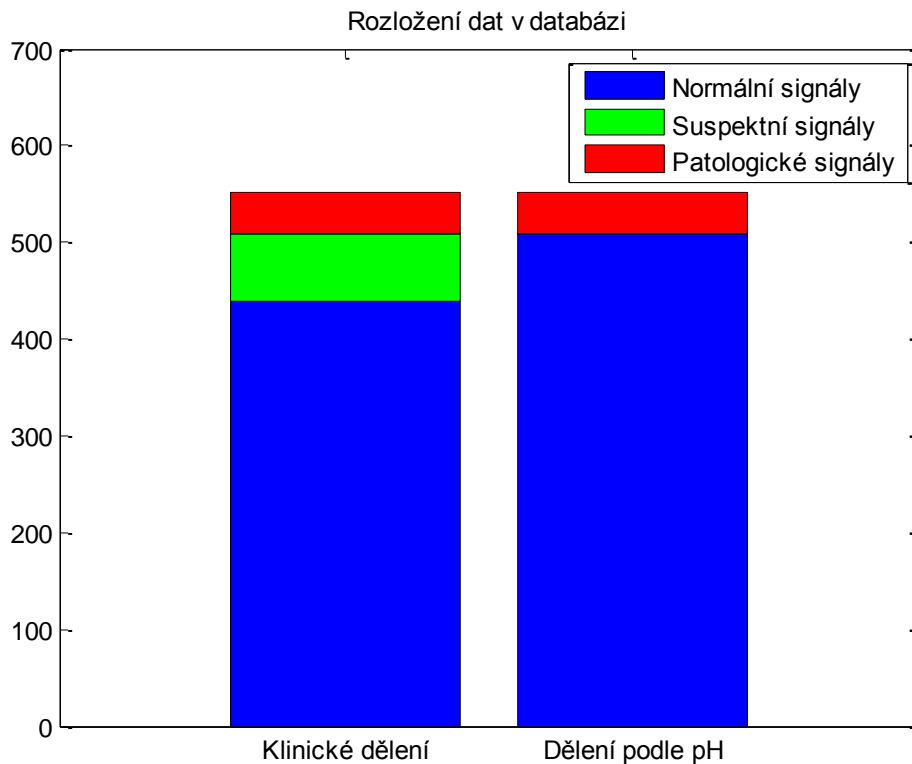
Obrázek 10 Filtrace signálu, červeně vyznačeny interpolované segmenty

Pro analýzu můžeme použít celou databázi signálů, to nám poskytne dostatečně velký soubor dat. V tomto případě ale trvají všechny výpočty poměrně dlouhou dobu, proto jsme vybrali 8 normálních a 8 patologických signálů, které budeme používat jako vzorové.

6.2. Rozdělení a seřazení signálů

Signály budeme klasifikovat do dvou skupin, na signály normální a patologické. Díky hodnotě pH v klinických datech u každého signálu předem známe výsledek porodu. Za patologické považujeme hodnoty pH menší než 7,05 (Chudáček a spol., 2014). Podle tohoto dělení je v databázi patologických 44 signálů z celkových 552. Je možné signály rozdělit do tří skupin, tj. normální, suspektní a patologické. Stejně dělení se používá u klinického hodnocení. V tomto případě je hranice pro suspektní nález hodnota pH 7,2. Celou databázi jsme si pro usnadnění další práce upravili tak, aby první

v pořadí byly normální signály a za nimi patologické. Máme k dispozici nevyváženou množinu dat, patologických signálů je výrazně méně. Stejná situace je i v klinické praxi, většina nálezů patří do normální skupiny.



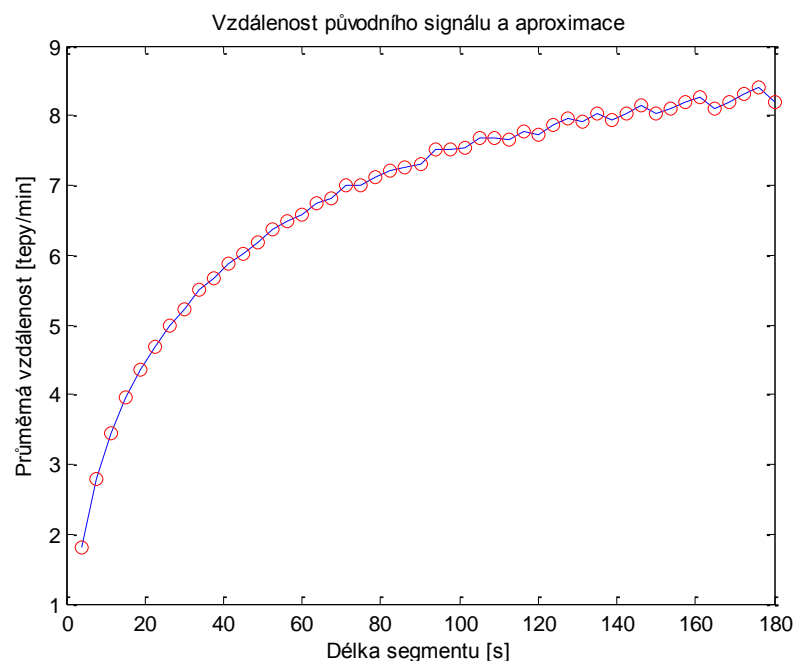
Obrázek 11 Obsah databáze

7. Diskretizace signálů

Diskretizace časových řad umožňuje efektivní získání informace z dat a vede k lepším výsledkům než zpracovávání řad v původním tvaru. Cílem diskretizace je redukovat počet příznaků v datech tak, aby tyto příznaky co nejlépe vyjadřovaly původní signál. Aproximace signálu je také nutná pro snížení výpočetní náročnosti. Řadu příznaků, kterou získáme z diskretizace, pak použijeme jako atributy do shlukování. Pro aproximaci KTG signálů budeme používat některé z nejběžnějších reprezentací.

7.1. Rozdělení signálu na okna

Metody reprezentace signálu PAA, PLA a SAX jsou založené na vytvoření stejně dlouhých segmentů. Pro tyto metody jsme signál rozdělili na časová okna. Délku okna je možné různě měnit, stejně jako je možné zvolit různý překryv oken. Pro získání dobrého výsledku klasifikace je nutné správně zvolit délku okna. Krátké okno sice umožní přesnější vyjádření signálu, ale nedojde k dostatečné redukci počtu příznaků. Ideální délka okna je kompromisem těchto dvou požadavků. Přesnost reprezentace lze určit výpočtem euklidovské vzdálenosti mezi aproximací a původním signálem. Ideální délku jsme tedy určili ze závislosti průměrné vzdálenosti na délce okna.

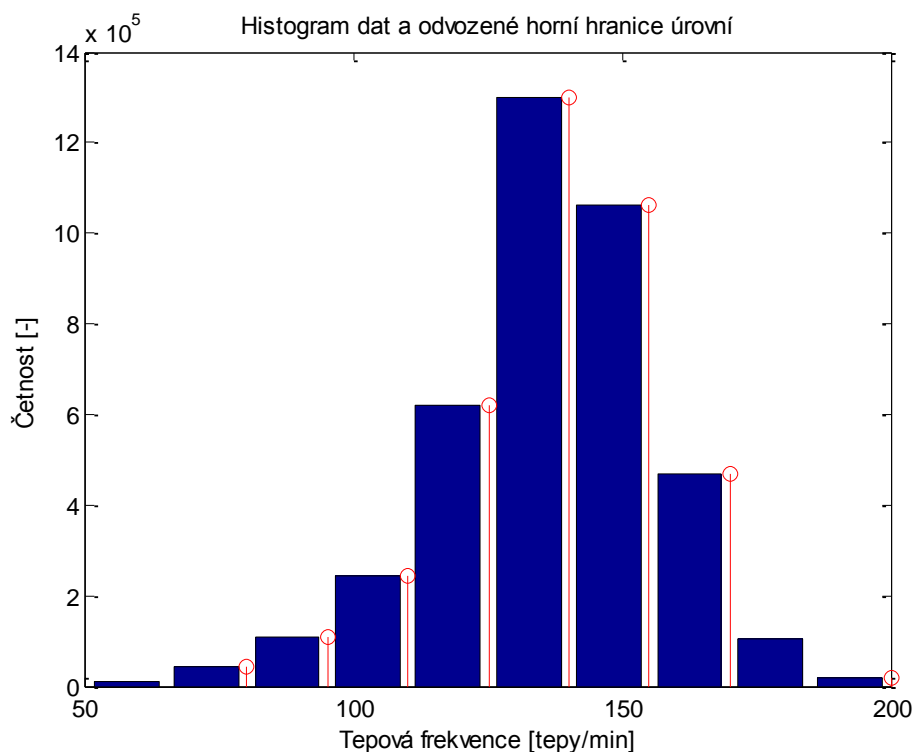


Obrázek 12 Přesnost aproximace PAA

Závislost přesnosti reprezentace metodou PAA má exponenciální průběh, pro délku okna 2 min se chyba aproximace blíží k 8 tepům/min. Pro PLA získáme podobný průběh, i průměrná vzdálenost této metody je srovnatelná. Platí, že s kratším oknem získáme přesnější aproximaci. Od délky okna více než 2 min už vzdálenost roste pomalu, měli bychom tedy dostat podobné výsledky i s velmi dlouhými okny. Přesto se zdá, že ideální délka okna by měla být menší než 30 s, z důvodu chyby menší než 5 tepů/min. Budeme počítat s oknem 15 s, to je doba jedné decelerace.

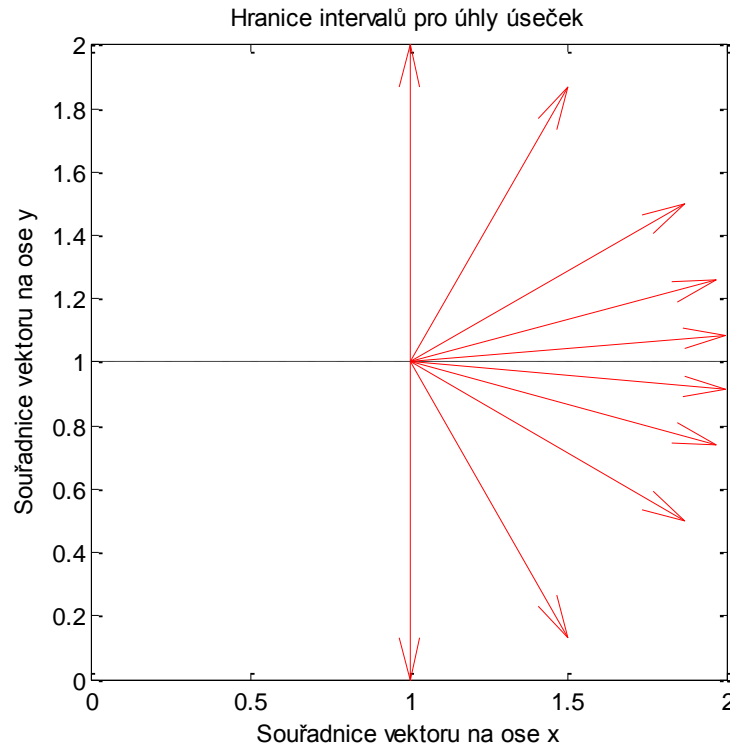
7.2. Definice hranic intervalů

Rozdělení signálu na okna poskytuje diskretizaci v časové ose, v ose hodnot se provede diskretizace rozdělením hodnot do intervalů. Pro metody PAA, SAX a APCA se jako aproximace používá průměr segmentu, případně modus nebo medián. Krajní hodnoty intervalů jsme získali z histogramu všech signálů. Histogram rozdělí data do požadovaného počtu tříd, odtud získáme horní hranice každé třídy. Naše analyzovaná data mají přibližně tvar normálního rozdělení, tudíž okrajové třídy mají nízké hodnoty četností. Proto jsme třídy s hranicemi menší než 80 tepů/min sloučili do jedné třídy. Stejně tak jsme sloučili i hranice větší než 180 tepů/min.



Obrázek 13 Histogram dat a hranice intervalů

U metod PLA a APLA je k aproximaci použita směrnice přímky, a tak jsou hranice určeny úhlem, který svírá přímka s časovou osou. Úhel může nabývat hodnot od -90° do 90° . Intervaly blízké 0° jsou zvoleny kratší, spadá do nich totiž velká část dat. Stejně jako u hranic z histogramu se při diskretizaci prvky menší než daná hranice aproximují jako střed tohoto intervalu.



Obrázek 14 Hraniční úhly intervalů

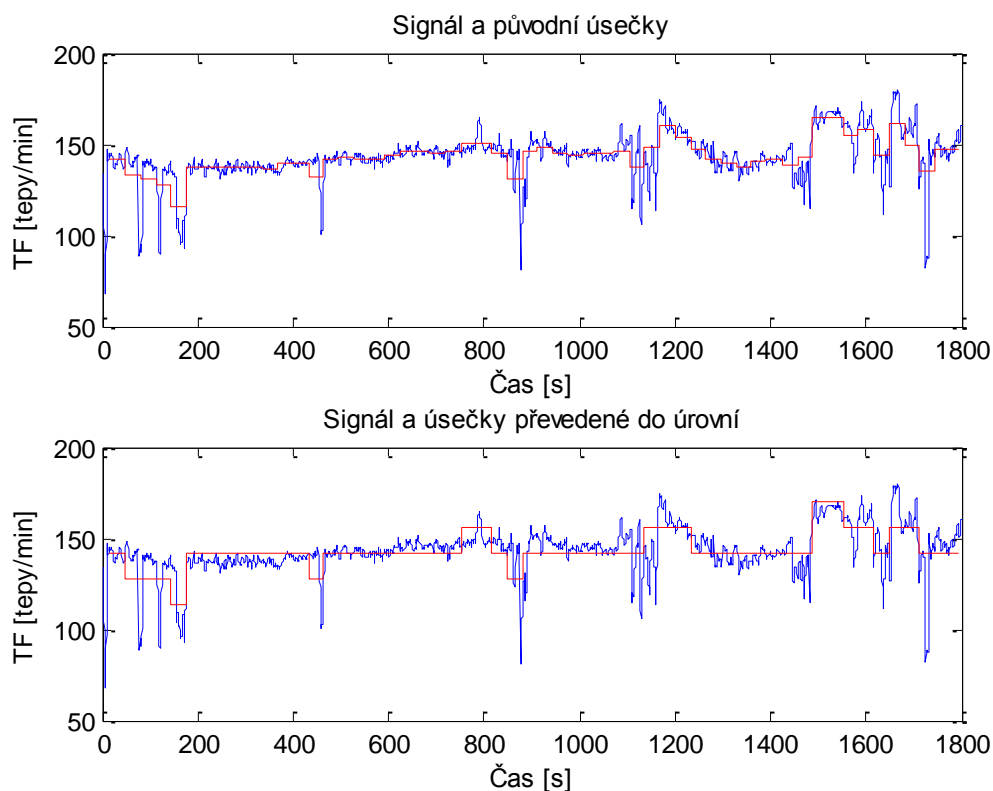
7.3. Reprezentace

Pro vyjádření signálu jsme použili pět metod. Reprezentace jsou založené buď na výpočtu průměru segmentu, nebo na lineární interpolaci v segmentu. Budeme moci následně porovnat jejich úspěšnost v klasifikaci. Další redukci příznaků a identifikaci podskupin v datech přinesl kompresní algoritmus LZ78.

7.3.1. PAA

Nejběžnější metodou aproximace časových řad je PAA. Po rozdělení signálu na segmenty v každém segmentu vypočítáme průměr, modus nebo medián. Při překryvu oken vypočítáme průměr daných charakteristik ze dvou oken. Toto číslo pak bude charakterizovat celý segment. Následně ho zařadíme do některého intervalu, podle už dříve vytvořených hranic. Intervalům jsou přiřazena přirozená čísla od 1 do K , kde K je

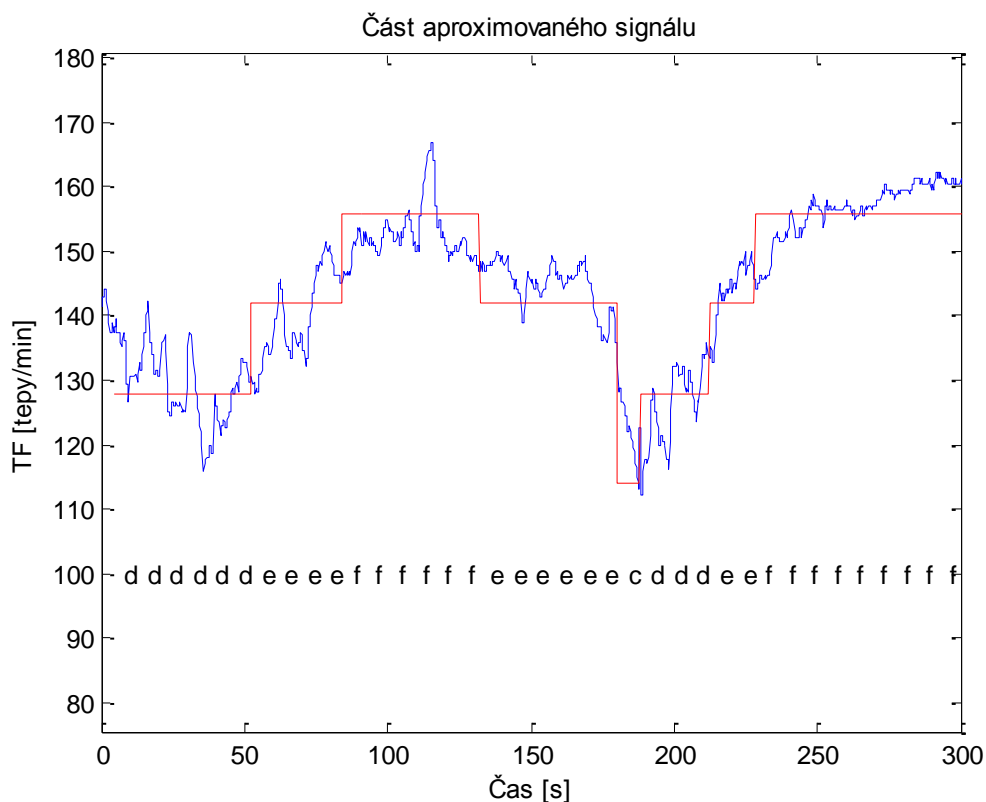
počet nadefinovaných intervalů. Celá databáze je pak vyjádřena jednou maticí přiřazených čísel. Sloupce určuje počet signálů, řádky jsou jednotlivé příznaky a hodnota prvku je číslo intervalu. Tato metoda je velmi jednoduchá na výpočet a dobře vystihuje změny v signálu. V aproximaci jsou zároveň potlačeny i artefakty, které se nepodařilo odstranit filtrací.



Obrázek 15 Repräsentace pomocí PAA

7.3.2. SAX

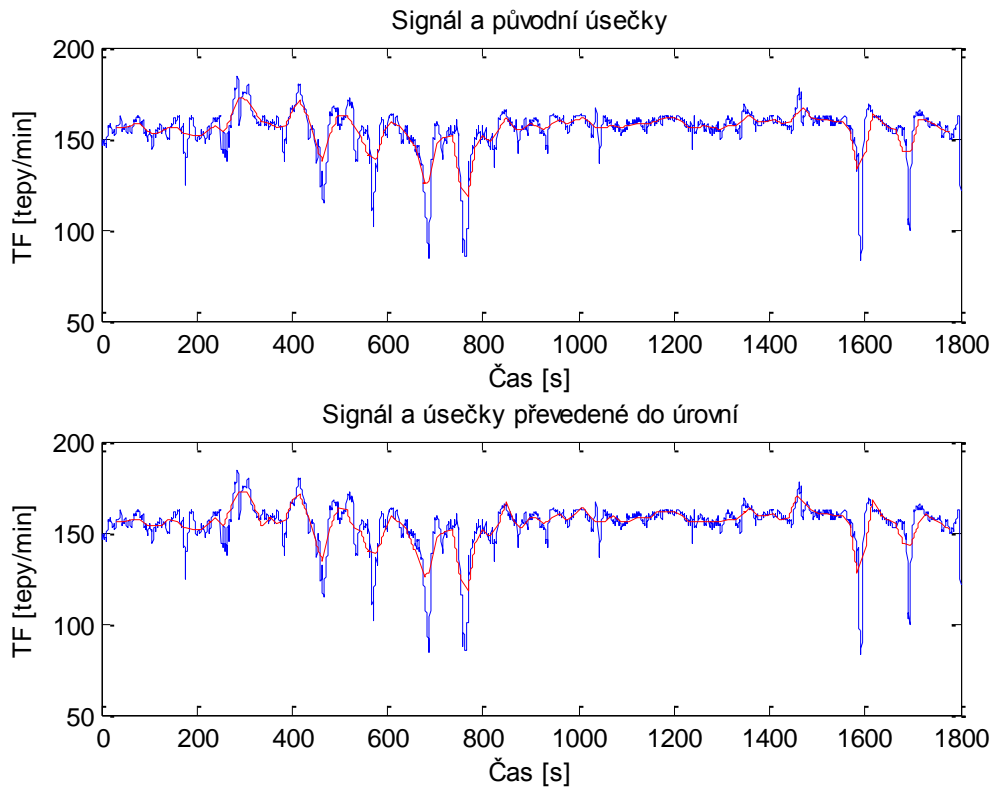
SAX vychází z PAA, segment však není vyjádřen číslem, ale písmenem. V základu používáme abecedu začínající písmenem A o délce K , tedy rovné počtu intervalů. Databáze je pak reprezentována jako matice znaků. Protože není možné provádět matematické výpočty s písmeny, pro většinu dalších operací převádíme znaky abecedy do ASCII kódu.



Obrázek 16 Repräsentace pomocí SAX

7.3.3. PLA

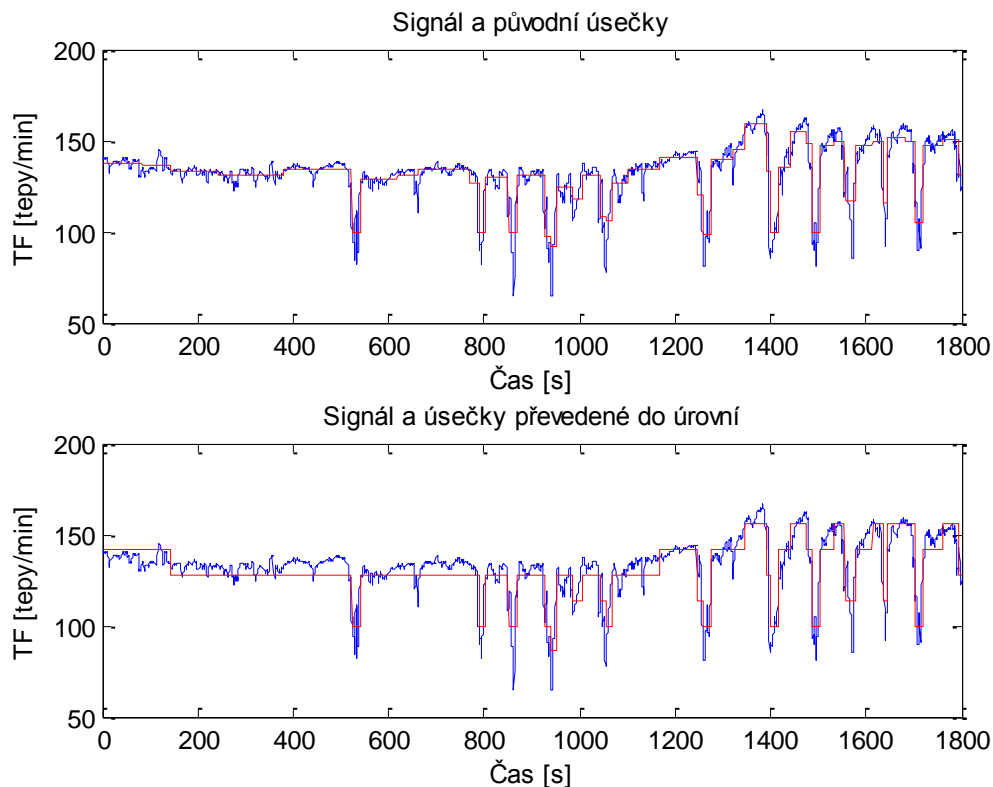
PLA vyjadřuje časovou řadu jako orientované úsečky. Úsečky můžeme určit několika způsoby. Nejjednodušší je provést lineární interpolaci mezi prvním a posledním vzorkem daného segmentu. Další možností je vypočítat průměr dvou po sobě jdoucích segmentů a provést interpolaci mezi těmito průměry. V tomto případě ale dochází k posunu, aproximace tedy nezačíná na začátku signálu. Segment můžeme vyjádřit i pomocí lineární regrese. Ta bere v úvahu všechny body v segmentu, ale je také výpočetně nejnáročnější. Úsečky vyjadřujeme směrnicovou rovnicí přímky, kde směrnice je tangens úhlu, který úsečka svírá s osou x . Vypočtené směrnice podle nadefinovaných hranic zařadíme do určených intervalů. V matici se pak nachází čísla od $-K/2$ do $K/2$. Záporná čísla vyjadřují záporné úhly s časovou osou, kladná čísla pak úhly kladné. Tato metoda vystihuje signál mnohem lépe než PAA, očekáváme proto lepší úspěšnost při klasifikaci.



Obrázek 17 Repräsentace pomocí PLA

7.3.4. APCA

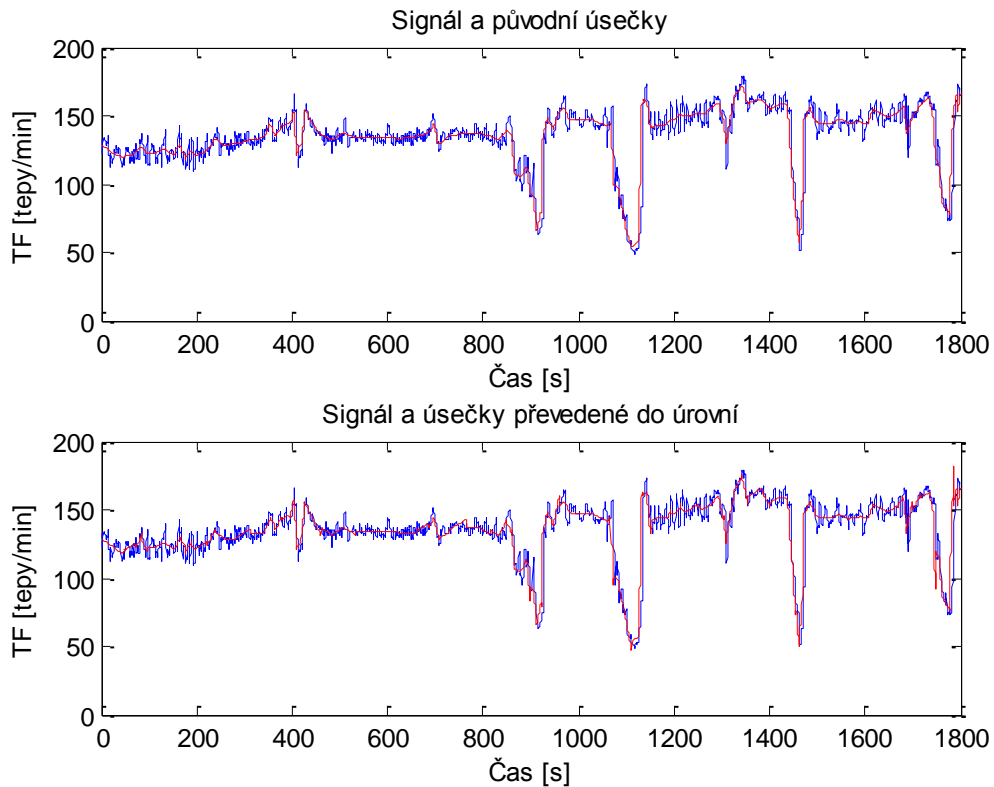
Nevýhoda předchozích metod jsou stejně dlouhá okna, signál mnohem lépe vyjadřují adaptivní metody. Úsek, kde je signál stacionární, je vyjádřen pouze jedním segmentem. Naopak v úsecích s častými změnami se vytvoří víc segmentů. Každý signál je vyjádřen dvěma sloupci, v prvním je číslo intervalu, do kterého spadá charakteristika segmentu, v druhém sloupci je délka daného segmentu. Adaptivní okna jsou vytvořena metodou sliding windows podle Keogha a spol. (2001). Začneme se segmentem o dvou po sobě jdoucích vzorcích. Vypočítáme aproximaci segmentu (průměr, modus, medián), následně vypočítáme chybu mezi aproximací a původním signálem. V dalším kroku zvětšíme segment o jeden vzorek a postup opakujeme. Segment stále narůstá, dokud je chyba aproximace menší než pevně zvolený práh. Při dosažení chyby se narůstání zastaví a vytvoří se konečný segment. Velikost maximální chyby jsme určili empiricky. Výhodou adaptivních metod je, že v místech stacionárního signálu vytvoří dlouhý segment, zatímco v úseku signálu s velkými změnami vytvoří více krátkých segmentů. Můžeme tak zkrátit části signálu s nedůležitou informací. Nevýhodou je, že pro aproximaci potřebujeme dvě hodnoty místo jedné.



Obrázek 18 Repräsentace pomocí APCA

7.3.5. APLA

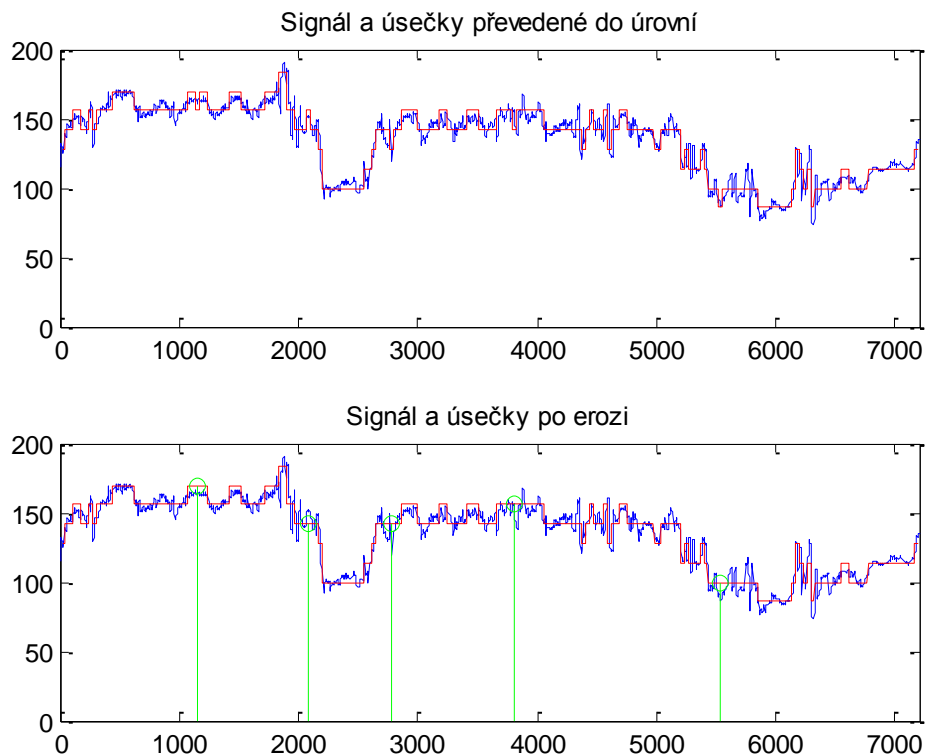
Stejným algoritmem jsme vytvořili i adaptivní PLA. Lineární interpolaci v tomto případě vytváříme buď jednoduše mezi prvním a posledním vzorkem segmentu, nebo body, mezi kterými budeme interpolovat, určíme průměrem. V prvním segmentu vypočítáme průměr první a druhé poloviny segmentu a interpolujeme mezi těmito průměry. Pro další segmenty už použijeme předchozí bod a průměr celého segmentu, takže zajistíme, aby úsečky na sebe navazovaly. Další možností je lineární regrese. Aproximace je tvořena různě dlouhými úsečkami, jako charakteristiku opět volíme směrnici. Nevýhodou adaptivních metod je rozdílná délka aproximace jednotlivých signálů, přestože analyzujeme stejně dlouhé signály. Tento problém řešíme tak, že upravíme délku všech aproximací podle nejkratší z nich.



Obrázek 19 Repräsentace pomocí APLA

7.4. Eroze

Klasifikace patologických signálů závisí na správném odhalení decelerací. Decelerace a akcelerace trvají alespoň 15 s, proto kratší úsečky můžeme zanedbat. U metody PAA a stejně tak i u SAXu jsme provedli erozi reprezentace. Změna v úrovních totiž nemusí znamenat změnu v srdeční frekvenci. Ve většině případů jde o podobné úrovně, které byly při diskretizaci zařazeny do jiných intervalů. Pokud je tedy délka okna kratší než 15s, dojde k odstranění krátké změny úrovně, pokud signál pokračuje na úrovni před změnou. V případě metody PLA bychom odstraněním úsečky porušili návaznost celé reprezentace a stejně tak u adaptivních metod mají i krátké úseky své opodstatnění, jde totiž o úseky s velkou změnou signálu.

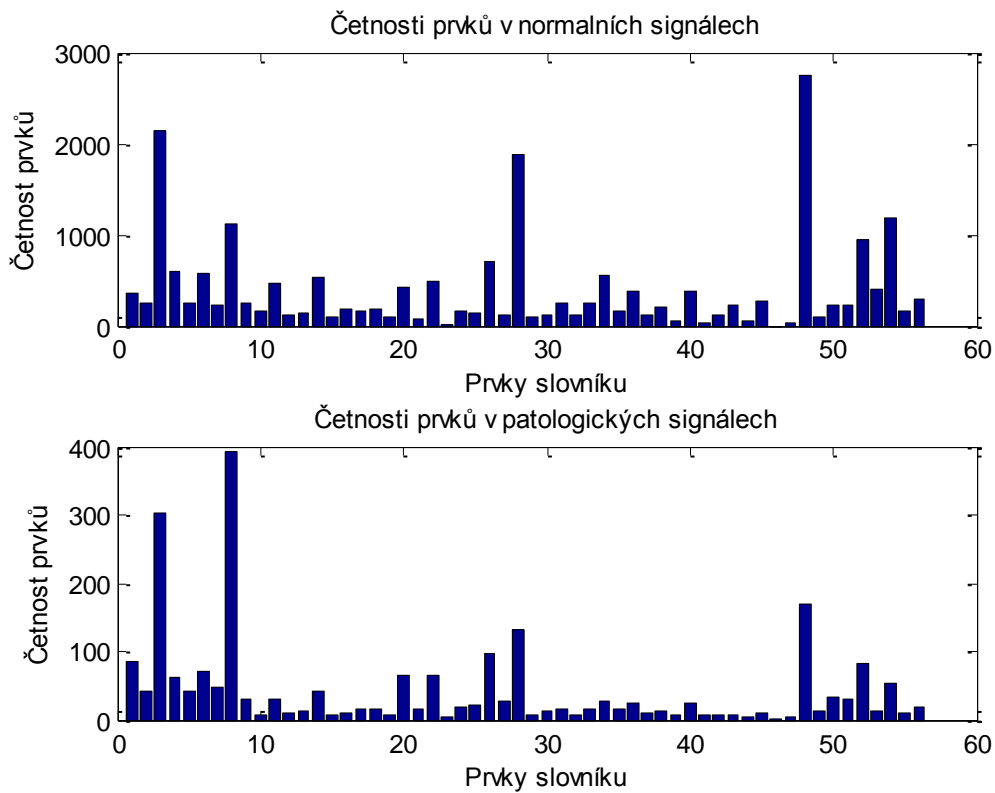


Obrázek 20 Eroze signálu s vyznačením odstraněných úseků

7.5. Komprese algoritmem Lempel-Ziv

Na aproximovaná data jsme aplikovali kompresní algoritmus LZ 78. Dosáhli jsme tím ještě většího snížení počtu příznaků. Vstupem do funkce je řetězec znaků, který chceme komprimovat a abeceda, která tvoří počáteční slovník. Do abecedy jsme přidali všechny jednotlivé znaky, které se v řetězci vyskytují. Algoritmus dále prochází řetězec od začátku a v každém kroku rozšiřuje skupinu analyzovaných znaků. Neznámá skupina znaků se přidá do slovníku. Pokud ale algoritmus narazí na skupinu, která již ve slovníku je, uloží se do nového řetězce odkaz na pozici této skupiny ve slovníku. Pokud je odkaz kratší, než skupina znaků, dochází k úspoře místa. Výstupem funkce je slovník obsahující všechny použité skupiny znaků a nový vektor s odkazy na polohu ve slovníku. Takto jsme zkomprimovali první signál, u ostatních signálů jsme už netvořili nový slovník, ale použili pro kompresi stále stejný slovník. Dekompresi je pak velice snadná, stačí místo odkazů dosadit zpět danou skupinu znaků ze slovníku. Stejně jako u adaptivních metod aproximace, i LZ78 tvoří různě dlouhé vektory. To nám dost znesnadňuje další práci s daty. Při následném porovnávání signálů jsme se rozhodli brát

v úvahu pouze četnosti výskytu prvků slovníku v jednotlivých vektorech. Signály, které si jsou spolu podobné, totiž mají více společných skupin znaků.



Obrázek 21 Histogram četností znaků ze slovníku v datech

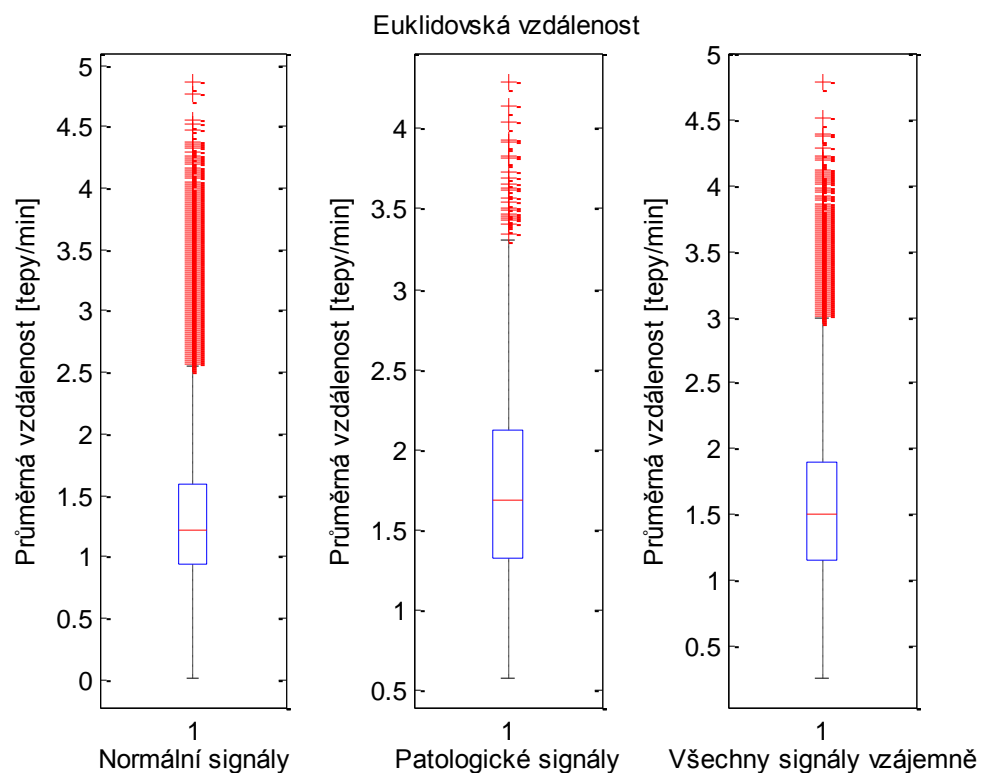
Z histogramu je zřejmé, že nejčastější skupinou v datech je dvojice úrovní 44, což odpovídá srdeční frekvenci kolem 120 tepů/min a čtveřice 6666, která označuje frekvenci přibližně 150 tepů/min. Prvních 8 prvků ve slovníku jsou jednotlivé znaky, i ty mají vysokou četnost. V patologických signálech se vyskytují vyšší úrovně než v normálních, jinak je rozložení skupin dost podobné.

8. Porovnávání podobnosti

Signály po aplikaci některé z aproximací a zkrácení na stejnou délku můžeme mezi sebou snadno porovnávat. K určení podobnosti jsme použili některé základní metody, jako je euklidovská vzdálenost a korelační koeficienty. Provedli jsme také frekvenční analýzu s cílem odhalit strukturu dat.

8.1. Euklidovská vzdálenost

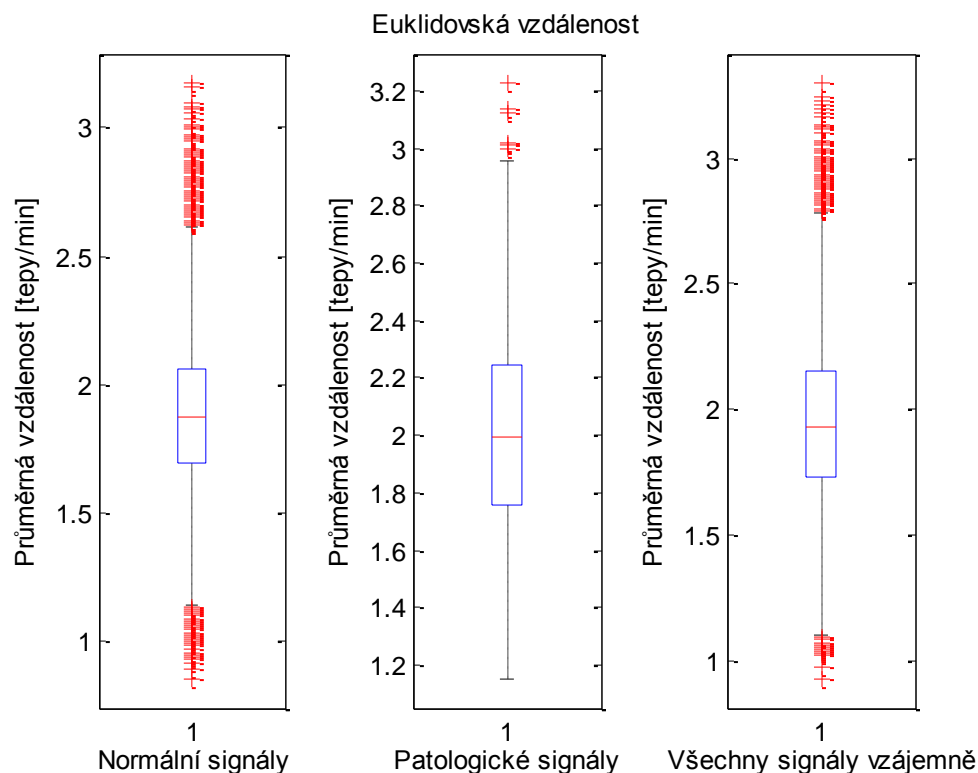
Nejběžnějším a jednoduchým způsobem porovnání dvou signálů je výpočet jejich vzdálenosti. Pro výpočet je nutné, aby byly aproximace stejně dlouhé. Celková vzdálenost je pak suma vzdáleností dvou odpovídajících bodů. Euklidovská vzdálenost není příliš vhodná pro KTG signály, nebere totiž v úvahu posunutí v časové ose ani rozdíl v hodnotách. Jako hrubý odhad podobnosti je ale dostatečná.



Obrázek 22 Vzdálenost signálů po aproximaci PAA

Počítali jsme vzdálenost mezi signály aproximovanými oknem dlouhým 16s, podle předchozího grafu pro určení délky okna by nám tedy měla vyjít vzdálenost kolem

2 tepů/min. To se potvrdilo. Z grafů je vidět, že normální signály mají mezi sebou nižší průměrnou vzdálenost, než signály patologické. To může být způsobeno větší variabilitou patologických signálů. Ovšem v obou skupinách se vyskytuje značný počet odlehlých hodnot, které označují signály s velkou vzdáleností. Vzdálenost mezi všemi signály vzájemně je průměrem těchto dvou skupin.

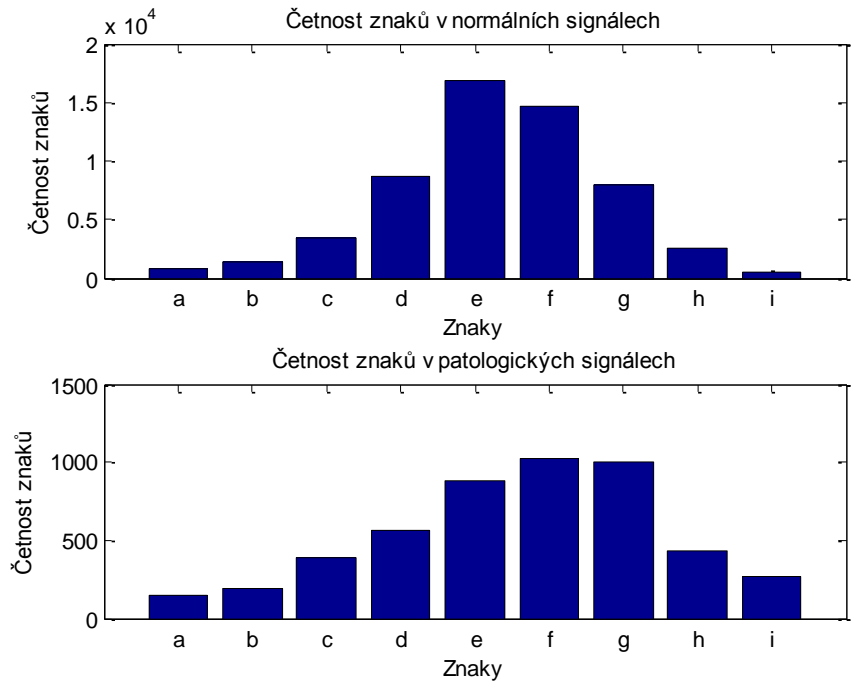


Obrázek 23 Vzdálenost signálů po aproximaci PLA

Pro metodu PLA jsou vzdálenosti obecně vyšší a více vyrovnané mezi oběma skupinami. Na rozdíl od PAA se objevují i odlehlé hodnoty směrem dolů. Nízkou vzdálenost mezi sebou mají signály se stejnou nebo podobnou bazální tepovou frekvencí, takže by mohly patřit do stejné skupiny.

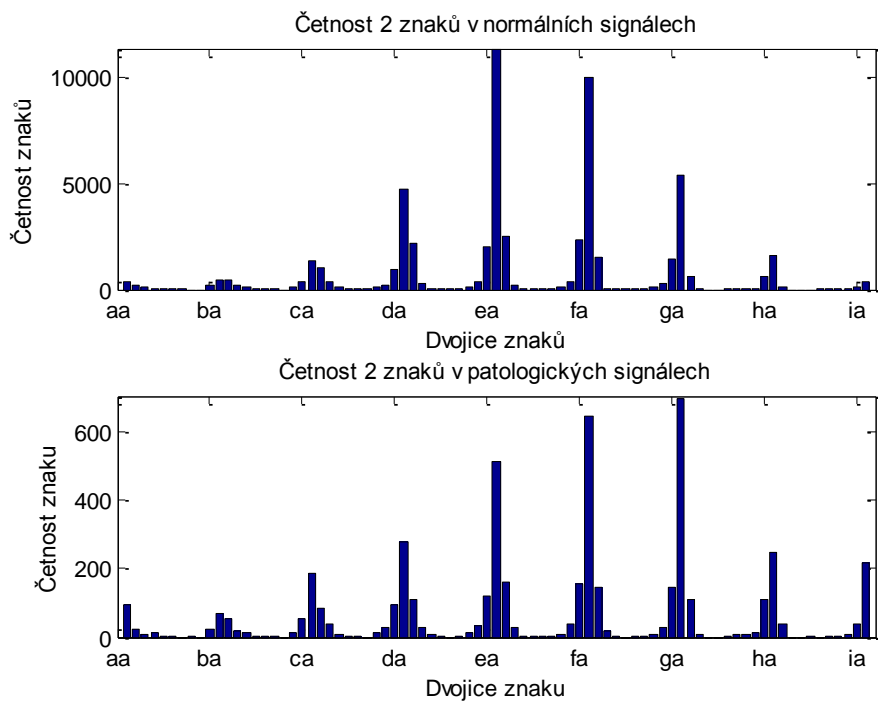
8.2. Frekvenční analýza

Ze získaných aproximací signálů můžeme frekvenční analýzou určit četnost jednotlivých znaků nebo skupin znaků. To nám pomůže zjistit, které úseky a vzory signálu jsou typické pro patologické záznamy. Zkoumali jsme skupiny od 1 do 5 znaků za sebou. Hledání skupin znaků provádí už algoritmus LZ77, ve frekvenční analýze ale budeme procházet všechny možné kombinace znaků.



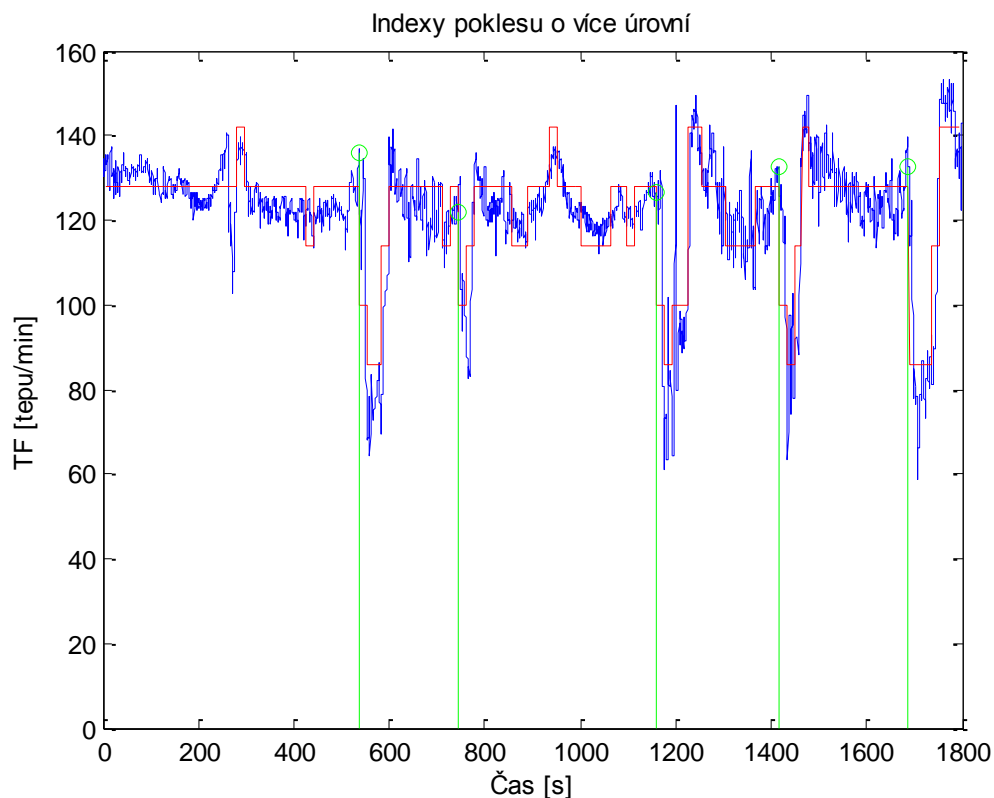
Obrázek 24 Četnost jednotlivých znaků – SAX

Z grafu četností vidíme, že data mají přibližně tvar normálního rozdělení. Patologické signály mají podle očekávání nejčetnější znaky ve vyšších tepových frekvencích. Četnosti dvojic znaků mají podobný tvar. Nejčastějšími dvojicemi jsou vždy dva stejné znaky.



Obrázek 25 Četnosti dvojice znaků – SAX

Hledáním podskupin znaků jsme se pokusili vyhledat v signálu decelerace a určit jejich počet. Detekovali jsme pokles v aproximaci signálu o více než dvě úrovně, to znamená přibližně 20 tepů/min.



Obrázek 26 Místa vyhodnocená jako decelerace

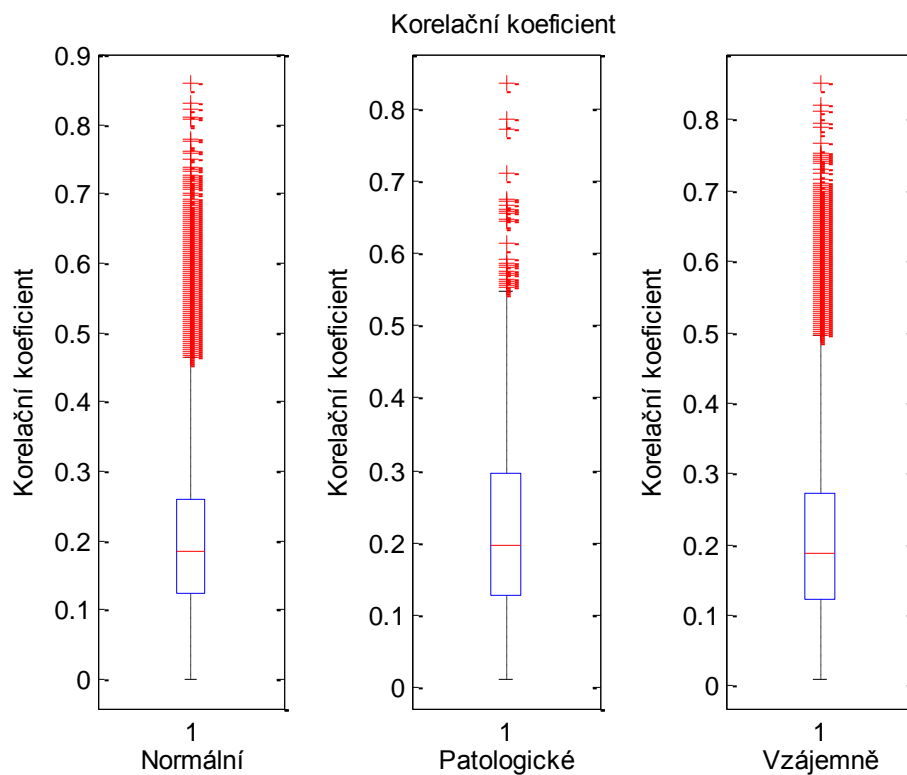
Tento postup jsme aplikovali na všechny signály aproximované různými metodami. Přestože postup není příliš přesný a nezaznamenává správně všechny decelerace, ve všech případech byl průměrný počet poklesů u patologických signálů vyšší než u normálních. V následné klasifikaci signálů se tedy zaměříme na identifikaci decelerací, jelikož pomocí nich lze zřejmě odlišit obě skupiny signálů.

Tabulka 3 Počty poklesů v úrovních pro jednotlivé metody

Metoda	Průměrný počet – normální signály	Průměrný počet – patologické signály
PAA	2.89	3.95
PLA	7.22	9.11
APCA	5.36	6.95
APLA	7.88	9.93

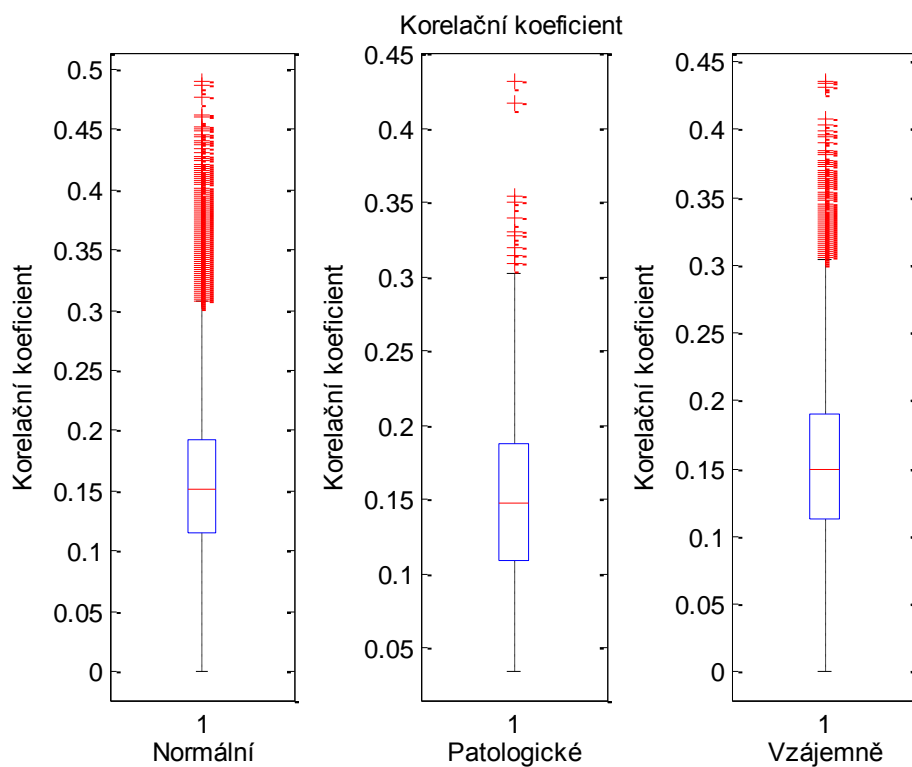
8.3. Korelační analýza

Korelace je přesnější mírou podobnosti signálů než klasická euklidovská vzdálenost. Korelační koeficient určuje míru lineární závislosti mezi signály. Nabývá hodnot od -1 do 1. Bereme v úvahu absolutní hodnotu koeficientu, jelikož chceme znát jakoukoli souvislost, i nepřímou. Koeficient jsme vypočítali jednak pro signály v původním tvaru, tak pro různé posuny signálů vůči sobě. V tomto případě jsme následně vybrali vždy nejlepší možnost. Pokud korelační koeficient překonal zvolenou prahovou hodnotu, označili jsme signály za vysoce korelované signály a přiřadili jsme je do jedné skupiny.



Obrázek 27 Korelační koeficienty PAA

Zatímco euklidovská vzdálenost vycházela menší u normálních signálů, korelační koeficient je o něco málo větší u patologických signálů. Z čehož lze vyvodit, že jsou si mezi sebou více podobné.



Obrázek 28 Korelační koeficienty – PLA

U metody PLA vychází korelace o něco nižší. Z grafu vidíme, že mezi korelačními koeficienty je množství odlehlých hodnot směrem nahoru. Signály s korelací větší než 0.4 jsou si podobné a pravděpodobně tak patří do stejné skupiny. Tímto způsobem je ale obtížné signály shlukovat, normální signály jsou totiž navzájem korelované i s patologickými, nelze tudíž vytvořit dvě oddělené skupiny.

9. Shlukování

Hlavní úkol této práce, klasifikaci signálů jsme provedli shlukováním. Vyzkoušeli jsme jak základní hierarchické shlukování, tak především algoritmus k-means. Jako vstup do shlukovacích metod jsme použili řady příznaků získané aproximací signálů, ale i četnosti výskytu podskupin znaků v aproximovaných signálech.

Úspěšnost klasifikátoru lze vyjádřit pomocí několika charakteristik, především specificity, senzitivity a přesnosti. Ty jsou definované následujícím způsobem (Štěpánková, 2013):

$$\text{Senzitivita} = \frac{\text{Počet správně pozitivních}}{\text{Počet správně pozitivních} + \text{Počet falešně negativních}}$$

$$\text{Specificita} = \frac{\text{Počet správně negativních}}{\text{Počet správně negativních} + \text{Počet falešně pozitivních}}$$

$$\text{Přesnost} = \frac{\text{Počet správně pozitivních} + \text{Počet správně negativních}}{\text{Počet všech prvků}}$$

Přesnost pro nás není příliš dobrým ukazatelem výkonnosti klasifikátoru, pokud dosáhneme např. vysoké senzitivity a nízké specificity, bude přesnost stejná jako u klasifikátoru s vyrovnanou senzitivitou a specificitou. Proto budeme celkovou výkonnost klasifikátoru hodnotit geometrickým průměrem specificity a senzitivity.

Pro zhodnocení výkonnosti klasifikátoru tedy použijeme geometrický průměr senzitivity a specificity. Geometrický průměr dvou čísel je definován jako odmocnina jejich součinu:

$$G(x_1, x_2, \dots, x_N) = \sqrt[N]{x_1 \cdot x_2 \dots x_N}$$

9.1. Shlukování metodou k-means

Algoritmus k-means je velice efektivní způsob shlukování. Lze ho použít i na vícerozměrná data, jako jsou časové řady. Každý vzorek signálu je považován za proměnnou příslušného prvku. Jedná se o iterativní algoritmus s poměrně vysokou výpočetní náročností. K-means obvykle volí počáteční centra shluků náhodně, proto se mohou shluky vytvořené ze stejných dat lišit. Lepších výsledků můžeme dosáhnout

správně předem zvolenými centry, nebo můžeme algoritmus vícekrát zopakovat a vybrat nejlepší výsledek. My jsme zvolili druhou možnost, výpočet vždy několikrát zopakujeme. K-means rovněž může využívat různé druhy vzdáleností, nám se jako nejúspěšnější jeví korelace a euklidovská vzdálenost.

Shlukovat můžeme buď aproximace signálů (upravené na stejnou délku), četnosti prvků ve slovníku komprese LZ78, nebo četnosti skupin znaků získaných z frekvenční analýzy.

Cílem shlukování je rozdělit signály do dvou skupin, na normální a patologické. Struktura dat může být tvořena více skupinami než jen dvěma, proto jsme postup vyzkoušeli s různými počty shluků. Patologické i normální signály tak může tvořit více než jeden shluk. V každém vytvořeném shluku vypočteme četnost prvků normálních a patologických. Shluky s největší četností normálních signálů pak označíme za normální a zbylé za patologické.

Jelikož je každý signál reprezentován více příznaky, pracuje k-means v multidimenzionálním prostoru. Abychom mohli výsledky vykreslit, použili jsme metodu hlavních komponent (PCA), která nám redukuje příznaky. Při použití dvou komponent pak jednotlivé signály můžeme vykreslit jako body.

9.1.1. Shlukování aproximací signálů

Algoritmem k-means budeme shlukovat především reprezentace signálů, které jsme získali metodami aproximace. Výpočet můžeme provádět s různým nastavením, a jelikož k-means určuje počáteční centra shluků náhodně, můžeme po každém spuštění zaznamenat mírně odlišný výsledek. Proto jsme výpočet zopakovali celkem 100x a vybrali výsledek s nejmenší celkovou vzdáleností uvnitř shluků. Tuto hodnotu budeme považovat za objektivní funkci, kterou určíme kvalitu vytvořených shluků.

PAA a SAX

Výsledky shlukování závisí na nastavení parametrů. Nejdůležitější je vhodně určit délku okna a jeho překryv, hranice intervalů a jejich počet a v neposlední řadě míru vzdálenosti. Jako první jsme shlukovali signály aproximované metodou PAA.

Patologické signály jsou označeny jako pozitivní nález, normální signály jsou tudíž nález negativní. Z vytvořených shluků jsme určili počty pozitivních a negativních prvků

a porovnali je s výsledky klasifikace podle hodnoty pH. Získali jsme následující kontingenční tabulku.

Tabulka 4 Kontingenční tabulka pro metody PAA a SAX

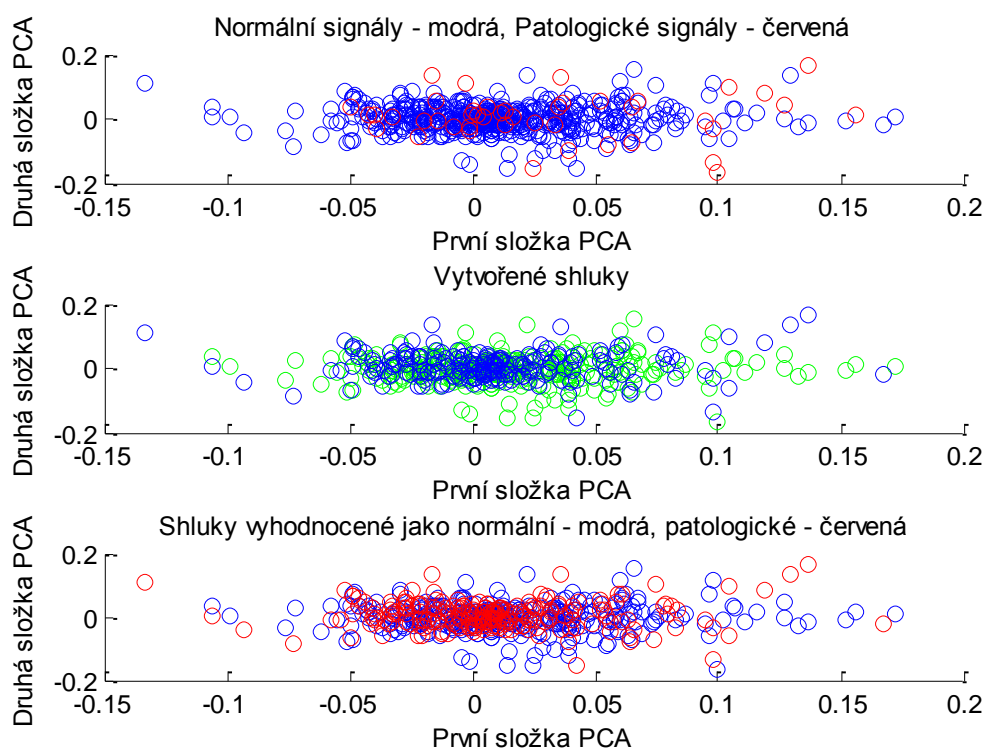
	Klasifikované pozitivní	Klasifikované negativní
Skutečně pozitivní	24	20
Skutečně negativní	192	316

Z této tabulky můžeme vypočítat veličiny určující úspěšnost klasifikátoru. Nejlépe klasifikátor charakterizují veličiny senzitivita a specificita.

$$\text{Senzitivita} = \frac{T(1,1)}{T(1,1)+T(1,2)} = 54.5 \%$$

$$\text{Specificita} = \frac{T(2,2)}{T(2,1)+T(2,2)} = 62.2 \%$$

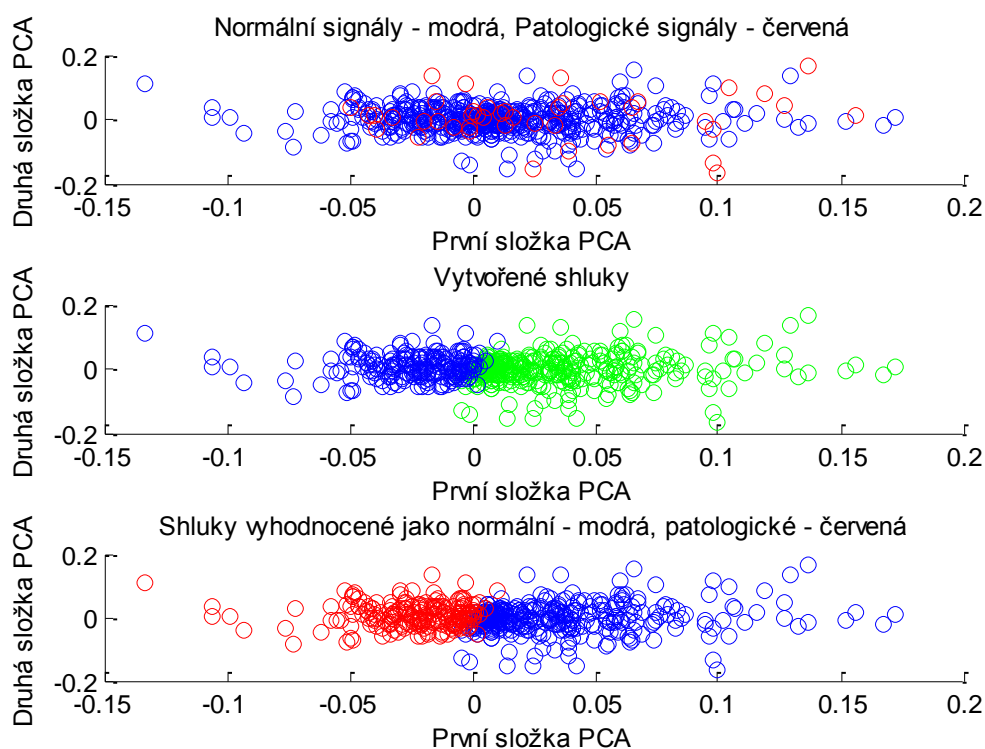
Celková vzdálenost uvnitř shluků: 46055



Obrázek 29 Výsledek k-means pro metody PAA a SAX

Nejllepších výsledků jsme dosáhli s délkou segmentu 16s a 50% překryvem oken. Při tomto nastavení je každý signál vyjádřen 112 příznaky, které spadají do 8 intervalů. Jako veličinu pro aproximaci jsme použili průměr. V k-means algoritmu byla podobnost určena euklidovskou vzdáleností.

Korelace jako míra podobnosti vytvořila oddělené nepřekrývající shluky. Protože jsou ale patologické signály rozptýleny mezi normální, není tato klasifikace pro patologické signály účinná. Korelace použitá jako míra podobnosti se vyznačuje poměrně nízkou senzitivitou. Shlukování metodou k-means pomocí korelace navíc nemůže být použito na prvky, které mají nízkou směrodatnou odchylku jednotlivých proměnných, proto jsme tyto signály vyřadili z analýzy. Počet vyřazených signálů z celé databáze závisí na velikosti okna.



Obrázek 30 Výsledek k-means pro metody PAA a SAX- korelace

Senzitivita: 34.1 %

Specificita: 62 %

Celková vzdálenost uvnitř shluků: 399.7

Postup s euklidovskou vzdáleností zopakovali i pro jiné počty shluků. Ukázalo se, že nejlepší počet shluků je 6, přičemž normální signály tvoří 3 shluky a 3 shluky tvoří patologické. Počet intervalů pro aproxiimaci je v tomto případě 10.

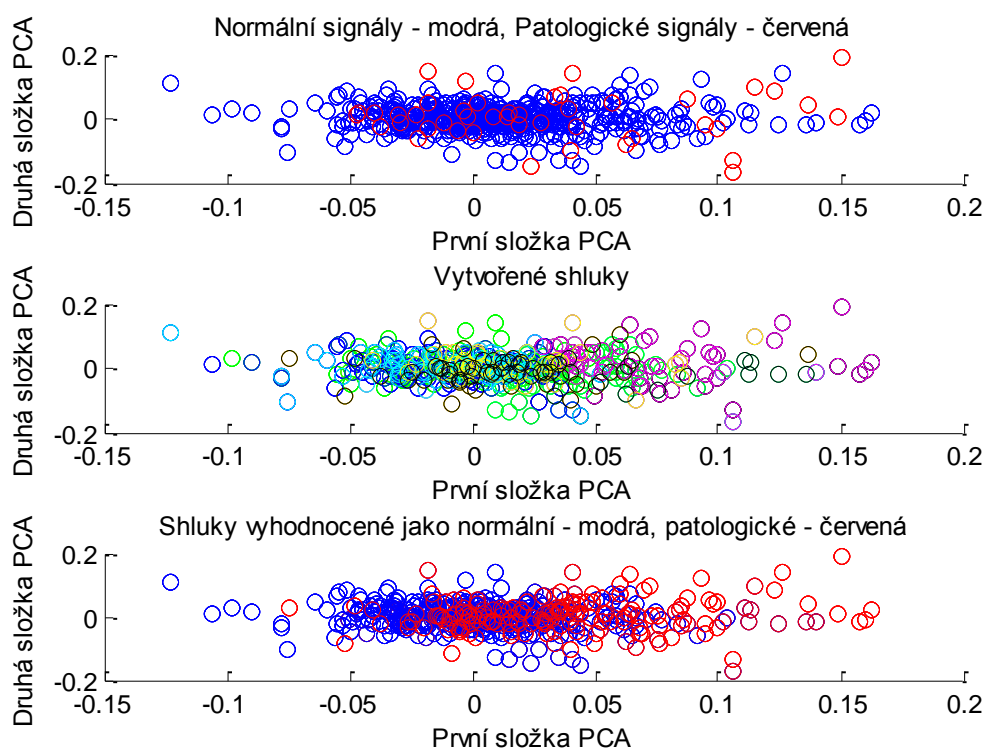
Tabulka 5 Kontingenční tabulka pro metody PAA a SAX - 6 shluků

	Klasifikované pozitivní	Klasifikované negativní
Skutečně pozitivní	28	16
Skutečně negativní	147	361

Senzitivita: 63.6 %

Specifická: 71 %

Celková vzdálenost uvnitř shluků: 55587



Obrázek 31 Výsledek k-means pro metody PAA a SAX - 6 shluků

PLA

U této metody očekáváme lepší výsledky než v předchozím případě, PLA totiž lépe charakterizuje data. Pro výsledek shlukování je v tomto případě důležitý hlavně

výběr počtu intervalů a jejich hranic. Stejně jako u PAA používáme délku okna 16s a poloviční překryv. Kratší okna poskytují také dobré výsledky, ale velmi se zvyšuje doba výpočtu. Příznaky jsou zařazeny do 11 intervalů Jako vhodnější míra podobnosti se na rozdíl od PAAjevila korelace.

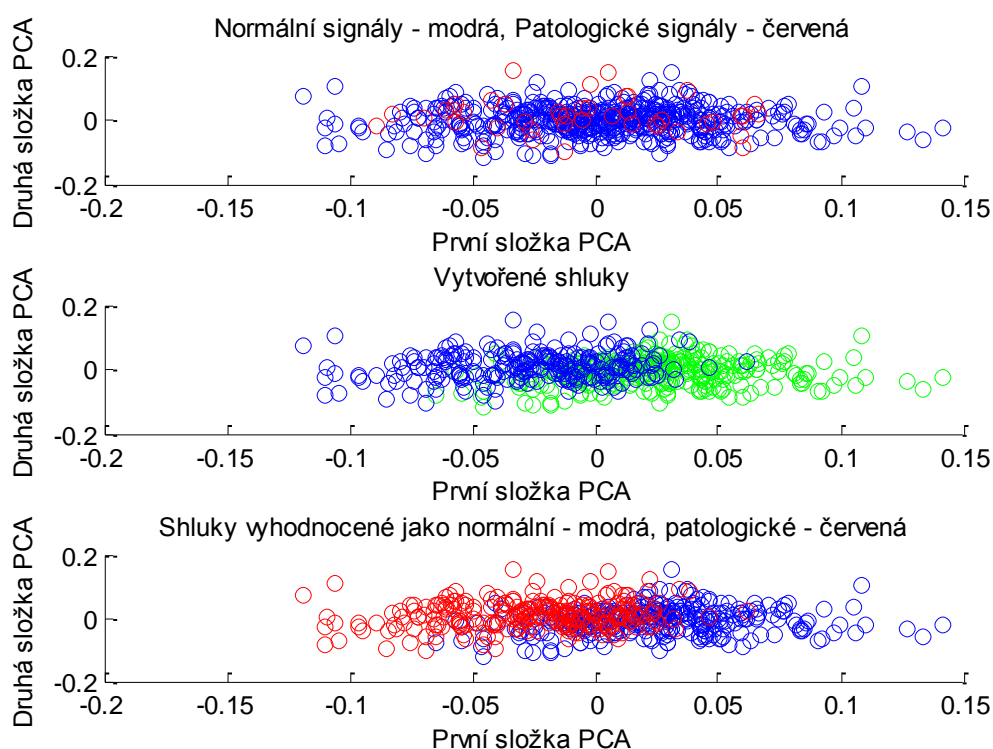
Tabulka 6 Kontingenční tabulka pro metodu PLA

	Klasifikované pozitivní	Klasifikované negativní
Skutečně pozitivní	28	16
Skutečně negativní	228	280

Senzitivita: 63.6 %

Specifická: 55.1 %

Celková vzdálenost uvnitř shluků: 481.2



Obrázek 32 Výsledek k-means pro metodu PLA

I v tomto případě jsme zkusili zvýšit počet shluků. Pokud normální signály tvoří dva shluky a dva patologické, dospějeme k stejnému výsledku. Více shluků sice výsledky příliš nezlepšilo, ale dochází k menším rozdílům při opakovaných iteracích.

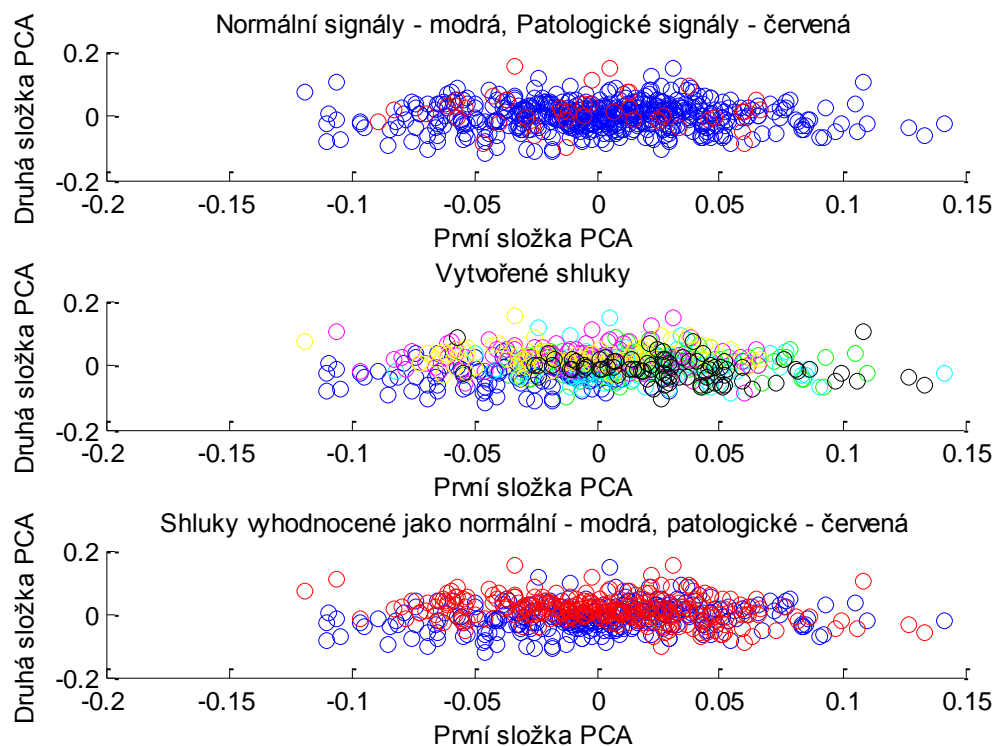
Tabulka 7 Kontingenční tabulka pro metodu PLA - 6 shluků

	Klasifikované pozitivní	Klasifikované negativní
Skutečně pozitivní	29	15
Skutečně negativní	240	268

Senzitivita: 65.9 %

Specifická: 52.8 %

Celková vzdálenost uvnitř shluků: 424.8



Obrázek 33 Výsledek k-means pro metodu PLA - 6 shluků

APCA

Reprezentace signálů adaptivními metodami poskytuje dvě informace pro každý segment. Interval, do kterého segment spadá, a také délka segmentu. Do shlukování

jako atributy použijeme pouze čísla intervalů. Jelikož je každá reprezentace jinak dlouhá, podle nejkratší z nich zkrátíme všechny ostatní. Jako parametr zde nastavujeme pouze velikost maximální chyby, při které se vytvoří segment. Vyzkoušeli jsme proto různé nastavení počtu shluků a velikosti chyby. Jako míru podobnosti jsme opět použili euklidovskou vzdálenost.

Tabulka 8 Výsledek klasifikace pro metodu APCA

Chyba: 200	2 shluky	4 shluky
Senzitivita	56.8%	45.5 %
Specifická	62 %	67.1 %
Celková suma vzdáleností	33060	28330
Chyba: 300	2 shluky	4 shluky
Senzitivita	61.4 %	45.5 %
Specifická	61 %	68.5 %
Celková suma vzdáleností	25181	21548
Chyba: 400	2 shluky	4 shluky
Senzitivita	59.1 %	43.2 %
Specifická	62.6 %	68.9 %
Celková suma vzdáleností	19917	16934

Z těchto výsledků jsme zjistili, že ideální velikost chyby je kolem 300 vzorků. Zvýšení počtu shluků opět výsledky příliš zlepšilo. Použili jsme u této metody 10 intervalů. Počet příznaků použitých pro každý signál (25 příznaků) byl u této metody výrazně nižší než u metod se stálou délkou okna. Úspěšnost je přitom srovnatelná.

APLA

Také tato metoda je adaptivní, proto reprezentace normujeme podle nejkratšího z nich. To značně zredukuje počet příznaků. Stejně jako u PLA byla použita korelace. Úspěšnosti klasifikace je v tomto případě nižší než u předchozí adaptivní metody. Nejlepší výsledky byly dosaženy s chybou 200 vzorků.

Tabulka 9 Výsledek klasifikace pro metodu APLA

Chyba: 200	2 shluky	4 shluky
Senzitivita	68.2 %	61.4 %
Specifická	50 %	50.9 %
Celková suma vzdáleností	462.8	419
Chyba: 300	2 shluky	4 shluky
Senzitivita	52.3 %	61.4 %
Specifická	51.2 %	54.5 %
Celková suma vzdáleností	452.1	403.7
Chyba: 400	2 shluky	4 shluky
Senzitivita	52.4 %	63.6 %
Specifická	50.7 %	51.9 %
Celková suma vzdáleností	443.3	389.2

9.1.2. Shlukování signálů po kompresi LZ78

Výstupem komprese algoritmem LZ78 jsou komprimované signály a slovník, pomocí něhož můžeme signály zpět rekonstruovat. Jelikož vektory komprimací pro jednotlivé signály nejsou stejně dlouhé, rozhodli jsme se shlukovat četnosti výskytů prvků slovníku v komprimacích. V signálech by se měly vyskytovat skupiny znaků charakteristické pro danou skupinu, mohlo by být tudíž možné odlišit patologické signály od normálních. Pro kompresi jsme použili signály aproximované metodou PAA.

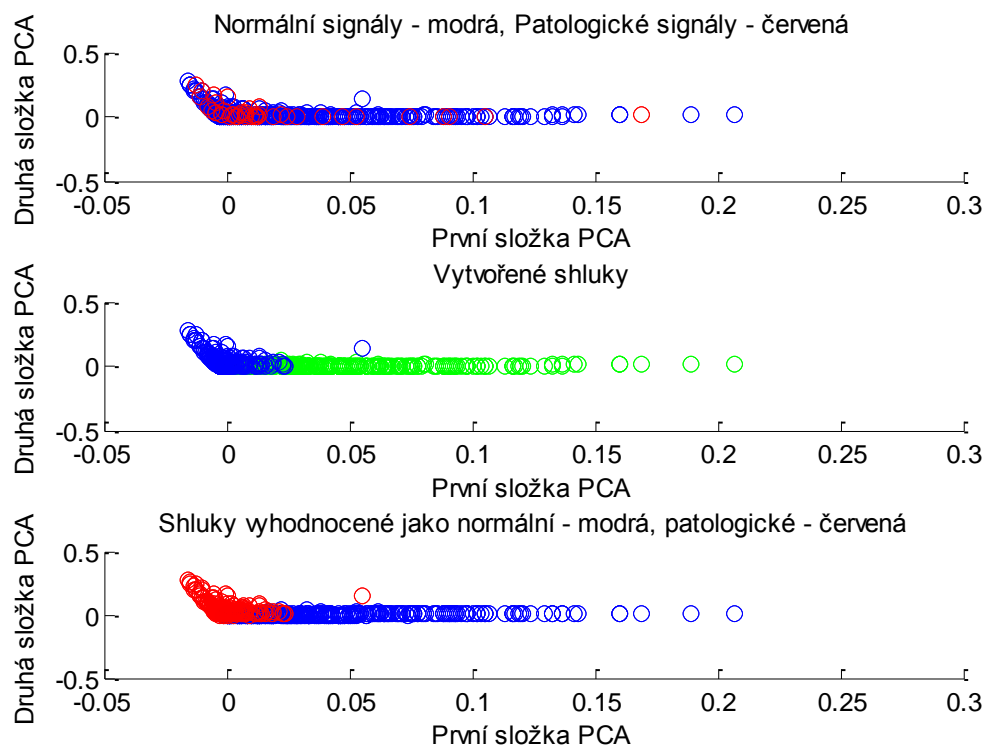
Tabulka 10 Kontingenční tabulka pro metodu LZ

	Klasifikované pozitivní	Klasifikované negativní
Skutečně pozitivní	26	18
Skutečně negativní	192	316

Senzitivita: 59.1 %

Specifická: 62.2 %

Celková suma vzdáleností uvnitř shluků: 184.1



Obrázek 34 Výsledek k-means pro metodu LZ - korelace

Algoritmus jsme použili i na další metody aproximace. Získali jsme následující výsledky. Z různých nastavení parametrů jsme vybrali tu nejlepší kombinaci.

Tabulka 11 Výsledek klasifikace pro metodu LZ

Metoda	Senzitivita	Specifická	Suma vzdáleností
PLA -2 shluky, korelace, segment 16s	54.5 %	55.3 %	97.6
APCA -2 shluky, korelace, chyba 300	56.8 %	61.4 %	162.8
APLA -2 shluky, euklidovská vzdálenost, chyba 200	47.7 %	54.5 %	18562

9.1.3. Shlukování strukturních příznaků

Poslední řadou příznaků, kterou jsme použili do algoritmu k-means, je reprezentace časových řad podle Lin a Yuan (2009). Tato reprezentace je založená

na symbolické aproximaci a metodě Bag – of - Words. Metoda efektivně popisuje strukturu dat a porovnává je na základě strukturních příznaků. Přístup Bag – of - Words se používá hlavně pro získávání informací z textových dokumentů, ale lze ho využít i ke zpracování medicínských dat. Biologické signály jsou obvykle dlouhé a nezarovnané, proto u nich standardní postupy pro porovnání časových řad nejsou úspěšné.

Způsob získání příznaků metodami SAX a Bag – of - Words je do jisté míry podobný algoritmu komprese LZ78. Ten vytváří skupiny znaků, jejichž četnosti můžeme shlukovat. Bohužel ale nevyčerpává všechny možné kombinace znaků. V této metodě nejdříve znormujeme vybrané 30 minutové sekvence signálu na nulovou střední hodnotu a jednotkovou směrodatnou odchylku. Normalizace signálů nemusí vést ke zlepšení výsledku, odečtením střední hodnoty totiž přijdeme o informaci o bazální tepové frekvenci, která může být charakteristická pro normální i patologické signály. Dále signály aproximujeme metodou SAX, přičemž použijeme abecedou o 4 znacích. Dále stejným postupem jako ve frekvenční analýze zjistíme četnosti všech možných kombinací 3 znaků v aproximacích. Výsledkem bude tabulka četností jednotlivých skupin znaků pro každý signál. Tu použijeme jako sadu příznaků pro shlukování. Získali jsme nejlepší výsledky pro délku okna 30s a euklidovskou vzdálenost, pro korelaci byla nejlepší délka okna 1 min.

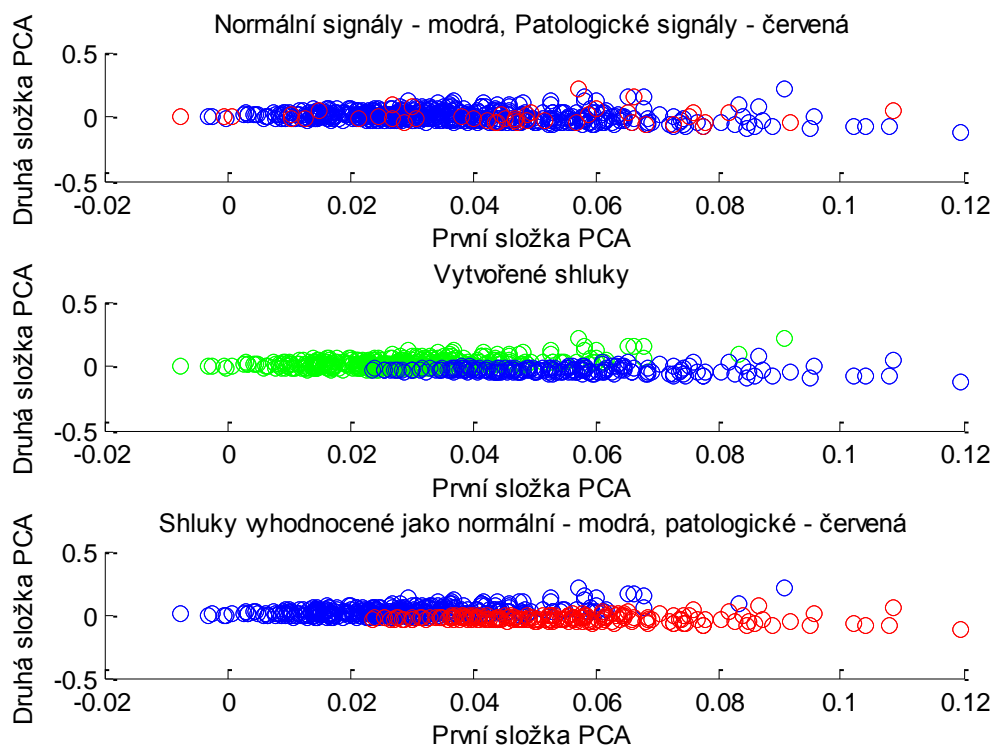
Tabulka 12 Kontingenční tabulka pro metodu Bag of Words

	Klasifikované pozitivní	Klasifikované negativní
Skutečně pozitivní	24	20
Skutečně negativní	169	339

Senzitivita: 54.5 %

Specificita: 66.7 %

Celková suma vzdáleností uvnitř shluků: 47737.9



Obrázek 35 Výsledek k -means pro metodu Bag - of - Words

9.2. Hierarchické shlukování

Hierarchické shlukování sdružuje signály s nejmenší vzájemnou vzdáleností. Postupujeme tedy aglomerativním způsobem. Vzdáleností mohou být některé z Minkovského vzdáleností (Euklidovská, Manhattan), korelace nebo Hammingova vzdálenost, která se často používá pro nenumernická data. Můžeme nastavit hranici, při které se prvky přiřadí k sobě. Výstupem hierarchického shlukování je diagram ve tvaru stromu - dendrogram. Je na uživateli, jak tento diagram interpretuje. Je ovšem možné nastavit počet shluků, které má diagram obsahovat.

9.2.1. Shlukování aproximací

Nejprve jsme hierarchicky shlukovali řadu příznaků získanou z aproximačních metod. Jako vzdálenosti jsme postupně vyzkoušeli euklidovskou vzdálenost a korelaci a porovnali jejich výsledky. Stromový diagram celé databáze je dost nepřehledný z důvodu velkého množství dat, proto jsme pro shlukování použili 16 vzorových signálů. V tomto případě dostaneme poměrně stručný diagram.

PAA, SAX a APCA

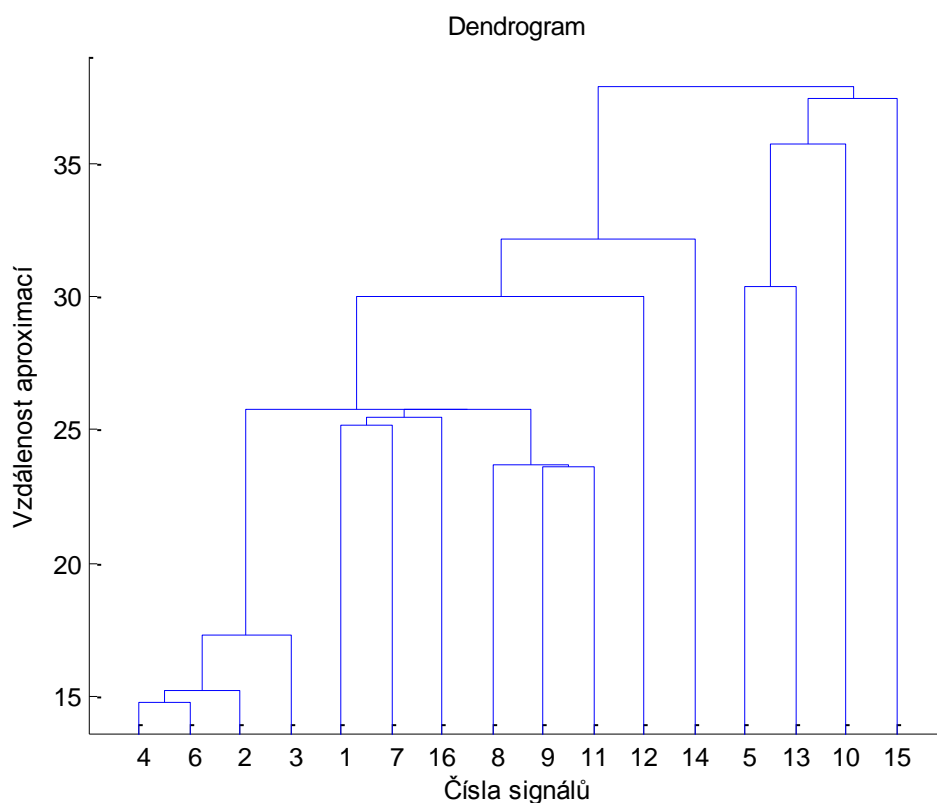
Pomocí euklidovské vzdálenosti jsme získali následující dendrogram. Z něho je patrné, že normální většina signálů označených čísly 1-8 byla správně zařazena do jednoho shluku a signály s vyššími čísly do druhého. Levou polovinu stromu můžeme označit za jeden shluk a pravou polovinu za druhý.

Tabulka 13 Kontingenční tabulka pro metodu PAA

	Klasifikované pozitivní	Klasifikované negativní
Skutečně pozitivní	7	1
Skutečně negativní	1	7

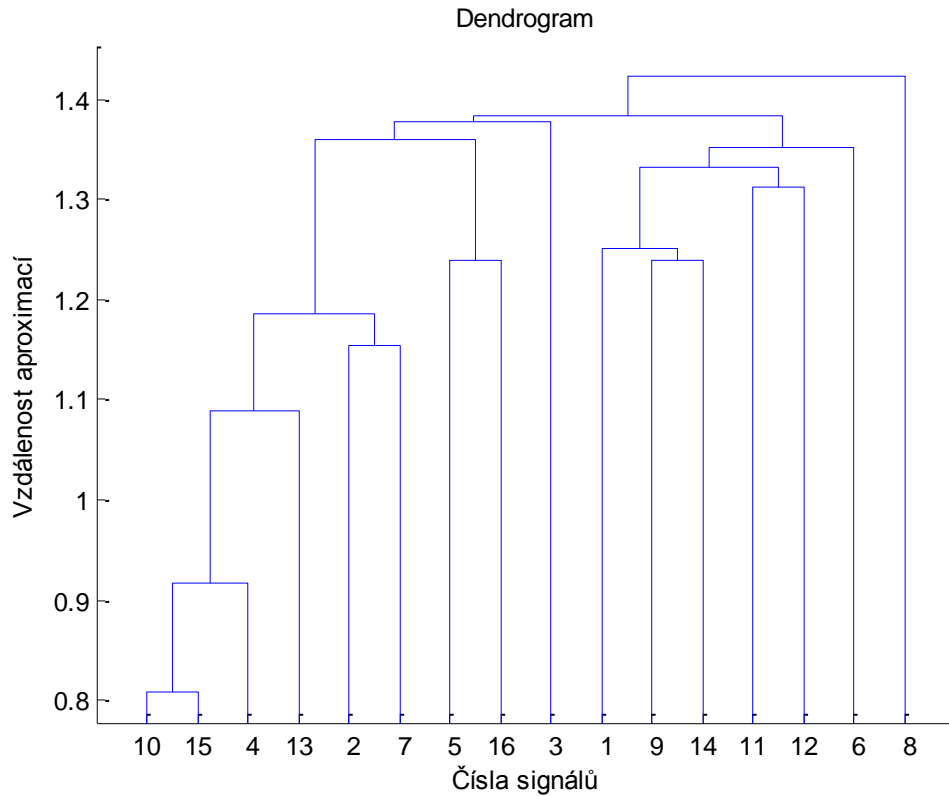
Senzitivita: 87.5 %

Specifická: 87.5 %



Obrázek 36 Diagram pro metodu PAA - euklidovská vzdálenost

Korelace jako míra podobnosti už tak dobré výsledky nepřináší. Výkonnost klasifikace by dosáhla pouze 50 %.

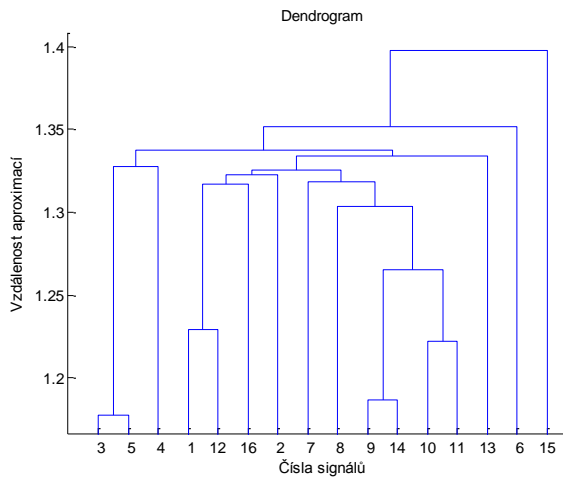


Obrázek 37 Diagram pro metodu PAA – korelace

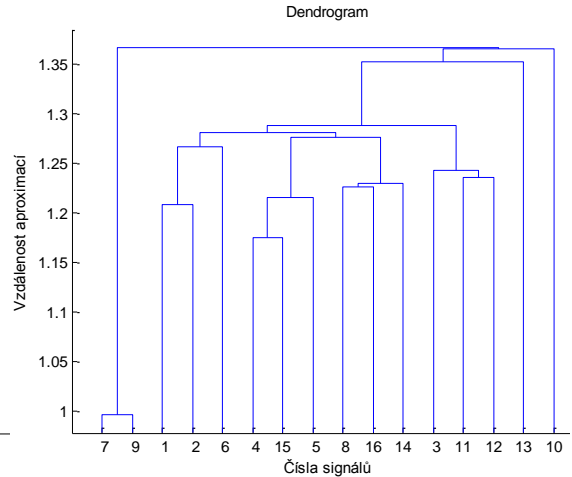
V případě adaptivní verze PAA máme k dispozici menší počet příznaků, proto jsou i výsledky o něco horší. Obě míry podobnosti zařazují do jedné skupiny normální signály s patologickými. Nelze také rozdělit signály na dvě stejně velké skupiny.

PLA a APLA

Obě tyto metody rozdělují signály poměrně správně. Stejně jako u shlukování k-means i zde pro tyto metody byla vhodnější korelace. Úspěšnost klasifikace je v obou případech 75%.



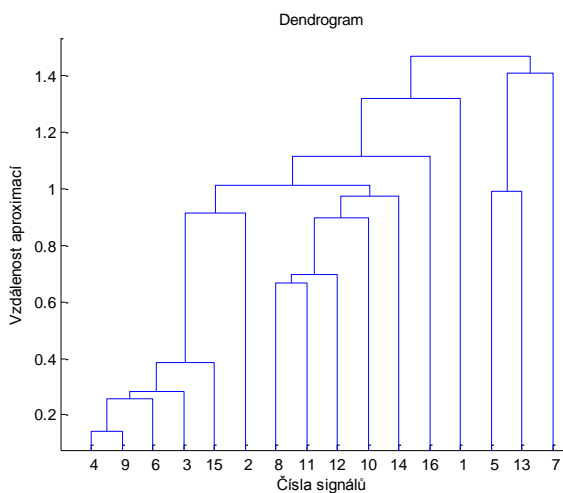
Obrázek 38 Diagram pro metodu APLA – korelace



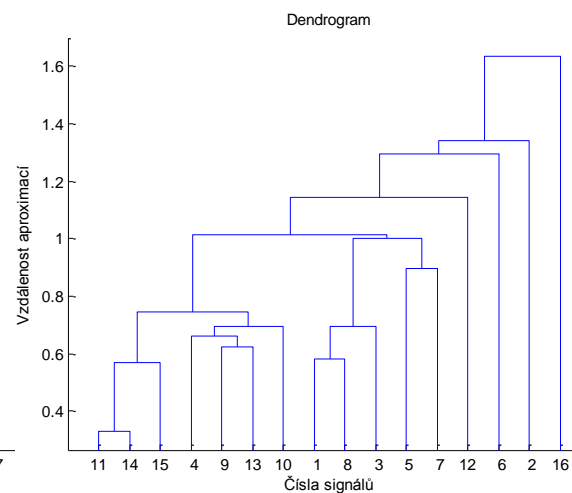
Obrázek 39 Diagram pro metodu PLA – korelace

9.2.2. Shlukování prvků slovníku LZ a strukturních příznaků

Signály aproximované metodou PAA jsme zkomprimovali algoritmem LZ78. Stejně jako u shlukování k-means jsme jako příznaky pro shlukování použili četnosti prvků slovníku v jednotlivých signálech. S korelací jsme dosáhli úspěšnosti 62,5%. Také pro metodu Bag of Words jsme zvolili korelaci s úspěšností 75%.



Obrázek 40 Diagram - Bag - of – Words



Obrázek 41 Diagram pro metodu LZ - korelace

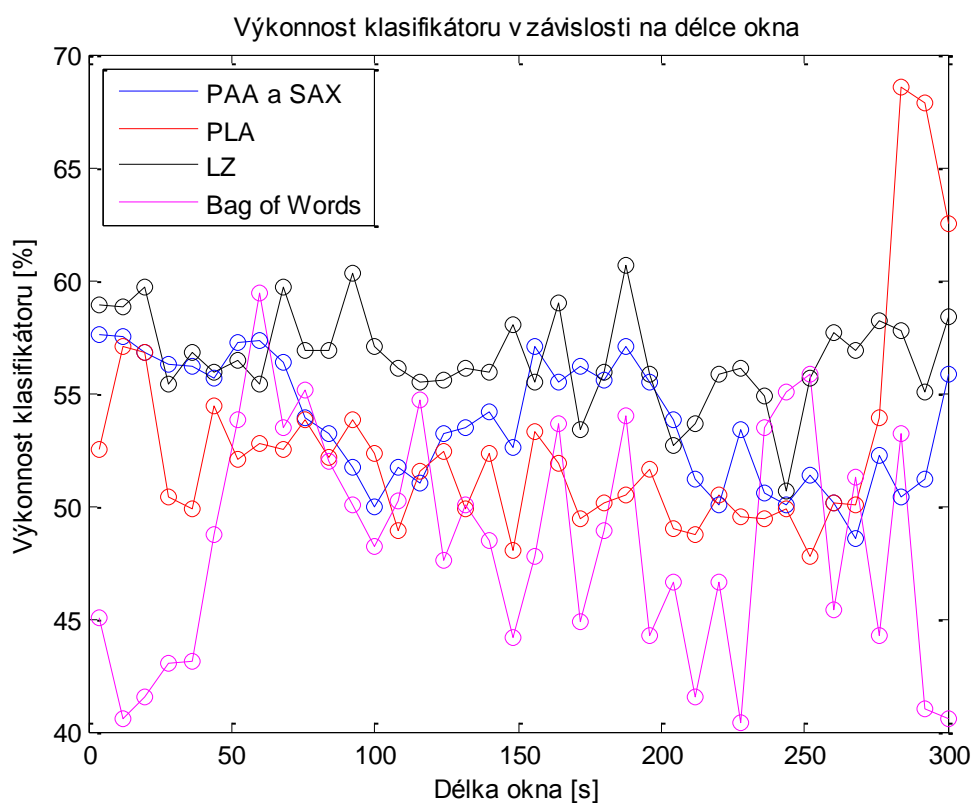
9.3. Závislost výsledků na parametrech

Předchozí výsledky jsme získali po vyzkoušení různých kombinací parametrů, vybrali jsme vždy nejlepší možné nastavení. Mohli bychom tedy zjistit, jakou závislost

má úspěšnost klasifikátoru na změně jednotlivých parametrů. Všechny grafy závislosti platí pro shlukovací metodu k-means.

9.3.1. Závislost na délce segmentu

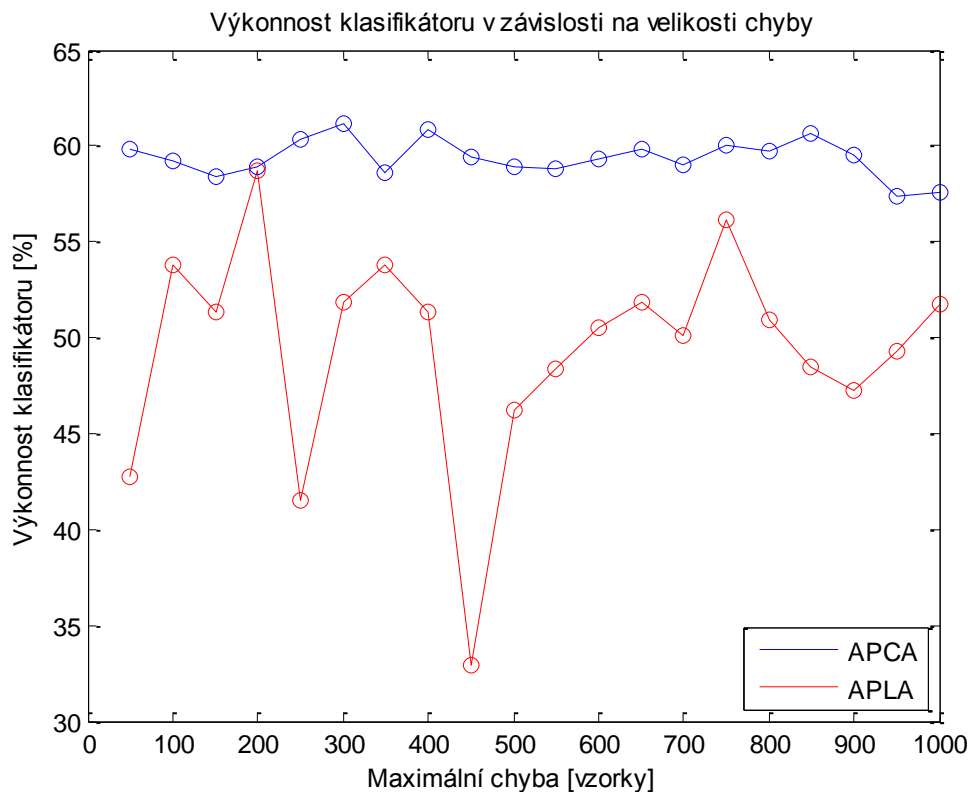
Délka okna, které používáme pro aproximaci, je pravděpodobně nejdůležitějším parametrem. Závislost jsme vykreslili pro délky segmentu od 4s do 5min. Použili jsme dva shluky a euklidovskou vzdálenost. Výkonnost klasifikátoru je zde reprezentována geometrickým průměrem specifity a senzitivity. Algoritmus jsme spustili vždy 100 krát za sebou a vybrali výsledek s nejmenší objektivní funkcí (celková vzdálenost uvnitř shluků).



Obrázek 42 Závislost úspěšnosti klasifikace na délce okna

Z grafu je zřejmé, že senzitivita a specifita se u většiny metod v závislosti na délce okna příliš nemění, pouze pozvolna klesá. U metody PLA se zvětšující se délkou okna úspěšnost výrazně stoupá, to je ovšem způsobeno tím, že při použití dlouhého okna je velká část prvků vyřazena z analýzy kvůli malé odchylce. Pro naši délku okna 16s dostáváme vysoké výsledky u všech metod kromě Bag - of - Words, kde je výkonnost nízká. Nejlepší volbou je v tomto případě okno o délce 1 min. Největší skoky v průběhu jsme zaznamenali u metody LZ.

Pro adaptivní metody jsme sledovali závislost úspěšnosti na velikosti chyby, při které se vytvoří segment. Testovali jsme chyby od 50 do 1000 vzorků.

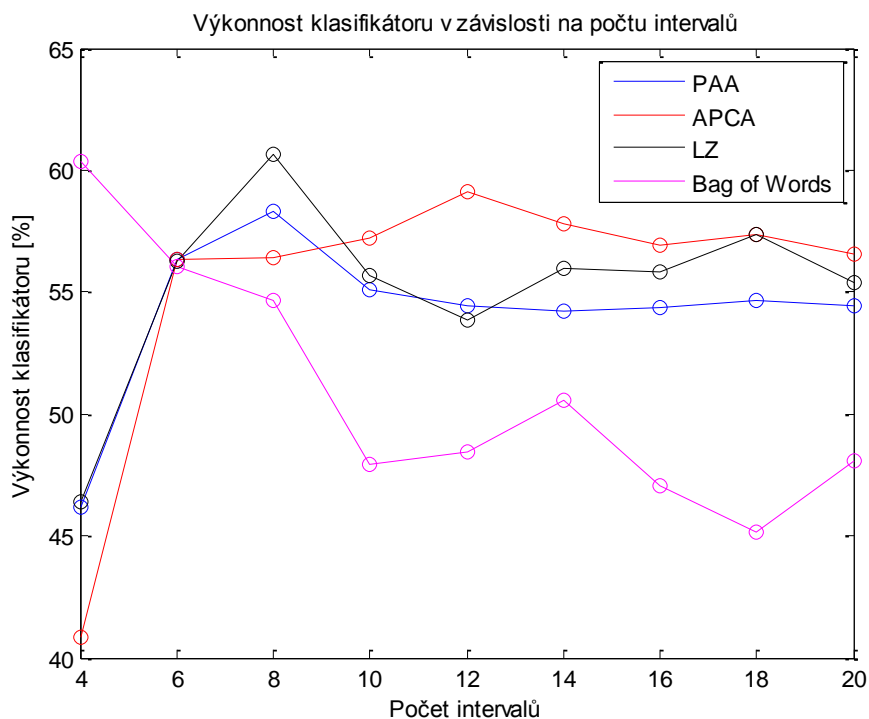


Obrázek 43 Závislost úspěšnosti klasifikace na velikosti chyby

Úspěšnost je téměř stejná pro všechny velikosti chyby u metody APCA, nejlepší výsledky jsou u chyby 300 a 400 vzorků, pro metodu APLA se v průběhu objevují poklesy u chyb 250 a 450 vzorků, které jsou způsobeny nízkou senzitivitou, tedy špatnou klasifikací patologických signálů. Nejlepší výsledek klasifikace je pro chybu 200 vzorků.

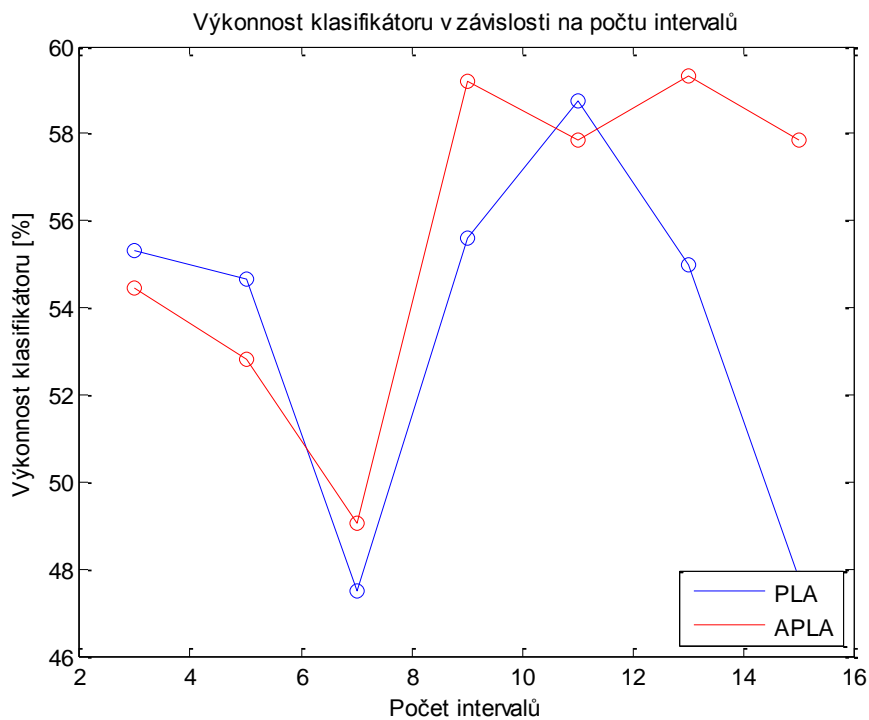
9.3.2. Závislost na počtu intervalů

U metod, kde je charakteristikou popisující segment průměr, jsme měnili počty intervalů, do nichž může průměr spadat. Vyzkoušeli jsme počty intervalů od 4 do 20. Téměř ve všech případech úspěšnost rostla do počtu 6 intervalů, pouze u metody Bag of Words vychází nejlépe 4 intervaly. Pro metodu LZ je nejlepších 8 intervalů, pro metodu APCA dokonce 12.



Obrázek 44 Závislost úspěšnosti klasifikace na počtu intervalů

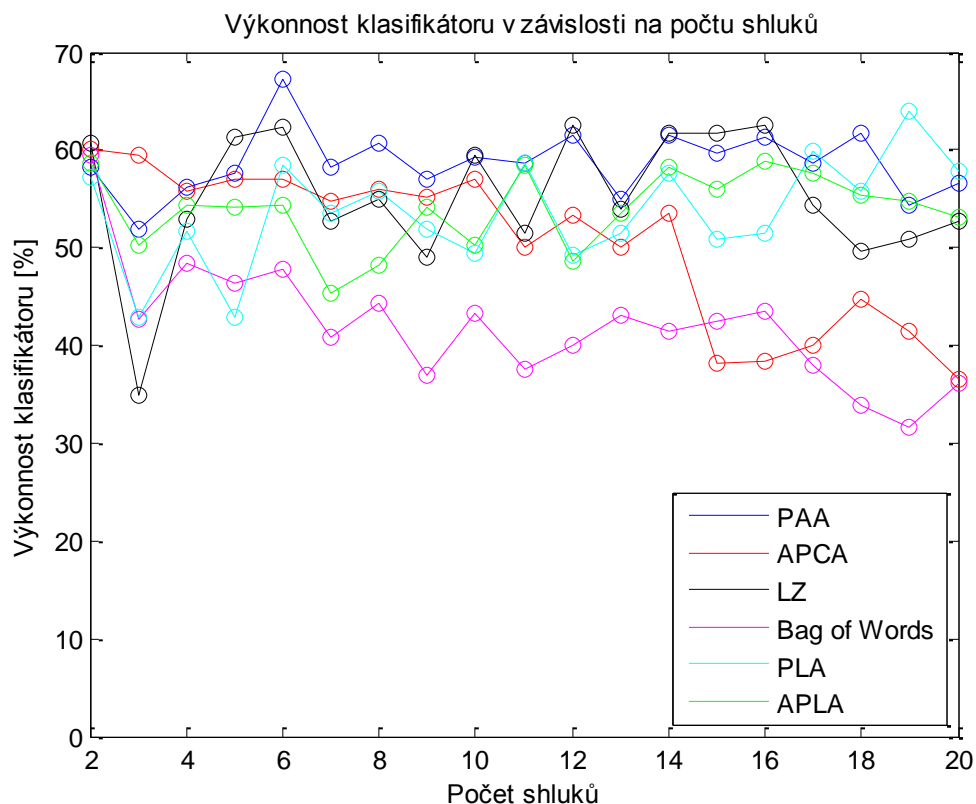
U metod lineární aproximace jsme měnili počty intervalů pro úhly od 3 do 15. Výrazně se mění pouze senzitivita u metody PLA a mírně i senzitivita metody APLA. Nejlepšího výsledku jsme dosáhli s 9 nebo 11 intervaly.



Obrázek 45 Závislost úspěšnosti klasifikace na počtu intervalů

9.3.3. Závislost na počtu shluků

Úspěšnost klasifikace nakonec závisí i na počtu shluků. Pro všechny metody jsme vyzkoušeli od minimálního počtu 2 shluků do 20. Téměř ve všech případech s počtem shluků roste specifita a klesá senzitivita. To je způsobeno tím, že shluky klasifikujeme podle četností normálních signálů ve shluku. Pro dva shluky dostaneme uspokojivý výsledek u všech metod. Úspěšnost je většinou nižší u lichých počtů shluků, protože v tom případě jsou normální signály tvořeny více shluky než signály patologické. Nejlepší výsledek vůbec pak dostaneme pro 6 shluků u metody PAA a SAX.



Obrázek 46 Závislost úspěšnosti klasifikace na počtu shluků

10. Závěr

Úkolem této práce bylo klasifikovat reálné FHR záznamy získané z kardiokografického vyšetření. KTG je v současné době nejběžnější metoda diagnostiky fetální hypoxie, která má význam obzvláště v monitoraci plodu během porodu. Nevýhodou této metody je vysoký počet falešně pozitivních nálezů, které vedou ke zbytečným zásahům. V interpretaci KTG záznamů se objevuje i velká inter a intrapersonální variabilita, vznikly proto snahy o vytvoření standardizovaného postupu hodnocení. To se snaží řešit metody automatické analýzy FHR signálů. V této práci jsme vytvořili klasifikátor založený na shlukovacích metodách. Shlukování patří mezi algoritmy s učením bez učitele, nepotřebujeme tedy znát výsledek klasifikace dat.

Zpracovávali jsme KTG signály z databáze CTU-UHB, která obsahuje 552 záznamů naměřených během porodu. Nás zajímaly pouze FHR signály a údaje o hodnotě pH novorozence, podle které jsme určili výsledek porodu. Pro výsledek klasifikace je důležité předzpracování signálů, tedy odstranění artefaktů a výběr úseku pro analýzu. Do shlukovacích metod vstupují signály popsané řadou příznaků. Pro extrakci příznaků jsme na signály aplikovali několik aproximačních metod založených na průměru segmentů (PAA) a lineární interpolaci (PLA). Hlavní použitou metodou je symbolická aproximace (SAX), která je v současné době často používána v analýze časových řad. Výhodou námi použitých metod je jejich jednoduchost a nenáročnost. Tento přístup nám umožňuje velmi efektivně popsat signál, aniž bychom znali jeho přesnou strukturu. Na aproximace signálů jsme dále aplikovali kompresní algoritmus LZ78 a metodu získávání příznaků Bag of Words.

Na takto zredukované signály jsme použili aglomerativní hierarchické shlukování a algoritmus k-means. Výsledek klasifikace jsme porovnali s hodnotou pH zjištěnou po porodu. S použitím dvou shluků se nám podařilo implementovat klasifikátor se senzitivitou 63.6% a specificitou 55.1%, výkonnost klasifikátoru je tedy srovnatelná s jinými metodami. Běžně se CTG signály klasifikují do dvou skupin, my jsme ale zjistili, že je výhodnější použít více shluků, které lépe reprezentují normální a patologické signály. S použitím 6 shluků (3 pro normální a 3 pro patologické) jsme dosáhli senzitivity 63.4% a specificity 71%. Nejlepších výsledků jsme dosáhli s příznaky získanými metodou PAA a z ní odvozeným SAXem. Výsledek shlukování

závisí především na nastavení parametrů, nejlépe vyšla segmentace signálu oknem dlouhým 16s. Jako míry podobnosti jsme použili euklidovskou vzdálenost a korelaci.

Výsledky této práce byly zaslány na konferenci Computing in Cardiology 2014. Podařilo se nám splnit vytyčené cíle a pomocí jednoduché extrakce příznaků z dat a základních metod shlukování vytvořit klasifikátor srovnatelný s mnohem složitějšími metodami.

Použitá literatura

AACH, J., CHURCH, G. M.: Aligning gene expression time series with time warping algorithms, *Oxford Journals, Bioinformatics*, svazek 17, číslo 6, str. 495-508, 2001, ISSN 1460-2059

AHMADPOUR-KACHO, M., ASNAFI, N., JAVADIAN, M., HAJIAHMADI, M., HOSSEINI TALEGHANI, N.: Correlation between Umbilical Cord pH and Apgar Score in High-Risk Pregnancy, *Iran J Pediatr*, svazek 20, číslo 4, str. 401-406, 2010, ISSN 2008-2142,

AMIT, G., GAVRIELY, N., INTRATOR, N.: Cluster analysis and classification of heart sounds, *Biomedical Signal Processing and Control*, svazek 4, číslo 1, str. 26-36, 2009

ANGELES-YRETA, A., SOLÍS-ESTRELLA, H., LANDASSURI-MORENO, V., FIGUEROA-NAZUNO, J.: Similarity Search In Seismological Signals, ENC 04 Proceedings of the Fifth Mexican International Conference in Computer Science, str. 50-56, 2004, ISBN:0-7695-2160-6

ARROYUELO, D., NAVARRO, G.: Space-efficient construction of Lempel-Ziv compressed text indexes, *Information and Computation*, svazek 209, číslo 7, str. 1070 - 1102, 2008

BARBU, T.: Unsupervised SIFT-based Face Recognition Using an Automatic Hierarchical Agglomerative Clustering Solution, *Procedia Computer Science*, svazek 22, str. 385-394, 2013

BERNARDES, J., MOURA, C., MARQUES DE SA, J. P., PEREIRA LEITE, L.: The Porto system for automated cadriocographic signal analysis, *J Perinat Med*, svazek 19, str. 61-65, 1991, ISSN: 0300-5577

CAHILL, A., GONZALEZ, C., DEYCH, E., SHANNON, W., MACONES, G.: Functional data analysis (FDA) of intrapartum fetal heart rate (FHR) patterns to predict outcomes, *American Journal of Obstetrics & Gynecology*, str. 321, leden 2014

CHAE, S. S., KIM, C., KIM, J. M., WARDE, W. D.: Cluster analysis using different correlation coefficients, *Statistical Papers*, svazek 49, číslo 4, str. 715 – 727, 2008, ISSN: 0932-5026

CHAKRABARTI, K., KEOGH, E., MEHROTRA, S., PAZZANI, M.: Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases, *ACM Transactions on Database Systems*, svazek 27, číslo 2, str. 188-228, 2002

CHANG, J. S., YOO, C. S., YI, S. H., HONG, K. H., OH, H. S., HWANG, J. Y., KIM, S. G., AHN, Y. M., KIM, Y. S.: Differential pattern of heart rate variability in patients with schizophrenia, *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, svazek 33, číslo 6, str. 991-995, 2009

- CHUDÁČEK, V., SPILKA, J., BURŠA, M., JANKŮ, P., HRUBAN, L., HUPTYCH, M., LHOTSKÁ, L.: Open access intrapartum CTG database, BMC Pregnancy and Childbirth, svazek 14, článek 16, leden 2014
- CHUDÁČEK, V., SPILKA, J., HUPTYCH, M., GEORGOULAS, G., LHOTSKÁ, L., STYLIOS, C., KOUCKÝ, M., JANKŮ, P.: Linear and Non-Linear Features for Intrapartum Cardiotocography Evaluation, Computing in Cardiology, str. 999-1002, 2010, ISSN 0276-6547
- CILIBRASI, R., VIT'ANYI, P. M. B.: Clustering by Compression, IEEE Transactions on Information Theory, svazek 51, číslo 4, str. 1523 – 1545, ISSN 0018-9448
- COSTA SANTOS, C., BERNADES, J., VITÁNIY, P. M.B., ANTUNES, L.: Clustering Fetal Heart Rate Tracings by Compression, 19th IEEE Symposium on Computer-Based Medical Systems, str. 685-690, 2006, ISSN 1063-7125
- DAOUD, M., RAVIER, P., HARBA, R., JABLOUN, M., YAGOUBI, B., BUTTELLI, O.: HRV spectral estimation based on constrained Gaussian modeling in the nonstationary case, Biomedical Signal Processing and Control, svazek 8, číslo 6, str. 483-490, listopad 2013
- DESHPANDE, R. A., RAMDASI, D. S.: HRV Analysis Using Wavelet Transform and Wigner – Ville Transform, IEEE EMBS Conference on Biomedical Engineering and Sciences, str. 52-56, prosinec 2012, ISBN: 978-1-4673-1664-4
- DI GIUSEPPE, M. G., TROIANO, A., TROISE, C., DE NATALE, G.: k-Means clustering as tool for multivariate geophysical data analysis. An application to shallow fault zone imaging, Journal of Applied Geophysics, svazek 101, str. 108-115, únor 2014
- DONOSO, F. I., FIGUEROA, R. L., LECANNELIER, E. A., PINO, E. J., ROJAS, A. J.: Clustering of Atrial Fibrillation Based on Surface ECG Measurements, 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, str. 4203 - 4206, červenec 2013, ISSN 1557-170X
- FERRARIO, M., SIGNORINI, M. G., CERUTTI, S.: Complexity Analysis of 24 Hours Heart Rate Variability Time Series, 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, svazek 2, str. 3956 – 3959, září 2004, ISBN: 0-7803-8439-3
- FIGO: Guidelines for the use of fetal monitoring, International Journal of Gynaecology and Obstetrics, 1985
- FIRA, C. M., GORAS, L.: An ECG Signals Compression Method and Its Validation Using NNs, Ieee Transactions on Biomedical Engineering, svazek 55, číslo 4, str. 1319-1326, duben 2008
- FRALEY, C., RAFTERY, A. E.: Model-Based Clustering, Discriminant Analysis, and Density Estimation, Journal of the American Statistical Association, svazek 97, číslo 458, str. 611, červen 2002
- GACEK, A., Data structure-guided development of electrocardiographic signal characterization and classification, Artificial Intelligence in Medicine, svazek 59, číslo 3, str. 197-204, listopad 2013

GEORGIEVA A., PAPAGEORGHIU, A. T., PAYNE, S. J., MOULDEN, M., REDMAN, C. W. G.: Phase-rectified signal averaging for intrapartum electronic fetal heart rate monitoring is related to acidaemia at birth, *BJOG: An International Journal of Obstetrics & Gynaecology*, únor 2014

GERMAN-SALLO, Z.: Wavelet transform based HRV analysis, *The 7th International Conference Interdisciplinarity in Engineering, Procedia Technology*, svazek 12, str. 105-111, 2013

GLOVER, D. R.: Picture Data Compression Icoder Using Subband/Transform Coding With A Lempel-Ziv-Based Coder, *The United States Of America As Represented By The Administrator Of The National Aeronautics And Space Administration*, květen 1995

GONÇALVES, H., BERNARDES, J., ROCHA, A. P., AYRES-DE-CAMPOS, D.: Linear and nonlinear analysis of heart rate patterns associated with fetal behavioral states in the antepartum period, *Early Human Development*, svazek 83, číslo 9, str. 585-591, 2007

GRANÅSEN, G., WIKLUND, U.: Multivariate Classification of Cardiac Autonomic Function and Echocardiographic Abnormalities, *Computing in Cardiology Conference*, str. 1151 – 1154, září 2013, ISSN: 2325-8861

GUO, C., LI, H., PAN, D.: An Improved Piecewise Aggregate Approximation Based on Statistical Features for Time Series Mining, *4th International Conference on Knowledge Science, Engineering & Management*, str. 233-244, 2010, ISBN 978-3-642-15280-1

HAUTAMÄKI, V., NYKÄNEN, P., FRÄNTI, P.: Time-series Clustering by Approximate Prototypes, *19th International Conference on Pattern Recognition*, str. 1-4, prosinec 2008, ISSN : 1051-4651

HOPKINS, P., OUTRAM, N., LÖFGREN, N., IFEACHOR, E. C., ROSÉN, K. G.: A Comparative Study of Fetal Heart Rate Variability Analysis Techniques, *Proceedings of the 28th IEEE EMBS Annual International Conference*, 2006

HORNIÁK, R.: Analýza EEG signálu, *Diplomová práce, Fakulta elektrotechnická, České vysoké učení technické v Praze*, 2010

HUI, Y., FANXING, M.: Application of Improved SAX Algorithm to QAR Flight Data, *International Conference on Applied Physics and Industrial Engineering, Physics Procedia*, svazek 24, část B, str. 1406-1413, 2012

JANKŮ, P.: Analýza ST úseku fetálního EKG v intrapartální diagnostice hypoxie plodu u rizikových gravidit, *Disertační práce, Lékařská fakulta Masarykovy univerzity*, 2007

JEŽOVÁ, M., FEIT, J.: Atlas patofyziologie novorozence, 2014, [cit. 2014-05-02], URL: http://atlases.muni.cz/atlases/novo/atlas_cz/main+novorozenec+novorasfyxcas.html

JUNEJO, I. N., AL AGHBARI, Z.: Using SAX representation for human action recognition, *Journal of Visual Communication and Image Representation*, svazek 23, číslo 6, str. 853-861, 2012

KANTELHARDT, J. W., BAUER, A., SCHUMANN, A. Y., BARTHEL, P., SCHNEIDER, R., MALIK, M., SCHMIDT, G.: Phase-rectified signal averaging for the

detection of quasi-periodicities and the prediction of cardiovascular risk, American Institute of Physics, Chaos, svazek 17, číslo 1, březen 2007

KANTOR, L.: ČR má jednu z nejnižších úmrtností novorozenců na světě, MEDICAL TRIBUNE CZ, 20.11. 2013, [cit 2014-07-03], URL:
<http://www.tribune.cz/clanek/31521-cr-ma-jednu-z-nejnizsich-umrtnosti-novorozencu-na-svete>

KARÁSEK, J.: Citlivost metod pro měření podobnosti kvantitativních proměnných, Access Server, 17.9. 2012, [cit 2014-04-03], ISSN 1214-9675, URL:
<http://access.feld.cvut.cz/view.php?cisloclanku=2012090003>

KEOGH, E., CHU, S., HART, D., PAZZANI, M.: An Online Algorithm for Segmenting Time Series, IEEE International Conference on Data Mining, str. 289-296, 2001, ISBN: 0-7695-1119-8

KEOGH, E., RATANAMAHATANA, C. A.: Exact indexing of dynamic time warping, Knowledge and Information Systems, svazek 7, číslo 3, str. 358-386, 2004, ISSN 0219-3116

KIKUCHI, A., SHIMIZU, T., HAYASHI, A., HORIKOSHI, T., UNNO, N., KOZUMA, S., TAKETANI, Y.: Nonlinear analyses of heart rate variability in normal and growth-restricted fetuses, Early Human Development, svazek 82, číslo 4, str. 217-226, duben 2006

KOVÁCS-VAJNA, Z. M.: A Fingerprint Verification System Based on Triangular Matching and Dynamic Time Warping, IEEE Transactions on Pattern Analysis and Machine Intelligence, svazek 22, číslo 11, str. 1266-1276, listopad 2000, ISSN : 0162-8828

KULAHCIOGLU, B., OZDEMIR, S., KUMOVA, B.: Application of Symbolic Piecewise Aggregate Approximation (Paa) Analysis to Ecg Signals, Applied Simulation and Modelling, červen 2008

LIN, J., KEOGH, E., LONARDI, S., CHIU, B.: A Symbolic Representation of Time Series, with Implications for Streaming Algorithms, 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, str. 2-11, 2003

LIN, J., YUAN, L.: Finding Structurally Different Medical Data, 22nd IEEE International Symposium on Computer-Based Medical Systems, str. 1-8, 2009, ISSN : 1063-7125

LKHAGVA, B., SUZUKI, Y., KAWAGOE, K.: New Time Series Data Representation ESAX for Financial Applications, 22nd International Conference on Data Engineering Workshops, str. 115, 2006, ISBN: 0-7695-2571-7

LÓPEZ-AVITIA, R., REYNA-CARRANZA, M. A., CETTO, L. A., BRAVO-ZANOQUERA, M. E., ARRIOLA-ZORRILLA, H. G.: QRS Complex Duration Variability Assessment Using Dynamic Time Warping and Piecewise Linear Approximation, Pan American Health Care Exchange, str. 10, 2010, ISBN 978-1-4244-6294-0

- MANNING, C. D., RAGHAVAN, P., SCHÜTZE, H.: An Introduction to Information Retrieval, Cambridge University Press, 2009, [cit. 2014-18-04], ISBN: 0521865719, URL: <http://www-nlp.stanford.edu/IR-book/>
- MEESRIKAMOLKUL, W., NIENNATTRAKUL, V., RATANAMAHAHATANA, C. A.: Shape-Based Clustering for Time Series Data, The 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining, část 1, str. 530-541, 2012
- NOVOTNÁ, M.: Současné možnosti detekce intrauterinní tísně plodu, Zdravotnické listy, číslo 23, 2001, [cit. 2014-16-02], URL: <http://zdravi.e15.cz/clanek/priloha-lekarske-listy/soucasne-moznosti-detekce-intrauterinni-tisne-plodu-136528>
- OHSAKI, M., ABE, H., YAMAGUCHI, T.: Numerical Time-Series Pattern Extraction Based on Irregular Piecewise Aggregate Approximation and Gradient Specification, New Generation Computing, svazek 25, číslo 3, str. 213-222, květen 2007
- OLVER, P. J.: Topics in Fourier Analysis: DFT & FFT, Wavelets, Laplace Transform, University of Minnesota, 2014
- PAN, Q., GONG, Y., GONG, S., HU, Q., ZHANG, Z., YAN, J., NING, G.: Enhancing the deceleration capacity index of heart rate by modified-phase-rectified signal averaging, Medical & Biological Engineering & Computing, svazek 48, str. 399-405, 2009
- PAPÍRNÍKOVÁ, P.: Kardiokografie, Moderní babičství, 2006, číslo 10, [cit. 2014-20-02], URL: <http://www.levret.cz/publikace/casopisy/mb/2006-10/?pdf=53>
- PHAM, N. D., LE, Q. L., DANG, T. K.: Two Novel Adaptive Symbolic Representations for Similarity Search in Time Series Databases, 12th International Asia-Pacific Web Conference, str. 181-187, 2010, ISBN : 978-1-4244-6600-9
- POSPÍŠIL, D., KLIMEK, M., KOZUMPLÍK, J.: Automatická analýza signálu EKG, Elektrovue, 2013, svazek 15, číslo 5, [cit. 2014-06-04], URL: <http://www.elektrovue.cz/file.php?id=200001203-e50a0e603d>
- ROSSINI, K., VERDUN, S., CARIOU, V., QANNARI, E. M., FOGLIATTO, F.S.: PLS discriminant analysis applied to conventional sensory profiling data, Food Quality and Preference, svazek 23, číslo 1, str. 18-24, 2012
- SANTIAGO-MOZOS, R., GARCIA-VIZUETE, B., LILLO-CASTELLANO, J. M., ROJO-ÁLVAREZ, J. L., MARTÍN-CABALLERO, C.: On the Early Detection of Perinatal Hypoxia with Information-Theory based Methods, Computing in Cardiology Conference, str. 425-428, září 2013, ISSN: 2325-8861
- SANTOSH, K. C.: Use of Dynamic Time Warping for Object Shape Classification through Signature, Kathmandu University Journal of Science, Engineering and Technology, svazek 6, číslo 1, str. 33-49, březen 2010
- SCHIERMEIER, S., PILDNER VON STEINBURG, S., THIEME, A., REINHARD, J., DAUMER, M., SCHOLZ, M., HATZMANN, W., SCHNEIDER, K. T. M.: Sensitivity and specificity of intrapartum computerised FIGO criteria for cardiotocography and fetal scalp pH during labour: multicentre, observational study, BJOG, svazek 115, číslo 12, str. 1557-1563, listopad 2008

- SHAH, G. H.: An Improved DBSCAN, A Density Based Clustering Algorithm with Parameter Selection for High Dimensional Data Sets, Nirma University International Conference on Engineering, str. 1-6, prosinec 2012, ISBN: 978-1-4673-1720-7
- SUN, Y., LI, J., LIU, J., SUN, B., CHOW, C.: An Improvement of Symbolic Aggregate Approximation Distance Measure for Time Series, Neurocomputing, svazek 138, str.1 189-198, srpen 2013
- SUNDSTRÖM, A. K., ROSÉN, D., ROSÉN, K. G.: Fetal surveillance, Neoventa Medical AB, 2000
- SYED, Z., SCIRICA, B. M., STULTZ, C. M., GUTTAG, J. V.: Comparing Symbolic Representations of Cardiac Activity to Identify Patient Populations with Similar Risk Profiles, Computers in Cardiology, str. 85-88, září 2008, ISSN : 0276-6547
- SZYMAŃSKA, E., SACCENTI, E., SMILDE, A. K., WESTERHUIS, J. A.: Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies, Metabolomics, svazek 8, str. 3-16, červen 2011
- ŠTĚPÁNKOVÁ O.: Hodnocení výkonnosti klasifikátoru, Dobývání a vizualizace znalostí, 2014
- TAO, Y., PI, D.: Unifying Density-Based Clustering and Outlier Detection, Second International Workshop on Knowledge Discovery and Data Mining, str. 644-647, 2009, ISBN: 978-0-7695-3543-2
- TROJAN, S. a kolektiv: Lékařská fyziologie, 4. vydání, Grada Publishing, 2013, ISBN 80-247-0512-5
- VALUCHOVÁ, H.: Hypoxie plodu – Intrapartální monitorace, Bakalářská práce, Univerzita Palackého v Olomouci, Fakulta zdravotnických věd, 2012
- WARREN LIAO, T.: Clustering of time series data—a survey, Pattern Recognition, svazek 38, číslo 11, str. 1857-1874, listopad 2005
- WELCH, T. A.: A Technique for High-Performance Data Compression, Computer, svazek 17, číslo 6, str. 8-19, červen 1984, ISSN : 0018-9162
- ZHAO, Y., KARYPIS, G.: Evaluation of Hierarchical Clustering Algorithms for Document Datasets, 1th Conference of Information and Knowledge Management, str. 515-524, 2002
- ZHU, Y., WU, D., LI, S.: A Piecewise Linear Representation Method of Time Series Based on Feature Points, Lecture Notes in Computer Science, svazek 4693, str. 1066-1072, 2007
- ZHANG, X. T., ZHANG, W., XIONG, X., WANG, Q. W., LI, C. Y.: A Model-Based Clustering For Time-Series With Irregular Interval, 2004 International Conference on Machine Learning and Cybernetics, svazek 5, str. 2883 – 2888, srpen 2014, ISBN: 0-7803-8403-2
- ZIV, J., LEMPEL, A.: A Universal Algorithm for Sequential Data Compression, IEEE Transactions on Information Theory, svazek 23, číslo 3, str. 337-343, květen 1977, ISSN : 0018-9448

ZIV, J., LEMPEL, A.: Compression of individual sequences via variable-rate coding, IEEE Transactions on Information Theory, svazek 24, číslo 5, str. 530-536, září 1978, ISSN : 0018-9448

Zdroje obrázků

[1] MOTIFOLIO: PPT Drawing Toolkit - Embryology, [cit. 2014-15-03], URL
<http://www.motifolio.com/1011181.html>

[2] KAZMI, T., RADFER, F., KHAN, S.: ST Analysis of the Fetal ECG, as an Adjunct to Fetal Heart Rate Monitoring in Labour: A Review, Oman Med J, svazek 26, číslo 6, str. 459-460, listopad 2011

[3] SRIVASTAVA, T.: Getting your clustering right (Part I), listopad 2013, [cit. 2014-01-04], URL: <http://www.analyticsvidhya.com/blog/2013/11/getting-clustering-right/>

[4] KEOGH, E.: Mining Shape and Time Series Databases with Symbolic Representations, SIGKDD 2007 Tutorial, srpen 2007, [cit. 2014-16-04], URL
<http://www.cs.ucr.edu/~eamonn/tutorials.html>

[5] BERSON, A., SMITH, S., THEARLING, K: An Overview of Data Mining Techniques, [cit. 2014-20-04], URL
<http://www.thearling.com/text/dmtechniques/dmtechniques.htm>

[6] VOGET, M.: Mining for Groups Using Clustering Algorithms, [cit. 2014-01-05], URL
http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/mvoget/cluster/cluster.html