

# Implementace grafů významnosti do FAKE GAME a porovnání s WEKou

Jaroslav Sýkora

**Abstrakt**— V první části textu je popsána implementace grafů významnosti do programu Fake Game. Ve druhé části je provedeno experimentální porovnání programů Weka a Fake Game při vyhodnocování významnosti třech různých datových sad.

## I. ZADÁNÍ

Do programu FAKE GAME implementujte graf významnosti vstupů. Proveďte srovnání vyhodnocování významnosti mezi Fake Game a Wekou na třech souborech dat: Iris [1], Mushrooms [2] a umělá data se známou významností.

## II. ÚVOD

Centrálním problémem strojového učení je identifikace reprezentativní podmnožiny atributů, ze kterých se dá pro danou úlohu sestavit klasifikační model. Algoritmy pro určování podmnožiny atributů typicky spadají do dvou kategorií: *feature ranging* a *subset selection*. Metody *feature ranging* ohodnotí atributy nějakou metrikou a následně eliminují všechny atributy, které nedosáhnou požadovaného skóre. Metody *subset selection* přímo vyhledávají v množině atributů optimální podmnožinu.

## III. IMPLEMENTACE GRAFU VÝZNAMNOSTI

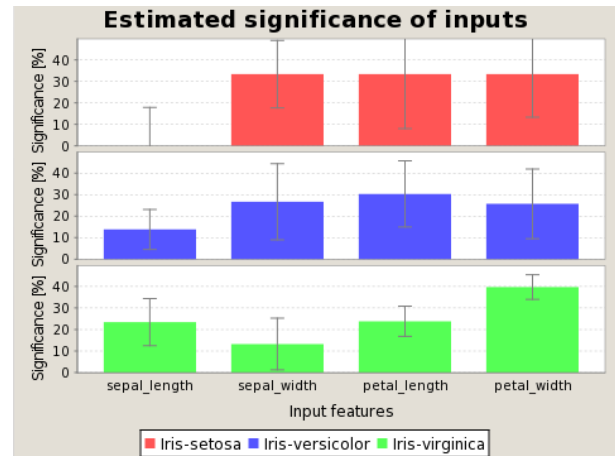
Semestrální práce se skládala z implementační a experimentální části.

V implementační části bylo úkolem naprogramovat grafy významnosti vstupů do programu FAKE GAME [3] s použitím knihovny JFreeChart [4]. Graf by mělo být možné zobrazit jak v okně, tak i v tiskové sestavě.

Data významnosti vstupů, tedy vstupní data grafu, bylo původně možné zobrazit jen v textové podobě v okně "Feature ranking" (obr. 4). Toto okno lze otevřít pomocí volby "Estimate significance of inputs" v menu "Model" (obr. 3). V textové podobě (obr. 4) je signifikance vytištěna v odstavcích (sekcích) podle jednotlivých výstupních proměnných. V rámci sekce jsou v řádcích vygenerované modely (Iris-setosa 0, Iris-setosa 1, ...) a ve sloupcích vstupní proměnné. Pro zobrazení v grafu bylo požadováno sloučit významnosti různých modelů pomocí mediánu, tj. každý sloupec seřadit a vybrat prostřední hodnotu. (V případě sudého počtu hodnot se vybere aritmetický průměr dvou prostředních hodnot.)

Mimo vlastní hodnoty se v grafu zobrazují i vymežující značky výběrové směrodatné odchylky [5], která je definována takto:

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$$



Obr. 1. Příklad grafu významnosti

kde  $x_i$  jsou čísla ve sloupci (tj. stejná vstupní a výstupní proměnná, různý model), a  $\bar{x}$  je aritmetický průměr. Prakticky je v implementaci použit vzorec:

$$s = \sqrt{\frac{1}{N-1} \left( \sum_{i=1}^N x_i^2 - N\bar{x}^2 \right)}$$

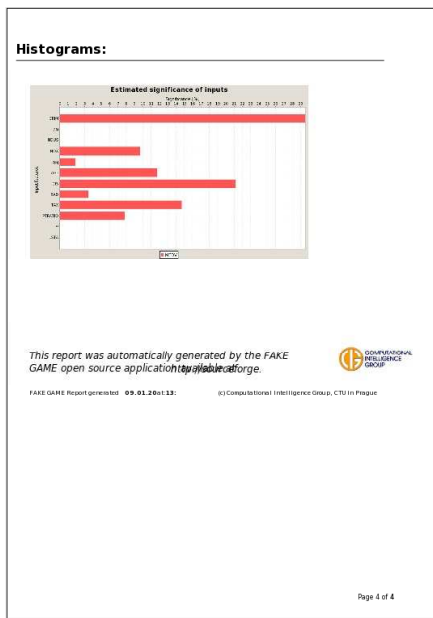
kteří nevyžaduje předběžný výpočet průměru. Druhý sčítanec pod odmocninou totiž lze počítat průběžně zároveň s výpočtem sumy čtverců  $x_i$  během jediného programového cyklu procházejícího vstupní data. Příprava grafu je v metodě `buildChart_EstimatedSignificanceOfInputs()`. Na obr. 1 je ukázka výsledného grafu pro dataset Iris.

Požadavek na zobrazení grafu v tiskové sestavě je vyřešen tak, že se metodou popsanou výše vygeneruje objekt `JFreeChart`, ten se uloží do obrázkového souboru `pic/generated/input_signif.png` v metodě `TreeData::saveChartPng` a obrázek se vytiskne v sestavě (obr. 2).

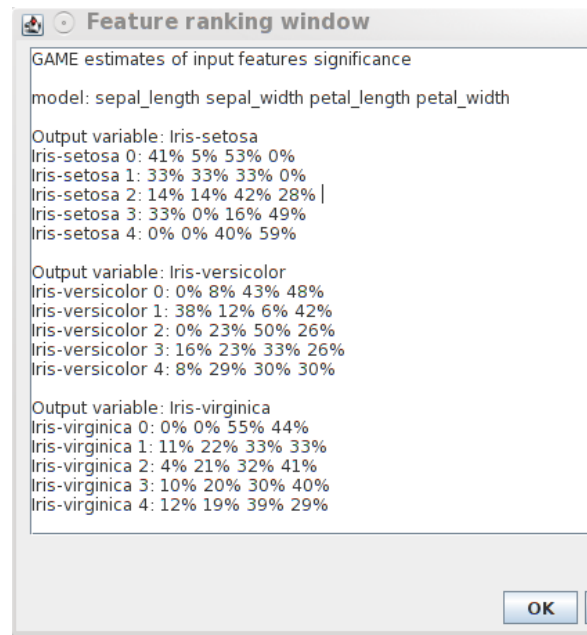
## IV. EXPERIMENTY

V druhé části semestrální úlohy bylo úkolem experimentálně porovnat určování významnosti mezi FAKE GAME [3] a Wekou [6]. Srovnání bylo provedeno na třech souborech dat:

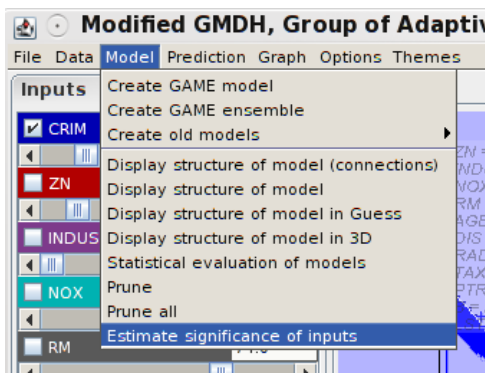
- Iris [1] - známý klasifikační dataset se 150 instancemi, 4 reálnými atributy a 3 třídami.
- UCI Mushrooms [2] - klasifikační dataset s 8124 instancemi, 22 atributy (multivariety) a 2 třídami.
- umělá data se známou významností (hypercube.csv) - 50 reálných atributů a 1 třída. Na obrázcích 5 a 6 je ukázka závislosti výstupu (třídy) na atributech  $i1$  a  $i20$ . Na první



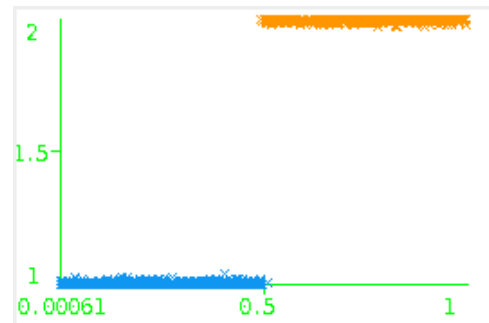
Obr. 2. Graf vložený do tiskové sestavy



Obr. 4. Textový výstup vyhodnocení významnosti



Obr. 3. Položka v menu Fake Game



Obr. 5. Rozložení atributu i1 vs třída (hypercube, jitter)

pohled je vidět, že pomocí atributu *i1* se dá vzorek 100% klasifikovat, zatímco atribut *i20* bude mít naopak téměř nulovou vypovídací hodnotu (malou významnost).

Dataset Mushrooms je multivarietní, tj. každý atribut nabývá několika diskretních hodnot. Odpovídá to např. datovému typu *enum*, který známe z běžných programovacích jazyků. Například první atribut *cap-shape* může nabývat hodnot: *bell=b*, *conical=c*, *convex=x*, *flat=f*, *knobbed=k*, *sunken=s*. Pro použití v neuronových sítích je nutno převést atributy do kódu 1 z *N*. Z jednoho atributu *cap-shape* se tak stane atributů několik: *cap-shape-b*, *cap-shape-c*, *cap-shape-x*, *cap-shape-f*, *cap-shape-k* a *cap-shape-s*. Tímto způsobem naroste celkový počet atributů z 22 na 126.

Významnost byla u každého datasetu určena třemi způsoby:

- 1) Weka Explorer,
- 2) Fake Game, pouze lineární neurony, ensemble 5 std. modelů,
- 3) Fake Game, všechny druhy neuronů, ensemble 5 std. modelů.

Ve Wece byla použita funkce "Select attributes" v okně Exploreru. Byl použit evaluátor *CfsSubsetEval* ([7]) a vyhledávací metoda *GreedyStepwise* (parametry  $-T -1.7976931348623157E308 -N -1$ ). Evaluátor *CfsSubsetEval* je založený na hypotéze, že dobrá množina atributů by měla obsahovat atributy, které vysoko korelují s výstupní proměnnou (třídou) a přitom málo sami mezi sebou.

Evaluátor *CfsSubsetEval* bohužel neposkytuje relativní významnost atributů, pouze jejich seznam. Pro nejobtížnější dataset Iris (tab. III, viz dále) bylo navíc provedeno vyhodnocení pomocí evaluátoru *InfoGainAttributeEval* s vyhledávací metodou *Ranker* (parametry  $-T -1.7976931348623157E308 -N -1$ ).

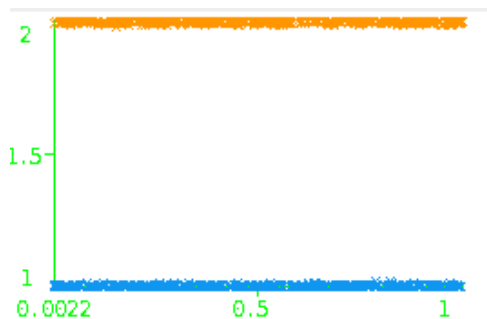
Výsledky jsou shrnuty v tabulkách I (Hypercube), II (Mushrooms) a III (Iris). Pro jednoduchost jsou v tabulkách uvedeny pouze první 4 vybrané atributy. U sloupců hodnot významnosti z Gamu mají uvedená procenta význam: "významnost % / směrodatná odchylka %".

TABULKA II  
VÝZNAMNOSTI PRO DATASET MUSHROOMS

Edible Mushrooms			
rank	Weka, CfsSubsetEval	GAME, linear	GAME, all types
1.	odor-f	odor-n 15%/2.5%	odor-n 11%/6%
2.	odor-n	odor-f 14%/10%	gill-size-b 9%/4.5%
3.	gill-size-n	gill-size-b 12%/8%	odor-f 8%/6%
4.	gill-color-k	stalk-surf-ab-k 7%/5.3%	gill-size-n 7%/5.3%
Poisonous Mushrooms			
rank	Weka, CfsSubsetEval	GAME, linear	GAME, all types
1.	odor-f	odor-f 12%/2%	odor-n 14%/3.2%
2.	odor-n	odor-n 10%/6.5%	odor-f 11%/4.2%
3.	gill-size-n	gill-size-n 9%/4.6%	gill-size-b 10%/4.8%
4.	gill-color-k	gill-color-b 9%/8.1%	stalk-surf-ab-k 6%/5%

TABULKA III  
VÝZNAMNOSTI PRO DATASET IRIS

Iris-Setosa				
rank	Weka, CfsSubsetEval	Weka, InfoGainAttributeEval	GAME, linear	GAME, all types
1.	sepal_width	petal_length 0.918	petal_length 59%/13.7%	petal_width 55%/26.6%
2.	petal_length	petal_width 0.918	sepal_length 19%/12.6%	petal_length 37%/17.5%
3.		sepal_length 0.594	petal_width 27%/9.7%	sepal_length 1%/12%
4.		sepal_width 0.337	sepal_width 0%/0%	sepal_width 0%/2.7%
Iris-Versicolor				
rank	Weka, CfsSubsetEval	Weka, InfoGainAttributeEval	GAME, linear	GAME, all types
1.	sepal_width	petal_length 0.751	petal_length 45%/14.7%	petal_length 40%/1.8%
2.		petal_width 0.712	petal_width 36%/13.4%	petal_width 26%/9.4%
3.		sepal_width 0.189	sepal_length 17%/2.7%	sepal_width 25%/12%
4.		sepal_length 0.166	sepal_width 3%/8.9%	sepal_length 13%/5.9%
Iris-Virginica				
rank	Weka, CfsSubsetEval	Weka, InfoGainAttributeEval	GAME, linear	GAME, all types
1.	petal_width	petal_width 0.75	petal_length 46%/17.9%	petal_width 50%/22.9%
2.		petal_length 0.744	petal_width 33%/18.7%	sepal_width 25%/11%
3.		sepal_length 0.366	sepal_length 30%/16.8%	petal_length 22%/17.4%
4.		sepal_width 0	sepal_width 15%/8.2%	sepal_length 17%/9.8%



Obr. 6. Rozložení atributu i20 vs třída (hypercube, jitter)

## V. DISKUSE

Nejjednodušší pro interpretaci výsledků je dataset Hypercube (tab. I). Weka vybrala 4 atributy: i1, i2, i3, i4, u Gamu získal v obou případech největší podíl atribut i1 (obr. 5).

U datasetu Mushrooms (tab. II) došlo též k pěkné shodě mezi Wekou a Gamem. Atributy odor-f a odor-n byly vybrány ve všech případech.

Nejproblématictější je dataset Iris (tab III). Při použití evaluátoru CfsSubsetEval dochází k rozporům: např. u třídy Iris-Setosa vybrala Weka atributy sepal\_width a petal\_length, zatímco u Gamu byl atribut sepal\_width ohodnocen jako nejméně významný.

Z tohoto důvodu bylo ve Wece provedeno nové vyhodnocení významnosti, tentokrát s evaluátorem InfoGainAttributeEval (tab. III, druhý sloupec). Tento evaluátor poskytuje oproti CfsSubsetEval i relativní významnosti atributů; ty jsou v tabulce uvedeny za jejich názvy. Při tomto druhém způsobu vyhodnocení významnosti ve Wece jsou již naměřené hodnoty v dobré shodě s Gamem.

TABULKA I  
VÝZNAMNOSTI PRO DATASET HYPERCUBE

Hypercube			
rank	Weka, CfsSubsetEval	GAME, linear	GAME, all types
1.	i1	i1 100%/32%	i1 43%/26%
2.	i2		i2 10%/18%
3.	i3		
4.	i4		

## VI. ZÁVĚR

V první části textu byla popsána implementace grafů významnosti do programu Fake Game spolu s některými jejími detaily.

Ve druhé části bylo provedeno experimentální porovnání programů Weka a Fake Game při vyhodnocování významnosti. Viděli jsme, že na jednodušších datech (Hypercube, Mushrooms) dochází k dobré shodě obou programů, u náročnějších dat (Iris) je však rozhodující vybrat ve Wece vhodný evaluátor.

## LITERATURA

- [1] UCI Machine Learning Repository: *Iris Data Set*,  
<http://archive.ics.uci.edu/ml/datasets/Iris>.
- [2] UCI Machine Learning Repository: *Mushroom Data Set*,  
<http://archive.ics.uci.edu/ml/datasets/Mushroom>.
- [3] Computational Intelligence Group: *Fake Game Project*,  
<http://neuron.felk.cvut.cz/game/>
- [4] *JFreeChart Project*,  
<http://www.jfree.org/jfreechart/>
- [5] Wikipedia: *Směrodatná odchylka*,  
[http://cs.wikipedia.org/wiki/Směrodatná\\_odchylka](http://cs.wikipedia.org/wiki/Směrodatná_odchylka)
- [6] The University of Waikato: *Weka 3: Data Mining Software in Java*,  
<http://www.cs.waikato.ac.nz/ml/weka/>
- [7] Mark A. Hall: *Correlation-based Feature Selection for Machine Learning*, 1998.