ZLOŽENÉ KLASIFIKÁTORY

(COMPOSITE CLASSIFIERS)

[*]Ján KRIŠTOF, [**]Eva OCELÍKOVÁ
[*]Univerzitná Knižnica Technickej univerzity v Košiciach, Letná 9,
042 00 Košice, tel. 095/602 2888, E-mail: kristof@lib.tuke.sk
[**]Katedra kybernetiky a UI, Fakulta elektrotechniky a informatiky Technickej univerzity v Košiciach, Letná 9,
042 00 Košice, tel. 095/602 2217, E-mail: ocelike@tuke.sk

SUMMARY

This paper deals with problems of composite classifier architectures. There are three primary architectures for combining classification algorithms: Stacked Generalization, Boosting and Recursive Partitioning. Our attention is focused on stacked generalization algorithm. In brief, stacked generalization is a recursive layered framework for classifier combination in which the layer of classifiers at each level is used to combine the predictions of the classifiers at the level immediately below..

Keywords: *composite classifier, k-NN, multilevel classifier*

1. ÚVOD

S rozvojom technológií sa ľudstvo snaží riešiť čoraz viac problémov prostredníctvom výpočtových prostriedkov, čím sa zvyšuje rýchlosť a presnosť riešenia, čo znamená aj ekonomický prínos.

Dôležitou súčasťou riešenia úloh je klasifikácia. Nie je podstatné o aký typ klasifikácie sa jedná alebo aké údaje sa spracúvajú. Podstatné však je, že k dispozícii je množina vstupných údajov, ktorých príslušnosti k triedam sú známe a na základe týchto je potrebné zaradiť do tried nové údaje, ktorých príslušnosť k triedam nepoznáme. Boli vytvorené rôzne klasifikačné technológie a algoritmy, ktorých snahou je zvýšiť presnosť, skrátiť čas klasifikácie a pod.

Jedným z možných riešení ako zvýšiť úspešnosť klasifikácie je využitie algoritmov tzv. *zložených klasifikátorov*. V článku je popísaný postup vytvorenia zloženého klasifikátora pomocou architektúry viacúrovňového klasifikátora, ktorého komponenty, ako aj zlučovací klasifikátor budú tvoriť klasifikátory pracujúce na princípe algoritmu *k*-NN (*k*–tého najbližšieho suseda).

2. ZLOŽENÉ KLASIFIKÁTORY

"Kombinovaním predikcií sady klasifikátorov môžeme dosiahnuť vyššiu presnosť ako dosiahne najpresnejší zo sady klasifikátorov" [1].

Kombinovanie výsledkov určitej množiny klasifikátorov sa skutočne javí ako účinná cesta k vytvoreniu zloženého klasifikátora, ktorý je presnejší ako jednotlivé klasifikátory, z ktorých sa tento skladá.

V prvom kroku zložený klasifikátor vytvoríme a natrénujeme. Trénovanie začína natrénovaním jednotlivých komponentov zloženého klasifikátora. Výstupy jednotlivých komponentov slúžia na vytvorenie trénovacej množiny pre zlučovací klasifikátor. Ak je zložený klasifikátor natrénovaný, privedieme na jeho vstup vzorku s neznámou triedou. Každý komponent klasifikuje túto vzorku osobitne a určí jej príslušnosť k určitej triede. Zlučovací klasifikátor zlúči získané predikcie o príslušnosti do triedy (použije ich ako svoj vstupný vektor) a určí triedu neznámej vzorky. Zlučovací klasifikátor môže pre klasifikáciu používať rôzne algoritmy, napr. jednoduché priame hlasovanie, klasifikačné pravidlo *k*-NN, ID3 a iné.

Prečo sa vlastne zaoberáme zloženými klasifikátormi, keď už jednoduché klasifikátory sú pomerne presné? Dôvodom je skutočnosť, že presnosť štandardného klasifikátora (napr. *k*-NN klasifikátora) môže byť ešte vylepšená použitím architektúry zloženého klasifikátora, v ktorom je zahrnutých niekoľko komponentov (napr. niekoľko *k*-NN klasifikátorov).

Existujú minimálne tri skutočnosti, ktoré podporujú toto tvrdenie:

1. Jeden komponent môže mať vysokú znalosť v určitej oblasti vzorkového priestoru a prejavuje sa vysokou lokálnou presnosťou. Zložený klasifikátor sa môže naučiť, v ktorej oblasti je ktorý komponent najlepší a to sa môže prejaviť vysokou presnosťou cez celý priestor vzoriek použitím tzv. lokálnych expertov.
2. Niektoré rozhodnutia môžu byť lepšie, ak sa zoberie do úvahy nielen individuálne rozhodnutie, ale skupinové rozhodnutie. Zatiaľ čo individuálne rozhodnutie môže viac podliehať chybe, skupinové rozhodnutia sú spoľahlivejšie.
3. Rozmanitosť výpočtových prostriedkov tiež podporuje kombináciu klasifikátorov. Môžeme sa pokúsiť vybrať sadu klasifikátorov v určitom zmysle komplementárnu a kombinovať ich výsledky. Vplyv chýb spôsobených niektorými klasifikátormi môže byť zmiernený zlúčením ich predikcií.

2.1. Konštrukcia zloženého klasifikátora

Kritériá pre návrh zloženého klasifikátora sú:

1. presnosť komponentov
2. rozmanitosť komponentov
3. efektívnosť zloženého klasifikátora.

Najdôležitejšie kritérium je *presnosť*. Základnou myšlienkou, z ktorej sa vychádza je skutočnosť, že výsledkom kombinácie nepresných predikcií nemôže byť predikcia s vyššou presnosťou. Preto sa vyvíja veľké úsilie na natrénovanie jednotlivých komponentov do vysokej presnosti, ako nezávislých klasifikátorov. Presnosť klasifikácie komponentov však nie je jediným dôležitým faktorom.

Druhým dôležitým kritériom je *rozmanitosť*. Toto kritérium vychádza z jednoduchého pozorovania, že kombinácia klasifikácií množiny klasifikátorov, ktoré vytvárajú tú istú chybu, nemôže viesť k zlepšeniu presnosti zloženého klasifikátora. Teda klasifikátory, ktoré kombinujeme, musia mať rozdielne chyby, ak chceme ich kombináciou dosiahnuť vyššiu presnosť.

Kritérium *efektívnosti* je postavené na základe dvoch hlavných požiadaviek a to, že klasifikátor by mal využívať minimálne množstvo času a pamäti pre trénovanie a aplikáciu. Máme dve možnosti ako sa vyhnúť neúmerne vysokým nárokom, či už časovým alebo pamäťovým:

1. použijeme menej komponentov
2. výpočtová náročnosť jedného komponentu pôjde na úkor ostatných výpočtovo menej náročných komponentov.

2.2. Počet komponentov zloženého klasifikátora

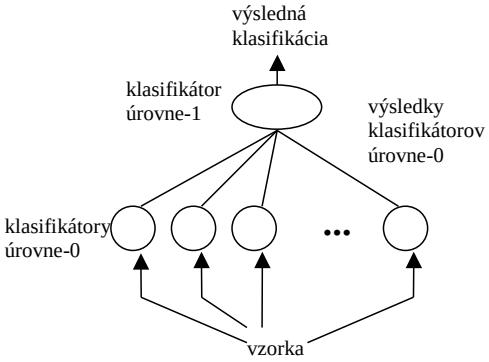
Určiť vhodný počet komponentov zloženého klasifikátora je pomerne náročný problém. Zdalo by sa, že čím viac bude komponentov, tým presnejší bude zložený klasifikátor. Avšak experimenty (napr. Battiti a Colla pri rozpoznávaní písmen zistili, že stačia dve alebo tri neurónové siete pre vyššiu presnosť ako je najlepšia presnosť získaná z jednej siete; Kwok a Carter pri projekte predpovede počasia využívali tri rozhodovacie stromy a menej, pri väčšom počte dokonca zaznamenali zhoršenie výsledkov) ukázali, že maximálna presnosť alebo presnosť blízko maximálnej, pre veľkú škálu úloh sa dosahuje s malým počtom klasifikátorov. Presný počet komponentov sa však určuje podľa konkrétneho špecifického problému.

Architektúry zložených klasifikátorov sa nazývajú sú rôzne, čo väčšinou závisí od komunity a od aplikácie, v ktorej sa tieto používajú napr. hybridné alebo zložené modely, zoskupenia, atď.

3. ARCHITEKTÚRA VIACÚROVŇOVÉHO KLASIFIKÁTORA

Táto architektúra sa sústreďuje väčšinou na dvojvrstvovú architektúru, v ktorej základné klasifikátory, ktoré kombinujeme, sú na úrovni - 0 a zlučovací klasifikátor t.j. klasifikátor, ktorý zlučuje ich predikcie, je klasifikátor na úrovni-1 Takéto vrstvenie môže pokračovať vytvorením klasifikátora na úrovni-2, atď.

Viacúrovňový klasifikátor je štruktúra pre kombináciu klasifikátorov, v ktorej je každá vrstva klasifikátorov použitá pre zlúčenie výsledkov klasifikátorov z najbližšej nižšej vrstvy. Klasifikátor, ktorý je na najvyššej úrovni dáva výslednú klasifikáciu. Vstupným vektorom klasifikátora na každej úrovni je vektor predikcií klasifikátorov vo vrstve bezprostredne pod ním. Kým informácia prejde z jednej vrstvy do nasledujúcej vrstvy, môže vektor predikcií dostať formu dôverných alebo iných dát. V článku sa zameriame len na systémy, v ktorých je z podriadenej vrstvy do nadriadenej odovzdávaná *iba predikcia triedy*. Architektúra viacúrovňového klasifikátora však nie je limitovaná touto voľbou. Architektúra dvojúrovňového klasifikátora je na obr. 1.



Obr. 1 Architektúra viacúrovňového klasifikátora
Fig. 1 Multilevel classifier architecture

3.1. Algoritmus viacúrovňového klasifikátora

Predpoklady pre viacúrovňový klasifikátor sú :

- množina učiacich sa algoritmov na úrovni–0
- učiaci sa algoritmus na úrovni –1
- množina trénovacích vzoriek.

Učiace sa algoritmy na úrovni–0 by mali byť rozdielne. Tým sa získa požadovaná rozmanitosť klasifikátorov na základnej úrovni–0. Algoritmus má dve fázy: *trénovaciu a testovaciu*.

Trénovacia fáza:

1. Trénovanie komponentov klasifikátora prebieha nasledovne: pre každú vzorku v trénovacej množine, trénujeme každý z *n* klasifikátorov na úrovni–0 pomocou zvyšných vzoriek. Po natrénovaní, klasifikujeme vzorku,

ktorá sa nezúčastnila trénovania a to každým z natrénovaných klasifikátorov na úrovni–0. Tvar kódovaného vektora je nasledovný: *n* predikcií klasifikátorov na úrovni–0 a trieda aktuálnej vzorky.

2. Klasifikátor na úrovni-1 trénujeme pomocou trénovacej množiny, ktorá pozostáva zo získaných kódovaných vektorov. Táto množina má rovnaký počet vzoriek ako trénovacia množina pre klasifikátory na úrovni–0, pretože každá trénovacia vzorka úrovne-1 korešponduje s jednou trénovacou vzorkou úrovne-0.

Aplikačná fáza:

Testovanie novej vzorky, ktorej trieda je neznáma, je nasledovné: klasifikujeme vzorku pomocou každého klasifikátora na úrovni–0, čím dostaneme vstupný kódovaný vektor pre klasifikátor úrovne-1. Tento vektor je potom klasifikovaný klasifikátorom na úrovni–1, ktorého výstupom je príslušnosť klasifikovanej vzorky do určitej triedy.

4. EXPERIMENTÁLNE VÝSLEDKY

Pre účely testov boli použité ako komponenty zloženého klasifikátora rôzne algoritmy klasifikátora *k*-NN. V prípade zlučovacieho klasifikátora sme použili klasifikátor *k*–NN alebo algoritmus priameho hlasovania. Testy boli realizované na dátových množinách z tab. 1.

Parametre \ Množina	kružnice	synth	Košice
Počet tried	2	2	7
Počet príznakov	2	2	7
Počet trénovacích vzoriek	2250	250	3166
Počet testovacích vzoriek	250	1000	3165

Tab. 1 Údaje o dátových množinách
Tab. 1 Information about date set

Typ \ Množina	Kosice	kruznice	synth
Zložený kl. *k*–NN	96,30%	100%	90,20%
Zložený kl. hlasovanie	96,40%	100%	87,50%
k–NN *k*=1	96,10%	99,20%	87,60%
k–NN, *k*=3	95,90%	99,60%	89,40%
ART	93,14%	-	-
Fuzzy ART	95,04%	-	-
Gauss ART	95,92%	-	-
FuzzyBP	95,01%	-	-

Tab. 2 Porovnanie dosiahnutých výsledkov
Tab. 2 Results comparison

Získané výsledky na dátovej množine Košice sú porovnané s výsledkami, ktoré boli dosiahnuté pomocou rôznych druhov neurónových sietí ART MAP [2], doprednou neurónovou sieťou s učením fuzzy backpropagation [3], 1-NN a 3–NN klasifikátormi s použitím redukčných algoritmov.

5. ZÁVER

Zložený klasifikátor založený na viacúrovňovej architektúre a vybudovaný z komponentov *k*–NN klasifikátorov, dosiahol výsledky, ktoré sme od neho očakávali. Presnosť zloženého klasifikátora bola vyššia ako presnosť samostatného klasifikátora pracujúceho podľa algoritmu *k*–NN.

V prípade dát Košice výsledky ukazujú, že zložený klasifikátor dosahoval lepšie výsledky klasifikácie ako to bolo v prípade rôznych druhov neurónových sietí ART MAP alebo v prípade doprednej neurónovej siete s učením fuzzy backpropagation. Ďalšie dva testy, rovnako ako v predchádzajúcom prípade, ukazujú, že zložený klasifikátor vytvorený viacúrovňovou architektúrou vykazuje vyššiu presnosť klasifikácie ako jednoduché *k*-NN klasifikátory.

LITERATÚRA

[1] Skalak B. D.: Prototype Selection for Nearest Neighbor Classifiers. CMPSCI Technical Report 1996

[2] Kopčo, N. - Sinčák, P.: ART neural networks for image classification : Influence of internal cluster representation on accuracy Method for evalution of confidence of clasification. http://neuron-ai.tuke.sk/~kopco/kosice.

[3] Fedor, M. - Zeťáková, Z.: Dopredná sieť fuzzy BP a realizácia klasifikácie obrazových dát. Projekt z predmetu Neurofuzzy systémy

BIOGRAPHY

Ján Krištof was born on 8.4.1972. In 1996 he graduated (MSc.) at the department of Cybernetics and Artificial Inteligence Faculty of Electrical Engineering and Informatics at Technical University in Kosice. His scientific research is focusing on statistical classification. In addition to this, he also investigates the questions related to the composite classifiers.

Eva Ocelíková defended her Ph.D. thesis, which dealt with multicriterial classification of situations in the complex system, in 1985 at the Slovak Technical University of Bratislava. She is working at the Department of Cybernetics and Artificial Intelligence of the Faculty of Electrical Engineering and Informatics at Technical University in Košice as associate professor. Her research work includes problems of decision processes, especially the problems of multicriterial classifications. designing and high dimensionality reduction of feature space of multidimensional data in decision.