Aproximace funkcí neuronovými sítěmi

Kůrková, Věra
1998

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Aproximace funkcí neuronovými sítěmi

Věra Kůrková

Technical report No. 770

prosinec 1998

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic
phone: (+4202) 6605 3231 fax: (+4202) 85 85 789
e-mail: vera@uivt.cas.cz

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Aproximace funkcí neuronovými sítěmi

Věra Kůrková [1]

Technical report No. 770
prosinec 1998

Abstract

V kapitole je podán přehled nejdůležitějších výsledků matematické teorie neuronových sítí a jejich důsledků pro metodologii volby architektury neuronové sítě a návrh algoritmů učení.

Keywords

Aproximace a interpolace, univerzální aproximační vlastnost, globální a lokální minima chybových funkcionálů, vlastnost nejlepší aproximace, složitost neuronových sítí, horní a dolní odhady přesnosti aproximace.

[1] Tato práce byla podporována GA AV A2030602.

Contents

1 Umělé neuronové sítě a teorie aproximace

1.1 Umělá inteligence a konekcionismus

Základním nástrojem a dominantní metaforou umělé inteligence je digitální počítač provádějící manipulaci se symboly na základě pravidel. Alternativy k tomuto přístupu k modelování vlastností inteligence byly uvažovány již ve 40.letech, kdy zakladatelé kybernetiky diskutovali o takových vlastnostech mozku jako je mnohonásobné propojení velkého počtu neuronů, nelokalizované ukládaní informací a jejich zpracování bez centrálního procesoru a bez pravidel.

Na základě myšlenek McCullocha a Pittse (1943), kteří uvažovali zjednodušený model mozku provádějící logické dedukce pomocí vzájemného propojení neuronů reprezentovaných spínači s nastavenými prahovými hodnotami, Rosenblatt a Wightman postavili v roce 1958 první neuropočítač Mark I Perceptron. Rosenblattův neuropočítač byl sestaven z výpočetních jednotek nazývaných perceptrony. *Perceptron* jakožto zjednodušený model neuronu počítá váženou sumu signálů přicházejících na jeho vstupy (váhy odpovídají síle synapsí), porovnává tuto sumu s prahem a výsledek je transformován ve výstupní signál pomocí tzv. aktivační funkce (Rosenblatt použil jako aktivační funkci nespojitou prahovou funkci; rozšíření gradientních metod učení vedlo později k jejímu nahrazení různými spojitými aproximacemi tzv. sigmoidami). Algoritmus učení

spočíval v adaptaci váhových a prahových parametrů tak, aby funkce neuropočítače odpovídala až na malou chybu dané množině dvojic vstup/výstup.

Meze tehdejších technologických možností (váhy Rosenblattova neuropočítače byly implementovány pomocí potenciometrů řízených elektromotory) společně s nesprávnou interpretací teoretického výsledku (důkaz, že jediný perceptron není schopen počítat logickou funkci XOR, Minsky & Papert, 1960, byl využit jako argument dokazující omezené výpočetní schopnosti počítačů složených z perceptronů) vedly ke značnému poklesu zájmu o konekcionistické modely.

K oživení zájmu o neuropočítače došlo až v osmdesátých letech díky rozvoji technologie (snadný přístup k rychlým počítačům) i teorie (znovuobjevení principů samoorganizace ve fyzice a v nelineární matematice). Avšak v době, kdy konekcionismus přežíval pouze v takových oborech jako adaptívní zpracování signálů nebo biologické modelování, se podařilo Werbosovi, 1970, rozšířit Rosenblattův algoritmus učení perceptronu na sítě složené z výpočetních jednotek obecného typu. Werbosův algoritmus zpětného šíření, který byl znovuobjevením metody stochastické regrese studované ve statistice od padesátých let, otevřel možnosti efektívního učení neuronových sítí mnoha typů. Název Werbosovy dizertace "Beyond regression" výstižně vyjadřuje možnosti konekcionistických modelů. Úloha nalezení optimálních parametrů křivky daného typu nejlépe aproximující určitou množinu dat je klasický problém, který řeší například lineární regrese nebo Fourierova reprezentace. Avšak tradiční metody se snaží nalézt pouze optimální parametry lineární kombinace *pevně dané množiny funkcí* (např. množiny trigonometrických funkcí s pevně danými frekvencemi), zatímco algoritmus zpětného šíření v procesu optimalizace kromě parametrů lineární kombinace vybírá rovněž *nejvhodnější množinu funkcí* (určenou vnitřním nastavením parametrů výpočetních jednotek daného počtu a typu). Optimalizace vnitřních parametrů umožňuje dosažení výrazně lepší aproximace, než jaké by bylo možno dosáhnout pouhým lineárním kombinováním jednotek s pevně danými parametry.

Od počátku 80.let byly konekcionistické modely s úspěchem použity pro řešení mnoha úloh klasifikace a rozpoznávání tvarů – např. vokalizace textu (Sejnowski & Rosenberg, 1987), rozpoznávání ručně psaných znaků a mluveného slova (Burr, 1988) a odezírání samohlásek (Sejnowski & Yuhas, 1991).

V expertních systémech a ve standartních metodách rozpoznávání tvarů je nejobtížnějším krokem nalezení pravidel či příznaků, na jejichž základě dochází ke klasifikaci či rozhodování. V konekcionistických modelech nejsou žádná pravidla ani příznaky explicitně implementovány do systému – informaci, kterou představují obsahuje celý systém (typ jednotek a způsob propojení a jejich parametrizace). Vhodná struktura je nalezena pomocí učení systému metodou pokus a omyl – po každém omylu jsou parametry systému pozměněny tak, aby se snížila chyba mezi aktuálním a požadovaným výstupem.

Konekcionistický přístup dobře ilustruje paradigmatický příklad využití neuronové sítě NETtalk. Sejnowski & Rosenberg (1987) jej vytvořili pro porovnání schopností konekcionistického systému s expertním systémem. NETtalk plní stejnou úlohu jako expertní systém DECtalk – konverzi grafických znaků na fonetické. Avšak zatímco DECtalk čte nahlas anglický text tak, že na základě složitých pravidel výslovnosti přiřazuje správný zvuk (foném) posloupnosti 7 znaků obsahující společně s vyslovo-

vaným znakem rovněž tři předcházející a tři následující znaky (výslovnost znaku v angličtině závisí na kontextu – viz např. prostřední "o" ve slovech "through" a "rough"), NETtalk se obejde bez explicitně formulovaných pravidel výslovnosti. Na základě učení na příkladech správné výslovnosti je NETtalk schopen vyslovovat anglický text stejně dobře jako DECtalk.

NETtalk využívá pro konverzi psaného textu v mluvené slovo dopřednou neuronovou síť s 203 vstupními jednotkami (každý ze 7 znaků může nabývat 26 hodnot odpovídajících písmenům anglické abecedy a oddělovacím znakům). Vstupní vektor představuje tedy binárně kódovaný řetězec 7 znaků čteného textu. Výstupy sítě odpovídají 30 anglickým fonémům. Ve skryté vrstvě je 80 sigmoidálních perceptronů a učení je prováděno standartním algoritmem zpětného šíření.

NETtalk je blízký lidskému způsobu rozpoznávání tvarů, které zpravidla není provázeno schopností explicitně popsat příznaky těchto tvarů či pravidla, na jejichž základě tyto tvary třídíme. Aspirace konekcionismu jsou mnohem skromnější než aspirace umělé inteligence – místo strojového dokazování matematických vět se pokouší napodobit schopnosti živých systémů orientovat se v záplavě smyslových dat. Vychází z předpokladu, že velké množství vzájemně propojených jednoduchých procesorů zvládne tyto úlohy rychleji a lépe než metody založené na symbolické reprezentaci a manipulaci se symboly pomocí jasně formulovaných pravidel.

1.2 Dopředné neuronové sítě

V průběhu minulého desetiletí bylo navrženo mnoho nových typů výpočetních jednotek, architektur jejich vzájemného propojení a algoritmů učení. Výpočetní jednotky používané v těchto modelech bývají sice stále často nazývány umělé neurony, ale mnohé z nich se vůbec nepodobají původním prahovým perceptronům. Typické pro současné modely neuropočítání nejsou vlastnosti jednotlivých výpočetních jednotek, ale jejich vzájemné propojení do sítí a schopnost řešit problémy na základě adaptace či učení. Modely neuropočítání vycházejí dnes spíše z vhodných matematických vlastností a technologických možností implementace než z biologických analogií.

V závislosti na způsobu vzájemného uspořádání výpočetních jednotek rozlišujeme různé typy modelů neuronových sítí: např. Booleovské obvody, dopředné a rekurentní sítě, Hopfieldovy sítě a Kohonenovy mapy.

Důležitým typem vhodným pro úlohy rozpoznávání a klasifikace jsou *dopředné neuronové sítě*. Z teoretického hlediska můžeme považovat tyto sítě za nástroje na počítání funkcí na podmnožinách vícedimenzionálních prostorů. Protože v konkrétních aplikacích jsou vstupní vektory vždy omezené, často se při studiu vlastností funkcí počítatelných těmito sítěmi omezujeme na funkce definované na d-dimenzionální Eukleidovské krychli $[0,1]^d$ nebo Boolovské krychli $\{0,1\}^d$. Funkce vstup/výstup dopředné neuronové sítě závisí na *architektuře* sítě a na *parametrech* výpočetních jednotek.

Architektura je dána typem a počtem jednotek (např. perceptrony nebo radiální jednotky), typem jejich vzájemného propojení (jednotky mohou být například uspořádány v několika vrstvách tak, že pouze jednotky v sousedních vrstvách jsou propojeny; někdy mohou být přidána i horizontální propojení v rámci téže vrstvy). Důležitým typem jsou *vrstevnaté sítě*, kde jsou jednotky umístěny v několika lineárně uspořádaných vrstvách

tak, že každá jednotka v předchozí vrstvě je propojena s každou jednotkou v následující vrstvě. V první tzv. *vstupní vrstvě* jsou *vstupní jednotky*, které identicky přenášejí jednotlivé složky vstupního vektoru. Poslední vrstva se nazývá *výstupní* a mezilehlým vrstvám se říká *skryté*. Nejčastěji používanými typy vrstevnatých sítí jsou sítě s jednou nebo se dvěmi skrytými vrstvami.

Výpočetní jednotky počítají funkce závisející na dvou vektorových proměnných: na *vstupním vektoru* a na *vektoru parametrů*. Obecně tyto jednotky počítají funkce tvaru $\phi : \mathcal{R}^p \times \mathcal{R}^d \to \mathcal{R}$, kde ϕ odpovídá typu jednotky, p a d dimenzím prostoru parametrů resp. vstupního prostoru a $\mathcal{R}$ značí množinu reálných čísel ($\mathcal{R}_+$ značí množinu nezáporných reálných čísel). Standartní typy výpočetních jednotek jsou lineární jednotky, perceptrony a radiální jednotky (RBF jednotky).

Lineární jednotka počítá váženou sumu svých vstupů, tj. funkci $\phi : \mathcal{R}^n \times \mathcal{R}^n$ tvaru $\phi(\mathbf{w}, \mathbf{x}) = \sum_{i=1}^{n} w_i x_i = \mathbf{w} \cdot \mathbf{x}$.

Perceptron s aktivační funkcí $\psi : \mathcal{R} \to \mathcal{R}$ počítá funkce tvaru $\phi((\mathbf{v}, b), \mathbf{x}) = P_\psi(\mathbf{v}, b, \mathbf{x}) = \psi(\mathbf{v} \cdot \mathbf{x} + b) : \mathcal{R}^{d+1} \times \mathcal{R}^d \to \mathcal{R}$, kde $\mathbf{v} \in \mathcal{R}^d$ je *vstupní váhový vektor* a $b \in \mathcal{R}$ je *práh* (bias). $\mathcal{P}_d(\psi) = \{f : [0,1]^d \to \mathcal{R}; f(\mathbf{x}) = \sum_{i=}^{n} \psi(\mathbf{v} \cdot \mathbf{x} + b), n \in \mathcal{N}_+, \mathbf{v} \in \mathcal{R}^d, b \in \mathcal{R}\}$ značí množinu funkcí na $[0,1]^d$, které lze počítat ψ-perceptronovými sítěm s libovolným počtem jednotek ve skryté vrstvě ($\mathcal{N}_+$ značí množinu kladných přirozených čísel).

Rosenblatt použil jako aktivační funkci nespojitou prahovou tzv. Heavisidovu funkci $\vartheta : \mathcal{R} \to \mathcal{R}$ definovanou $\vartheta(t) = 0$ pro všechna $t < 0$ a $\vartheta(t) = 1$ pro všechna $t \geq 0$. Dnes bývají používány její spojité aproximace, tzv. sigmoidy. Termínem *sigmoidální funkce* bývá zpravidla označována spojitá neklesající funkce $\sigma : \mathcal{R} \to \mathcal{R}$ s limitami $\lim_{t \to -\infty} \sigma(t) = 0$ a $\lim_{t \to \infty} \sigma(t) = 1$. Nejčastěji používané sigmoidy jsou *logistická sigmoida* $\frac{1}{1+e^{-t}}$ a hyperbolický tangens (po vhodné změně počátku souřadnic a měřítka lze jednu obdržet z druhé, takže z teoretického hlediska je jejich použití ekvivalentní).

Radiální (RBF) jednotka s radiální funkcí $\psi : \mathcal{R} \to \mathcal{R}_+$ počítá funkce tvaru $\phi((\mathbf{v}, b), \mathbf{x}) = B_\psi(\mathbf{v}, b, \mathbf{x}) = \psi(b\|\mathbf{x} - \mathbf{v}\|)$, kde $\mathbf{v} \in \mathcal{R}^d$ je *střed*, $b \in \mathcal{R}_+$ je *šířka* a $\|.\|$ je norma na $\mathcal{R}^d$. $\mathcal{B}_d(\psi) = \mathcal{B}_d(\psi, \|.\|) = \{f : [0,1]^d \to \mathcal{R}; f(\mathbf{x}) = \sum_{i=1}^{n} \psi(b\|\mathbf{x} - \mathbf{v}\|), n \in \mathcal{N}_+, \mathbf{v} \in \mathcal{R}^d, b \in \mathcal{R}\}$ značí množinu funkcí na $[0,1]^d$, které lze počítat RBF sítěmi s radiální funkcí ψ. Nejčastěji používanou radiální funkcí je Gaussova funkce $\gamma(t) = e^{-t^2}$.

Sigmoidální perceptrony a RBF jednotky jsou geometricky opačné: Perceptrony aplikují sigmoidální funkci na váženou sumu vstupů, k níž je přičten práh – odpovídají tedy *nelokalizovaným* oblastem vstupního prostoru, který rozkládají pomocí více či méně rozmazaných nadrovin (v závislosti na strmosti sigmoidy) na dva poloprostory, zatímco radiální jednotky počítají vzdálenost od středu, násobí ji šířkou a pak aplikují radiální funkci – odpovídají tedy *lokalizovaným* oblastem. Funkce počítané perceptrony patří do třídy funkcí nazývaných *rovinné vlny* nebo též *hřebenové funkce* – jejich hodnota je totiž konstantní podél všech nadrovin rovnoběžných s nulovou nadrovinou afinní funkce $\mathbf{v} \cdot \mathbf{x} + b$ dané vnitřními parametry perceptronu vahou $\mathbf{v}$ a prahem b, tj. nadrovinou $H_{\mathbf{v},b} = \{\mathbf{x} \in \mathcal{R}^d; \mathbf{v} \cdot \mathbf{x} + b = 0\}$.

Sítě s jedinou lineární výstupní jednotkou a s jednou skrytou vrstvou s jednotkami počítajícími funkci ϕ se nazývají *ϕ-sítě*. ϕ-sítě počítají tedy funkce tvaru $\sum_{i=1}^{n} w_i \phi(\mathbf{a}_i, .)$, kde $\mathbf{a}_i \in \mathcal{R}^p$ jsou parametry ϕ-sítě a n je libovolné nezáporné přirozené číslo (odpovídající počtu jednotek sítě).

Algoritmy učení (např. gradientní, genetické nebo inkrementální) hledají lokální nebo globální minima chybových funkcionálů udávajících, jak se liší požadovaná funkce vstup/výstup od funkce, kterou síť s konkrétními parametry počítá. Chyba se měří pomocí standartních norem používaných ve funkcionální analýze. Volba normy závisí na konkrétní aplikaci sítě, např. na tom, zda je požadována stejná přesnost aproximace pro všechny vstupní vektory nebo zda je předpokládána nějaká pravděpodobnost výskytu či důležitost některých vstupů.

1.3 Aproximace v normovaných lineárních prostorech

Počátky teorie aproximace funkcí sahají do druhé poloviny 19. století, kdy se Weierstrass a Čebyšev zabývali aproximací spojitých funkcí pomocí algebraických a trigonometrických polynomů. Na základě jejich práce byla postupně vybudována *teorie lineární aproximace*, tj. aproximace pomocí funkcí z nějakého konečně dimenzionálního lineárního podprostoru (např. n-dimenzionálního lineárního prostoru tvořeného všemi polynomy stupně menšího než n) (viz např. Singer, 1970). Pokud množina aproximujících funkcí netvoří lineární podprostor (tj. není uzavřená na sčítání funkcí nebo na jejich násobení skalárem), nemusí platit výhodné vlastnosti lineární aproximace jako je jednoznačnost nejlepší aproximace, spojitost aproximačního operátoru apod. Avšak nelineární aproximační množiny mohou zaručit aproximaci s mnohem menší chybou, než jakou lze dosáhnout lineárními aproximačními množinami stejné složitosti. Příkladem takové nelineární aproximační množiny jsou racionální funkce tvaru $\frac{p}{q}$, kde p je polynom stupně nejvýše n a q je polynom stupně nejvýše m (viz např. Braess, 1986). Funkce, které lze počítat neuronovými sítěmi daného typu s nejvýše n skrytými jednotkami rovněž netvoří lineární podprostor, takže zkoumání aproximace funkcí neuronovými sítěmi patří do teorie nelineární aproximace.

V teorii aproximace funkcí neuronovými sítěmi se zabýváme aproximací funkcí z nějakého vhodného *reálného lineárního prostoru*, nejčastěji prostoru $\mathcal{C}(K)$ všech spojitých funkcí na nějaké kompaktní podmnožině $K \subset \mathcal{R}^d$ (nejčastěji d-dimenzionální krychli $[0,1]^d$) nebo některého z prostorů $\mathcal{L}_p(K) = \{f : K \to \mathcal{R}; (\int f^p d\lambda)^{\frac{1}{p}} \leq \infty\}$ pro $p \in [1,\infty]$ (λ značí Lebesgueovu míru, mohou být ale použity i jiné míry).

Aproximační chybu měříme pomocí vhodné normy, zabýváme se tedy aproximací v *normovaných lineárních prostorech*. V prostoru $\mathcal{C}(K)$ se zpravidla užívá supremová norma označovaná $\|.\|_{\mathcal{C}}$ (definovaná $\|f\|_{\mathcal{C}} = \sup_{x \in K} |f(x)|$); v prostorech $\mathcal{L}_p(K)$ se užívají $\mathcal{L}_p$-normy označované $\|.\|_p$ (definované $\|f\|_p = (\int f^p d\lambda)^{\frac{1}{p}}$). Supremová norma je vhodná v případě, kdy je požadována stejná přesnost aproximace pro všechny vstupní vektory, $\mathcal{L}_p$-normy vystihují pravděpodobnost výskytu či důležitost některých vstupů.

Některé vlastnosti aproximace funkcí je vhodné formulovat pro obecné normované lineární prostory splňující určité předpoklady, např. pro Banachovy nebo Hilbertovy prostory a pro striktně konvexní prostory.

Připomeňme, že *Banachův prostor* je normovaný lineární prostor, který je úplný a *Hilbertův prostor* je Banachův prostor, jehož norma je generována skalárním součinem, tj. $\|f\| = \sqrt{f \cdot f}$. Příkladem Hilbertova prostoru je konečně dimenzionální Eukleidovský prostor $\mathcal{R}^d$ s l_2-normou $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^d x_i^2}$ a prostor $\mathcal{L}_2(K)$ s normou $\|f\|_2 =$

$\sqrt{\int f^2 d\lambda}$.

$B_r(\|.\|)$ značí kouli o poloměru r vzhledem k normě $\|.\|$, tj. $B_r(\|.\|) = \{f \in X; \|f\| \leq r\}$. Normovaný lineární prostor je *striktně konvexní*, jestliže pro $x, y \in X$ takové, že $x \neq y$ a $\|x\| = \|y\| = 1$ platí $\|\frac{x+y}{2}\| < 1$ (to znamená, že jednotková koule $B_1(\|.\|)$ nemá rovné plochy). Např. každý Hilbertův prostor je striktně konvexní.

Podmnožina normovaného lineárního prostoru, která má tu vlastnost, že pomocí jejích prvků lze s libovolnou přesností aproximovat každou funkci z tohoto prostoru, se nazývá hustá. Pro matematickou definici pojmu hustoty potřebujeme topologický pojem uzávěru množiny $Y \subseteq X$ označovaný $cl\,Y$: $f \in cl\,Y$, jestliže pro všechna $\varepsilon > 0$ existuje $g \in Y$ takové, že $\|f - g\| < \varepsilon$. Podmnožina $Y \subseteq X$ je tedy *hustá*, jestliže $cl\,Y = X$.

Je-li G podmnožina lineárního prostoru X, pak *lineární obal* G označovaný $span\,G$ je množina všech lineárních kombinací prvků G, tj. $span\,G = \{\sum_{i=1}^{n} w_i g_i; w_i \in \mathcal{R}, g_i \in G, n \in \mathcal{N}_+\}$; $span_n G$ značí množinu všech lineárních kombinací nejvýše n prvků G, tj. $span_n G = \{\sum_{i=1}^{n} w_i g_i; w_i \in \mathcal{R}, g_i \in G\}$. $conv\,G$ značí *konvexní obal* množiny G, tj. množinu všech konvexních kombinací jejích prvků, tj. $conv\,G = \{\sum_{i=1}^{n} a_i g_i; a_i \in [0,1], \sum_{i=1}^{n} a_i = 1, n \in \mathcal{N}_+\}$; $conv_n G$ značí množinu všech konvexních kombinací nejvýše n prvků G, tj. $conv_n G = \{\sum_{i=1}^{n} w_i g_i; w_i \in [0,1], \sum_{i=1}^{n} w_i = 1, g_i \in G\}$.

V prostoru $\mathcal{C}([0,1]^d)$ je tedy množina funkcí, které lze počítat P_ψ-sítěmi rovna množině $span\ P_\psi$.

Teorie aproximace studuje vlastnosti chybových funkcionálů udávajících chybu aproximace měřenou pomocí vzdálenosti od množiny aproximujících funkcí. Pro podmnožinu Y normovaného lineárního prostoru $(X, \|.\|)$ je chybový funkcionál $e_Y : X \to \mathcal{R}_+$ definován formulí $e_Y(f) = \|f - Y\| = \inf_{g \in Y} \|f - g\|$. Připomeňme, že pro každou množinu Y je e_Y spojité zobrazení, které ale nemusí být lineární; dokonce ani v případě, kdy Y je konečně dimenzionálním lineárním podprostorem X (viz Singer, 1970).

Množina těch prvků Y, které jsou nejblíže k dané funkci $f \in X$ se nazývá *projekce* f *na* Y a je označována $P_Y(f) = \{g \in Y; \|f - g\| = \|f - Y\|\}$. Množinové zobrazení $P_Y : X \to \mathcal{P}(Y)$, kde $\mathcal{P}(Y)$ značí množinu všech podmnožin Y, přiřazující každému prvku f prostoru X jeho projekci na Y se nazývá *metrická projekce* nebo jen *projekce* X *na* Y. Jestliže $P_Y(f)$ je pro všechna $f \in X$ neprázdná, potom Y se nazývá *proximinální* (nebo též *existenční množina*); jestliže pro všechna $f \in X$ je projekce $P_Y(f)$ jednobodová, pak se Y nazývá *Chebyševova množina*. V tomto případě označujeme $p_Y : X \to Y$ jednoznačně definované *projekční zobrazení* z X do Y.

Nechť $A : X \to \mathcal{P}(Y)$ je množinové zobrazení. *Výběr z* A je zobrazení $\alpha : X \to Y$ takové, že pro všechna $f \in X$ platí $\alpha(f) \in A(f)$. Zobrazení $p : X \to Y$, které je výběrem z P_Y, se nazývá *operátor nejlepší aproximace* pomocí prvků množiny Y. Množina Y je tedy proximinální právě tehdy, když má operátor nejlepší aproximace, a je Čebyševova právě tehdy, když je takový operátor jediný. *Spojitý výběr* z P_Y je operátor nejlepší aproximace, který je spojitý v každém $f \in X$.

V teorii aproximace nestačí studovat aproximaci jedné funkce, ale je třeba posoudit vhodnost aproximačního schematu pro aproximaci celé množiny funkcí (zpravidla množiny funkcí majících shora omezenou některou z norem, např. Sobolevovu normu omezující velikost parciálních derivací). Vhodnost aproximační množiny lze měřit chybou aproximace nejhoršího případu, tj. nejhůře aproximované funkce, což vysti-

huje pojem *odchylky (deviace)* množiny funkcí, které mají být aproximovány, $B \subseteq X$ od množiny aproximujících funkcí Y. Odchylka B od Y je definována $\delta(B,Y) = \sup_{f \in B} \|f - Y\|$.

2 Univerzální aproximace

Matematická teorie aproximace funkcí pomocí neuronových sítí teoreticky zdůvodnila úspěšné využití těchto sítí v širokém spektru aplikací tím, že ukázala, že neuronové sítě mnoha typů jsou "univerzální" v tom smyslu, že mohou být nastaveny na základě vhodných učících algoritmů pro takové úlohy rozpoznávání obrazců či klasifikace, které lze formulovat jako zobrazení mezi podmnožinami vícedimenzionálních prostorů. Mezi univerzální sítě patří nejen sítě všech standartních architektur užívané v současné době v aplikacích, ale i mnohé další, které doposud nebyly používány.

2.1 Univerzální aproximace pro perceptronové a radiální sítě

První teoretická otázka týkající se daného typu architektury neuronové sítě je, zda-li dostatečně velká síť tohoto typu je schopna počítat s požadovanou přesností libovolnou funkci, která se může vyskytnout v aplikacích. V terminologii neuronových sítí se tato schopnost třídy neuronových sítí nazývá *univerzální aproximační vlastnost*.

Matematicky lze univerzální aproximační vlastnost definovat pomocí topologického pojmu hustoty v některém ze standartních normovaných lineárních prostorů. Vzhledem k tomu, že v praktických úlohách jsou vstupní vektory vždy omezené, stačí studovat hustotu množin funkcí vstup/výstup definovaných na nějakém omezeném a uzavřeném (tj. kompaktním) podprostoru $\mathcal{R}^d$, kde d odpovídá počtu vstupních jednotek. Nejčastěji jsou výsledky formulovány pro funkce definované na $[0,1]^d$.

Třída neuronových sítí má univerzální aproximační vlastnost, jestliže pro každé kladné přirozené číslo d je množina funkcí, které lze počítat sítěmi z této třídy s d vstupními jednotkami a s libovolným počtem skrytých jednotek, hustá v množině $\mathcal{C}([0,1]^d)$ nebo $\mathcal{L}_p([0,1]^d)$.

Následující věta o redukci vstupní dimenze dokázaná Stinchcombem a Whitem (1990) ukazuje, že pro perceptronové sítě stačí ověřit univerzální aproximační vlastnost pouze pro sítě s jednou vstupní jednotkou.

Věta 2.1 *Nechť $\psi : \mathcal{R} \to \mathcal{R}$ je libovolná funkce. Potom $\mathcal{P}_1(\psi)$ je hustá v $\mathcal{C}([0,1])$ právě když pro každé kladné přirozené číslo d je $\mathcal{P}_d(\psi)$ hustá v $\mathcal{C}([0,1]^d)$.*

Důkaz této věty je založen na existenci aktivační funkce určitých vhodných vlastností – funkci exp. Množina funkcí, které lze počítat perceptronovými sítěmi s touto aktivační funkcí $\mathcal{P}_d(\exp)$ totiž pro všechna kladná přirozená čísla d splňuje předpoklady Stone-Weierstrassovy věty o hustotě množin funkcí.

Stoneovo rozšíření klasické Weierstrassovy věty o hustotě polynomů dává dvě postačující podmínky pro to, aby množina spojitých funkcí $\mathcal{A}$ na kompaktní podmnožině K prostoru $\mathcal{R}^d$ byla hustá v prostoru $\mathcal{C}(K)$ se supremovou normou: první z nich požaduje, aby $\mathcal{A}$ byla algebra (tj. aby byla uzavřená na sčítání a násobení funkcí),

druhá požaduje, aby $\mathcal{A}$ oddělovala body K (tj. aby pro každé dva různé body $x, y \in K$ existovala funkce $f \in \mathcal{A}$ taková, že $f(x) \neq f(y)$).

Množiny funkcí, které lze počítat ϕ-sítěmi, jsou vždy uzavřené na sčítání, pro mnohé funkce ϕ též oddělují body, ale nebývají uzavřené na násobení. Pouze v případě perceptronů s exponencielou jako aktivační funkcí tvoří množina $\mathcal{P}_d(\exp)$ pro libovolné d algebru a tedy je dle Stone-Weierstrassovy věty hustá v $\mathcal{C}([0,1]^d)$.

Jestliže $\mathcal{P}_1(\psi)$ je hustá v $\mathcal{C}([0,1])$, lze pomocí funkcí z $\mathcal{P}_1(\psi)$ aproximovat s libovolnou přesností funkci exp. Pomocí superpozice této aproximace lze ukázat, že pro libovolné d je $\mathcal{P}_d(\psi)$ hustá v $\mathcal{P}_d(\exp)$ a tedy je hustá i v $\mathcal{C}([0,1]^d)$.

Stone-Weierstrassova věta o hustotě polynomů je standartním nástrojem pro ověřování hustoty množin spojitých funkcí. Hornik, Stinchcombe & White (1989) ji použili pro důkaz univerzální aproximační vlastnosti sítí s jednou skrytou vrstvou sigmoidálních perceptronů. Stone-Weierstrassovu větu rovněž využili Leshno et al. (1993), kteří výsledek Hornika et al. rozšířili a podali úplnou charakterizaci aktivačních funkcí, pro které perceptronové sítě mají univerzální aproximační vlastnost. Ukázali, že univerzální aproximační vlastnost není omezena na perceptronové sítě se sigmoidální aktivační funkcí vycházející z biologické motivace – s výjimkou polynomů zaručí univerzalitu téměř jakákoliv aktivační funkce.

Věta 2.2 *Nechť $\psi : \mathcal{R} \to \mathcal{R}$ je lokálně omezená po částech spojitá funkce. Potom $\mathcal{P}_d(\psi)$ je hustá v $\mathcal{C}([0,1]^d)$ pro všechna kladná přirozená čísla d právě když ψ není algebraický polynom.*

Důkaz této věty využívá možnost vyjádřit všechny mocniny, tj. funkce x^k, jako limity parciálních derivací dostatečně vysokého řádu vzhledem k proměnné v funkce $\psi(vx+b)$ (za předpokladu, že ψ je analytická funkce, která není polynomem) a na pozorování, že formule definující parciální derivaci vzhledem k parametru v $\lim_{\eta \to 0} \frac{(\psi((v+\eta)x+b) - \psi(vx+b)}{\eta}$ je limitou posloupnosti funkcí, které lze počítat perceptronovými sítěmi s aktivační funkcí ψ. Těmito sítěmi lze tedy aproximovat s libovolnou přesností funkci $\frac{\delta^k \psi(vx+b)}{\delta v^k} = x^k \psi^{(k)}(vx+b)$. Pokud ψ není polynom, derivace jakéhokoliv řádu funkce ψ nemůže být identicky rovna nule. Po dosazení parametrů $v = 0$ a b_k, pro které $\psi^{(k)}(b_k) = c_k \neq 0$, dostaneme posloupnost sítí, které aproximují funkci $c_k x^k$. Jestliže všechny mocniny mohou být aproximovány sítěmi daného typu, pak polynomy, které jsou lineárními kombinacemi mocnin, mohou být aproximovány sítěmi, které jsou lineární superpozicí sítí počítajících mocniny. Jestliže lze pomocí funkcí z $\mathcal{P}(\psi)$ aproximovat s libovolnou přesností všechny polynomy a pomocí polynomů všechny spojité funkce na $[0,1]^d$, je $\mathcal{P}_d(\psi)$ hustá v $\mathcal{C}([0,1]^d$.

Hornik (1993) si povšiml, že důkaz, který podali Leshno et al. platí i tehdy, jsouli všechny parametry sítě omezeny shora, a to dokonce libovolně malou konstantou. To znamená, že univerzální aproximační vlastnost platí i pro perceptronové sítě s omezenými vahami (pokud je aktivační funkce nepolynomiální a analytická).

Avšak parametry sítě nemohou být omezené současně zdola i shora – k tomu abychom zachovali hustotu, je totiž třeba, aby množina parametrů skrytých jednotek měla buďto konečný nebo nekonečný akumulační bod.

Další standartní metoda ověřování hustoty množin funkcí je založena na Hahn-Banachově větě: stačí ověřit, že každý lineární funkcionál, který je nulový na této

množině, musí být roven nule na celém lineárním prostoru. Tuto metodu použil pro důkaz univerzální aproximační vlastnosti sigmoidálních perceptronových sítí Cybenko (1989). Jiné metody důkazu hustoty perceptronových sítí jsou založeny na vhodných funkcionálních reprezentacích – např. integrální reprezentace jako Radonova transformace (Carroll & Dickenson, 1989) nebo Kolmogorova reprezentace spojitých funkcí několika proměných pomocí superpozice spojitých funkcí jedné proměnné (Kůrková, 1992).

Ačkoliv důkazová technika výše uvedené věty o redukci vstupní dimenze neumožňuje rozšíření na sítě s radiálními jednotkami, pro mnohé radiální funkce lze odvodit univerzální aproximační vlastnost radiálních sítí pomocí konvolucí, které lze použít pro všechny vstupní dimenze. Girosi & Poggio (1990) dokázali univerzální aproximační vlastnost pro Gaussovské radiální sítě a Park & Sandberg (1993) rozšířili jejich výsledek na další radiální funkce. Důkaz následující věty je založen na klasické metodě aproximace funkcí pomocí posloupnosti konvolucí s jádry, která konvergují k Diracově delta funkci.

Věta 2.3 *Pro každé kladné přirozené číslo d a pro každou funkci $\psi : \mathcal{R} \to \mathcal{R}_+$ která má konečný nenulový integrál a pro každou normu $\|.\|$ on $\mathcal{R}^d$ je $\mathcal{B}_d(\psi, \|.\|)$ hustá v $\mathcal{C}(\mathcal{R}^d)$.*

Univerzální aproximační vlastnost dokonce platí i pro Gaussovské radiální sítě s pevnou konstantní šířkou (proměnné jsou pouze parametry odpovídající středům a výstupním vahám) – pomocí vlastností Hermitovských polynomů ji dokázal Mhaskar (1997).

2.2 Nejlepší aproximace a reprezentace funkcí s konečným definičním oborem

V praktických aplikacích neuronových sítí je velikost parametrů a počet jednotek sítě vždy omezený. Nechť $\mathcal{P}_d(\psi, n, c)$, $\mathcal{B}_d(\psi, n, c)$, resp., značí množinu funkcí, které lze počítat perceptronovými, radiálními, resp., sítěmi s jednou skrytou vrstvou s aktivační, radiální resp., funkcí ψ s nejvýše n skrytými jednotkami a se všemi parametry omezenými shora konstantou c, tj. splňujícími $|w_i| \leq c$, $|b_i| \leq c$ a $\|\mathbf{v}_i\| \leq c$ pro všechna $i = 1, \ldots, n$.

Následující věta (Kůrková, 1995) ukazuje, že množiny funkcí, které lze počítat sítěmi s jednou skrytou vrstvou s omezeným počtem jednotek i velikostí parametrů jsou proximinální v prostorech $\mathcal{C}([0,1]^d)$. To znamená, že existuje *globální minimum chybového funkcionálu*. V terminologii teorie aproximace funkcí se tato vlastnost nazývá *vlastnost nejlepší aproximace*.

Věta 2.4 *Pro libovolná kladná přirozená čísla d, n, pro každé reálné číslo c a pro každou omezenou spojitou funkci $\psi : \mathcal{R} \to \mathcal{R}$, $\psi : \mathcal{R} \to \mathcal{R}_+$, resp., je množina $\mathcal{P}_d(\psi, n, c)$, $\mathcal{B}_d(\psi, n, c)$, resp., proximinální podmnožina $\mathcal{C}([0,1]^d)$.*

Neuronové sítě bývají používány pro aproximaci funkcí daných konečnými množinami dvojic vstup/výstup. Jedná se tedy o funkce s konečným definičním oborem

a proto se na ně vztahují výsledky teorie interpolace. Tyto výsledky ukazují, že v případě funkcí definovaných pouze na konečné podmnožině $\mathcal{R}^d$ můžeme nahradit univerzální aproximaci přesnou *reprezentací*. Jeden z hlavních výsledků interpolační teorie, Micchelliho věta (Micchelli, 1986), ukazuje, že každá reálná funkce na konečné podmnožině $\mathcal{R}^d$ může být interpolována lineární kombinací Gaussovských funkcí s vhodně nastavenými středy a šířkami. V terminologii neuronových sítí to znamená, že každá funkce s konečným definičním oborem může být přesně reprezentována sítí s Gaussovskými radiálními jednotkami. Obdobný výsledek pro sigmoidální perceptronové sítě dokázal Ito (1992). Avšak tyto přesné reprezentace vyžadují lineární kombinaci stejně velkého počtu funkcí jako je definiční obor reprezentované funkce, což v případě neuronových sítí znamená stejný počet jednotek ve skryté vrstvě jako je počet dvojic dat vstup/výstup.

Vybudování interpolační teorie bylo motivováno potřebou sestrojovat povrchy určitých vhodných vlastností tak, aby procházely daným malým počtem bodů. Přestože výsledky této teorie platí pro libovolný konečný počet bodů, jejich využití v oblasti neuronových sítí je značně omezeno, neboť konstrukce sítí se stejným počtem skrytých jednotek jako je počet dvojic vstup/výstup v případě velkých množin dat naráží na meze dané implementací. Sítěmi s menším počtem jednotek je možno získat pouze aproximaci, jejíž přesnost lze odhadnout pomocí obdobných metod jako pro funkce s nekonečným definičním oborem.

3 Odhady rychlosti aproximace

Univerzalita nemůže být nikdy dosažena v rámci prakticky realizovatelných mezí složitosti. Každý univerzální výpočetní model určuje jinou hierarchii složitosti výpočetních úloh. Složitost neuronových sítí lze měřit pomocí různých měr složitosti definovaných na základě možností implementace. V případě simulace sítí na klasických počítačích je důležitou mírou složitosti počet jednotek sítě.

3.1 Lineární a nelineární aproximace

V teorii aproximace funkcí je studována rychlost konvergence aproximační chyby vypovídající o závislosti přesnosti aproximace na složitosti aproximující funkce. Pokud aproximující funkce patří do nějaké parametrické množiny funkcí, pak lze složitost aproximujících funkcí měřit délkou vektoru parametrů (odpovídajícího např. stupni polynomu nebo racionální funkce, počtu uzlů splinu nebo počtu skrytých jednotek v neuronové síti). Takové parametrické množiny mohou být reprezentovány jako posloupnosti navzájem do sebe vnořených množin funkcí s postupně rostoucí délkou parametrických vektorů. V tradičních aproximačních schematech (jako jsou polynomy nebo řady funkcí) tvoří tyto množiny do sebe vnořené lineární podprostory rostoucí dimenze. Aproximací pomocí funkcí z lineárních podprostorů se zabývá teorie *lineární aproximace*.

Je-li $\{Y_n; n \in \mathcal{N}_+\}$ posloupnost do sebe vnořených podmnožin normovaného lineárního prostoru $(X, \|.\|)$, pak *rychlost aproximace* funkce f pomocí $\{Y_n; n \in \mathcal{N}_+\}$ je měřena

pomocí rychlosti poklesu hodnoty chybových funkcionálů $e_{Y_n}(f)$. Rychlost aproximace množiny funkcí K je charakterizována chybou nejhůře aproximované funkce odpovídající odchylce $\delta(K, Y_n) = \sup_{f \in K} e_{Y_n}(f) = \|f - Y_n\|$.

Pokud je $\bigcup_{n \in \mathcal{N}_+} Y_n$ hustá podmnožina X, máme sice teoreticky zaručeno, že pro každou funkci f posloupnost $e_{Y_n}(f)$ konverguje k 0, ale pro praktické aplikace je třeba, aby tato konvergence byla dostatečně rychlá, tj. aby požadovaná přesnost aproximace byla dosažena pro takové n, pro které je ještě možné všechny funkce z množiny Y_n implementovat.

V případě funkcí několika proměnných, tj. když množina aproximovaných funkcí je tvořena funkcemi na $\mathcal{R}^d$, se stává, že odchylka je řádově $\mathcal{O}(\frac{1}{\sqrt[d]{n}})$. To znamená, že pro dosažení přesnosti aproximace ε pro všechny funkce z aproximované množiny je třeba použít aproximující funkce složitosti řádově $\left(\frac{1}{\varepsilon}\right)^d$. Složitost tedy závisí na d exponencionálně. Tento jev se nazývá "prokletí dimenzionality", neboť omezuje užití takové aproximační metody na funkce malého počtu proměnných.

Je známo, že k exponencionální závislosti složitosti aproximujících funkcí na počtu proměnných dochází při lineární aproximaci mnoha množin funkcí (viz Pinkus, 1985). Pro nelineární aproximační množiny sice neplatí mnohé z výhodných vlastností (jako např. jednoznačnost, spojitost, homogeneita) projekčních operátorů do lineárních podprostorů, ale tato ztráta je v některých případech kompenzována výrazným zlepšením rychlosti aproximace, někdy dokonce bez exponencionální závislosti složitosti aproximujících funkcí na počtu proměnných.

Aproximace dopřednými neuronovými sítěmi patří do oblasti *nelineární aproximace*. Například množiny funkcí, které lze počítat neuronovými sítěmi s jednou lineární výstupní jednotkou netvoří lineární prostor, jsou totiž sjednocením mnoha lineárních podprostorů (všech konečně dimenzionálních prostorů generovaných funkcemi, které počítají výpočetní jednotky ve skryté vrstvě).

Jestliže $G_\phi = \{\phi(\mathbf{a}, .) : \mathcal{R}^d \to \mathcal{R}; \mathbf{a} \in \mathcal{R}^p\}$ je parametrická množina funkcí odpovídající typu výpočetní jednotky, potom ϕ-síť s n jednotkami ve skryté vrstvě je schopna počítat jako funkce vstup/výstup všechny lineární kombinace n funkcí z množiny G_ϕ, tj. všechny funkce z množiny $span_n G_\phi$. Tato množina je sjednocením všech n-dimenzionálních podprostorů generovaných n-ticemi prvků G_ϕ a je mnohem větší než jediný n-dimenzionální podprostor. Lze tedy očekávat, že pro některé třídy funkcí bude rychlost konvergence při aproximaci neuronovými sítěmi v závislosti na počtu prvků n výrazně rychlejší než v případě aproximace funkcemi pouze z jednoho n-dimenzionálního lineárního podprostoru. V řadě aplikací byly skutečně s potřebnou přesností aproximovány funkce několika stovek proměnných sítěmi s malým počtem jednotek (viz např. Sejnowski & Rosenberg, 1987, Sejnowski & Yuhas, 1991).

3.2 Variace vzhledem k množině funkcí

Popsat vlastnosti množin funkcí mnoha proměnných, které lze aproximovat neuronovými sítěmi s jednotkami daného typu, jejichž počet neroste exponencionálně v závislosti na počtu proměnných, je možné pomocí normy "šité ma míru" typu jednotek neuronové sítě.

Pro ϕ-sítě je rychlost klesání aproximační chyby nejhůře aproximované funkce z

množiny K měřena odchylkou $\delta(K, span_n G_\phi)$ množiny K od množiny $span_n G_\phi$ odpovídající množině funkcí vstup/výstup ϕ-sítě s n jednotkami. Na základě následující věty, která je Barronovým (1993) upřesněním Jonesova (1992) výsledku, lze pro množiny K splňující určité podmínky odvodit horní odhad této odchylky který je řádově $\mathcal{O}(\frac{1}{\sqrt{n}})$ nezávisle na počtu proměnných. Věta je formulována pro aproximaci funkcemi z množiny $conv_n G = \{\sum_{i=1}^n a_i g_i; a_i \in [0,1], \sum_{i=1}^n a_i = 1\}$, avšak plynou z ní důsledky pro aproximaci pomocí $\{span_n G; n \in \mathcal{N}_+\}$.

Věta 3.1 *Nechť $(X, \|.\|)$ je Hilbertův prostor, b kladné reálné číslo a G podmnožina X taková, že pro každé $g \in G$ $\|g\| \leq b$. Potom pro každou $f \in cl\, conv\, G$ a pro každé kladné přirozené číslo n platí $\|f - conv_n G\| \leq \sqrt{\frac{b^2 - \|f\|^2}{n}}$.*

Jones-Barronova věta dává horní odhad rychlosti aproximace pomocí posloupnosti množin $\{conv_n G; n \in \mathcal{N}_+\}$ a to pouze pro funkce v konvexním uzávěru množiny G. Pokud množina G není uzavřená na násobení skalárem, je množina $conv\, G$ vlastní podmnožinou množiny $span\, G$, obdobně též $cl\, conv\, G$ je vlastní podmnožinou $cl\, span\, G$, takže hustota množiny $span\, G$ (tj. $cl\, span\, G = X$) nezaručuje, že všechny prvky X splňují předpoklady věty Věty 3.1. Pokud ale chceme odhadnout rychlost aproximace pomocí posloupnosti $\{span_n G; n \in \mathcal{N}_+\}$, můžeme postupným zvětšováním množiny G tím, že přidáváme násobky všech jejích prvků skaláry až do velikosti c (tj. G nahradíme množinou $G(c) = \{wg; |w| \leq c, g \in G\}$), docílit toho, že libovolný prvek množiny $span\, G$ lze vyjádřit pro dostatečně velké c jako prvek množiny $conv\, G(c)$. Platí totiž $span\, G = \bigcup_{c \in \mathcal{R}_+} conv\, G(c)$. Pro každé $c > 0$ takto odhadneme rychlost aproximace množinami $conv_n G(c)$; vzhledem k tomu, že $conv_n G(c) \subseteq span_n G$, dostaneme tak horní odhad pro všechna $f \in \bigcup_{c \in \mathcal{R}_+} cl\, conv\, G(c)$.

Jones-Barronovu větu tedy můžeme aplikovat i na funkce, které nejsou v konvexním uzávěru množiny G, ale které jsou v konvexním uzávěru množiny $G(c)$ pro nějaké $c > 0$; pak ovšem konstanta b ve jmenovateli horního odhadu musí být vynásobena c. Nejmenší c, které stačí pro danou funkci f k tomu, aby $f \in cl\, conv\, G(c)$, můžeme charakterizovat pomocí normy "šité na míru" množině G (například množině G_ϕ obsahující funkce, které počítají jednotky neuronové sítě).

Připomeňme, že každá norma je jednoznačně určená svou jednotkovou koulí: je-li totiž $B_1(\|.\|) = \{f \in X; \|f\| \leq 1\}$ jednotková koule vzhledem k normě $\|.\|$, potom pro každé $f \in X$ platí $\|f\| = \inf\{\lambda \in \mathcal{R}_+; \frac{1}{\lambda} f \in B_1(\|.\|)\}$. Má-li tedy množina $B \subset X$ všechny vlastnosti jednotkové koule, můžeme pomocí formule $\|f\| = \inf\{\lambda \in \mathcal{R}_+; \frac{1}{\lambda} f \in B\}$ definovat normu $\|.\|$ na lineárním prostoru X (tato formule se nazývá *Minkowského funkcionál* množiny B). K tomu, aby množina mohla být jednotkovou koulí nějaké normy, musí být tato množina uzavřená, konvexní a balancovaná, tj. musí platit $cl\, B = B$, $conv\, B = B$ a $\lambda g \in B$ pro všechna $\lambda \in [-1, 1]$ a všechna $g \in B$.

Norma nazvaná *G-variace* (variace vzhledem k množině G) je definována jako Minkowského funkcionál uzavřené konvexní balancované množiny $cl\, conv\, G(1)$. G-variace je tedy norma na podprostoru $\{f \in X; \|f\|_G < \infty\} \subseteq X$. Snadno lze ověřit, že $\|f\|_G = \inf\{c > 0; f \in cl\, conv\, G(c)\}$ a že pro každé $f \in X$ platí $\|f\| \leq \|f\|_G \sup_{g \in G} \|g\|$.

Pojem variace zavedl Barron (1992) pro množimu charakteristických funkcí poloprostorů odpovídající množině funkcí, které počítají perceptrony s nespojitou prahovou

aktivační funkcí. Nazval tento funkcionál *variace vzhledem k poloprostorům*, protože pro funkce jedné proměnné se tento pojem až na konstantní faktor shoduje s pojmem totální variace studované v teorii integrace. Kůrková (1997) zavedla obecný pojem variace vzhledem k množině funkcí G neboli G-variaci.

Pokud je množina funkcí G ortonormální bází separabilního Hilbertova prostoru $(X, \|.\|)$, pak je G-variace rovna l_1-normě vzhledem k této bázi, tj. je-li $f = \sum_{g \in G} f \cdot g$, pak platí $\|f\|_G = \sum_{g \in G} |f \cdot g|$. Například variace vzhledem k Fourierově bázi je spektrální norma. Pojem G-variace je tedy kromě zobecnění pojmu totální variace také zobecněním pojmu l_1-normy.

Následující věta je ekvivalentem Jones-Barronovy věty formulovaným pomocí pojmu G-variace (Kůrková, 1997, 1998). G^0 značí množinu normalizovaných prvků množiny G, tj. $G^0 = \{\frac{g}{\|g\|}; g \in G\}$ (používáme ji zde proto, abychom nemuseli konstantu ve jmenovateli násobit $\sup_{g \in G} \|g\|$).

Věta 3.2 *Nechť $(X, \|.\|)$ je Hilbertův prostor a G je jeho podmnožina. Potom pro všechna $f \in X$ a pro každé kladné přirozené číslo n platí $\|f - span_n G\| \leq \frac{\|f\|^2_{G^0} - \|f\|^2}{n}$.*

Jonesův důkaz i Barronova modifikace tohoto důkazu jsou konstruktívní – jsou založeny na horním odhadu chyby $\|f - conv_n G\|$ vyjádřeném pomocí rekurzívní formule. Stejný horní odhad $\|f - conv_n G\|$ jako v Jones-Barronově větě odvodil Maurey pomocí pravděpodobnostního argumentu (viz Barron, 1993) považujícího reprezentace funkce f jakožto prvku množiny $conv\, G$, $f = \sum_{i=1}^m a_i g_i$, za konečné pravděpodobnostní rozdělení na množině G definované $P(g = g_i) = a_i$. Je-li $f_n = \sum_{j=1}^n \frac{1}{n} h_j$ náhodná proměnná odpovídající barycentru n-tice prvků G vybraných z G s pravděpodobností P, pak lze ukázat že střední hodnota $\|f - f_n\|^2$ je omezená shora $\frac{b^2 - \|f\|^2}{n}$. Musí tedy existovat nějaká n-tice $h_1, \ldots, h_n \in G$ taková, že $\|f - \sum_{j=1}^n \frac{1}{n} h_j\| \leq \sqrt{\frac{b^2 - \|f\|^2}{n}}$. Rozšíření tohoto odhadu na $cl\, conv\, G$ dává tutéž horní mez jako Jones-Barronova věta. Takže Jones-Barronova věta odhaduje shora rychlost konvergence nejlepších aproximantů z množiny $conv_n G$ pomocí rychlosti konvergence barycentrů průměrných n-tic prvků G.

Darken et al. (1993) rozšířili Jones-Barronovu větu na $\mathcal{L}_p$-prostory pro $p \in (1, \infty)$ – v tomto případě je rychlost konvergence o něco pomalejší, řádově $\mathcal{O}(n^{-\frac{1}{q}})$, kde $q = \max(p, \frac{p}{p-1})$. Darken et al. také ukázali, že Jonesův konstruktívní důkaz nelze použít v případě norem, které nejsou hladké (v normovaných lineárních prostorech, v nichž má jednotková koule ostrý roh). Avšak Maurayho pravděpodobností důkaz může být modifikován tak, aby platil i pro supremovou normu (Barron, 1992, Girosi, 1995, Gurvits & Koiran, 1997 , Kůrková, Savický & Hlaváčková, 1998).

Jonesův konstruktívní důkaz věty 3.1 a jeho různá rozšíření dávají navíc kromě horního odhadu aproximační chyby také důkaz konvergence inkrementálních aproximantů. Tento důkaz slouží jako teoretický základ pro studium *inkrementálních algoritmů* učení neuronových sítí, tj. algoritmů určujících parametry sítě pomocí posloupnosti kroků, kde v každém kroku je přidána nová skrytá jednotka (viz Kůrková, 1998).

3.3 Dimenzionálně nezávislá aproximace

Jones-Barronova věta dává horní odhad rychlosti aproximace funkcemi z posloupnosti množin tvaru $\{span_n G; n \in \mathcal{N}_+\}$ vyjádřený pomocí dvou norem aproximované funkce: její G-variaci $\|f\|_G$ a její normě $\|f\|$, kde $\|.\|$ je norma, v níž je měřena chyba aproximace. Následující důsledek Věty 3.2 ukazuje, že pro všechny funkce s G-variací omezenou danou konstantou r (tj. pro funkce z koule $B_r(\|.\|_G)$ o poloměru r) je tato rychlost shora omezená $\frac{r}{\sqrt{n}}$.

Důsledek 3.3 *Nechť $(X, \|.\|)$ je Hilbertův prostor a G je jeho podmnožina. Potom pro každé kladné přirozené číslo n platí $\delta(B_r(\|.\|_G), span_n G) \leq \frac{r}{\sqrt{n}}$.*

Aplikujeme-li tento horní odhad na aproximaci funkcí několika proměnných (například z Hilbertova prostoru $\mathcal{L}_2([0,1]^d)$, dostaneme popis množin funkcí několika proměnných, které lze aproximovat množinami typu $\{span_n G; n \in \mathcal{N}_+\}$ bez "prokletí dimenzionality". Důsledek 3.3 totiž zaručuje pro všechny funkce z $B_r(\|.\|_G)$ přesnost ε při aproximaci funkcemi ze $span_n G$, kde $n = \left(\frac{1}{\varepsilon}\right)^2$ nezávisle na počtu proměnných d. Složitost aproximujících funkcí tedy neroste exponencionálně s počtem proměnných, ale závisí na $\frac{1}{\varepsilon}$ pouze kvadraticky bez ohledu na to, kolik proměnných mají aproximované funkce. Například, všechny funkce s G_ϕ-variací menší nebo rovnou r lze aproximovat ϕ-sítěmi s počtem jednotek nerostoucím exponencionálně s počtem vstupních proměných – chyba nejhůře aproximované funkce mezi nimi je při aproximaci ϕ-sítěmi s n jednotkami ve skryté vrstvě nejvýše $\frac{r}{\sqrt{n}}$. Rychlost aproximace tedy nezávisí na vstupní dimenzi d; na vstupní dimenzi ale závisí velikost množin $B_r(\|.\|_G)$.

Abychom mohli použít Jones-Barronovu větu pro odhad chyby při aproximaci neuronovými sítěmi, je ovšem třeba odhadnout G_ϕ-variaci pro ϕ odpovídající standartním výpočetním jednotkám. Jednou z nejčastěji používaných skrytých jednotek je perceptron se sigmoidální aktivační funkcí. Nejjednodušší sigmoida je nespojitá prahová Heavisidova funkce ϑ. Vzhledem k tomu, že množina P_ϑ funkcí, které počítají perceptrony s Heavisidovou aktivační funkcí, je totožná s množinou charakteristických funkcí poloprostorů $\mathcal{R}^d$ omezených na $[0,1]^d$, nazývá se P_ϑ-variace *variace vzhledem k poloprostorům* (přesněji variace vzhledem k charakteristickým funkcím poloprostorů).

Variace vzhledem k poloprostorům (P_ϑ-variace) se v $\mathcal{L}_p$-prostorech rovná P_σ-variaci pro libovolnou spojitou sigmoidální funkci σ (Kůrková, Kainen & Kreinovich, 1997). Stačí se tedy při studiu rychlosti aproximace sigmoidálními perceptronovými sítěmi omezit na variaci vzhledem k poloprostorům.

Jedna z možností, jak odhadnout shora variaci vzhledem k množině funkcí výpočetních jednotek, je odhadnout ji pomocí variace nějaké vhodné ortogonální množiny, pro niž variace odpovídá $\mathcal{L}_1$-normě. Barron (1993) takto odhadl variaci vzhledem k poloprostorům pomocí varianty spektrální normy ($\mathcal{L}_1$-normy vzhledem k Fourierově bázi vynásobené vhodnými vahami) a ukázal příklady funkcí, pro které je aproximace perceptronovými sítěmi výrazně lepší než lineární aproximace. Popsal totiž množiny funkcí, pro které při aproximaci perceptronovými sítěmi počet prvků potřebných pro danou přesnost aproximace roste jen kvadraticky, zatímco lineární aproximace vyžaduje dimenzi lineárního podprostoru rostoucí exponencionálně s počtem proměnných.

Pojem variace lze také popsat geometricky na základě úhlu mezi aproximovanou funkcí a aproximujícími funkcemi. Následující věta, kterou dokázali Kůrková, Savický & Hlaváčková (1998), dává takovou geometrickou interpretaci. $G^\perp$ značí *ortogonální doplněk* množiny G.

Věta 3.4 *Nechť $(X, \|.\|)$ je Hilbertův prostor a G je jeho neprázdná podmnožina. Potom pro všechna $f \in X$ platí $\|f\|_G = \sup_{h \in S} \frac{|f \cdot h|}{\sup\limits_{g \in G} |g \cdot h|}$, kde $S = \{h \in X - G^\perp; \|h\| = 1\}$.*

Pokud tedy f není ortogonální ke G, pak $\|f\|_G \geq \frac{\|f\|}{\sup_{g \in G} |f \cdot g|}$, což znamená, že čím větší je úhel $\arccos |f \cdot g|$, tím větší je G-variace funkce f. Velkou G-variaci tedy mají funkce, které jsou "téměř ortogonální" k množině G.

Kůrková, Savický & Hlaváčková (1998) využili tuto geometrickou charakterizaci G-variace pro popis třídy funkcí s variací vzhledem k poloprostorům rostoucí s počtem proměnných exponencionálně. Tyto příklady popisují typy funkcí, které bude patrně obtížné aproximovat perceptronovými sítěmi. Není však známo, zda pro funkce s velkou variací vzhledem k poloprostorům je rychlost aproximace pomocí neuronových sítí skutečně tak pomalá, že dosahuje horní odhad plynoucí z Jones-Barronovy věty. Pro tyto funkce totiž není znám žádný dolní odhad rychlosti aproximace neuronovými sítěmi. Jedná se obtížný otevřený problém související se složitostí Boolovských obvodů (viz Hajnal et el., 1987).

Pokud umíme vyjádřit aproximovanou funkci pomocí integrální rovnice odpovídající ϕ-síti s kontinuem skrytých jednotek, můžeme alespoň teoreticky odhadnout variaci této funkce vzhledem k množině funkcí výpočetních jednotek. Následující věta, kterou dokázali Kůrková, Kainen & Kreinovich (1997), dává takový horní odhad variace.

Věta 3.5 *Nechť d, p jsou kladná přirozená čísla, $J \subseteq \mathcal{R}^d$ a $f \in (\mathcal{C}(J), \|.\|_\mathcal{C})$ je funkce, kterou lze reprezentovat jako $f(\mathbf{x}) = \int_Y w(\mathbf{a})\phi(\mathbf{a}, \mathbf{x})d\mathbf{a}$, kde $A \subseteq \mathcal{R}^p$, $w \in \mathcal{C}(A)$ má kompaktní nosič a nechť $G_\phi = \{\phi(\mathbf{a}, .) : J \to \mathcal{R}; \mathbf{a} \in Y\}$. Potom $\|f\|_{G_\phi} \leq \int_A |w(\mathbf{a})|d\mathbf{a}$.*

Integrální rovnice, jejíž platnost je předpokládána ve větě 3.5, se nazývá *lineární integrální rovnice prvního řádu s jádrem ϕ*. Takže G_ϕ-variace (vzhledem k supremové normě) integrální transformace w je omezená shora $\mathcal{L}_1$-normou váhové funkce w. Pomocí vhodných integrálních reprezentací odvodili různí autoři odhady G_ϕ-variace pro některé typy výpočetních jednotek: např. Barron (1993) použil k odhadu variace vzhledem k poloprostorům variaci vzhledem k určité variantě Fourierovy báze, kterou odhadl pomocí vážené Fourierovy transformace, Girosi & Anzellotti (1993) odhadli variaci vzhledem ke Gaussovským radiálním funkcím pomocí Fourierovy reprezentace funkcí s omezenými frekvencemi, Kůrková, Kainen & Kreinovich (1997) odvodili integrální reprezentaci ve tvaru neuronové sítě s kontinuem perceptronů s Heavisidovou aktivační funkcí $\{\vartheta(\mathbf{e} \cdot \mathbf{x} + b); \mathbf{e} \in S^{d-1}, b \in \mathcal{R}\}$ s výstupními vahami $w(\mathbf{e}, b)$ odpovídajícími ortogonálním "průtokům řádu d" funkce f nulovou nadrovinou $H_{\mathbf{e}b}$ perceptronu s vahou $\mathbf{e}$ a prahem b (vektor $\mathbf{e}$ je kolmý k nadrovině $H_{\mathbf{e},b}$, určuje tedy její směr, zatímco práh b určuje její posun vzhledem k počátku souřadnic).

3.4 Spojitost a přesnost aproximace

V klasické lineární aproximační teorii lze využít pro odhady rychlosti aproximace výhodné vlastnosti aproximačních operátorů jako je jednoznačnost, homogeneita (tj. pro všechna $f \in X$ a pro všechna $\lambda \in \mathcal{R}$ platí $p_Y(\lambda f) = \lambda p_y(f)$) a spojitost. Následující věta, kterou dokázali Kainen, Kůrková & Vogt (1999), ukazuje, že v některých normovaných prostorech geometrické vlastnosti množin aproximujících funkcí způsobují, že neexistuje žádný spojitý operátor nejlepší aproximace.

Věta 3.6 *Nechť $(X, \|.\|)$ je striktně konvexní Banachův prostor se striktně konvexním duálem a nechť Y je jeho nekonvexní podmnožina. Potom $P_Y : X \to \mathcal{P}(Y)$ nemá spojitý výběr.*

Předpoklady této věty splňují například všechny prostory $\mathcal{L}_q([0,1]^d)$ pro $q \in (1, \infty)$. Sjednocení všech n-dimenzionálních podprostorů množiny G, $span_n G$, je konvexní jen v krajním případě, pokud $span_n G$ tvoří lineární podprostor. Pro mnohé výpočetní jednotky ϕ jsou množiny $span_n G_\phi$ nekonvexní. Pro sítě s takovými jednotkami nelze dosáhnout nejlepší aproximaci spojitým způsobem a to dokonce ani v případě, když se omezíme na libovolně malé okolí nuly. Navíc za určitých dodatečných podmínek na množinu výpočetních jednotek nelze pomocí spojitého aproximačního operátoru získat ani aproximaci, která se od nejlepší aproximace odchyluje o předem danou libovolně malou konstantu ε (viz Kainen, Kůrková & Vogt, 1999).

Geometrické vlastnosti množin funkcí počítatelných neuronovými sítěmi tedy způsobují nespojitost aproximačních operátorů, což má za následek, že odhady přesnosti aproximace neuronovými sítěmi odvozené pomocí spojitých aproximačních operátorů nemohou být přesné. Rovněž metody odhadu dolních mezí přesnosti aproximace, které značně omezují lineární aproximační metody, se nedají použít pro neuronové sítě – tyto meze vyznačující se exponencionálním růstem složitosti s počtem proměnných jsou totiž odvozeny na základě spojitosti (viz Pinkus, 1985, DeVore at al., 1989).

Literatura

Barron, A. R. (1992). Neural net approximation. In *Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems* (pp. 69–72).

Barron, A. R. (1993). Universal approximation bounds for superposition of a sigmoidal function. *IEEE Transactions on Information Theory*, **39**, 930–945.

Burr, D. J. (1998). Experiments on neural net recognition of spoken and written text. *IEEE Trans. Acoust. Speech and Signal Processing*, **36**, 1162–1168.

Braess D.(1986). *Nonlinear Approximation Theory.* Berlin: Springer.

Carroll, S. M. & Dickinson, B. W. (1989). Construction of neural nets using the Radon transform. In *Proceedings of IJCNN'89* (pp. I. 607–611). New York: IEEE Press.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics Control Signals Systems*, **2**, 303–314.

Darken, C., Donahue, M., Gurvits, L. & Sontag, E. (1993). Rate of approximation results motivated by robust neural network learning. In *Proceedings of the 6th Annual ACM Conference on Computational Learning Theory* (pp. 303–309). New York: ACM.

Girosi, F., & Anzellotti, G. (1993). Rates of convergence for radial basis function and neural networks. In *Artificial Neural Networks for Speech and Vision* (pp.97–113). London: Chapman & Hall.

Girosi, F. (1995). Approximation error bounds that use VC-bounds. In *Proceedings of ICANN'95* (pp. 295–302). Paris: EC2 & Cie.

Girosi, F. & Poggio, T. (1990). Networks and the best approximation property. In *Biological Cybernetics*, **63**, 169–176.

Gorman, R. P., & Sejnowski, J.: Learned classification of sonar targets using massively parallel network. *IEEE Trans. Acoust. Speech and Signal Processing* **36** (1988), 1135-1140.

Gurvits, L. & Koiran, P. (1997). Approximation and learning of convex superpositions. *Journal of Computer and System Sciences*, **55**, 161–170.

Hajnal, A., Maass, W., Pudlák, P., Szegedy, M., & Turán, G. (1987). Threshold circuits of bounded depth. In *Proceedings of ASFCS* (pp. 99-110). IEEE.

Hornik, K. (1993). Some new results on neural network approximation. *Neural Networks*, **6**, 1069–1072.

Hornik, K., Stinchcombe M. & White H. (1989). Multi-layer feedforward networks are universal approximators. *Neural Networks*, **2**, 359–366.

Hecht-Nielsen R. (1990). *Neurocomputing*. New York: Addison-Wesley.

Ito, Y. (1992). Finite mapping by neural networks and truth functions. *Mathematical Scientist*, **17**, 69–77.

Jones, L. K. (1992). A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Annals of Statistics*, **20**, 608–613.

Kainen, P. C., Kůrková, V. & Vogt, A. (1999). Approximation by neural networks is not continuous. *Neurocomputing* (in press).

Kůrková, V. (1992). Kolmogorov's theorem and multilayer neural networks. *Neural Networks*, **5**, 501–506.

Kůrková, V. (1995). Approximation of functions by perceptron networks with bounded number of hidden units. *Neural Networks*, **8**, 745–750.

Kůrková, V. (1997). Dimension-independent rates of approximation by neural networks. In *Computer-Intensive Methods in Control and Signal Processing: Curse of Dimensionality* (Eds. K. Warwick, M. Kárný) (pp. 261–270). Boston: Birkhauser.

Kůrková, V. (1998). Incremental approximation by neural networks. In *Complexity: Neural Network Approach.* (Eds. K. Warwick, M. Kárný, V. Kůrková) (pp. 177–188). London: Springer.

Kůrková, V., Kainen, P. C. & Kreinovich, V. (1997). Estimates of the number of hidden units and variation with respect to half-spaces. *Neural Networks*, **10**, 1061–1068.

Kůrková, V., Savický , P. & Hlaváčková, K. (1998). Representations and rates of approximation of real-valued Boolean functions by neural networks. *Neural Networks*, **11**, 651–659.

Leshno, M., Lin, V. Y, Pinkus, A. & Schocken, S. (1993). Multilayer feedforward networks with a non-polynomial activation can approximate any function. *Neural Networks*, **6**, 861–867.

McCulloch, W. S., Pitts, W.: A logical calculus immanent in nervous activity. *Buletin of Mathematical Biophysics*, **5** (1943).

Mhaskar H. N. & Micchelli, C. A. (1992). Approximation by superposition of sigmoidal and radial basis functions. *Advances in Applied Mathematics*, **13**, 350–373.

Mhaskar, H. N. & Micchelli, C. A. (1994). Dimension-independent bounds on the degree of approximation by neural networks. *IBM Journal of Research and Development*, **38**, 277–284.

Mhaskar, H. N. (1995). Versatile Gaussian networks. In *Proceedings of IEEE Workshop of Nonlinear Image Processing* (pp. 70–73).

Micchelli, C. A. (1986). Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive Approximation*, **2**, 11–22.

Minsky, M. & Papert, S. (1969). *Perceptrons.* Cambridge: MIT Press.

Park, J., & Sandberg, I. W. (1993). Approximation and radial-basis-function networks. *Neural Computation*, **5**, 305–316.

Pinkus, A.(1985) *n-Width in Approximation Theory.* Berlin: Springer.

Rosenblatt F. (1958). The perceptron: A probabilistic model for information storage and organization of the brain. *Psychological Review*, **65**, 386–408.

Sejnowski, T. J. & Rosenberg, C. (1987). Parallel networks that learn to pronounce english text. *Complex Systems* **1**, 145–168.

Sejnowski, T. J., & Yuhas, B. P. (1991). Mapping between hight-dimensional representation of acoustic and speech signals. In *Computation and Cognition* (pp. 52–68). Philadelphia: Siam.

Singer I. (1970) *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces.* Berlin: Springer.

Stinchcombe, M. & White, H. (1990). Approximating and learning unknown mappings using multilayer networks with bounded weights. In Proceedings of IJCNN'90 (pp. III. 7–16). New York: IEEE Press.