Technická univerzita v Košiciach
Fakulta elektrotechniky a informatiky Katedra kybernetiky a umelej inteligencie

Zadanie z predmetu Hybridná výpočtová inteligencia Využitie HVI v medicínskych aplikáciách

Meno a priezvisko: Dávid Gajdoš
Tomáš Juščík
Richard Kačur
Tomáš Lichanec

Odbor: Inteligentné systémy

Ročník: 4.

Akademicky rok: 2018/2019

Zadanie projektu

Cieľom zadania je využiť nejakú z kombinácii fuzzy systémov a neurónových sietí pri klasifikácií reálnych dát.

Analýza

V rámci analýzi sme si zvolili dáta zo stránky Kaggle (https://www.kaggle.com/jsphyg/weather-dataset-rattle-package?fbclid=IwAR0SXLyz3O4H4CaVWVQaSACSYkdMu-NigQkPglXT5j6QcYSx5O-bRFAmSAU)ktorá zaznamenáva dáta o počasí v priebehu skoro desiatich rokov (1.11.07 – 25.6.17). Záznamy sú merané vždy o 9am a 3pm. Dokopy má náš vybraný dataset približne 3,5 milióna záznamov z 49 rôznych miest naprieč celou Austráliou.
Každé meranie obsahovaloúdaje rozdelené do týchto kategórií:

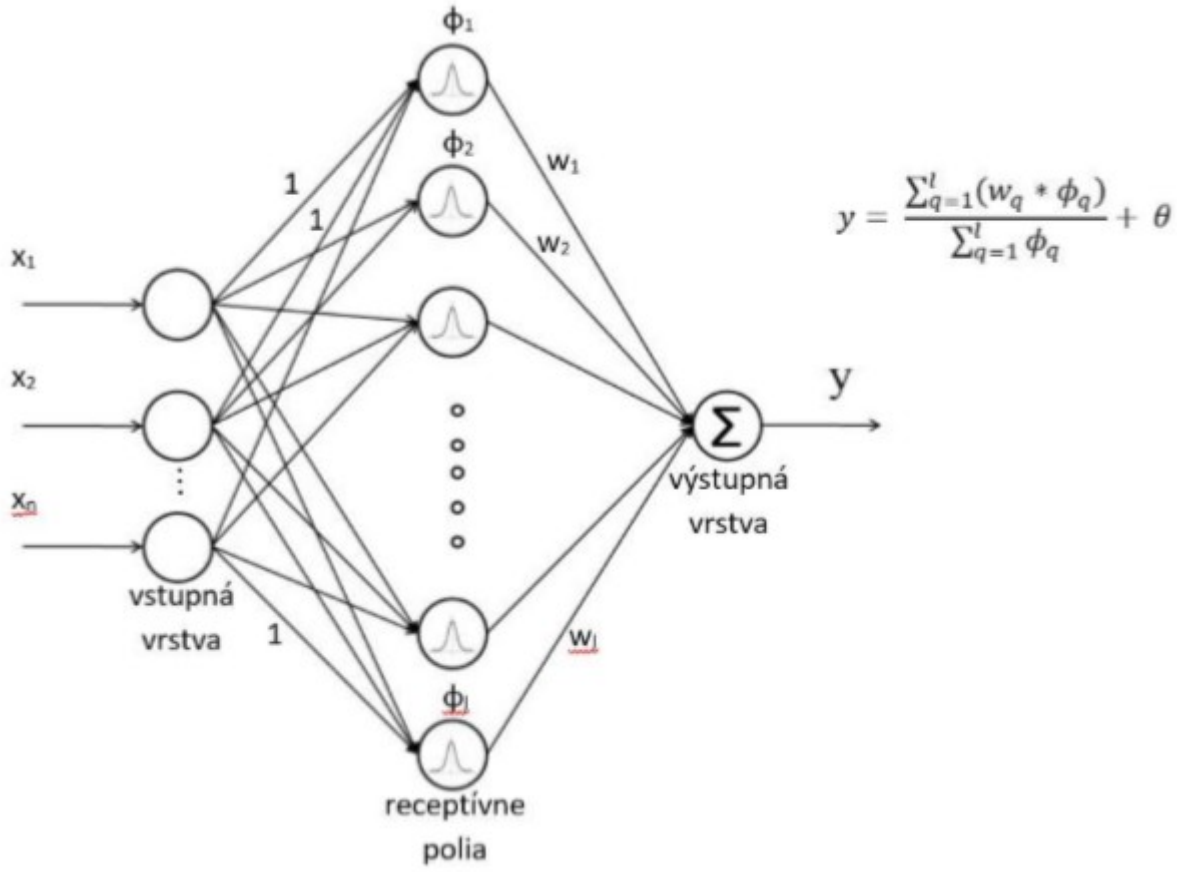
Date yyyy-mm-dd,
Location string,
MinTemp float,
MaxTemp float,
Rainfall float,
Evaporation float,
Sunshine float,
WindGustDir float,
WindGustSpeed int,
WindDir9am string,
WindDir3pm string,
WindSpeed9am int,
WindSpeed3pm int,
Humidity9am int,
Humidity3pm int,
Pressure9am float,
Pressure3pm float,
Cloud9am int,
Cloud3pm int,
Temp9am float,
Temp3pm float,
RainToday bool (written as yes/no),
RISK_MM,
RainTommorow bool (yes/no).

Na červeno vyznačené dáta neboli využité kvôli relevantnosti, alebo v prípade RISK_MM samotnej predpovede toho čo chceme vyrátať RBF sieťou.

Nakoniec prebehla diskusia o samotnej technickej realizácií zadania. Vzhľadom na to, že väčšina tímu uprednostnila programovací jazyk Python, bolo výsledné rozhodnutie vytvoriť webovú aplikáciu pomocou Python a Flask-u. To bude slúžiť na zadávanie príznakov od používateľa, a na pozadí bude volať nami pripravené funkcie pre klasifikáciu vstupu. Výsledok klasifikácie potom zobrazí používateľovi.

Riešenie projektu

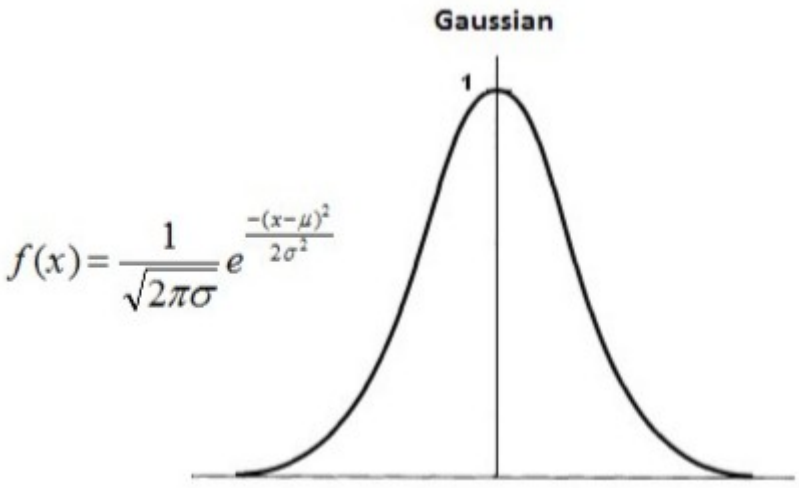
Koncept RBF neurónovej siete prvý krát uviedli Moody a Darken v 1989. Štruktúru siete môžete vydieť na nasledujúcom obrázku:



Tento prostriedok je reprezentovaný ako výlučne vždy trojvrstvová neurónová sieť, ktorá sa vyznačuje nasledujúcimi vlastnosťami:
- vstupná vrstva neovplyvňuje vstup žiadnym spôsobom
- synaptické váhy idúce od vstupných neurónov k neurónom vrstvy receptívnych polí sú rovné 1 a počas učenia sa ich hodnoty nemenia
- neuróny receptívnych polí používajú ako aktivačnú funkciu niektorú z foriem mediánových funkcií (RBF), napríklad Gaussovskú funkciu
- synaptické váhy idúce od neurónov receptívnych polí k výstupnému neurónu sa nastavujú počas parametrickej fázy učenia
- neurón výstupnej vrstvy agreguje vážené výstupy z neurónov receptívnych polí

V rámci učenia RBF neurónovej siete je potrebné nastaviť hodnoty synaptických váh medzi neurónmi receptívnych polí a parametre aktivačných funkcií receptívnych polí, ktoré sú závislé od použitej RBF funkcie(napríklad pri Gaussovej funkcii je potrebné nastaviť parameter strednej hodnoty a variancie). Nakoľko sme pri všetkých množinách trénovacích dát mali vždy riešiť úlohu binárnej klasifikácie, vo výstupnej vrstve sme mali vždy práve jeden výstupný neurón.
Existujú principiálne dva spôsoby učenia RBF siete. Prvý spôsob pozostáva len z jednej fázy a realizuje sa ako kontrolované učenie, pri ktorom sa synaptické váhy wi nastavujú na základe delta pravidla pomocou metódy spätného šírenia chyby. Podobný princíp sa aplikuje aj pri výpočte zmien hodnôt parametrov RBF funkcie.

My sme sa rozhodli použiť druhý spôsob učenia, ktorý kombinuje kontrolované a nekontrolované učenie a pozostáva z dvoch fáz. Ako typ RBF funkcií pre neuróny receptívnych polí sme si vybrali Gaussovu funkciu, ktorá sa riadi predpisom:



Z hľadiska teórie fuzzy množín ide o funkciu príslušnosti s nekonečne veľkým nosičom. Táto funkcia má len dva parametre: strednú hodnotu μ a varianciu σ. V prípade viacrozmernej Gaussovej funkcie sa hodnota μ stáva vektorom a výraz (x - μ) nadobúda podobu vzdialenosti medzi vektorom daným hodnotami vzorky a vektorom μ. My sme použili Euklidovskú vzdialenostnú normu, ktorá sa vypočíta podľa vzťahu:

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

V prvej fáze učenia sa realizuje nekontrolované učenie, ktoré pozostáva z dvoch krokov:

1) Určenie centier zhlukov cq

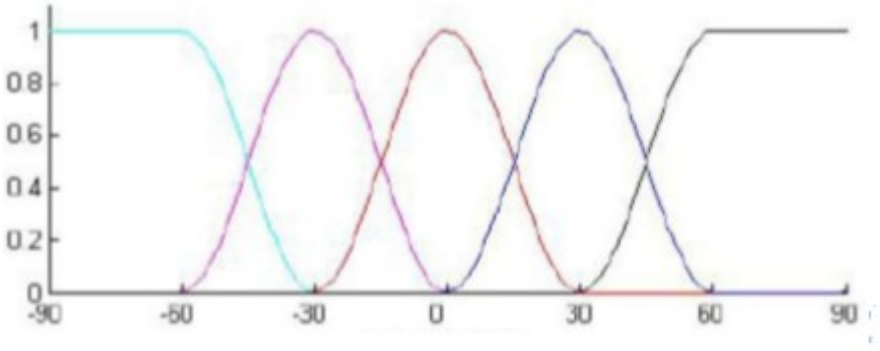
Pre určovanie centier zhlukov sme použili zhlukovaciu metódu K-Means v nasledujúcej implementácii:

a) určenie počtu zhlukov k
b) náhodná inicializácia centier k zhlukov
c) výpočet vzdialenosti každej trénovacej vzorky od centra každého zhluku a priradenie vzorky ku najbližšiemu zhluku
d) prepočítanie nových centier zhlukov ako priemer súradníc trénovacích vzoriek, ktoré boli k zhluku priradené v bode c)
e) určíme najväčšie zmenu polohy δmax v rámci prepočítania centier zhlukov v d)
f) ak je δmax > ε, koniec, inak návrat na c)

Ako presnosť zhlukovania ε sme použili pevne nastavenú hodnotu 0.001, počet zhlukov k bol volený metódou pokus-omyl, pričom pre naše použité dáta sa používali hodnoty z intervalu <5; 25>.

2) Určenie variancií σq tak, aby vzniklo hladké prekrývanie uzlov

Cieľom tohto kroku je stanoviť priebeh RBF funkcií na základe určenia priebehu poklesu funkčnej hodnoty z vrcholu smerom nadol tak, aby úroveň prieniku s ďalšími RBF funkciami bola v rozumnej výške tak, aby bol rovnomerne pokrytý priestor, v ktorom sa realizuje klasifikácia.



Nastavovanie variancií σq bolo realizované heuristikou najbližšieho suseda, ktorá je daná vzťahom:

$$\sigma_q = \frac{\| c_q - c_{closest} \|}{\beta}$$

β reprezentuje parameter prekrytia funkcií. Jeho hodnoty sme volili empiricky, pričom sme používali najmä hodnoty z intervalu <1; 1,5>. Uvedené dva kroky nekontrolovaného učenia nám nastavia parametre RBF funkcií pre neuróny vo vrstve receptívnych polí. Po nich nasleduje fáza kontrolovaného učenia, v ktorej je cieľom nastaviť hodnoty synaptických váh medzi neurónmi vo vrstve receptívnych polí a výstupným neurónom. Pre určenie optimálnych hodnôt synaptických váh sa snažíme minimalizovať chybovú funkciu danú vzťahom:

$$E(w_q) = \frac{1}{2}\sum_{k=1}^{p}(d^k - y^k)^2$$

p - počet trénovacích vzoriek, dk - očakávaný výstup, yk - vypočítaný výstup

Aplikovaním delta pravidla dostaneme vzťah pre výpočet zmeny váh:

$$\Delta w_q = y^*\left(d^k - y^k\right)$$

γ je parametrom učenia, ktorý sme volili experimentálne z intervalu hodnôt <0,05; 1>.

Uvedená štruktúra RBF neurónovej siete má nevýhodu v tom, že nevieme presne určiť, koľko neurónov má byť vo vrstve receptívnych polí. Zvolená stratégia učenia však zabezpečuje relatívne dobrú rýchlosť učenia, avšak presnosť siete je nižšia než u klasického viacvrstvového perceptrónu.

Výsledok a zhodnotenie experimentu

Pri parametroch siete ktorá mala 15 neurónov na skrytej vrstve, s 20 iteráciami, pri 0.02 learning rate a vzorkách o 500 údajoch sme po 10minútovom učení dosiahli presnosti 0,79759519.

Experiment preukázal že je možné predikovať počasie na základe historických dát s využitím RBF sieťe na natrénovanie. Jedná sa o klasifikačný algoritmus. Systém bol naprogramovaný ako webová aplikácia s python + flask-om a deploynuté na Azure cloude.