Univerzita Pardubice

Fakulta ekonomicko-správní

Modelování predikce ozónu pomocí SVM

Bc. Jiří Kolín

Diplomová práce

2011

Univerzita Pardubice
Fakulta ekonomicko-správní
Akademický rok: 2010/2011

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Bc. Jiří KOLÍN**
Osobní číslo:	**E09833**
Studijní program:	**N6209 Systémové inženýrství a informatika**
Studijní obor:	**Informatika ve veřejné správě**
Název tématu:	**Modelování predikce ozónu pomocí SVM**
Zadávající katedra:	**Ústav systémového inženýrství a informatiky**

Zásady pro vypracování:

- Analyzujte vstupní data (parametry) pro následující predikci.
- Charakterizujte SVM z hlediska aproximace a predikce.
- Navrhněte model na predikci ozónu.
- Verifikujte navrhnutý model.
- Uskutečněte analýzu výsledků.

Rozsah grafických prací:

Rozsah pracovní zprávy:

Forma zpracování diplomové práce: **tištěná/elektronická**

Seznam odborné literatury:

KVASNIČKA, V. a kol.: Úvod do teórie neurónových sietí. Iris, Bratislava, 1997.

HAYKIN, S.: Neural Networks: A Comprehensive Foundation. 2nd edition, New Jersey, Prentice-Hall, Inc., 1999, 842s.

Vedoucí diplomové práce: **prof. Ing. Vladimír Olej, CSc.**
Ústav systémového inženýrství a informatiky

Datum zadání diplomové práce: **5. října 2010**

Termín odevzdání diplomové práce: **6. května 2011**

L.S.

doc. Ing. Renáta Myšková, Ph.D.
děkanka

doc. Ing. Jiří Křupka, Ph.D.
vedoucí ústavu

V Pardubicích dne 5. října 2010

Prohlášení autora

Prohlašuji:
Tuto práci jsem vypracoval samostatně. Veškeré literární prameny a informace, které jsem v práci využil, jsou uvedeny v seznamu použité literatury.
Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Souhlasím s prezenčním zpřístupněním své práce v Univerzitní knihovně.

V Pardubicích dne 22.6.2011

Bc. Jiří Kolín

Poděkování

Rád bych na tomto místě poděkoval vedoucímu mé diplomové práce, panu prof. Ing. Vladimíru Olejovi, Csc. za jeho podporu, mnoho cenných návrhů, rad a především za čas, který mi věnoval. Dále bych rád poděkoval svým rodičům a přátelům za všestrannou podporu při studiu.

SOUHRN

Diplomová práce se zabývá modelováním predikce ozonu pomocí SVM neuronových sítí. Cílem je popis základních charakteristik ozonu a oblastí s ním spojených, spolu s vytvořením modelu v programu Statistica. Jako vstupní data byly použity hodnoty naměřené na Dukle v Pardubicích. Tato diplomová práce se skládá ze dvou základních částí.

V první z nich, v rámci seznámení se s tématem, budou popsány vlastnosti a druhy ozonu následované procesem jeho vzniku a možností měření. V poslední části této kapitoly zmíním také látky, které se podílejí na poškozování a naopak úmluvy a závazky přijaté na ochranu ozonové vrstvy.

Druhou, a významnější částí této práce, je charakteristika SVM neuronových sítí a vlastní tvorba modelu. K tomu je nejprve potřeba analýza vstupních dat následovaná návrhem modelu a jeho verifikací. Poslední kapitola se zabývá analýzou výsledků a charakteristik navržené struktury.

KLÍČOVÁ SLOVA

Ozonová vrstva, ultrafialové záření, freon, Support Vector Machine, jádro, klasifikace, regrese, predikce, učení s učitelem.

TITLE

Modelling of ozone predictions by SVM

ABSTRACT

This thesis deals with the modelling of ozone predictions by SVM neural networks. The aim is to describe the basic characteristics of ozone and associated areas, together with the creation of model in Statistica. As input data were used values measured in Dukla in Pardubice. This thesis consists two main parts.

The first one, the introduction to the topic, describing the characteristics and types of ozone and it is followed by a process of its formation and measurement options. In the last section of this chapter I will describe the substances that involving in the damaging of ozone layer and contrary to them I will describe some conventions and commitments made to protect the ozone layer.

The second and more important part of this thesis describing characteristics of SVM neural networks and creation of model. For this is required input data analysis followed by model design and verification. The last chapter analyzes the results and characteristics of the proposed structure.

KEYWORDS

Ozone layer, ultraviolet radiation, CFC, Support Vector Machine, kernel, classification, regression, prediction, supervised learning.

Obsah

Úvod

Ozon je jedním z přirozeně se vyskytujících stopových plynů, které tvoří naši atmosféru. Z chemického hlediska lze říci, že se jedná o kyslík s extra přidanou molekulou navíc. Z elektrického hlediska lze označit ozon jako kyslík s vyšší energetickou úrovní. V závislosti na tom, kde se v atmosféře nachází, ovlivňuje život na Zemi buď dobrém, nebo ve špatném slova smyslu. Stratosférický ozon vzniká přirozeně prostřednictvím interakce slunečního ultrafialového záření s molekulárním kyslíkem a snižuje množství škodlivého UV záření dopadajícího na zemský povrch. Troposférický ozon je tvořen hlavně z reakcí mezi znečišťujícími látkami v ovzduší. Ty jsou závislé na přítomnosti tepla a slunečního záření, což přináší větší tvorbu ozonu v letních měsících. Každé čtyři roky Světová meteorologická organizace a Organizace spojených národů pro životní prostředí přezkoumají stav ozónové vrstvy za účelem stanovení jejího zeslabení a kvůli potvrzení účinnosti Vídeňské úmluvy a Montrealského protokolu. Přestože byly již v minulém století podniknuty razantní kroky ke snížení nebo ke kompletnímu potlačení emisí škodlivých látek v ovzduší a i nadále dochází každoročně k jejímu dalšímu snižování po celém světě, potrvá obnovení ozonové vrstvy ještě mnoho desetiletí a nikdo nemůže s jistotou říci, že dojde k obnovení do původního stavu, neboť tato vrstva je ovlivňována celou řadou složitých procesů, které člověk ještě stále nemá nebo nemůže mít pod kontrolou.

Na základě posloupnosti zjištěných datových řad lze predikovat velikost ozonové vrstvy s určitou přesností i do budoucna. K tomu můžeme využít SVM sítě představené na konferenci v roce 1992, jejichž jádro, i když v jiné podobě, je známo již od roku 1964. Tato metoda, založená na principu učení s učitelem, může být použita k vyřešení jak klasifikačního, tak i regresního problému. Hlavním cílem modelů je nalezení optimální nadroviny definující rozhodující hranice, které následně slouží k oddělení kategorií cílové proměnné. Support Vector Machines můžeme označit za výkonný algoritmus se silnými teoretickými základy z teorie Vapnik-Chervonenkis představující rozšíření nelineárních modelů. Jeho algoritmus je založen na statistické teorii učení a VC dimenzi, kterou zavedl společně Vladimir Vapnik a Alexey J. Chervonenkis. Tato metoda založená na neuronových sítích je v současnosti používána jako jeden ze standardních nástrojů strojového učení a data miningu.

Cílem diplomové práce je vytvoření modelu pro predikci ozonu pomocí výše zmíněných SVM neuronových sítí. K jeho tvorbě bude využit program Statistika od společnosti StatSoft CR s.r.o., ve kterém bude následně tento vytvořený model testován na poskytnutých datech. Vstupním souborem dat jsou hodnoty ozonu a jiných látek ovlivňujících jeho výskyt naměřené v městské části Dukla, v Pardubicích.

Vlastní práce se skládá ze 2 stěžejních částí - teoretické, ve které je probírána problematika ozonové vrstvy, spolu s látkami, které tuto vrstvu poškozují a také úmluvami, které byly uzavřeny, aby předcházely jejímu dalšímu poškozování a praktické, která se již zabývá vývojem a historií SVM neuronových sítí, jejichž znalost je následně využita pro návrh modelu prediktoru ozonové vrstvy. Práce je zakončena analýzou a zhodnocením dosažených výsledků, spolu s porovnáním jednotlivých použitých metod.

1. Problematika ozonové vrstvy

Pravděpodobně prvním člověkem, který zjistil přítomnost ozonu, byl holandský chemik Van Marum. V popisu svého experimentu zmínil charakteristický zápach, který cítil. I přesto byl objev ozonu zmíněn až o desetiletí později v zápiscích německo-švýcarského chemika Christiana Friedricha Schönbeina, které jsou z roku 1840. Jeho objev byl předložen k prostudování univerzitě v Mnichově. Christian Schönbein si během svých experimentů všiml stejného charakteristického zápachu jako Van Marum. Nazval tento plyn ozon, slova odvozeného od „ozein“, které je řeckého původu a znamená vůni. Po roce 1840 následovalo mnoho studií na možnosti jeho použití jako dezinfekčního přípravku. První generátor na produkci ozonu byl zhotoven v Berlíně společností Von Siemens. Tento výrobce také napsal knihu o možnostech jeho využití ve vodě, což způsobilo několik projektů, během nichž byly zkoumány dezinfekční vlastnosti této sloučeniny [11, 25].

Když se přesuneme do současnosti, tak ohrožení ozonové vrstvy patří mezi jeden z nejvýznamnějších problémů dnešní doby. Vlivem zásahů lidské populace do přírodního ekosystému a jeho změnou dochází mimo jiné ke snižování tloušťky ozonové vrstvy a tím i většímu průchodu nebezpečného záření UV-B a UV-C. Z nedávné historie víme, že toho téma bylo velmi populární již okolo 85. roku minulého století, když se hojně používaly halogenové uhlovodíky, též známé jako freony. Tyto látky totiž nejen že poškozovaly ozonovou vrstvu, ale také zvyšovaly vliv tzv. skleníkového jevu.

1.1. Vlastnosti ozonu

Chemická značka ozonu O_3 značí, že se jedná o sloučeninu tří atomů kyslíku. Už zde je možné vidět jeho rozdílnost od běžného, atmosférického kyslíku, který je tvořen pouze dvěma molekulami.

Obr. 1: Molekulová struktura ozonu [21]

Z obr. 1 je možné vidět strukturu ozonu, ve které je úhel svíraný mezi molekulami 116.8°. Další charakteristické vlastnosti jsou uvedeny v tab. 1.

Molární hmotnost	47.998 g·mol^{-1}
Vzhled	plyn namodralé barvy
Hustota	2.144 g/L (0 °C)
Teplota tání	80.7 K, −192.5 °C
Teplota varu	161.3 K, −111.9 °C
Kritická teplota	−12 °C
Kritický tlak	5.4 MPa
Rozpustnost ve vodě	0.105 g/100mL (0 °C)
Index lomu	1.2226
Dipólový moment	0.53 D

Tab. 1: Charakteristické vlastnosti ozonu [23]

Vznik ozonu probíhá ve dvou stupních pomocí ultrafialového záření UV-C, nebo při elektrických výbojích. Nejdříve dojde k rozštěpení dvouatomové molekuly kyslíku na dva jednotlivé atomy. Každý z nich se následně ihned spojuje s dalšími dvouatomovými molekulami. Ozon je poměrně nestabilní sloučenina, jejíž střední doba života je při teplotě -25 °C jen přibližně 18 dní a poté se opět rozkládá. Při této reakci dochází ke zvyšování teploty a snižování tlaku. Při jeho výrobě prochází několika fázemi. Nejdříve je ozon namodralým plynem, ale při jeho postupném ochlazování dochází ke změně skupenství na tmavě modrou kapalinu a poté na pevnou látku stejného zabarvení [5].

Zastoupení na Zemi má velmi malé, jen okolo 10^{-5} procenta. Téměř všechen se vyskytuje ve stratosféře a jen přibližně 10% se nachází ve výškách do 10km od

zemského povrchu. Jelikož má velmi vysokou absorpci, tak i při malém zastoupení prvků na Zemi stačí k zabránění průchodu převážného množství UV záření.

Množství ozonu je měřeno v Dobsonových jednotkách (D.U.). Gordon Miller Bourne Dobson byl meteorolog a fyzik, který prozkoumal ozonovou vrstvu země a vyvinul první jednoduchý spektrometr pro jeho měření. Jedna D.U. je rovna $2.69*10^{16}$ ozonových molekul, které se nacházejí na čtverečním metru. Toto množství za standardních podmínek odpovídá přibližně vrstvě o výšce 10 μm [24].

1.2. Druhy ozonu

Jelikož se ozon nachází v různých nadmořských výškách a je vytvářen různými způsoby, lze jej dělit na [14, 22]:

Stratosférický ozon

Asi 90% ozónu v zemské atmosféře se nachází v oblasti nazývané stratosféra. Jedná se atmosférickou vrstvu mezi 20 a 30 km nad zemským povrchem. Ozón tu tvoří vrstvu, ve které je více koncentrovaný než kdekoliv jinde. Zatímco kyslík a ozón spolu absorbují 95 až 99.9% ultrafialového záření Slunce, tak pouze ozon účinně pohlcuje ultrafialové světlo známé jako UV-B (280 nm – 315 nm) a UV-C (100 nm – 280 nm). Ultrafialové záření s vlnovou délkou od 320 do 400 nanometrů (UVA) není absorbováno.

Troposférický ozon

Vzniká především za horkých letních dnů a bezvětří, při působení intenzívního slunečního záření, reakcí uhlovodíků a oxidů dusíku. Nachází se v přízemní vrstvě zemské atmosféry (0 - 2 km). Jeho hlavním rozdílem od ozonosféry je, že při zvýšení své koncentrace má negativní dopady, především na živé organismy.

Endogenní ozón

Tento druh představuje nejméně rozšířený typ ozonu. Nachází se v těle teplokrevných živočichů a je vytvářen bílými krvinkami, ze kterých pak dochází k jeho uvolňování do krve a tkání. Tím přispívá k odstraňování choroboplodných zárodků.

1.3. Vlastnosti látek poškozujících ozonovou vrstvu

Mezi látky, které nejvíce poškozují ozonovou vrstvu, patří freony a halony. Freony jsou označovány halogen-deriváty uhlovodíků, které ve své stavbě obsahují minimálně 2 vázané halogeny, přičemž alespoň jeden z nich musí být fluor. Tyto látky byly využívány ve formě plynů, popřípadě kapalin, pro chladící zařízení, hasící prostředky a v neposlední řadě jako hnací plyny ve sprejích. Mezi jejich vlastnosti patří, že jsou výborné izolanty a rozpouštědla, která nemají zápach, barvu a jsou nehořlavé. Nejznámějšími představiteli freonů jsou například chlortrifluormethan a trichlorfluormethan.

Jako halony jsou označovány látky podobající se chlorofluorovaným uhlovodíkům. Mezi hlavní prvky těchto sloučenin patří uhlík, fluor, brom a někdy i chlor. Za normálních podmínek jsou velmi stálé a netoxické. Vyskytují se ve formě nízkovroucích kapalin nebo plynů. Jejich hlavními představiteli je například tribromfluormethan nebo methylbromid.

Dříve byly freony velmi využívány a tak dostaly zvláštní označení. Jako CFC látky začínaly symbolem R s dalšími 2 až 4 číslicemi. Písmeno R značí anglické slovo refrigerant = osvěžující, chladící. U cyklických freonů je využito písmeno C, a jestliže obsahovala molekula brom, tak bylo využito písmeno B. Notace halonových prvků, se od této freonové, liší.

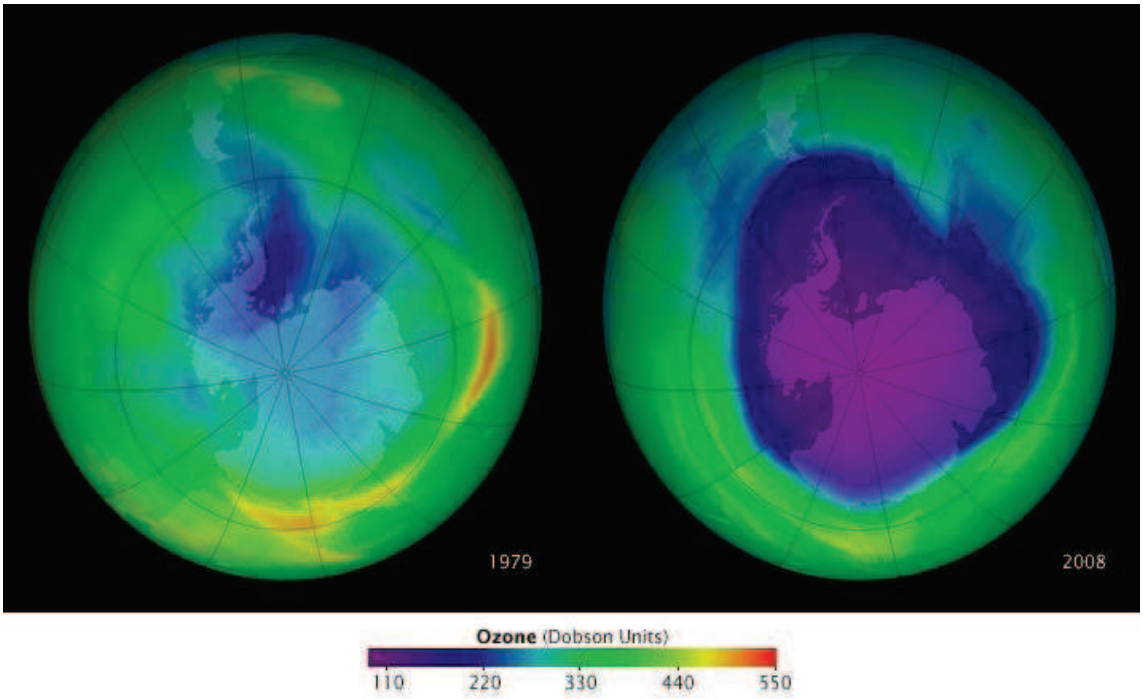
Poškození ozonu je měřeno pomocí jednotky ODP (Ozone Depletion Potential). Tato veličina byla stanovena v roce 1983 jako míra destruktivních účinků určité látky vzhledem k CFC trichlorfluormethanu (R11). ODP sloučeniny R11 je tedy rovna jedné. Hodnoty jednotlivých látek jsou zjišťovány buď pomocí výpočtů, nebo experimentů. Při výpočtech je potřeba znát kinetiku fotochemických dějů zkoumané látky. U experimentů je naopak potřeba dodržet potřebné podmínky (vlastnosti stratosféry) [26].

1.4. Vznik ozonové vrstvy

Jelikož je množství ozonu velmi úzce spjato s množstvím kyslíku na naší planetě, je potřeba se za jeho vznikem vydat okolo 3 miliard let do minulosti. V té době byl obsah kyslíku v atmosféře přibližně 0.1%. S postupem let se ale začal jeho poměr zvyšovat a tak rostlo i množství ozonu. To se odvíjí od nadmořské výšky, se kterou

souvisí i počet atomů kyslíku O_2. Postupem času se procento kyslíku zvyšovalo, až v období před 420 miliony let dosáhlo úrovně 10% a tím došlo k posunu oblasti pro tvorbu ozonu i do větších nadmořských výšek [5].

Přítomnost ozonu byla vědci odhalena v 19. stol. a již tehdy bylo pomocí pokusů dokázáno, že při zemi je jeho koncentrace velmi nízká a se zvyšující se nadmořskou výškou roste. Bohužel v té době neexistovala dostatečně vyspělá technika, aby byly domněnky o ozonosféře potvrzeny. Postupem času a rozvojem techniky však byla dokázána její existence. Nachází se ve výškách 20-30 km nad zemským povrchem. Působením UV záření je zde rozkládán dvouatomový kyslík, který ihned reaguje s okolními molekulami. Po určité době se opět rozkládá zpět, takže se jeho množství nehromadí. Těmito reakcemi dochází jednak k pohlcovaní UV záření, a také k uvolňovaní energie, což ovlivňuje teplotu této vrstvy. Množství slučovaných i štěpených atomů je v rovnováze, které je ve vrchní stratosféře docíleno poměrně rychle (v řádu sekund), naopak v dolní stratosféře může nastolení rovnováhy trvat v řádu dnů. Cirkulaci vzduchu a tím i ozonu zajišťuje vzdušné proudění, které ho přenáší mezi jednotlivými vrstvami. Tím dochází při jeho měření ke kolísaní hodnot. Tímto jevem je vysvětlován i vyšší výskyt ozonu v rovníkových oblastech a nižší v oblasti pólů.



Obr. 2: Ozonová vrstva v letech 1979 a 2008 [2]

Asi nejdiskutovanějším a nejznámějším tématem poslední doby je termín „ozonová díra“. Takto je označováno místo, kde množství ozonu dočasně pokleslo pod 50% svého průměrného stavu. Jeden z největších úbytků za poslední roky byl zaznamenán 14. 03. 2011 nad Antarktidou [6]. Na předchozím obr. 2 je zobrazeno porovnání vývoje množství ozonu z let 1979 a 2008.

1.5. Způsoby měření stavu ozonu

Přelom v měření ozonu přišel po 1. sv. válce. Gordon M. B. Dobson v roce 1924 vyvinul přístroj pro měření změny UV záření po průchodu ozonosférou a tím bylo umožněno vypočítat její velikost. Tento přistroj byl po svém vynálezci nazván Dobsonův spektrometr. Je jím měřena intenzita ultrafialového záření různých vlnových délek UV záření, z nichž lze odvodit koncentrace ozonu přítomného atmosféře. Spektrometr obsahuje fotoelektrickou buňku a filtrační zařízení, které umožňuje UV záření na čtyřech vlnových délkách dopadnout postupně na danou buňku. Z těchto 4 vlnových délek jsou dvě absorbovány ozonem a dvě propuštěny. Poměr mezi těmito intenzitami je měřítkem koncentrace ozonu. Z toho je pak možno pomocí spektrálních absorpčních koeficientů a rovnic přenosu elektromagnetického záření určit jeho množství v atmosféře. Přístroj je možno používat k měření UV záření vyzařované Měsícem a hvězdami, stejně tak jako sluneční UV záření, ale u něj mohou být hodnoty zkresleny aerosoly a znečišťujícími plyny v ovzduší. V dnešní době již Dobsonův spektrometr nahradily modernější přístroje. Jedním z jejich příkladů je Brewerův spektrometr, který je již plně automatizovaný a řízený počítačem [5, 7].

Důležité jsou při měření ozonové vrstvy údaje vztahující se k její vertikální výšce. Měření probíhá pomocí sond připevněných na například na meteorologických balonech. Ten je obvykle naplněný buď vodíkem, nebo héliem. Kvůli nižším nákladům je ale převážně používán vodík. Rychlost výstupu lze ovládat množství plynu, s nímž je naplněn. Vlastní měření probíhá ve výšce přibližně 30-40km. Ta je omezená snižujícím se tlakem působícím na balón a tak dochází k jeho rozpínání. Běžně i 100:1.

Jednou z dalších metod měření je Light Detection and Ranging. LIDAR pracuje na principu radaru, ovšem s jinou vlnovou délkou. Je to optická metoda mapování a monitorování vzdálených objektů tím, že na měřeném objektu zaznamenáváme zpětně

rozptýlené záření. Při metodách dálkové detekce umožňuje mimořádná spektrální intenzita laserů provádět analýzu na vzdálenost až několika kilometrů.

Předchozí metody mají jednu velkou nevýhodu. Tou je měření jen v určité výšce a na jednom místě. Proto jsou používána speciální stratosférická letadla. Tato metoda je na rozdíl od ostatních komplexní, protože zde probíhá kompletní analýza chemického složení. Její velkou nevýhodou je ale veliká nákladnost a tak se využívá převážně jen v polárních oblastech.

Pro pozorování vývoje ozonu v globální míře nám slouží již od roku 1970 družice. V současné době jsou využívány například družice Aura, Envisat, Goes, Meteor, Nimbus 7, NOAA, a další. Český hydrometeorologický ústav využívá k měření ozonu data z družice NOAA.

Jako měřící systémy jsou u těchto metod využívány TOMS (Total Ozone Mapping Spectrometer) a SBUV (Solar Backscatter Ultraviolet), které byly oba umístěny na satelitu Nimbus 7. Ten provádí měření pro celou zeměkouli již od roku 1978. Jedná se o družici s prostorovým rozlišením 50km, což slouží ke sledování velkých jevů, například ozonových děr [15].

Obr. 3: Meteorologický balon [3], LIDAR [4], družice NOAA [20]

Na obr. 3 jsou zobrazeny jako názorná ukázka příklady meteorologického balonu, LIDARu a družice NOAA.

Pro celosvětové mapování ozonu a spolupráci mezi jednotlivými státy vnikla organizace GO3OS (Global Ozone Observing System), která sdružuje celosvětovou síť měřících stanic. Je zde zapojen i Český hydrometeorologický ústav.

Pravidelné a spolehlivé sledování atmosférického ozónu je základní zdroj informací pro hodnocení změn ozónové vrstvy a pro tvorbu strategie jeho mezinárodní

ochrany. Globální systém pro pozorování ozonu (GO3OS), který je součástí Global Atmosphere Watch (GAW), poskytuje mimo jiné také informace o množství celkového ozonu, získané od pozemní monitorovacích stanic po celém světě. Měření je prováděno Dobsonovými spektrofotometry používanými k získání základního souboru údajů, který je často používán jak pro zkoumání dlouhodobých změn atmosférického ozónu, tak i pro hodnocení kvality ostatních monitorovacích systémů.

S měřením ozonu je velmi těsně spjata i oblast měření škodlivého UV-B záření. To ale zabírá asi jen 1% z celkového spektra a je jen možno měřit například pomocí filtrových UV-radiometrů a kapesních UV-B dozimetrů.

1.6. Úmluvy a závazky na ochranu ozonové vrstvy

V roce 1974 došlo ke zjištění škodlivosti freonu pro ozonovou vrstvu. Některé státy, jako například Norsko, Kanada a USA začali ihned přijímat patřičná opatření. První celosvětovou akcí byla v roce 1985 Vídeňská úmluva, následovaná Montrealským protokolem v roce 1987. Oba dokumenty se týkaly omezení výroby a používaní látek na bázi freonů. Jelikož stále nedocházelo ke zlepšování situace, tak byly přijaty doplňující dodatky (Londýnský, Kodaňský a další), týkající se rozšíření počtu látek, které budou označeny jako škodlivé. Dále došlo u sledovaných látek k jejich doplnění o další, u kterých není riziko ohrožení ozonosféry příliš veliké, ale i tak představují hrozbu. Tímto je dosaženo zpřísnění, zavedené i do určitých článků daných protokolů. Česká republika, tehdy Československo, se také zapojila. V roce 1990 vláda přistoupila na Vídeňskou úmluvu a Montrealský protokol [5, 17].

Vídeňská úmluva

Cílem této dohody je ochrana lidského zdraví a životního prostředí proti nepříznivým účinkům v důsledku lidských činností, které mění nebo by mohly změnit ozonovou vrstvu. Úmluva vyzývá státy, aby přijaly vhodná opatření v souladu s jejími ustanoveními a protokoly. K dosažení výše uvedených cílů se očekává, že budou spolupracovat, aby lépe pochopily a tím mohly posoudit dopady lidské činnosti na ozonovou vrstvu a vlivu jejích změn. Zavazují se k přijmutí vhodných opatření, ke kooperaci s příslušnými mezinárodními subjekty a spolupráci na tvorbě dohodnutých opatření.

Montrealský protokol

Protokol byl podepsán dne 16. 9. 1987 s platností od 1. 1. 1989. Tehdejší Československá republika se k němu připojila v roce 1990, s tím, že dokument začal platit od 1. 1. 1991. Pro ČR platí od 1. 1. 1993. Cílem bylo pro každou ze stran zajistit, aby úroveň spotřeby regulovaných látek nepřekročila vypočtenou úroveň spotřeby z roku 1986.

Londýnský dodatek

Tento dokument byl přijat 29. 6. 1990 při příležitosti 2. konference států podepsaných na Montrealském protokolu. V platnost vstoupil dnem 1. 1. 1992 a tím byl rozšířen počet sledovaných a regulovaných látek. Bylo ujednáno, že všechny strany zabezpečí, aby v průběhu roku 1995 a ve všech následujících ročních obdobích, nepřesahovala vypočítaná hladina roční spotřeby kontrolovaných látek 50 % spotřeby z roku 1986. To se týkalo se to zejména freonů a halonů.

Kodaňský dodatek

Dodatek byl schválen 25. 11. 1992. Jeho platnost byla stanovena ale až od 14. 6. 1994. Představuje další doplnění a zpřísnění Londýnského dodatku, zvláště o časové a objemové údaje. Jedná se převážně o halony, CFC a jiné ozonu škodlivé látky.

Montrealský dodatek

Tento dokument se stal dalším v řadě zpřísňujících norem a omezení výroby ozonu nebezpečných látek. Jeho schválení proběhlo 17. září 1997 se vstupem v platnost 10. 10. 1999. Montrealský dodatek upravuje regulace vývozu, dovozu a udělování licencí pro dovoz, nebo vývoz.

Pekingský dodatek

Dodatek podepsaný v Pekingu je zatím posledním vydaným doplněním stanovených norem. K jeho podpisu došlo 3. prosince 1999 a změna předpisů vstoupila v platnost 1. 1. 2001. Jelikož předešlá opatření nestačila k dostatečnému zlepšení situace tak dochází k dalšímu zpřísnění regulace výroby a spotřeby sledovaných látek. Mezi tyto látky byl nově zařazen i bromchlormethan.

V den, kdy Česká republika vstoupila do Evropské unie, vstoupilo v platnost nové nařízení (ES) 2037/2000 o látkách poškozujících ozonovou vrstvu. To však bylo 1. ledna 2010 zrušeno a nahrazeno novým nařízením Evropského parlamentu a rady (ES) č. 1005/2009. Platit zůstala jen určitá doplňující národní právní úprava, která je zabezpečena zákonem o ochraně ovzduší a vyhláškou č. 279/2009 Sb., o předcházení emisím regulovaných látek a fluorovaných skleníkových plynů. Tyto dva národní právní nástroje pokrývají oblasti vymezené výše jmenovaných nařízení ES pro povinnou právní úpravu ze strany členských zemí [18].

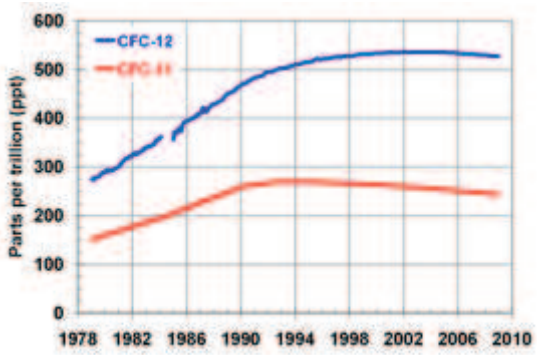
Montrealský protokol stanovuje přesné termíny vyloučení produkce a spotřeby až 100 chemických látek poškozující ozón. Den jeho založení slavíme jako Mezinárodní den ochrany ozonové vrstvy. Díky tomuto protokolu se do konce roku 2009 povedlo snížit 98% chemických látek jím omezovaných. V minulém roce byly, díky jeho doplnění, podniknuty další významné kroky, a to zákaz používání nových tvrdých freonů v rozvojových státech světa (v rozvinutých zemích zakázaný již v roce 1996) a výrazné omezení měkkých freonů v rozvinutých zemích platící od 1. ledna 2010. I přesto je obnova ozónu hudbou budoucnosti. Jen ke stavu z roku 1980 bude potřeba 40 až 65 let přísného dodržování pravidel. S tímto souvisí i zlepšování problému s globálním oteplováním. V období let 1990-2000 se díky regulaci produkce a spotřeby látek snížily emise v přepočtu na CO_2 ve výši 25 miliard tun. Smluvní strany Montrealského protokolu se snaží neustále diskutovat nad dalšími postupy zlepšujícími ochranu klimatu ještě více. Například nad regulací fluorovaných skleníkových plynů sloužících jako náhrada za měkké freony, což bylo hlavním tématem na 22. zasedání v listopadu roku 2010 [16].

Česká republika se také aktivně zapojuje do ochrany ozónové vrstvy a pomáhá s ní i státům, které nejsou schopny své závazky plynoucí z této smlouvy plnit. Jedná se obzvláště o země střední Asie, východní Evropy a Afriky. Příkladem může být například pomoc s instalací a kalibrací Dobsonových ozónových spektrofotometrů odborníky ze Solární a ozonové laboratoře ČHMÚ v Keni, Egyptě, Jihoafrické republice a Botswaně.

Všechny výše zmíněné legislativní prostředky slouží k implementaci požadavků uvedených ve Vídeňské úmluvě a Montrealském protokolu. Mimo těchto úmluv v současnosti platí Kjótský protokol, který se vztahuje na fluorované skleníkové plyny,

občas nazývané také jako F-plyny. Ten je součástí rámcové úmluvy OSN o změně klimatu. Tyto plyny jsou využívány zejména v oblasti chladírenství, klimatizačních zařízení, tepelných čerpadel a systémech požární ochrany. Základní požadavky vytvářející systém předcházení emisím fluorovaných skleníkových plynů stanovuje nařízení Evropského parlamentu a Rady k němuž byly vydány další prováděcí nařízení.

Freony dosáhly nejvyšších hodnot v roce 2000. Od té doby je jejich množství převážně stejné, občas mírně klesá. Průběh vývoje hodnot CFC-11 a CFC-12 je na grafu 1.



Graf 1: Vývoj hodnot freonů mezi roky 1978-2010 [27]

2. Metoda SVM

Support Vector Machines jsou také někdy nazývány jako kernel machines. Jsou velmi podobné neuronovým sítím a využívají schopnost univerzálních aproximátorů jakékoli vícerozměrné funkce, při libovolném stupni přesnosti. V důsledku toho mají zvláštní význam pro modelování systémů, nebo procesů.

SVM byly vyvinuty v pořadí od teorie k realizaci a experimentům. Pro jejich metodu učení je možné říct, že existuje neznámá závislost $y = f(x)$ mezi vstupním vektorem x a výstupem y. Nejsou u nich k dispozici informace o pravděpodobnostní funkci, a proto patří SVM do technik učení s učitelem.
Při navrhování modelů existují dva základní konstruktivní přístupy [19]:

1. vybrat vhodné struktury modelu (pořadí polynomů, počet neuronů, počet pravidel ve fuzzy logice modelu) a tím minimalizovat chybu učení.
2. udržet hodnotu chyby učení fixní (přibližně rovnou nule nebo nějaké přijatelné úrovni), a minimalizovat interval spolehlivosti.

Při tvorbě modelů neuronových sítí je využíván první a u SVM druhý přístup. Při použití metody SVM by měl výsledný model řešit kompromis mezi nedostatečným naučením a přeučením.

2.1. Historie

Původní algoritmus optimální nadroviny navržený Vladimirem Vapnikem v roce 1963 byl lineárním klasifikátorem. Nicméně, v roce 1992, Bernhard Boser, Isabelle Guyon a Vladimir Vapnik navrhli způsob, jak vytvořit nelineární klasifikátory použitím metody jádrové transformace (*Kernel Trick*) na nadrovinu s maximálním rozpětím (*Maximum-Margin Hyperplane*), čímž je každý skalární součin nahrazen nelineární funkcí jádra. To umožňuje algoritmu přizpůsobit maximum-margin nadrovinu v transformovaném prostoru (F*eature Space*). Toto řešení bylo představeno na konferenci COLT-92 (Computational Learning Theory) [30].

V roce 1995, Corinna Cortes a Vladimir Vapnik navrhli upravenou nadrovinu s maximálním rozpětím, která počítá i se špatně označenými příklady (M*islabeled Examples*). Pokud neexistuje nadrovina, která může rozdělit příklady na ty z množiny x

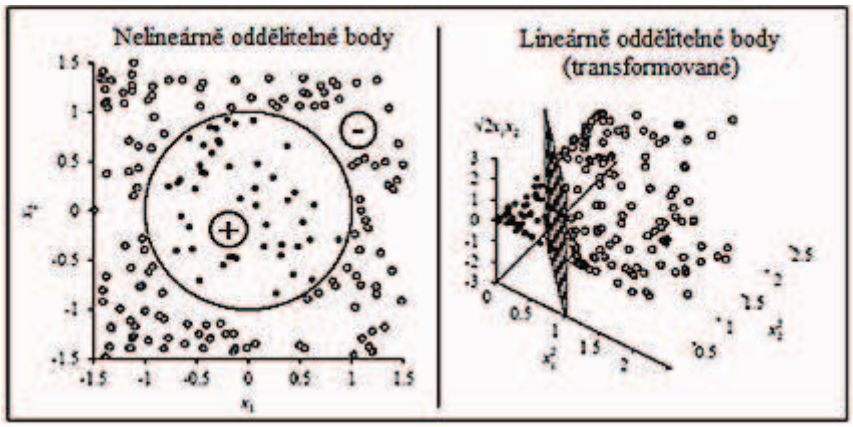
a množiny y, metoda měkkého okraje (*Soft Margin Method*) vybere nadrovinu, která oddělí příklady tak, jak je to jen možné. Přitom je maximalizována vzdálenost k nejbližšímu oddělitelnému příkladu. Tato práce popularizovala výraz Support Vector Machine [29].

SVM verze pro regresy, s názvem Support Vector Regression (SVR), byla navržena v roce 1996 Vapnikem. Na této metodě se společně podílel i Alex Smola, Harris Drucker, Chris Burges a Linda Kaufman [8].

SVM je lineární stroj (L*inear Machine*) založený na metodě SRM (*Structural Risk Minimization*) a na teorii statistického učení. V důsledku toho poskytuje dobré zobecnění problému při rozpoznávání vzorů.

2.2. Základní pohled na SVM

Ze znalosti neuronových sítí víme, že perceptron, nebo složitější jednovrstvá neuronová síť, má omezenou možnost klasifikace. Dochází zde pouze k učení lineárních rozhodovacích hranic ve vstupním prostoru. Oproti tomu vícevrstvé sítě mají mnohem vyšší klasifikační výkon, ale jejich nevýhodou je nesnadná natrénovatelnost, kvůli velkému množství lokálních minim a vysokému rozměru vektoru vah. Řešením tohoto problému jsou sítě Support Vector Machines, nebo obecně řečeno jádrové stroje (*Kernel Machines*), které mají efektivní učící algoritmy, a můžou tak představovat komplexní nelineární hranice. Pomocí nich je řešen problém oddělení bodů, které nejsou lineárně oddělitelné. Základní myšlenkou a jednou z klíčových složek SVM je použití metody s názvem „kernel trick“. Je to způsob, jak vyřešit nelineární problém separace tím, že mapujeme původní nelineárně oddělitelné body do vyšších dimenzionálních prostorů, kde je následně používán lineární klasifikátor. Ačkoli matematické principy této metody jsou téměř jedno století staré, bylo to až mnohem později co byly uplatněny ve strojovém učení. Díky této metodě je lineární zařazení do nových prostorů ekvivalentní k nelineární klasifikaci v původním prostoru. Grafický příklad metody kernel trick je na obr. 4 [9].



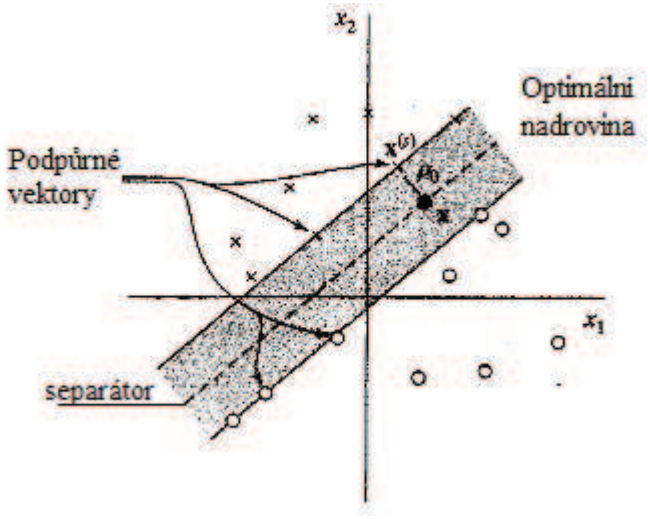
Obr. 4: Použití metody "kernel trick" [9]

Levá část obr. 4 ukazuje dvou-rozměrný vstupní prostor, ve kterém je objekt označen pomocí $x = (x_1, x_2)$.

V tomto případě máme tedy dvě třídy objektů:

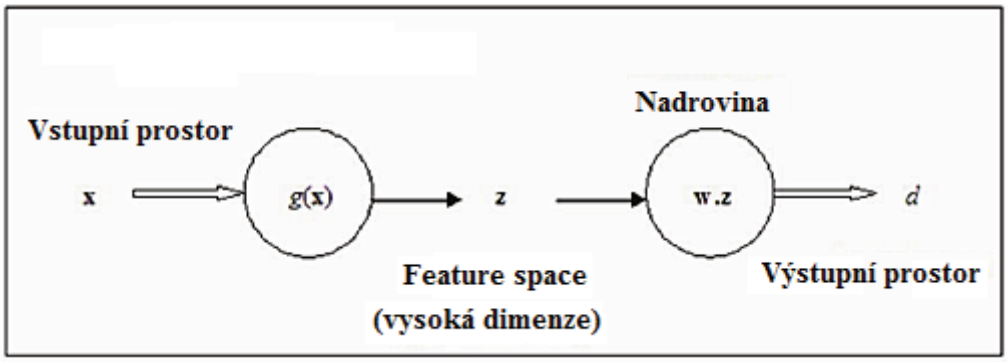
- Pozitivní (+), které jsou umístěny uvnitř kruhu a pro které platí d = +1.
- Negativní (-), které se nacházejí vně kruhu a pro které platí d = -1.

V případech lineárního oddělení množin můžeme sestrojit separační přímku k jejich snadnému oddělení. Když se ale podíváme na levou část obr. 4, lze konstatovat, že lineární oddělení mezi dvěma soubory dat nelze použít. Jedinou možností je převedení tohoto $\mathbb{R}^2$ prostoru (x) na $\mathbb{R}^3$ prostor (z). Obecně lze říci, že převádíme prostor $\mathbb{R}^n \rightarrow \mathbb{R}^m$, kde $m > n$. Tím dostaneme zobrazení bodů uvedené v pravé části na obr. 4. Tuto transformaci provedeme pomocí funkce $z = g(x)$. Nyní již jsou objekty lineárně separovatelné. Pokud bychom promítli tento prostor, dostaneme detailní obraz této metody, znázorněný na obr. 5 níže. Lze na něm vidět optimální nadrovinu s maximálním rozpětím. Body, které se jí nacházejí nejblíže, značí okraj separační oblasti a nazývají se podpůrné vektory (*Support Vectors*).



Obr. 5: Zobrazení separačních přímek a jejich vzdáleností [10]

Na obr. 5 je zobrazeno rozpětí oddělení. Separační přímka se nachází v polovině šířky pásma. Lineární oddělovač (*Linear Separator*) určuje vzdálenost mezi pozitivními a negativními objekty. Rozpětí oddělení měří vzdálenost mezi křivkou separátoru a nejbližším bodem trénovacího souboru. Jestliže je původní prostor mapován do prostoru s dostatečně vysokou dimenzí, pak se takovéto nelineárně oddělitelné objekty stávají lineárně oddělitelné. Vždy je hledán separátor s největší šířkou oddělení.



Obr. 6: Obecné schéma kernel machine [9]

SVM založena na dvou po sobě jdoucích činnostech:

1) nelineární zobrazení prostoru vstupu do vícerozměrného prostoru (*High-Dimensional Feature Space*);
2) konstrukce optimální oddělující nadroviny.

Klíčová myšlenka tvorby SVM je reprezentována použitím vnitřního jádra (*Inner-Product Kernel*) mezi podpůrnými vektory a libovolnými vektory vstupního prostoru, na kterém bude popsán základní koncept vnitřního jádra [9].

Označíme x určité vektory vstupního prostoru a $\mathrm{g(x)} = \{\mathrm{g_j(x)}, \mathrm{j} = 1,2,\dots,\mathrm{m_1}\}$ soubor nelineárních transformací od vstupního prostoru x do prostoru z, kde m_1 je jeho dimenze. K dané lineární transformaci g definujeme nadrovinu, označenou jako rozhodující plochu (*Decision surface*) pomocí vztahu 2.1

$$\sum_{j=1}^{m_1} w_j g_j(x) + b = 0, \tag{2.1}$$

kde $\mathrm{w} = \{\mathrm{w_j}, \mathrm{j} = 1,2,\dots,\mathrm{m_1}\}$ představuje vektor vah spojující feature space z s výstupním prostorem d. Parametr b označuje zkreslení. Když $\mathrm{g_0(x)} \equiv 1$ a $\mathrm{w_0} = \mathrm{b}$ můžeme psát

$$\sum_{j=1}^{m_1} w_j g_j(x) = 0, \tag{2.2}$$

čímž $\mathrm{g(x)} = \left(\mathrm{g_0(x)}, \mathrm{g_1(x)}, \dots, \mathrm{g_{m_1}(x)}\right)$ představuje transformaci vstupního vektoru ve feature space vážený pomocí $\mathrm{w} = (\mathrm{w_0}, \mathrm{w_1}, \dots, \mathrm{w_{m_1}})$ a rozhodující plocha ve feature space je pak dána vztahem

$$wg^T(x) = 0. \tag{2.3}$$

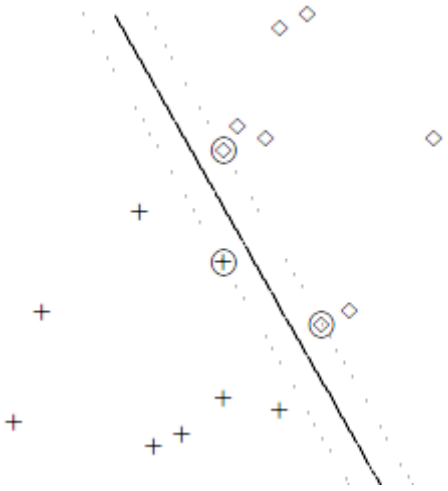
Označením vstupního trénovacího vektoru pomocí $\mathrm{x_i} = 1,2,\dots,\mathrm{N}$, můžeme vytvořit vnitřní jádro definované jako

$$\mathrm{K(x, x_i)} = \mathrm{g(x)g^T(x_i)} = \sum_{\mathrm{j}=1}^{\mathrm{m_1}} \mathrm{g_j(x)g_j(x_i)}, \mathrm{i} = 1,2,\dots,\mathrm{N}. \tag{2.4}$$

Jádro použité při hledání optimálně oddělené nadroviny zdůrazňuje skutečnost, že explicitní znalosti feature space nejsou vůbec zapojeny do tohoto procesu. Jako výpočetně efektivní postup k získání optimální nadroviny je možno použít například kvadratické optimalizace [9].

2.3. Základní principy tvorby a fungování SVM

Nechť je dán soubor trénovacích dat $T = \{x_i, d_i; i = 1,2, \dots, N\}$, kde x_i představuje vstupní vzorek a d_i k němu odpovídající výstup. Prvotním předpokladem lineární oddělitelnost pozitivních vzorů ($d_i = +1$) a negativních vzorů ($d_i = -1$) jak je zobrazeno na obr. 7.



Obr. 7: Lineárně separovatelná data [32]

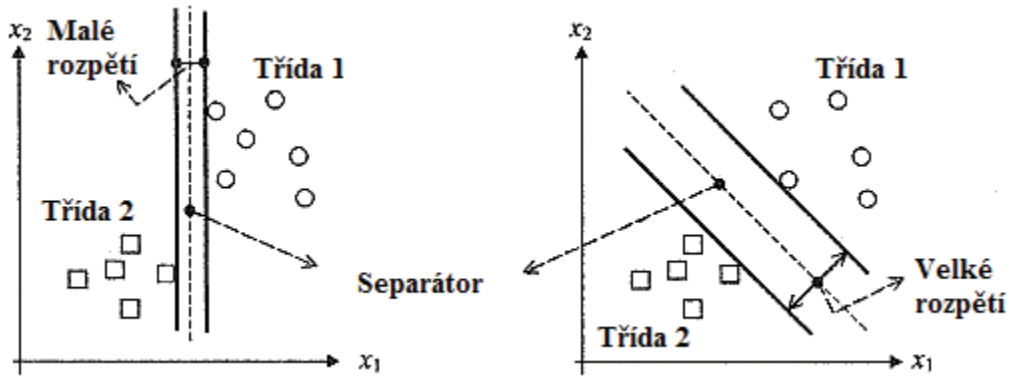
Poté je rovnice oddělující nadroviny dána vztahem 2.5

$$wx^T + b = 0, \tag{2.5}$$

kde *x* je vstupní vektor, *w* jsou nastavitelné váhy, a *b* je zkreslení. Dále víme, že

$$\begin{aligned} wx_i^T + b \geq 0 \text{ když } d_i = +1, \\ wx_i^T + b < 0 \text{ když } d_i = -1. \end{aligned} \tag{2.6}$$

Pro známý vektor vah w a zkreslení b je oddělení dané nadrovinou dáno rovnicí 2.6 a nejbližší místo ze dvou oblastí je právě rozpětí oddělení. Cílem SVM je najít tuto nadrovinu, která maximalizuje tuto vzdálenost. Rozhodující je zde plocha nazvána optimální nadrovinou (*Optimal Hyperplane*).



Obr. 8: Porovnání vzdálenosti přímek od separační přímky [12]

Jak je vidět na předchozím obr. 8, optimální nadrovinou je ta v jeho pravé části, protože má větší rozpětí (*Margin*).

Nechť w_oa b_o označují hodnoty odpovídající optimální nadrovině, určené vztahem

$$w_0 x^T + b_0 = 0, \tag{2.7}$$

potom můžeme diskriminační funkci definovat jako

$$g(x) = w_0 x^T + b_0, \tag{2.8}$$

měřící vzdálenost od vektoru x k optimální nadrovině. Pokud označíme r požadovanou vzdálenost, pak dostaneme vztah

$$g(x) = w_0 x^T + b_0 = r\|w_0\|, \text{nebo} \tag{2.9}$$

$$r = \frac{g(x)}{\|w_0\|}. \tag{2.10}$$

Problém je odhadnout parametry w_0 a b_0, odpovídající optimální nadrovině na základě tréninkového souboru dat. Musí splňovat podmínky:

$$\begin{aligned} w_0 x_i^T + b_0 \geq 1 \; když \; d_i = +1, \\ w_0 x_i^T + b_0 \leq -1 \; když \; d_i = -1. \end{aligned} \tag{2.11}$$

Body (x_i, d_i) definované následující vztahem

$$w_0 x_i^T + b_0 = \pm 1 \tag{2.12}$$

jsou nazývány podpůrné vektory a proto je tato metoda nazvána Support Vector Machine. Tyto vektory jsou umístěny na rozhodovací hranici, oddělující dvě kategorie, a jsou proto velmi těžko zařaditelné. Na jejich základě jsme schopni konkrétně oddělit dvě kategorie [9]. Jak je vidět na obr. 8, hodnota rozpětí oddělení je získaná maximalizací vzdálenosti

$$\rho = \frac{2}{\|w\|}, \tag{2.13}$$

která je ekvivalentní k minimalizaci funkce:

$$L(w) = \frac{\|w\|^2}{2}, \tag{2.14}$$

s omezením:

$$f(x_i) = \begin{cases} +1, když\, wx_i^T + b \geq 1, \\ -1, když\, wx_i^T + b \leq -1. \end{cases} \tag{2.15}$$

Řešení tohoto problému lze získat například pomocí kvadratického programování.

SVM jsou účinné stroje učení (*Learning Machines*), považované za přibližnou implementaci techniky SRM (S*tructural Risk Minimization*), která má kořeny ve Vapnik-Chervonenkis (VC) teorii dimenzí. Rozměr VC stroje učení určuje způsob, jakým je vnořená struktura aproximace funkcí používána v rámci structural risk minimization. Pro používání SRM musíme vytvořit soubor oddělujících nadrovin různých dimenzí VC takovým způsobem, že budou jak trénovací chyba, tak i rozměr VC současně omezovány. VC dimenze je zde definována jako maximální počet vzorků, které lze rozdělit na jakoukoli kombinaci dvou souborů pomocí sady funkcí. Protože pomocí m-rozměrných nadrovin lze oddělit nejvýše m+1 vzorků, tak bude VC dimenze souboru rovna m+1. Tento přístup je založen na Vapnikovu teorému, uvedeném níže [9, 32].

Nechť $\mathbb{R}^n$ je n- rozměrný Euklidovský prostor a $H_\rho$soubor lineárních klasifikátorů, které oddělují $\mathbb{R}^n$ za použití nadrovin s šířkou ρ a nechť $H_{\rho+}$ je soubor lineárních klasifikátorů s tloušťkou větší nebo rovnou ρ. Potom $X_r = \{x_1, x_2, \dots, x_k\} \subset \mathbb{R}^n$ označuje soubor bodů obsažených v oblasti o poloměru r. Rozměr VC parametru $H_{\rho+}$, omezený na X_r, splňuje nerovnost:

$$VC_{dim}(H_{\rho+}) \leq min\left(\left[\frac{4r^2}{\rho^2}\right], n\right) + 1, \tag{2.16}$$

kde $\left[\frac{4r^2}{\rho^2}\right]$ představuje nejbližší celé číslo funkce.

Nechť D je nejmenší průměr obsahující všechny trénovací vektory $x_1, x_2, \ldots, x_N$. VC dimenze hranice pro soubor možných oddělujících nadrovin je dána nerovností

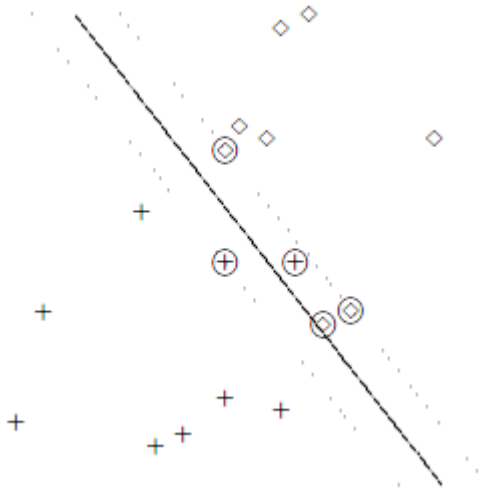
$$h \leq min\left(\frac{D^2}{\rho_0^2}, m_0\right) + 1, \tag{2.17}$$

kde h je dimenze VC, $\rho_0 = \frac{2}{\|w_0\|}$ je rozpětí oddělení mezi dvěma třídami a m_o reprezentuje dimenzi vstupního prostoru.

Tento výsledek nám umožňuje kontrolu rozměru VC h nezávisle na dimenzi vstupního prostoru m_0, je-li rozpětí oddělení vhodně zvoleno. SVM poskytuje metodu pro kontrolu složitosti modelu bez přihlédnutí k rozměrnosti.

2.4. Nelineárně oddělitelné vzory

V reálném světě je velmi málo dat reálně oddělitelných. Tyto případy lze běžně lehko vyřešit. Ale co když se jedná o nelineárně oddělitelná data? V tomto ohledu jsou Support Vector Machines velmi významné. Dokáží základní lineární rámec snadno rozšířit na případ, kdy soubor dat není lineárně oddělitelný. Základní myšlenkou tohoto rozšíření je transformovat vstupní prostor do vyššího dimenzionálního prostoru (*High-Dimensional Space*) nazvaného feature space, kde jsou již data lineárně oddělitelná. Pokud vybíráme tyto transformace pečlivě, všechny výpočty související s tímto převodem mohou být provedeny ve vstupním prostoru. To znamená, že i když transformujeme data vstupního prostoru na lineárně separovatelná, tak se nezvyšuje výpočetní náročnost. Funkce související s těmito přeměnami se nazývá jádrová funkce, a proces tento proces je nazýván jako „kernel trick“.



Obr. 9: Nelineárně separovatelná data [32]

Na rozdíl od předchozí situace, není možné pro tento trénovací soubor údajů vytvořit oddělující nadrovinu bez chybné klasifikace. A tak bereme v potaz soubor dalších nezáporných skalárních proměnných $\{\xi_i, i = 1,2, \dots, N\}$ nazývaných doplňkové proměnné (*Slack Variables*). Ty měří odchylku bodu od ideálního stavu oddělitelnosti vzoru. V tomto případě je cílem minimalizovat nákladovou funkci

$$L(w, \xi) = \frac{\|w\|^2}{2} + C \sum_{i=1}^{N} \xi_i, \tag{2.18}$$

s omezením

$$f(x_i) = \begin{cases} +1, když\, wx_i^T + b \geq 1 - \xi_i, \\ -1, když\, wx_i^T + b \leq -1 + \xi_i, \end{cases} \tag{2.19}$$

kde C je tzv. regularizační parametr, který je volen uživatelem v jednom ze dvou způsobů [9]:

- Experimentální určení, přes standardní použití trénovacího/testovacího souboru dat.
- Analytické určení prostřednictvím odhadu dimenze VC a pak použitím hranice na generalizaci výkonu stroje, založeném na VC dimenzi.

2.5. Typy jádra u SVM

Pro metodu Support Vector Machine existuje mnoho možných jader, která lze použít. Proto je vždy potřeba dopředu vědět jaký problém budeme řešit a výstup, který očekáváme. Z celkového počtu téměř 30-ti je v této kapitole uvedeno 5 základních a nejpoužívanějších typů.

1. Lineární jádro

$$K(x, x_i) = x x_i^T. \tag{2.20}$$

2. Polynomické jádro

$$K(x, x_i) = (\gamma x x_i^T + r)^q, \tag{2.21}$$

kde q označuje stupeň polynomického jádra. Tento parametr je volen uživatelem.

3. RBF jádro

$$K(x, x_i) = exp\left(-\frac{1}{2\sigma^2}\|x - x_i\|^2\right). \tag{2.22}$$

Parametr σ^2 je společný pro všechna jádra a je definován uživatelem.

Zlomek $\frac{1}{2\sigma^2}$ lze také nahradit parametrem γ.

4. Dvou-vrstvý perceptron

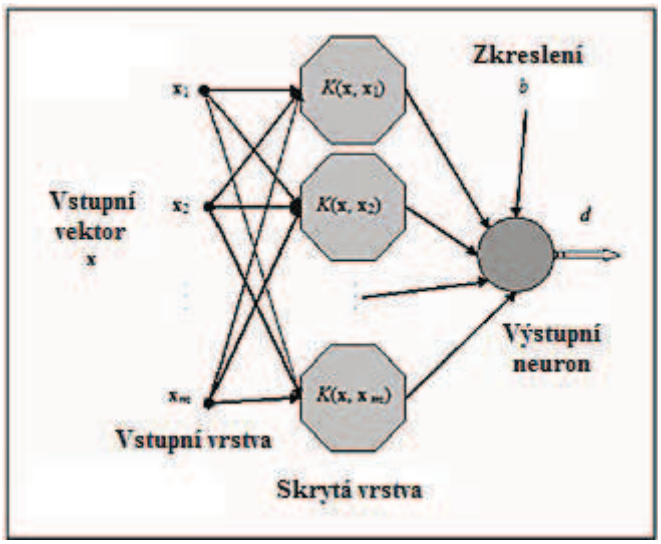
$$K(x, x_i) = \tanh(\beta_0 x x_i^T + \beta_1), \tag{2.23}$$

kde některé hodnoty parametrů β_0a β_1 splňují Mercerův teorém.

5. Sigmoidální jádro

$$K(x, x_i) = \tanh(x x_i^T + r). \tag{2.24}$$

Přestože SVM jsou zvláštním typem lineárních učících strojů, zařazujeme je do oblasti neuronových sítí (NN), protože jejich architektura se řídí stejnou filozofií, jak je možno vidět z následujícího obr. 10.



Obr. 10: Architekrura SVM [9]

2.6. XOR problém

Exkluzivní disjunkce je typ logické operace zahrnující dvojici parametrů (p, q). Tvrzení obsahující XOR problém lze zapsat například ve tvaru „buď …, anebo …“, takže můžeme považovat situaci, kdy parametr p nebo q je pravdivý, ale ne oba současně. Tím zde vzniká problém nelineární oddělitelnost (*Non-Linear Separability*) a proto nemůže být vyřešen jednovrstvou sítí. Jednotlivé vstupy a výstupu pro tento problém jsou zapsány v tab. 2.

Vstupní vektor x	Požadovaný výstup d
(-1,-1)	-1
(-1,+1)	+1
(+1,-1)	-1
(+1,+1)	+1

Tab. 2: Pravdivostní tabulka logické operace XOR

Jestliže označíme $x = (x_1, x_2)$ a $x_i = (x_{i1}, x_{i2})$, pak je vnitřní jádro

$$K(x, x_i) = (xx_i^T + 1)^2, \tag{2.25}$$

vyjádřené jako

$$K(x, x_i) = 1 + x_1^2 x_{i1}^2 + 2x_1 x_2 x_{i1} x_{i2} + 2x_1 x_{i1} + 2x_2 x_{i2}. \quad (2.26)$$

Obraz ze vstupního vektoru x, vytvořeného ve feature space, je dán vztahem

$$z = g(x) = \left(1, x_1^2, \sqrt{2x_1 x_2}, x_2^2, \sqrt{2x_1}, \sqrt{2x_2}\right). \quad (2.27)$$

Podobně tak i

$$z_i = g(x_i) = \left(1, x_{i1}^2, \sqrt{2x_{i1} x_{i2}}, x_{i2}^2, \sqrt{2x_{i1}}, \sqrt{2x_{i2}}\right). \quad (2.28)$$

Označíme-li K matici $\left[K(x_i, x_j)\right]$, pak $K(x_i, x_j)$ označuje vnitřní jádro. Matici K můžeme zapsat následovně

$$K = \begin{bmatrix} 9 & 1 & 1 & 1 \\ 1 & 9 & 1 & 1 \\ 1 & 1 & 9 & 1 \\ 1 & 1 & 1 & 9 \end{bmatrix}. \quad (2.29)$$

Použití kvadratické optimalizace s ohledem na Lagrangeovy multiplikátory, získáme optimální vektor vah

$$w_0 = \left(0{,}0, -\frac{1}{\sqrt{2}}, 0{,}0{,}0\right) \quad (2.30)$$

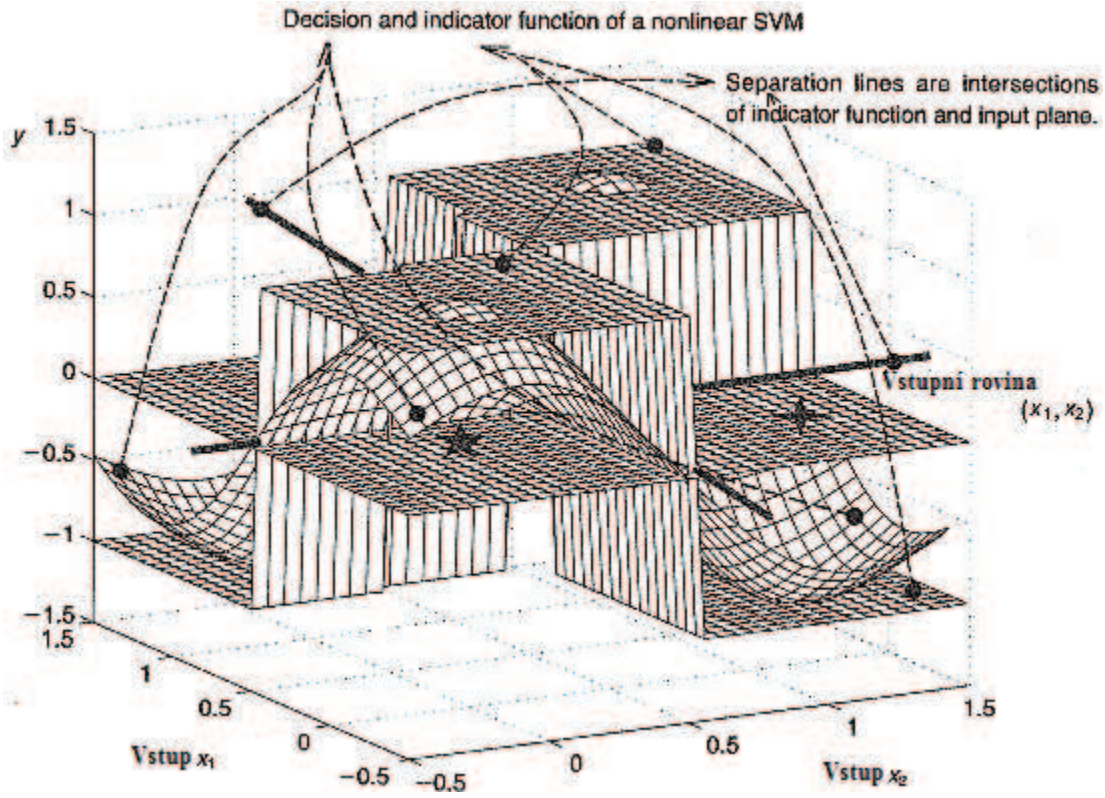
a rovnici optimální nadroviny

$$w_0 g^T(x) = \left(0{,}0, -\frac{1}{\sqrt{2}}, 0{,}0{,}0\right)(1, x_1^2, \sqrt{2x_1 x_2}, x_2^2, \sqrt{2x_1}, \sqrt{2x_2})^T = 0, \quad (2.31)$$

čímž dostaneme

$$x_1 x_2 = 0. \quad (2.32)$$

Na následujícím obr. 11 je uvedeno grafické řešení tohoto problému pomocí nelineárního SVM klasifikátoru s Gaussovou jádrovou funkcí.



Obr. 11: Grafické řešení XOR problému [12]

2.7. Support vector regression

Kromě použití v rozpoznávání vzorů, jsou SVM využívány i pro nelineární regresi, pod názvem Support Vector Regression (SVR).

Je dán nelineární regresní problém, popsaný regresní rovnicí

$$d = f(x) + v, \tag{2.33}$$

kde d je závislá skalární proměnná, x je vektor nezávislé proměnná, f je skalární nelineární funkce definované za podmínek $E[D|x]$, kde D je náhodná proměnná s realizací označenou jako d, a v představuje 'šum'.

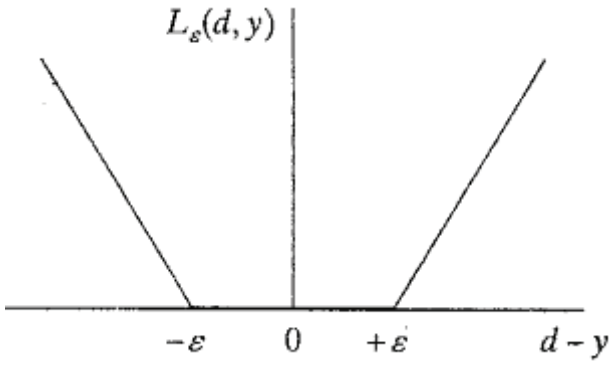
Potom na základě souboru trénovacích dat $T = \{x_i, d_i; i = 1,2, \dots, N\}$ můžeme odhadnout závislost d na x. Označíme-li parametrem y odhad d, a pomocí $g(x) = \{g_j(x), j = 0,1,2, \dots, m_1\}$ soubor nelineárních bázových funkcí, můžeme považovat rozšíření y v rámci $g(x)$ vztahem

$$y = \sum_{j=0}^{m_1} w_j g_j(x) = w g^T(x), \tag{2.34}$$

kde stejně jako dříve, w představuje vektor vah a zkreslení b se rovná váze w_0.

Chceme-li vytvořit SVR pro odhad závislosti d na x, ε-necitlivá ztrátová funkce (*ε-Insensitive Loss Function*), je dána vztahem

$$L_\varepsilon(d, y) = \begin{cases} |d - y|, |d - y| \geq \varepsilon, \\ 0 \quad\quad , |d - y| < \varepsilon. \end{cases} \tag{2.35}$$



Obr. 12: ε-insensitive loss function [10]

To znamená, že problém, který má být vyřešen, se nachází v minimalizaci empirického rizika:

$$R = \frac{1}{N} \sum_{i=1}^{N} L_\varepsilon(d_i y_i), \tag{2.36}$$

za podmínky, že

$$\|w\|^2 \leq c_0, \tag{2.37}$$

kde c_0 značí konstantu.

Vzhledem k tomu, že používáme dvě sady nezáporných doplňkových proměnných $\{\xi_i, i = 1,2, \dots, N\}$ a $\{\acute{\xi}_i, i = 1,2, \dots, N\}$ minimalizujeme nyní účelovou funkci

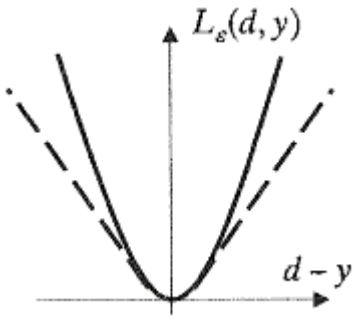
$$L(w) = \frac{\|w\|^2}{2} + C\sum_{i=1}^{N}(\xi_i + \acute{\xi}_i), \tag{2.38}$$

s následujícími omezeními

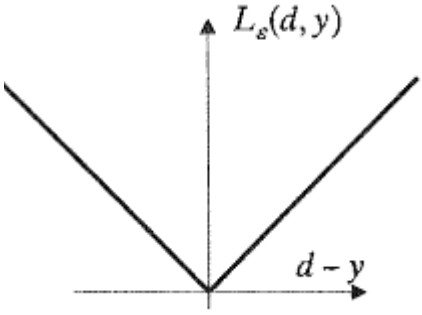
$$\begin{aligned} d_i - wg^T(x_i) &\leq \varepsilon + \xi_i\,, i = 1, \dots, N\,, \\ wg^T - d_i &\leq \varepsilon + \acute{\xi}_i\,, i = 1, \dots, N\,, \\ \xi_i &\geq 0, i = 1, \dots, N\,, \\ \acute{\xi}_i &\geq 0, i = 1, \dots, N\,. \end{aligned} \tag{2.39}$$

Mimo výše zmíněných, existují také jiné algoritmy pro SVM. Kromě výchozího Langrangian Support Vector Machine (LSVM) je to například Finite Newton Langrangian SVM (NLSVM) nebo Finite Newton Support Vector Machine (NSVM).

Kromě Vapnikovy ε-insensitive ztrátové funkce patří mezi další známé například Huberova (obr. 13), kvadratická (obr. 13) a absolutní chyby (obr. 14).



Obr. 13: Kvadratická a Huberova ztrátová funkce [12]



Obr. 14: Ztrátová funkce absolutní chyby [12]

3. Návrh modelu prediktoru ozonové vrstvy

Cílem této kapitoly je navrhnout model pro predikci ozonu. V jednotlivých částech se zaměřím na vlastní tvorbu modelu, postup, jakým byla data upravována a potřebné činnosti pro jejich úpravu před vlastním použitím. Následovat bude postup jejich rozdělení na trénovací a testovací soubor. V poslední části se zaměřím na testování a analýzu výsledků.

SVM patří mezi metody strojového učení. Pokud jde o stroje, je možné říci, že se učí kdykoliv, když se změní jejich struktura, program, nebo data takovým způsobem, že dochází ke zlepšení jejich budoucí výkonnosti. Strojové učení se obvykle vztahuje ke změnám v systémech provádějících úkoly, které související s umělou inteligencí [8].

Support Vector Machines se řadí mezi metody učení s učitelem. Tento typ úloh je založen na principu hledání takové hypotézy, pro kterou je minimální rozdíl mezi výstupními daty a požadovaným výstupem. Podle jeho velikosti jsou pak upraveny nastavení vah, čímž je docíleno snížení této chyby. V případě druhého typu učení, bez učitele, naopak nemáme funkční hodnoty pro množinu trénovacích dat. Ty jsou následně rozdělena do jednotlivých podskupin na základě vzájemné podobnosti. Zde je tedy systém schopen vytvořit, na rozdíl od metody učení s učitelem, výstup pouze na základě použitých vstupních vzorů [10].

3.1. Návrh modelu

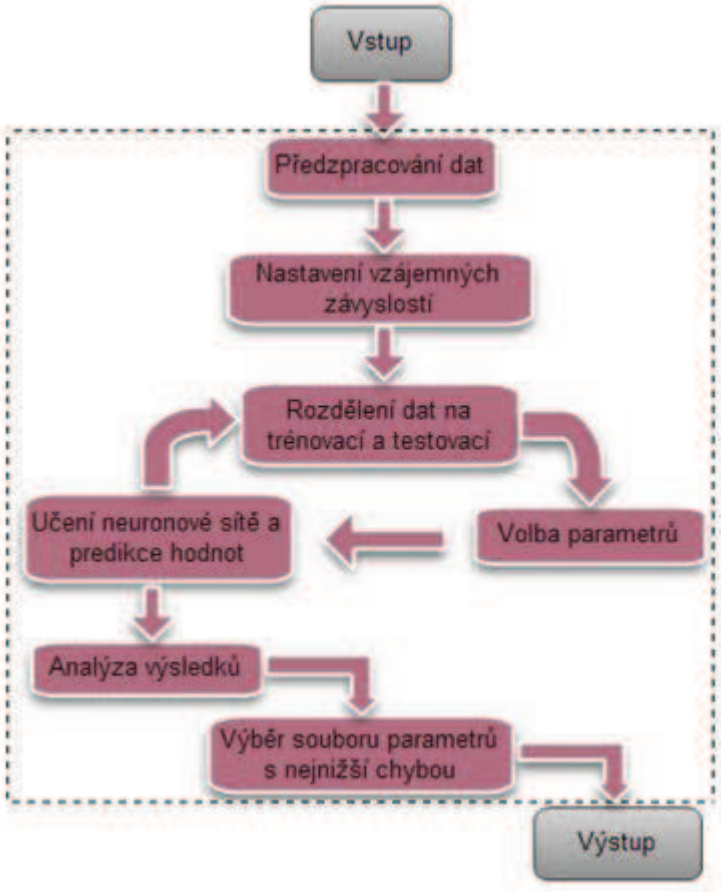
Na obr. 15 je uveden model, použitý na predikci ozonu pomocí SVM neuronových sítí. Na vstup jsou přivedena data týkající se měření hladiny ozonu a dalších látek, které jeho úbytek, nebo přírůstek ovlivňují.

První blok se zabývá předzpracováním dat. Tato část má klíčový význam pro dosažení správných výsledků. Je zde potřeba data upravit z hlediska standardizace, normalizace a určit možné korelované veličiny. Díky těm pak můžeme rozhodnout, které proměnné v souboru ponecháme, a které naopak vyloučíme.

Nastavení vzájemných závislostí tvoří část, kde dochází k definování jednotlivých vztahů mezi daty. Hlavním úkolem je zde jejich rozdělení na nezávislé a závislé proměnné.

Pro získání velikosti chyby učení potřebujeme následující 3 prvky. Prvním z nich je rozdělení dat na trénovací a testovací množinu. Dále je pro správné nastavení modelu potřeba určit hodnoty jednotlivých parametrů následované učením neuronové sítě a s tím spojený výpočet výsledné chyby. Tento proces je opakován s každou změnou poměru rozdělení trénovacích a testovacích dat a jednotlivých parametrů ovlivňujících přesnost učení.

Předposledním prvkem je analýza výsledků, ve které dochází k porovnání naměřených hodnot a určení závislostí mezi jednotlivými parametry. Správným provedením této části získáme celkový obraz o kvalitě jednotlivých modelů, jejich nastavení a velikosti naměřených chyb. Výstupem systému, a tou nejdůležitější částí, je predikovaná hodnota hladiny ozonu závislá na dříve zvolených parametrech.



Obr. 15: Návrh modelu

3.2. Charakteristika vstupních dat

V rámci mé diplomové práce jsem měl k dispozici data naměřená v městské části Dukla v Pardubicích. Data obsahují 679 hodnot pro každý z 16 parametrů ovlivňujících hodnotu ozonu a dalších 7 parametrů obsahujících přímo jeho naměřené hodnoty. Jejich názvy, společně s popisem, jsou uvedeny v tab. 3.

Parametr	Název	Popis	Parametr	Název	Popis
P_1	SO_2	oxid siřičitý	P_{13}	tlak	tlak
P_2	PM_{10}	prachové částice menší než 10 µm	P_{14}	sluneční svit	sluneční svit
P_3	$PM_{2,5}$	prachové částice menší než 2.5 µm	P_{15}	teplota	teplota
P_4	NO_x	oxidy dusíku	P_{16}	směr větru	směr větru
P_5	NO	oxid dusnatý	P_{17}	O_3-1	ozon v čase t-1 dni
P_6	NO_2	oxid dusičitý	P_{18}	O_3-2	ozon v čase t-2 dni
P_7	CO	oxid uhelnatý	P_{19}	O_3-3	ozon v čase t-3 dni
P_8	den	den v týdnu	P_{20}	O_3-4	ozon v čase t-4 dni
P_9	měsíc	měsíc v roce	P_{21}	O_3-5	ozon v čase t-5 dni
P_{10}	pracovní den	pracovní den v týdnu	P_{22}	O_3-6	ozon v čase t-6 dni
P_{11}	rychlost větru	rychlost větru	P_{23}	O_3	hladina ozonu
P_{12}	vlhkost	vlhkost			

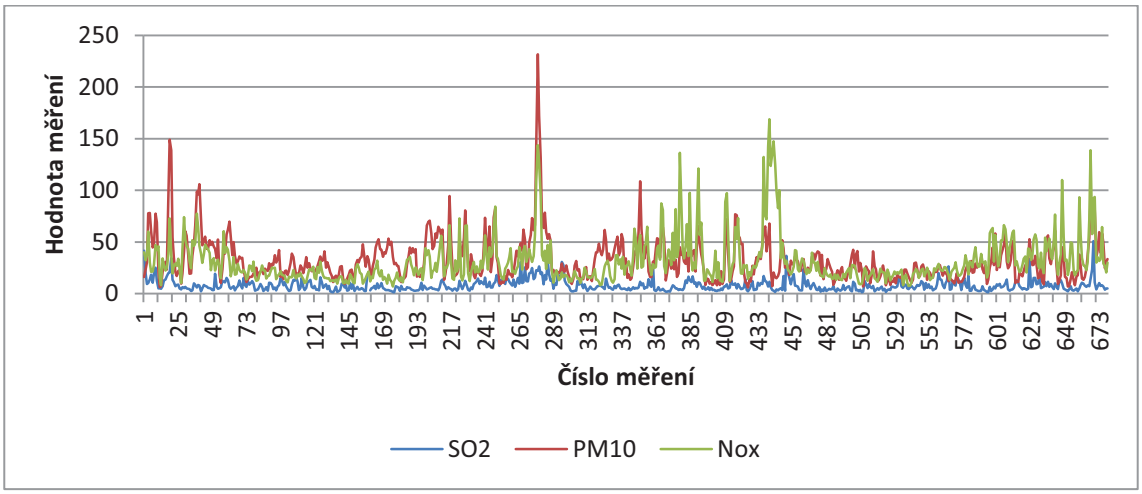
Tab. 3: Přehled parametrů

Tato data jsou s využitím programu Microsoft Excel popsána v následující tab. 4. Ta obsahuje jednotlivé parametry ze vstupního souboru dat a k nim vypočtené základní charakteristiky odrážející jejich vlastnosti.

	Střední hodnota	Medián	Modus	Směr. odchylka	Rozptyl	Špičatost	Šikmost	Rozdíl max-min
SO_2	8.025	6.5	4.2	5.642	31.830	9.821	2.448	49.5
PM_{10}	32.333	26.6	16.2	21.550	464.391	16.293	2.892	225.5
$PM_{2.5}$	20.565	16.8	14.8	14.307	204.699	15.008	2.844	139.3
NO_x	31.135	24.5	15.1	22.274	496.12	9.056	2.658	161.8
NO	6.889	3.7	2.6	8.943	79.975	13.461	3.355	65.9
NO_2	20.087	18.3	19.9	9.457	89.431	3.663	1.454	65
CO	571.685	555.1	658.2	248.089	61548.1	1.149	0.903	1497.8
den	4	4	2	2.001	4.006	-1.250	0	6
měsíc	6.779	7	3	3.257	10.609	-1.141	-0.043	11
pracovní den	0.714	1	1	0.452	0.204	-1.099	-0.951	1
rychlost větru	2.731	2.3	2	1.708	2.916	2.406	1.383	11
vlhkost	76.511	78	82	11.557	133.557	-0.327	-0.388	64
tlak	990.452	990.4	982.7	8.334	69.457	0.197	-0.172	50.2
sluneční svit	4.718	4.4	0	4.095	16.767	-1.353	0.295	12.6
teplota	10.023	108	11.3	7.667	58.777	-0.730	-0.300	35.2
směr větru	17.860	18	26.3	7.885	62.180	-1.228	-0.074	34
O_3	50.4	48.8	56.8	23.7	561.8	-0.6	0.2	121.8

Tab. 4: Popisná statistika dat

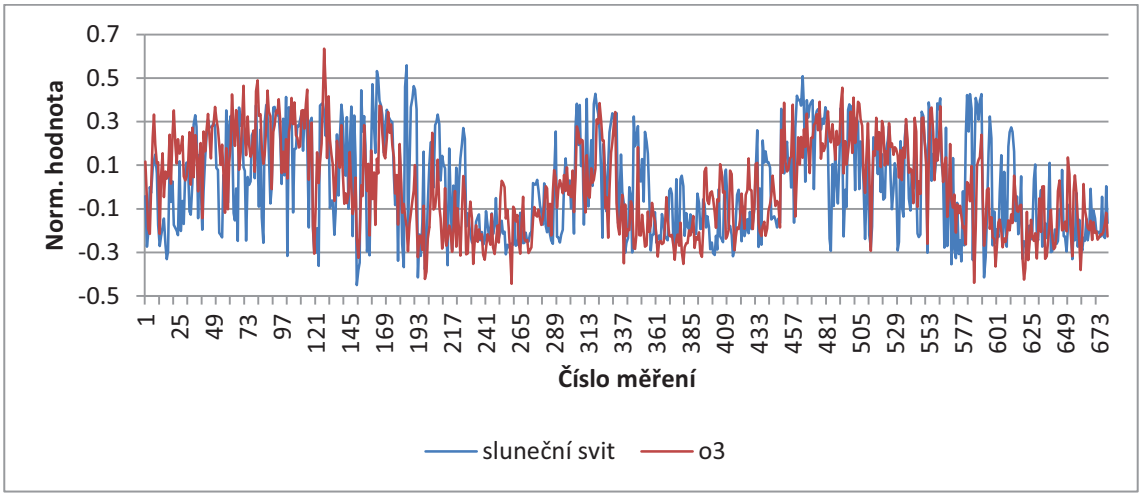
Jak je vidět podle tab. 4, tak se jedná o proměnné s velmi rozmanitými hodnotami. Ty se liší často i v řádu stovek. Na základě toho je pro demonstraci této rozdílnosti uveden následující graf 2, kde jsou srovnány naměřené hodnoty SO_2, PM_{10} a NO_X. Jsou na něm vidět jednak značné velikosti jednotlivých měření, ale patrná je i závislost mezi těmito veličinami.



Graf 2: Zobrazení vztahu mezi SO_2, PM_{10} a NO_X

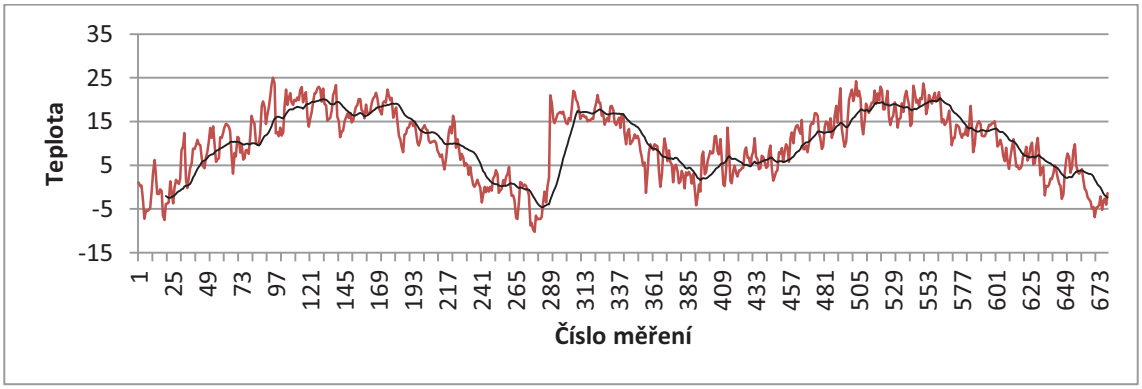
Jistě neméně zajímavým je graf 3, ve kterém je provedeno srovnání množství slunečního svitu s množstvím naměřeného ozonu. V grafu 3 je dobře vidět vzájemný

vliv obou veličin. Jelikož jsou jednotky obou parametrů velmi rozdílné, jsou zobrazeny normalizované hodnoty.



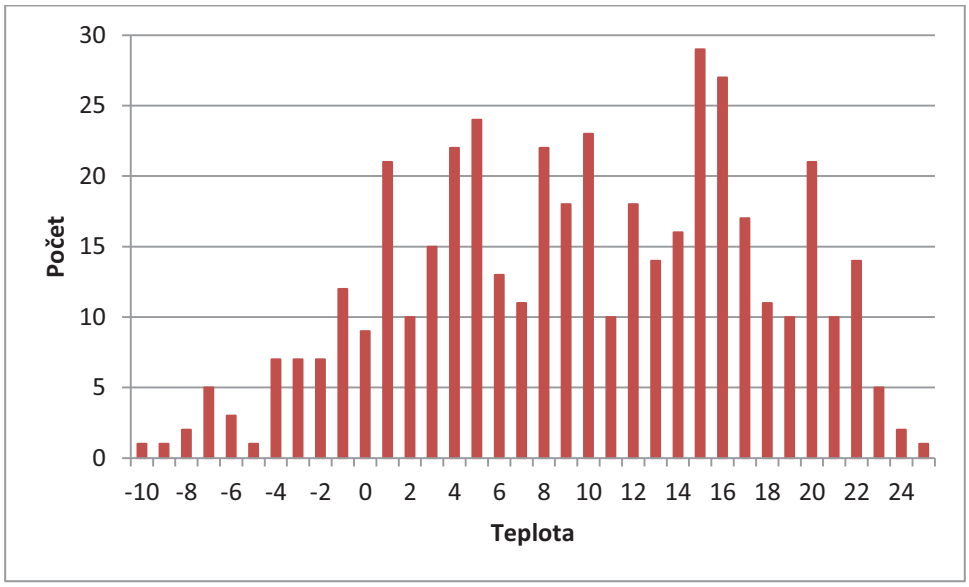
Graf 3: Vzájemná závislost slunečního svitu a hladiny ozonu

Při zkoumání průběhu jednotlivých parametrů mě též zaujaly naměřené velikosti teploty. Je možné z nich poměrně dobře odlišit jednotlivá roční období. Pro lepší názornost je vývoj proložen spojnicí trendu, konstruovanou jako klouzavý průměr s počtem 20 měření.



Graf 4: Vývoj teploty

Pro výše uvedený graf 4 bylo, jako pro jednu z mála proměnných, možné vytvořit histogram hodnot (graf 5). Na jeho základě lze snadno vyčíst nejen minimální a maximální hodnoty v průběhu měření, ale i to, jaké teploty jsou v dané oblasti nejčastější. V tomto případě byla nejvíce krát naměřena teplota 15°C.



Graf 5: Histogram naměřených teplot

3.3. Předzpracování dat

Před samotným začátkem práce s daty bylo zapotřebí udělat několik základních úprav. Jako první z nich jsem provedl standardizaci, protože jednotlivé veličiny byly vyjádřeny v různých jednotkách. Jednalo se tedy o úpravu ve smyslu odstranění dominujících znaků mezi nimi. Soubor dat je možné chápat jako matici dat $M = (m_{ij})$ typu n×p, kde symbol n vyjadřuje počet řádků a symbol p počet sloupců matice. Výpočet standardizace se skládal z následujících 2 kroku:

- Výpočet střední hodnoty a směrodatné odchylky pro každý parametr.
- Přepočet původních hodnot na standardizované.

Pro výpočet střední hodnoty a směrodatné odchylky byly použity vztahy 3.1 a 3.2.

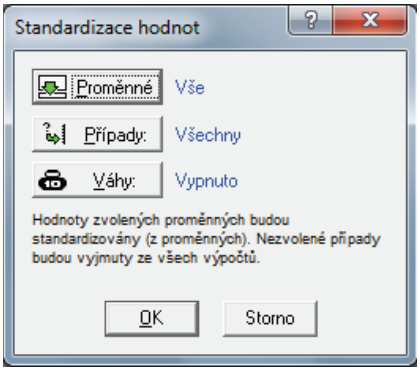
$$\bar{m}_j = \frac{1}{n}\sum_{i=1}^{n} m_{ij} \tag{3.1}$$

$$s_j = \left[\frac{1}{n}\sum_{i=1}^{n}\left(m_{ij} - \bar{z}_j\right)^2\right]^{1/2} \tag{3.2}$$

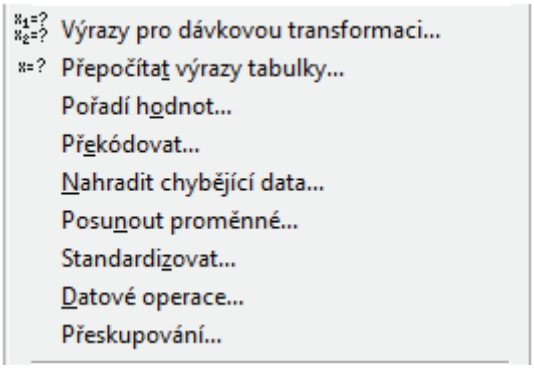
Standardizovaná hodnota je určena pomocí vztahu 3.3.

$$x_{ij} = \frac{m_{ij} - \bar{z}_j}{s_j} \tag{3.3}$$

Jelikož program Statistika umožňuje širokou paletu možností práce s daty, je jednou z nich právě možnost standardizace dat. Na následujícím obr. 16 je zobrazeno okno používané pro tuto standardizaci. Na obr. 17 jsou další možnosti práce s daty.



Obr. 16: Standardizace hodnot



Obr. 17: Další možnosti práce s daty

Dalším krokem v úpravě dat je normalizace. Týká rozdělení vícerozměrných souborů dat a odstranění vlivu jednotlivých proměnných a jednotek měření, což umožňuje porovnání základních charakteristik. Tím je docíleno možnosti srovnávání, protože je použito společné měřítko. Existuje mnoho nožných způsobu normalizace. Mezi nejznámější metody patří Max-Min, lineární, Soft-Max a Z-score transformace. V tomto případě byla použita poslední zmíněná a tím dochází k transformaci dat do nových hodnot se střední hodnotou 0 a směrodatnou odchylkou 1.

Pro vlastní výpočet je potřeba si nejdříve spočítat normu vektoru. Ta v lineární algebře měří jeho délku, která se rovná Euklidově vzdálenosti od koncového bodu tohoto vektoru po začátek vektorového prostoru. Tato vzdálenost je počítána jako druhá odmocnina ze součtu čtverců všech prvků vektoru. Například pro vektor $y = [x_1, x_2, \dots, x_n]$ bude tento součet vyjádřen vztahem

$$\|y\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}. \tag{3.4}$$

Při vlastní normalizaci dělíme každé číslo hodnotou $\|y\|$ a tím dostaneme výslednou hodnotu $\tilde{y}$. Pro kontrolu je možné si ověřit výpočet podle následujícího vztahu

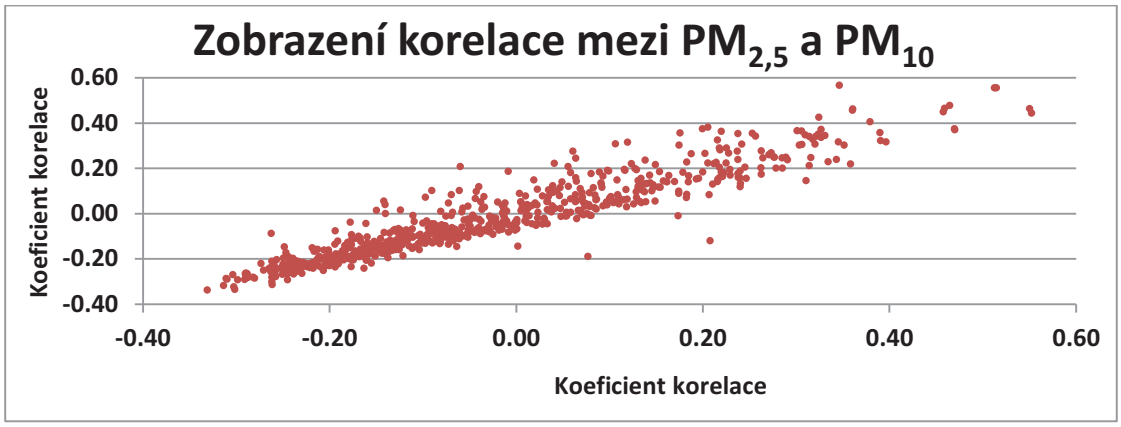
$$\sqrt{\tilde{y}_1^2 + \tilde{y}_2^2 + \cdots + \tilde{y}_n^2} = \sqrt{1}. \tag{3.5}$$

Poslední provedenou úpravou bylo zjištění, zda jsou mezi sebou některé parametry závislé. Tím je dosaženo pomocí Pearsonova korelačního koeficientu, občas také nazývaného jako momentový korelační koeficient. Jedná se o parametrický statistický test, který předpokládá normální rozdělení pravděpodobnosti náhodných veličin. Tento koeficient nabývá hodnot od –1 (negativní závislost) do +1 (pozitivní závislost), přičemž jednička označuje dokonalou lineární závislost. V případě dokonalé nezávislosti proměnných je hodnota r rovna nule. Když mluvíme o kladné korelaci, je tím míněno, že hodnoty obou proměnných současně stoupají. V opačném případě jedna stoupá a druhá klesá, což je stejné i pro zápornou korelaci. Pearsonův korelační koeficient je bezrozměrná veličina, která je nezávislá na jednotkách původních dat.
K výpočtu Pearsonova korelačního koeficientů byl použit vztah 3.6.

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}, \tag{3.6}$$

kde: - x_i, y_i vyjádřují i-té proměnné x, y.
- $\bar{x}$, $\bar{y}$ reprezentují střední hodnoty veličin x, y.
- s_x, s_y představují směrodatné odchylky proměnných x, y.
- n vyjadřuje počet hodnot.

Na následujícím grafu 6 je znázorněna korelace mezi prachovými částicemi menšími než 2.5 μm a prachovými částicemi menšími než 10 μm.



Graf 6: Korelace mezi $PM_{2,5}$ a PM_{10}

Pomocí výše uvedené metody jsem spočítal hodnoty korelace pro všechny kombinace vstupních proměnných a přehled těch nejvýznamnějších z nich je uveden v následující tab. 5.

	PM_{10}	NO_x	den	O_3-1	O_3-2	O_3-3	O_3-4	O_3-5	O_3-6
$PM_{2.5}$	**0.94**	0.566	0.036	-0.232	-0.226	-0.208	-0.172	-0.161	-0.166
NO	0.399	**0.917**	0.101	-0.329	-0.321	-0.3	-0.294	-0.312	-0.4
NO_2	0.552	**0.835**	0.076	-0.283	-0.236	-0.213	-0.207	-0.209	-0.268
pracovní den	-0.042	-0.058	**-0.761**	-0.012	-0.022	-0.034	-0.015	0.039	0.048
O_3-2	-0.179	-0.308	0.016	**0.762**	1				
O_3-3	-0.167	-0.284	0.046	0.637	**0.759**	1			
O_3-4	-0.133	-0.272	0.021	0.586	0.624	**0.76**	1		
O_3-5	-0.119	-0.283	0.005	0.55	0.591	0.631	**0.761**	1	
O_3-6	-0.121	-0.361	-0.035	0.543	0.546	0.583	0.618	**0.756**	1
O_3	-0.097	-0.473	-0.043	0.578	0.549	0.548	0.57	0.63	**0.747**

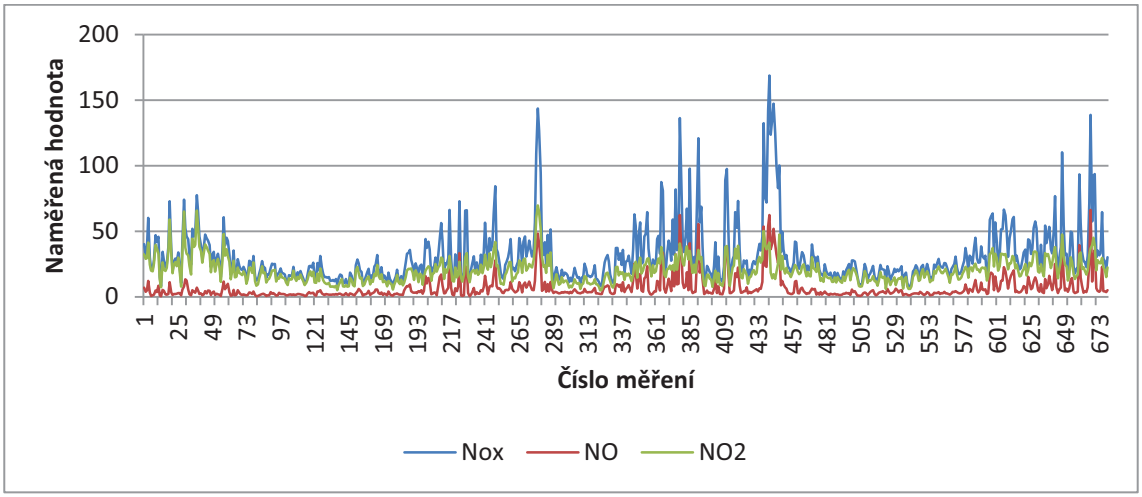
Tab. 5: Nejvýznamnější korelace mezi proměnnými

Na základě výše uvedených hodnot jsem se rozhodl pro vynechání proměnných, které jsou ve vzájemné korelaci a jejichž koeficient r je větší než 0.7.

Když se zaměříme na NO, NO_2 a NO_x tak je na první pohled vidět, že se jedná o velmi korelované veličiny. Jejich závislost je zobrazena na následujícím grafu 7. Hodnoty nejsou normalizovány, protože takto je lépe vidět vzájemná závislost. Na základě tohoto grafu a vypočtených hodnot jsem se rozhodl pro:

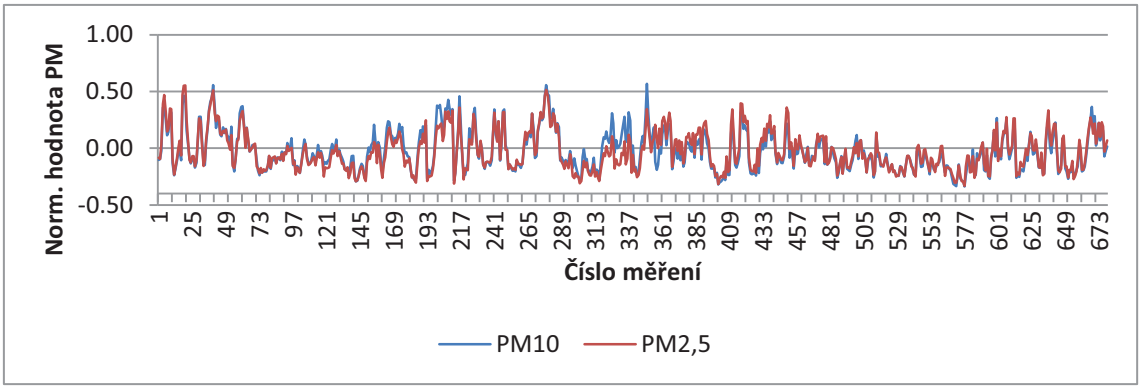
- Ponechání hodnoty NO_x z korelované dvojice NO a NO_x.
- Ponechání hodnoty NOx z korelované dvojice NO_2 a NO_x.

Tímto krokem došlo k nahrazení 3 proměnných jednou a tím pádem i snížení složitosti navrhovaného modelu.



Graf 7: Vzájemné porovnání parametrů NO, NO_2 a NO_X

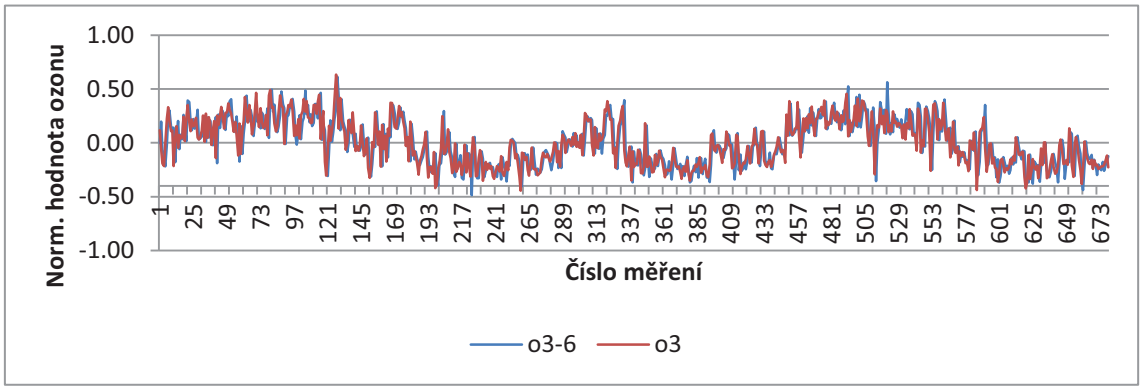
Dalšími veličinami s velikou korelací (0.94) jsou PM_{10} a $PM_{2.5.}$ Jak je vidět na následujícím grafu 8, hodnota 0.94 značí velikou závislost, protože se křivky téměř překrývají. Graf 8 je sestaven na základě standardizovaných a normalizovaných hodnot. Z parametrů PM_{10} a $PM_{2.5}$ jsou pro tvorbu modelu ponechány pouze hodnoty proměnné PM_{10}.



Graf 8: Zobrazení vzájemné korelace mezi parametry $PM_{2,5}$ a PM_{10}

Poslední závislostí, kterou je možné v tab. 5 vidět, je korelace hodnot O_3 až O_3-6. Ta je v rozmezí od 0.747 do 0.762. Pro názorné srovnání jsem zvolil parametry O_3 a O_3-6, jejichž průběh je vidět na grafu 9. Opět v důsledku vysoké korelace dochází k převážnému překrývaní křivek obou proměnných. Na základě výpočtu jednotlivých

závislostí mezi parametry O_3, O_3-1, O_3-2, O_3-3, O_3-4, O_3-5 a O_3-6 jsou pro tvorbu a testování modelu ponechány pouze hodnoty parametru O_3.



Graf 9: Porovnání průběhu měření hodnot O_3 a O_3-6

3.4. Rozdělení dat na testovací a trénovací

V oblasti umělé inteligence často hledáme závislosti mezi daty. Cílem je sestavit takový model, který je obvykle tvořen funkcí s množstvím parametrů, nastavených podle trénování na pozorovaných datech. Potom jsme na základě vstupních hodnot schopni vypočítat ty výstupní. Důležitou vlastností modelu je přesnost. Ta je zjišťována tak, že nejprve dojde k rozdělením dat na trénovací a testovací (v některých případech ještě validační). Na základě trénovacích dat dochází k učení modelu a použití testovacích dat slouží k prověření získaných znalostí a k vyhodnocení jeho přesnosti. S postupným učením dochází k jejímu zlepšování. Na druhou stranu je ale potřeba sledovat chyby modelu, protože v určité fázi může dojít k přeučení a pak by přesnost začala opět klesat. V tom případě je potřeba se vrátit do bodu, kdy byla tato hodnota největší a v tu chvíli trénování ukončit. Přetrénování lze také v určitých případech zabránit snížením počtu parametrů.

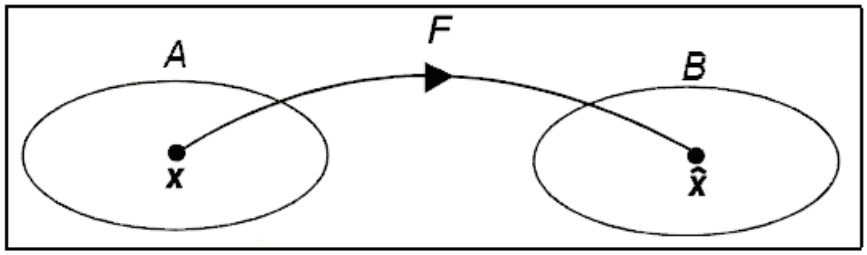
Pro následující rozklad množiny objektů A a B na tréninkovou a testovací množinu bude zavedeno označení A_{train} a A_{test} pro objekt A, a B_{train} a B_{test} pro objekt B, kde

$$A=A_{train} \cup A_{test} \text{ a } B=B_{train} \cup B_{test} . \qquad (3.7)$$

Všeobecná formulace klasifikačního problému [13] lze vyjádřit pomocí pojmu zobrazení, který značí funkci definovanou nad dvěma množinami A a B. Díky tomu, lze

neuronové sítě použít jako klasifikátory nebo prediktory. Nechť je *F(x)* funkce definovaná nad množinou, která přiřadí každému elementu $x \in A$ funkční hodnotu z množiny B, $\hat{x} = F(x) \in B$. Tato závislost je definována následujícím vztahem 3.8 a graficky vyobrazena na obr. 18 uvedeném níže.

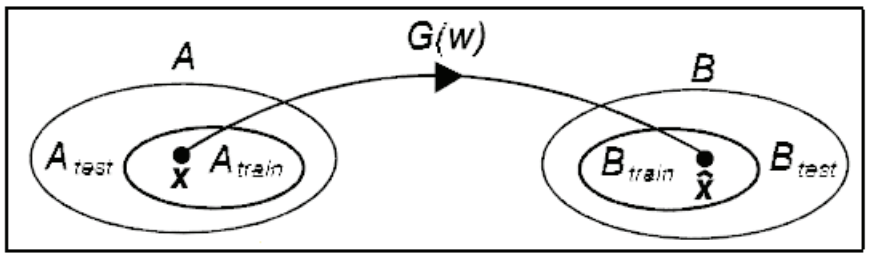
$$F: A \rightarrow B \tag{3.8}$$



Obr. 18: Grafické zobrazení klasifikačního problému [13]

Nechť *G(x,w)* je funkce, jejíž argumenty jsou z konečné podmnožiny $A_{train} = \{x_1, x_2, \ldots . x_r\} \subset$ A a *w* je parametr použitý pro zobrazení *G*, pak $\hat{x} = G(x, w) \in B_{train} \subset$ B. Tuto skutečnost popisuje níže uvedený vztah 3.9 spolu s grafickým vyjádřením na obr. 19.

$$G(w): A_{train} \rightarrow B_{train} \tag{3.9}$$



Obr. 19: Zobrazení funkce G(w) [13]

Zobrazení *G(w)* můžeme nazvat restrikcí zobrazení *F(x)* nad množinou $A_{train} \subset A$. Doplňkem množiny A_{train} vzhledem k množině *A* je A_{test}, která odpovídá podílu $A \backslash A_{train}$. Předpokládejme, že pro každé $x_i \in A_{train}$ poznáme požadovanou funkční hodnotu $\hat{x}_i$. Například pro hodnotu x_1 dokážeme určit hodnotu $\hat{x}_1$. Požadované funkční hodnoty $\hat{x}_i$ jsou interpretovány jako obrazy funkce *F*.

$$\hat{x}_i = F(x_i) \;\; pro\; i = 1, 2, \ldots, r \tag{3.10}$$

Cílem je tedy nalézt parametr *w* funkce *G(x,w)* tak, aby funkční hodnoty argumentů z tréninkové množiny A_{train} byly co nejblíže požadovaným hodnotám.

3.5. Použití metody SVM

Jádrové funkce

Program Statistica podporuje při použití modelů Support Vector Machines celou řadu jader. Patří mezi ně lineární, polynomiální, sigmoidální jádra a jádro radiální bázové funkce (RBF). Tyto typy byly již vysvětleny v předcházející kapitole. Mimo jiné ještě existuje mnoho dalších typů jako například exponenciální, Laplaceovo, multikvadratické, Bayesovské, atd.

Regrese

Učení modelu SVM na trénovacích datech je proces, který zahrnuje, stejně jako klasifikace, postupnou optimalizaci ztrátové funkce. V závislosti na definici této funkce existují dva typy SVM modelů:

Ztrátová funkce typu 1

$$\phi\left(w, \xi, \dot{\xi}\right) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i + C \sum_{i=1}^{N} \dot{\xi}_i, \tag{3.11}$$

za podmínek že $\xi_i, \dot{\xi}_i \geq 0, i = 1, \ldots, N.$

Ztrátová funkce typu 2:

$$\phi\left(w, \xi, \dot{\xi}\right) = \frac{1}{2} w^T w - C \left(\nu\varepsilon + \frac{1}{N} \sum_{i=1}^{N} \left(\xi_i + \dot{\xi}_i\right) \right), \tag{3.12}$$

za podmínek že $\xi_i, \dot{\xi}_i \geq 0, i = 1, \ldots, N, \varepsilon \geq 0,$

kde parametr *w* představuje váhy

ξ,$\dot{\xi}$ doplňkové proměnné popisující ztrátovou funkci

ε označuje parametr necitlivosti

3.6. Volba parametrů

Při testování modelu máme na výběr z různých možností nastavení. V první řadě je zapotřebí nastavit nezávislé a závislé proměnné. V tomto případě je jedinou závislou proměnnou hodnota ozonu nazvaná O_3. Dalším možným nastavením, které software umožnuje, je rozdělení dat na trénovací a testovací s výběrem jedné ze 3 možností použití:

1) Náhodný výběr. Zde je možno nastavit procentuální podíl trénovacích dat a počet seedů.
2) Vzorová proměnná.
3) Prvních N hodnot.

Podle použitého typu regrese se opět mění možnost zadání parametrů. Při použitém typu 1 regresní funkce (vztah 3.11) je dále potřeba zadat parametry C a ε. V případě, že se rozhodneme pro typ 2 (vztah 3.12), zadáváme parametr C a ν. Následující nastavení jsou již závislá na typu jádra, které je zvoleno. Je možné vybrat si z následujících 4 typů:

Lineární jádro

V případě použití lineárního jádra není možno nastavit žádné další parametry.

Polynomické jádro

Při zvolení této možnosti můžeme pro výpočet modelu volit ještě hodnoty stupně polynomického jádra (q), parametr gama (γ) a hodnotu koeficientu (r).

RBF jádro

Když se rozhodneme pro možnost využití RBF jádra, tak je potřeba nastavit hodnotu parametru gama (γ).

Sigmoidální jádro

Poslední variantou je použití sigmoidálního jádra. Zde se vyplňují hodnoty parametru gama (γ) a koeficientu (r).

Program Statistika dále nabízí využití možnost křížové validace. Po zaškrtnutí „Apply v-fold cross-validation“ jsme vyzváni k zadání hodnoty V a počtu seedů. Dále zde můžeme omezit hodnotu parametru C a ε. Pro každý z nich je možno nastavit minimum, maximum a velikost přírůstku.

Poslední důležitou volbu, kterou program nabízí před započetím vlastních výpočtů je nastavení maximálního počtu iterací a velikosti chyby. Tato kombinace umožňuje vytvořit dostatečně naučený model, jehož doba učení je omezena buď maximálním počtem iterací, nebo přesností. Jeho učení je opakovaný proces při kterém je cílem minimalizace funkce chyby. Čím větší provedeme počet iterací, tím více je model naučen. Volba parametru „Maximální počet iterací“ určuje horní hranici počtu iterací, kterými je procházeno. Pokud je ale dosaženo předem nastavené hodnoty chyby, tak považujeme model za dostatečně naučený, a proces je ukončen, i když tohoto počtu nebylo dosaženo.

4. Analýza výsledků

Pro vytvoření modelu bylo zvoleno softwarové prostředí programu Statistika od společnosti Statsoft. Testování parametrů probíhalo formou změny jeho jednotlivých parametrů. Mezi základní rozdíly modelů patří typ použité jádrové funkce. Pro účely této diplomové práce byly vybrány 3 typy. Prvním z nich je lineární jádrová funkce, druhou polynomická a poslední funkcí je RBF. Každá se liší počtem jednotlivých parametrů. U lineární funkce byly použity pouze 2 parametry, na rozdíl od polynomické, kde jich bylo 5. Celkově docházelo ke změně hodnot parametrů C a ε, stupně polynomického jádra (q), parametru gama (γ) a hodnoty koeficientu (r). Testy byly prováděny pro různá rozdělení dat na trénovací a testovací množinu. V rámci učení modelu byly použity poměry dat testovacích vůči trénovacím 50:50, 60:40, 70:30, 80:20, 90:10.

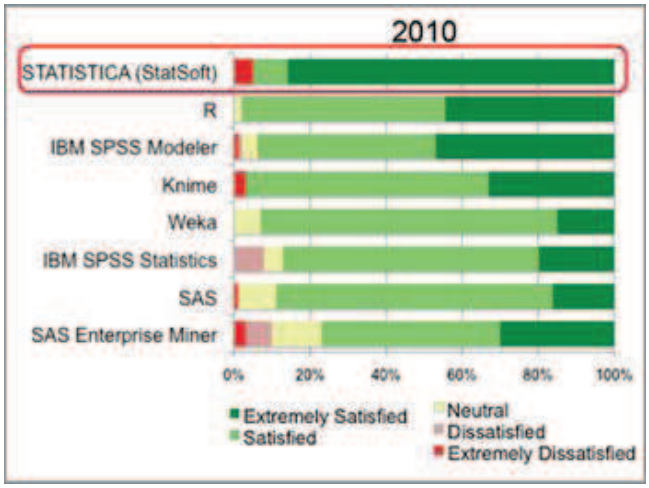
Cílem této kapitoly je analyzovat navržené modely a podle velikosti jejich chyby z nich vybrat ten nejvhodnější. Pro jejich porovnávání slouží hodnoty MSE a RMSE. Chybu MSE (*Mean Squared Error*), je možné považovat za kritérium pro výběr vhodného odhadu. Rovná se totiž součtu rozptylu a druhé mocniny vychýleného odhadu. Nikdy není možné docílit nulové hodnoty Mean Square Error, protože potom by se jednalo o ideální model s dokonalou přesností. Druhá možnost porovnání modelů je podle RMSE (*Root Mean Squared Error)*. Tato chyba je vypočítána jako druhá odmocnina MSE a udává míru přesnosti modelu. Pro hodnocení v této diplomové práci byla vybrána u všech modelů hodnota RMSE.

4.1. Použité softwarové prostředí

Pro testování jsem použil program od společnosti StatSoft, která byla založena v roce 1984 a je jedním z největších producentů softwaru použitelného pro analýzu dat, data mining a řízení kvality. Její produkty jsou používány po celém světě na většině hlavních univerzit, podniků, vládních agentur a jsou podporovány vzdělávací a poradenské služby prostřednictvím celosvětové sítě poboček. Ty jsou umístěny ve 24 zemích na všech kontinentech. Společnost je mimo jiné softwarový partner společnosti Intel a jejího Intel® Software Partner Programu, v rámci kterého byla vyvinuta

technologie, kterou využívá architektura Intel CPU a dodává tím vyšší výkon při paralelním zpracování dat.

Program Statistica byl zařazen analytickými odborníky v nezávislých průzkumech mezi ty s nejvyšší spokojeností a nejsnazším ovládáním. STATISTICA Data Miner a STATISTICA Text Miner obdrželi v roce 2010 nejvyšší průměrné hodnocení, včetně nejvyššího ratingu jako primární prediktivní analytický nástroj a také získaly ocenění nejvyšší spokojenosti uživatelů ze všech konkurenčních softwarových produktů. Následující obr. 20 ukazuje porovnání vydané v únoru roku 2011 společností Rexer Analytics v rámci 4. ročníku Rexer Analytics Data Miner Survey.



Obr. 20: Porovnání programů na základě průzkumu. [28]

4.2. Načtení dat

Načtení dat do programu lze provést několika způsoby:

- Importem souboru podporovaného formátu uloženého v počítači, popřípadě na externím médiu,
- Připojení k databázi – Oracle, MS SQL Server, atd.,
- Zkopírováním a vložením dat do nové tabulky,

- Otevřením souboru MS Excel přímo, bez importu,
- Načítáním dat přímo z měřicího přístroje.

4.3. Volba parametrů

V průběhu testování byly měněny hodnoty v následujících mezích:

- parametr $C \in \langle 1, 15 \rangle, \Delta C = 1.$
- parametr $\varepsilon \in \langle 0.1, 1 \rangle, \Delta \varepsilon = 0.1.$
- stupeň polynomického jádra $q \in \langle 1, 5 \rangle, \Delta q = 1.$
- parametr $\gamma \in \langle 0.1, 3 \rangle, \Delta \varepsilon = 0.001.$
- hodnota koeficientu $r \in \langle 0, 5 \rangle, \Delta r = 1.$

Pro každý model bylo nastaveno 600 iterací algoritmu se zastavením učení při dosažení trénovací chyby 0.001.

4.4. Model s lineární jádrovou funkcí

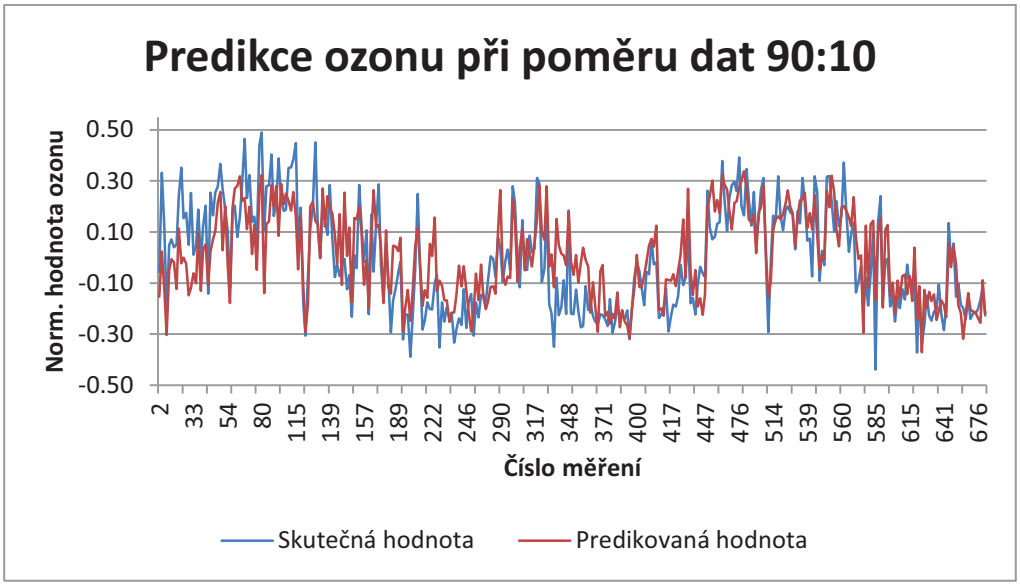
Pro tento typ jádra je charakteristické, že nastavujeme pouze 2 parametry. Těmi jsou C a ε. V následující tab. 6 jsou zobrazeny výsledky jednotlivých chyb pro různé poměry trénovací a testovací množiny. Nejlepší výsledky vykazoval model pro stejný poměr dat v obou množinách, tedy 50:50, hodnotou parametr C rovnou 1 a parametru ε = 0.1. Všimněme si, že ačkoli byl měněn poměr dat, tak docházelo ke změně hodnoty pouze v řádu setin. Výsledky testování modelů pro lineární jádrovou funkci zobrazuje tab. 6.

C	ε	RMSE	$O_{train} : O_{test}$
1	0.1	**0.122863**	50:50
4	0.2	0.126025	60:40
4	0.2	0.125038	70:30
11	0.3	0.125633	80:20
10	0.4	0.125291	90:10

Tab. 6: Zobrazení výsledných hodnot RMSE pro model s lineární jádrovou funkcí

Na následujícím grafu 10 je zobrazen vývoj predikce ozonu při použití lineární jádrové funkce. Pro demonstraci byl použit poměr trénovacích a testovacích dat 90:10 s

ostatními nastaveními modelu zobrazenými v předchozí tab. 6. I když se v tomto případě jedná o chybu modelu o hodnotě pouze 0.12591, tak jsou na grafu vidět občasné velmi velké rozdíly mezi naměřenými a predikovanými daty.



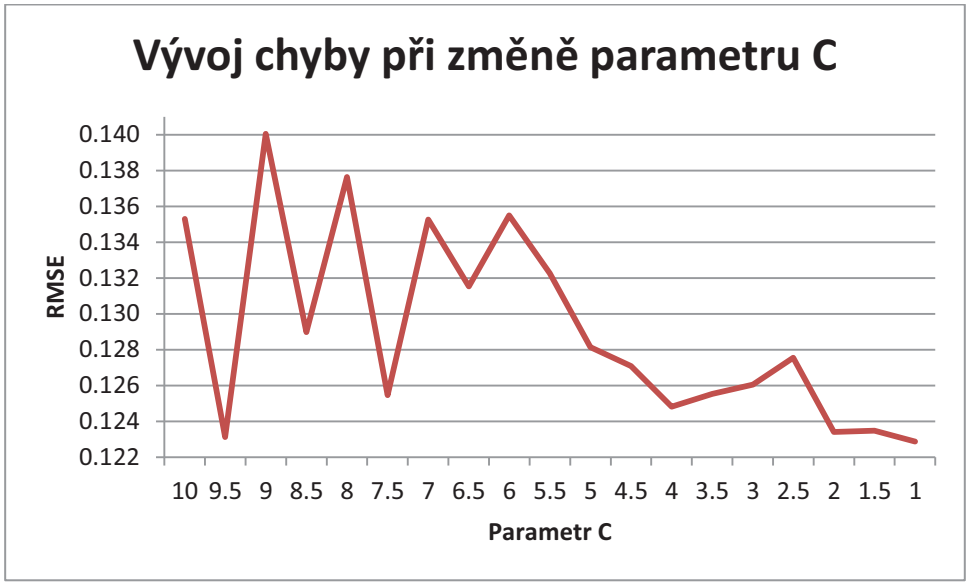
Graf 10: Porovnání skutečné hodnoty a hodnoty predikované modelem s lineární jádrovou funkcí

Při testování modelů změnou parametrů docházelo ke značným kolísáním velikosti chyby v závislosti na jednotlivých parametrech. U parametru C platilo v převážné většině případů pravidlo, že čím byla jeho hodnota menší, tím klesala i chyba modelu. I když je pravdou, že i zde se našly výjimky. Vývoj chyby v závislosti na změnách parametrů je zobrazen v následující tab. 7, pro kterou byly vybrány hodnoty pro poměr trénovacích dat vůči testovacím 90:10.

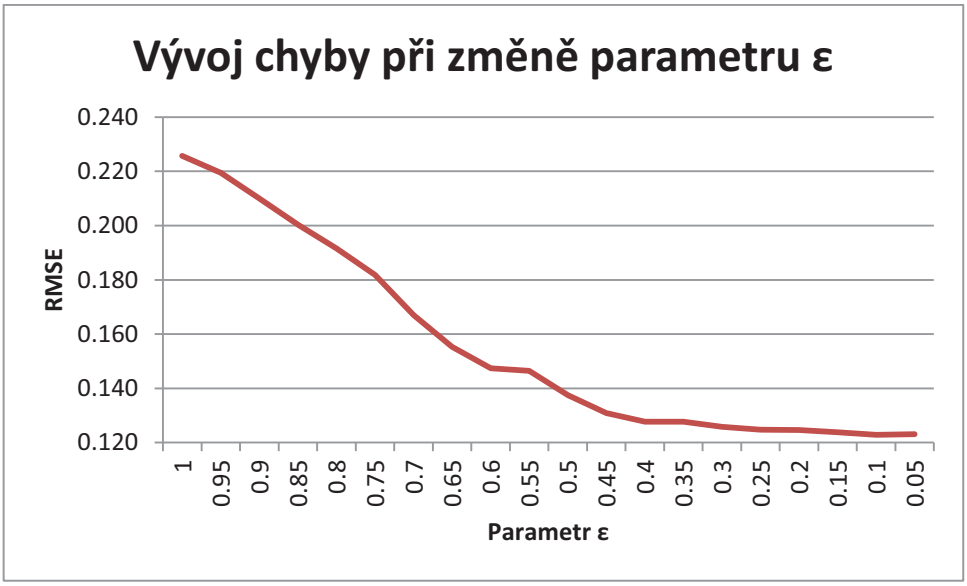
C	ε	RMSE	$O_{train} : O_{test}$
5	0.1	0.131911	90:10
1	0.2	0.130733	90:10
2	0.2	0.129638	90:10
12	0.2	0.128337	90:10
2	0.3	0.128326	90:10
3	0.3	0.127875	90:10
6	0.3	0.125712	90:10
10	0.4	**0.125291**	90:10

Tab. 7: Vývoj chyby pro různé parametry modelu s lineární jádrovou funkcí

Pro názornou ukázku vývoje chyby tohoto modelu slouží následující graf 11. Ten ukazuje vývoj RMSE chyby v průběhu změny parametru C při konstantním parametru ε. Jak je z průběhu křivky vidět, tak mezi těmito veličinami není žádná zřejmá závislost. Jedinou pozorovanou vlastností je snižující se hodnota chyby při postupném snižování hodnoty C pod úroveň čísla 6. Tato měření probíhala při konstantní velikosti parametru ε=0.1.



Graf 11: Vývoj chyby při změně parametru C u modelu s lineární jádrovou funkcí



Graf 12: Vývoj chyby při změně parametru ε u modelu s lineární jádrovou funkcí

Předchozí graf 12 zobrazuje testování, které bylo provedeno pro změnu hodnoty ε při konstantní velikosti C. Nejvyšší hodnotu RMSE byla vykazována pro ε=1. S postupným snižováním o Δε=0.05 docházelo také k jejímu snižování, které skončilo na hodnotě 0.122863.

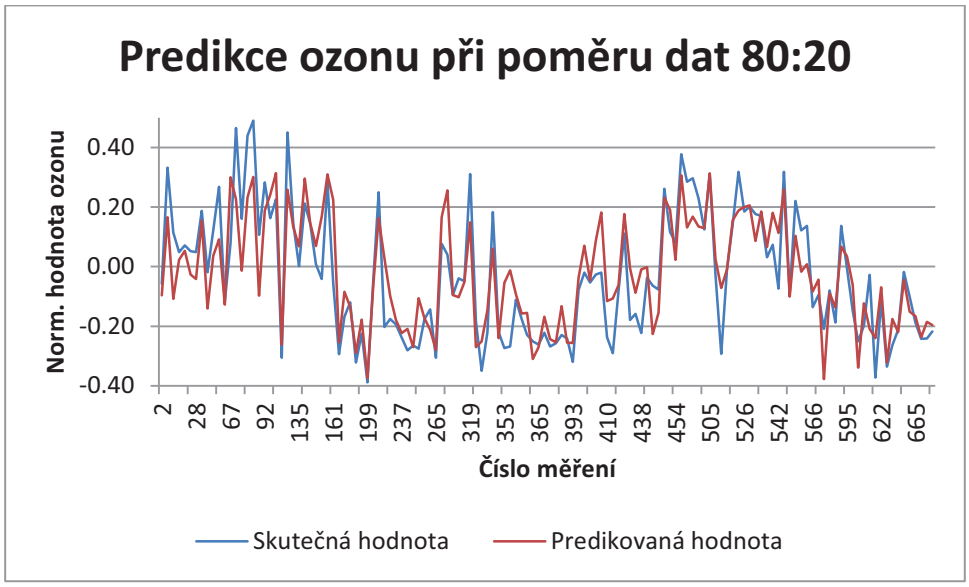
4.5. Model polynomickou jádrovou funkcí

Tento model je ze všech zmíněných nejnáročnější na testování, protože je zde potřeba nastavovat všech 5 parametrů. Na druhou stranu vykazuje oproti modelu s lineární jádrovou funkcí menší velikost chyby. Jak je vidět v následující tab. 8, nejmenší hodnota RMSE je pro poměr trénovacích a testovacích dat 80:20. Zajímavostí je zde hodnota parametru C, která je vůči těm ostatním neobvykle vysoká. Naopak stereotypní se zde jeví velikost parametru ε, který při hodnotě 2 zajišťoval nejlepší vlastnosti modelu. Při testování velikosti koeficientu r v rozmezí hodnot 0 až 5, byla RMSE chyba ve všech případech nejmenší pro hodnotu rovnou 0.

C	ε	q	γ	r	RMSE test	$O_{train} : O_{test}$
1	0.1	2	0.835	0	0.111159	50:50
1	0.1	2	0.71	0	0.113593	60:40
1	0.1	3	0.502	0	0.110675	70:30
11	0.1	2	0.283	0	**0.108130**	80:20
2	0.2	2	0.8	0	0.112030	90:10

Tab. 8: Zobrazení výsledných hodnot RMSE pro model s polynomickou jádrovou funkcí

Pro zobrazení vývoje chyby tohoto modelu jsem vybral poměr trénovacích a testovacích dat 80:20. Ostatní parametry jsou zapsány v předchozí tab. 8. Jak je vidět na následujícím grafu 13, jediný větší rozdíl mezi skutečnými a predikovanými hodnotami je mezi hodnotami 67 a 92 na ose x, jinak dochází téměř ke shodě mezi vykreslenými křivkami.



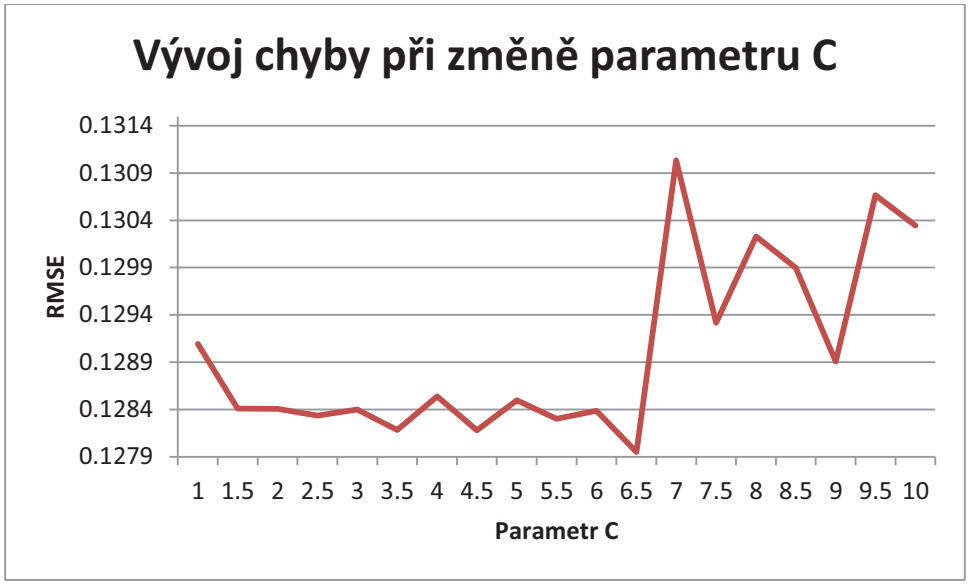
Graf 13: Porovnání skutečné a predikované hodnoty modelu s polynomickou jádrovou funkcí

Jelikož tato jádrová funkce obsahuje mnoho parametrů, tak je její chyba velmi ovlivňována ať nepatrnou změnou každého z nich. Jak je vidět v následující tab. 9, pro nalezení dalších nižších a nejnižší hodnoty chyby modelu bylo potřeba měnit současně i několik parametrů současně, protože s téměř stejnou hodnotou je q. V jeho případě měla převážná část modelů nejnižší chybu pro hodnotu 2. Pro demonstraci vývoje učení modelu je zde použit poměr trénovacích a testovacích dat 60:40 s nejnižší hodnotou chyby RMSE 0.113593.

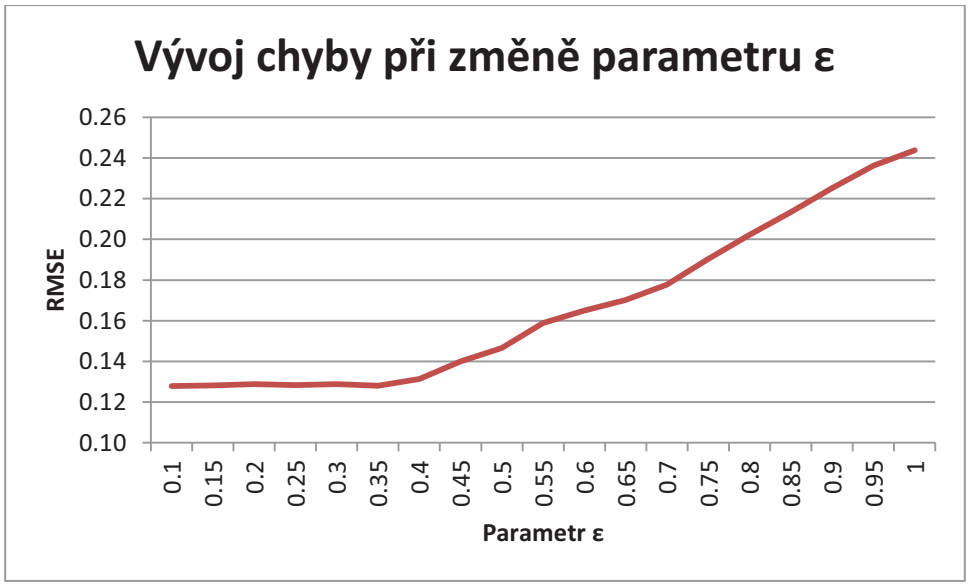
C	ε	q	γ	r	RMSE test	$O_{train} : O_{test}$
10	1	2	1	0	0.230015	60:40
10	0.3	2	1	0	0.126061	60:40
1	0.1	2	0.7	2	0.124264	60:40
4	0.2	2	1	0	0.122284	60:40
1	0.1	3	0.7	0	0.121552	60:40
4	0.3	2	1	0	0.121127	60:40
2	0.3	2	1	0	0.119009	60:40
4	0.2	2	1.1	0	0.117438	60:40
1	0.1	2	0.5	0	0.116868	60:40
1	0.1	2	0.8	0	0.116836	60:40
1	0.1	2	0.71	0	**0.113593**	60:40

Tab. 9: Vývoj chyby pro různé parametry modelu s polynomickou jádrovou funkcí

Prvním parametrem, který je v tomto modelu měněn, je proměnná C. Jak je vidět z následujícího grafu 14, vývoj chyby z počátku téměř nereagoval na změnu hodnoty, což se změnilo od C=6.5, kde bylo zaznamenáno minimum křivky. Ta pak začala prudce stoupat. Testování končí při C=10 s RMSE chybou 0.130345.



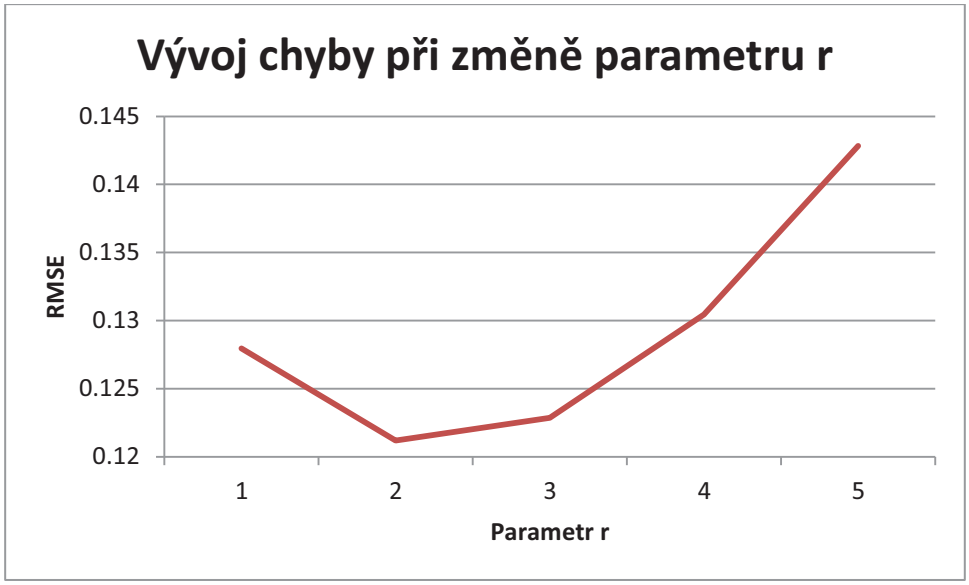
Graf 14: Vývoj RMSE chyby při změně parametru C u modelu s polynomickou jádrovou funkcí



Graf 15: Vývoj RMSE chyby při změně parametru ε u modelu s polynomickou jádrovou funkcí

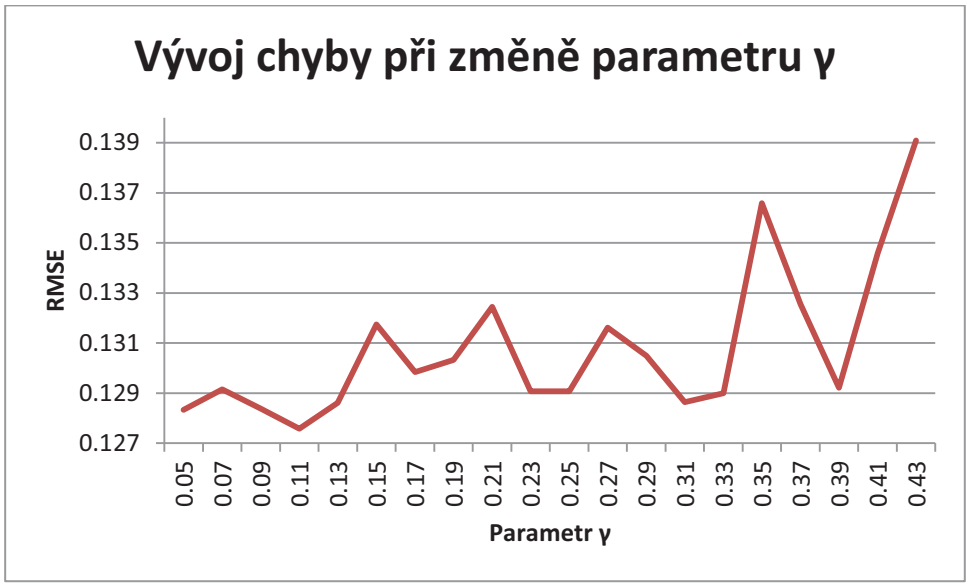
Neméně důležitou hodnotou pro vlastnosti navrhovaného modelu je parametr ε. Jeho změna je testována na předchozím grafu 15. Všeobecně lze říci, že čím nižší je jeho hodnota, tím nižší je i výsledná chyba modelu. To potvrdil i tento graf, kde lze vidět, že postupným snižováním od hodnoty 1 po krocích $\Delta\varepsilon=0.05$ dochází k téměř lineárnímu snižování RMSE. Okolo hodnoty RMSE=0.35 dochází ke zlomu a pak už křivka klesá pomaleji. I tak je ale chyba minimální při nejnižší testované hodnotě $\varepsilon=0.1$.

Průběh parametru r, používaného jako hodnoty koeficientu je zobrazen na následujícím grafu 16. Jeho rozpětí je malé, pouze v hodnotách 1 až 5. Nejlepší vlastnosti model vykazoval při $r=2$. Poté již docházelo k velkému růstu chyby.



Graf 16: Vývoj RMSE chyby při změně parametru r u modelu s polynomickou jádrovou funkcí

Posledním parametrem, který umožňoval model s polynomickou jádrovou funkcí nastavit je hodnota γ. Jak je vidět na grafu 17, mezi hodnotami chyby modelu a velikostí tohoto parametru není žádná zřejmá závislost. Jediné co k tomuto grafu lze poznamenat je, že když bychom proložili křivku spojnicí trendu, tak by značila rostoucí hodnotu chyby modelu se zvyšujícím se parametrem γ. Při pohledu je zřejmé, že u tohoto parametru bylo velmi těžké určit globální minimum vzhledem k jeho velikostem, změnám $\Delta\gamma=0.01$, testovaném intervalu a velkému počtu lokálních minim.



Graf 17: Vývoj RMSE chyby při změně parametru γ modelu s polynomickou jádrovou funkcí

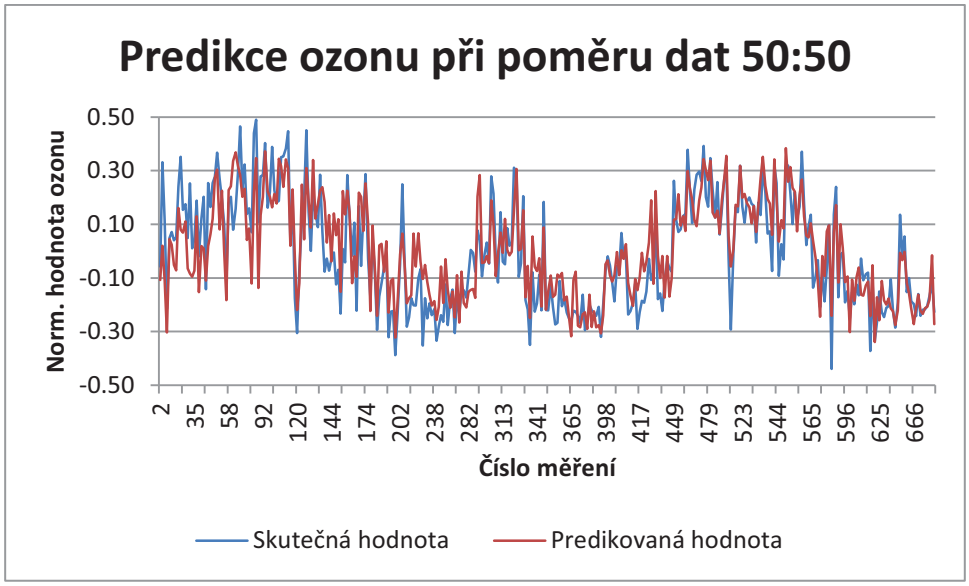
4.6. Model s RBF jádrovou funkcí

U posledního testovaného modelu je použita funkce RBF. Když se podíváme do následující tab. 10, je možné říci, že v porovnání s předchozím typem funkce vykazuje lepší výsledky. Ty se místy liší i o 2 setiny. Tento model vyžaduje zadání 3 parametrů, a byl testován na základě jejich změn. Nejlepší výsledky vykazuje pro nastavení trénovacích a testovacích dat 70:30 s chybou 0.099767. Jak je možné si všimnout při pohledu na tuto tabulku, tak pro jednotlivé poměry dat byla RMSE chyba nejmenší téměř za shodných parametrů C a ε. Jedinou proměnnou, která má velmi rozdílné hodnoty je γ.

C	ε	γ	RMSE test	O_{train} : O_{test}
1	0.1	1.05	0.104808	50:50
1	0.1	2.714	0.105981	60:40
1	0.1	2.97	**0.099767**	70:30
5	0.1	1.005	0.101154	80:20
1	0.1	0.876	0.103800	90:10

Tab. 10: Zobrazení výsledných hodnot RMSE pro model s RBF jádrovou funkcí

Na následujícím grafu 18 je již vidět, že při poměru dat 50:50 a chybě 0.010985 dochází u křivky značící predikovaná data k téměř kopírování průběhu křivky pro skutečně naměřené hodnoty ozonu.



Graf 18: Porovnání skutečné a predikované hodnoty modelu s RBF jádrovou funkcí

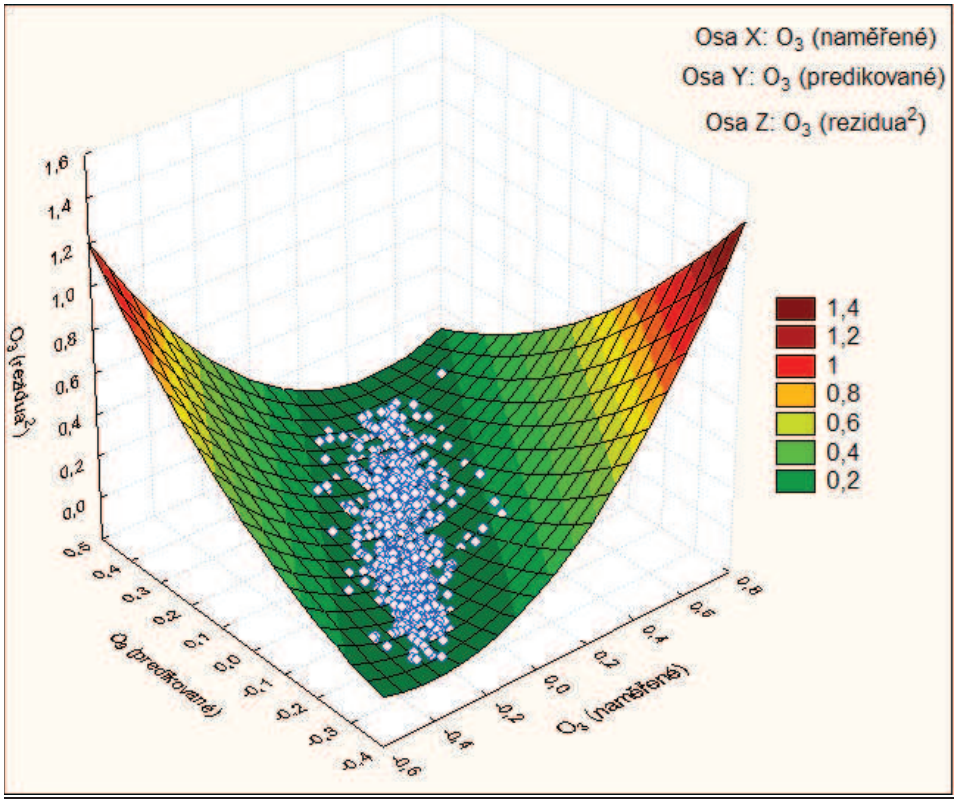
V závěrečné tabulce k podkapitole RBF jádrové funkce je zde uveden vývoj chyby při testování jednotlivých nastavení parametrů modelu (tab. 11). Pro docílení postupného snižování chyby bylo potřeba mnoha testů, při kterých docházelo ke značným změnám v hodnotách jednotlivých proměnných.

C	ε	γ	RMSE test	$O_{train} : O_{test}$
2	0.5	0.5	0.140650	50:50
3	0.3	1	0.116156	50:50
5	0.2	1	0.112670	50:50
0.1	10	1	0.112235	50:50
6	0.2	1.5	0.110507	50:50
7	0.1	1	0.110187	50:50
1	0.1	0.7	0.107889	50:50
3	0.1	0.6	0.107739	50:50
1	0.1	0.93	0.104855	50:50
1	0.1	1.05	**0.104808**	50:50

Tab. 11: Vývoj chyby pro různé parametry modelu s RBF jádrovou funkcí

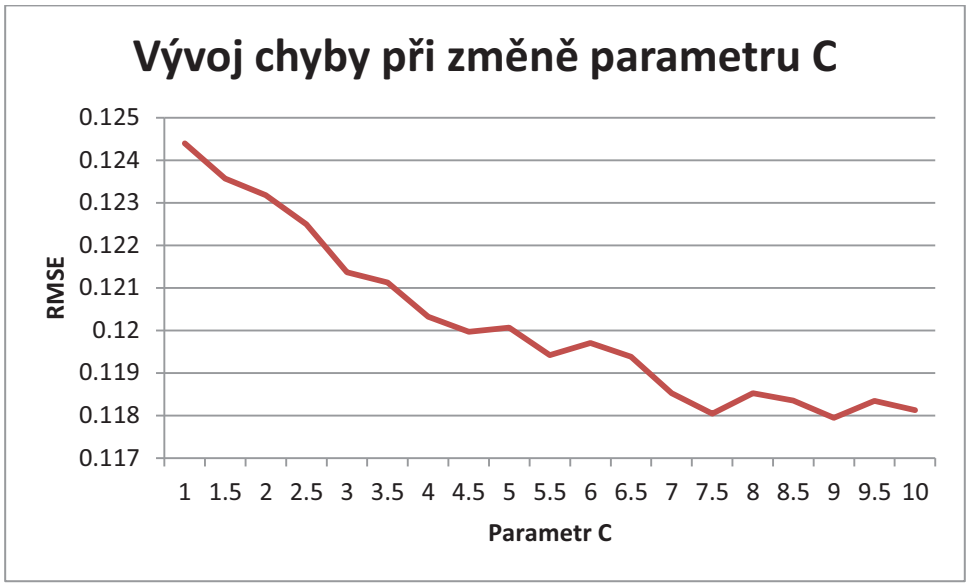
Jak je vidět v tab. 11, mezi parametry neexistuje žádný lineární vztah, pomocí kterého by se dalo určit, že změnou dané hodnoty bude docíleno snížení chyby. Naopak, často se stávalo, že při změně prvního parametru došlo k jejímu výraznému zvýšení a až úpravou dalších hodnot opět klesla pod úroveň té, doposud nalezené chyby.

Jelikož program Statistika umožňuje tvorbu nepřeberného množství grafů jak načtených, tak vypočtených hodnot, je do této podkapitoly vložen následující graf 19. Vyjadřuje vztah mezi naměřenými (osa X), predikovanými (osa Y) a čtverci reziduí (osa Z) hodnot ozonu.



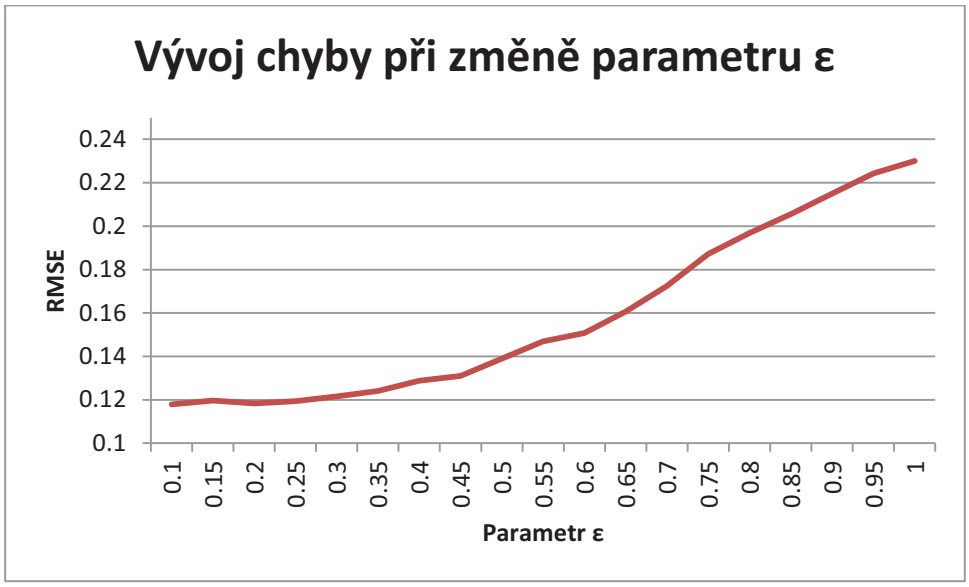
Graf 19: Závislost naměřených a predikovaných hodnot O_3 a reziduí^2 u modelu s RBF jádrovou funkcí

Prvním parametrem, u něhož docházelo u modelu s RBF funkcí ke změnám je hodnota proměnné C. Jak lze vidět z následujícího grafu 20, tak se při postupném zvyšování hodnot parametru C snižuje RMSE chyba. V bodě kdy C dosáhne hodnoty 10 se nachází globální minimum. Při následujících testech pro vyšší C již začala chyba modelu opět stoupat.



Graf 20: Vývoj chyby při změně parametru C u modelu s RBF jádrovou funkcí

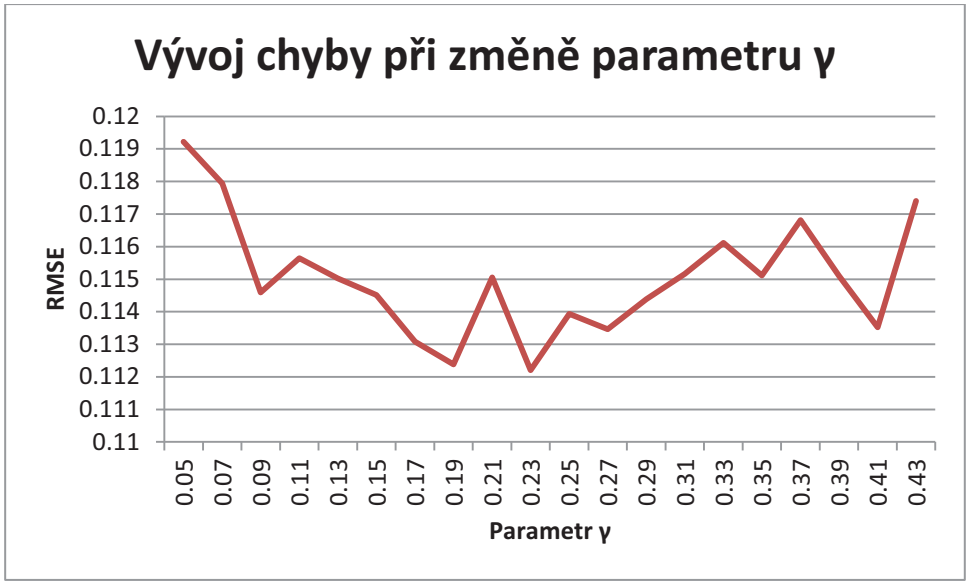
Křivka vývoje reakce modelu na změnu parametru ε je zobrazena na následujícím grafu 21. Lze říci, že s postupným zvyšováním tohoto parametru dochází v případě modelu s RBF jádrem ke zvyšování jeho chyby.



Graf 21: Vývoj chyby při změně parametru ε u modelu s RBF jádrovou funkcí

Pro hodnotu proměnné γ je charakteristické, že mezi jeho hodnotou a vývojem chyby modelu není lehké nalézt závislost. Pro jeho změně nelze jednoznačně říci, že při konkrétní hodnotě bylo nalezeno globální minimum chybu modelu, protože ta dokáže

po několik měření postupně stoupat/klesat, nebo se také mění při každém měření. Na následujícím grafu 22 je zobrazeno testování pro interval γ=<0.05; 0.43> s Δγ =0.02. Při proložení křivky bychom viděli, že vývoj chyby modelu má přibližně tvar písmene U. V této názorné ukázce bylo nalezeno minimum při hodnotě 0.11.



Graf 22: Vývoj chyby při změně parametru γ u modelu s RBF jádrovou funkcí

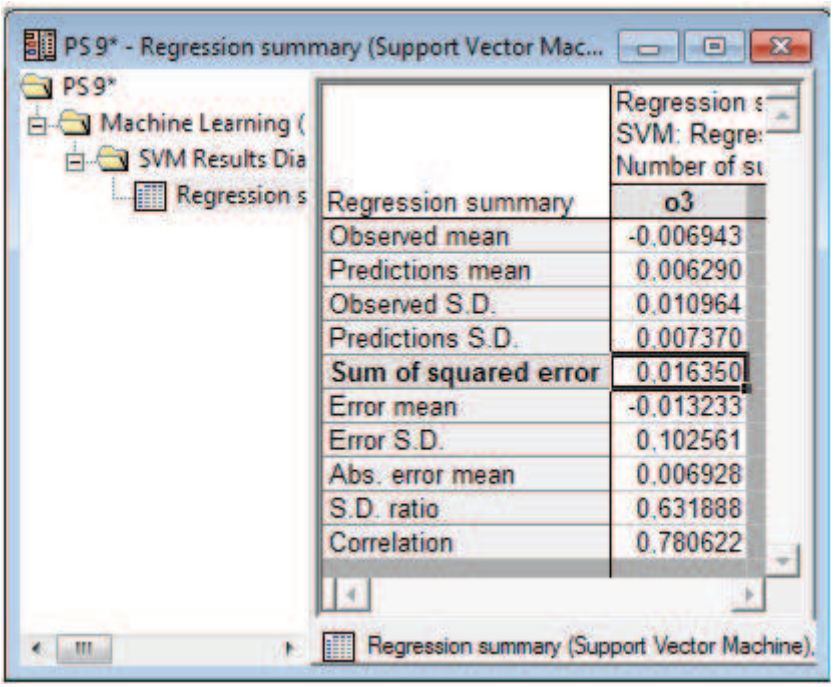
4.7. Závěrečné srovnání

Cílem této kapitoly bylo nalézt nejlepší model pro predikci hodnot ozonu. Celkem k tomu byly využity 3 druhy jádrových funkcí (lineární, polynomické a RBF). Každá z nich představovala rozdílní přístup a počet zadávaných parametrů. Jejich srovnání je zobrazeno v následující tab. 12.

Typ jádra	C	ε	q	γ	r	RMSE test	$O_{train} : O_{test}$
Lineární	4	0.2	-	-	-	0.125038	70:30
Lineární	1	0.1	-	-	-	0.122863	50:50
Polynomické	1	0.1	3	0.502	0	0.110675	70:30
Polynomické	11	0.1	2	0.283	0	0.108130	80:20
RBF	1	0.1	-	0.876	-	0.103800	90:10
RBF	5	0.1	-	1.005	-	**0.101154**	80:20

Tab. 12: Porovnání naměřených RMSE chyb pro modely s jednotlivými jádrovými funkcemi

Nejlepší vlastnosti a tím i nejmenší RMSE chybu měl model s použitím Radial Basis Function. Je možné si všimnout odlišnosti jednotlivých modelů i v tom, pro jaký poměr trénovacích a testovacích dat vykazovaly nejlepší výsledky. Zajímavostí na konec kapitoly, a spíše jen k zamyšlení, bych zde uvedl pravděpodobně chybu v popisu hodnoty při zobrazení popisné statistiky. Na níže uvedeném obr. 21 má označená hodnota 0.016350 popisek jako „Sum of squared error“. Ale když se podíváme na obr. 22, tak z něho lze vyčíst identickou hodnotu označenou jako „Mean error squared“.



Obr. 21: Popisná statistika modelu

```
Dataset upravena_data_ozon:
  Dependent: o3
  Independents: SO2, PM10, Nox, CO, den, měsíc, pracovní den, rychlost
  Sample size = 338 (Train), 339 (Test), 677 (Overall)

Support Vector machine results:
  SVM type: Regression type 1 (capacity=1,000, epsilon=0,100)
  Kernel type: Polynomial (degree=2, gamma=0,076, coefficient=0,000)
  Number of support vectors = 223 (215 bounded)

  Mean error squared = 0,015(Train), 0,016(Test), 0,016(Overall)
  S.D. ratio = 0,576(Train), 0,632(Test), 0,604(Overall)
  Correlation coefficient = 0,832(Train),  0,781(Test),  0,809(Overall)
```

Obr. 22: Výsledky modelu zobrazené ve Statistice

Závěr

Cílem této diplomové práce bylo vytvoření modelu sloužícího pro predikci ozonu. Pro zasvěcení do problematiky jsem v úvodních kapitolách popsal jednotlivé vlastnosti a druhy ozonu včetně látek, které ohrožují ozonovou vrstvu. V rámci teoretické části jsem poté objasnil metodu Support Vector Machines, která je poté v praktické části použita pro tvorbu modelu pro predikci. Protože se jedná o velmi rozsáhlou oblast, byla zaměřena pozornost hlavně na vysvětlení základních pojmů a principů včetně nejpoužívanějších typů jader, společně s ozřejměním problému XOR. Jelikož lze SVM použít kromě rozpoznávání vzorů i pro nelineární regresi (SVR), je věnována část práce i této oblasti.

Pro učení a následné testování modelu jsem použil data naměřená v městské části Dukla v Pardubicích. Pro každou proměnnou se jedná o 679 hodnot, rozdělených v poměrech 50:50 až 90:10 na trénovací a testovací množinu. Pro zjištění RMSE chyby byly postupně měněny parametry lineární, polynomické a RBF jádrové funkce, z nichž každá představuje rozdílní přístup a počet zadávaných parametrů. Při porovnání těchto funkcí musím říci, že největším úskalím a nejsložitější bylo nalezení nejnižší chyby u modelu s polynomickým jádrem, protože zde, pro dosažení nejlepšího výsledku, dochází k současné změně 5-ti parametrů. Na základě porovnání jednotlivých modelů v kapitole analýzy výsledků jsem zjistil, že nejmenší chybu vykazuje model s RBF jádrovou funkcí s hodnotou RMSE 0.101154.

Z výsledků vyhodnocení chyb jednotlivých navržených modelů s různými jádrovými funkcemi lze říci, že použitá metoda Support Vector Machines je vhodná pro predikci ozonové vrstvy.

Použitá literatura a zdroje

[1] *Activities in Ozone* [online]. 2011 [cit. 2011-06-22]. United Nations Environment Programme. Dostupné z WWW: <http://www.unep.org/themes/ozone/?page=info>.

[2] *Antarctic Ozone Hole* [online]. 2009 [cit. 2011-06-22]. NASA Earth Observatory. Dostupné z WWW: <http://earthobservatory.nasa.gov/images/imagerecords/38000/38835/ozone_1979-2008.png>.

[3] *Bear* [online]. 2011 [cit. 2011-06-22]. SABLE-3. Dostupné z WWW: <http://bear.sbszoo.com/sable/images/sbl3/bart65.jpg>.

[4] *Centre for Atmospheric Science* [online]. 2011 [cit. 2011-06-22]. Mobile boundary-layer LiDAR. Dostupné z WWW: <http://www.cas.manchester.ac.uk/images/photos/themes/600x400/Capel-Dewi-02.jpg>.

[5] CÍLEK, V., et al. *Ozonová vrstva Země*. Praha: Vesmír, 1995. 154 s. ISBN 80-901131-5-X.

[6] DOBROVOLNÝ, P., et al. *Problematika ztenčování ozónové vrstvy* [online]. [cit. 2011-06-22]. Ozon. Dostupné z WWW: <http://www.sci.muni.cz/~dobro/ozon_1.htm>.

[7] *Dobson Ozone Spectrophotometer* [online]. 2011 [cit. 2011-06-22]. Museum of Learning. Dostupné z WWW: <http://www.museumstuff.com/learn/topics/Dobson_ozone_spectrophotometer>.

[8] DRUCKER, H., BURGES, CH.J.C., KAUFMAN, L., SMOLA, A.J., VAPNIK, V. *Support Vector Regression Machines*. NIPS, 1996. pp.155~161.

[9] GORUNESCU, F. *Data Mining: Concepts, Models and Techniques*. Berlin: Springer, 2011. 360 s.

[10] HAYKIN, S. *Neural Networks: A Comprehensive Foundation*. New Jersey: Prentice-Hall, Inc., 1999. 842 s.

[11] *History of Ozone* [online]. 2009 [cit. 2011-06-22]. Lenntech. Dostupné z WWW: <http://www.lenntech.com/library/ozone/history/ozone-history.htm>.

[12] KECMAN, V. *Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models*. Cambridge: MIT Press, 2001. 541 s.

[13] KVASNIČKA, V., et al. *Úvod do teórie neurónových sietí*. Bratislava: Iris, 1997. 262 s.

[14] *Learn*: *Introduction to Ozone* [online]. [cit. 2011-06-22]. Atmospheric Science Explorers. Dostupné z WWW: <http://www.ucar.edu/learn/1_5_1.htm>.

[15] *Measured Ozone Depletion* [online]. 2006 [cit. 2011-06-22]. Ozone measurement. Dostupné z WWW: <http://www.albany.edu/faculty/rgk/atm101/ozmeas.htm>.

[16] *Mezinárodní den ochrany ozonové vrstvy* [online]. 2011 [cit. 2011-06-22]. Ministerstvo životního prostředí. Ozonový den. Dostupné z WWW: <http://www.mzp.cz/cz/ozonovy_den>.

[17] MORRISETTE, P.M. *The Evolution of Policy Responses to Stratospheric Ozone Depletion* [online]. [cit. 2011-06-22]. Dostupné z WWW: <http://www.ciesin.org/docs/003-006/003-006.html>.

[18] *Národní a mezinárodní legislativa* [online]. 2008 [cit. 2011-06-22]. Ministerstvo životního prostředí. Dostupné z WWW: <http://www.mzp.cz/cz/narodni_mezin%C3%A1rodni_legislativa>.

[19] NILSSON, N.J., et al. *Introduction to Machine Learning*. Stanford: Stanford University, 1998. 188 s.

[20] *NOAA* [online]. 2005 [cit. 2011-06-22]. News Online. Dostupné z WWW: <http://www.ozonelayer.noaa.gov/action/poessat.gif >.

[21] *Otevřená encyklopedie* [online]. 2011 [cit. 2011-06-22]. NAVAJO. Dostupné z WWW: <http://kyslik.navajo.cz/kyslik-4.png>.

[22] *Ozone (O3): Stratospheric and Ground-Level* [online]. [cit. 2011-06-22]. Clean Air Strategic Alliance. Dostupné z WWW: <http://dwb.unl.edu/teacher/nsf/c09/c09links/www.casahome.org/ozone.htm>.

[23] *Ozone* [online]. 2011 [cit. 2011-06-22]. Wikipedia. Dostupné z WWW: <http://en.wikipedia.org/wiki/Ozone>.

[24] *Ozone Facts: What is a Dobson Unit?* [online]. 2008 [cit. 2011-06-22]. Ozone Hole Watch. Dostupné z WWW: <http://ozonewatch.gsfc.nasa.gov/facts/dobson.html>.

[25] PARSON, E. *Protecting the Ozone Layer: Science and Strategy*. Oxford University Press, 2003. 377 s

[26] PYLE, J.A., et al. *Ozone Depletion and Chlorine Loading Potentials* [online]. 1992 [cit. 2011-06-22]. Scientific Assessment of Ozone Depletion. Dostupné z WWW: <http://www.ciesin.org/docs/011-551/011-551.html>.

[27] *Skleníkový efekt* [online]. 2011 [cit. 2011-06-22]. Meteocentrum.cz. Dostupné z WWW: <http://www.meteocentrum.cz/zmeny-klimatu/sklenikovy-efekt-dalsi-plyny.php>.

[28] *Statistica* [online]. 2011 [cit. 2011-06-22]. Published Reviews. Dostupné z WWW: <http://www.statsoft.com/company/reviews/2011-published-reviews/#computerwoche>.

[29] STEINWART, I., CHRISTMANN, A. *Support Vector Machines*. Berlin: Springer, 2008. 601 s.

[30] *Support Vector Machine Regression* [online]. [cit. 2011-06-22]. Kernel SVM. Dostupné z WWW: <http://kernelsvm.tripod.com/>.

[31] WANG, L. *Support Vector Machines: Theory and Applications*. Berlin: Springer, 2005. 431 s.

[32] WEBB, A.R. *Statistical Pattern Recognition*. Chichester: John Wiley and Sons, 2002. 496 s.

Slovník pojmů a zkratek

UV-A, UV-B, UV-C	Ultrafialové záření typu A, B a C
D.U.	Dobsonova jednotka (Dobson unit)
CFC	Označení pro freony
ODP	Ozone Depletion Potential
LIDAR	Light Detection and Ranging
TOMS	Total Ozone Mapping Spectrometer
SBUV	Solar Backscatter Ultraviolet
GO3OS	Global Ozone Observing System
GAW	Global Atmosphere Watch
COLT	Computational Learning Theory
NN	Neuronová síť (Neural Network)
RBF	Radial Basis Function
SRM	Structural Risk Minimization
SVM	Support Vector Machine
VC	Vapnik-Chervonenkis
XOR	Exkluzivní disjunkce (Exclusive OR)
LSVM	Langrangian SVM
NLSVM	Finite Newton Langrangian SVM
MSE	Střední čtvercová chyba
RMSE	Střední kvadratická chyba

Seznam obrázků

Seznam grafů

Seznam tabulek