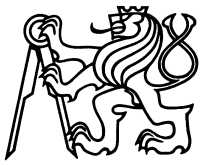
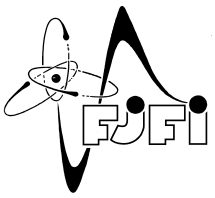


České vysoké učení technické v Praze
Fakulta jaderná a fyzikálně inženýrská



Statistická a strojová klasifikace signálů akustické emise pro detekci defektů v materiálech

Classification of acoustic emission signals in material defectoscopy based on statistics and machine learning

Bakalářská práce

Autor: **Jan Zavadil**

Vedoucí práce: **Ing. Václav Kůs, Ph.D.**

Akademický rok: 2020/2021

Katedra: matematiky

Akademický rok: 2020/2021

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Student:	Jan Zavadil
Studijní program:	Aplikace přírodních věd
Studijní obor:	Matematické inženýrství
Studijní zaměření:	Aplikované matematicko-stochastické metody
Název práce (česky):	Statistická a strojová klasifikace signálů akustické emise pro detekci defektů v materiálech
Název práce (anglicky):	Classification of acoustic emission signals in material defectoscopy based on statistics and machine learning

Pokyny pro vypracování:

1) Seznamte se s paletou defektoskopických metod materiálů, speciálně se zaměřte na akustickou emisi (AE) včetně struktury detekovaných signálů a jejich charakteristik.
2) Prostudujte vhodné statistické a klasifikační přístupy pro klastrování těchto signálů jako jsou např. 'model based' klasifikátory založené na distribučních směsích (MBC), jádrové odhady (KDE), rozhodovací stromy (DT), a to včetně divergenčního rozhodovacího stromu s učitelem (SDDT).
3) Prozkoumejte poskytnutá data z laboratorních zkoušek AE na tenkém plechu, statisticky předzpracujte signály a najděte vhodné atributy pro úspěšnou klasifikaci.
4) Aplikujte vybrané metody (MBC, SDDT,...) na poskytnuté signály AE, klasifikujte různé typy signálů do tříd, tzn. oddělte různé zdroje AE. Srovnejte výsledky statisticky orientovaných metod s metodami strojovými. Zkuste funkčnost on-line klasifikace nově vygenerovaného signálu AE.
5) Seznamte se se základy dalších moderních metod strojového učení jako např. náhodné lesy (RF), hluboké neuronové sítě (DNN) nebo konvoluční sítě (CNN) a jejich způsoby optimálního učení.

Doporučená literatura:

1) B. Kopec, et al., Nedestruktivní zkoušení materiálů a konstrukcí. Akademické nakladatelství CERM, Brno, 2008.

2) S. Aghabozorgi, A. S. Shirkhorshidi, T. Y. Wah, Time-series clustering – A decade review. Elsevier, Information Systems 53, 2015, 16–38.

3) L. Pardo, Statistical inference based on divergence measures. Chapman & Hall/CRC, Taylor-Francis, London, 2006.

4) P. Bouř, V. Kůs, J. Franc, Statistical classification techniques in high energy physics (SDDT algorithm). Journal of Physics: Conference Series 738, 012034, IOP Publishing Ltd., 2016.

5) A. Zhang, Z. C. Lipton, M. Li, A. J. Smola, Dive into Deep Learning. Release 0.14.4, Sep 18, 2020, https://d2l.ai/d2l-en.pdf. Možná varianta:
J. Quinn, J. J. McEachen, M. Fullan, M. Gardner, M. Drummy, Dive Into Deep Learning: Tools for Engagement. Corwin publisher, 1st Edition (July 15), 2019.

Jméno a pracoviště vedoucího bakalářské práce:

Ing. Kůs Václav, Ph.D.
Katedra matematiky, Fakulta jaderná a fyzikálně inženýrská, České vysoké učení technické v Praze, Trojanova 13, 120 00 Praha 2

Jméno a pracoviště konzultanta:

Datum zadání bakalářské práce: 31.10.2020

Datum odevzdání bakalářské práce: 7.7.2021

Doba platnosti zadání je dva roky od data zadání.

V Praze dne 30.10.2020

garant oboru

vedoucí katedry

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
Fakulta jaderná a fyzikálně inženýrská
-1-

děkan

Poděkování:

Chtěl bych zde poděkovat svému školiteli Ing. Václavu Kůsovi, Ph.D. za zodpovědné, vstřícné, odborné a trpělivé vedení mé bakalářské práce.

Čestné prohlášení:

Prohlašuji, že jsem tuto práci vypracoval samostatně a uvedl jsem všechnu použitou literaturu.

V Praze dne 7. července 2021 Jan Zavadil

Název práce:

Statistická a strojová klasifikace signálů akustické emise pro detekci defektů v materiálech

Autor: Jan Zavadil

Obor: Matematické inženýrství

Zaměření: Aplikované matematicko-stochastické metody

Druh práce: Bakalářská práce

Vedoucí práce: Ing. Václav Kůs, Ph.D., České vysoké učení technické v Praze, Fakulta jaderná a fyzikálně inženýrská, Katedra matematiky

Abstrakt: Spolehlivá klasifikace signálů akustické emise je klíčová pro praktické využití této defektoskopické metody. Signály jsou při klasifikaci reprezentovány pomocí vhodné nízkodimenzionální skupiny atributů. V této práci se zabýváme výběrem vhodných atributů a následně popisem a srovnáním několika metod klasifikace, jmenovitě Divizivní metody, Model Based klasifikace, metody KDE a klasifikace za pomoci Divergenčního stromu s učitelem. Součástí práce je návrh vlastního atributu a vlastní klasifikační metody. Testování metod provádíme na laboratorně naměřených datech. Jako nejspolehlivější klasifikační metoda se jeví metoda KDE s učitelem.

Klíčová slova: Akustická emise, atributy signálu, ϕ-divergence, KDE klasifikace, MBC, shluková analýza

Title:

Classification of acoustic emission signals in material defectoscopy based on statistics and machine learning

Author: Jan Zavadil

Abstract: Reliable classification of acoustic emission signals is crucial for practical use of this non-destructive testing technique. During our classification, signals are represented by a convenient, low-dimensional set of attributes. This paper addresses the problem of selecting appropriate atributes and consequently describes and compares several classification methods, specifically Division methods, Model Based clustering, KDE method and classification using Supervised Divergence Decision Tree. The paper proposes new attribute and classification method. The methods were tested and compared on a set of laboratory measured data. The most reliable method seems to be the supervised KDE classification method.

Key words: Acoustic emission, clustering, KDE clustering, ϕ-divergence, MBC, signal attributes

Obsah

Úvod

Akustická emise (AE) jakožto nedestruktivní diagnostická metoda prošla v nedávné době významným rozvojem, především díky rychlému rozvoji výpočetní techniky. S tímto rozvojem přichází přirozená potřeba klasifikace signálů AE podle fyzikálního zdroje. Kvalitní klasifikátory AE nacházejí uplatnění v mnoha odvětvích průmyslu a slouží jak k zvýšení bezpečnosti provozu, tak k snížení nákladů výroby.

Tato práce se zabývá extrakcí příznaků (atributů) ze signálu akustické emise a několika klasifikačními metodami, které tyto atributy využívají. Cílem práce bylo nalezení vhodných atributů, které umožňují efektivní klasifikaci signálů a následně porovnání úspěšnosti implementovaných klasifikačních metod. K otestování metod byla použita laboratorní data pocházející z měření na tenké plechové desce a na hliníkové eloxované konvici.

V první kapitole jsme shrnuli základní informace o akustické emisi, jejím šíření v pevných tělesech, způsobu detekce a aplikacích AE v průmyslu. Dále jsou v této části představena námi použitá data. Druhá kapitola popisuje podobu signálu akustické emise a představuje použité atributy, jak ze signálu samotného, tak z jeho spektra. Celkem jsme použili šest různých atributů, které signály akustické emise reprezentují ve výběrovém prostoru. V kapitolách tři až šest jsou popsány implementované klasifikační metody. První dvě - divizivní metoda a model based clustering (MBC) jsou klasifikátory bez učitele, tzv. *unsupervised*. Divizivní metoda je nejjednodušší a k rozdělení pozorování využívá pouze součtů vzdáleností mezi nimi, MBC metoda funguje na principu odhadu hustot pravděpodobnosti, kdy pozorování ve výběrovém prostoru prokládá směsí hustot normálních rozdělení. Zbylé metody pracují s trénovací množinou dat a řadí se tedy mezi *supervised* klasifikátory. Metody SMBC (Supervised MBC) a GMMC (Gaussian Mixture Model Clustering) vycházejí z MBC a snaží se jí vylepšit. Metoda SKDEC (Supervised Kernel Desity Estimation Clustering) klasifikuje na základě beparametrických hustot odhadnutých jádrovým odhadem a konečně metoda SDDT (Supervised Divergence Decision Tree) využívá ke klasifikaci divergenčního rozhodovacího stromu, vyvíjeného při katedře matematiky FJFI. V sedmé kapitole jsou uvedeny výsledky klasifikace a porovnání jednotlivých metod. Všechny klasifikační metody byly implementovány v prostředí Matlab.

Kapitola 1

Akustická emise

Akustická emise je libovolný fyzikální jev při kterém dochází k uvolnění nahromaděné deformační energie v tělese a její přeměně na energii kinetickou. Tato energie se šíří tělesem v podobě elastické vlny. S akustickou emisí se můžeme setkat i u kapalin a plynů, typickým zdrojem jsou však různé mechanismy deformace a lomu v pevných látkách. K zaznamenání emisních událostí se používají piezoelektrické snímače.

Pokud dokážeme v signálu rozpoznat jednotlivé oddělené emisní události, hovoříme o praskavé akustické emisi (*burst type emission*), v opačném případě pak emisi nazýváme spojitou (*continuous emission*). Příkladem praskavé akustické emise je většina lomových procesů v materiálu nebo přetržení vlákna kompozitu. Příkladem spojité emise je únik látky z tlakové nádoby či potrubí nebo obrábění materiálu. V této práci se zabýváme pouze praskavou akustickou emisí.

Přirozeným zdrojem nazýváme děj při kterém dochází ke vzniku akustické emise uvnitř tělesa, např. při tahových zkouškách. Tyto zdroje mají blíže k reálným zdrojům s nimiž se setkáváme v technické praxi. V laboratorním prostředí je častější variantou umělý zdroj akustické emise, kdy je emise vynucena kontaktem zkoumaného předmětu a vnějšího budiče, např. pen-test. Vynucená emise vzniká pouze na povrchu tělesa, což příliš nevadí u tenkých předmětů. Umožňuje nám to navíc alespoň částečně kontrolovat parametry emisní události a její polohu.

1.1 Šíření elastických vln

Jelikož akustická emise je projevem šíření elastických vln, bude užitečné uvést si zde základní výsledky z teorie mechaniky kontinua, která se jimi zabývá.

Při popisu pevných látek vyjdeme z Lagrangeova popisu kontinua, kde v čase t_0 má každý bod kontinua souřadnice $\boldsymbol{x} = (x_1, x_2, x_3)$, v čase $t > t_0$ potom souřadnice $\boldsymbol{x} + \boldsymbol{u}, \boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x}, t)$, kde $\boldsymbol{u}$ se nazývá *vektor posunutí*. Pro popis zatíženého tělesa se zavedou dva tenzory: *tenzor deformace* e_{ij} a *tenzor napětí* τ_{ij}. Při omezení na malé transformace platí mezi oběma tenzorovými veličinami tzv. zobecněný Hookův zákon

$$e_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \quad i, j = 1, 2, 3,$$

$$\tau_{ij} = c_{ijkl} e_{kl},$$

za použití Einsteinovi sumační konvence v druhé rovnici, kde konstanty úměrnosti c_{ijkl} jsou závislé na vlastnostech materiálu tělesa. Nazývají se *elastické moduly* a obecně se jedná o tenzory čtvrtého řádu, které mají 81 prvků. V izotropních materiálech dochází k maximální redukci

koeficientů c_{ijkl}, pro jejich popis stačí pouze Lamého konstanta λ a smykový modul μ. Pro izotropní látky platí

$$c_{ijkl} = \lambda\delta_{ij}\delta_{kl} + \mu\left(\delta_{il}\delta_{jk} + \delta_{ik}\delta_{jl}\right),$$

kde δ_{ij} značí Kroneckerovo delta.

Šíření elastických vln v pevných látkách se řídí elastodynamickou (vlnovou) rovnicí

$$\frac{\partial\tau_{ij}}{\partial x_j} + f_i = \rho\frac{\partial^2 u_i}{\partial t^2}, \quad i = 1, 2, 3,$$

kde ρ značí hustotu materiálu a f_i vnitřní síly. Řešením elastodynamické rovnice získáme příslušné dynamické posuvy $\boldsymbol{u}$. Výsledky se zásadně liší pro homogenní, nehomogenní, izotropní a anizotropní látky. V nejjednodušším případě izotropního homogenního prostředí se látkou šíří dva druhy vln s rozdílnou rychlostí:

- podélné P-vlny, s rychlostí

$$c_1 = \sqrt{\frac{\lambda + 2\mu}{\rho}} = \sqrt{\frac{E\left(1-\nu\right)}{\rho\left(1+\nu\right)\left(1-2\nu\right)}}, \tag{1.1}$$

- smykové (příčné) S-vlny, s rychlostí

$$c_2 = \sqrt{\frac{\mu}{\rho}} = \sqrt{\frac{E}{2\rho\left(1+\nu\right)}}, \tag{1.2}$$

kde λ je Lamého koeficient, μ modul smyku, ρ hustota, E modul pružnosti a ν Poissonovo číslo.

V nehomogenních látkách je šíření elastických vln výrazně komplikovanější, jednotlivé vlny nejsou separovatelné a materiálem se šíří celý vlnový balík. V případě pouze slabě nehomogenního materiálu a vysokofrekvenční vlny se však hustota a elastické moduly lokálně příliš neliší a vzorce (1.1, 1.2) platí alespoň přibližně. Pokud uvažujeme ohraničené kontinuum, případně kontinuum s rozhraním, objevují se další druhy vln, např. povrchové Rayleighovy vlny. Výsledné vlnové pole v tělese je tak často velmi složité.

Na tenkých deskách předpovídá teorie navíc k S-vlnám, P-vlnám a povrchovým vlnám vznik dvou módů deskových vln, symetrického a antisymetrického. Jejich rychlosti lze určit řešením transcendentní rovnice

$$\frac{\tanh\left(\frac{\pi h}{\lambda}\sqrt{1-\frac{c^2}{c_1^2}}\right)}{\tanh\left(\frac{\pi h}{\lambda}\sqrt{1-\frac{c^2}{c_2^2}}\right)} = \left(\frac{\left(\frac{c^2}{c_1^2}-2\right)^2}{\sqrt[4]{1-\frac{c^2}{c_2^2}}\sqrt{1-\frac{c^2}{c_2^2}}}\right)^{\pm 1}, \tag{1.3}$$

kde h je tloušťka desky, λ vlnová délka, c_1 rychlost P-vlny a c_2 rychlost S-vlny. Kladný exponent odpovídá symetrickému a záporný antisymetrickému módu.

Rovnice (1.3) nelze obecně analyticky vyřešit, nicméně pokud h/λ nabývá malých hodnot, je možné odvodit vztahy

$$c_s = \sqrt{\frac{E}{\rho\left(1-\nu^2\right)}}, \tag{1.4}$$

$$c_a = \sqrt{\frac{\omega^2 E h^2}{12\rho\left(1-\nu^2\right)}}, \tag{1.5}$$

kde c_s je rychlost šíření symetrického módu, c_a je rychlost šíření antisymetrického módu deskových vln a ω je úhlová frekvence. U tenkých desek používaných pro buzení akustické emise a pro frekvenční rozsah běžných snímačů platí $c_s > c_a$, což v praxi může komplikovat určení času příchodu signálu AE k jednotlivým snímačům.

1.2 Aplikace AE

Nejvýznamnější oblastí aplikace akustické emise je nedestruktivní defektoskopie, neboť deformační či fázové změny uvnitř materiálu jsou jejími přirozenými zdroji. Akustická emise se ukazuje jako nesmírně užitečná metoda především díky tomu, že umožňuje kontinuální monitorování zkoumaného objektu a je poměrně snadné umístit aparaturu pro snímání signálů i na hůře přístupná místa. Nedestruktivní metody nám na základě provedených měření umožňují provádět odhady zbytkové životnosti, případně nás upozorní na vznik nebezpečné trhliny. Typickým příkladem takto kontrolovaného objektu je jaderný reaktor, kdy akustickou emisi snímáme za provozu bez nákladné odstávky a jsme včas upozorněni na vznik potenciálně nebezpečného defektu. Ten je následně během odstávky podrobně prozkoumán dalšími defektoskopickými metodami, například pomocí ultrazvukové detekce, či pomocí rentgenových a gama paprsků.

S akustickou emisí se dále můžeme setkat rovněž ve stavebnictví, kde slouží především k diagnostice velkých železobetonových konstrukcí, např. mostů. Moderní je využití akustické emise při obrábění, kdy nás při monitorování obráběcího procesu může upozornit na poškození obráběcího nástroje a umožňuje optimálně řídit celý obráběcí proces (např. rychlost řezání či tloušťku třísky). Také se osvědčila při detekcích poškození kuličkových ložisek strojů a užitkových vozů nebo v letectví. Další oblastí využívající akustickou emisi je také geologie a seismologie. Sledováním šíření elastických vln tělesem Země je možné odhadovat složení Země či hledat ložiska surovin.

1.3 Detekce AE

Nedestruktivní metody pracující s jevem akustické emise jsou odkázány na detekci elastických vln na povrchu těles. To se děje za pomoci elektromechanických měničů. Standardně jsou používány piezoelektrické snímače akustické emise, jejichž maximální citlivost leží v pásmu nízkých až středních ultrazvukových frekvencí (20kHz až 2MHz).

Na základě zaznamenaných elektrických signálů je prováděna lokalizace a identifikace zdroje dané emisní události. Emisních událostí obvykle přichází značné množství a jsou navíc zkresleny množstvím odrazů, elektronickým šumem a spektrálními vlastnostmi snímačů. Ke zpracování signálů lze přistoupit dvěma způsoby:

- "On-line zpracování" využívá jednoduché charakteristiky signálů, které jsou poskytovány v reálném čase aparaturou pro detekci akustické emise. Jedná se např. o amplitudu emisní události, čas do maxima (rise time) či celkový počet překmitů přes uživatelem nastavený práh šumu. Výhodou tohoto přístupu je možnost okamžitě reagovat na vznikající kritické situace, záznam signálu je však značně komprimován a dochází tedy ke ztrátě informace obsažené v signálu,

- "Off-line zpracování" emisní události v reálném čase pouze zaznamená a další vyhodnocování probíhá až následovně. Tento postup se využívá především proto, že je možné zaznamenané signály dále zpracovat a využít k vyhodnocení výpočetně náročnějších algoritmů.

Nemožnost reagovat v reálném čase je však zásadní překážkou ve využitelnosti a proto je jasné, že budoucnost patří on-line klasifikaci. Díky rostoucímu výkonu mikroprocesorů je možné aplikovat stále komplikovanější algoritmy už při on-line zpracování.

1.4 Klasifikace zdroje AE

Určení typu zdroje akustické emise nám má v praxi pomoci odhalit, zda se jedná o diagnosticky významný signál nebo pouze o hluk způsobený okolím. Klasifikací lze též zjistit, o jak závažný problém se jedná a jak k němu přistoupit. Je tedy nutné vytvořit skupiny (shluky) odpovídající různým typům akustické emise (např. velká prasklina, malá prasklina, únik média...) a následně přiřazovat neznámé signály k daným skupinám. Klasifikace probíhá ve dvou zásadních krocích:

1. volba vhodných atributů (parametrů) signálů, které budou dané shluky co nejlépe oddělovat,

2. volba vhodné klasifikační metody.

Klasifikační metody se dají rozdělit na metody bez učení, u kterých na začátku nemáme žádnou informaci o příslušnosti jednotlivých signálů k daným třídám, a metody s učením, u kterých máme na začátku tzv. trénovací množinu signálů, jejichž třídu známe. Metody s učením jsou přesnější, ale v praxi je často nereálné získat dostatečně velkou trénovací množinu. Existuje také riziko přetrénování metody s učením, které může vést k horším výsledkům klasifikace.

Pro otestování a porovnání implementovaných klasifikačních metod jsme využili již dříve laboratorně naměřených dat. Použité signály pocházejí z měření piezoelektrickými snímači na tenké čtvercové desce z válcovaného plechu. Do počítače byly zaznamenány pomocí přístroje DAKEL-XEDO. Akustická emise byla buzena různými způsoby uprostřed desky a snímána čtyřmi snímači v rozích desky ve vzdálenosti 10 cm od nejbližších hran. V tabulce 1.1 jsou uvedeny zdroje akustické emise včetně počtu dostupných pozorování.

Zdroj AE	Počet pozorování
Pentest	65
Kartáček	173
Křída	103
Řetízek	126
Houbička	135

Tabulka 1.1: Zdroje použitých signálů AE

K těmto datům jsme navíc připojili ještě signály měřené na eloxované konvici. Jednalo se o pentestové signály - lámání tuhy o konvici. Bylo použito několik druhů tuhy, které, jak je patrno z kapitoly 7, budily velmi podobné signály. Pro metody s učitelem jsme jako trénovací množinu využili vždy 30 náhodně vybraných signálů z daného zdroje.

V praxi není vždy jasné kolik shluků signálů máme v datech hledat. V této práci pracujeme pouze s laboratorními daty a problematikou určení optimálního počtu shluků pro klasifikaci se zatím nezabýváme, bude následovat ve výzkumném úkolu. Více k tomuto tématu viz [1]. Pro klasifikaci dat používáme celkem šest atributů popsaných v následující kapitole.

Kapitola 2

Parametry signálů pro klasifikaci

Použití kompletních naměřených signálů pro klasifikaci se nejeví jako efektivní [3], proto je pro snížení dimenze úlohy reprezentujeme pomocí několika atributů extrahovaných ze signálu. Důvodem pro snížení dimenze je nižší paměťová náročnost než u syrových signálů a nižší výpočetní náklady. V dnešní době strojového učení lze již klasifikovat celé signály AE, tento přístup plánujeme aplikovat v dalším výzkumu.

Klíčovým krokem v úspěšné klasifikaci je nalezení vhodných atributů (parametrů) z emisních událostí, které použijeme k třídění signálů do skupin odpovídajících fyzikálnímu zdroji. Tyto parametry se získávají pomocí určitých zobrazení, jak ze signálu samotného, tak z jeho spektra.

Uspořádaná $d-tice$ těchto atributů signálu bude tvořit prvek výběrového prostoru $\mathbb{R}^d$, mezi těmito prvky budeme hledat shluky pomocí různých klasifikačních metod. Výběr atributů tedy přímo určuje strukturu shluků. Je proto žádoucí vybírat je tak, aby v rámci jednoho fyzikálního zdroje měly co nejmenší rozptyl a co nejlépe od sebe oddělovaly prvky pocházející z různých fyzikálních zdrojů.

Po atributech chceme aby každý přinášel novou informaci o rozdílnostech mezi jednotlivými signály a proto se je snažíme volit tak, aby na sobě nebyly závislé. S vyšším počtem použitých atributů se zvyšuje dimenze výběrového prostoru a intuitivně se tak zvětšují vzdálenosti mezi jeho prvky, což by mělo vést ke snazšímu oddělení jednotlivých typů detekované AE.

2.1 Signál

Měřící aparatura nepřetržitě monitoruje dění na snímačích, při překročení prahové hodnoty se přístroj v bufferu vrátí o 3000 záznamů zpět (pretrigger) a zaznamená emisní událost o délce 10 000 záznamů. Vzorkovací frekvence je 4 MHz.

Označme x_t naměřenou posloupnost v milivoltech úrovně signálu. Spektrum signálu $\tilde{S}_f$ spočteme pomocí *diskrétní Fourierovy transformace* jako

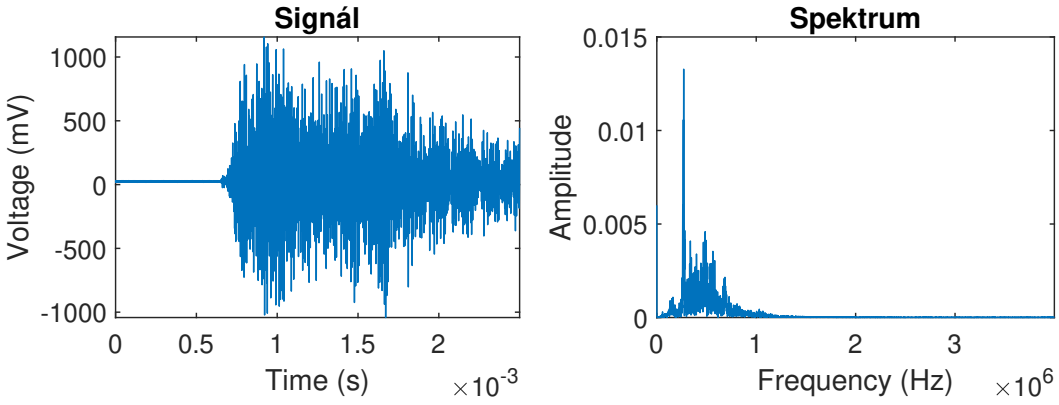
$$\tilde{S}_f = \sum_{t=0}^{T-1} x_t \exp\left(-\frac{2\pi i f t}{T}\right), \quad f = 1, \ldots, T, \tag{2.1}$$

kde T je délka záznamu emisní události. Aby člen f získal fyzikální smysl frekvence, bylo by ho třeba vynásobit faktorem $\frac{f_{vz}}{T}$, kde f_{vz} je vzorkovací frekvence. Jelikož je $\tilde{S}_f$ komplexní, platí $\tilde{S}_f = |\tilde{S}_f|e^{i\varphi}$, kde $\frac{1}{T}|\tilde{S}_f|$ nazýváme amplitudou harmonické složky, φ její fází. Dále ztotožňujeme pojmy spektrum a amplitudy spektra, fáze pro nás nejsou významné.

Pro výpočet klasifikačních atributů navíc spektrum emisní události normujeme na jedničku

$$S_f = \frac{|\tilde{S}_f|}{\sum_{f=0}^{T-1} |\tilde{S}_f|},$$

abychom potlačili závislost amplitud spektra na energii signálu a tím se pokusili odbourat vliv vzdálenosti zdroje emise od snímače.



Obrázek 2.1: Příklad signálu akustické emise a jeho spektra

2.2 Použité atributy signálů

1. Atribut Z_c:

$$Z_c = \sum_{t=\tilde{t}}^{T-1} \delta(x_t),$$

$$\text{kde } \tilde{t} = \min J, J = \left\{ j \in [0, T-1] : x_j \geq c \max_{t \in [0,T-1]} |x_t| \right\}, \quad c \in (0,1),$$

$$\delta(x_t) = \begin{cases} 1 & \text{když } \operatorname{sgn}(x_t x_{t+1}) = -1, \\ 0 & \text{jinak.} \end{cases}$$

   Z_c počítá průchody signálu nulovou hodnotou po prvním překročení prahové hodnoty $c \max_{t \in [0,T-1]} x_t$.

2. Atribut M:

$$M = \arg\max_{t \in [0,T-1]} |x_t|.$$

   M určuje polohu maximální amplitudy signálu.

2.3 Použité atributy ze spektra signálu

1. Atribut W_α:

$$W_\alpha = \arg\min_{l \in [0,T-1]} \sum_{f=0}^{T-1} |l - f| \left| |S_f| - \left| \overline{S_T} \right| \right|^\alpha,$$

$$\text{kde } \left| \overline{S_T} \right| = \frac{1}{T} \sum_{f=0}^{T-1} |S_f| \quad a \quad \alpha \in [1, \infty).$$

W zachycuje polohu významných amplitud ve smyslu frekvence.

2. Atribut Q_β:

$$Q_\beta = \min\Big\{ F \in [0, T-1] : \sum_{f=0}^{F} |X_f| \geq \beta \Big\} \quad \text{pro } \beta \in (0, 1).$$

Q_β představuje z teorie pravděpodobnosti převzatý $\beta - kvantil$.

3. Atribut S_γ:

$$S_\gamma = \max J - \min J,$$
$$\text{kde } J = \Big\{ j \in [0, T-1] : |X_j| \geq \gamma \max_{f \in [0, T-1]} |X_f| \Big\}, \quad \gamma \in (0, 1).$$

S zachycuje největší vzdálenost poloh vyšších hodnot spektra.

4. Atribut P:

$$P = \frac{1}{T} \sum_{f=0}^{T-1} f S_f.$$

P představuje střední hodnotu spektra, které lze považovat za hustotu pravděpodobnosti.

Část parametrů byla převzata z [4], příklady dalších lze nalézt v [5], parametry P a M jsou vlastní.

Kapitola 3

Hierarchické metody klasifikace

Hierarchické metody patří mezi nejjednodušší klasifikační metody. K rozdělení objektů (vzorků) výběrového prostoru $\mathbb{R}^d$ do jednotlivých shluků využívají, stejně jako ostatní klasifikační metody, rozdílnost (též vzdálenost) a podobnost těchto vzorků.

Definice. Rozdílností $\varrho(x, y)$ mezi vzorky x a y, $x, y \in \mathbb{R}^d$, nazveme funkci $\varrho : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, pokud splňuje následující podmínky:

$$\begin{aligned} &\varrho(x, y) \geq 0, && \text{pro každé } x, y \in \mathbb{R}^d, \\ &\varrho(x, x) = 0, && \text{pro každé } x \in \mathbb{R}^d, \\ &\varrho(x, y) = \varrho(y, x), && \text{pro každé } x, y \in \mathbb{R}^d. \end{aligned}$$

Rozdílnost není totéž co metrika, v definici metriky navíc požadujeme, aby platila trojúhelníková nerovnost, tj. $\forall x, y, z \in \mathbb{R}^p$ platí

$$\varrho(x, y) + \varrho(y, z) \geq \varrho(x, z),$$

dvojice $(\mathbb{R}^d, \varrho)$ se pak stává metrickým prostorem.

Dále uvádíme několik vybraných příkladů vzdáleností mezi vzorky x a y na prostoru $\mathbb{R}^p$.

- *Eukleidovská vzdálenost*

$$\varrho(x, y) = \| x - y \| = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2},$$

- *Manhattanská vzdálenost*

$$\varrho(x, y) = \sum_{i=1}^{d} |x_i - y_i|$$

- *Minkovského vzdálenost*

$$\varrho(x, y) = \sqrt[r]{\sum_{i=1}^{d} |x_i - y_i|^r}, \quad r \in (0, \infty), \text{ pro } r \geq 1 \text{ metrická vzdálenost},$$

- *Čebyševova vzdálenost*

$$\varrho(x, y) = \max_{i \in \hat{d}} |x_i - y_i|$$

Na základě vzdálenosti mezi vzorky dochází při hierarchickém shlukování buď k postupnému spojování menších shluků (aglomerativní metody) nebo k rozdělování velkých shluků na menší (divizivní metody). V každém kroku hierarchické metody se obvykle hledá lokální optimum bez ohledu na další postup, tyto metody proto nemohou zaručit nalezení globálního optimálního řešení.

3.1 Aglomerativní metody

Na začátku aglomerativní metody shlukování reprezentuje každý vzorek jeden samostatný shluk, v každém kroku se sloučí dva nejbližší shluky. Tento proces se opakuje, dokud není splněna předdefinovaná prahová podmínka, např. je dosažen jistý počet shluků. Pro chod tohoto algoritmu je nutné definovat vzdálenost mezi jednotlivými shluky, která je přirozeně závislá na použité vzdálenosti mezi dvěma objekty.

Uvažujme konečnou množinu objektů $\{x_i\}_{i=1}^N$ na prostoru $\mathbb{R}^d$. Označíme-li R, Q dva různé shluky a $card(R)$,resp. $card(Q)$, počet objektů ve shluku R, resp. Q, potom můžeme vzdálenost mezi shluky R a Q vyjádřit např. následujícími způsoby:

- *Group average method*

$$\varrho(R,Q) = \frac{1}{card(R)card(Q)} \sum_{\substack{x_i \in R \\ x_j \in Q}} \varrho(x_i, x_j),$$

- *Single linkage*

$$\varrho(R,Q) = \min_{\substack{x_i \in R \\ x_j \in Q}} \varrho(x_i, x_j),$$

- *Complete linkage*

$$\varrho(R,Q) = \max_{\substack{x_i \in R \\ x_j \in Q}} \varrho(x_i, x_j).$$

Aglomerativní metody jsme v rámci naší práce neimplementovali, další informace v [2].

3.2 Divizivní metody

Divizivní metoda pracuje v opačném směru než aglomerativní, na počátku přísluší všechny objekty jednomu velkému shluku. V každém kroku dojde k rozštěpení shluku s největším průměrem a k přesunutí jednoho objektu do jiného shluku. Průměr shluku Q se spočte jako

$$diam(Q) = \max_{\substack{x_i \in R \\ x_j \in Q}} \varrho(x_i, x_j),$$

proces se opakuje dokud nedojde k překročení definovaného prahového kritéria.

Nyní zde podrobněji popíšeme binární verzi divergenčního klasifikátoru, který byl implementován a vyzkoušen na naměřených datech. Mějme tedy na začátku jeden shluk $A \subset \mathbb{R}^d$ obsahující všechny objekty a prázdný shluk B. Pro každý objekt $x_i \in A$ spočteme následující výraz

$$\varrho(x_i, A \setminus \{x_i\}) = \frac{1}{card(A) - 1} \sum_{\substack{x_j \in A \\ x_j \neq x_i}} \varrho(x_i, x_j). \tag{3.1}$$

Objekt x_i, pro který výraz (3.1) dosáhne maximální hodnoty, přemístíme z A do B

$$A_{new} = A_{old} \setminus \{x_i\},$$
$$B_{new} = B_{old} \cup \{x_i\}.$$

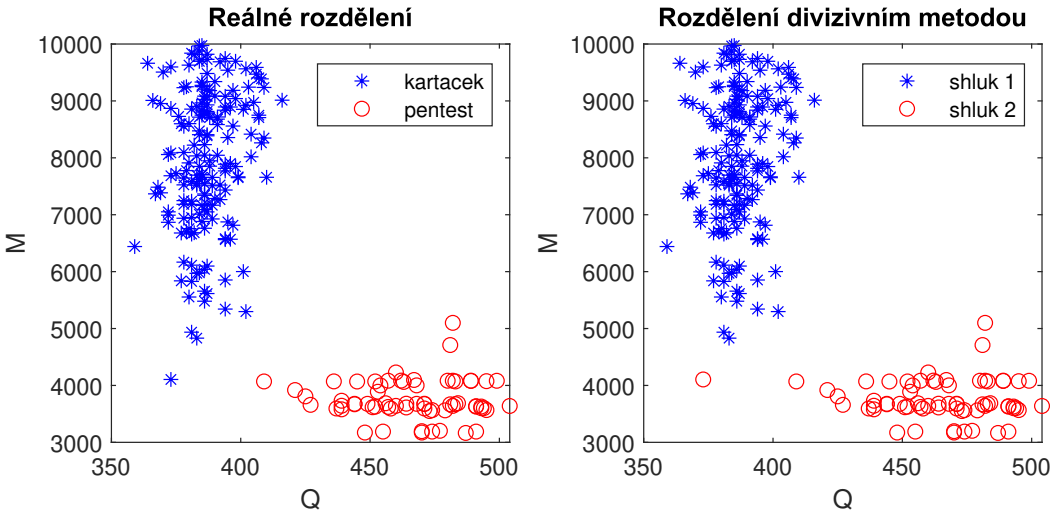
Dále budeme v každém kroku hledat objekt, který bychom přemístili z A do B. Pro každý objekt $x_i \in A$ spočteme výraz

$$\nabla_i \varrho = \varrho(x_i, A \setminus \{x_i\}) - \varrho(x_i, B) = \frac{1}{card(A) - 1} \sum_{\substack{x_j \in A \\ x_j \neq i}} \varrho(x_i, x_j) - \frac{1}{card(B)} \sum_{x_k \in B} \varrho(x_i, x_k).$$

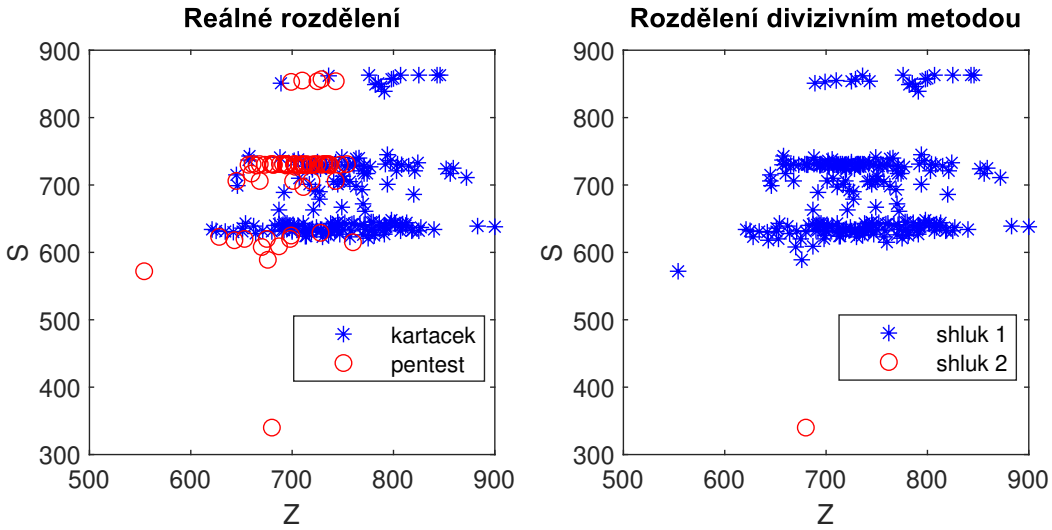
Pokud je $max_{x_i \in A} \nabla_i \varrho > 0$, přemístíme prvek x_i pro který se nabývá maxima, z A do B a postup opakujeme. Je-li $\nabla_i \varrho \leq 0$ pro všechna $x_i \in A$, je klasifikace dokončena.

Příklad 3.1. Na obrázku 3.1 je znázorněn příklad klasifikace dvou zdrojů akustické emise - pentestu a kartáčku. Byly použity atributy M a $Q_{0,33}$, tedy výběrový prostor je $\mathbb{R}^2$. Vybrané atributy dokázaly efektivně odhalit rozdílnost mezi dvěma typy signálu a oddělit tak příslušné prvky výběrového prostoru, díky čemuž divizivní klasifikátor zaznamenal pouze jednu chybu v klasifikaci.

Naproti tomu na obrázku 3.2 je vidět neúspěšný pokus o divizivní klasifikaci stejné sady dat za použití atributů $S_{1/3}$ a $Z_{1/20}$. Zde se prvky výběrového prostoru neseskupily podle příslušnosti k typu zdroje signálu, ale vytvořily nesourodé uskupení. Jedno z odlehlých pozorování dokonce způsobilo, že se algoritmus divizivního dělení zastavil hned po první iteraci a jako jeden ze shluků označil jednoprvkovou množinu obsahující pouze tento odlehlý prvek.



Obrázek 3.1: Úspěšná klasifikace divizivní metodou



Obrázek 3.2: Neúspěšná klasifikace divizivní metodou

Kapitola 4

Klasifikace na základě modelu

Klasifikace na základě modelu (Model Based Clustering - MBC) je klasifikační metoda, ve které na naměřená data pohlížíme jako na realizace náhodných veličin, jejichž hustota je ve formě směsi normálních rozdělení, kde každá komponenta směsi reprezentuje jiný shluk. Předpokládáme tedy, že reprezentace signálů jednoho fyzikálního původu budou ve výběrovém prostoru tvořit více-méně eliptické shluky.

4.1 Distribuční směsi

Definice. Distribuční směsí $p(x)$ rozumíme každou konvexní kombinaci hustot pravděpodobnosti $p_j(x)$

$$p(x) = \sum_{j=1}^{M} \alpha_j p_j(x), \qquad \sum_{j=1}^{M} \alpha_j = 1,$$

kde $M \in \mathbb{N}$ je počet komponent směsi, α_j jsou váhy jednotlivých komponent.

Ve zbytku této kapitoly uvažujeme za komponenty směsi pouze hustoty d-dimenzionálního normálního rozdělení. Jedná se o parametrickou hustotu, můžeme tedy psát

$$p(x|\Theta) = \sum_{j=1}^{M} \alpha_j p_j(x|\theta_j).$$

Odhadování rozdělení (4.1) se tedy redukuje na odhad parametrů

$$\Theta = (\alpha_1, \ldots, \alpha_M, \theta_1, \ldots, \theta_M), \tag{4.1}$$

které distribuční směs plně určují.

4.2 EM algoritmus

EM algoritmus (*expectation maximization*) je iterační metoda pro hledání maximálně věrohodného odhadu z pozorovaných dat v modelu, který obsahuje skryté proměnné. Tento algoritmus budeme při klasifikaci používat pro odhadnutí skupiny parametrů (4.1). V následující části textu zavedeme EM algoritmus a uvedeme jeho vybrané základní vlastnosti.

Uvažujme statistický model, v němž označíme množinu *pozorovaných* dat $X \subset \mathbb{R}^d$ a množinu *skrytých* dat $Z \subset \mathbb{R}^g$. Společný soubor pozorování $(X, Z) \subset \mathbb{R}^{d+g}$ nazveme *kompletní* data. Nechť je dále na (X, Z) definován systém hustot pravděpodobnosti

$$f(x, z|\Theta), \quad x \in X, \ z \in Z, \ \Theta \in \Omega,$$

kde Ω je parametrický prostor daných rozdělení. Zvolme $x \in X$ libovolně pevně, pak můžeme vyjádřit podmíněnou hustotu pravděpodobnosti $z \in Z$ při daném x jako

$$f_Z(z|x, \Theta) = \frac{f(x, z|\Theta)}{f_X(x|\Theta)}, \tag{4.2}$$

kde

$$f_X(x|\Theta) = \int_{\mathbb{R}^g} f(x, z|\Theta) \ \mathrm{d}z.$$

Po zlogaritmování vztahu (4.2) a úpravě dostáváme

$$\ln f_X(x|\Theta) = \ln f(x, z|\Theta) - \ln f_Z(z|x, \Theta). \tag{4.3}$$

Výraz (4.3) přeznačíme do formalismu logaritmických věrohodnostních funkcí, upravujeme tedy jako $\ln f_X(x|\Theta) = \ln L(\Theta|x) = l_X(\Theta|x)$, u ostatních členů analogicky,

$$l_X(\Theta|x) = l(\Theta|x, z) - l_Z(x, \Theta|z). \tag{4.4}$$

Maximalizovat věrohodnost $l(\Theta|x, z)$ na Ω není možné, jelikož neznáme hodnoty z. Budeme proto hledat maximum podmíněné střední hodnoty $l(\Theta|x, z)$ při daném x.

Definice. Pro všechny dvojice $(\Theta, \Phi) \in \Omega \times \Omega$ zavedeme funkci $Q(\Theta, \Phi)$ předpisem

$$Q(\Theta, \Phi) = E_z \left(l(\Theta|x, z)|x, \Phi \right).$$

Abychom zajistili konečnost výrazu (4.3), požadujeme $f(x, z|\Theta) > 0$ pro skoro všechna $(x, z) \in \mathbb{R}^{d+g}$. Nyní už můžeme definovat iteraci EM algoritmu následujícím schématem.

Definice. Pro $k \in \mathbb{N}_0$ definujeme k-tou iteraci EM algoritmu $\Theta^k \to \Theta^{k+1}$ následovně:

- E-krok: výpočet $Q(\Theta, \Theta^k)$,
- M-krok: nalezení Θ^{k+1} tak, že platí $\Theta^{k+1} = \arg\max_{\Theta \in \Omega} Q(\Theta, \Theta^k)$.

Věta 1. *EM algoritmus v každé iteraci zvyšuje hodnotu věrohodnosti* $l(\Theta|x)$,

$$l(\Theta^{k+1}|x) \geq l(\Theta^k|x), \quad \forall k \in \mathbb{N}_0.$$

Rovnost nastává právě tehdy, když platí

$$Q(\Theta^{k+1}, \Theta^k) = Q(\Theta^k, \Theta^k),$$

a zároveň

$$f_Z(z|x, \Theta^{k+1}) = f_Z(z|x, \Theta^k) \quad \textit{skoro všude v } \mathbb{R}^g.$$

Důkaz. Označíme

$$H(\Theta, \Phi) = E_z(l_Z(x, \Theta|z)|x, \Phi).$$

Aplikací podmíněné střední hodnoty na rovnici (4.4) získáváme vztah

$$l_X(\Theta|x) = Q(\Theta, \Phi) - H(\Theta, \Phi), \tag{4.5}$$

kde $E(l_X(\Theta|x)|x, \Phi) = l_X(\Theta, x)$, protože $l_X(\Theta, x)$ nezávisí na $z \in Z$. Do vztahu (4.5) dosadíme za $(\Theta, \Phi) = (\Theta^{k+1}, \Theta^k)$ resp. $(\Theta, \Phi) = (\Theta^k, \Theta^k)$, tedy

$$l_X(\Theta^{k+1}|x) = Q(\Theta^{k+1}, \Theta^k) - H(\Theta^{k+1}, \Theta^k), \tag{4.6}$$

$$l_X(\Theta^k|x) = Q(\Theta^k, \Theta^k) - H(\Theta^k, \Theta^k). \tag{4.7}$$

Po odečtení (4.7) od (4.6) pak dostáváme

$$l_X(\Theta^{k+1}|x) - l_X(\Theta^k|x) = Q(\Theta^{k+1}, \Theta^k) - Q(\Theta^k, \Theta^k) - [H(\Theta^{k+1}, \Theta^k) - H(\Theta^k, \Theta^k)]. \tag{4.8}$$

V M-kroku EM algoritmu je Θ^{k+1} voleno jako $\Theta^{k+1} = \arg\max_{\Theta \in \Omega} Q(\Theta, \Theta^k)$ a tedy platí

$$Q(\Theta^{k+1}, \Theta^k) - Q(\Theta^k, \Theta^k) \geq 0. \tag{4.9}$$

Stačí tedy dokázat, že

$$H(\Theta^{k+1}, \Theta^k) - H(\Theta^k, \Theta^k) \leq 0$$

a že při $f_Z(z|x, \Theta^{k+1}) = f_Z(z|x, \Theta^k)$ nastává v tvrzení věty rovnost. Z předpokladu $f(x, z|\Theta) > 0$ plyne, že $f_Z(z|x, \Theta) > 0$ skoro všude na (X, Z) a pro všechna $\Theta \in \Omega$. Za pomoci Jensenovi nerovnosti můžeme psát

$$\begin{aligned} H(\Theta^{k+1}, \Theta^k) - H(\Theta^k, \Theta^k) &= \int \ln \frac{f_Z(z|x, \Theta^{k+1})}{f_Z(z|x, \Theta^k)} f_Z(z|x, \Theta^k) \mathrm{d}z \\ &\overset{\text{Jen.}}{\leq} \ln \int \left(\frac{f_Z(z|x, \Theta^{k+1})}{f_Z(z|x, \Theta^k)} \right) f_Z(z|x, \Theta^k) \mathrm{d}z \\ &= \ln \int f_Z(z|x, \Theta^{k+1}) \mathrm{d}z \\ &= \ln 1 = 0. \end{aligned}$$

Platí tedy $H(\Theta^{k+1}, \Theta^k) - H(\Theta^k, \Theta^k) \leq 0$. S použitím (4.9) vyplývá ze vztahu (4.8)

$$l_X(\Theta^{k+1}|x) \geq l_X(\Theta^k|x).$$

Pokud navíc $f_Z(z|x, \Theta^{k+1}) = f_Z(z|x, \Theta^k)$, potom z definice funkce H platí $H(\Theta^{k+1}, \Theta^k) = H(\Theta^k, \Theta^k)$. Při splnění druhého předpokladu věty $Q(\Theta^{k+1}, \Theta^k) = Q(\Theta^k, \Theta^k)$ tedy skutečně z (4.8) plyne rovnost

$$l_X(\Theta^{k+1}|x) = l_X(\Theta^k|x).$$

□

Víme tedy, že posloupnost parametrů $\{\Theta^k\}_{k \geq 0}$ daná EM algoritmem zvyšuje nebo zachovává v každém kroku hodnotu věrohodnostní funkce. Otázkou konvergence ke stacionárnímu bodu věrohodnostní funkce se zabývá následující věta.

Věta 2. *Nechť posloupnost iterací EM algoritmu* $\left\{\Theta^k\right\}_{k\geq 0}$ *splňuje*

- $\left.\frac{\partial Q(\Theta,\Theta^k)}{\partial\Theta}\right|_{\Theta=\Theta^{k+1}} = 0$
- Θ^k *konverguje k* $\Theta^* \in \Omega$

a nechť $f_Z(z|x,\Theta)$ *je třídy* $C^{(1)}$. *Pak platí*

$$\left.\frac{\partial l_X(\Theta|x)}{\partial\Theta}\right|_{\Theta=\Theta^*} = 0.$$

Poznámka. *Tato věta tedy říká, že pokud posloupnost* Θ^k *konverguje, potom konverguje k stacionárnímu bodu věrohodnostní funkce* $l_X(\Theta|x)$. *Nic nám však nezaručuje, že v případě existence více stacionárních bodů věrohodnosti bude EM algoritmus konvergovat ke globálnímu maximu.*

4.3 MBC klasifikace

Mějme dána pozorování $\boldsymbol{x} = (x_1,\ldots,x_n), x_i \in \mathbb{R}^d$, která chceme rozdělit do M shluků. Při MBC klasifikaci nahlížíme na data jako na realizace nezávislé a stejně rozdělené (i.i.d.) náhodné veličiny, jejíž hustota je ve formě distribuční směsi, kde každá složka směsi má normální rozdělení a reprezentuje jeden shluk. Distribuční směs má v tomto případě tvar

$$p(x_k|\Theta) = \sum_{i=1}^{M} \alpha_i f_i(x_k|\theta_i), \tag{4.10}$$

$$\alpha_i \geq 0 \quad \forall i \in \widehat{M} \quad \wedge \quad \sum_{i=1}^{M} \alpha_i = 1,$$

kde f_i je hustota d-rozměrného normálního rozdělení

$$f_i(x_k|\theta_i) = f_i(x_k|\mu_i,\mathbb{C}_i) = \frac{1}{(2\pi)^{\frac{d}{2}}|\mathbb{C}_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_k-\mu_i)^T\mathbb{C}_i^{-1}(x_k-\mu_i)\right).$$

Pro úspěšnou klasifikaci je třeba maximalizovat věrohodnostní funkci směsi (4.10)

$$L(\theta_1,\ldots,\theta_M,\alpha_1,\ldots,\alpha_M|\boldsymbol{x}) = \prod_{k=1}^{n}\sum_{i=1}^{M} \alpha_i f_i(x_k|\theta_i),$$

resp. logaritmickou věrohodnostní funkci

$$l(\theta_1,\ldots,\theta_M,\alpha_1,\ldots,\alpha_M|\boldsymbol{x}) = \sum_{k=1}^{n} \ln \sum_{i=1}^{M} \alpha_i f_i(x_k|\theta_i).$$

Hledáme tedy maximálně věrohodný odhad Θ^* parametru $\Theta = (\theta_1,\ldots,\theta_M,\alpha_1,\ldots,\alpha_M) \in \Omega$, tzn $\Theta^* = \arg\max_{\Theta\in\Omega} l(\Theta|\boldsymbol{x})$.

Najít Θ^* analyticky není možné, využijeme proto v sekci 4.2 popsaný EM algoritmus. Za kompletní data při shlukování budeme považovat data $y_k = (x_k, z_k)$, kde $z_k = (z_{k1},\ldots,z_{kM})$ jsou chybějící data, která určují příslušnost pozorování x_k k jednomu z M shluků, tedy

$$z_{ki} = \begin{cases} 1 & \text{pokud } x_k \text{ patří ke shluku } i, \\ 0 & \text{jinak.} \end{cases}$$

Při daném z_k je hustota pravděpodobnosti x_k rovna výrazu

$$p(x_k|z_k) = \prod_{i=1}^{M} f_i(x_k|\theta_i)^{z_{ki}}.$$

Jestliže předpokládáme, že $\boldsymbol{x}$ je i.i.d. pak i vektor $\boldsymbol{z}$ je i.i.d. a platí

$$p(z_k) = \prod_{i=1}^{M} \alpha_i^{z_{ki}}.$$

Věrohodnost pro kompletní data potom můžeme psát ve tvaru

$$L(\Theta|\boldsymbol{y}) = L(\theta_1, \ldots, \theta_M, \alpha_1, \ldots, \alpha_M|\boldsymbol{x}, \boldsymbol{z}) = \prod_{k=1}^{n}\prod_{i=1}^{M} \alpha_i^{z_{ki}} f_i(x_k|\theta_i)^{z_{ki}}.$$

Po zlogaritmování dostáváme přívětivější tvar logaritmické věrohodnosti

$$l(\Theta|\boldsymbol{y}) = \sum_{k=1}^{n}\sum_{i=1}^{M} z_{ki} \ln\left[\alpha_i f_i(x_k|\theta_i)\right]. \tag{4.11}$$

Výraz (4.11) nelze maximalizovat přímo, protože závisí na hodnotách z_{ki}, které neznáme. Využijeme proto EM algoritmu, který radí maximalizovat podmíněnou střední hodnotu z (4.11), tzn.

$$\begin{aligned} E_z\left(l(\Theta|\boldsymbol{z}, \boldsymbol{x})|\boldsymbol{x}, \boldsymbol{\Theta}\right) &= \sum_{k=1}^{n}\sum_{i=1}^{M} \gamma(z_{ki}) \ln\left[\alpha_i f_i(x_k|\theta_i)\right] \\ &= \sum_{k=1}^{n}\sum_{i=1}^{M} \gamma(z_{ki}) \ln \alpha_i + \sum_{k=1}^{n}\sum_{i=1}^{M} \gamma(z_{ki}) \ln f_i(x_k|\theta_i), \end{aligned} \tag{4.12}$$

kde $\gamma(z_{ki})$ je posteriorní pravděpodobnost příslušnosti prvku x_k k i-té komponentě směsi. Spočte se pomocí Bayesovy věty jako

$$\gamma(z_{ki}) = E(z_{ki}|\Theta, x_k) = P(z_{ki} = 1|\Theta, x_k) = \frac{\alpha_i f_i(x_k|\theta_i)}{\sum_{j=1}^{M} \alpha_j f_j(x_k|\theta_j)}.$$

K nalezení maxima výrazu (4.12) se využije iterativní EM algoritmus. Nejprve inicializujeme parametry $\Theta^0 = (\theta_1^0, \ldots, \theta_M^0, \alpha_1^0, \ldots, \alpha_M^0) = (\mu_1^0, \ldots, \mu_M^0, \mathbb{C}_1^0, \ldots, \mathbb{C}_M^0, \alpha_1^0, \ldots, \alpha_M^0)$. Dále v cyklu provádíme následující kroky.

- **E krok:** Spočteme $\gamma(z_{ki})$ pomocí současných hodnot parametrů Θ^t

$$\gamma(z_{ki}) = \frac{\alpha_i^t f_i(x_k|\theta_i^t)}{\sum_{j=1}^{M} \alpha_j^t f_j(x_k|\theta_j^t)}.$$

- **M krok:** Odhadneme nové hodnoty parametrů Θ^{t+1}

$$\begin{aligned} \mu_i^{t+1} &= \frac{1}{n_i}\sum_{k=1}^{n} \gamma(z_{ki}) x_k, \\ \mathbb{C}_i^{t+1} &= \frac{1}{n_k}\sum_{k=1}^{n} \gamma(z_{ki})(x_k - \mu_i^{t+1})(x_k - \mu_i^{t+1})^T, \\ \alpha_i^{t+1} &= \frac{n_k}{n}, \quad n_k = \sum_{k=1}^{n} \gamma(z_{ki}). \end{aligned}$$

V každé iteraci algoritmu kontrolujeme, zda není splněna konvergenční podmínka. Ta je obvykle formulována jako rozdíl podmíněné střední hodnoty věrohodnostní funkce ve dvou po sobě jdoucích iteracích. Cyklus tedy přerušujeme např. pokud platí

$$1 - \frac{E_z(l(\Theta^t|\boldsymbol{z},\boldsymbol{x})|\boldsymbol{x},\Theta^t)}{E_z(l(\Theta^{t+1}|\boldsymbol{z},\boldsymbol{x})|\boldsymbol{x},\Theta^{t+1})} < \varepsilon.$$

Hodnotu ε obvykle volíme $\varepsilon = 10^{-4}$. Poslední vektor parametrů Θ^{t+1} po kterém zastavujeme iteraci označíme Θ^*.

Klasifikaci následně provádíme pomocí posteriorních pravděpodobností $\gamma(z_{ki})$, které vypočteme za dosazení sady parametrů Θ^*. Následně je k-té pozorování přiřazeno j-té komponentě, právě tehdy když

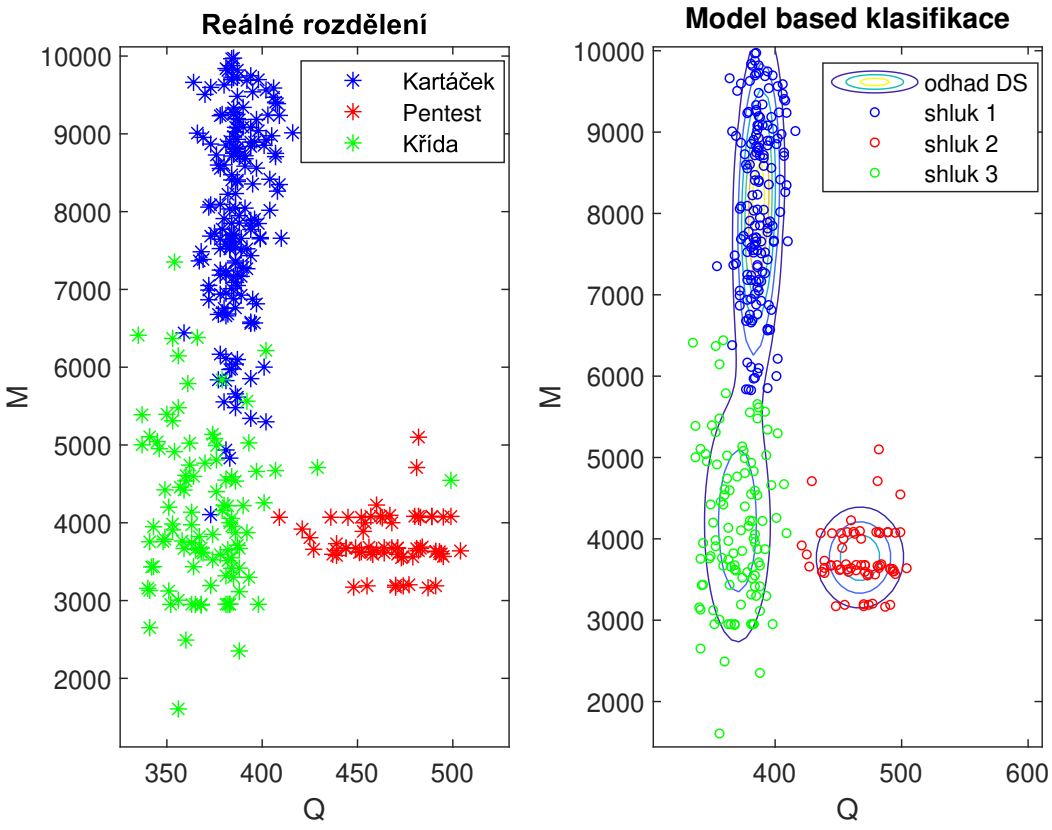
$$\arg\max_{i\in\widehat{M}} \gamma(z_{ki})\Big|_{\Theta=\Theta^*} = j.$$

Ze sekce o EM algoritmu víme, že algoritmus buď diverguje do nekonečna, nebo nutně konverguje ke stacionárnímu bodu věrohodnosti. Pokud bude věrohodnostní funkce hladká, omezená s jediným stacionárním bodem, pak algoritmus najde maximálně věrohodný odhad Θ^* bez ohledu na volbu počátečních podmínek. V praxi má však věrohodnost distribuční směsi často mnoho lokálních maxim a sedlových bodů, obzvláště při vyšším počtu komponent směsi a nižším počtu dostupných dat. Volba počátečních hodnot parametrů Θ^0 tak hraje zásadní roli v tom zda najdeme použitelné lokální maximum. Nejjednodušší způsob, jak se ujistit, že najdeme použitelný výsledek, je nechat EM algoritmus proběhnou několikrát s náhodně vybranými inicializačními parametry, které reprezentují různé části parametrického prostoru Ω, následně vyřadit divergující řešení a z řešení, která zkonvergovala, vybrat to s nejvyšší hodnotou věrohodnostní funkce.

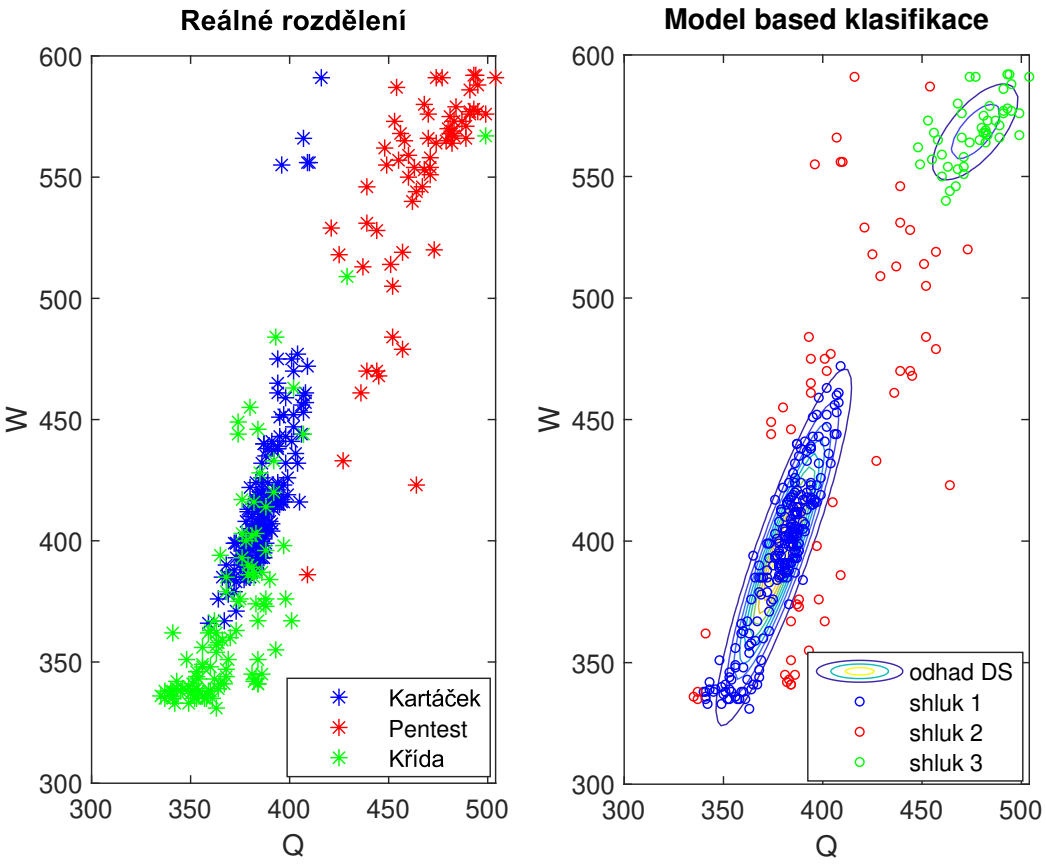
Pro klasifikaci směsi s vyšším počtem komponent se tento naivní přístup již příliš neosvědčil. Algoritmus často diverguje, případně považuje za shluk jen několik málo odlehlých pozorování a tím efektivně snižuje počet rozklasifikovaných shluků. V další části textu se pokoušíme MBC klasifikaci vylepšit zapojením trénovací množiny dat.

Příklad 4.1. Provedli jsme klasifikaci dat pocházejících ze tří různých budících zdrojů - pentestu, kartáčku a křídy. Při volbě atributů M a $Q_{0,33}$ dokázal EM algoritmus správně určit centra shluků a výsledná distribuční směs, jak je vidět na obrázku 4.1, dobře odpovídá reálnému rozdělení naměřených signálů. MBC metoda tak v tomto případě klasifikuje s velmi dobrou úspěšností přes 95%.

Při volbě atributů W_2 a $Q_{0,33}$, viz. obrázek 4.2 však prvky výběrového prostoru reprezentující signály kartáčku a křídy tvoří dva, přímo na sebe navazující shluky, a dochází k překryvu těchto dvou druhů dat. EM algoritmus je tak považuje za jediný veliký shluk, správně sice určuje signály pocházející z pentestů za samostatnou komponentu směsi, ale třetí komponentu volí jako rozdělení s velikým rozptylem a přiřazuje k němu ta pozorování napříč celým výběrovým prostorem, která se vychylují z prvních dvou rozdělení. Vzniká tak "shluk" dat připomínající spíše určitý šum, než samostatnou skupinu pozorování.



Obrázek 4.1: Úspěšná klasifikace metodou MBC



Obrázek 4.2: Neúspěšná klasifikace metodou MBC

4.4 MBC s učitelem

Tato klasifikační metoda, kterou nazýváme SMBC (Supervised Model Based Clustering) využívá trénovací data pouze k nalezení vhodných inicializačních parametrů Θ^0 v EM algoritmu, jinak se od MBC metody neliší.

Nechť máme trénovací množinu shluků $T = (T_1, \ldots, T_M) \subset \mathbb{R}^d$, kde $x_{ik} \in T_i$ právě tehdy když pozorování x_{ik} náleží i-tému shluku a M je počet shluků, které budeme hledat. Dále máme množinu testovacích dat $\boldsymbol{x} = (x_1, \ldots, x_n)$, $x_i \in \mathbb{R}^d$, pocházejících ze stejného měření jako trénovací množina, kterou budeme chtít klasifikovat. Tvar distribuční směsi popisující data je stejný pro trénovací i testovací množiny. Díky tomu můžeme využít znalost rozdělení pozorování pocházejících z jednotlivých množin T_i pro nastavení počátečních parametrů v EM algoritmu.

O datech z T_i předpokládáme pro všechna i, že pocházejí z normálního rozdělení. Nejprve tedy pro každé i najdeme maximálně věrohodný odhad p-rozměrného normálního rozdělení za předpokladu pozorování $(x_{i1}, \ldots, x_{in_i}) = T_i$, kde n_i je počet pozorování v T_i. Věrohodnostní funkce bude mít tvar

$$L(\mu_i, \mathbb{C}_i | T_i) = \prod_{k=1}^{n_i} f(x_{ik} | \mu_i, \mathbb{C}_i),$$

kde $f(x_{ik}|\mu_i, \mathbb{C}_i)$ je hustota normálního rozdělení. Po dosazení za f a zlogaritmování

$$\begin{aligned} l(\mu_i, \mathbb{C}_i|T_i) &= \sum_{k=1}^{n_i} \ln\left(\frac{1}{(2\pi)^{\frac{d}{2}}|\mathbb{C}_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_{ik}-\mu_i)^T\mathbb{C}_i^{-1}(x_{ik}-\mu_i)\right)\right) \\ &= -\frac{n_i}{2}\ln|\mathbb{C}_i| - \frac{1}{2}\sum_{k=1}^{n_i}(x_{ik}-\mu_i)^T\mathbb{C}_i^{-1}(x_{ik}-\mu_i), \end{aligned}$$

kde jsme v poslední úpravě vypustili výrazy nezávisející na parametrech μ_i a $\mathbb{C}_i$. Derivováním podle μ_i

$$\frac{\partial l}{\partial \mu_i} = \sum_{k=1}^{n_i}(x_{ik}-\mu_i)^T\mathbb{C}_i^{-1},$$

a porovnáním s nulou dostáváme tvar MLE odhadu parametru μ_i

$$\widehat{\mu_i} = \frac{1}{n_i}\sum_{k=1}^{n_i} x_{ik}.$$

Provedení derivace podle $\mathbb{C}_i$ je technicky náročnější a vyžaduje využití několika závěrů pokročilé lineární algebry, jak je uvedeno v [6]. Zde jen ve zkratce

$$\frac{\partial l}{\partial \mathbb{C}_i^{-1}} = \frac{n_i}{2}\mathbb{C}_i - \frac{1}{2}\sum_{k=1}^{n_i}(x_{ik}-\mu_i)(x_{ik}-\mu_i)^T.$$

Po porovnání s nulou a úpravě získáváme odhad

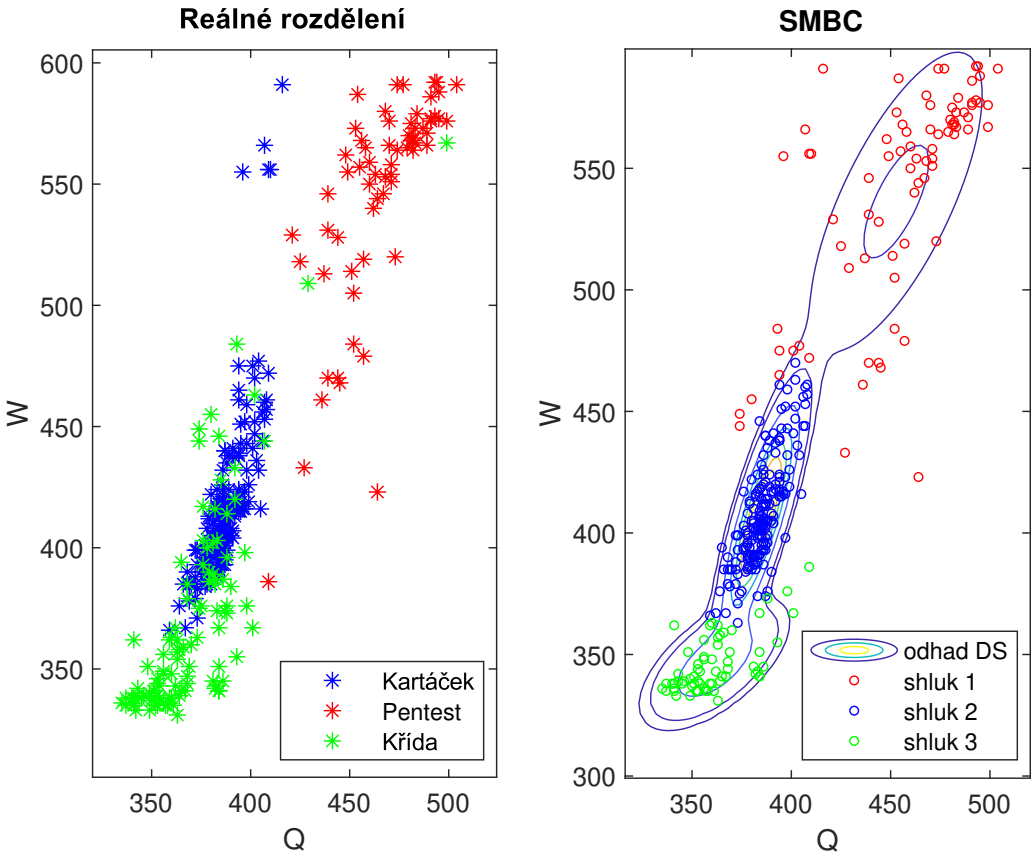
$$\widehat{\mathbb{C}_i} = \frac{1}{n_i}\sum_{k=1}^{n_i}(x_{ik}-\widehat{\mu_i})(x_{ik}-\widehat{\mu_i})^T.$$

Po spočtení odhadů $\widehat{\mu_i}$ a $\widehat{\mathbb{C}_i}$ ze všech trénovacích množin máme určenou první část parametru $\Theta^0 = (\mu_1^0, \ldots, \mu_M^0, \mathbb{C}_1^0, \ldots, \mathbb{C}_M^0, \alpha_1^0, \ldots, \alpha_M^0)$, kde za μ_i^0 klademe $\widehat{\mu_i}$ a za $\mathbb{C}_i^0$ klademe $\widehat{\mathbb{C}_i}$. Zbývá nastavit hodnoty $\alpha_1^0, \ldots, \alpha_M^0$. Ty volíme buď rovnoměrně jako $\alpha_i^0 = \frac{1}{M} \quad \forall i$, nebo tak, aby reflektovaly velikost trénovacích množin

$$\alpha_i^0 = \frac{n_i}{\sum_{j=1}^{M} n_j}.$$

Tím máme kompletní vektor inicializačních parametrů Θ^0, který použijeme při aplikaci MBC metody na testovací množinu dat tak, jak je popsána v předchozí kapitole .

Příklad 4.2. Metodu SMBC jsme aplikovali na stejná data jako v druhé části příkladu 4.1, na obrázku 4.3 je vidět výsledek. Úspěšnost klasifikace se oproti metodě MBC výrazně zlepšila, protože EM algoritmus dosáhl díky lepšímu nastavení inicializačních parametrů jiného lokálního maxima věrohodnostní funkce, takto vzniklá distribuční směs už rozdělí signály na tři kompaktní shluky. Signály pocházející z křídy a kartáčku se při této volbě atributů silně překrývají, použití této sady atributů je tedy nepraktické a výrazně lepších výsledků klasifikace lze dosáhnout pouze použitím jiné sady charakteristik signálu.



Obrázek 4.3: Klasifikace metodou SMBC

4.5 Odhad distribuční směsi bez použití EM algoritmu

Další navrženou klasifikační metodou je GMMC (Gaussian Mixture Model Clustering). Tato metoda opět předpokládá, že data jsou realizací náhodné veličiny s hustotou p ve tvaru distribuční směsi, jejíž komponenty jsou hustoty d-rozměrného normálního rozdělení f,

$$p(x|\Theta) = \sum_{i=1}^{M} \alpha_i f(x|\mu_i, \mathbb{C}_i),$$

kde M je počet komponent směsi. Pro odhad parametru Θ není využit EM algoritmus jako u MBC metody, místo něj spoléháme pouze na trénovací množinu dat $T = (T_1, \ldots, T_M)$. Pro všechna $i \in M$ odhadneme parametry μ_i a $\mathbb{C}_i$ stejně jako v metodě SMBC, viz. sekce 4.4.

$$\widehat{\mu_i} = \frac{1}{n_i} \sum_{k=1}^{n_i} x_{ik},$$

$$\widehat{\mathbb{C}_i} = \frac{1}{n_i} \sum_{k=1}^{n_i} (x_{ik} - \widehat{\mu_i})(x_{ik} - \widehat{\mu_i})^T,$$

kde n_i je počet pozorování v množině T_i. Parametry $(\alpha_1, \ldots, \alpha_M)$ nastavíme jako

$$\alpha_i = \frac{n_i}{\sum_{j=1}^{M} n_j}.$$

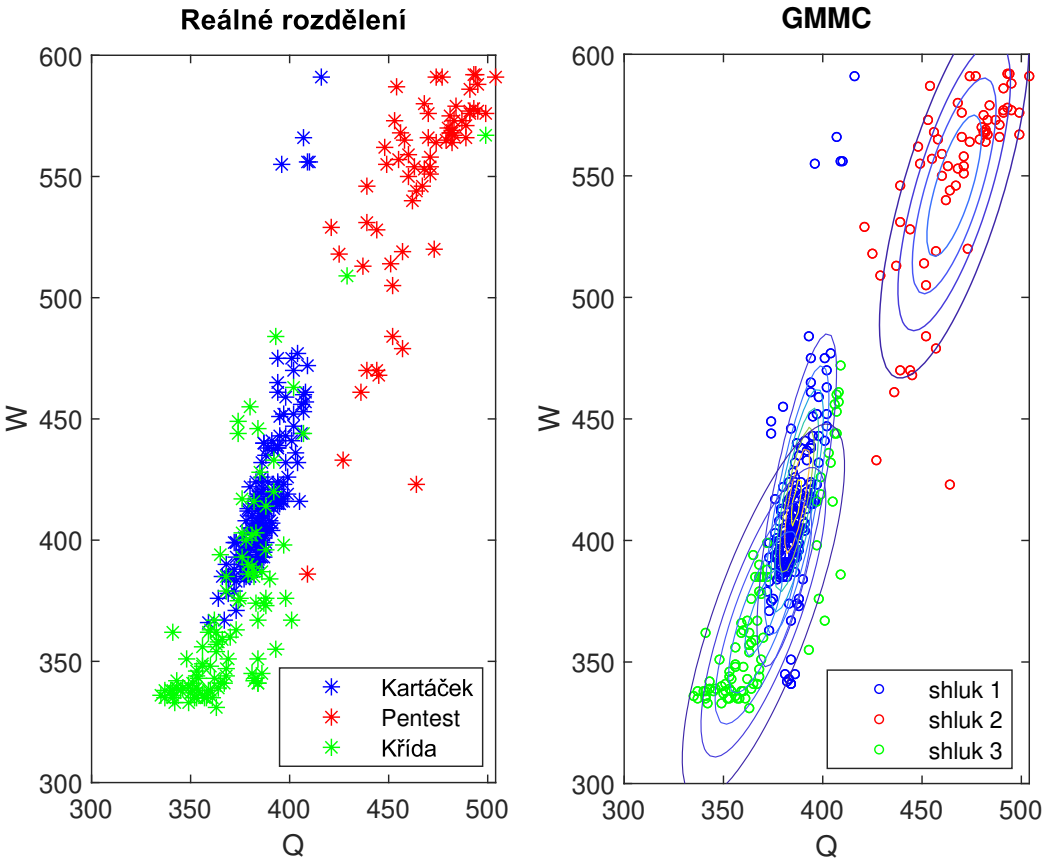
Tím máme definován vektor parametrů Θ, který plně charakterizuje distribuční směs $p(x|\Theta)$. Nyní spočteme pro pozorování x_k, které chceme klasifikovat, vektor $t_k = (t_{k1}, \ldots, t_{kM})$ posteriorních pravděpodobností jeho příslušnosti ke všem uvažovaným shlukům, tzn.

$$t_{ki} = \frac{\alpha_i f(x_k|\mu_i, \mathbb{C}_i)}{\sum_{j=1}^{M} \alpha_j f(x_k|\mu_j, \mathbb{C}_j)}.$$

Následně klasifikujeme podle pravidla, že x_k náleží k j-tému shluku právě tehdy, když

$$\arg\max_{i \in \widehat{M}} t_{ki} = j.$$

Příklad 4.3. Na obrázku 4.4 jsou znázorněny výsledky klasifikace již známé množiny dat metodou GMMC za použití parametrů W_2 a $Q_{0,33}$. V pravé části obrázku jsou znázorněny dílčí hustoty normálního rozdělení generované trénovacími daty T_i. Ani tato metoda není v tomto případě schopna spolehlivě oddělit signály kartáčku a křídy, které tvoří jeden kompaktní shluk, jak je patrno z blízkosti a podobné orientace rozdělení příslušných trénovacích dat.



Obrázek 4.4: Klasifikace metodou GMMC

Kapitola 5

Jádrové odhady

5.1 Jednorozměrné KDE

Jádrový odhad (kernel density estimate - KDE) je neparametrická metoda pro odhad hustoty pravděpodobnosti, pouze na základě naměřených dat, tedy bez nutnosti specifikovat parametrickou rodinu distribucí. To z něj činí velmi užitečný nástroj při odhadování hustot multimodálních, případně atypických rozdělení. Klasický jádrový odhad definujeme následovně.

Definice. Nechť $X_1, \ldots, X_n$ jsou nezávislé a stejně rozdělené náhodné veličiny s hustotou pravděpodobnosti $f : \mathbb{R} \to \mathbb{R}_0^+$. Jádrový odhad $\hat{f}$ hustoty f v bodě $t \in \mathbb{R}$ definujeme

$$\hat{f}(t) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{t - X_j}{h}\right),$$

kde $h > 0$ nazveme vyhlazovacím parametrem (bandwidth) a funkci $K : \mathbb{R} \to \mathbb{R}_0^+$ splňující podmínku $\int_{\mathbb{R}} K(t)\,\mathrm{d}t = 1$ nazveme jádrem.

Jádro K je tedy díky požadavkům nezápornosti a integrovatelnosti na jedničku samo o sobě hustotou pravděpodobnosti. Z definičního předpisu $\hat{f}$ plyne nezápornost a navíc

$$\int_{\mathbb{R}} \hat{f}(t)\,\mathrm{d}t = \frac{1}{nh} \sum_{j=1}^{n} \int_{\mathbb{R}} K\left(\frac{t - X_j}{h}\right)\,\mathrm{d}t = \frac{1}{nh} \sum_{j=1}^{n} h \cdot 1 = 1.$$

Máme tedy jistotu, že i odhad $\hat{f}$ je hustotou pravděpodobnosti. V tabulce (5.1) jsou uvedeny příklady běžně používaných jader.

Název	$K(t)$		eff(K)
Epanechnikovo jádro	$\frac{3}{4}(1-t^2)$ 0	pro $\lvert t\rvert \leq 1$ jinak	1
Normální jádro	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$	$t \in \mathbb{R}$	0,9512
Obdélníkové jádro	$\frac{1}{2}$ 0	pro $\lvert t\rvert \leq 1$ jinak	0,9295
Trojúhelníkové jádro	$1 - \lvert t\rvert$ 0	pro $\lvert t\rvert \leq 1$ jinak	0,9859

Tabulka 5.1: Příklady jednorozměrných jader včetně jejich eficience

Odchylku odhadu $\hat{f}$ od hustoty f v konkrétním bodě $t \in \mathbb{R}$ vyjadřujeme pomocí střední kvadratické chyby (MSE)

$$\mathrm{MSE}_t(\hat{f}) = E(\hat{f}(t) - f(t))^2, \tag{5.1}$$

kde střední hodnotu počítáme přes náhodné veličiny $X_1, \ldots, X_n$. Rovnici (5.1) lze upravit do tvaru

$$\mathrm{MSE}_t(\hat{f}) = \big(\underbrace{E\big(\hat{f}(t) - f(t)\big)}_{\mathrm{Bias}\hat{f}(t)} \big)^2 + \mathrm{Var}\hat{f}(t).$$

$\mathrm{Bias}\hat{f}(t)$ nazýváme vychýlením. Pro určení celkové chyby odhadu definujeme střední integrovanou kvadratickou odchylku (MISE) jako

$$\mathrm{MISE}\,(\hat{f}) = E \int_{\mathbb{R}} \Big(\hat{f}(t) - f(t)\Big)^2 \mathrm{d}t.$$

Díky nezápornosti integrandu lze MISE psát ve tvaru

$$\mathrm{MISE}\,(\hat{f}) = \int_{\mathbb{R}} \mathrm{MSE}_t(\hat{f})\mathrm{d}t = \int_{\mathbb{R}} \Big(E(\hat{f}(t) - f(t))\Big)^2 \mathrm{d}t + \int_{\mathbb{R}} \mathrm{Var}\hat{f}(t)\mathrm{d}t.$$

Pokud na jádro K naklademe navíc následující podmínky

$$\int_{\mathbb{R}} K(t)\mathrm{d}t = 1, \quad \int_{\mathbb{R}} tK(t)\mathrm{d}t = 0, \quad \int_{\mathbb{R}} t^2K(t)\mathrm{d}t = k_2 \neq 0, \tag{5.2}$$

pak podle [8] platí

$$\mathrm{Bias}\hat{f}(t) \approx \frac{1}{2}h^2k_2f''(t),$$
$$\mathrm{Var}\hat{f}(t) \approx \frac{1}{nh}f(t)) \int_{\mathbb{R}} K^2(t)\mathrm{d}t,$$

za předpokladu existence spojité derivace skutečné hustoty f do druhého řádu. Odhad MISE má potom tvar

$$\mathrm{MISE}(\hat{f}) \approx \frac{1}{4}h^4k_2^2 \int_{\mathbb{R}} \big(f''(t)\big)^2 \mathrm{d}t + \frac{1}{nh} \int_{\mathbb{R}} K^2(t)\mathrm{d}t. \tag{5.3}$$

Z této aproximace vyplývá, že odchylka jádrového odhadu $\hat{f}$ od skutečné hustoty f pro pevně danou množinu pozorovaných dat závisí pouze na volbě jádra K a velikosti vyhlazovacího parametru h. Volba optimální hodnoty h jež minimalizuje MISE má podle [7] tvar

$$h_{opt} = n^{-\frac{1}{5}}k_2^{-\frac{2}{5}} \left(\frac{\int_{\mathbb{R}} K(t)^2\mathrm{d}t}{\int_{\mathbb{R}} f''(t)^2\mathrm{d}t} \right)^2. \tag{5.4}$$

Po dosazení (5.4) do (5.3) dostáváme

$$\mathrm{MISE}(\hat{f}) \approx \frac{5}{4}n^{-\frac{4}{5}}C(K) \left(\int_{\mathbb{R}} f''(t)^2\mathrm{d}t \right)^{\frac{1}{5}} \quad \text{pro } C(K) = k_2^{\frac{2}{5}} \left(\int_{\mathbb{R}} K(t)^2\mathrm{d}t \right)^{\frac{4}{5}},$$

kde $C(K)$ závisí jen na volbě jádra K. Pro minimalizaci odchylky je tedy třeba vybrat jádro K splňující podmínky (5.2) s co nejnižší hodnotou $C(K)$. Optimální volbou je dle [9] tzv. *Epanechnikovo* jádro

$$K_e(t) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}t^2\right) & \text{pro } |t| \leq \sqrt{5}, \\ 0 & \text{jinak.} \end{cases}$$

V tabulce 5.1 je Epanechnikovo jádro uvedeno v přeškálovaném, používanějším tvaru. Pro porovnání jader zavádíme eficienci jádra K jako

$$\text{eff}(K) = \left(\frac{C(K_e)}{C(K)} \right)^{\frac{5}{4}}.$$

Jak je patrné z tabulky 5.1, hodnoty eficience jsou u běžně používaných jader blízké jedné. Větší vliv na kvalitu odhadu tak má volba vyhlazovacího parametru h. Optimální parametr h_{opt} však podle (5.4) závisí na odhadované hustotě f, kterou neznáme. Pro nastavení parametru proto užíváme následujících tzv. *Rule-of-thumb* vzorců (pro normální rozdělení).

- Po dosazení normální hustoty $f(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$ do výrazu (5.4) a použití Normálního jádra dostáváme optimální volbu vyhlazovacího parametru vhodnou pro data, jejichž rozdělení se podobá normálnímu, ve tvaru

  $$h = 1,06 \cdot \sigma n^{-\frac{1}{5}} \qquad \text{(Scott)},$$

  kde σ nahradíme výběrovou směrodatnou odchylkou s_n.

- Pro data pocházející z unimodálního rozdělení s vyšší šikmostí nebo špičatostí, než má Gaussovo rozdělení je vhodnější použít

  $$h = 0,79 \cdot IQR\, n^{-\frac{1}{5}} \qquad \text{(IQR)},$$

  kde $IQR = x_{0,75} - x_{0,25}$ je interkvartilové rozpětí. Tato volba h však poskytuje horší výsledky pro multimodální rozdělení.

- Jako kompromis mezi předchozími dvěma variantami je proto navržena volba umožňující odhadovat hustotu dat jak z multimodálního rozdělení, tak z rozdělení s vyšší šikmostí/špičatostí

  $$h = 0,9 \cdot \min \left\{ s_n, \frac{IQR}{1,34} \right\} n^{-\frac{1}{5}} \qquad (Silverman).$$

Otázkou asymptotických vlastností KDE pro počet pozorování $n \to \infty$ se zabývají následující dvě věty.

Věta 3. *Předpokládejme, že vyhlazovací parametr h_n závisí na počtu pozorování n a jádro K je omezená borelovská funkce splňující*

$$\int_{\mathbb{R}} |K(t)| \mathrm{d}t < \infty \qquad a \qquad |tK(t)| \to 0 \quad pro\ |t| \to \infty$$

Dále nechť

$$h_n \to 0 \wedge nh_n \to \infty \qquad pro\ n \to \infty$$

a nechť f je spojitá v bodě t. Potom dle [11] *je $\hat{f}$ konzistentním odhadem f v bodě t, tzn.*

$$\hat{f}(t) \xrightarrow{P} f(t) \qquad \forall t \in \mathbb{R}.$$

Věta 4. *Nechť K je omezená funkce s omezeným rozptylem, která je skoro všude vzhledem k Lebesgueově míře spojitá. Nechť dále h_n splňuje*

$$h_n \to 0 \wedge \frac{nh_n}{\ln n} \to \infty \qquad \textit{pro } n \to \infty$$

a f je stejnoměrně spojitá na $\mathbb{R}$. Potom dle [7]

$$\sup_t \left| \hat{f}(t) - f(t) \right| \xrightarrow{s.j.} 0 \qquad \textit{pro } n \to \infty,$$

tedy $\hat{f}$ je stejnoměrně konzistentním odhadem skutečné hustoty f.

5.2 Vícerozměrné KDE

Při klasifikaci signálů akustické emise budeme vždy používat alespoň dva atributy, zavedeme proto jádrový odhad pro vícerozměrná data. Předpokládejme, že $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ jsou i.i.d. náhodné veličiny s hustotou pravděpodobnosti $f : \mathbb{R}^d \to \mathbb{R}_0^+$.

Definice. Nechť $K : \mathbb{R}^d \to \mathbb{R}_0^+$ splňuje podmínku $\int_{\mathbb{R}^d} K(\boldsymbol{t})\mathrm{d}\boldsymbol{t} = 1$, a $h > 0$. Potom definujeme d-rozměrný jádrový odhad funkce f jako

$$\hat{f}(\boldsymbol{t}) = \frac{1}{nh^d} \sum_{j=1}^{n} K\left(\frac{\boldsymbol{t} - \boldsymbol{X}_j}{h} \right).$$

Funkci K a parametr h nazýváme stejně jako v jednorozměrném případě.

V definici vícerozměrného jádrového odhadu figuruje parametr h jako konstantní pro všechny dimenze. Pokud mají naměřená data v různých dimenzích výrazně jiné rozsahy je třeba buď použít vektor vyhlazovacích parametrů $\boldsymbol{h}$, nebo na datech provést lineární transformaci s následným vyhlazením za použití radiálně symetrického jádra - tzv. pre-whitening. Jádrový odhad s pre-whiteningem můžeme podle [10] zapsat jako

$$\hat{f}(\boldsymbol{t}) = \frac{1}{nh^d \sqrt{\det(\hat{\Sigma})}} \sum_{j=1}^{n} K\left(\left(\hat{\Sigma}^{-\frac{1}{2}} \right) \frac{\boldsymbol{t} - \boldsymbol{X}_j}{h} \right),$$

kde $\hat{\Sigma}$ je odhad kovarianční matice

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{j=1}^{n} (\boldsymbol{X}_j - \overline{\boldsymbol{X}}_n)(\boldsymbol{X}_j - \overline{\boldsymbol{X}}_n)^T, \qquad \overline{\boldsymbol{X}}_n = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{X}_j.$$

Stejně jako v jednorozměrném případě určujeme odchylku odhadu $\hat{f}$ od skutečné hustoty f pomocí MISE

$$\mathrm{MISE}\,(\hat{f}) = \int_{\mathbb{R}^d} \left(\mathrm{Bias}\hat{f}(\boldsymbol{t}) \right)^2 \mathrm{d}\boldsymbol{t} + \int_{\mathbb{R}^d} \mathrm{Var}\hat{f}(\boldsymbol{t})\mathrm{d}\boldsymbol{t}. \tag{5.5}$$

Za předpokladu, že K je radiálně symetrická funkce a f má omezené a spojité druhé derivace můžeme odhadnout

$$\mathrm{Bias}(\hat{f}(\boldsymbol{t})) \approx \frac{1}{2} h^2 \alpha \nabla^2 f(\boldsymbol{t}), \text{ kde} \qquad \alpha = \int_{\mathbb{R}^d} t_1^2 K(\boldsymbol{t})\mathrm{d}\boldsymbol{t},$$

$$\mathrm{Var}\hat{f}(\boldsymbol{t}) \approx \frac{1}{nh^d} \beta f(\boldsymbol{t}), \text{ kde} \qquad \beta = \int_{\mathbb{R}^d} \left(K(\boldsymbol{t}) \right)^2 \mathrm{d}\boldsymbol{t}.$$

Dosazením do (5.5) získám odhad MISE

$$\mathrm{MISE}\,(\hat{f}) \approx \frac{1}{4}h^4\alpha^2 \int_{\mathbb{R}^d} \left(\nabla^2 f(\boldsymbol{t})\right)^2 \mathrm{d}\boldsymbol{t} + \frac{\beta}{nh^d}.$$

Optimální hodnotou vyhlazovacího parametru h je potom

$$h_{opt} = \left(\frac{\beta d}{n\alpha^2} \frac{1}{\int_{\mathbb{R}^d} (\nabla^2 f(\boldsymbol{t}))^2 \mathrm{d}\boldsymbol{t}}\right)^{\frac{1}{d+4}} = A(K, f) n^{-\frac{1}{d+4}},$$

kde A závisí pouze na volbě jádra a odhadované hustotě f. Jádro minimalizující MISE opět nazýváme Epanechnikovo a jeho vícerozměrná varianta má tvar

$$K_e(\boldsymbol{t}) = \begin{cases} \frac{d+2}{2V_d}(1 - \boldsymbol{t}^T\boldsymbol{t}) & \text{pro } \boldsymbol{t}^T\boldsymbol{t} < 1, \\ 0 & \text{jinak,} \end{cases}$$

kde V_d značí objem d-rozměrné koule o poloměru jedna. V tabulce 5.2 jsou uvedeny příklady vícedimenzionálních jader a příslušných funkcí $A(K, f)$, kde za f byla dosazena hustota d-rozměrného normálního rozdělení.

Název	$K(t)$		Dimenze	A(K)
Epanechnikovo jádro	$\frac{d+2}{2V_d}(1 - \boldsymbol{t}^T\boldsymbol{t})$ 0	pro $\boldsymbol{t}^T\boldsymbol{t} < 1$ jinak	d	$\left(\frac{8(d+4)}{V_d}(2\sqrt{\pi})^d\right)^{\frac{1}{d+4}}$
Normální jádro	$\frac{1}{\sqrt{2\pi}^d} e^{-\frac{\boldsymbol{t}^T\boldsymbol{t}}{2}}$	$\boldsymbol{t} \in \mathbb{R}^d$	d	$\left(\frac{4}{d+2}\right)^{\frac{1}{d+4}}$
K_2	$\frac{3}{\pi}(1 - \boldsymbol{t}^T\boldsymbol{t})^2$ 0	pro $\boldsymbol{t}^T\boldsymbol{t} < 1$ jinak	2	2,78

Tabulka 5.2: Příklady vícerozměrných jader včetně funkcí $A(K, f)$ pro normální rozdělení

5.3 Klasifikace pomocí KDE

Klasifikační metodu využívající jádrových odhadů nazýváme SKDEC (Supervised Kernel Density Estimation Clustering). Tato metoda využívá stejného přístupu k datům jako metoda GMMC, viz. sekce 4.5. Mějme tedy opět trénovací množinu $T = (T_1, \ldots, T_M) \subset \mathbb{R}^d$, kde $x_{ik} \in T_i$ právě tehdy, když pozorování x_{ik} náleží i-tému shluku a M je počet shluků, které budeme hledat. Mějme dále množinu testovacích dat $\boldsymbol{x} = (x_1, \ldots, x_n)$, $x_i \in \mathbb{R}^d$ pocházejících ze stejného měření jako trénovací data. Tedy hustota pravděpodobnosti, ze které pocházejí naměřená data je stejná jak pro trénovací množinu dat, tak pro data, která chceme klasifikovat. Analogicky k distribučním směsím odhadneme celkovou hustotu všech dat jako vážený součet M hustot dílčích skupin signálů

$$\hat{f}_{celk}(t) = \sum_{j=1}^{M} \alpha_i \hat{f}_i(t),$$

kde $\hat{f}_i$ je jádrový odhad hustoty dat z trénovací množiny T_i a α_i je váhový koeficient zohledňující velikost množiny T_i, tzn.

$$\hat{f}_i(t) = \frac{1}{n_i h_i} \sum_{k=1}^{n_i} K\left(\frac{t - x_{ik}}{h_i}\right),$$

$$\alpha_i = \frac{n_i}{\sum_{j=1}^{M} n_j},$$

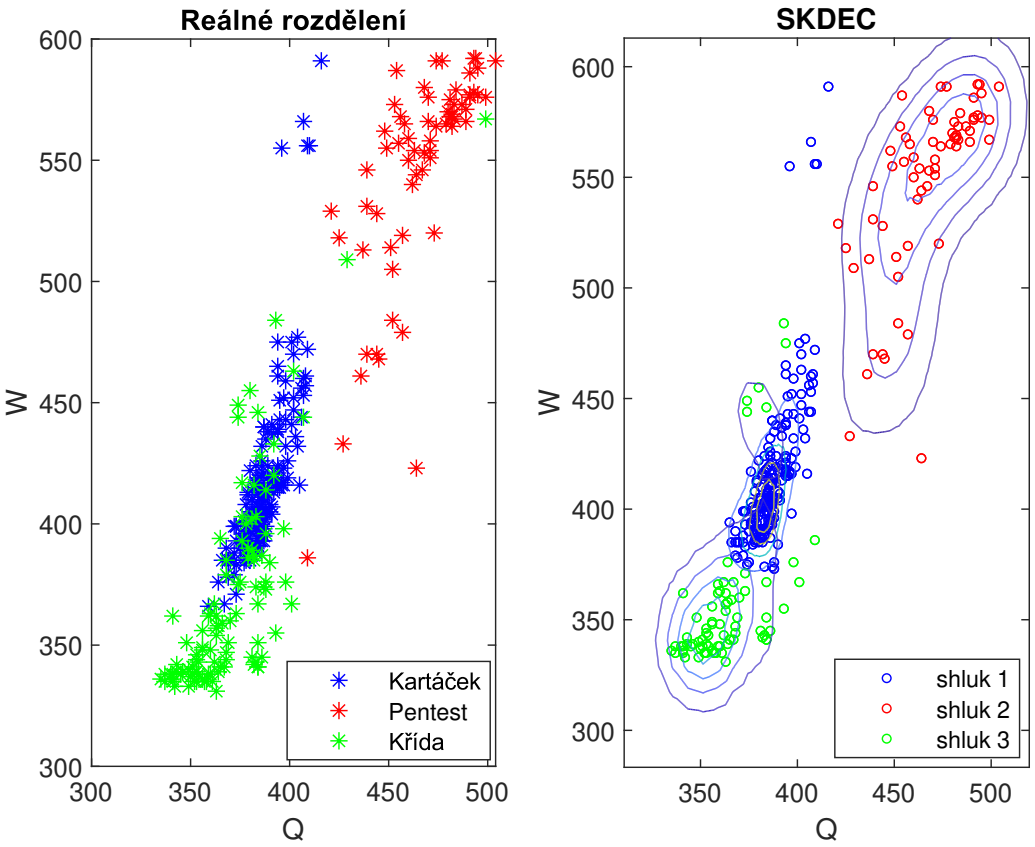
n_i zde opět značí počet pozorování v i-té trénovací množině T_i. Následně pro všechna pozorování x_k, která chceme klasifikovat, spočteme vektor $t_k = (t_{k1}, \ldots, t_{kM})$ pravděpodobností jeho příslušnosti ke všem uvažovaným shlukům

$$t_{ki} = \frac{\alpha_i \hat{f}_i(x_k)}{\hat{f}_{celk}(x_k)}.$$

Klasifikujeme podle pravidla, že x_k náleží do j-tého shluku právě tehdy, když

$$\arg\max_{i \in \widehat{M}} t_{ki} = j.$$

Příklad 5.1. Na obrázku 5.3 je znázorněna klasifikace stejného souboru dat jako v předchozích kapitolách v proměnných W_2 a $Q_{0,33}$ metodou SKDEC. Díky schopnosti jádrového odhadu vytvářet neeliptické a nesymetrické odhady hustot dokáže tato metoda lépe popsat reálné rozložení trénovacích dat a tím výrazným způsobem zvýšit přesnost klasifikace v případech, kdy se pozorování, pocházející z různých zdrojů na výběrovém prostoru, překrývají. Metoda SKDEC dosáhla na tomto datovém souboru nejlepších výsledků s úspěšností klasifikace přesahující 88%.



Obrázek 5.3: Klasifikace metodou SKDEC

Kapitola 6

Divergenční rozhodovací strom s učitelem

Divergenční rozhodovací strom s učitelem (Supervised Divergence Decision Tree - SDDT) je binární klasifikátor vyvinutý v [12] a dále upravovaný v [8]. Jedná se o binární rozhodovací strom, který využívá divergenční míry a jádrové odhady. Byl navržen pro klasifikaci dat pocházejících z experimentu DØ prováděného v laboratořích Fermilab.

Zabýváme se klasifikací dat, která mají různá rozdělení. V následující části textu proto zavedeme několik způsobů jak měřit vzdálenost mezi pravděpodobnostními rozděleními. To nám při klasifikaci umožní vybrat vhodnou kombinaci atributů, pro kterou bude daná vzdálenost co nejvyšší a bude tedy snazší oddělit dvě skupiny pozorování.

6.1 Vzdálenosti pravděpodobnostních rozdělení a ϕ-divergence

Uvažujme v této sekci měřitelný prostor statistických pozorování $(\mathcal{X}, \mathcal{A})$, kde $\mathcal{X} \subset \mathbb{R}^n$ a $\mathcal{A}$ je σ-algebrou nad $\mathcal{X}$. $\mathcal{P}$ bude značit třídu pravděpodobnostních měr na $(\mathcal{X}, \mathcal{A})$.

Definice. Nechť $P, Q \in \mathcal{P}$ a k nim příslušné distribuční funkce F, G. Potom definujeme *Kolmogorovu vzdálenost* mezi distribucemi P a Q

$$\rho_K(P, Q) = \sup_{x \in \mathcal{X}} |F(x) - G(x)| .$$

Dále pro f, g hustoty rozdělení P, Q definujeme *totální variaci*

$$\rho_{TV}(P, Q) = \int_{\mathcal{X}} |f(x) - g(x)| \mathrm{d}x.$$

Pro $\mathcal{X} = \mathbb{R}$ zavádíme *Lévyho vzdálenost*

$$\rho_L(P, Q) = \inf \{\epsilon > 0 : F(x - \epsilon) - \epsilon \leq G(x) \leq F(x + \epsilon) + \epsilon, \quad \forall x \in \mathbb{R}\} .$$

Věta 5. *Pro Kolmogorovu, Lévyho vzdálenost a totální variaci na $\mathcal{X} = \mathbb{R}$ platí*

$$\rho_L \leq \rho_K \leq \frac{\rho_{TV}}{2}.$$

Dále zavádíme ϕ-divergenční míry. Jedná se o třídu vzdáleností, které lze jednotně zapsat pomocí generující funkce ϕ a které se pro speciální volbu ϕ mohou stát metrikami.

Definice. Nechť $P, Q \in \mathcal{P}$ a f, g jsou k nim příslušné hustoty pravděpodobnosti vzhledem k Lebesgueově míře $(\mathrm{d}x)$. Nechť Φ^* je třída konvexních funkcí $\phi : (0, \infty) \mapsto \mathbb{R}$, které splňují $\phi(1) = 0$. Pak pro $\phi \in \Phi^*$ definujeme $\phi - divergenci$ pravděpodobnostních rozdělení P, Q

$$D_\phi(P, Q) = \int_{\mathcal{X}} g(x)\phi\left(\frac{f(x)}{g(x)}\right) \mathrm{d}x,$$

kde v integrálu přijímáme konvence

$$0\,\phi\left(\frac{0}{0}\right) = 0, \qquad 0\,\phi\left(\frac{p}{0}\right) = \lim_{t\to\infty} \frac{\phi(t)}{t}.$$

Pokud vezmeme $\phi \in \Phi^*$, která je diferencovatelná v $t = 1$ potom funkce ψ

$$\psi(t) = \phi(t) - \phi'(1)(t - 1)$$

také leží v Φ^* a navíc $\psi'(1) = 0$. Díky konvexnosti ϕ lze odvodit [13]

$$D_\phi(P, Q) = D_\psi(P, Q).$$

Třídu Φ^* díky tomu lze ekvivalentně nahradit množinou

$$\Phi = \Phi^* \cap \left\{\phi \mid \phi'(1) = 0\right\}$$

a dále uvažovat pouze zúženou množinu generujících funkcí. Je snadné nahlédnout, že ϕ-divergence obecně nesplňuje axiomy metriky, jelikož ji však budeme využívat k určování vzdálenosti, bylo by vhodné najít podmínky, za kterých alespoň nějaké vlastnosti metriky splňovat bude.

Věta 6.(reflexivita). *Nechť $\phi \in \Phi$ je striktně konvexní v bodě $t = 1$. Potom*

$$D_\phi(P, Q) = 0 \iff P = Q.$$

Definice. Pro $\phi \in \Phi$ definujeme tzv. konjugovanou funkci

$$\phi^*(t) = t\,\phi\left(\frac{1}{t}\right) \qquad \text{pro } t \in (0, \infty).$$

Nechť $\phi \in \Phi$ a ϕ^* je její konjugovaná funkce, potom platí

$$\phi^* \in \Phi, \qquad (\phi^*)^* = \phi, \qquad \phi^*(0) = \lim_{t\to\infty} \frac{\phi(t)}{t}.$$

Věta 7. *(symetrie). Nechť $\phi \in \Phi$, potom pro kterákoliv $P, Q \in \mathcal{P}$ platí*

$$D_{\phi^*}(P, Q) = D_\phi(Q, P). \tag{6.1}$$

Důkazy obou vět jsou provedeny v [13]. Pokud tedy chceme sestrojit reflexivní i symetrickou divergenci volíme funkci $\phi \in \Phi$ s $\phi(1) = 0$, která je striktně konvexní v $t = 1$ a sestrojíme $(\phi + \phi^*)$-divergenci. Ta je podle (6.1) symetrická, neboť je sama sobě konjugovanou funkcí.

V tabulce 6.1 jsou uvedeny některé ϕ-divergence včetně příslušných generujících funkcí. Generující funkce Power divergence $\phi_\alpha(t) = \frac{t^\alpha - \alpha t + \alpha - 1}{\alpha(\alpha - 1)}$ je striktně konvexní $\forall t > 0$ a platí $\phi(1) = 0$, jsou tak splněny předpoklady věty 6 a Power divergence je tedy reflexivní. Zároveň je patrné, že $\phi^*_\alpha = \phi_{1-\alpha}$ a z věty 7 tak vyplývá, že Power divergence je symetrická pouze pro $\alpha = \frac{1}{2}$, označme ji $I_{\frac{1}{2}}(P, Q)$. Lze navíc ukázat, že $\sqrt{I_{\frac{1}{2}}(P, Q)}$ splňuje trojúhelníkovou nerovnost, jedná se tedy o metriku na $\mathcal{P}$.

Definice. Po znormování definujeme *Hellingerovu vzdálenost*

$$H(P,Q) = \sqrt{\frac{1}{2} I_{\frac{1}{2}}(P,Q)} = \left(\int_{\mathcal{X}} \left(\sqrt{f(x)} - \sqrt{g(x)} \right)^2 \mathrm{d}x \right)^{\frac{1}{2}},$$

která je metrikou na $\mathcal{P}$.

$\phi(t)$	Název ϕ-divergence
$t \ln t$	Kullback-Leibler
$(t-1)^2$	Pearsonova χ^2
$\lvert 1-t \rvert^\alpha, \quad \alpha \geq 1$	χ-divergence řádu α
$\frac{t^\alpha - \alpha t + \alpha - 1}{\alpha(\alpha-1)}, \quad \alpha \neq 0, \alpha \neq 1$	Power divergence řádu α

Tabulka 6.1: Vybrané generující funkce ϕ-divergencí

Integrál vyskytující se ve výpočtu Power divergence nazveme *Hellingerův integrál řádu* α, značíme

$$H_\alpha(P,Q) = \int_{\mathcal{X}} f^\alpha(x)\, g^{1-\alpha}(x)\, \mathrm{d}x.$$

Definice. Definujeme normovanou *Rényiho vzdálenost* řádu α

$$R_\alpha(P,Q) = \frac{\ln H_\alpha(P,Q)}{\alpha - 1} = \frac{1}{\alpha - 1} \ln \int_{\mathcal{X}} f^\alpha(x)\, g^{1-\alpha}(x)\, \mathrm{d}x, \tag{6.2}$$

pro $\alpha \neq 0, \alpha \neq 1$.

6.2 Binární klasifikace

V této části shrneme základní poznatky o binární klasifikaci a kritériích její kvality. Uvažujme $\boldsymbol{X} \in \mathbb{R}^d$ náhodnou veličinu, označme x její realizaci. Označme w_1, w_2 dvě disjunktní třídy pozorování, které splňují $P_1 + P_2 = 1$, kde $P_i = P(w_i) > 0$ je apriorní pravděpodobnost třídy w_i. Hustotu pravděpodobnosti náhodné veličiny $\boldsymbol{X}$ můžeme psát ve tvaru

$$p(x) = P_1\, p(x|w_1) + P_2\, p(x|w_2),$$

kde $p(x|w_i)$ značí podmíněnou hustotu pravděpodobnosti příslušnosti x k třídě w_i. S použitím Bayesova vzorce vyjádříme pravděpodobnost třídy w_i za předpokladu pozorování x

$$P(w_i|x) = \frac{P_i\, p(x|w_i)}{p(x)}. \tag{6.3}$$

Pozorování x přiřadíme do $w_k \Leftrightarrow P(w_k|x) > P(w_i|x)$ pro každé $i \neq k$. Tuto podmínku můžeme za pomoci (6.3) pro $k = 1$ přepsat do tvaru

$$\frac{P_1\, p(x|w_1)}{p(x)} > \frac{P_2\, p(x|w_2)}{p(x)},$$

odkud díky předpokladu $p(x) > 0$ dostáváme

$$l(x) = \frac{p(x|w_1)}{p(x|w_2)} > \frac{P_2}{P_1} = \delta, \tag{6.4}$$

kde funkci $l : \mathbb{R}^d \to \mathbb{R}_0^+ \cup \{\infty\}$ nazveme *věrohodnostní poměr* a parametr $\delta \in \mathbb{R}_0^+$ *dělící bod*. Pozorování x tedy přiřadíme třídě w_1 pokud platí $l(x) > \delta$, v opačném případě ho přiřadíme třídě w_2. Při klasifikaci pozorování x tedy může dojít právě k následujícím případům

- $x \in w_1 \quad \wedge \quad x$ klasifikujeme jako w_1 — True Positive
- $x \in w_1 \quad \wedge \quad x$ klasifikujeme jako w_2 — False Positive
- $x \in w_2 \quad \wedge \quad x$ klasifikujeme jako w_1 — False Negative
- $x \in w_2 \quad \wedge \quad x$ klasifikujeme jako w_2 — True Negative.

Pokud klasifikujeme množinu pozorování $\boldsymbol{x} = (x_1, \ldots, x_n)$ a pro všechny čtyři možnosti výsledku sečteme pozorování, pro která daný výsledek nastal, dostáváme postupně hodnoty TP, FP, FN, TN charakterizující úspěšnost klasifikace. Za pomoci těchto hodnot definujeme několik kritérií kvality klasifikace

$$\begin{aligned}
ACC &= \frac{TP+TN}{TP+FN+FP+TN} && \text{accuracy,}\\
ERR &= \frac{FP+FN}{TP+FN+FP+TN} && \text{globální chyba,}\\
\varepsilon &= \frac{TP}{TP+FN} && \text{sensitivita (true positive rate),}\\
\rho &= \frac{TN}{TN+FP} && \text{specificita,}\\
FPR &= \frac{FP}{TP+TN} = 1-\rho && \text{false positive rate.}
\end{aligned}$$

Všechny tyto indikátory kvality klasifikace jsou závislé na hodnotě dělícího bodu δ (6.4). Ten pro danou klasifikační úlohu volíme jako argument maxima vhodně vybraného kritéria kvality (figure of merit - FOM)

$$\delta^* = \underset{\delta \in \mathbb{R}_0^+}{\arg\max}(FOM).$$

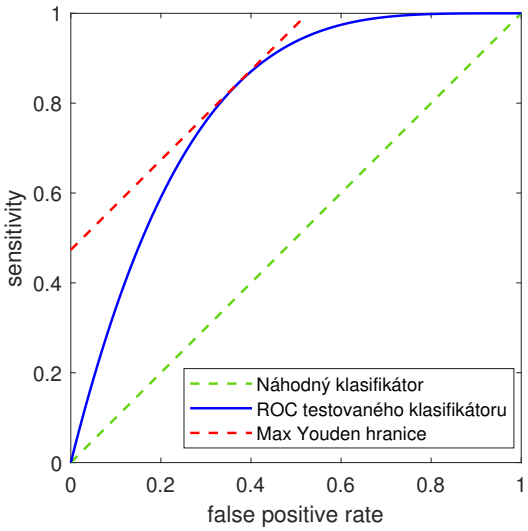
Další používanou charakteristikou kvality kklasifikace je *ROC křivka* (Receiving Operating Characteristic Curve), ta umožňuje grafické znázornění rozhodovacího pravidla. Konstruuje se napočítáním hodnot false positive rate a sensitivity pro všechny hodnoty dělícího parametru $\delta \in \mathbb{R}_0^+$ a jejich následným vynesením na osu x, resp. y. ROC křivku je tedy možné definovat jako množinu bodů

$$\left\{ [FPR(\delta), \varepsilon(\delta)] \,|\delta \in \mathbb{R}_0^+ \right\}.$$

Na obrázku 6.2 je zobrazen příklad takovéto křivky. Veličina AUC (*Area Under Curve*) značí obsah plochy pod ROC křivkou a redukuje tak kvalitu ROC křivky na jednu skalární veličinu. Pro hledání optimálního dělícího bodu δ^* zavedeme tzv. *Youden index*

$$Youden\ index = \varepsilon + \rho - 1 = \varepsilon(\delta) - FPR(\delta), \tag{6.5}$$

a následně δ^* volíme tak, aby byl tento index maximální. Získáváme tak dělící bod, pro který je signál kvalitně klasifikován a zároveň je pravděpodobnost chybné klasifikace pozadí držena na relativně nízké úrovni.



Obrázek 6.2: Ukázka ROC křivky

6.3 Princip SDDT

Popíšeme nyní podobu a algoritmus Divergenčního rozhodovacího stromu s učitelem. SDDT je speciální případ binárního rozhodovacího stromu, čemuž odpovídá jeho základní struktura. Jedná se o orientovaný graf, tvořený z uzlů a jejich propojení. Uzly jsou děleny na kořen, vnitřní uzly a listy. Kořen nemá žádného rodiče, na rozdíl od ostatních uzlů, které mají vždy právě jednoho. Každý uzel má buď žádného, nebo dva potomky. Uzly v nejnižší vrstvě, které žádného potomka nemají nazýváme listy. Jak už bylo zmíněno, SDDT je binární klasifikátor, předpokládejme tedy že chceme pozorování rozdělit do dvou tříd: signál - w_S a pozadí -w_B. Do kořene vstupuje množina pozorování $\boldsymbol{x} = (x_1, \ldots, x_n), x_i \in \mathbb{R}^d$, ke každému $x_i \in \boldsymbol{x}$ je přiřazena jedna třída podle specifikovaného dělícího kritéria. Množina $\boldsymbol{x}$ se tak rozdělí na dvě disjunktní podmnožiny $\boldsymbol{x}_S$ s vyšším zastoupením signálu a $\boldsymbol{x}_B$ s vyšším zastoupením pozadí. Následně $\boldsymbol{x}_S$ vstupuje do pravého potomka a $\boldsymbol{x}_B$ do levého potomka kořene. V každém uzlu dále probíhá stejným způsobem dělení množin do nich vstupujících, a to až dokud se všechna pozorování nedostanou do listů stromu. Tvar stromu, jeho hloubka a dělící parametry v každém uzlu se nastavují ve fázi učení podle trénovací množiny pozorování $T = (T_S, T_B)$, kde $T_S \subset w_s$ a $T_B \subset w_B$.

Nyní detailněji popíšeme průběh učení a rekurzivní výstavby stromu, začneme vložením trénovací množiny T do kořene stromu.

1. Zkontrolujeme, zda aktuální uzel nesplňuje podmínku pro zastavení redukce. Pokud pro pevně nastavené $\gamma > 0$ platí

$$\sum_{x_i \in T} \chi_{w_S}(x_i) < \gamma \qquad \vee \qquad \sum_{x_i \in T} \chi_{w_B}(x_i) < \gamma,$$

   kde χ_{w_S}, resp. χ_{w_B}, je charakteristická funkce w_S, resp. w_B, prohlásíme uzel za čistý a data již dále nedělíme. V opačném případě pokračujeme dalším krokem.

2. Nechť $k < d$ je pevně dané přirozené číslo určující kolik atributů použijeme k dělení dat. Pro libovolnou k-tici atributů spočteme za použití *kvantilových histogramů*, jejichž

přesný popis lze nalézt v [12], odhad hustoty pravděpodobnosti zvlášť pro T_S, označíme $\hat{f}_S^{hist}$, zvlášť pro T_B, označíme $\hat{f}_B^{hist}$. Pro tyto odhady vyčíslíme hodnotu jejich Rényiho divergence řádu $\alpha = \frac{1}{2}$, $R_{\frac{1}{2}}(\hat{f}_S^{hist}, \hat{f}_B^{hist})$ dle (6.2).

3. Krok 2. zopakujeme pro všechny možné kombinace k atributů. Dostáváme tak $\binom{d}{k}$ hodnot Rényiho divergence $R_{\frac{1}{2}}^{(j)}$.

4. Najdeme j^*-tou kombinaci k proměnných pro kterou Rényiho divergence nabyla nejvyšší hodnoty
$$j^* = \arg\max_{j \in \widehat{\binom{d}{k}}} R_{\frac{1}{2}}^{(j)}.$$

5. Vybraných k proměnných z j^*-té kombinace využijeme při klasifikaci dat z T za pomoci jádrového odhadu. Množinu T tak rozdělíme na T_S^e, obsahující data klasifikovaná jako signál a T_B^e, obsahující data klasifikovaná jako pozadí. Množinu T_S^e vložím do pravého potomka aktuálního uzlu, množinu T_B^e do levého potomka a v obou se vracíme ke kroku 1. tohoto algoritmu.

Parametr γ z prvního kroku jsme pro naši klasifikaci nastavili na hodnotu $\gamma = 7$. V pátém kroku získáváme pro i-té pozorování t_i z trénovací množiny T a pro každý uzel, kterým toto pozorování prochází, číslo $p_i^l \in (0, 1)$, $l \in \widehat{l_i}$, kde l_i je počet uzlů, které t_i navštíví. Číslo p_i^l vyjadřuje pravděpodobnost, že dané pozorování patří v l_i-tém navštíveném uzlu do třídy w_S. Hodnota p_i^l odpovídá věrohodnostnímu poměru l z (6.4) vyčíslenému v bodě t_i. Dělící bod δ potom volíme tak, aby byl maximalizován Youden index (6.5). Tím máme díky trénovacím datům nastavenou jak strukturu a tvar stromu, tak vhodnou k-tici parametrů a optimální dělící bod δ^* v každém uzlu.

Do kořene připraveného stromu vkládáme testovací data $\boldsymbol{x}$ a pro každé pozorování $x_i \in \boldsymbol{x}$ spočítáme pomocí nastavených parametrů hodnoty p_i^l. Výstupem průchodu pozorování x_i stromem je tzv. *diskriminant pravděpodobností* $D_i = (p_i^1, \ldots, p_i^{l_i})$. Pomocí D_i vyjádříme celkové skóre příslušnosti pozorování x_i ke třídě w_S jako

$$s_i = \frac{1}{l_i} \sum_{j=1}^{l_i} p_i^j.$$

Hodnota s_i opět odpovídá funkci l (6.4) vyčíslené v bodě x_i, dělící bod volíme pomocí trénovací množiny dat tak, aby maximalizoval Youden index.

Při klasifikaci signálů akustické emise často potřebujeme rozdělit signály do M shluků, kde $M > 2$. V takovém případě použijeme SDDT celkem M-krát, kdy při každém běhu volíme jako signál jinou skupinu signálů a ostatní považujeme za pozadí. Pozorování x_i následně přiřadíme do shluku, pro který bylo skóre příslušnosti k w_S nejvyšší.

Z důvodu nízkého množství trénovacích dat jsme byli nuceni si vygenerovat dostatečné množství dat pro natrénování klasifikačního stromu, což je v aplikacích strojového učení běžná praxe. Předpokládali jsme, že data z každé trénovací množiny T_i mají normální rozdělení. Odhadli jsme jeho hustotu a umělá trénovací data jsme generovali z tohoto rozdělení. Pro každou skupinu signálů jsme generovali 1000 pozorování.

Kapitola 7

Porovnání použitých klasifikačních metod

V této kapitole aplikujeme výše popsané klasifikační metody na naměřená data. Pro atributy závislé na svém vnitřním parametru volíme pro celou tuto kapitolu

$$Z = Z_{\frac{1}{20}},\ W = W_2,\ Q = Q_{0,33},\ S = S_{\frac{1}{3}}.$$

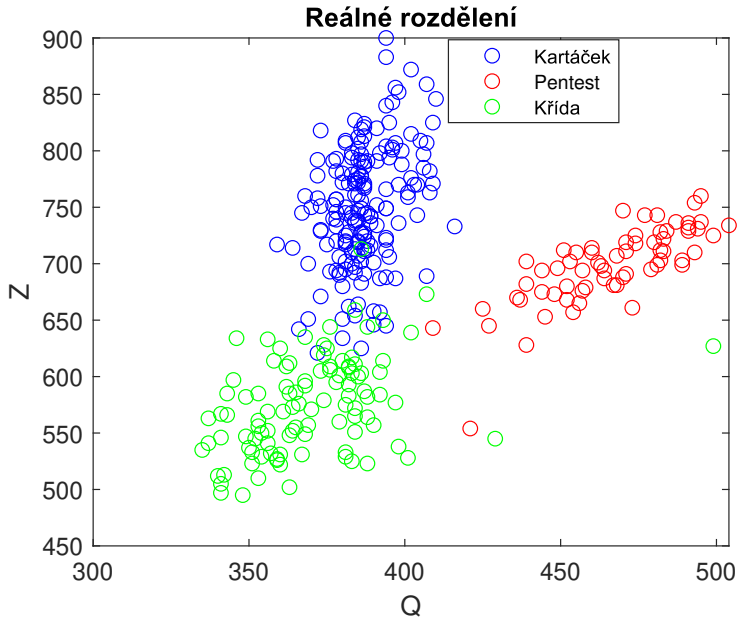
Klasifikace tří různých zdrojů AE

Nejprve budeme klasifikovat tři druhy signálů naměřené na plechu, které byly buzeny pomocí kartáčku, pentestu a křídy. Pro klasifikaci jsme použili několik dvojic parametrů, ty jsou včetně výsledků uvedeny v tabulce 7.1.

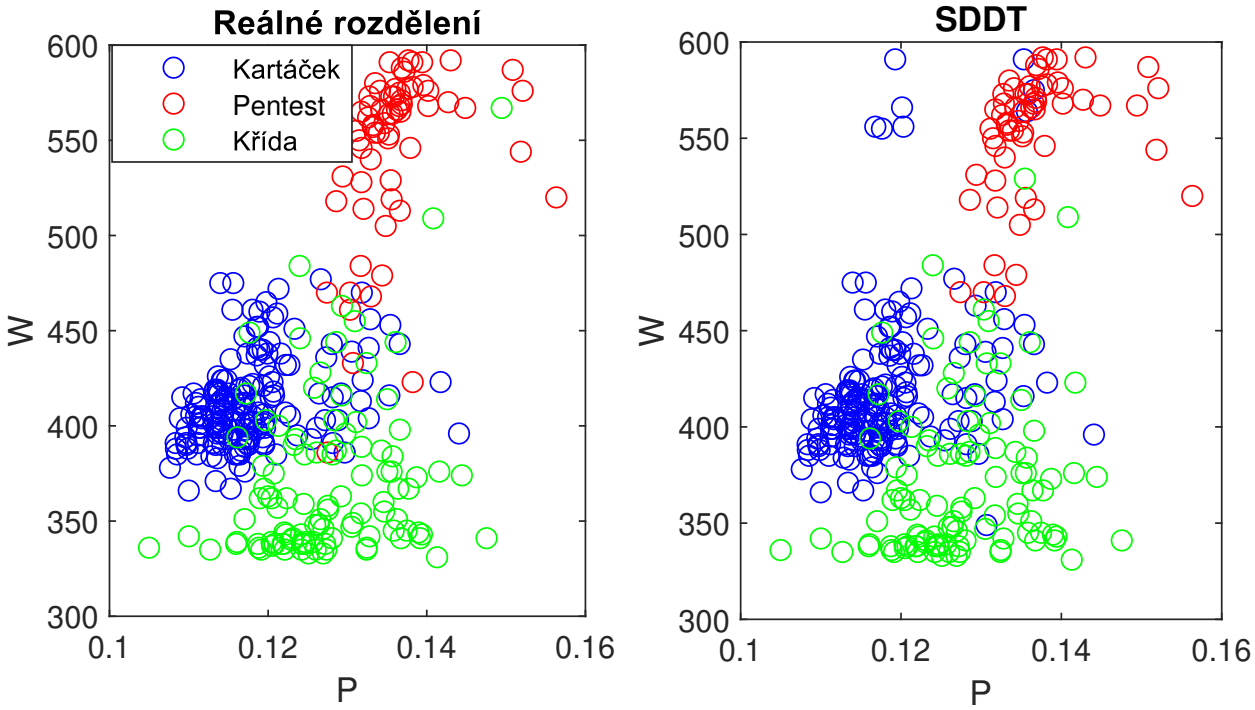
Atributy	MBC	SMBC	GMMC	SKDEC
P-Q	88,0	88,0	90,0	92,1
Q-Z	**96,5**	**96,5**	96,2	**96,5**
Q-W	67,7	70,4	80,4	88,6
W-M	87,7	91,5	93,0	95,3
S-P	70,1	73,6	79,2	83,6
Z-P	89,4	88,9	94,1	93,5

Tabulka 7.1: Úspěšnost klasifikace tří zdrojů akustické emise uvedená v % správně klasifikovaných signálů

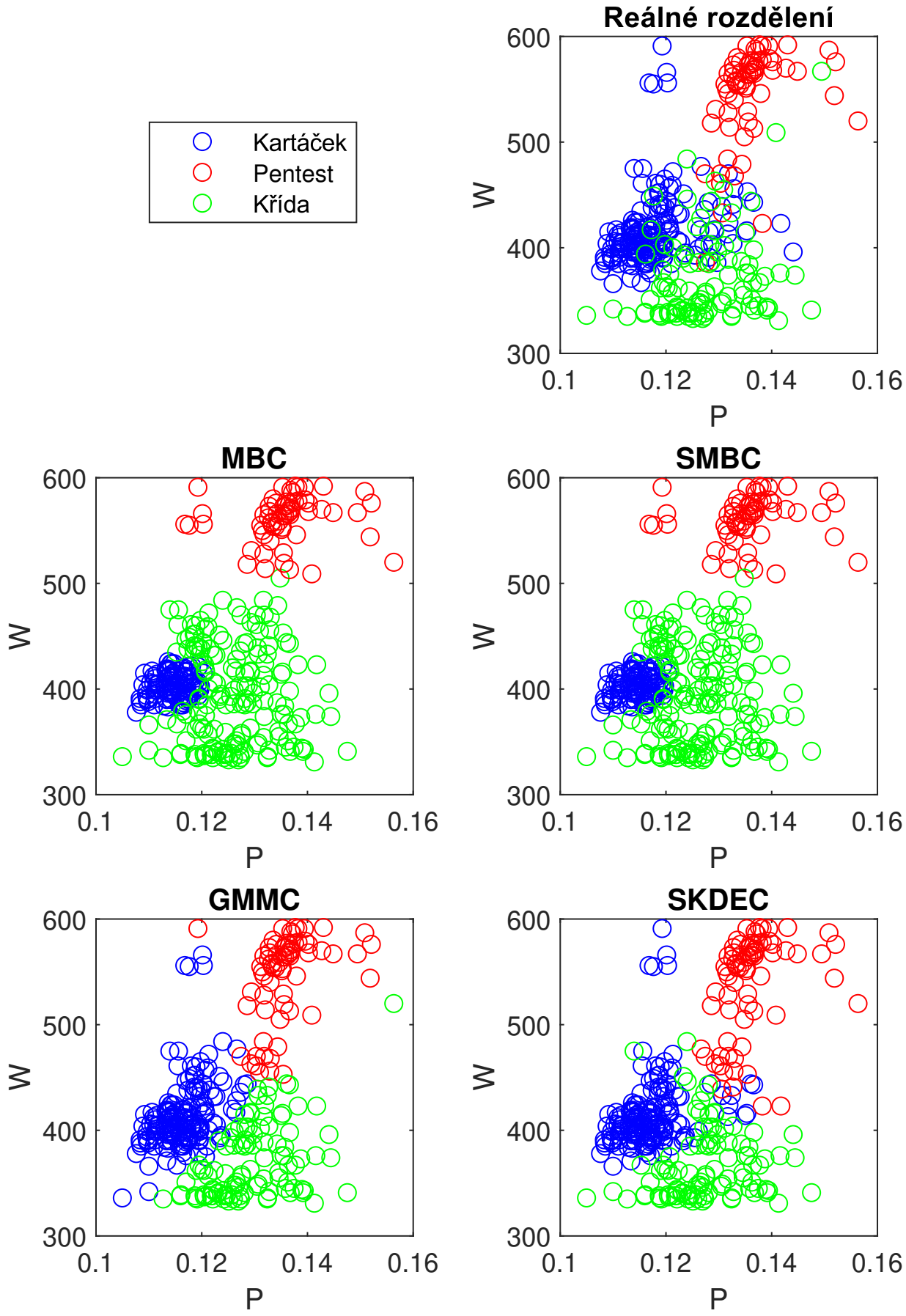
Z výsledků je patrné, že úspěšnost klasifikace závisí především na volbě atributů. Při volbě Q-Z se výsledky použitých metod významně neliší, to je dáno tím, že jsou pozorování ve výběrovém prostoru jasně oddělena viz obrázek 7.2. Pro ostatní volby atributů, kdy jsou shluky komplikovanější a překrývají se, je patrné výrazné zlepšení u metod s učitelem, zejména pro metody GMMC a SKDEC. Metoda SDDT není zahrnuta v tabulce, protože na rozdíl od ostatních pracuje vždy se všemi atributy najednou, má tak oproti ostatním klasifikačním metodám k dispozici více informací. Metoda SDDT dosahuje na těchto datech úspěšnosti 97,3 %, její výsledek je zobrazen na obrázku 7.3, kde jsou pozorování jen pro porovnání s ostatními metodami vyneseny do proměnných P-W. Výsledky klasifikace ostatními metodami v těchtýž proměnných jsou na obrázku 7.4. Je tu dobře patrná drobná nevýhoda EM algoritmu, využívaného v metodách MBC a SMBC, kdy se odhadovaná hustota příliš zúžila na koncentrovaný střed modrého shluku a přišla tak o pozorování na řidčeji obsazených krajích. K tomu u metod GMMC a SKDEC nedochází.



Obrázek 7.2: Příklad atributů Q-Z kvalitně oddělujících shluky

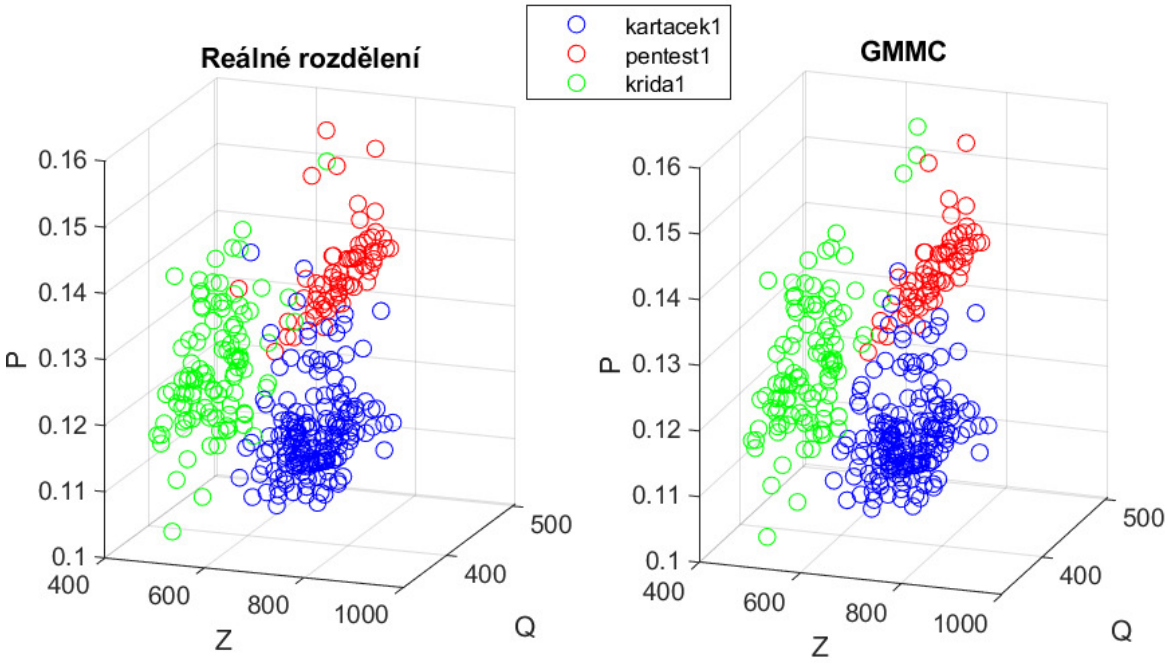


Obrázek 7.3: Klasifikace tří zdrojů metodou SDDT ve srovnávacím zobrazení na prostoru P-W



Obrázek 7.4: Klasifikace tří zdrojů signálu v prostoru P-W

Pokud bychom alespoň částečně kompenzovali deficit nižšího množství informací, který ostatní metody oproti SDDT mají, přidáním dalšího atributu a tedy rozšířením výběrového prostoru na $\mathbb{R}^3$, jsme schopni dosáhnout srovnatelných úspěšností klasifikace jako SDDT. Například pro metodu GMMC a výběrový prostor Z-Q-P dostáváme úspěšnost klasifikace 97,4 %, viz obrázek 7.5



Obrázek 7.5: Klasifikace tří zdrojů AE na prostoru Z-Q-P metodou GMMC

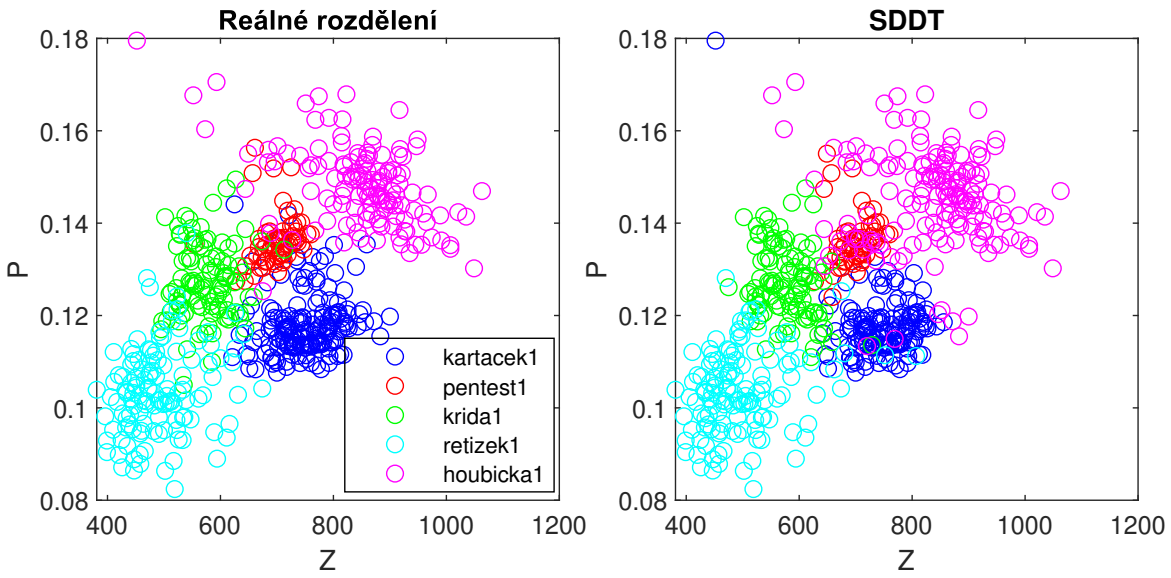
Klasifikace pěti různých zdrojů AE

Nyní budeme klasifikovat všech pět zdrojů signálu měřených na plechu. Opět se omezíme na dvourozměrný výběrový prostor a použijeme stejné dvojice atributů jako v předchozí sekci. Výsledky klasifikace jsou uvedeny v tabulce 7.6. Pro grafické zobrazení výsledků klasifikace, viz obrázek 7.8, jsme zvolili kombinaci atributů Z-P. Signály jednotlivých zdrojů tu vytváří pět relativně oddělených elipticky tvarovaných shluků, díky čemuž už i metoda MBC dosahuje velmi kvalitních výsledků. Zde stojí za zmínku fakt, že z důvodu náhodné inicializace počátečních parametrů v EM algoritmu neposkytne MBC metoda vždy stejný výsledek, a to i přesto, že EM algoritmus proběhne pětkrát s různými inicializačními parametry a je vybrán pouze běh maximalizující hodnotu věrohodnostní funkce. Pokud bychom tedy MBC metodu spouštěli stále znovu dostaneme eventuelně stejný výsledek jako u metody SMBC. Dále si v tabulce 7.6 lze povšimnout, že SKDEC je pro všech prvních pět kombinací atributů nejlepší a překonává stejné množství dat využívající metodu SMBC o 8-34 %.

Atributy	MBC	SMBC	GMMC	SKDEC
P-Q	73,3	78,9	85,2	87,4
Q-Z	74,6	79,7	89,9	91,0
Q-W	49,3	59,3	72,4	74,3
W-M	62,9	65,3	73,1	79,6
S-P	50,5	50,7	79,2	84,1
Z-P	86,7	88,7	**92,0**	91,1

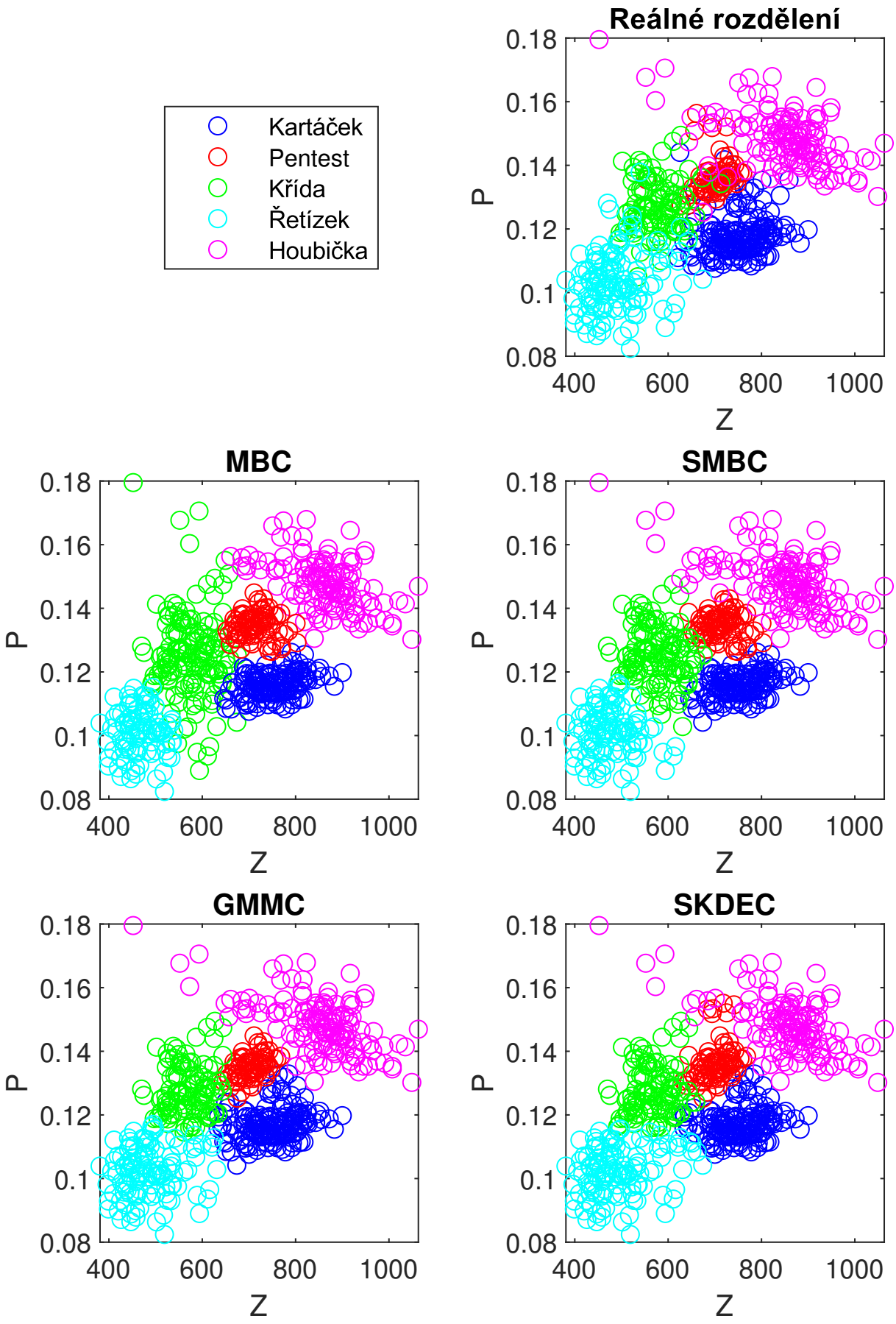
Tabulka 7.6: Úspěšnost klasifikace pěti zdrojů akustické emise uvedená v % správně klasifikovaných signálů

Klasifikace metodou SDDT s úspěšností 89,9 % je zobrazena na obrázku 7.7, vynesena je pro porovnání opět v proměnných Z-P. Výsledek metody SDDT se v tomto případě o 1-2 % nevyrovná metodám GMMC ani SKDEC a to ani při použití pouze dvou atributů.

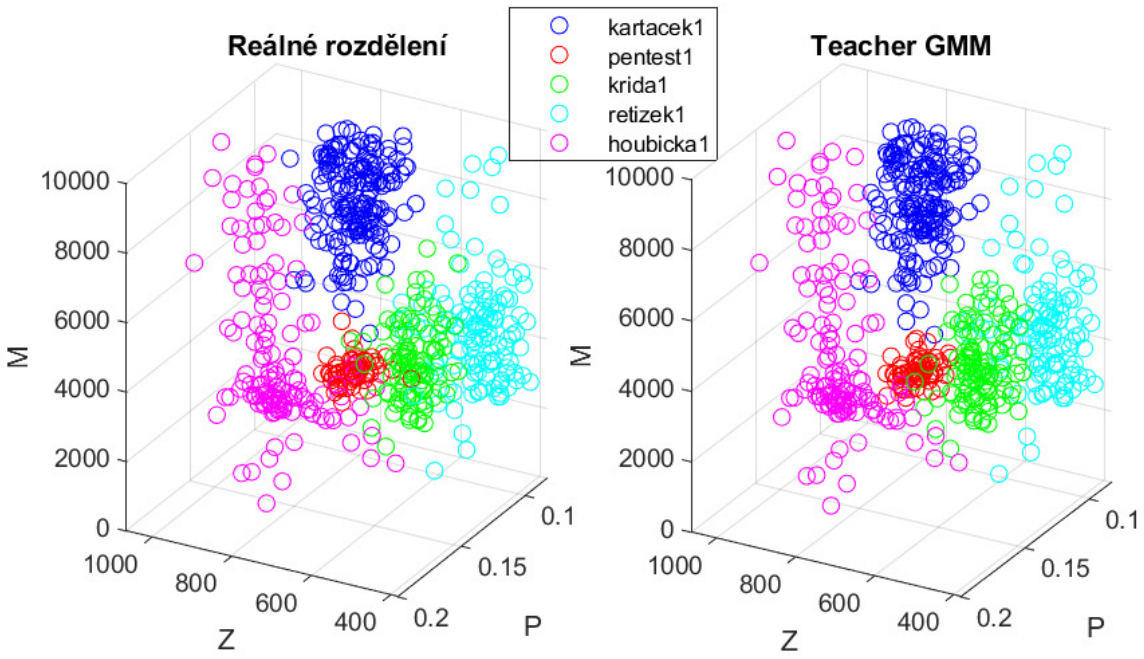


Obrázek 7.7: Klasifikace pěti zdrojů AE metodou SDDT ve srovnávacím zobrazení na prostoru Z-P

Při použití tří atributů Z-P-M, viz obrázek 7.9, lze metodou GMMC dosáhnout úspěšnosti 94,7 %, jedná se o nejlepší dosažený výsledek pro klasifikaci těchto pěti druhů signálu.



Obrázek 7.8: Klasifikace pěti zdrojů signálu AE na prostoru Z-P



Obrázek 7.9: Klasifikace pěti zdrojů na prostoru Z-P-M metodou GMMC

Oddělení podskupiny pentestových signálů

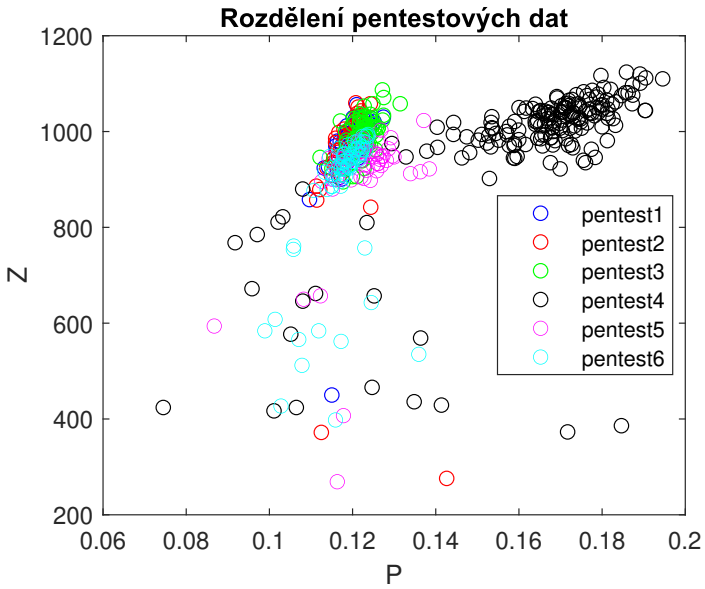
V této části použijeme i signály pocházející z měření na hliníkové eloxované konvici. Jedná se o pět typů pentestových signálů, které byly buzeny lámáním tuh různého průměru a tvrdosti. Pokusy o oddělení signálů různých tuh selhaly, protože signály, a tedy i atributy z nich extrahované, byly příliš podobné. Až na data pentest4 tvoří jediný shluk a přímo se překrývají, viz obrázek 7.11, kde jsme volili proměnné Z-P, protože je v nich alespoň skupina pentest4 dobře odseparovatelná. Tyto vlastnosti dat nám dávají možnost pokusit se oddělit všechny naměřené signály podskupiny pentestů od pozadí, které budou tvořit ostatní typy signálů naměřené na plechu. V tabulce 7.10 jsou uvedeny výsledky této klasifikace.

Atributy	DIVIZ	MBC	SMBC	GMMC	SKDEC
Q-Z	84,1	91,3	91,3	89,1	90,7
Q-W	88,2	89,5	89,6	90,0	92,6
Z-P	83,0	78,3	92,4	86,5	**93,7**

Tabulka 7.10: Úspěšnost separace pentestů oproti pozadí uvedená v % správně klasifikovaných signálů

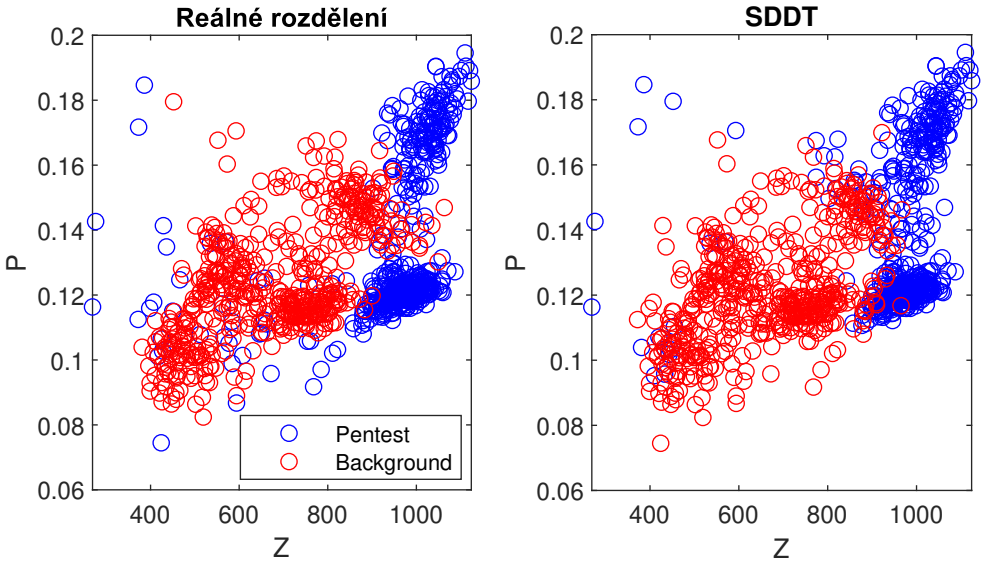
Pro grafické zobrazení klasifikace jsme opět zvolili kombinaci atributů Z-P na obrázku 7.12, zejména z důvodu, že pozorování pentestů mají v tomto případě bimodální hustotu a je zajímavé pozorovat jak se s ní jednotlivé metody vypořádají. Divizivní metoda se nehodí na případy plynule mezi sebou přecházejících shluků, metoda MBC v tomto případě také v podstatě selhává, neboť naprosto ignoruje druhé ohnisko pentestových signálů. To je nejspíše způsobeno tím, že MBC byla nucena hledat pouze dva shluky. V praxi se tento problém překonává tak, že se hledá vícero komponent, což ale vede k potížím s interpretací jednotlivých detekovaných shluků a podshluků. Překvapivě kvalitního výsledku dosáhla SMBC, která spojila obě ohniska

do elipsoidního shluku o něco šikovněji než metoda GMMC. Zdaleka nejlepší výsledek poskytuje metoda SKDEC, která má jako jediná potenciál odhalit bimodální povahu dat.

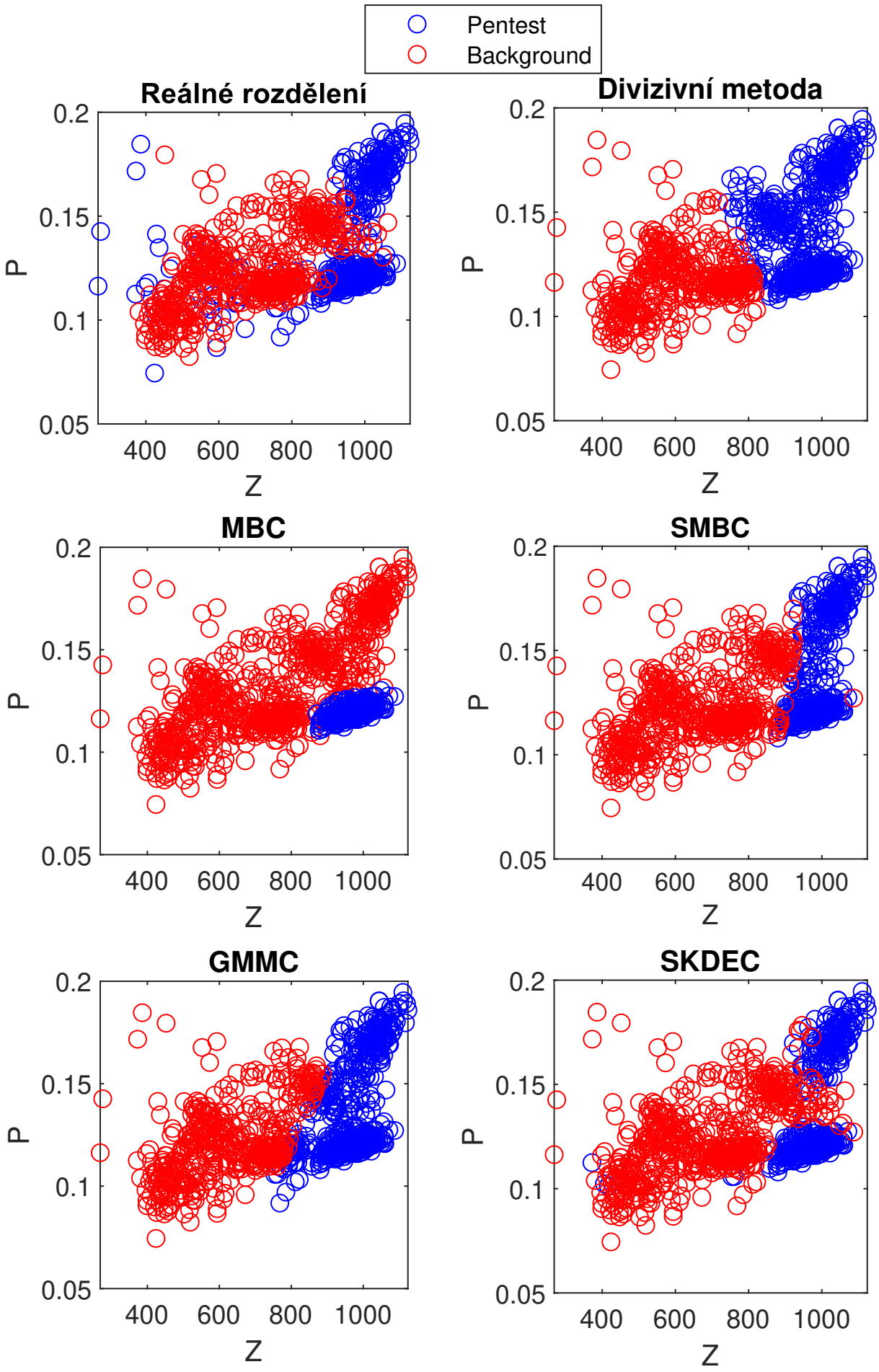


Obrázek 7.11: Rozložení signálů pentestů naměřených na konvici ve výběrovém prostoru P-Z

Zbývá nám použít metodu strojového učení SDDT, která byla pro takovýto úkol binární klasifikace původně navržena. Ta klasifikuje tyto data s úspěšností 90,9 %, výsledek je znázorně na obrázku 7.13. Metoda SDDT tedy o 1-2,7 % nedosahuje takové úspěšnosti jako triviálnější supervizované metody, předpokládáme, že to je dáno především nedostatečným množstvím trénovacích dat, která tak musela být uměle generována, což mohlo vést k její kompromitaci.



Obrázek 7.13: Klasifikace dvou zdrojů metodou SDDT ve srovnávacím zobrazení na prostoru Z-P



Obrázek 7.12: Klasifikace dvou zdrojů signálů pentest vs. ostatní na prostoru Z-P

Závěr

Hlavním cílem této bakalářské práce bylo nalezení vhodných atributů signálu akustické emise a následné porovnání klasifikačních metod na laboratorně naměřených datech. Navrhli jsme dva atributy (P a M) z celkových šesti použitých. Při klasifikaci se osvědčily různé kombinace atributů Z, P, Q, W a M, vždy v závislosti na konkrétních datech. Parametr S nerozděloval kvalitně pozorování z různých zdrojů a při klasifikaci se neosvědčil.

Celkem bylo implementováno a odzkoušeno šest klasifikačních metod. Nejjednodušší divizivní metoda se ukázala jako použitelná jen v ideálních případech dvou znatelně oddělených shluků. Znatelně sofistikovanější metoda MBC, která byla popsána ve 4. kapitole, nabídla díky odhadu hustot distribuční směsi možnost klasifikovat více shluků. Její slabiny se projevily při klasifikaci silně se překrývajících shluků, kdy měla tendenci překryv sloučit v jeden klastr, kvůli svému omezení pouze na normální rozdělení hledaných shluků. Po analýze tohoto chování jsme navrhli vylepšení v podobě metody SMBC, která na základě trénovací množiny dat lépe inicializuje EM algoritmus, jež tvoří základ metody MBC. Ukázalo se, že použití SMBC může výrazně zvýšit úspěšnost, zejména v případě klasifikace kolidujících shluků. Metoda GMMC byla navržena jako zjednodušení metody MBC, kdy k odhadu distribuční směsi jsou použita trénovací data. Paradoxně navzdory své relativní jednoduchosti a výpočetní nenáročnosti dosahuje často výrazně lepších výsledků než metody MBC či SMBC. Stále je však omezena na normální rozdělení a tedy na více-méně eliptický tvar shluků, což se v některých případech ukázalo jako nepraktické. Dále je představena teorie jádrových odhadů a na základě nich navržená klasifikační metoda SKDEC. Ta už netrpí omezeními metody GMMC na elipticky symetrické shluky a jsme s její pomocí schopni klasifikovat i nekonvexní shluky. V šesté kapitole jsme shrnuli základní poznatky o ϕ-divergencích a popsali převzatý stromový klasifikátor SDDT, který využívá Rényiho divergenci k optimalizaci klasifikace.

V poslední kapitole jsme pomocí implementovaných metod klasifikovali celkem tři různé soubory dat z laboratorních měření. V prvním případě šlo o tři zdroje akustické emise, v druhém o pět zdrojů, které všechny pocházely z měření na plechové desce. Třetí soubor dat odděloval pentestové signály měřené na eloxované konvici od pozadí tvořeného signály z plechu. Z výsledků je patrné, že úspěch klasifikace závisí především na volbě atributů - pro kvalitní atributy jsou všechny klasifikační metody schopny dosáhnout výborných výsledků. Jako nejspolehlivější klasifikátor se ukázala metoda SKDEC, především kvůli problémům, které měly metody založené na MBC s klasifikací dat s bimodální nebo nesymetrickou hustotou.

Relativní neúspěch metody SDDT přisuzujeme nedostatečnému množství trénovacích dat. V další práci se budeme zabývat úpravou SDDT algoritmu tak, aby byl použitelnější pro klasifikaci signálů akustické emise.

Literatura

[1] Jan Tláskal: *Statistické klasifikační metody v akustické emisi* . Diplomová práce, KM FJFI ČVUT, Praha, 2009.

[2] Zuzana Farová: *Statistické metody odhadu hustot a klasifikace signálů*. Diplomová práce, KM FJFI ČVUT, Praha, 2010.

[3] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, Teh Ying Wah: *Time-series clustering – A decade review*. Information Systems, 53:16-38, 2015. *https://doi.org/10.1016/j.is.2015.04.007*

[4] Zuzana Dvořáková: *Classification of Acoustic Emission Stochastic Signals for Defects Imaging (Signal Deconvolution Principle by Means of Time Reversal Symmetries with Biomedical and Industry Applications)*. Study of the dissertation thesis, KM FJFI ČVUT, Praha, 2013.

[5] Yolanda L. Hinton: *Problems Associated With Statistical Pattern Recognition of Acoustic Emission Signals in a Compact Tension Fatigue Specimen*. Langley Research Center Hampton, Virginia 23681-2199, 1999.

[6] Michael Jordan: *The Multivariate Gaussian (Bayesian Modeling and Infernce)*. Lecture notes, EECS Berkeley, Spring 2010. *https://people.eecs.berkeley.edu/ jordan/courses/260-spring10/other-readings/chapter13.pdf*

[7] Kristina Jarůšková: *Statistická separace a identifikace s využitím divergenčních technik pro vícerozměrná data*. Bakalářská práce, KM FJFI ČVUT, Praha, 2018.

[8] Kristina Jarůšková: *Identifikace energetických prongů při oscilaci neutrin v experimentu NOvA*. Výzkumný úkol, KM FJFI ČVUT, Praha, 2019.

[9] Bernard W. Silverman: *Density estimation for statistics and data analysis*. Boca Raton: Chapman & Hall/CRC, 1998.

[10] Petr Bouř, et al: *Kernel and divergence techniques in high energy physics separations*. J. Phys.: Conf. Ser. 898 072004, 2017.

[11] Emanuel Parzen: *On Estimation of a Probability Density Function and Mode*. The Annals of Mathematical Statistics [online]. 1962, 33(3), 1065-1076 [cit. 2021-06-16]. DOI: 10.1214/aoms/1177704472.

[12] Petr Bouř: *Divergenční metody ve statistických separacích*. Bakalářská práce, KM FJFI ČVUT, Praha, 2014.

[13] Igor Vajda: *Information-theoretic methods in Statistics*. Research Report, ÚTIA AV ČR, 1995.