VĚDECKÉ SPISY VYSOKÉHO UČENÍ TECHNICKÉHO V BRNĚ

Edice PhD Thesis, sv. 372

ISSN 1213-4198

Ing. Petr Honzík

Robustní chybová funkce pro regresní klasifikátory

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
Fakulta elektrotechniky a komunikačních technologií
Ústav automatizace a měřicí techniky

Ing. Petr Honzík

ROBUSTNÍ CHYBOVÁ FUNKCE PRO REGRESNÍ KLASIFIKÁTORY

Robust Loss Function for Regression Classifiers

ZKRÁCENÁ VERZE PH.D. THESIS

Obor:	Kybernetika, automatizace a měření
Školitel:	Prof. Ing. Petr Vavřín, DrSc.
Oponenti:	Prof. Dr. Ing. Jiří Maryška, CSc. Doc. Ing. Lenka Lhotská, CSc.
Datum obhajoby:	31. 3. 2006

Klíčová slova

Klasifikátor, klasifikace, regresní model, regresní klasifikátor, chybová funkce, robustní chybová funkce, AUC, ROC, neparametrická charakteristika, chybná klasifikace, váha, modifikovaná váha

Keywords

Classifier, classification, regression model, regression classifier, loss function, robust loss function, AUC, ROC, nonparametric characteristic, misclassification, weight, modified weight

Disertační práce je k dispozici na adrese VUT v Brně, Fakulta elektrotechniky a komunikačních technologií, Vědecké a zahraniční oddělení, Údolní 244/53, 60200, Brno.

ISBN 80-214-3186-5
ISSN 1213-4198

Obsah

1 Úvod

klasifikace, klasifikátor, regresní model

Každý den provádíme celou řadu rozhodnutí, mnohdy téměř podvědomě (který ručník v koupelně použijeme). Nad jinými rozhodnutími trávíme více času (výběr politické strany při volbách). Existují však i velice závažná profesní rozhodnutí, jako volba vhodného typu matriálu pro vnější plášť lodi nebo rozhodnutí o vině či nevině žalovaného. Hlavním společným rysem výše uvedených příkladů je skutečnost, že výsledek je vybrán z konečného počtu možných odpovědí, které často nelze jednoznačně uspořádat (např. politické strany). Proces výběru odpovědi z konečného počtu tříd je označován slovem klasifikace, matematický model provádějící takovou rozhodovací činnost pak slovem klasifikátor. Pokud je výstupem spojitá (kvantitativní) veličina, bude používáno označení regresní model nebo regresní funkce.

regresní klasifikátor

Jedním z atributů používaným k dělení klasifikátorů je jejich vnitřní struktura. Regresní klasifikátor je složen ze dvou částí. Tou první je regresní funkce, jejímž vstupem jsou nezávislé veličiny a výstupem je spojitá veličina. Právě spojitý výstup je důvodem pro použití termínu *regresní*. Regresní funkce bývá nastavována pomocí tradičních chybových funkcí (MNČ – metoda nejmenších čtverců, ML – maximální věrohodnost). Druhou část tvoří kritické (prahové) hodnoty, jejichž porovnání s výstupem regresní funkce určuje konečnou klasifikaci. Jednoduchým příkladem regresního klasifikátoru je např. vážený průměr známek studenta na vysoké škole použitý k rozhodnutí, zda student bude či nebude absolvovat souhrnnou zkoušku. Výpočet průměru na základě známek a kreditů tvoří regresní funkci. Kritickou hodnotou je pak průměrná známka, při které je studentovi souhrnná zkouška ještě prominuta.

robustní chybová funkce

V praxi bývá řešen problém, jak získat z datových souborů znalost v podobě parametrů modelu. Informaci o tom, jak přesný v daný okamžik model je, udává svou hodnotou chybová funkce. Vyjadřuje, jak významně se liší výstup modelu od požadované hodnoty. Chybu je zpravidla možné vyjádřit pro každý dílčí prvek.

Robustností je obecně rozuměna necitlivost vůči malé odlišnosti od modelovaného předpokladu [11,14] (přičemž odlišností může být rozuměna malá odchylka výstupu modelu od očekávané hodnoty nebo velká odchylka malého počtu dat – tedy robustnost vůči tzv. outliers). Robustní chybovou funkcí je v disertační práci rozuměna chybová funkce typu R (R-Estimate), která principiálně vychází z pořadových neparametrických testů. Její hlavní rozdíl v porovnání s tradiční chybovou funkcí spočívá v tom, že ji nelze vyjádřit pro každý prvek zvlášť, ale pouze pro celý soubor analyzovaných dat.

téma disertační práce

Tématem práce je nalezení vhodné robustní chybové funkce, jejíž použití místo tradičních chybových funkcí povede k lepšímu nastavení regresní funkce v regresním klasifikátoru a tím pádem i k přesnější klasifikaci.

2 Současný stav

regresní klasifikátory

Tato rešerše uvádí základní přehled o typech regresních klasifikátorů a matematických nástrojích používaných při jejich nastavování a vyhodnocování. Smyslem je vyznačit základní rysy problematiky s odkazy na literaturu, která se jimi podrobněji zabývá.

2.1 Regresní modely a klasifikátory

regresní model, klasifikátor, regresní klasifikátor

Typickými příklady klasifikátorů jsou metody typu IBL (instance based learning), k-NN (k nearrest neighbourhood) nebo třeba rozhodovací stromy (použité pro účely klasifikace). Typickým příkladem regresního klasifikátoru je logitový model. Výstupní binární veličina (kvalitativní) je interpretována jako diskrétní, výstup samotného logitového modelu je pak spojitá veličina. Nastavený regresní model (logitová funkce) se ve finále používá ke klasifikaci na základě porovnání jeho výstupní hodnoty s tzv. prahovou hodnotou. Až v této fázi dochází ke klasifikaci. Mezi další lineární klasifikační metody patří např. lineární regrese nebo lineární či kvadratická diskriminační analýza [6].

porovnání výkonnosti metod

Srovnání výkonnosti různých typů lineárních regresních klasifikátorů ilustruje tabulka (2.1). Rozdíl mezi nejslabší metodou (lineární regresí) a nejpřesnější (logistická regrese) činí v ilustrativním příkladu více než 10% [6].

Tabulka 2.1: Porovnání výkonnosti 4 lineárních klasifikátorů

Typ modelu	**Chyba modelu**	
	Trénovací data	**Testovací data**
Lineární regrese	0,48	0,67
Lineární diskriminační analýza	0,32	0,56
Kvadratická diskriminační analýza	0,01	0,53
Logistická regrese	0,22	0,51

2.2 Chybové funkce v regresních modelech

definice chybové funkce

Zápisem $(x,y,f(x)) \in X \times Y \times Y$ rozumíme uspořádanou trojici, ve které veličina x představuje vstupní hodnotu, y představuje požadovanou výstupní hodnotu a $f(x)$ představuje predikovanou výstupní veličinu. Funkce $L: X \times Y \times Y \rightarrow (0;\infty)$, pro kterou platí, že pro $\forall x \in X$ a $\forall y \in Y$ je $L(x,y,y)=0$, je označována jako chybová funkce [16].

metoda nejmenších čtverců

Metoda nejmenších čtverců (MNČ) je aproximační metoda, která spočívá v tom, že hledáme takové parametry zvolené funkce, pro které je součet čtverců odchylek vypočtených hodnot od hodnot naměřených minimální [18].

Výpočet chyby *Err* podle uvedené definice vyjadřuje následující vztah:

$$Err = \sum_{i=1}^{N} \left[y_i - f\left(x_i, \boldsymbol{b}\right) \right]^2 \qquad (1)$$

kde y je požadovaná výstupní hodnota, x je hodnota vstupní veličiny a $\boldsymbol{b}$ je vektor parametrů modelu $f(x,\boldsymbol{b})$.

Dále jsou používány různé varianty této metody. Mezi nejpoužívanější patří tzv. lineární MNČ, nelineární MNČ, váhová MNČ, absolutní chyba, polynomické varianty a robustní verze s pásmem necitlivosti.

maximální věrohodnost

Mějme pravděpodobnostní funkcí f_p. Její hodnota v daném bodě $\boldsymbol{x_i}$ vyjadřuje pravděpodobnost nastolení události A. Typický je výpočet věrohodnosti hypotézy, že veličina $\boldsymbol{x}$ je daného rozložení (pro konkrétní parametry) nebo že

vztah veličin $\boldsymbol{x}$ a $\boldsymbol{y}$ popisuje určitá funkce za předpokladu známého rozložení chyby měření těchto veličin [14,4]. V takových případech odpovídá funkce f_p funkci hustoty použitého rozložení. Funkce f_p může také v binárních klasifikátorech vyjadřovat pravděpodobnost, se kterou veličina $\boldsymbol{x}$ v konkrétní hodnotě $\boldsymbol{x_0}$ náleží do třídy 0 nebo 1 (logitový model). Vztah pro výpočet věrohodnosti vypadá následovně:

$$L = \prod_{i=1}^{N} f_p(x_i)^{y_i} \left[1 - f_p(x_i)\right]^{1-y_i} \tag{2}$$

společné vlastnosti MNČ a MLE

Základní společnou vlastností obou uvedených metod je skutečnost, že vycházejí z transformace stejné vstupní informace – diference mezi predikovanou a skutečnou hodnotou. Lze tedy vypočíst ohodnocení modelu pro jednotlivý prvek a celková chyba je tvořena souhrnem chyb dílčích. Mějme ohodnocení modelu na základě dvou libovolných prvků tvořených uspořádanou dvojicí $(\boldsymbol{x};\boldsymbol{y})$. Ordinální relace mezi těmito ohodnoceními bude stejná pro obě uvedené metody. V mnoha případech vedou obě metody ke stejnému řešení [14].

rozdíly mezi MNČ a MLE

Základní rozdíl mezi uvedenými přístupy spočívá v typu transformační funkce, způsobu výpočtu souhrnné chyby modelu a informaci obsažené v celkové chybě modelu.

Transformace vstupní diference Δy je v případě MNČ funkce rostoucí, v případě MLE funkce klesající. Zatímco u MNČ je funkce označována za chybu a je minimalizována, v případě MLE se hovoří o věrohodnosti a její velikost je maximalizována. Souhrnná chyba je u MNČ získána součtem všech dílčích chyb, u MLE je použit součin jednotlivých věrohodností. Z toho také vyplývá, že u MLE má význam absolutní hodnota chyby, která vyjadřuje podmíněnou pravděpodobnost za předpokladu předem zvoleného typu rozložení chyby.

robustní chybové funkce

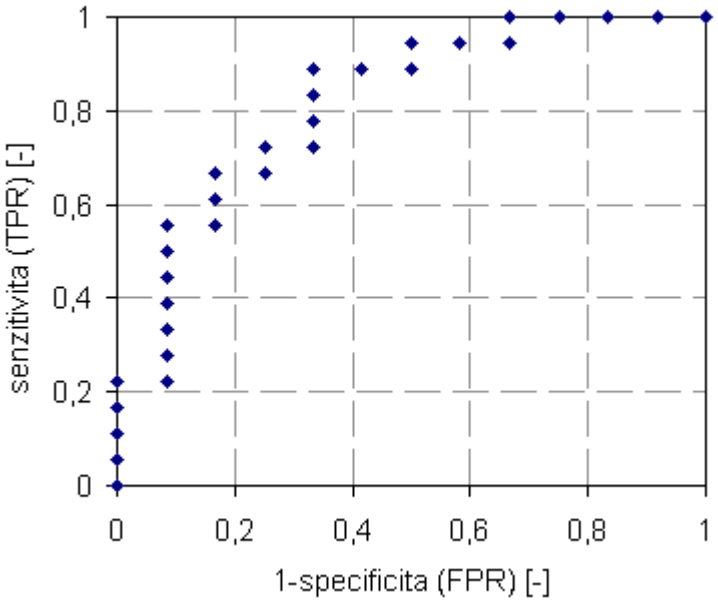
Robustností je obecně rozuměna necitlivost vůči malé odchylce od idealizovaných předpokladů [14,11]. Z toho vyplývá, že robustní chybovou funkcí je např. varianta MNČ s pásmem necitlivosti ε. Obecně se robustní metody dělí do tří skupin. První vychází z maximální věrohodnosti (M-estimate), druhá používá lineárních kombinací pořadových statistik jako je např. medián (L-estimate) a poslední je založena na použití pořadových testů jako jsou neparametrické korelace a regresní koeficienty (R-estimate).

2.3 Vyhodnocení klasifikátorů – ROC analýza

graf ROC

Kvalitu dichotomního klasifikátoru lze vyjádřit pomocí čtyřpolní tabulky a z ní vypočtených parametrů (senzitivita, specificita, pozitivní a negativní prediktivní hodnota atd.). Používaným grafickým znázorněním kvality dichotomního modelu je graf ROC. Na ose x je FPR (False Posivite Rate), tedy *1-specificita* a na ose y TPR (True Positive Rate), tedy *senzitivita*. Jeden klasifikátor odpovídá v grafu ROC jednomu bodu. Zjednodušeně pak v grafu ROC platí, že čím je bod blíže hornímu levému rohu, tím je klasifikátor lepší.

Mezi významné vlastnosti grafu ROC patří nezávislost tvaru na pravděpodobnostní funkci i na typu rozložení. Často bývá tato křivka kvůli zjednodušení popsána parametrickou funkcí. Existují také její vícerozměrné varianty, které počítají s klasifikací do více než dvou tříd.



Obr.2.1: Graf ROC

AUC – plocha pod ROC

Pro ROC obecně platí, že čím blíže se body charakteristiky nacházejí u levého horního rohu, tím je model přesnější. Jako charakteristika „predikčního potenciálu“ modelu se proto používá plocha pod ROC (AUC). Její hodnota se pohybuje v intervalu $\langle 0;1 \rangle$. Hodnota 0,5 odpovídá náhodné klasifikaci. AUC vyjadřuje kromě plochy pod křivkou také míru uspořádanosti prvků. „AUC je ekvivalentní pravděpodobnosti, že náhodně vybraný pozitivní prvek bude zařazen výše než náhodně vybraný negativní prvek“ [3]. Z toho také vyplývá podobnost s některými neparametrickými statistickými testy.

Body tvořící ROC křivku bývají aproximovány parametrickou funkcí. Popis ROC se tak zjednoduší na typ funkce a její parametry. Pokud je např. předpokládáno, že analyzovaná data mají normální rozložení, ROC křivka bude mít binormální rozložení [17], které je určeno dvěma parametry. Jedná se však o aproximaci ROC podmíněnou předpokladem o rozložení dat.

Ekvivalenty a aproximace AUC

AUC je interpretována jako plocha pod ROC křivkou. Existuje však hned několik jiných charakteristik a testů, které mají obdobné vlastnosti a jsou vzájemně ekvivalentní. Patří sem Gini index, Somersovo D_{xy}, Mann-Whitneyův test a Wilcoxonův test.

Existují však i algoritmy určené k aproximaci skutečné hodnoty AUC. Obecně je lze rozdělit na parametrické a neparametrické [2]. Parametrické přístupy vycházejí z předpokladu, že X a Y jsou vzájemně nezávislé veličiny s normálním nebo log-normálním rozložením [15,19]. Neparametrické přístupy se pak zaměřují na zjednodušené formy numerických algoritmů vycházející buď z Mann-Whitneova U testu [1,5] nebo kernelových metod [12,13,19].

3. Cíle disertační práce

1. **Popis a vysvětlení rozporu plynoucího z použití tradičních chybových funkcí v regresních klasifikátorech.**

   Postupy běžně používané k nastavení regresních klasifikátorů mohou vést k chybným řešením. Cílem je přesně specifikovat příčiny těchto chyb a ukázat je na konkrétních příkladech.

2. **Určení a definice chybové funkce, jejíž vlastnosti odstraní rozpory vyplývající z použití tradičních chybových funkcí v regresních klasifikátorech.**

   Kromě definování vhodné robustní chybové funkce je cílem také popis algoritmu pro její efektivní výpočet.

3. **Popis vlivu nové chybové funkce na strukturu regresních klasifikátorů.**

4. **Ověření výše navrhované metody na konkrétních příkladech a vyhodnocení dosažených výsledků.**

   Cílem je zpracovat dostatečné množství generovaných i reálných dat, aby bylo možné o použití robustní chybové funkce učinit statisticky významné závěry.

5. **Navržení dalších oblastí a aplikací, ve kterých je možné novou charakteristiku využít.**

   Cílem je nejen charakterizovat oblasti, ve kterých může nový přístup rozšířit stávající metodiku, ale také aplikovat nový parametr při řešení konkrétního praktického problému.

4 Zvolené metody zpracování

dvě části práce

Celá práce je tvořena dvěma hlavními částmi, teoretickou a praktickou. Teoretická část se zabývá výpočtem nové neparametrické statistiky a jejím vlivem na lineární model, je-li použita jako chybová funkce. Praktická část je pak věnována experimentálnímu ověření předpokladu, že použití robustní chybové funkce povede ke zlepšení kvality predikce. Nakonec je uvedeno další možné použití modifikované váhy pro účely nastavení šířky fuzzy množiny.

4.1 Rozpor ve způsobu použití regresních klasifikátorů

interpretace kvalitativní výstupní proměnné

Důvodem záměny kvalitativní proměnné za kvantitativní v regresních klasifikátorech je umožnit **výpočet odchylky ΔY**, ze které je dále počítána chyba predikce (ML, MNČ). Sama záměna je však **přidáním informace** do výstupních dat a může způsobit vznik ve skutečnosti neexistujících souvislostí.

chybová funkce vs. vyhodnocení modelu

Regresní klasifikátor se skládá ze dvou částí – regresního modelu a klasifikátoru. Regresní klasifikátor je nastavován tak, aby byla **funkce** jeho **regresního modelu co nejlépe proložena nominálními daty nahrazenými čísly**. Klasifikace je v procesu nastavení regresního modelu zcela opomenuta. Celý **model je** tak vlastně **nastavován na něco jiného, než na co je** nakonec **používán**.

závěry plynoucí z příkladů

- Chyba regresního modelu *Err* nevypovídá nic o klasifikačním potenciálu modelu vyjádřeném parametrem *AUC*.
- Některé z koeficientů $\boldsymbol{b}$ (*b0*, *b1*, ...) jsou z hlediska binární klasifikace v popsaných modelech zbytečné.
- V případě jednorozměrné vstupní veličiny *X* je z hlediska binární klasifikace transformace libovolnou ryze monotónní funkcí zbytečná.
- V případě vícerozměrné vstupní veličiny $\boldsymbol{X}$ jsou z hlediska binární klasifikace všechny ryze monotónní lineární nebo linearizovatelné modely ekvivalentní.

cíl disertační práce

Cílem disertační práce je pokusit se nalézt vhodnou chybovou funkci, kterou by bylo možno použít na nastavení regresních klasifikátorů a jejíž vlastnosti by řešily rozpory uvedené v předešlých bodech.

4.2 Robustní chybová funkce pro regresní klasifikátory

výchozí situace

Mějme nezávislou kvantitativní veličinu *X* a závislou binární veličinu *G* (třídy označme „*×*“ a „*o*“). Seřaďme uspořádané dvojce (*G;X*) podle velikosti *X*. Význam z hlediska binární klasifikace má pak informace o tom, jak kvalitně jsou data uspořádána z hlediska veličiny *G*. Charakteristikou vyjadřující míru uspořádání souboru dat číslem z intervalu $\langle 0;1 \rangle$ je váha *w* [7,9].

neuspořádaná sekvence

Definice: neuspořádaná sekvence *G* podle *X* – *S(G,X)* je vektor prvků veličiny *G* seřazených podle veličiny *X*.

Poznámka: mějme množinu uspořádaných dvojic (g_i,x_i). Seřaďme dvojice podle velikostí prvků x_i. Vytvořme vektor z prvků *G*, kde pro každé dva prvky g_i a g_j platí relace odpovídající relaci mezi odpovídajícími prvky veličiny x_i, x_j. Tímto vektorem je rozuměna sekvence *G* podle *X*, *S(G,X)*.

Poznámka: pojmem sekvence bude v dalším textu rozuměna neuspořádaná sekvence.

Poznámka: zápisem *{S(G,X)}* rozumíme množinu všech různých sekvencí a zápisem *|{S(G,X)}|* kardinalitu (počet všech sekvencí) množiny *{S(G,X)}*.

Příklad 4.1: mějme množinu uspořádaných dvojic (*G,X*). Uspořádané dvojice (×;1), (o;2), (×;2), (×;2), (o;3) jsou seřazeny podle veličiny *X*. Pak sekvencí *S(G,X)* mohou být všechny vektory z množiny permutací s opakováním *{S(G,X)} = {(×o××o),(××o×o),(×××oo)}*.

subsekvence

Definice: subsekvence $S_S(G,x)$ je vektor všech prvků v sekvenci takových, že pro libovolný prvek tohoto vektoru g_i platí, že $x_i=x$.

Poznámka: subsekvence je tvořena prvky *G*, které nelze na základě veličiny *X* seřadit, protože hodnoty veličiny *X* jsou pro všechny tyto prvky stejné.

Poznámka: sekvence se skládá z disjunktních subsekvencí.

Poznámka: množinu subsekvencí $\{S_S(G,x)\}$ tvoří permutace s opakováním ze všech prvků subsekvence.

Příklad 4.2: pro prvky z předešlého příkladu platí, že sekvence z nich sestavená obsahuje např. subsekvenci $S_S(G,2)$. Množina subsekvencí je $\{S_S(G,2)\} = \{(o××),(×o×),(××o)\}$. Počet subsekvencí $|\{S_S(G,2)\}|=3!/(1!2!)=3$.

Poznámka: Počet sekvencí (kardinalita) $|\{S(G,X)\}|$ je dána součinem kardinalit všech subsekvencí.

Příklad 4.3: kardinalita množiny sekvencí vyplývající z předešlého příkladu $|\{S(G,X)\}| = |\{S_S(G,1)\}|.\ |\{S_S(G,2)\}|.\ |\{S_S(G,3)\}| = 1.3.1 = 3$.

uspořádaná subsekvence

Definice: mějme určeno ordinální pořadí tříd veličiny G. Pak uspořádanou subsekvencí $S_{SU}(G,x)$ rozumíme takovou subsekvenci, jejíž prvky jsou uspořádány podle samotné veličiny G.

Poznámka: pro uspořádaná subsekvenci platí, že $S_{SU}(G,x) = S(S_S(G,x),G)$, $S_{SU}(G,x) \in \{S_S(G,x)\}$.

Poznámka: ordinální pořadí veličiny G je určeno buď na základě apriorního předpokladu nebo na základě rozložení tříd vyplývajícího z uspořádání podle veličiny X, tedy $S(G,X)$.

Příklad 4.4: v předešlém příkladě vyplývá z rozložení veličiny G v sekvencích $S(G,X)$, že ordinalita mezi třídami dána pořadím tříd $(\times,o)$, tedy $\times < o$. Uspořádanou subsekvencí takovou, že $S_{SU}(G,x) = S(S_S(G,x),G)$, je pouze subsekvence $(\times\times o)$.

uspořádaná sekvence

Definice: uspořádaná sekvence $S_U(G,X)$ je taková sekvence, jejíž všechny subsekvence jsou uspořádané.

Příklad 4.5: v předešlém příkladě byly uvedeny 3 různé subsekvence. Jediná uspořádaná subsekvence je $(\times\times o)$. Uspořádaná sekvence $S_U(G,X)$ je tedy sekvence $(\times\times\times oo)$.

krok, počet kroků

Poznámka: jedním krokem při změně uspořádání prvků ve vektoru je rozuměna vzájemná záměna dvou sousedních prvků různé třídy (hodnoty).

Definice: počtem kroků $K(V_1,V_2)$ je rozuměn minimální počet kroků nutných ke změně uspořádání vektoru V_1 na V_2.

Příklad 4.6: mějme $V_1 = (\times\times oo\times)$ a $V_2 = (\times\times\times oo)$. Pak $K(V_1,V_2) = 2$.

maximální a minimální počet kroků

Poznámka: mějme dánu klasifikační veličinu G se zadanou ordinalitou mezi jednotlivými třídami. Ordinalitou obrácenou pak značíme jako G'.

Poznámka: maximálním počtem kroků rozumíme číslo

$$max\_steps = K(S(G,G),S(G,G')) \tag{3}$$

Poznámka: minimální počet kroků $min\_steps = K(V_1,V_1) = 0$.

průměrný počet kroků

Poznámka: průměrný počet kroků je určen vztahem

$$\overline{K}(\{V_1,\ldots,V_n\},V_k) = [K(V_1,V_k) + \ldots + K(V_n,V_k)]/n\ . \tag{4}$$

počet kroků

Věta: počet kroků nutných k vytvoření uspořádané sekvence je určen vztahem

$$no\_steps = \overline{K}\left(\{S(G,X)\},S_U(G,X)\right) + K\left(S_U(G,X),S(G,G)\right). \tag{5}$$

Důkaz věty pro případ binární a obecné klasifikace je uveden v následujících dvou kapitolách.

váha – definice

Definice: váha w je definována jako rozdíl mezi maximálním počtem kroků a počtem kroků nutných k přeuspořádání $S(G,X)$ na $S(G,G)$ dělený maximálním počtem kroků.

$$w = \frac{max\_steps - no\_steps}{max\_steps} \tag{6}$$

Poznámka: pro výpočet váhy platí následující vztah:

$$w = \frac{K\left(S(G,G), S(G,G')\right) - \overline{K}\left(\{S(G,X)\}, S_U(G,X)\right) - K\left(S_U(G,X), S(G,G)\right)}{K\left(S(G,G), S(G,G')\right)} \tag{7}$$

Poznámka: zjednodušeně řečeno, mějme nějaký vektor veličiny *G*, ve kterém jsou prvky uspořádány podle veličiny *X*. Váha vyjadřuje, do jaké míry veličina *X* asociuje (předpovídá) veličinu *G*. Tuto „míru asociace“ lze zjistit tak, že prvky *G*, které jsou uspořádány podle veličiny *X*, přeuspořádáme podle *G*. V závislosti na tom, jak moc byla tato úprava náročná (počet kroků odpovídá počtu záměn prvků provedených např.algoritmem Bubble-Sort), vyjadřuje váha, jak moc si jsou veličina *X* asociuje veličinu *G*.

váha pro binární klasifikaci

Binární váhu *w* lze vypočítat v případě uspořádání k-pravém podle vztahu

$$w_P = \frac{n \cdot m - \left(\sum_{i}^{subsekvencí} \left(\frac{m_i \cdot n_i}{2} \right) + K_P \right)}{n \cdot m} \tag{8}$$

v případě k-levém

$$w_L = \frac{n \cdot m - \left(\sum_{i}^{sekvencí} \left(\frac{m_i \cdot n_i}{2} \right) + K_L \right)}{n \cdot m} \tag{9}$$

váha pro vícerozměrnou klasifikaci

Vícerozměrnou váhu w_P a w_L určují následující rovnice:

$$w_P = \frac{\sum_{i=1}^{C-1} \sum_{j=i+1}^{C} n_i \cdot n_j - \sum_{h=1}^{subsekvencí} \frac{\sum_{i=1}^{C-1} \sum_{j=i+1}^{C} n_{h,i} . n_{h,j}}{2} - \sum_{k=1}^{C-1} \sum_{i=1}^{n_k} \sum_{j=k+1}^{C} M_k(i,j)}{\sum_{i=1}^{C-1} \sum_{j=i+1}^{C} n_i \cdot n_j} \tag{10}$$

$$w_L = \frac{\sum_{i=1}^{C-1} \sum_{j=i+1}^{C} n_i \cdot n_j - \sum_{h=1}^{subsekvencí} \frac{\sum_{i=1}^{C-1} \sum_{j=i+1}^{C} n_{h,i} . n_{h,j}}{2} - \sum_{k=2}^{C} \sum_{i=1}^{n_k} \sum_{j=1}^{k-1} M_k(i,j)}{\sum_{i=1}^{C-1} \sum_{j=i+1}^{C} n_i \cdot n_j} \tag{11}$$

modifikovaná váha - s koeficientem w_k

Protože algoritmus váhy využívá řazení prvků na základě vzájemných záměn chybně zařazených prvků, stačí jednotlivé kroky odlišně ohodnotit v závislosti na tom, kolik prvků dané třídy se v trénovacích datech vyskytuje. Potom tedy záleží na tom, prvky jakých dvou tříd jsou v kroku zaměňovány a tento krok je vynásoben příslušným koeficientem. Méně zastoupené třídy jsou násobeny větším koeficientem než třídy zastoupené větším počtem prvků, což má v konečném důsledku za následek, že je algoritmus na počtu prvků v jednotlivých třídách nezávislý.

Pokud C je počet klasifikačních tříd, n_i je počet prvků v i-té třídě a C_M je počet prvků v nejpočetněji zastoupené třídě, krok mezi třídou i a j je násoben koeficientem

$$\frac{C_M^2}{n_i \cdot n_j} \tag{12}$$

Jinak platí původní algoritmus pro výpočet váhy. Pro maximální počet kroků platí následující zjednodušený vztah:

$$max\_steps = \sum_{i=1}^{C-1} \sum_{j=i+1}^{C} n_i \cdot n_j \cdot \frac{C_M^2}{n_i \cdot n_j} = C_M^2 \cdot \frac{C \cdot (C-1)}{2} \tag{13}$$

algoritmus pro výpočet váhy

```
VSTUP: X,Y
VYSTUP: AUC

seřaď(Y,X)// Fáze 1: seřaď veličinu Y podle veličiny X
zaměň(Y)  // Fáze 2: zaměň nominální proměnné Y za ordinální veličiny

          // získej informace z proměnných X a Y
cs        // počet tříd
C         // C(i) počet prvků ve třídě i
N         // počet všech prvků - např. velikost vektoru Y

E=C;
U=false;  // Náleží stávající prvek do neuspořádané subsekvence > 1 (US)?
UE()=0;   // UE(i) počet prvků třídy i v neuspořádané subsekvenci
NS=0;     // počet kroků
ANS=0;    // průměrný počet kroků

// Fáze 3: urči MNS - maximální počet kroků
MNS=0;
for i = 1 to cs-1
  for j = (i+1) to cs
    MNS=MNS+C(i)*C(j);
  end for
end for

// Fáze 4: urči NS - počet kroků nutných k uspořádání Y podle tříd Y
for i = 1 to N
  E(Y(i))=E(Y(i))-1;

  // rozezná začátek a střed US
  if (i<N)&&(X(i)==X(i+1))
    U=true;
    UE(Y(i))=UE(Y(i))+1;

  // rozezná konec US
  elseif (U==true)
    UE(Y(i))=UE(Y(i))+1;
    for j = 1 to (cs-1)
      for k = (j+1) to cs
        ANS=ANS+UE(j)*UE(k)/2;
      end for
    end for
    NS=NS+ANS;
    for j = 2 to cs
      if (UE(j)>0)
        for k = 1 to (j-1)
          NS=NS+UE(j)*E(k);
        end for
      end if
    end for
    ANS=0;
    UE()=0;
    U=false;
```

```
  // rozezná uspořádanou část sekvence
  else
    for j = 1 to (Y(i)-1)
      NS=NS+E(j);
    end for
  end if
end for

AUC = (MNS-NS)/MNS
```

algoritmus pro výpočet modifikované váhy

```
VSTUP: X,Y
VYSTUP: AUC

seřaď(Y,X)// Fáze 1: seřaď veličinu Y podle veličiny X
zaměň(Y)  // Fáze 2: zaměň nominální proměnné Y za ordinální veličiny

          // získej informace z proměnných X a Y
cs        // počet tříd
C         // C(i) počet prvků ve třídě i
N         // počet všech prvků - např. velikost vektoru Y
Cmax      // počet prvků v nejpočetněji zastoupené třídě

E=C;
U=false;  // Náleží stávající prvek do neuspořádané subsekvence > 1 (US)?
UE()=0;   // UE(i) počet prvků třídy i v neuspořádané subsekvenci
NS=0;     // počet kroků
ANS=0;    // průměrný počet kroků

// Fáze 3: urči MNS - maximální počet kroků
MNS=Cmax^2*cs*(cs-1)/2


// Fáze 4: urči NS - počet kroků nutných k uspořádání Y podle tříd Y
for i = 1 to N
  E(Y(i))=E(Y(i))-1;

  // rozezná začátek a střed US
  if (i<N)&&(X(i)==X(i+1))
    U=true;
    UE(Y(i))=UE(Y(i))+1;

  // rozezná konec US
  elseif (U==true)
    UE(Y(i))=UE(Y(i))+1;
    for j = 1 to (cs-1)
      for k = (j+1) to cs
        ANS=ANS+UE(j)*UE(k)/2*Cmax^2/C(j)/C(k);
      end for
    end for
    NS=NS+ANS;
    for j = 2 to cs
      if (UE(j)>0)
        for k = 1 to (j-1)
          NS=NS+UE(j)*E(k)*Cmax^2/C(j)/C(k);
        end for
      end if
    end for
    ANS=0;
    UE()=0;
    U=false;

  // rozezná uspořádanou část sekvence
  else
    for j = 1 to (Y(i)-1)
      NS=NS+E(j)*Cmax^2/C(j)/C(Y(i));
    end for
  end if
end for

AUC = (MNS-NS)/MNS
```

P&D vs. modifikovaná váha

Výpočetní náročnost algoritmu P&D (Provost and Domingos) je $O(C.N.log_2N)$. Algoritmus pro výpočet váhy má nižší komplexitu, pokud platí, že $C.log_2N>(log_2N+C^2)$, což platí v situaci, kdy $log_2N>(C+2)$. Lze předpokládat, že mezi počtem dat a tříd, do kterých je klasifikováno, bude platit vztah $N>>C$. Pak je algoritmus pro váhu efektivnější. Vzhledem k tomu, že algoritmus P&D v určitých případech selhává, není jeho srovnání s novou metodou věnována další pozornost

H&T vs. modifikovaná váha

Výpočetní náročnost algoritmu H&T (Hand and Till) je $O(C^2.N.log_2N)$. Výpočet pomocí váhy je efektivnější, pokud platí, že $C^2.log_2N>(log_2N+C^2)$, což jest v případě, že $log_2N>2$, tedy $N>4$. V tabulce 4.1 je na základě poměru komplexity obou algoritmů vyjádřeno, kolik procent náročnosti vyžaduje algoritmus váhy oproti algoritmu H&T (a to i v případě algoritmů modifikované váhy). Jak jest z tabulky 4.1 patrno, nový algoritmus pro výpočet váhy představuje v průměru několikanásobné zlepšení oproti algoritmu H&T.

Tabulka 4.1: Výpočetní náročnost nového algoritmu pro AUC oproti původnímu algoritmu vyjádřená v procentech v závislosti na počtu tříd a prvků

Počet prvků	Počet tříd						
	3	4	5	8	10	15	20
100	26	21	19	17	16	16	15
200	24	19	17	15	14	14	13
300	23	18	16	14	13	13	12
400	23	18	16	13	13	12	12
600	22	17	15	12	12	11	11
1000	21	16	14	12	11	10	10
1500	21	16	13	11	10	10	10
2000	20	15	13	11	10	10	9
4000	19	15	12	10	9	9	9
10000	19	14	12	9	9	8	8
20000	18	13	11	9	8	7	7
50000	18	13	10	8	7	7	7
100000	17	12	10	8	7	6	6
500000	16	12	9	7	6	6	6

4.3 Změna struktury regresního klasifikátoru

sférická transformace

Použití modifikované váhy jako chybové funkce mělo za následek snížení počtu parametrů nutných k nastavení lineárního modelu. Pro N-rozměrný model pak platí po sférické transformaci, že:

$$Y = \sin(\varphi_1)\sin(\varphi_2)\cdot\ldots\cdot\sin(\varphi_{N-1})X_1 + \cos(\varphi_1)\sin(\varphi_2)\cdot\ldots\cdot\sin(\varphi_{N-1})X_2 + \\ + \cos(\varphi_2)\sin(\varphi_3)\cdot\ldots\cdot\sin(\varphi_{N-1})X_3 + \ldots + \cos(\varphi_{N-1})X_N \qquad (14)$$

Skutečný počet parametrů je tedy $N-1$ a je dán úhly $\varphi_1,\ldots,\varphi_{N-1} \in \langle 0,2\pi \rangle$ oproti klasickému modelu s $N+1$ parametry $\boldsymbol{b}$ z intervalu $(-\infty,\infty)$.

4.4 Aplikace robustní chybové funkce

aplikace váhy jako chybové funkce

Bylo zpracováno 782 vygenerovaných datových souborů (cca. 20 milionů hodnot) a 7 reálných datových souborů (cca. 40 tisíc hodnot). Vyhodnocení experimentů tvoří dvě části. První je posouzení několika nulových hypotéz o stejnosti modelů nastavených tradičním způsobem a pomocí nové chybové funkce. Ptáme se, jestli se kvalita modelů liší natolik, že toto již nelze vysvětlit pouhou náhodou. Druhá část vyhodnocení vyjadřuje, o kolik se liší jednotlivé modely v predikční kvalitě.

První část vyhodnocení experimentů zamítla hypotézu, že všechny tři modely jsou stejně kvalitní, na hladině významnosti 97,5%. Na hladinách 95% a 90% pak byly zamítnuty hypotézy o stejnosti modelu AUC a logitového, dále pak AUC a lineárního. Nejdůležitější bylo závěrečné srovnání nového a tradičního přístupu. Zde bylo možno hypotézu a stejnosti výsledků zamítnout na hladině významnosti 50%. Hypotézu tedy zamítnout nelze. Zjištěné výsledky lze shrnout do následujícího tvrzení: „Přestože je nový model významně lepší než jednotlivé modely lineární či logitový, použijeme-li vždy přesnější z uvedených dvou modelů, není nový přístup statisticky významně lepším."

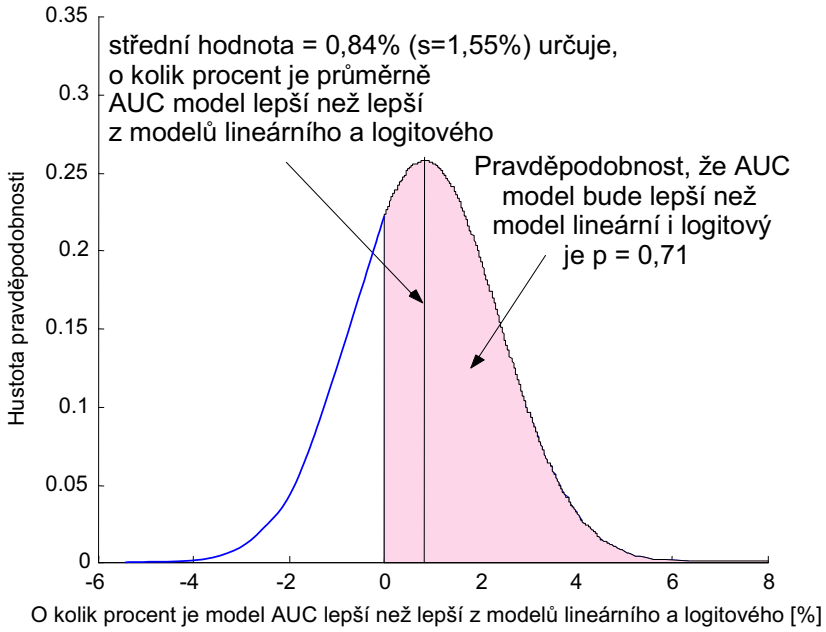
Porovnání nového modelu s lineárním a logitovým modelem znázorňuje tabulka 4.2. Její hodnoty vyjadřují v procentech rozdílnost v přesnosti nalezených řešení mezi modelem s AUC a modelem lineárním, logitovým a lepším z těchto dvou modelů.

Tabulka 4.2. Rozdíl přesnosti predikce nového modelu oproti modelu lineárnímu nebo logitovému vyjádřený v procentech.

Datové soubory	Lineární	Logitový	Lepší(Log,Lin)
CA+	6,13	-0,13	-0,13
CB+	0,18	-0,96	-0,96
BC+	-0,04	0,38	-0,04
CS+	5,57	0,11	0,11
IO+	3,81	3,85	3,81
LD+	5,95	2,42	2,42
ID+	0,69	0,68	0,69
Průměr	**3,18**	**0,91**	**0,84**
Rozptyl	**2,62**	**1,53**	**1,55**

Zkratky viz. text nebo abecedně seřazené vysvětlené zkratky.

Nejdůležitější porovnání, tedy model s AUC vs. lepší z modelů (lineární a logitový), je patrné z následujícího grafu. Rozdíl není statisticky významný.

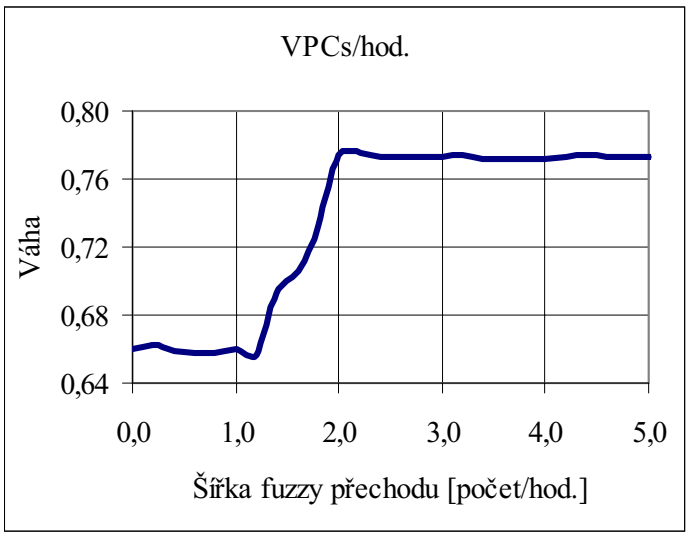


Obr. 4.1: Rozdíl přesnosti nově nastaveného lineárního modelu ve srovnání s lepším z modelů lineárního a logitového.

4.5 Použití váhy pro nastavení fuzzy množin

aplikace váhy na nastavení šířky fuzzy množin

V praktické aplikaci byla váha použita pro nastavení šířky fuzzy množin v medicínském diagnostickém modelu [8,10]. Podstatou problému byla skutečnost, že rozhodování o rizikovosti vychází z určitého počtu nezávisle naměřených veličin, přičemž na základě každé je stav pacienta posouzen dvouhodnotově – je/není rizikový. V publikované práci byly ostré přechody jednotlivých faktorů fuzzifikovány a na určení optimální šířky byla použita váha. Na obr. 4.2 je patrný průběh veličiny VPCs/hod. Se zvětšováním šířky fuzzy přechodu se zvětšovala váha samotného parametru, avšak jen od určité hodnoty, od které již k významnému zlepšení nedocházelo. Z grafu byla pak odečtena optimální šířka pro daný parametr.



Obr. 4.2: změna kvality predikce v závislosti na šířce fuzzifikované přechodové funkce

Takto bylo zpracováno 6 parametrů, pro každý z nich byla stanovena nová šířka fuzzy přechodu. Dále byl vytvořen jeden konečný parametr ze součtu dílčích rizik určených z jednotlivých faktorů. Celkově pak došlo ke zlepšení kvality konečné binární predikce je/není rizikový. Výsledky jsou uvedeny v následující tabulce:

Tabulka 4.3: Výsledky dosažené fuzzifikací kritických parametrů

Parametr		**BRS**	**EF**	**LP**	**SDANN**	**SDNN**	**VPCs**	**Suma r.f.**	**Fuzzy suma r.f.**
Senzitivita		66,7	66,7	50,0	61,1	38,9	44,4	83,3	77,8
Specificita		69,1	84,9	15,4	71,3	86,0	87,9	75,7	86,4
PPV		12,5	22,6	3,8	12,4	15,6	19,5	18,5	27,5
Senz.	**PPV=50%**	5,6	5,6	-	11,1	16,7	-	38,9	44,4
Spec.		98,9	99,6	-	99,3	98,2	-	97,4	97,1
Váha faktoru / váha fuzzy faktoru		0,979 / 0,708	0,758 / 0,761	0,673 / -	0,619 / 0,663	0,625 / 0,719	0,660 / 0,772	0,832	0,843
Optimální šířka fuzzy přechodu		3,5	5,0	-	21,0	7,0	2,5	-	-

5 Závěr

hlavní výsledky disertační práce

Významným výsledkem je nový algoritmus pro stanovení hodnoty charakteristiky AUC (Area Under the Receiver Operating Characteristic). Je výpočetně několikanásobně méně náročný než současné postupy. Na mezinárodní konferenci WSEAS získala publikace představující tyto výsledky ocenění za nejlepší studentskou prezentaci [7].

Aplikace robustní chybové funkce je aktuálním tématem. Dosažené výsledky prezentované v této disertační práci úspěšně konkurují současným řešením, která využívají aproximací charakteristiky AUC. Otevřeno zůstává jak pokračování v prezentovaném výzkumu, tak využití nové charakteristiky v řadě dalších oblastí matematického modelování.

teoretická část

Teoretickou část práce otevírá podrobný popis rozporu plynoucího z tradičního pojetí regresních klasifikátorů. Hlavními argumenty jsou výhrady vůči převodu nominálních proměnných (např. ano/ne) na diskrétní hodnoty (např. 0/1), čímž je do výstupních dat přidána nová informace, a dále pak následné použití chybových funkcí (metoda nejmenších čtverců, maximální věrohodnost), které prokládají těmito diskrétními hodnotami zvolenou parametrickou funkci. Na konkrétních příkladech je ukázáno, že uvedené postupy vedou v případě použití stejných dat a modelů k jejich rozdílným nastavením, která nejsou ekvivalentní; nejlepší nalezené řešení se liší při použití různých chybových funkcí. Na druhou stranu dva identicky klasifikující modely různého typu jsou těmito chybovými funkcemi hodnoceny jako zcela odlišné. Dochází dokonce k tomu, že nulová chyba není v modelu ani teoreticky dosažitelná. Tyto skutečnosti plynou z faktu, že běžné chybové funkce zohledňují vzdálenost jednotlivých bodů od prokládané funkce, což je však smysluplné při predikci veličiny kvantitativní. Konečné kritérium používané pro vyjádření klasifikačního potenciálu spojitých prediktorů (AUC) však zohledňuje něco odlišného – míru uspořádanosti prvků. Další práce je proto zaměřena na určení vhodné robustní chybové funkce, jejíž vlastnosti budou lépe odpovídat skutečným požadavkům kladeným na klasifikátory.

Byla zavedena nová charakteristika *váha*, která vyjadřující míru asociace závislé veličiny Y na nezávislé veličině X. Její hodnota je v případě binární klasifikace shodná s parametrem AUC (nebo např. Somersovým D_{xy}, Man Whitneyovým pořadovým testem, Gini indexem). Odlišnost její interpretace se projevuje až při vícerozměrné klasifikaci. Váha je citlivá na počet prvků jednotlivých tříd v trénovacích datech. Dále byla definována charakteristika *modifikovaná váha*, která na četnostech ve třídách závislá není a je ekvivalentní AUC. Hlavním přínosem modifikované váhy oproti tradičnímu algoritmu AUC je odlišný přístup k výpočtu, který se ve výsledku projevuje několikanásobným snížením výpočetní náročnosti (konkrétní hodnota závisí na dvou parametrech – počtu tříd a počtu prvků). Práce publikované v současné době se právě z důvodu velké výpočetní náročnosti zaměřují na aproximaci AUC. Prezentovaný algoritmus nabízí zrychlení výpočtu bez ztráty přesnosti nalezeného řešení.

Dalším přínosem je analýza vlivu robustní chybové funkce na strukturu lineárního (linearizovatelného) modelu. Praktickým závěrem je snížení stupně volnosti z N+1 na N-1 (kde N je počet vstupních veličin X). Pro zjednodušení prohledávání definičního oboru všech možných řešení modelu je provedena jeho sférická transformace, na které je také názorně vysvětleno, proč a jak je možné, že uvedené snížení počtu parametrů nemá na kvalitu klasifikace žádný vliv.

praktická část

Úspěšnost aplikace robustní chybové funkce byla ověřena na datech generovaných (782 datových souborů, 20 miliónů hodnot) i reálných (7 datových souborů, 40 tisíc hodnot). Statistické srovnání nového přístupu s přístupy tradičními je založeno na několika hypotézách. Posuzována byla nulová hypotéza, že modely lineární, logitový a model lineární nastavený pomocí robustní chybové funkce jsou ekvivalentní. Výsledky na reálných datech statisticky významně dokazují, že robustní chybová funkce dosáhne lepšího nastavení lineárního modelu regresního klasifikátoru, přestože má menší počet parametrů. To však platí při srovnání s jednotlivými typy modelů zvlášť. Pokud je nový postup porovnán s výběrem vždy lepšího z modelů tradičních, zlepšení sice patrné je, není však již statisticky významné. Výsledek experimentu, jehož smyslem bylo dokázat, že lineární model v regresních klasifikátorech lze nastavit robustní chybovou funkcí lépe než pomocí běžných chybových funkcí, je pozitivní, zlepšení však není dostatečné, aby bylo prohlášeno za statisticky významné.

Druhou aplikací je použití váhy k nastavení šířky fuzzy množiny. Byl řešen konkrétní problém z oblasti medicíny, který se týká fuzzifikace kritických hodnot na faktorech indikujících zvýšené riziko náhlé srdeční smrti po infarktu myocardu. Metodika vedla ke zpřesnění predikce. Příčinou byla fuzzifikace samotná, avšak určení optimální šířky přechodu bylo podmíněno využitím váhy jako chybové funkce.

hlavní úskalí a další směry výzkumu

K nastavení lineárního modelu robustní chybovou funkcí je použit genetický algoritmus. Jeho slabinou je skutečnost, že negarantuje nalezení nejlepšího řešení, negarantuje ani nalezení stejného řešení při opakovaném výpočtu. Vlastností robustní chybové funkce je navíc fakt, že připouští existenci celých podprostorů řešení, které jsou chybovou funkcí považovány za ekvivalentní. Spojení těchto dvou vlastností tak otevírá prostor pro řadu nepřesností. Protože v postupu nastavení modelu existují určité rezervy, je jejich eliminace vzhledem k současným pozitivním výsledkům hlavním směrem další práce.

Možným rozšířením algoritmu váhy jsou implementace matice nákladů a umožnění vzájemného porovnání více veličin najednou. Z hlediska aplikace váhy se nabízí prostor v algoritmech rozhodovacích stromů, fuzzy logiky nebo nelineárních modelů.

Seznam zkratek

AUC	Area Under ROC (Receiver Operating Characteristic). Plocha pod křivkou ROC.
EF	Ejekční frace.
FPR	False positive rate. Chybovost.
H&TA	Hand and Till approach. Postup výpočtu AUC, který publikovali autoři Hand a Till.
IBL	Instance based learning. Učení založené na zapamatování vybraných vzorů.
ID3	Algoritmus pro tvorbu rozhodovacích klasifikačních stromů.
k-NN	k Nearrest Neighbourhood. Klasifikační algoritmus k-nejbližších sousedů.
LP	Pozdní potenciály.
ML	Maximální věrohodnost.
MLE	Maximum likelihood estimator. Postup používaný k určení parametrů modelu za využití maximální věrohodnosti.
MNČ	Metoda nejmenších čtverců.
NPV	Negative predictive value. Negativní prediktivní hodnota.
P&DA	Provost and Domingos approach. Algoritmus pro výpočet AUC, který publikovali autoři Porovst a Domingos.
PPV	Positive predictive value. Pozitivní prediktivní hodnota.
ROC	Reciever Operating Characteristic. Křivka ROC.
SDNN	Variabilita srdeční frekvence.
TPR	True positive rate. Úplnost, senzitivita.
VPCs	Extrasystoly.
WSEAS	World scientific and engineering academy and society.

Literatura

[1] BAMBER, D.C.: The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology 1975; 12: 387-415.

[2] FARRAGI, D., REISER, B.: Estimation of the area under the ROC curve. Statistics in Medicine. 2002, 21:3093-3106.

[3] Fawcett T.: ROC Graphs: Notes and Practical Considerations for Researchers. HP Laboratories, © 2004 Kluwer Acadaemic Publisher.

[4] FRANK, E., HARRELL, J. Regression Modeling Strategies. NY: Springer, 2001. 568 pages. ISBN 0-387-95232-2.

[5] Hanley, J.A., McNeil, B.: The meaning and use of the area under the Receiver operating Characteristic (ROC) curve. Radiology. 1982, p.29-36.

[6] Hastie T., Tibshirani R., Friedman J.: The Elements of Statistical Learning. Springer, 2001. ISBN 0-387-95284-5.

[7] HONZÍK, P. Area under the ROC Curve by Bubble-Sort Approach (BSA) In Automatic Control, Modeling and Simulation (ACMOS'05). 7th WSEAS International Conference on AUTOMATIC CONTROL, MODELING AND SIMULATION (ACMOS '05). Praha: WSEAS, 2005, s. 494 - 499, ISBN 960-8457-12-2

[8] HONZÍK, P., HRABEC, J., LÁBROVÁ, R., SEMRÁD, B., HONZÍKOVÁ, N. Fuzzification, weight and summation of risk factors in a patient improves the prediction of risk for cardiac death. Scripta medica, Brno, Masaryk University in Brno, ISSN 0211-3395, 2003, roč. 76, č. 3, s. 141 - 148

[9] HONZÍK, P., HRABEC, J., SEMRÁD, B,. HONZÍKOVÁ, N. Risk Stratification Of Patients After Myocardial Infarction By The Fuzzy And Weighted Methods. Analysis of Biomedical Signals and Images. 2002, vol. 16, no. 6, p. 463-465. ISSN 1211-412X.

[10] HONZÍKOVÁ, N., FIŠER, B., SEMRÁD, B., LÁBROVÁ, R., HONZÍK, P., HRABEC, J. Nonlinear analysis of inter-beat data in patients after myocardial infarction. Acta Physiologica Hungarica, ISSN 0231-424X, 2002, roč. 89, č. 1-3,

[11] Huber, P.J. 1981, Robust Statistics (New York: Wiley).

[12] LLOYD, C.J., YONG, Z.: Kernel estimators of the ROC curves are better than empirical. Statistics and Probability Letters 1999; 44:221-228.

[13] LLOYD, C.J.: Using smoothed receiver operating characteristic curve to summarize and compare diagnostic systems. *Journal of the American Statistical Association* 1998; 93:1356-1364.

[14] Press W.H., Teukolsky A.S., Vetterling W.T., Flannery B.P.: Numerical Recepies in C, 1992. ISBN 0-521-43108-5.

[15] REISER, B., FARAGGI, D.: Confidence intervals for the generalized ROC criterion. Biometrics 1997; 53: 644-652.

[16] Schölkopf B., Smola A.J.: Learning with Kernels. MIT Press, Cambridge, MA, 2002. ISBN 0-262-19475-9.

[17] Tilbury J.B.: Evaluation of Intelligent Medical Systems. PhD Thesis 2002.

[18] Weisstein, E.W. "Least Squares Fitting." From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/LeastSquaresFitting.html

[19] ZOU, K.H., TEMPANY, C.M., FIELDING J.R., SILVERMAN, S.G.: Original smooth receiver operating characteristic curve estimation from continuous data: statistical methods for analyzing the predictive value of spiral CT of ureteral stones. Academic Radiology 1998; 5:680-687.

Petr Honzík - stručný životopis

OSOBNÍ ÚDAJE:

Příjmení, jméno, titul:	Honzík Petr, Ing., Dipl.-Ing.
Ulice, číslo:	Ukrajinská 13/24
Město, PSČ:	Brno, 62500
Země:	Česká republika
E-mail:	honzikp@feec.vutbr.cz
Datum narození:	27.10.1977

VZDĚLÁNÍ:

1992-1996	gymnázium tř. Kpt. Jaroše 14, Brno, CZ, *zaměření: matematika*
1996-2001	Vysoké učení technické v Brně, Fakulta elektrotechniky a informatiky Ústav automatizace a měřicí techniky
2000-2001	FernUniversität v Hagenu, Německo, Fakulta elektrotechniky Ukončeno získáním německého diplomu a titulu Dipl.-Ing.
od 2001	Ph.D. studium, Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav automatizace a měřicí techniky

PRAXE, GRANTY:

1996	Orgrez, měření a zpracování dat v tepelné elektrárně Dětmarovice
1999	Siemens – programování fuzzy regulátorů na PLC S5
2001	BD-Sensors, programování v Javě, server zabudovaný v senzoru
2002	Získán grant FRVŠ, internetové technologie
2003/4	Vývoj a správa virtuálních laboratoří v projektu distančního vzdělávání
2005	Technický asistent na VUT Brno

VÝBĚR PUBLIKACÍ:

HONZÍKOVÁ, N., FIŠER, B., SEMRÁD, B., LÁBROVÁ, R., HONZÍK, P., HRABEC, J. Nonlinear analysis of inter-beat data in patients after myocardial infarction. *Acta Physiologica Hungarica*, ISSN 0231-424X, 2002, roč. 89, č. 1-3.

HONZÍK, P., ŠEDIVÁ, S., BRADÁČ, Z. Internet Technologies for Use in Virtual Laboratories. *WSEAS Transactions on Computers, Malta*, ISSN 1109-2750, 2003, roč. 2, č. 2, s. 481 – 485

HONZÍK, P., ŠEDIVÁ, S., HONZÍK, B. Software Tools for Use in Virtual Laboratories. *The 10th EDS 2003 Electronic Devices and Systems Conference*, Brno 9.-10.9.2003. Brno, Czech Republic: VUT Brno, 2003, s. 126 - 129, ISBN 80-214-2452-4

HONZÍK, P., HRABEC, J., LÁBROVÁ, R., SEMRÁD, B., HONZÍKOVÁ, N. Fuzzification, weight and summation of risk factors in a patient improves the prediction of risk for cardiac death. *Scripta medica, Brno, Masaryk University in Brno*, ISSN 0211-3395, 2003, roč. 76, č. 3, s. 141 - 148.

JIRSÍK, V., HONZÍK, P. Hybrid Expert System. *WSEAS TRANSACTIONS on INFORMATION SCIENCE & APPLICATIONS*, Austria, Salzburg. ISSN 1790-0832.

HONZÍK, P. Area under the ROC Curve by Bubble-Sort Approach (BSA). *7th WSEAS International Conference on AUTOMATIC CONTROL, MODELING AND SIMULATION (ACMOS '05)*. Praha: WSEAS, 2005, s. 494 - 499, ISBN 960-8457-12-2

JAZYKOVÉ ZNALOSTI:

Angličtina	6 let, dobrá znalost
Němčina	5 let, studium a absolvování zkoušek v Německu, diplomová práce a její obhajoba v němčině, dobrá znalost

Abstract

The aim of the thesis is to improve the accuracy of the regression classifiers by the use of the robust loss function. *Weight*, a new nonparametric characteristic and loss function is described in the theoretical section of the thesis. Furthermore the *modified weight* is introduced. It equals to the AUC (Area Under Receiver Operating Characteristic). The computational complexity of the *modified weight* is several times lower compared to the complexity of traditional algorithms used to AUC evaluation. The result is a meaningful one, since the high computational complexity is one of the reasons the approximation of AUC rather then AUC proper are being commonly employed. In the applied section, the *weight* and genetic algorithms were applied to setup the regression classifier. The experimental results are better in comparison with the results of the current methods but they are not statistically significant. The next research is focused on the improvement of the teaching algorithm with the aim to achieve better results that will be statistically significant.