VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
KOMPIUTERIJOS KATEDRA

Magistro baigiamasis darbas

Klaidų kainoms jautrūs klasifikavimo algoritmai

Cost-sensitive classification algorithms

Atliko: Vardaitis Pavardaitis .....
(parašas)

Darbo vadovas:
prof. hab. dr. Vardenis Pavardenis .....
(parašas)

Recenzentas:
dr. Vardonis Pavardonis .....
(parašas)

Vilnius
2014

Santrauka

Praktiniuose taikymuose, kaip kad medicinos diagnostikoje ir veidų atpažinime, pripažintas santykinis klasifikavimo klaidų kainų, priklausančių nuo tikrosios ir priskirtosios klasių, skirtumas. Šis darbas apima klaidų kainoms jautraus hibridinio klasifikatoriaus, sudaryto iš sprendimo medžio ir daugiasluoksnio perceptrono, kūrimą ir analizę. Hibridiniam klasifikatoriui konstruoti buvo realizuoti kainoms jautrių *C4.5* sprendimo medžių variantai bei klaidų kainoms jautrus daugiasluoksnis perceptronas, jiems kombinuoti panaudota *Banerjee* metodika. Atlikti eksperimentai su realaus pasaulio ir sintetiniais duomenimis leidžia teigti, kad hibridinis klasifikatorius gali būti naudingas, t. y. sumažinti jį inicializavusio sprendimo medžio klasifikavimo kainą ir pasiekti šį rezultatą greičiau nei atsitiktiniais svoriais inicializuotas daugiasluoksnis perceptronas.

Raktiniai žodžiai: nesubalansuoti duomenys, jautrus kainoms, klasifikavimas, dirbtiniai neuroniniai tinklai, sprendimų medis, hibridinis, daugiasluoknis perceptronas.

Summary

In practical applications including medical diagnosis and face detection it has been admitted that classification errors might differ in relative cost depending on the real and predicted classes. The work comprises implementation and analysis of a cost-sensitive hybrid classifier, consisting of a decision tree and a multi-layer perceptron. The hybrid classifier was constructed using several varieties of a cost-sensitive *C4.5* decision tree and a cost-sensitive multi-layer perceptron, which were combined using the *Banerjee* method. Conducted experiments with real world and synthetic data allow to conclude that the hybrid method might be useful, namely, decrease the misclassification error cost of the initializing tree and achieve this result faster than a randomly initialized multi layer perceptron.

Keywords: imbalanced dataset, cost-sensitive, classification, artificial neural network, decision tree, hybrid, multilayer perceptron.

Turinys

Terminai

1. ANT - klasikinis atgalinio perdavimo neuroninis tinklas (angl. back-propagation neural network)
2. DNT - dirbtinis neuroninis tinklas (angl. artificial neural network)
3. DSP - daugiasluoksnis perceptronas (angl. multilayer perceptron)
4. ...

Įvadas

Klasifikavimo uždavinį naudojant induktyvaus pobūdžio mokymąsi plačiai mėginta spręsti orientuojantis į klasifikavimo klaidų skaičiaus minimizavimą. Panaudojant aibę mokymosi duomenų - vektorių, kuriems įvardyta priklausomybė tam tikrai klasei, - konstruojamas algoritmas, besistengiantis kuo didesnį skaičių elementų priskirti teisingai klasei.

...

Keliama tokia **hipotezė**:

Įmanoma panaudoti pavienį sprendimo medį, kurio apmokymas greitas, tačiau jautrumas kainoms silpnas, inicializuoti tinklui, kurio architektūra ir parametrai nežinomi, be to, apmokymas lėtas, tačiau jautrumas kainoms geras, kad būtų gautas greitai apmokomas gero jautrumo kainoms klasifikatorius.

...

1. Klasifikavimo uždavinys

Klasifikavimo problema - kaip pagal turimus duomenis[1] $\boldsymbol{D}$, kurie susideda iš n duomenų taškų (vektorių[2]) $\boldsymbol{x}_i \in \mathbb{R}^p$, $i = 1, ..., n$, bei žinomas jų klases $class(\boldsymbol{x_i}) \in \{1, 2, ..., m\}$, $i = 1, ..., n$, sudaryti metodą, kuris galėtų nustatyti vektoriaus $\boldsymbol{x'} \in \mathbb{R}^p$ klasę.

Nagrinėjant realaus pasaulio duomenis iškyla dvi problemos [Rou06]:

...

[1] Pusjuodės didžiosios raidės žymi matricas.

[2] Pusjuodės mažosios raidės žymi vektorius.

2. Dirbtinių neuroninių tinklų apžvalga

2.1. Bendrieji dirbtinių neuroninių tinklų principai

2.1.1. Perceptronas

Perceptronas[3] yra iteraciškai apmokomas tiesinis klasifikatorius. Įvedami žymėjimai: duomenų vektorių $\boldsymbol{x} = (x_1, x_2, ..., x_p)$ praplečiame vienetu, $\boldsymbol{z} = (1, x_1, x_2, ..., x_p)$, perceptrono įėjimų svorių vektorių $\boldsymbol{w} = (w_1, w_2, ..., w_p)$ praplečiame w_0, $\boldsymbol{v} = (w_0, w_1, w_2, ..., w_p)$. Naudosime sigmoidinę glodninimo funkciją:

$$f(x) = \frac{1}{1 + e^{-x}} \, . \tag{1}$$

Tada i-tajam duomenų vektoriui perceptrono išėjimas apskaičiuojamas taip:

$$o_i = f(\boldsymbol{x}_i \boldsymbol{w}^T + w_0) = f(\boldsymbol{z}_i \boldsymbol{v}^T) \, . \tag{2}$$

...

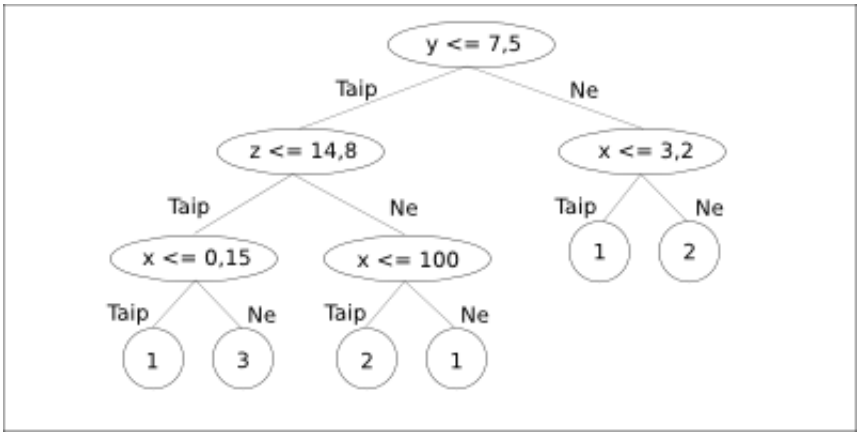
[3]Perceptronu vadinsime McCulloch-Pitts neuroną su sigmoidine aktyvavimo funkcija.

3. Sprendimo medžių apžvalga

3.1. Bendrieji sprendimo medžių principai

Sprendimo medžiu vadinamas medžio pavidalo klasifikatorius, priskiriantis klasėms daugiamačius vektorius, kurių požymiai gali būti tiek kategoriniai, tiek tolydieji kintamieji.

Medį sudaro arba lapas, pažymėtas klasės etikete, arba struktūra, apimanti su dviem ar daugiau pomedžių sujungtą sprendimo priėmimo mazgą [**?**]. Pastarosios rūšies mazgai apibrėžiami testo pavidalu, o jų pomedžiai atitinka visas įmanomas šio testo baigtis (pvz., žr. 1).



1 pav.: Sprendimo medžio pavyzdys

...

4. Algoritmų realizacija

4.1. Dirbtinių neuroninių tinklų realizacija

Norint geriau susipažinti su klasifikavimo klaidų kainų įvedimo metodais bei jų savybėmis, prieš kuriant SM ir DNT kombinuojantį klasifikatorių, buvo nuspręsta atskirai išsinagrinėti kainų įvedimo metodus į SM ir DNT klasifikatorius. Atlikus kainų įvedimo į DNT metodų analizę (žr. 2.1.1 skyrių) paaiškėjo, kad geriausias duomenų subalansavimo metodas yra (1), o praktikoje nusistovėjęs kainų įvedimo į DNT metodas yra (2), kuris ir realiuotas šiame darbe.

...

5. Hibridinių klasifikatorių veikimo eksperimentinis tyrimas

5.1. Bendrieji eksperimentų nustatymai

Eksperimentais šiame skyrelyje siekiama nustatyti, ar apskritai yra prasminga kurti kainoms jautrų hibridinį klasifikatorių, panaudojant SM ir DNT hibridizacijos metodiką [Ban97] ir įvedant jautrumą kainoms kiekvienoje klasifikatoriaus dalyje nepriklausomai. Kad būtų prasminga, reikėtų parodyti, kad hibridinis klasifikatorius sugeba pasiekti mažesnę klasifikavimo kainą su testiniais duomenimis nei jį inicializavęs sprendimo medis per mažiau iteracijų nei tokios pat architektūros, tačiau atsitiktiniais svoriais inicializuotas DSP.

...

	Pavadinimas	Inicializuojantis medis	Kainoms jautrus algoritmo aspektas
1.	*Hybrid_C4.5*	C4.5 medis	DSP
2.	*Hybrid_Laplace*	C4.5 medis, Laplace genėjimas	Genėjimas, DSP
3.	*Hybrid_MetaCost*	MetaCost medis, naudojantis C4.5 medį kaip bazinį klasifikatorių	MetaCost, DSP
4.	*Hybrid_Mod_prob*	C4.5 medis, pakeitus apriorines klasių tikimybes ir lapų klases pagal kainų matricą	Tikimybių modifikacija ir lapų pernumeravimas, DSP
5.	*Hybrid_Mod_prob_err*	C4.5 medis, pakeitus apriorines klasių tikimybes pagal kainų matricą, klaidomis grįstas genėjimas	Tikimybių modifikacija, DSP
6.	*Hybrid_Mod_prob_lap*	C4.5 medis, pakeitus apriorines klasių tikimybes pagal kainų matricą, Laplaso genėjimas	Tikimybių modifikacija, genėjimas, DSP
7.	*Hybrid_C5.0*	C5.0 medis	C5.0 medis, DSP
8.	*Banerjee*	C4.5 medis	Kainoms nejautrus

1 lentelė.: Lyginami hibridiniai klasifikatoriai.

...

Išvados

Šiame darbe realizuota:

1. Sukurta bazinė C4.5 algoritmo realizacija ir keli jautrumo kainoms joje užtikrinimo metodai: gaubiamasis MetaCost algoritmas, pakeistosios klasių tikimybės, Laplace genėjimas.

...

Atlikus eksperimentus su sintetiniais ir realaus pasaulio duomenimis, gautos tokios išvados:

1. Parodyta, kad hibridinis kainoms jautrus klasifikatorius, paremtas [**?**] hibridizacijos metodika ir jautrumo kainoms įvedimu į sprendimo medį bei daugiasluoksnį perceptroną atskirai, gali sumažinti inicializavusio sprendimo medžio klasifikavimo klaidų kainą su testiniais duomenimis. Taip pat parodyta, kad hibridas pasiekia geriausios kainos iteraciją greičiau nei analogiškos architektūros, tačiau atsitiktinių pradinių svorių daugiasluoksnis perceptronas.

...

Literatūros sąrašas

[Ban97] Arunava Banerjee. Initializing neural networks using decision trees. *Computational learning theory and natural learning systems*, IV:3–15, 1997.

[Rou06] Nathan Rountree. *Initialising Neural Networks with Prior Knowledge*. PhD thesis, University of Otago, Dunedin, New Zealand, September 2006.

Priedas Nr. 1.
Papildomų eksperimentų rezultatų lentelės

2 lentelė.: Rezultatai su *Tae* duomenimis. *Banerjee* ir hibridiniai klasifikatoriai, taip pat medžiai ir jų hibridai lyginami pagal klaidų kainą, o DSP ir hibridai - pagal mokymosi iteracijų skaičių.

	$\bar{x}$	σ^2	$a > b$			$a < b$		
Algoritmas				p/2	T		p/2	T
a) Banerjee	1.6335	0.5584	N	0.0000	-5.1173	T	0.0000	5.1173
b) Hybrid_C50	1.7395	0.5647						
a) C50	1.7490	0.5819	N	0.3047	0.5109	N	0.3047	-0.5109
b) Hybrid_C50	1.7395	0.5647						
a) DSP	82.1200	244.7042	T	0.0112	2.2845	N	0.0112	-2.2845
b) Hybrid_C50	63.2800	202.4601						
a) Banerjee	1.6335	0.5584	T	0.0000	35.0309	N	0.0000	-35.0309
b) Hybrid_Laplace	0.7147	0.1800						
a) Laplace	0.6894	0.1159	N	0.0045	-2.6250	T	0.0045	2.6250
b) Hybrid_Laplace	0.7147	0.1800						
a) DSP	82.1200	244.7042	T	0.0000	12.4643	N	0.0000	-12.4643
b) Hybrid_Laplace	0.6100	1.3107						