PREDIKCE POČTU UCHAZEČŮ O STUDIUM S VYUŽITÍM NEURONOVÝCH SÍTÍ

P. Matušík

Evropský polytechnický institut, s.r.o, Osvobození 699, 686 04 Kunovice

Abstract

Neuronové sítě se v době využívají v řadě vědních oborů a praktických řešení pro předvídání nejrůznějších událostí jako je například predikce burzovních indexů, předvídání cen, finančních kurzů atd. Neuronové sítě se využívají nejen v ekonomické sféře, ale i v dalších průmyslových odvětvích, např. pro predikci spotřeby energií, v navigacích, ve zdravotnictví atd. Předkládaný článek se zabývá využitím neuronových sítí pro predikci počtu uchazečů o studium na našem Soukromém gymnáziu a střední odborné škole s právem státní jazykové zkoušky, s.r.o. Kunovice. Neuronové sítě mají pro predikci výhodu v tom, že jsou schopny vystihnout i silně nelineární závislosti a to pouze z předložené trénovací množiny vstupních hodnot a ověřením jejích schopností na množině testovacích hodnot.

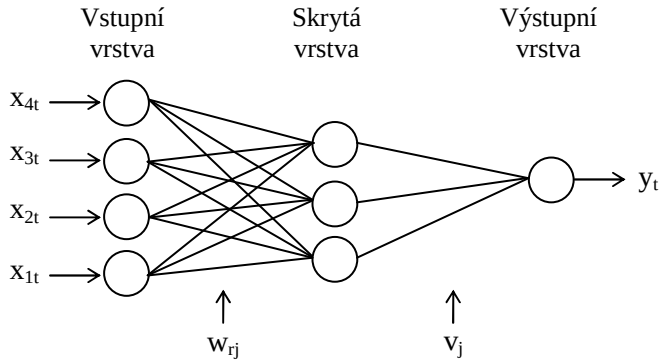
1 Úvod

Cílem této úlohy bylo navrhnout predikční model pomocí neuronové sítě, sloužící k předběžnému určování počtu nově nastupujících studentů ke studiu na naši střední školu. Tato predikce se řídí na základě definovaných vstupních parametrů. Výhodou v použití neuronové sítě je to, že se síť může učit na předložených příkladech a po adaptaci může vystihnout nelineární závislosti. Nevýhodou ovšem je, že se tato síť může naučit predikovat na základě závislosti v trénovací a testovací množině dat, ale pro předpověď dalších neznámých hodnot získáváme data s velkou chybou.

Vzhledem k povaze řešené úlohy byla vybrána asi nejčastěji používaná neuronová síť s dopředným šířením signálu. Jde o síť, kterou tvoří jedna skrytá vrstva s jedním výstupním neuronem a vrstvou obsahující vstupní neurony.

2 Architektura sítě a vstupní data

Jak již bylo řečeno výše, použili jsme pro náš model neuronovou síť s dopředným šířením signálu j jednou skrytou vrstvou viz. obr. 1.



Obr. č. 1: Topologie neuronové sítě

V nejnižší vstupní vrstvě jsou umístěny neurony, pomocí kterých vstupují do sítě data. Ty jsou označeny jako vektor $x_t = (x_{1t}, x_{2t}, x_{3t}, x_{4t})$, t = 1, 2, … , N. N určuje počet sérií vstupních dat. Neurony

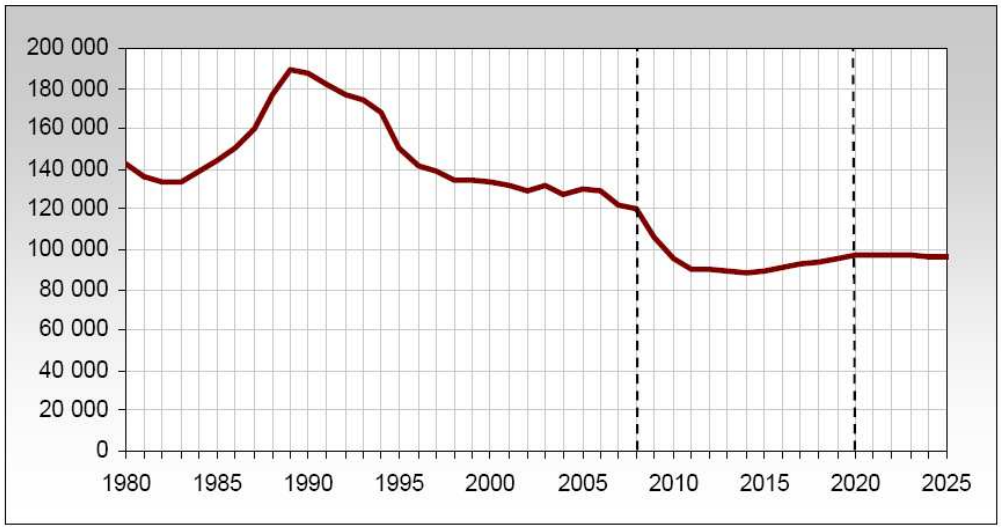
ze skryté střední vrstvy jsou zcela propojeny s neurony ze vstupní vrstvy. Těmto propojením se říká synaptické spojení. V tomto propojení na data z první vstupní vrstvy působí tzv. synaptické váhy w_{rj}. Neurony ve střední vrstvě potom vytvoří skalární součin vektorů x_t a w_{rj}. Obdobné je to mezi skrytou a výstupní vrstvou. Vektor synaptických vah je definován počtem neuronů ve vstupní vrstvě a vrstvě skryté, respektive počtem ve vrstvě skryté a vrstvě výstupní.

2.1 Vstupní data

V posledních letech narůstají obavy z výrazného poklesu počtu osob v populačních ročnících přicházejících do středních škol a v souvislosti s tím i z dramatických změn vzdělanostní struktury žáků ve středním a vyšším vzdělávání. Zejména zaměstnavatelé projevovali obavy z možného výrazného poklesu podílu vyučených, což může být důsledkem výrazného zájmu mladých lidí o maturitní úroveň vzdělávání. S poklesem počtu žáků v populačním ročníku bude vlastně relativně narůstat vzdělávací nabídka v této oblasti, zároveň roste počet studijních příležitostí nabízených ve vyšším odborném a vysokoškolském vzdělávání. [3]

Počet obyvatel ve věku ukončení povinné základní docházky (za tento věk budeme považovat stáří 15 let) je jedním z nepřímých faktorů ovlivňujících jak vzdělanostní strukturu žáků středních škol, tak početnost středních škol vůbec. [3]

Vývoj počtu 15letých osob v ČR uvádí obr. č. 2. Údaje od roku 2008 jsou výsledky prognózy ČSÚ vytvořené v roce 2003.



Obr. č. 2: Vývoj počtu 15letých osob v ČR v letech 1980 – 2025

Zdroj: [4]

Z uvedeného grafu je patrné, že rekordního maxima dosáhl počet obyvatel České republiky ve věku 15 let v roce 1989, kdy v tomto věku žilo téměř 190 tisíc osob. Tato výrazná vlna byla spjata s vysokými počty narozených dětí v počátcích 70. let, kdy tehdejší vláda zavedla úspěšná propopulační opatření. [3]

Mezi další faktory ovlivňující počty uchazečů na naši střední školu patří i počet státních i soukromých škol s podobným vzdělávacím zaměřením, počet námi nabízených studijních oborů a velikost platby školného v porovnání k průměrné hrubé mzdě v jednotlivých letech. Z globálního hlediska na nepříznivý vliv počtu studentů na soukromých školách má bezesporu i probíhající finanční krize. Lidé, zejména rodiče, nastupujících studentů na střední školy ovlivňuje ve velké míře i to, zda budou muset platit školné či nikoli. Velký vliv na počet nově nastupujících studentů má i kvalita a úspěšnost náboru marketingového oddělení na veletrzích a školních burzách a úspěšnost tzv. „Nultého ročníku“ určeného žákům 8. a 9. tříd základních škol. Celkově shrnuto kvalita propagace školy.

Nemalou míru úspěšnosti náboru má i pověst školy jako takové a pověst kvality a odbornosti jednotlivých vyučujících na škole a atraktivnost studijních oborů. V našem případě nabízíme studium v oborech Komerční právo, Zahraniční obchod, Jazykové gymnázium a Počítačové elektronické systémy se zaměřením na počítačové sítě, programování nebo počítačovou grafiku. Jako navazujícího studia mohou absolventi využít studia na naší Vyšší odborné škole právní a vysoké škole s bakalářským stupněm studia v oborech Finance a daně, Management a marketing zahraničního obchodu, Elektronické počítače a Ekonomická informatika.

Pro náš případ studie jsme vybrali pouze nějakým způsobem definovatelné údaje, se kterými se dá dále pracovat. Tím jsme dospěli k modelu sítě se čtyřmi vstupními neurony.

Již v úvodu musím říci, že výsledky a kvalitu učení sítě také ovlivnil počet dostupných vstupních údajů s ohledem na délku působení naší školy. První studenti na školu nastupovali ve školním roce 1991/1992 a v průběhu let došlo k velkým vnitřním změnám, např. v počtech a typech studijních oborů.

V podstatě bylo k dispozici pouze 19 sad vstupních hodnot, ze kterých bylo potřeba vytvořit trénovací a testovací množinu dat, což není mnoho. Z tohoto důvodu je otázka, jak se bude neuronová síť chovat pro predikci počtů uchazečů v následujících letech.

3 Učení a simulace neuronové sítě

Před tím než začneme pracovat se vstupními daty, tak pro zjednodušení početní náročnosti a docílení vyšší výpočtové přesnosti je vhodné tyto data transformovat. Transformaci dat vytvoříme pomocí normalizování pomocí vzorce (1), a to tak, že od každé hodnoty odečteme aritmetický průměr množiny dat a podělíme standardní odchylkou dat. [1]

Tedy pro normalizovanou hodnotu platí, že

$$x_n = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \qquad \text{kde} \quad x_n \in \langle 0, 1\rangle, \quad x \in \langle x_{\min}, x_{\max}\rangle \tag{1}$$

Pro zpětný převod z normalizovaných hodnot použijeme vztah

$$x = x_n \cdot \left(x_{\max} - x_{\min}\right) + x_{\min} \qquad \text{kde} \quad x_n \in \langle 0, 1\rangle, \quad x \in \langle x_{\min}, x_{\max}\rangle \tag{2}$$

kde x_n je normalizovaná hodnota a x je hodnota vstupního údaje nijak neupravovaná (původní)

Po převodu vstupních dat můžeme vytvořit soubor vstupních dat pro trénování a následné otestování naučené sítě.

Pro adaptaci vícevrstvé neuronové sítě byla vyvinuta metoda nazývaná back-propagation (metoda zpětného šíření). Tento algoritmus v podstatě umožňuje šíření chyby výstupu až na vstup sítě. Nevýhodou back-propagation algoritmu je jeho schopnost s konvergencí a může se rychle „přeučit“. To znamená, že se neuronová síť naučí téměř přesně generovat požadované výstupy ze zadaných hodnot vstupní trénovací množiny, ale pro testovací množinu vykazuje velké rozdíly mezi generovanou hodnotou a hodnotou reálnou. „Přeučení“ u neuronových sítí znamená něco podobného jako použití příliš velkého stupně polynomu v polynomické regresi. [5]

Naše testovací síť měla následující parametry:

Network Type: *Feed – forvard backprop*

Training function: *Traingdx*

Adaptation learning function: *learngdm*

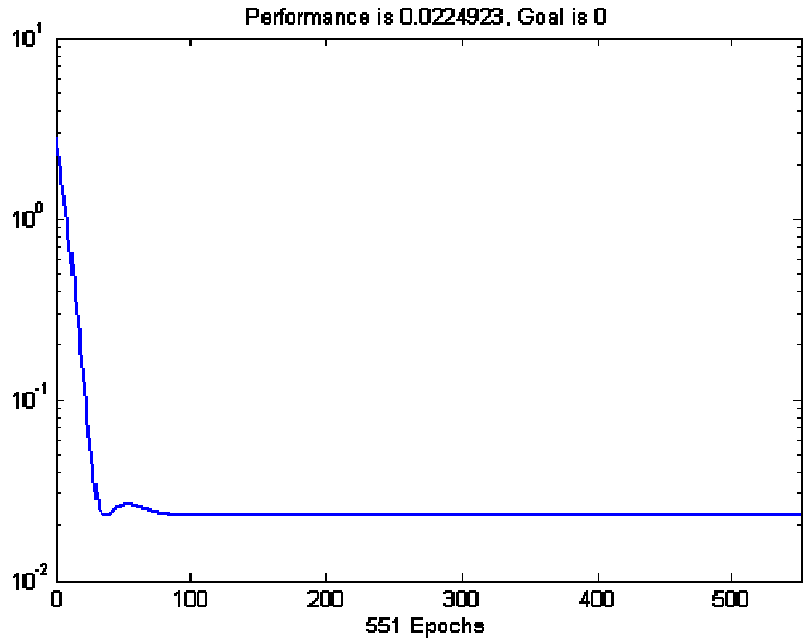
Performance function: *MSE*

Number of layers: *3*

Pro jednotlivé vrstvy byly použity následující aktivační funkce: *tansig, logsig, purelin*

Nastavování jednotlivých parametrů probíhalo experimentálně, proto byla tato část časově nejnáročnější. Samotné učení a výpočty již probíhají docela rychle. Záleží na počtu nastavených epoch učení.

Na následujícím obrázku obr. č. 3 je znázorněna střední kvadratická chyba MSE (mean square error), která byla využita jako vyhodnocovací funkce v závislosti na počtu proběhnutých epoch.



Obr. č. 3: Vývoj MSE v závislosti na počtu vykonaných epoch

4 Závěr

V další fázi testování využití neuronových sítí pro tento problém bude porovnání generovaných výsledků s výsledky založenými na statistických metodách. Bohužel úspěšnost a přesnost predikce se dozvíme cca za rok, po ukončení zápisu studentů ke studiu.

Je jasné, že pro praktické využití mají určitou váhu pouze předpovědi ex ante, tedy ty, u kterých předem neznáme výsledek. Vzhledem k tomu ale nemůžeme hodnotit jejich chyby. Musíme se však uspokojit s tím, že když nám predikční systém generuje prognózy ex post (hodnoty, u kterých předem známe výsledek) s přijatelnou chybou, bude nám generovat v nejbližším časovém horizontu i prognózy ex ante. Kvalitní a správně navržené predikční systémy založené na neuronových sítích to potvrzují.

Literatura

[1] MARČEK, D. – MARČEK. M. *Neuronové siete a ich aplikacie.* EDIS : Žilina, 2006.

[2] GORR, W. L. - NAGIN, D. - SZCZYPULA, J. Comparative study of artificial neural network and statistical models for predicting student grade point averages. *International Journal of Forecasting*. 1994, no. 10, s. 17-34.

[3] VOJTĚCH, J. – CHAMOUTOVÁ, D. – SKÁCELOVÁ, P. *Vývoj vzdělanostní a oborové struktury žáků a studentů ve středním a vyšším odborném vzdělávání v ČR a v krajích ČR a postavení mladých lidí na trhu práce ve srovnání se stavem v Evropské unii 2008/09.* Praha: Národní ústav odborného vzdělávání, 2009.

[4] *Český statistický úřad* [online]. 2009 [cit. 2009-09-12]. Dostupný z WWW: <www.czso.cz>.

[5] KVASNIČKA, V. – POSPÍCHAL, J. – TIŇO, P. *Evolučné algoritmy.* STU Bratislava : Bratislava, 2000.

Author1
Ing. Petr Matušík
matusik@edukomplex.cz