Mendelova univerzita v Brně
Provozně ekonomická fakulta

Klasifikace a shlukování ekonomických dat pomocí metod strojového učení

Diplomová práce

Vedoucí práce:
prof. RNDr. Ing. Jiří Šťastný, CSc.

Bc. Andrea Kosová

Brno, 2012

Ráda bych poděkovala svému vedoucímu diplomové práce prof. RNDr. Ing. Jiřímu Šťastnému, CSc. za hodnotné rady a čas, který mi věnoval.

Prohlašuji, že jsem tuto diplomovou práci vypracovala samostatně pod dohledem vedoucího diplomové práce, za použití zdrojů a informací, které jsou uvedeny v literatuře.

V Brně, 20.května 2012 .....

Abstract

Kosová, A. Classification and clustering of economic data by using the methods of machine learning. Diploma thesis. Brno, 2012

The thesis is research dependencies between attributes, including editing data, analysis and determine possible results using different machine learning methods. This thesis consists of theoretical and practical part. Theoretical part includes introduction to neural networks, algorithms used for classification and clustering, work environment Weka and Matlab including Neural Network toolbox and also marketing research. The practical part focus on the treatment of data, dependencies of attributes selection and classification in the Weka software and clustering in Matlab. The results of different methods, which were calculated on the basis of the above technologies are compared at the close.

Abstrakt

Kosová, A. Klasifikace a shlukování ekonomických dat pomocí metod strojového učení. Diplomová práce. Brno, 2012

Předmětem diplomové práce je výzkum závislosti mezi atributy včetně úpravy dat, analýza a stanovení možných výsledků pomocí různých metod strojového učení. Diplomová práce se skládá z teoretické a praktické části. Teoretická část zahrnuje seznámení s neuronovými sítěmi, algoritmy použité ke klasifikaci a shlukování, pracovní prostředí Weka a Matlab včetně Neural Network toolbox a také charakteristiku marketingového výzkumu. Praktická část je zaměřena na úpravu dat, výběr závislosti atributů a klasifikování v programu Weka, dále shlukování v programu Matlab. Na závěr jsou porovnány výsledky různých metod, které byly vypočítány na základě výše uvedených technologií.

Obsah

1 Úvod

Marketingový výzkum se zabývá získáváním, úpravou, zpracováváním, analýzou a prezentací informací, sloužících k řešení různých marketingových okolností v organizaci či podniku pomoci statistických nástrojů. Data jsou získaná od respondentů v různých věkových kategoriích žijící v České republice formou dotazníku. Jedná se o výzkum chování spotřebitelů na trhu s potravinami. Dotazník obsahuje mnoho instancí a atributu, pro než je vhodné použít moderní metody strojového učení a následně porovnat dosažené výsledky.

Pro použití metod strojového učení je nejprve potřeba dotazník upravit. Je třeba zkontrolovat výsledky, zda odpovídají dotazované otázce a také zjistit, zda byly zodpovězeny na všechny otázky z pohledu spotřebitele. Rovněž je nezbytné zakódovat příslušné odpovědi, aby se daly zpracovat v softwaru MATLAB.

Ke klasifikaci byl použit rozhodovací strom J48, neuronová síť Multilayer Perceptron a rozhodovací pravidlo PART. U pravidla zeroR budou výsledky jen informativní, posléze budou použity pro porovnání ostatních metod, aby neměly horší výsledky, než je právě toto pravidlo. Pomocí metody PART budou data analyzovaná a budou zjišťovány závislosti atributů. Klasifikace těchto metod bude prováděna v programu Weka. Následně bude vytvořena samoorganizující mapa, určena pro shlukování, spojena s učící vektorovou kvantizací. Tyto výpočty budou uskutečněné v programu MATLAB. Nejdříve budou nasazeny klasifikační metody v programu Weka za použití všech atributů. U všech atributů budou srovnány algoritmy zeroR, j48 a MLP. Dále bude použita filtrace k vyběru atributů, které dávají dobré závislosti a výsledky. U těchto proměnných budou srovnaný metody MLP, J48 a LVQ.

2 Cíl a metodika

Cílem práce je analýza daných ekonomických dat pomocí metod strojového učení. Bude proveden výběr vhodných kombinací tříd ze vstupních dat a shlukování pomocí metody SOM a LVQ v programu Matlab. Dále budou tato data klasifikována v programu Weka.

Pro splnění daného cíle práce bude nutné data upravit a navrhnout postup pro výběr vhodných kombinací tříd. Následně pak bude použita metoda Kohonenovy samoorganizované mapy pro učení bez učitele, dále označení výstupních neuronů kategoriemi a doučení sítě jedním z algoritmu LVQ pro učení s učitelem v programu Matlab.

3 Marketingový výzkum

3.1 Marketingový výzkumný systém

Je chápán jako podsystém marketingového informačního systému, v jehož oblasti je ustanoveno, jaké data jsou potřebná, jakým způsobem jsou kvantifikovatelná, jakým způsobem získaná.(Simová,2005) Data marketingového výzkumu lze dělit dle (Jandová, Přibová, 2006)na:

- Primární
  - aktuální data z trhu, od konkurence nebo od spotřebitele
  - unikátní
  - získání pomocí profesionálního výzkumu v terénu
  - v čase měnné
- Sekundární
  - Vnitřní (údaje o nákladech, obratu, počtu zákazníků, počtu prodaných služeb a výrobků, aj.)
  - Externí (získané data ze státní statistiky, publikací, novin, časopisů, výroční zprávy atd.)

Úloha marketingového informačního systému v podniku

Na základě požadavků marketingových pracovníků se získávají, zpracovávají a také vyhodnocují potřebné informace, které se předávají zpět marketingovému oddělení k dalšímu využití. Marketingový informační systém pracuje s různými zdroji a typy dat. Těmito typy jsou interní, operativní a externí data a také data získaná prostřednictvím marketingového výzkumu. Může mít různou podobu, jako například jednoduchá tabulka v Excelu nebo komplexní počítačový systém umožňující získání aktuálních informací, které se zpracovávají, analyticky vyhodnocují a modelují možné situace, chování a trendy trhu. (Simová,2005)

3.2 Sestavování dotazníku

Dotazník je nástroj pro sběr údajů pro různé typy výzkumů. Je složen z posloupnosti otázek, které se liší počtem a typem otázek, způsobem tvorby. Dotazník by měl být sestaven tak, aby vyhovoval jak potřebám, tak i cílům výzkumu a získal hodnotné informace od respondentů.(Simová,2005)

Sestavování dotazníku se musí řídit pravidly dle (Jandová, Přibová, 2006)

- Otázky musí být
  - jednoduché,

  - srozumitelné
  - sociálně přijatelné
- Otázky nesmí být
  - příliš dlouhé
  - sugestivní
- Formulace otázky
  - jednoznačná,
  - nedvojmyslná

Typy otázek dle (Jandová, Přibová, 2006)

- uzavřené(nabídka možných variant odpovědi)
- polouzavřené(po nabídce možných variant může respondent doplnit i něco navíc)
- otevřené(odpověď je spontánní)
- přímé
- nepřímé
- identifikační

3.3 Zpracování a analýza dat

Zpracování dat dle následujících kroku:

- kontrola dotazníků (ujištění, že data jsou v pořádku)
- popřípadě úprava dat (prověření, zda data vyjadřují to, co mají)
- klasifikace (přesné definování třídicích vlastností)
- kódování dat (přiřazení číselných hodnot slovním výrazům a kategoriím, pro vyhodnocování počítačem)(Simová,2005)

Tři přístupy v analýze dat:

1. vyhodnocení jedné proměnné - využívají se metody popisné statistiky průměry, medián, modus, četnost, procenta, směrodatné odchylky, rozptyly (zjišťuje rozdělení četností zjištěných hodnot znaků a úrovně)

2. vyhodnocení dvou proměnných - zjišťuji závislosti dvou proměnných, zda se navzájem ovlivňují (regresní a korelační analýzy)
3. vyhodnocení několika proměnných - zkoumá se zde vzájemné působení několika proměnných (metody mnohonásobné regrese, multidimensionální analýzy, shlukové analýzy, analýzy časových řad) (Simová,2005)

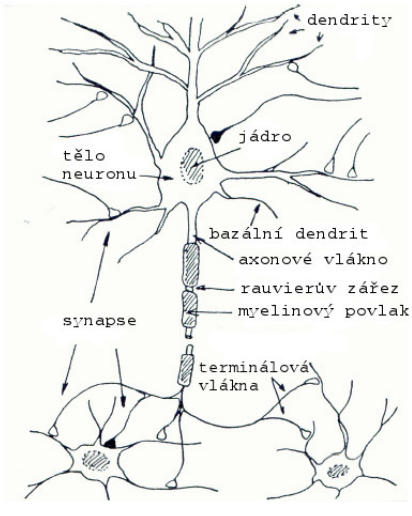
4 Neuronové sítě

Neuronové systémy pracují bez algoritmu. Jejich činnost je založena na procesu učení, kdy se neuronová síť postupně co nejlépe adaptuje k řešení dané úlohy. Neuronovými sítěmi lze řešit mnoho složitých úloh dle Šnorek (Šnorek, 2004) jako je:

- Rozpoznávání písma
- Identifikace podpisů
- Převod mluvené řeči na psaný text
- Predikce v pojišťovnictví a bankovnictví
- Identifikace radarových údajů
- Detekce explozivních látek na letištích
- Chybová diagnostika strojů
- Třídění zásilek podle PSČ
- Aplikace v oblasti data mining
- jiné

4.1 Biologický model neuronu

Neuron je základním stavebním prvkem centrálního nervového systému specializovaný na zpracování, uchování a přenos informací.



Obrázek 1: Biologický neuron(Churý,2005)

Struktura biologického neuronu

Biologické neurony se skládají ze čtyř částí:

- Soma – tělo neuronů s buněčným jádrem
- Dendrity – tvoří vstupy. Jsou krátké, tenké s počtem okolo 10^4 (Tučková, 2009)
- Axon – rozvětvený výstup z neuronu je spojen s dendrity (vstupy) dalších neuronů pomocí synapse(spojky) (Šnorek, 2004)
- Synapse – zakončují axony a tvoří informační rozhraní. Jsou vyznačované velkou plasticitou, která souvisí s průchodností synapsí. Ta se adaptivně přizpůsobuje potřebám a situacím. Její průchodnost se mění při procesu učení.(Tučková, 2009)

4.2 Matematický model neuronu

Matematické modely lze dělit dle Tučkové (Tučková, 2003)

- Podle povahy vstupních dat:
  1. Binární (s nespojitými aktivačními (přenosovými) funkcemi)
  2. Spojité (se spojitými aktivačními (přenosovými)funkcemi)
- Podle složitosti na modely:
  1. první generace (popisuje jednodušší biologické neurony)
  2. druhé generace (popisuje složitější chování neuronů s impulsním nebo chaotickým charakterem)
- podle aktivačních funkcí a charakteru algoritmu učení
  - lineární
  - nelineární
  - statické
  - dynamické
  - deterministické
  - stochastické
- dělení algoritmu učení dle (Zelinka, 1998)
  1. učení s učitelem – Umělá neuronová síť se učí srovnáváním skutečného výstupu s žádaným výstupem. Váhy se mění dle nějakého algoritmu, jenž zabezpečuje snižování chyby mezi aktuálním a požadovaným výstupem. (Šnorek, 2004)

2. učení bez učitele - Zde není žádné vnější kritérium. Učení je založeno na informacích, které samotná síť během celého procesu učení získala (hledá vzorky, které mají společné vlastnosti). (Šnorek, 2004)

5 Vícevrstvé sítě – ML- perceptron

Vícevrstvé sítě tvoří početnou skupinu paradigmat. Paradigma je souhrnný název pro topologii a algoritmus učení. Základem vícevrstvé sítě typu Perceptron je model neuronu s lineárně váženou obvodovou funkci a aktivační skokovou funkcí (treshold function). Vstup je tvořen vektorem vzorků x =(x_1,x_2,…,x_n) a vektorem cílových hodnot c=(c_1,c_2,…,c_n) v n-dimensionálním prostoru. (Tučková, 2003)

5.1 Učící algoritmus Perceptronu

dle Tučkové: (Tučková, 2003)

- V jedné iteraci pro obvodovou funkci u (x)a výstupní funkci platí

$$u(x) = w_0 + \sum_{i=1}^{n} w_i \, x_i$$

$$y(x) = \begin{cases} 1, & u(x) > 0 \\ 0, & u(x) \leq 0 \end{cases}$$

- Výstup je porovnán s požadovanou (cílovou) hodnotou a je spočítána chyba e:

$$E = c - y$$

  - Je-li y = c a chyba e = 0 (váhy se nezmění)
  - Je-li y = 0 a c =1, chyba je e = 1(k váhám se přičte vstupní hodnota) – váhový vektor se přiblíží ke vstupnímu vektoru, zvýší se klasifikace 1
  - Je-li y = 1 a c = 0, chyba e = -1 (od vah se odečte vstupní hodnota)- váhový vektor se vzdálí od vstupního vektoru, klasifikace = 0

- Pro adaptaci vah a prahů platí rovnice

$$W_{ij}\ (t + 1) = w_{ij}\ (t) + e_j\ x_i$$
$$\Theta_j{=}(t + 1) = \Theta_j + e_j$$

5.2 ADALINE A MADALINE

ADALINE(Adaptive Linear Neuron) používá lineární aktivační funkci. Jako algoritmus učení je použit LMS algoritmus (Least Mean Squares). Je to minimalizace Euklidovy vzdálenosti mezi požadovanou odezvou sítě a vnitřním potencionálem neuronu.

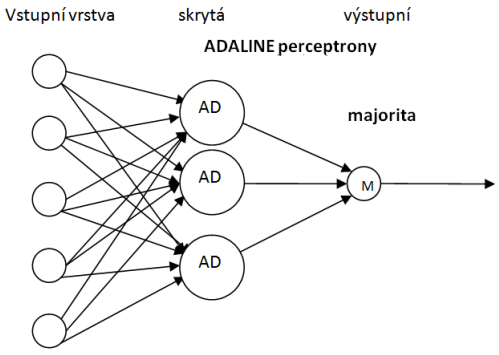
$$w_{ij}\ (t+1) = w_{ij}(t) + \ e_j\ (t)\ x_{\ i}\ (t)$$

$$e_j(t) = y_j\ (t) - c_j\ (t)$$

- e_j (t)... chyba učení,
- c_j (t)... požadovaná hodnota,
- y_j (t)... výstupní hodnota
- Výstup z neuronu může nabývat libovolnou hodnotu. Tato hodnota není omezena na 0 a 1.(Tučková, 2003)

MADALINE

Multiple Adaptive Linear Element se označuje jako síť s jednou skrytou a jednou výstupní vrstvou.



Obrázek 2: MADALINE

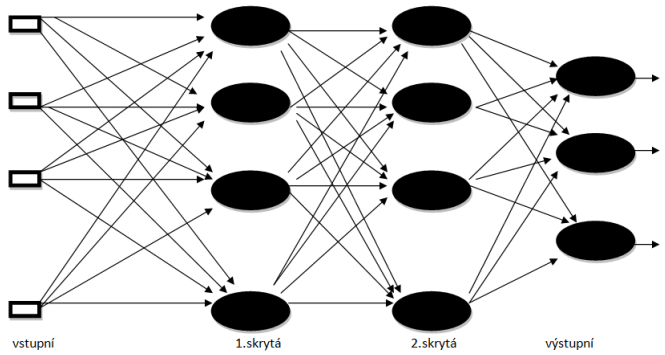
MADALINE se učí s učitelem. Pokud je u poloviny perceptronů skryté vrstvy nastaven výstup na +1 je i takový výstup celé sítě. Pokud není alespoň polovina perceptronů je výstup sítě -1. Jestliže nesouhlasí skutečný a požadovaný výstup sítě, adaptují se váhy perceptronu skryté vrstvy. Jedná se o váhy, jejichž vnitřní potenciál je nejblíže k nulové hodnotě. Postup se při učení opakuje do té doby, dokud nedojde na výstupu sítě ke shodě (Šnorek, 2004)

5.3 Architektura sítě back-propagation

Struktura je tvořena několika vrstvami neuronů. Dle Tučková (Tučková, 2003) se síť skládá ze tři typu vrstev:

- jedna vstupní vrstva
- jedna nebo více skrytých vrstev
- jedna výstupní vrstva

Ze vstupu příjde signál a vstupní vrstva tento signál rozvětvuje na všechny neurony první skryté vrstvy. Tato vrstva má se vstupní úplné propojení(propojen je každý s každým).(Šnorek, 2004)



Obrázek 3: struktura sítě

Algoritmus Backpropagation

Nejčastější algoritmus učení pro vícevrstvé sítě MLP založený na minimalizaci chyby neuronové sítě.(Štencl,Šťastný, 2011) Tento algoritmus obsahuje dvě fáze dle Zelinka (Zelinka, 1998)

1. Aktivační fáze: Tato fáze je používaná při učení a vybavování sítě. Síť se inicializuje a váhy se nastaví na hodnotu náhodných čísel. Váhy jsou v rozmezí 0,5 až -0,5. Nastavení vah probíhá v opačném směru, než jakým se šíří vstupní informace. Obsažená informace ve vstupním vektoru se šíří ze vstupu na výstup. Vstupní vektor je ve vstupní vrstvě rozvětven a každý spoj je ohodnocen vahou. Součet ohodnocených spojů je argument přenosové funkce, kde výsledná hodnota je výstupem z neuronu sloužící jako vstupní hodnota do dalších neuronů ve vyšší vrstvě.(Zelinka, 1998)
2. Adaptační fáze: Výstupní vektor je porovnán s požadovaným vektorem a tento rozdíl je použit pro výpočet nových vah. Nejprve se opraví váhy u spojů, které vstupují do výstupní vrstvy. Poté jsou opravené váhy u nižší vrstvy. Pokud se dosáhne vrstvy vstupní, tato fáze je ukončena a opakuje se

fáze aktivační. Po každém porovnání vstupním a požadovaným vektorem se tato hodnota uchová v paměťové proměnné a sumarizuje se s dalšími postupně získanými rozdíly. Tato hodnota za celou trénovací množinu se nazývá globální chyba.(Zelinka, 1998)

Popis algoritmu učení dle(Šnorek, 2004)

```
Nahodna_inicializace_vah
  Repeat
     Repeat
        Vyber_vzor_z_trenovaci mnoziny
        Priloz_vybrany_vzor_na_vstupy_site
        Vypocti_vystupy_site
        Porovnej_vystupy_s_pozadovanymi_hodnotami
        Modifikuj_vahy
     Until
   Until globalní_chyba < kriterium
```

6 Kohonenova síť

Je druhem umělých neuronových sítí, které nepotřebují k trénování učitele.(Šnorek, 2004) Kohonenová síť je jednovrstvá a je uspořádaná v řádku nebo v ploše. Tato metoda pracuje na základě shlukování. Základem shlukové analýzy je hledání vzájemných závislostí a společných vlastností v množině předkládaných vzorů.(Tučková, 2003)
Kohonenova síť, nazývaná Self Organizing Map(samoučící se neuronová síť) dále SOM, byla vyvinuta finským profesorem T. Kohonenem začátkem 80. let v Helsinkách. (Šnorek, 2004)Kohonen se zabýval výzkumy odhalení procesů samoorganizace v mozku, ale také i adaptivní schopnosti učit se.(Šnorek, 2004) Vymezil dva základní mechanizmy vyznačující se síti s prostorovou samoorganizaci:(Zelinka, 1998)

- nalezení výstupního (vítězného) neuronu, který nejlépe odpovídá předloženému vstupnímu vzoru
- modifikace vah a spojů vítěze a jeho nejbližšího (vítězného) okolí

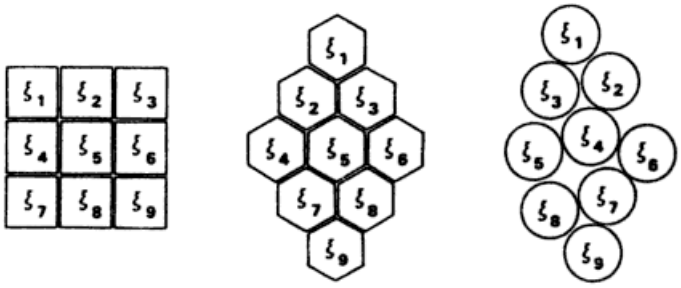
6.1 Architektura sítě SOM

Fyziologové vypracovali mapy lidského mozku, z kterých je patrné, jaké oblasti mozku řídí nebo zpracovávají podněty z jednotlivých orgánů. Data, která jsou přijatá z okolí (senzorů) lze na ně pohlížet jako na vektory, které vstupují do neuronové sítě a ta je transformuje. Transformace převádí vícedimenzionální data do nižší dimenze. Tento proces se nazývá komprese dat. Základní myšlenkou SOM vychází tedy z komprese dat, kdy lidský mozek používá pro zpracování a uchování informace vnitřní reprezentaci dat, která má nižší dimenzi než dimenze původní.(Šnorek, 2004)
SOM má jedinou výkonnou vrstvu. Tato vrstva se obvykle prezentuje jako plošné uspořádání neuronů. Každý neuron je zároveň výstupem a je spojen se svými sousedy (Tučková, 2003).
Základní matematickou operací je hledání minimální vzdálenosti mezi vstupními vektory a souřadnicemi neuronů v mapě. Nejčastěji se používá Euklidovská vzdálenost. Neuron, který má minimální vzdálenost, je vítězem (neuron nejbližší zkoumanému vzoru) (Tučková, 2003).
Nechť je x $\in R^n$náhodně zvolený datový vektor. Funkce p(x) je hustota pravděpodobnosti n-rozměrného datového vektoru x do dvou dimenzí. Vstupní vektor x se porovnává se všemi w_i jakýkoliv matice.

$$j^* = \arg\min||x - w_i||$$

$||x - w_i||$...Minimální euklidovská vzdálenost

Existují různé topologie map (např. kruhové, čtvercové, hexagonální), které je možno vidět na obrázku níže.(Kohonen, 2001)



Obrázek 4: topologie map

Neurony Kohonenovy vrstvy jsou uspořádány. Uspořádání slouží k tomu, aby každý neuron měl definované sousedy. Existuje dvojdimenzionální uspořádání, které se používá nejčastěji, ale také lineární, či jiné. Počet neuronů je libovolný. Čím je počet větší, tím je pokrytí prostoru vstupních dat lepší. V praxi se pohybuje rozsah neuronů od 10 x 10 do 25 x 25.(Šnorek, 2004)

6.2 Učení Kohonenovy sítě

Proces učení je autonomní (bez vnější pomocné informace) a probíhá iterativně. V každém kroku probíhá adaptace vah, která je založena na porovnání vstupních vzorů a reprezentantů představovaným každým neuronem. Po nalezení neuronu, který se nejlépe shoduje se vstupním vzorem, jsou upraveny jeho váhy a dále váhy všech neuronů, které tvoří jeho okolí. Na začátku trénování jsou všechny neurony uspořádány náhodně, až po učícím procesu dostanou konkrétní tvar. (Šnorek, 2004)

Proces učení lze rozdělit do několika kroků:

- Krok 1.Inicializace
  Váhy jsou nastaveny na malá náhodná čísla. Parametr učení je nastaven na libovolnou hodnotu v intervalu 0 - 1. Dále je také potřeba nastavit kolem každého výstupního neuronu počáteční velikost všech okolí. (Smith, Gupta, 2003)Tato velikost se volí pro všechny neurony stejně tak, aby je okolí pokrylo. (Šnorek, 2004)

- Krok 2. Předložení vzoru

Předložení nového trénovacího vektoru na vstupní vrstvu neuronové sítě(Šnorek, 2004)

$$X(t) = x_0(t), x_1(t), \ldots, x_{n-1}(t)$$

x_1(t)...vstup uzlu i v čase t.(Vondrák, 2000)

- Krok 3. Výpočet vzdálenosti vzorů
  Výpočet všech vzdálenosti d_j mezi výstupními neurony j dle vztahu (Zelinka, 1998)

$$d_j = \sum_{i=0}^{n} -1[x_i(t) - w_{ij}(t)]^2$$

  X_i(t)...jednotlivé složky vektoru vstupního vzoru

  W_i(t)...váhy mezi i-tým vstupem a j-tým výstupním neuronem

  (Šnorek, 2004)

- krok 4. Výběr nejbližšího neuronu (určení vítěze)
  Určení vítěze prostřednictvím minima dj jako j*.

$$d_j^* = min_j(d_j)$$

(šnorek)

- Krok 5. Adaptace vah(Zelinka, 1998)
  Při nalezení neuronu, k jenž má vstupní vzor nejblíže, se vykoná adaptace vah pro vítězný neuron, ale také pro jeho okolí. Vykoná se rozdíl mezi vstupním a referenčním vektorem:

$$w_{ij}(t+1) = w_{ij}(t) + (t)[x_i(t) - w_{ij}(t)]$$

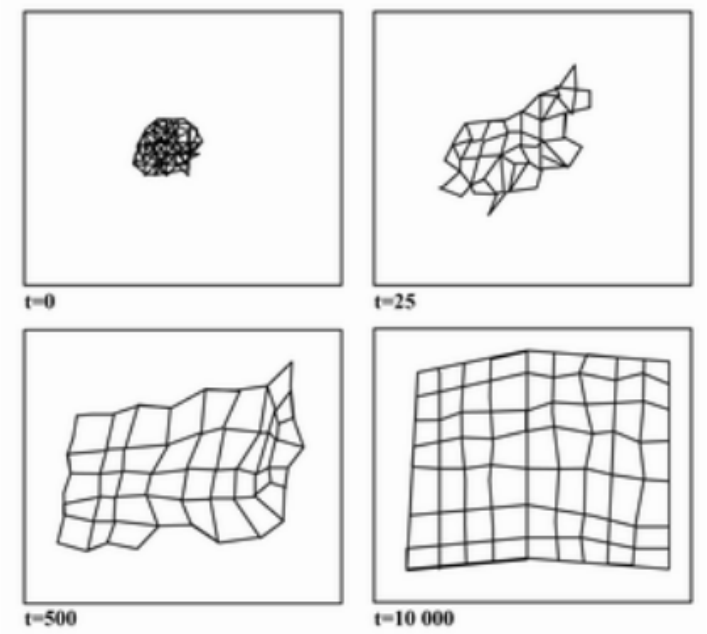
  w_{ij}(t)...váhy mezi i-tým vstupem a j-tým výstupním neuronem

  j...všechny neurony v okolí vybraného neuronu j*

  $\eta(t) \ldots zahrnujeuckroka\ velikostokol$

Velikost okolí se postupně sníží až na minimum. Váhy neuronu jsou adaptovány dle algoritmu učení. Jsou tvořeny shluky – třídy, odpovídající průměru všech vektorů, které do této třídy patří. Vstupní data jsou rovnoměrně rozložena v dvojdimenzionálním prostoru, konkrétně ve čtvercové oblasti. Na (obrázku 5) lze vidět rozložení neuronů během režimu učení. Je zde ukázáno 0, 25, 500 a 10 000 kroků. Z obrázku, který se nachází dole vpravo, je zřejmé, že neurony zde byly optimálně rozloženy tak, aby pokryly datový prostor.(Vondrák, 2000)

- Krok 6. Pokračování učícího procesu Přejdeme ke kroku 2. Pokud jsme nevyčerpali všechny vzory, které chceme síť naučit (Šnorek, 2004)



Obrázek 5: rozložení neuronů

7 Metoda LVQ

Je to metoda, která kombinuje učení bez učitele a učení s učitelem. Může řešit nejen klasifikační úlohy, ale také jednoduché rozpoznávání. Učení je založeno na definici kvantizační oblasti mezi sousedními vektory kódové knihy. LVQ je obdoba Voronoiových množin u vektorové kvantizace(VQ). (Tučková, 2009) Kvantizace je aproximací analogové hodnoty jednou z konečného počtu číselných hodnot. Pokud se jedná o více parametrů současně, mluvíme o vektorové kvantizaci. Používá se ke snížení možných stavů. Vektorový kvantizér zachycuje množinu vektorů do předem neznámého počtu konečných tříd. Při VQ se rozděluje trénovací množina na n oblastí, které jsou reprezentované n centroidy. Množina kódových vektorů (centroidů) tvoří tzv. kódovou knihu (codebook). (Tučková, 2003) Veronoiova mozaika je dvoudimenziónální prostor, který má konečný počet kódových vektorů odpovídající souřadnicím. (Tučková, 2009)

Princip LVQ:

- Nejprve se použije na trénovací vzory klasická Kohonenova síť, která neobsahuje informaci o příslušnosti ke třídám. Tím se dosáhne hrubého nastavení vektorů do pravděpodobných budoucích poloh(shluků). Poté se všem vektorům přiřadí symbol třídy, kterou reprezentuje dle nejbližšího centroidu.(Šnorek, 2004)
- Dále se upraví polohy třídy pro možnost zařazení nového vzoru a pokud vzor nelze zařadit do stávající třídy, vytvoří se nová třída.(Tučková, 2003)
- Fáze učení je o něco komplikovanější.Varianty tohoto učení je LVQ1,LVQ2,LVQ3, které se liší způsobem hledání optimální hranice mezi třídami.(Šnorek, 2004)

7.1 LVQ1

Cílem této metody je nalezení dostatečně vzdálených tříd od hraničních přímek.(Tučková, 2003)

Pro daný trénovací vzor X se známou klasifikací se vybere ten neuron j^* respektive jeho váhový vector w_{j^*}, který je dle Euklidovské metriky nejblíže. Jen pro tento neuron provedeme změnu váhového vektoru dle následujícího vztahu, ostatní vektory zůstanou nezměněny dle Šnorek (Šnorek, 2004) :

1. Pokud byl vzor klasifikován dobře (učení):

$$W_{ij}{}^*(t+1)=w_{ij}{}^*(t)+ (t)[x_i(t) - w_{ij}{}^*(t)]$$

2. Zda byl vzor klasifikován špatně (odučení):

$$W_{ij}{}^*(t+1)=w_{ij}{}^*(t)- (t)[x_i(t) - w_{ij}{}^*(t)]$$

a) Pro všechny ostatní vektory, kde 0<=i <=N -1 probíhá pro všechny elementy vstupního vzoru. Proměnná 0<= (t)<=1 se nazývá parametr učení.

$$W_{ij}(t+1)= w_{ij}(t)$$

Tento algoritmus redukuje počet rozložení w_i v blízkosti hraničních ploch a také se tímto minimalizuje chybná klasifikace.

7.2 LVQ2

Dle Tučková (Tučková, 2009) rozdělení do tříd je stejné jako u LVQ1, avšak při učení existují kódové knihy $w_{ij}{}^1$ a $w_{ij}{}^2$, které reprezentují dvě třídy T_1 a T_2 nacházející se ve vektorovém prostoru blízko sebe. Vektor X se musí dobře klasifikovat do správné třídy, ale také je třeba, aby patřil do oblasti hodnot stanovených okénkem

$$min\left\{\frac{d_i, d_j}{d_j, d_i}\right\} > s$$

$$s = \frac{1 - win}{1 + win}$$

d_i,d_j ...Euklidovská vzdálenost X od $w_{ij}{}^1$ a $w_{ij}{}^2$

win...relativní šířka okénka v rozmezí 0.2 < win < 0.3.

7.3 LVQ3

Dle Šnorek(Šnorek, 2004) má metoda LVQ2 nevýhody:

a) Postupnou adaptaci obou vektorů se snižuje vzdálenost, místo aby se zvyšovala.

b) Postupně se dá dostat z nalezené optimální polohy do jiného rovnovážného stavu, který není optimální.

Počáteční nastavení váhových vektorů se provede stejně jako u metody LVQ1 Pokud třídy T_1 a T_2 jsou dvě nejbližší třídy a vzor navíc patří do třídy T_2 a ne do třídy T_1, potom se mohou provést tyto úpravy:

$$W_{ij}{}^1(t+1)=w_{ij}{}^1(t) - \ (t)[x_i(t) - w_{ij}{}^1(t)]$$

, kdy se odsune vector od chybné třídy T_1

$$W_{ij}{}^2(t+1)=w_{ij}{}^2(t) - (t)[x_i(t) - w_{ij}{}^2(t)]$$

, kdy se přisune vector ke správné třídě T_2 Tato metoda LVQ3 je stabilní (po nalezení optimální pozice váhového vektoru se už umístění váhového vektoru nebude měnit)

8 Vývojové prostředí Matlab

Nejprve se seznámíme s programem Matlab a také s jeho pracovním prostředím. Výhody sw a dále bude představen nástroj neuronových sítí.
MATLAB je programovací jazyk a také komplexní interaktivní prostředí pro vývoj algoritmů, vizualizaci dat a všeobecné numerické výpočty. Lze zde pracovat se signály a systémy, vytvářet aplikace s podporou grafického rozhraní, provádět reálná měření atd.
Software MATLAB je produktem americké společnosti The MathWorks.Inc.
Matlab obsahuje velké množství vestavěných funkcí, proto je řešení technických problémů v tomto prostředí mnohem jednodušší než v jiných programovacích jazycích.(Karban, 2006)
Mezi důležité vlastnosti Matlabu patří vysokoúrovňový jazyk pro technické výpočty. Obsahuje velké množství aplikačních knihoven a má otevřený a rozšiřitelný systém. Podporuje vícerozměrná pole a datové struktury. Zahrnuje interaktivní nástroj pro tvorbu grafického uživatelského rozhraní (GUI). Je možné importovat a exportovat data do mnoha formátů. Lze rozšířit moduly jazyky C, C++, Java či Fortran.(Karban, 2006)

8.1 Pracovní prostředí

Pracovní prostředí je složeno z několika oken. Nejdůležitější okno má název Command Window. Zapisují se zde příkazy a sleduje se odezva systému. Dalším oknem je workspace, který zobrazuje vytvořené proměnné a také jejich obsah, kde lze provádět různé činnosti jako je tvorba grafu, kopirování atd. Okno s názvem Command History zobrazuje všechny potvrzené příkazy. Jednou použitý příkaz lze znovu použít pomocí klávesnice (nahoru, dolů).(Zaplatílek, 2011)
Menu Current Folder slouží k výběru složky (adresáře), kde je uložena vytvořená práce. V souborech s příponou *.m (m-soubory) jsou uloženy zdrojové texty algoritmů.(Zaplatílek, 2011)

8.2 M-soubory

obsahují seznam použitých příkazů. Za pomoci příkazu edit se otevře editor. Po zápisu příkazů se soubor uloží a kdykoliv se pak může editovat či spustit. Editor poskytuje mnoho služeb, např. upozornění na chyby, hlídání párovosti závorek atd. Dále je možnost při práci používat několik tlačítek jako je například symbol červené tečky. Tento symbol je zarážka zdrojového textu. Spustí-li se kód, algoritmus se začne provádět a zarazí se právě na této zarážce(breakpoint). Pokud je zdrojový text delší je možnost využít záložku(bookmark), pak pomocí klávesy F2 lze rychle mezi záložkami přepínat v textu a tím ušetřit také čas.(Zaplatílek, 2011)

8.3 Toolbox neuronové sítě

Neural Network Toolbox ™ poskytuje funkce a aplikace pro modelování složitých nelineárních systémů, které nejdou snadno modelovat. Neural Network Toolbox podporuje hlídané učení s učitelem, ale také učení bez učitele jako samo-organizující mapu a kompetitivní vrstvu. Pomocí nástroje lze navrhnout, trénovat, vizualizovat a simulovat neuronové sítě. Nástroj pro tvorbu neuronových sítí je možno také použit pro aplikace, jako je rozpoznání, shlukování, časové řady predikce a dynamické modelování systémů a řízení.
Pro urychlení přípravy a zpracování velkých objemů dat lze distribuovat výpočty a data napříč vícejádrových procesorů, GPU, a počítačových clusterů pomocí paralelních výpočtů Toolbox ™. (MathWorks, 2012).

9 Vývojové prostředí Weka

Je open source software vydaný pod licencí GNU General Public License. "WEKA" je zkratkou pro životní prostředí Waikato (Weka je nelétavý pták zvídavé povahy žijící na ostrovech Nového Zélandu)pro znalostní analýzy, který byl vyvinut na Univerzitě Waikato na Novém Zélandu. Weka obsahuje algoritmy pro data předzpracování, klasifikace, regrese, shlukování, asociační pravidla, a vizualizace. Je vhodný pro vývoj nových systému strojového učení. Weka je využívána pro výzkum, vzdělávání a aplikací. Je to rozšiřitelný software a stal se sbírkou strojového učení algoritmů pro řešení problémů reálného světa dolování dat. Je napsán v Jave a běží na téměř každé platformě.(Mihaescu,2011)
Podporované formáty pro weku jsou dle(Scuse, 2012)

- ARFF (Attribute-Relation File Format), která se skládá z hlavičky a datové části.
  - hlavička obsahuje jméno relace
  - v datové části jsou data k příslušným atributům
- XRFF (eXtensible attribute-Relation File Format), reprezentuje data ve formátu, který může uložit komentáře, atributy a instance vah.
- Databáze (MS Acces, JDBC)
- CSV, URL, BSI, C4.5 a binární soubory

Rozhraní se skládá ze čtyř tlačítek, které jsou hlavní aplikace ve Wece dle (Mihaescu,2011):

a) explorer - Okno Explorer zobrazuje pro zvolený atribut rozložení jeho hodnot (zde formou histogramu). Lze vybrat nástroj pro předzpracování dat, filtraci, editaci apod. Dále lze vybrat klasifikaci, shlukování, aj. (Hřebíček, Žižka 2007)

b) Simple CLI – je jednoduché rozhraní příkazového řádku, který umožňuje přímé provádění příkazů pro operační systémy, které neposkytují vlastní rozhraní příkazového řádku

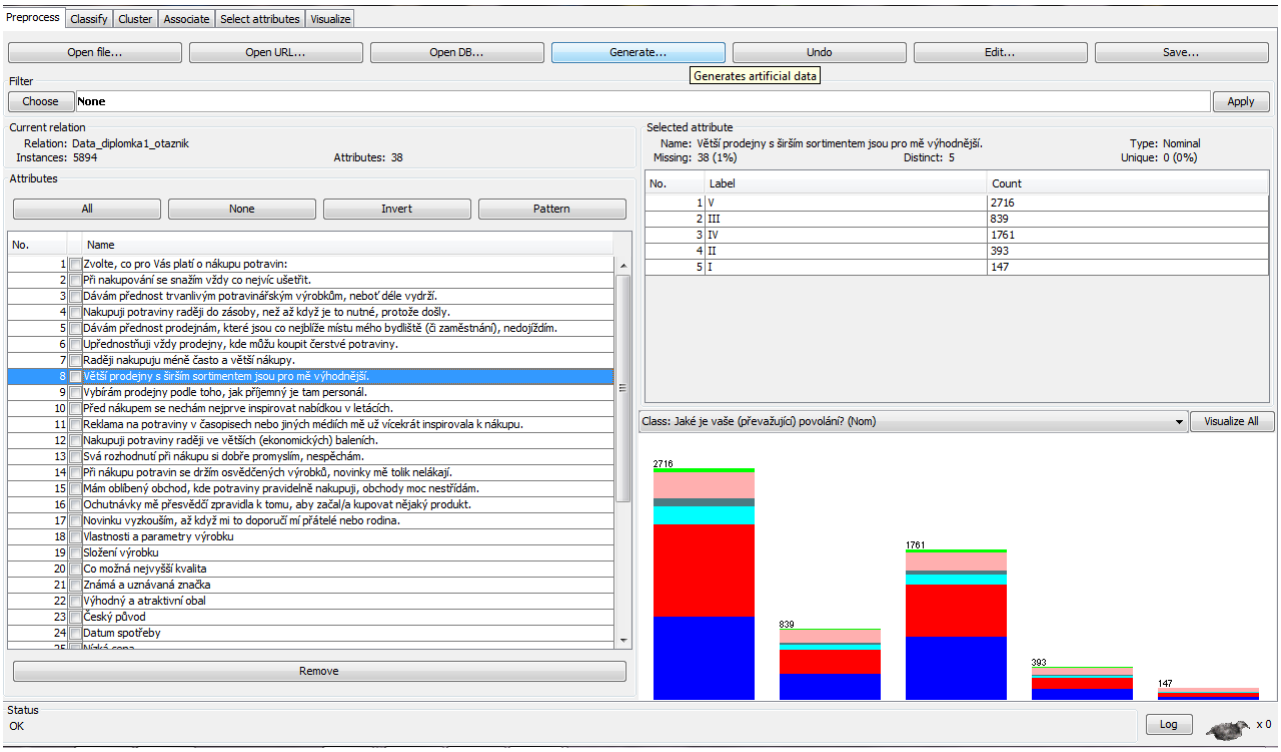
c) Experimenter – je to prostředí pro objevování dat

d) KnowledgeFlow – toto prostředí podporuje stejné funkce jako Explorer, ale drag-and-drop rozhraním.

Okno Explorer

Pomocí Open File se vybere dataset ve formátu arff nebo csv. Pro každý atribut lze vidět rozložení hodnot dle histogramu. Data se dají předzpracovat (filtrace,

editace, konverze a jiné). Nyní si lze zvolit činnost jako je například klasifikace, shlukování a jiné. Data lze také vizualizovat, podrobněji níže.(Hřebíček, Žižka 2007)



Obrázek 6: okno Explorer

Způsoby vizualizace dat dle manuálu(Scuse, 2012):

- Plot : pro vykreslení 2D plot datasetu
- ROC: uloží dříve uložené ROC křivky
- Vizualizace stromu : orientovaný graf, např. Rozhodovací strom
- Vizualizace grafu: zobrazuje XML BIF nebo DOT formát grafu, např. pro Bayesovské sítě

9.1 Rozhodovací strom J48

Rozhodovací stromy mají za úlohu klasifikovat příklady do dvou či více tříd(Kubik, 2004). Patří k účinným algoritmům, ale mají svá principiální omezení. Výchozí množina tréninkových příkladů se dělí na podmnožiny. V ideálním případě obsahují pouze členy jediné třídy a entropie je nulová. (Hřebíček, Žižka

2007) Je nutné hledat takové dotazy na určité atributy, které vedou ke snížení entropie vzhledem k předchozímu vztahu. Je hledán atribut, který nejlépe rozdělí prvotní příklady do podmnožin. Poté se dělí dále podmnožiny až do možnosti, dokud nelze získat lepšího výsledku. Může vznikat přeučení a generalizace, potom podmnožiny obsahují pouze jeden prvek. Bylo by dosaženo dokonalých výsledků na trénovacích datech, ale špatně na testovacích. (Hřebíček, Žižka 2007)

Struktura rozhodovacího stromu

(Kubik, 2004)

a) listy - pojmenované dle tříd, do kterých entity začleňujeme

b) uzly - zvolené atributy

c) hrany - hodnoty atributů

Algoritmus rozhodovací strom J48 používá program Weka. J48 je verze jednoduššího algoritmu C4.5 vytvořený J. Ross Quinlanem. Strom je používán ke klasifikaci intervalů, kde chybí hodnoty spojitých atributů.(Pejčoch, 2011)
Rozhodovací strom má ve wece implicitně nastaveny parametry J48 -C 0.25 -M 2,

- C je konfidenční faktor, který ovlivňuje prořezávání úplného stromu v míře 0,25. Když je C rovno nule, potom je maximální prořezání na základě minimální důvěry v informaci, který poskytuje strom. C roven jedné nese naopak důvěru maximální, minimalizující prořezání
- M – kolik instancí může být minimálně v listu stromu. Defaultně jsou nastavené dva trénovací příklady (umožňuje vyhnout se přeučení stromu).

(Hřebíček, Žižka 2007)

10 Popis dat

Data byla shromážděna pomocí dotazníku od různých osob. Tento výzkum zahrnoval všechny kraje České republiky. Výzkumu se zúčastnilo 2335 respondentů. Tito lidé byli v různé věkové kategorii, s rozdílným vzděláním či povoláním. Tyto dotazníky obsahují 29 otázek. Sedm charakterizuje respondenty, a dalších 22 je zaměřeno na chování spotřebitele na trhu s potravinami jako je např. den, jak často chodí lidé nakupovat a co jim vadí nebo podle čeho si vybírají prodejnu s potravinami, a jakou prodejnu mají mezi oblíbenými atd. Celkově bylo 105 atributů, protože některé otázky obsahovaly podotázky. Jinak řečeno obsahovaly možnou odpověď, na kterou se odpovídalo např. ano/ne. Příkladem může být třeba otázka: V jaké dny chodíte nejčastěji nakupovat. Tato otázka obsahuje sloupce (pondělí, úterý, středa,...) a na tyto dny lze odpovědět ano/ne, jak bylo zmíněno výše. Jindy byla potřeba označit odpovědi na stupnici od 1 do 10 podle důležitosti při rozhodování o zakoupení konkrétního produktu. Stupnice 10 určuje nejvyšší důležitost.
V datech byla použita kombinace binárních, nominálních a ordinálních hodnot. Např. u atributů pohlaví (žena/muž), otázky s odpověďmi (ano/ne) byla použita binární data. Bylo použito také velké množství nominálních dat např. u atributu vzdělání (SS s maturitou/Ss bez maturity/Vš/Zš/vyšší odborná) a ordinální hodnoty jako je výše uvedená stupnice důležitosti.
I po různých úpravách nám data dávala velmi vysoké chyby, které se nedaly použít. Tomuto se přisuzuje, že na mnoho otázek respondent odpovídal podle sebe, tedy bylo mnoho různých odpovědí na otázku. Dále bylo mnoho atributů, které neměli mezi sebou velkou souvislost.
Po různém promazání bylo použito 73 atributů, které byly různě zkombinované. Provádět klasifikaci nad tímto počtem atributů nebylo přínosné, protože výsledky byly velmi špatné. Trénovací a testovací data byla rozdělena pomocí krosvalidace.
Po neúspěchu těchto dat bylo použito jiných s podobným tvarem odpovědí. Tento výzkum se také zabýval stejným tématem, jak u prvního dotazníku. Velký rozdíl mezi předchozím a tímto dotazníkem byl v tom, že zde odpovídalo 5894 respondentů a otázek bylo poměrně méně. Celkový počet atributu byl 29. Jelikož respondentů bylo více a otázek méně, měly zde výsledky větší úspěch.

11 Weka

Shromážděná data se klasifikovala v programu Weka. Nejprve se data musela upravit.

11.1 Úprava dat

Dataset byl připraven ve formátu xls v nominální a číselné podobě pro použití datasetu v softwaru weka 3.6[1], který byl zvolen pro klasifikaci dat. Tyto data byla potřeba přeuložit ze souboru xls do csv, protože program weka pracuje se soubory csv. Dále se musela data formátovat a pročistit, protože mnoho atributů obsahovalo velkou míru nezodpovězených otázek. Tyto atributy musely být odstraněny, protože prázdné buňky se přisuzovaly ke špatným výsledkům s vysokým procentem chyb. Jelikož někteří respondenti odpovídali na otázky zřídka, bylo nutné promazat i právě některé instance.
V programu weka byly vybrány dva klasifikační algoritmy. Mezi těmito algoritmy jsou strom J48 a MLP. Rozhodovací strom J48 umí klasifikovat jen s kvalitativními daty. Některá data, která obsahují číselné hodnoty 1 až 10 označující stupnici důležitosti, musela být změněna právě na nominální hodnoty a tím se stupnice zkrátila a přeměnila do této podoby:

- 1,2 = I
- 3,4 = II
- 5,6 = III
- 7,8 = IV
- 9,10 = V

3. Algoritmy použité pro klasifikaci:
   - Rozhodovací pravidlo zeroR
   - Rozhodovací pravidlo PART
   - Rozhodovací strom J48
   - Neuronová síť MultilayerPerceptron

   Metoda zeroR je nejjednodušší metoda klasifikace. Identifikuje nejčastější hodnoty třídy v datasetu. Tato metoda je pro nás základní. Správně zařazené instance jsou informativní a další algoritmy by neměly poskytnout horší výsledky. U rozhodovacího stromu J48 byl nastaven parametr konfidenční faktor C, který ovlivňuje prořezávání úplného stromu, jehož rozmezí je od 0.0 do 1.0. C = 0.0

[1] software weka 3.6.0 lze stahnout na: http://www.cs.waikato.ac.nz/ml/weka/

je maximální prořezání a C = 1.0 minimalizuje prořezání stromu. Dalším parametrem je počet prvků v listu, kde bylo nastavováno i 60 prvků. Nastavení záviselo na počtu atributů. Při zanechání všech 29 atributů se dalo nastavit až 60 a více prvků. V tomto případě čím více prvků tím lepší výsledky. Samozřejmě že, pokud četnost atributů byla už velmi vysoká (100 a více) výsledky se zhoršovaly. Pokud byl počet atributů nižší např. 7 a méně, bylo nastaveno nejvíce třech prvků v listu. Při zvýšení se hodnoty velmi zhoršily.
Dále bylo použito rozhodovací pravidlo PART, které obsahuje parametry jako je prořezání stromu a kolik bude prvků v listu. Jsou to stejné parametry, které se udávají u stromu J48. U této metody byly vyhodnoceny zajímavé závislosti mezi atributy. Byly zde klasifikované různé třídy.
Neuronová síť MultilayerPerceptron byla pro vytvoření modelu velmi pomalá. Proto byla použita klasifikace jen pro třídy, které měly vyšší úspěšnost v metodě j48. Dále po použití filtrace, která je zmíněna níže, kde bylo vyfiltrováno 6 atributů, byl čas zpracování 238.9 seconds.

11.2 Použití všech atributů

Tabulka 1: Tabulka hodnot

Atributy	**zeroR**	**j48**	**MLP**
Věk	31,73%	74,14%	67,75 %
V jaké domácnosti žijete?	35,63%	61,16%	53,48%
Před nákupem se nechám nejprve inspirovat nabídkou v letácích	49,15%	55,92%	49,46%
Jak hodnotíte příjem vaší domácnosti z pohledu pokrytí potřeb a kvality života?	49,24%	51,24%	49,40 %
Nakupuji potraviny raději do zásoby, než když je to nutné, protože došly.	25,25%	34,78%	31,25%
Nízká cena	34,41%	54,76%	49,85%
Výrobek je v akční nabídce	35,83%	55,87%	46,32%
Dávám přednost prodejnám, které jsou co nejblíže místu mého bydliště(či zaměstnání), nedojíždím	29,91%	37,21%	34,18%
Reklama na potraviny v časopisech nebo jiných médiích mě už vícekrát inspirovala k nákupu.	22,49%	47,57%	37,81
Dávám přednost trvanlivým potravinářským výrobkům, neboť déle vydrží.	29,91%	37,21%	32,42%

Atributy	**zeroR**	**j48**	**MLP**
V jakém kraji bydlíte?	40,41%	40,65%	30,32%
Český původ	26,55%	33,66%	31,8%
Pohlaví	59%	63,40%	63%
Co možná nejvyšší kvalita	40,45%	55%	48,91%
Známá a uznávaná značka	35,46%	41,67%	35,49%
Jak velká je obec, ve které bydlíte?	40,5%	47,5%	29,74%
Novinku vyzkouším, až když mi to doporučí mí přátelé nebo rodina	28,54%	35,53%	29,74
Ochutnávky mě přesvědčí zpravidla k tomu, aby začal/a kupovat nějaký produkt.	34,93%	39,67%	36,51%
Energetická hodnota	27,54%	45,88%	38,82%
Složení výrobku	33,42%	49,12%	42%
Výhodný a atraktivní obal	29,59%	36,97%	33,13%
Doporučení odborníků	29,57%	47,47%	39,45%
Patří mezi výrobky zdravé výživy	25,93%	43,32%	39,53%
Datum spotřeby	50,58%	56,24%	51,12%
Upřednostňuji vždy prodejny, kde můžu koupit čerstvé potraviny	48,34%	50,56%	48,5%
Vlastnosti a parametry výrobku	38,30%	48,47%	40,64%
Doporučení známých	31,96%	47,05%	39,68%
Jaké je vaše (převažující) povolání?	37,55%	71,31%	49,79%

V tabulce lze vidět, že metoda zeroR měla horší výsledky ve všech případech. U metody j48 bylo nastaveno prořezání v hodnotě 0,5 a 50 prvků v listu. Při nižším zvolení počtů prvků byly výsledky horší. U tříd, které charakterizují respondenty, byly výsledky pozitivní, avšak nejsou pro nás moc zajímavé. Proto jsme se více zaměřily na atributy, které nám sdělí něco zajímavého. Třídy, jejichž správné zařazení se pohybuje okolo 55% byly použity pro metodu PART. Těmito atributy jsou:

- Nízká cena
- Datum spotřeby
- Co možná nejvyšší kvalita
- Výrobek je v akční nabídce

Lze si také povšimnout, že u metody J48 bylo dosažených lepších výsledků ve všech případech než u metody MLP. Nejlépe klasifikovaná byla třída věk se 74% a nejnižší naopak český původ výrobku 33,66%. Mnoho výsledků se pohybovalo při nižší hranici úspěšnosti, toto zařazení nemá pro nás žádný přínos.

11.3 Závislosti mezi atributy:

U výše čtyř zmíněných atributů bylo použito pravidlo PART, který vyhodnotil zajímavé závislosti a výsledky.

Nízká cena

U atributu nízká cena bylo nastaveno 35 prvků v listu a průřez stromu 0,7. Výsledek správně zařazených prvků byl 55,39

Výrobek je v akční nabídce = V AND

Dávám přednost trvanlivým potravinářským výrobkům, neboť déle vydrží. = V AND

Jak hodnotíte příjem vaší domácnosti z pohledu pokrytí potřeb a kvality života? = A: V (110.0/6.0)

Zde bylo správně klasifikováno 110 instancí a 6 špatně. Tento výsledek je tedy pro nás přínosný. Tímto lze říci, že lidé, kteří preferují nízkou cenu, nakupují potraviny v akční nabídce, ale také dávají přednost trvanlivým potravinářským výrobkům, jejichž příjem domácnosti je nízký (základní potřeby domácnost pokryje, ale musí v nich šetřit a eventuelně se omezovat)

Výrobek je v akční nabídce = V AND

Dávám přednost trvanlivým potravinářským výrobkům, neboť déle vydrží. = V AND

Jaké je vaše (převažující) povolání? = důchodce AND

V jakém kraji bydlíte? = Jihomoravský kraj : V (57.0/5.0)

Další závislosti atributů poukazují na to, že výrobky s nízkou cenou nakupují důchodci, kteří bydlí v Jihomoravském kraji. Výrobky opět nakupují v akční nabídce a dávají přednost výrobkům, které déle vydrží. Správně bylo klasifikováno 57 instancí a 5 špatně.

Datum spotřeby

Zde bylo nastaveno 10 prvků v listu a průřez stromu 0,7. Kde byly správně klasifikované objekty v 53,19 procentuální úspěšnosti. Se zvyšujícím počtem prvků v listu byl i lepší klasifikační výsledek. Pro 30 prvků byla úspěšnost 56%. Pro 60 prvků 57,63%. Závislost atributů pro toto nastavení si uvedeme zde:

Upřednostňuji vždy prodejny, kde můžu koupit čerstvé potraviny. = V AND

Pohlaví = žena AND

Český původ = V AND

Co možná nejvyšší kvalita = V AND

V jakém kraji bydlíte? = Kraj Vysočina : V (65.0/4.0)

Tato závislost ukazuje, že pro ženy z kraje Vysočina je velmi důležité kupovat čerstvé potraviny českého původu a v nejvyšší kvalitě. Správně bylo klasifikováno 65 objektů a špatně 4.

Upřednostňuji vždy prodejny, kde můžu koupit čerstvé potraviny. = V AND

Pohlaví = žena AND

Český původ = V: V (663.0/93.0)

Zde ženy upřednostňuji prodejny s čerstvými potravinami českého původu. Správně bylo zařazeno 663 objektů a špatně 93.

Co možná nejvyšší kvalita

Nastaveno bylo 10 prvků a průřez stromu 0,5 s úspěšnosti 52,65%. Po zvolení vyššího počtu prvků, který činil 50, byla úspěšnost vyšší o 3%

Složení výrobku = V AND

Upřednostňuji vždy prodejny, kde můžu koupit čerstvé potraviny. = V AND

Vlastnosti a parametry výrobku = V AND

Známá a uznávaná značka = V: V (203.0/18.0)

Pro třídu co možná nejvyšší kvalita bylo klasifikováno, že lidé se zaměřují na složení výrobku, vlastnosti a parametry výrobku, ale také preferují uznávanou značku a upřednostňují vždy prodejny, kde můžou koupit čerstvé potraviny. Klasifikováno bylo správně 203 objektů a 18 špatně.

Složení výrobku = V AND

Upřednostňuji vždy prodejny, kde můžu koupit čerstvé potraviny. = V: V (1117.0/250.0)

Zde byla závislost jen mezi atributy složení výrobku a prodejny s čerstvými potraviny. Tedy v tomto případě jsou pro 1117 lidí důležité tyto 2 atributy.

Složení výrobku = V AND

Upřednostňuji vždy prodejny, kde můžu koupit čerstvé potraviny. = V AND

Patří mezi výrobky zdravé výživy = V AND

Známá a uznávaná značka = III: V (95.0/13.0)

Další klasifikace této třídy vytvořila závislost mezi těmito atributy, která poukazuje na to, že složení výrobků, výrobky zdravé výživy, známá značka a prodejny s čerstvými potraviny jsou důležité pro zvolení výrobku s co nejvyšší kvalitou.

Složení výrobku = V AND

Upřednostňuji vždy prodejny, kde můžu koupit čerstvé potraviny. = V AND

Vlastnosti a parametry výrobku = V AND

Český původ = V: V (193.0/43.0)

Lidé, pro které je důležitá kvalita výrobků preferují u zboží jeho složení, parametry a také český původ a nakupují tam, kde se prodávají čerstvé potraviny.

Výrobek je v akční nabídce

nejlepších výsledků bylo dosaženo 57.676 %

Nízká cena = V AND

Před nákupem se nechám nejprve inspirovat nabídkou v letácích. = V: V (728.0/113.0)

Respondenti nakupující v akční nabídce upřednostňující nízkou cenu a před nákupem se nechají inspirovat v letácích. Zde bylo správně zařazeno 728 objektů a špatně 113.

11.4 Použití atributů po filtraci

V softwaru weka byl otevřen soubor se všemi daty, kde se pro výběr atributů použila metoda hodnocení příznaků ChiSquaredAttributeEval. Tato metoda s ohledem na vstupní hodnotu volí příznaky na základě výpočtu chí statistiky. A metoda vyhledávání Ranker seřadí příznaky dle individuálních ohodnocení. Mohou být vybrány pouze hodnotitelem. Po seřazení atributů bylo vybráno prvních 6 atributů.

Po použití výše uvedených metod byly vybrány tyto atributy:

- Věk
- Před nákupem se nechám nejprve inspirovat nabídkou v letácích
- Jak hodnotíte příjem vaší domácnosti z pohledu pokrytí potřeb a kvality života
- Nakupuji potraviny raději do zásoby, než až když je to nutné, protože došly
- Nízká cena
- Výrobek je v akční nabídce

V příloze B (obrázek 23) lze vidět rozložení hodnot všech atributů. Na ukázku zde byla zobrazena třída ”výrobek je v akční nabídce”, kde byly rozloženy vstupní vzorky dle toho, do jaké třídy patří.

U těchto atributů byly použity metody MLP, rozhodovací pravidlo PART a rozhodovací strom j48. U metody MLP byla použita dvoudílná crossvalidace a u metody Part a j48 byla použita desetidílná a dvacetidílná crossvalidace. Byl zvolen rozhodovací strom J48 z důvodu největší úspěšnosti klasifikace mezi ostatními rozhodovacími stromy. Při použití stromu ID3 byla úspěšnost například u nízké ceny nižší o více než 10%. Z tohoto důvodu byla použita metoda J48.

Výsledky neuronové sítě MLP a rozhodovacího stromu jsou velmi podobné. U třídy nízká cena a výrobek je v akční nabídce bylo klasifikováno přes 56% v obou případech. Z tabulky lze vidět, že u třídy ”Nakupuji potraviny raději do

Tabulka 2: Výsledky MLP a J48

Třída	**MLP**[%]	**J48**[%]
Nízká cena	56,40	56,70
Výrobek je v akční nabídce	56,42	56,50
Před nákupem si nechám nejprve inspirovat nabídkou v letácích	41,07	36,87
Nakupuji potraviny raději do zásoby, než když je to nutné, protože došly	33,72	35
Jak hodnotíte příjem vaší domácnosti z pohledu pokrytí potřeb a kvality života	46,06	51,80
Věk respondenta	32,37	36

zásoby,.."byla u algoritmu J48 dosaženo výsledků s 51,71% úspěšnosti, je tedy o více než 5% vyšší klasifikace. Naopak u třídy "Před nákupem si nechám nejprve inspirovat nabídkou v letácích"je úspěšnější metoda MLP, která má výsledek 41,07%. Celkově tyto výsledky nejsou moc pozitivní. Nejvíce se tedy zaměříme na na první dvě třídy, které mají úspěšnost 56%.

Z vykresleného stromu J48 viz. příloha B(obrázek 24), kde byla klasifikovaná třída "výrobek je v akční nabídce" lze shlédnout, že tyto lidé, kteří nakupují výrobky v akční nabídce, preferují nízkou cenu. Správně bylo zařazeno 2143 respondentů, kteří nízké ceně dávají velkou přednost. Špatně bylo klasifikováno 640 instancí.

Osoby ve věku 18 až 24 let (správně klasifikovány 4 vzorky, špatně 1)a lidé ve věku 35-54 let (správně klasifikováno 14 instancí, špatně 7), pro které je středně důležitá nízká cena, se před nákupem inspirují nabídkou v letácích a preferují nákup potravin raději do zásoby.

U algoritmu PART byly nastaveny tyto parametry: 2 prvky v listu, jak bylo zmíněno výše, při zvyšování počtu prvku úspěšnost klasifikace klesala a průřez stromu 0,5. Metoda vyhodnotila tyto závislosti s 60% úspěšnosti a relativní hodnota absolutní chyby byla 66%.

Výrobek je v akční nabídce = V AND

Jak hodnotíte příjem vaší domácnosti z pohledu pokrytí potřeb a kvality života? = A AND

Před nákupem se nechám nejprve inspirovat nabídkou v letácích. = V: V (88.0/5.0)

Lidé, kteří nakupují levné výrobky, hodnotí příjem domácnosti nízký (základní potřeby domácnost pokryje, ale musí v nich šetřit eventuelně se omezovat) a zaměřují se na výrobky v akční nabídce a také se právě proto nechají inspirovat nabídkou v letácích. Klasifikování bylo správně 88 instancí a špatně 5. Výrobek je v akční nabídce

U této třídy byla úspěšnost metody j48 v 56,50% a pravidlo Part mělo úspěšnost 61%, kde byly provedeny tyto závislosti:

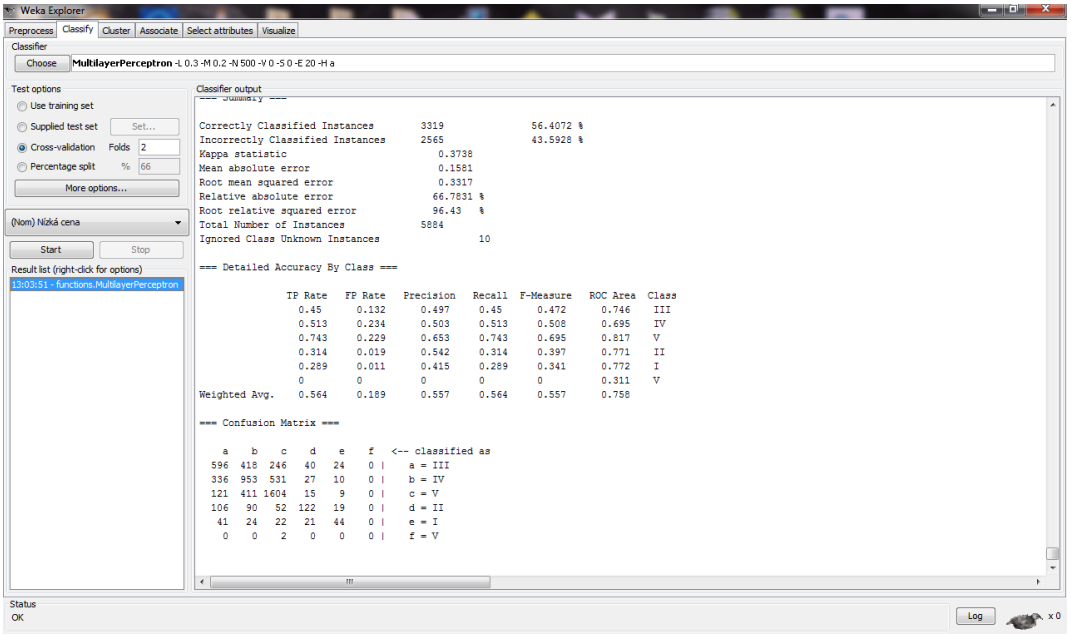
Nízká cena = V AND

Před nákupem se nechám nejprve inspirovat nabídkou v letácích. = V AND

Nakupuji potraviny raději do zásoby, než až když je to nutné, protože došly. = IV AND

Z tohoto poznatku vyplívá, že lidé kupující levné výrobky se nechají inspirovat letáky a také je pro ně důležité nakupovat potraviny do zásoby. Správně bylo klasifikováno 50 objektů a nesprávně 7.

Výsledky metody MLP



Obrázek 7: vystup: nizka cena

Lze vidět, že nejvíce bylo opět správně zařazeno do třídy V (nejvyšší ve stupnici důležitosti)

12 Matlab

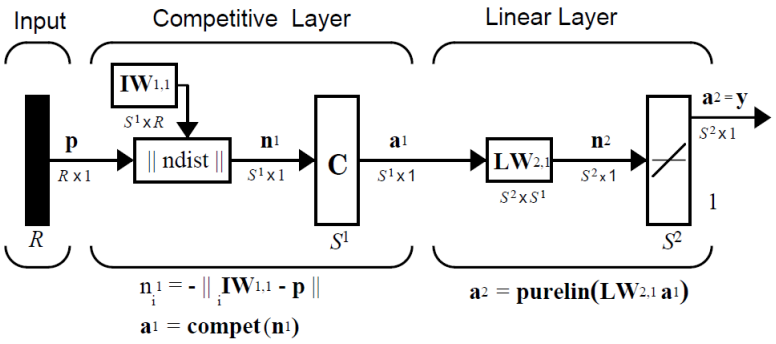
Pro shlukování byl použit software Matlab.

12.1 Úprava dat

Data musela být opět upravena. Program Matlab pracuje se soubory xls, proto se nemuselo nic přeukládat. Avšak neumí pracovat s nominálními hodnotami, proto se musela data vybraných atributů změnit z nominální hodnoty na číselné. K těmto datům je poskytnuta legenda, kterou lze vidět níže na obrázku. Hodnoty jsou od 1 do 5. Jednička znázorňuje nejmenší souhlas a pětka naopak největší. Dále byl vybrán atribut věk, který byl rozřazen do 6 skupin. A nakonec také prázdné buňky byly označeny nulou.

12.2 Učící vektorová kvantizace

Učící vektorová kvantizace dále jen LVQ byla vytvořena v prostředí Matlab, jehož verze je Rb2010b. Tato verze má k dispozici neural network toolbox. Zde je mnoho funkcí pro SOM a LVQ. Na následujícím obrázku lze vidět architektura.



Obrázek 8: architektura - síť učící vektorové kvantizace(MathWorks, 2012)

LVQ se skládá ze dvou vrstev. První vrstva je kompetitivní a druhá lineární. Vrstva kompetitivní se učí klasifikovat vstupní vektory stejným způsobem jako konkurenční vrstva samoorganizujících map. Lineární vrstva transformuje třídy kompetitivní vrstvy do cílové (target) klasifikace definované uživatelem. Třídy učené kompetitivní vrstvou jsou označovány jako podtřídy a třídy lineární vrstvy jako cílové třídy. Obě vrstvy (komp., lineární) mají jeden neuron na (sub nebo cíl)třídy. Kompetitivní vrstva tak může naučit až S1 podtříd. Lineární vrstva S2 cílových tříd. S1 je vždy větší než S2.

Nejdříve si popíšeme Self organizing feature map, dále SOFM

SOFM

SOFM učí klasifikovat vstupní vektory podle toho, jak jsou seskupeny ve vstupním prostoru. Od kompetitivních vrstev se liší v tom, že okolní neurony v self-organizační mapě se naučí rozpoznat sousední částí vstupního prostoru.

Data byla normalizovaná od 0 do 1.

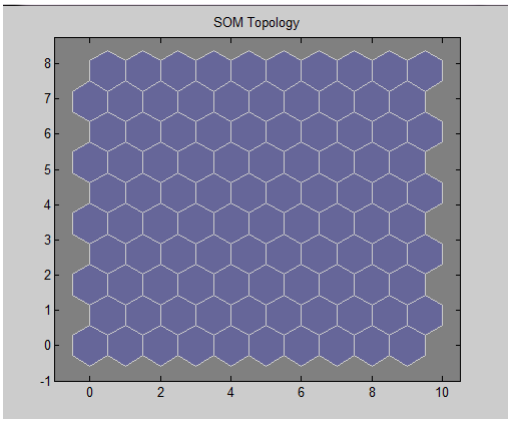
Tvorba sítě selforgmap

Po vytvoření skriptu a normalizaci dat bylo potřeba nastavit síť. K tomuto účelu byla použita funkce selforgmap. SOM učí shluky dat na základě podobnosti a topologie.

```
selforgmap(dimensions,coverSteps,initNeighbor,topologyFcn,distanceFcn)
```

Je potřeba zvolit topologii a dimenzi, do které budou uspořádané neurony. Existují čtyři typy funkcí topologie map, které vypočítávají pozici neuronu ve vrstvě:

- Hextop: je hexagonální vrstva topologie. Tato topologie byla zvolena, protože byla osvědčena jako optimální. Množina vstupních vektorů v ní byla nejlépe uspořádána
- Gridtop: neboli čtvercová topologie, uspořádání neuronů zde nebylo optimální.
- Randtop: neurony jsou uspořádány v N – rozměrným náhodným vzorem
- Tritop: neurony jsou uspořádány v N – rozměrné trojúhelníkové síti



Obrázek 9: topologie som

Zvolena byla 2-dimenzionální síť:

```
dimension1 =9;
dimension2 =9;
dimensions= [dimension1  dimension2]
```

Dále byla potřeba také nastavit coverSteps, což značí počet trénovacích kroků pro počáteční pokrytí vstupního prostoru. Defaultně je nastavena hodnota 100. Po různém nastavení, byla nakonec použita hodnota 130. Parametr initNeighbor je počáteční vzdálenost mezi sousedy. Defaultně je nastavena hodnota 3. Tato hodnota zůstala nezměněna.

Posledním parametrem funkce selforgmap je distanceFcn(funkce vzdálenosti). Vstupní vektory jsou porovnávány prostřednictvím zvolených metrik s každým referenčním vektorem. Jinak řečeno tyto metriky vypočítají vzdálenost od určitého neuronu k jeho sousedům.

- Typy metrik:
- Dist : euklidova váhová funkce.
- Linkdist: vzdálenost pozic mezi neurony. Vzdálenost od jednoho neuronu je jen počet vazeb či kroků.
- Mandist : vzdálenost mezi dvěma vektory x a y se vypočítá jako

$$\text{D = sum(abs(x-y))}$$ (MathWorks, 2012)

Pro výpočet vzdálenosti byla vybrána euklidovská vzdálenost. Tudíž byla použita metrika dist, která byla aplikovaná společně s hextop.

$$\text{D = sum((x-y).2).0.5}$$

(MathWorks, 2012)

Pro inicializaci vstupních vah byla použita funkce midpoint, která inicializuje váhy vektorů na střední hodnotu intervalu.

```
net.inputWeights{1,1}.initFcn = 'midpoint'
net.layers{1}.transferFcn = 'compet'
net = configure (net,inputt)
```

Při simulaci sítě, jsou negativní vzdálenosti mezi každým neuronem váhového vektoru a vstupního vektoru vypočteny pomocí negdist. Kompetitivní přenosová funkce produkuje 1 pro všechny výstupní prvky odpovídající vítězství neuronu. Všechny ostatní výstupní prvky jsou 0.

```
net.inputWeights{1,1}.weightFcn = 'negdist'
```

Učení (learnsomb)

Jako učící algoritmus byla zvolena funkce learnsomb, která je přímo určená pro SOM.

```
net.inputWeights{1,1}.learnFcn ='learnsomb'
```

Nejdříve síť identifikuje vítězný neuron pro každý vstupní vektor. Každá váha vektoru se pak přesune na průměrné polohy všech vstupních vektorů, pro které je vítěz nebo je v sousedství vítěze.

Dále bylo potřeba nastavit počet iterací (epoch) pro provedení trénování. Po různém zkoušení počtů epoch bylo zjištěno, že lepší je větší počet epoch. Nastaveno bylo 1500 epoch. Trénování proběhlo pomocí funkce trainbu, která spolupracuje s learnsomb, obě tyto funkce jsou dávkové (batch). Byl zvolen dávkový režim, protože má znatelně vyšší rychlost učení.

```
net.trainParam.epoch=1500;

net.trainFcn = 'trainbu'

[net,tr] = train(net,input);
```

Testování je provedeno pomocí funkce sim. Funkce vec2ind(a) konvertuje vstupní vektory do tříd. Tato funkce umožňuje, aby indexy byly zastoupeny budˇ samostatně nebo jako vektory obsahující jedničku v řádku indexu.

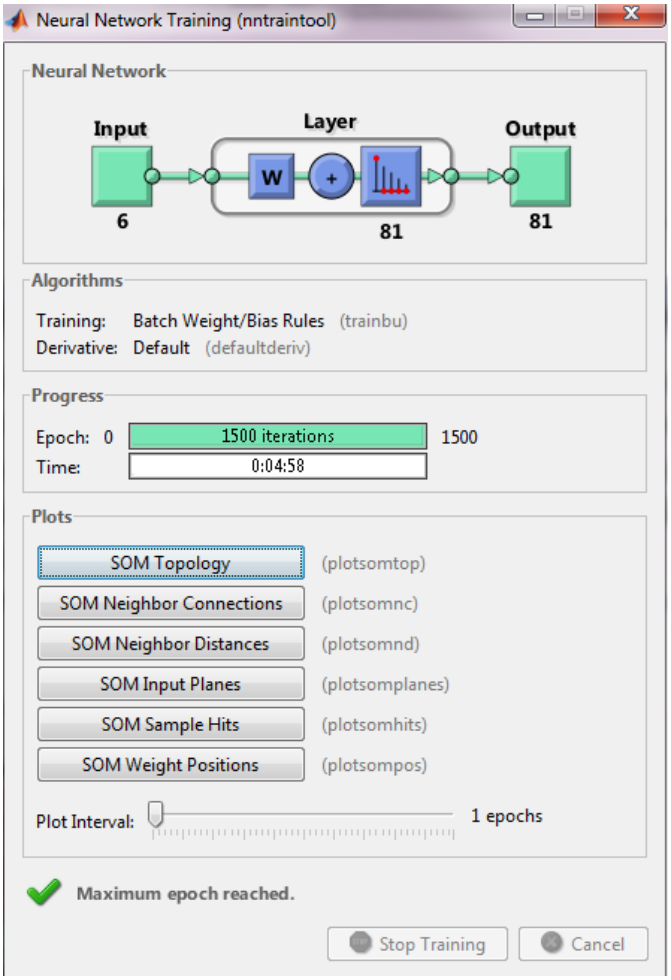
```
a = sim(net,input);

a1 = vec2ind(a);
```

- Proměnná a je výstup z neuronové sítě, která obsahuje jedničky a nuly. V každém sloupci je uvedena jedna jednička, která reprezentuje jednu ze tříd
- Proměnná a1 obsahuje pouze jeden řádek, kde je každý člověk zastoupen v nějaké třídě.

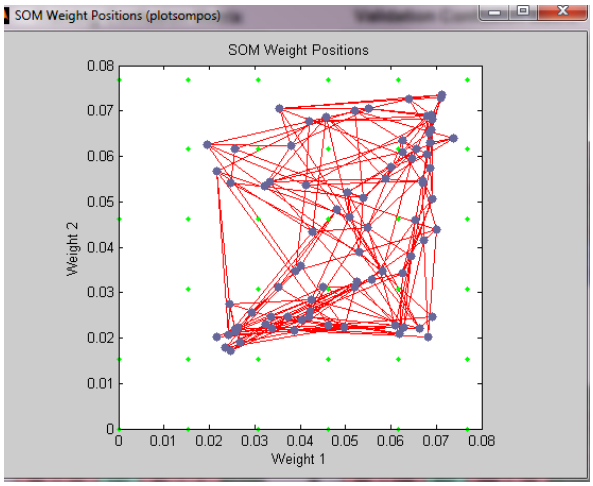
Vykreslení výsledků SOFM

Na obrázku lze vidět nástroj, který trénuje v tomto případě 9x9 dvourozměrné mapy 81 neuronů.



Obrázek 10: trénovací nástroj

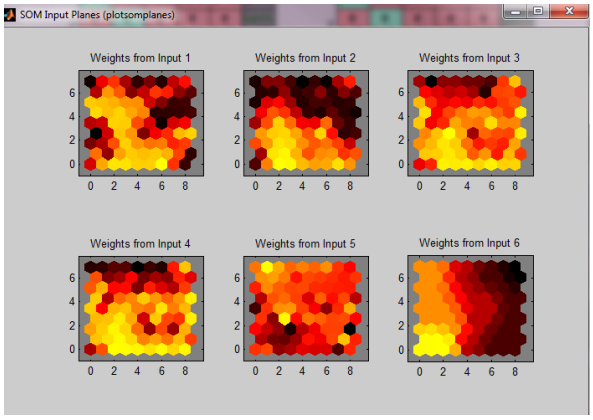
Existuje několik vizualizací, které mohou přistupovat právě z okna (nástroj trénování). Plotsompos ukazuje umístění datových bodů a váhy vektorů.



Obrázek 11: plotsompos

Další obrázek(plotsomnd)zobrazí vzdálenosti mezi sousedními neurony. Používá různě barevné kódování:

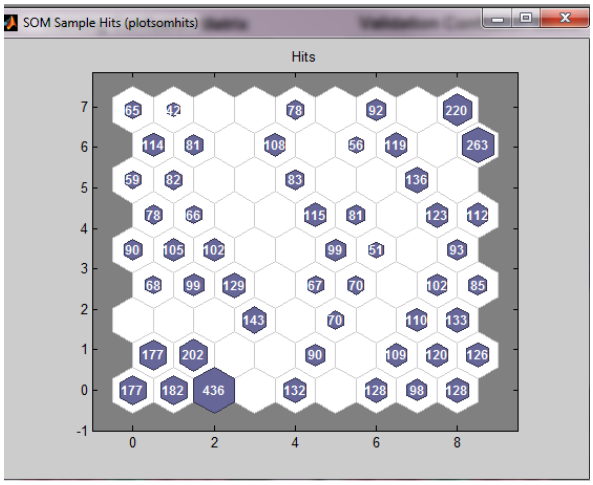
- červené linky spojují sousední neurony
- modré šestiúhelníky představují neurony
- barvy v regionech, které obsahují červené čáry označující vzdálenost mezi neurony(MathWorks, 2012):
  1. tmavší barvy představují větší vzdálenost
  2. světlejší barvy naopak menší vzdálenost



Obrázek 12: plotplanes

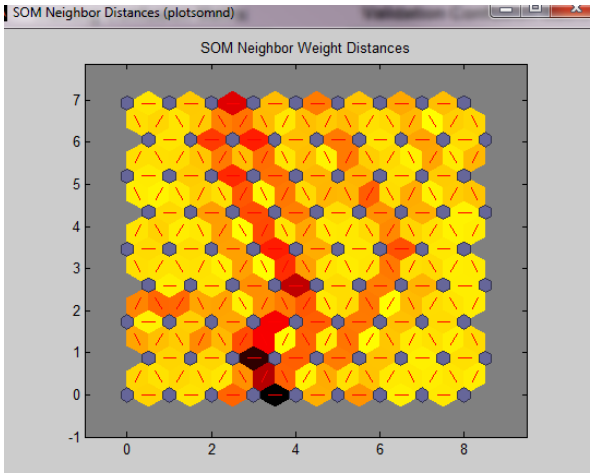
z obrázku lze vidět, že na mnoha místech jsou znázorněné lehké segmenty. V těchto oblastech jsou váhy blíže u sebe. Tmavé segmenty rozdělují právě světlé segmenty, kde jsou váhy ve vyšší vzdálenosti a vytváří tím několik shluků.

Níže je uveden obrázek plotsomhits, který ukazuje rozložení dat mezi neurony. Data jsou téměř rovnoměrně rozložená mezi neurony. Jen o něco více jsou soustředěna v levém dolním rohu. Neurony vytváří 5 shluků.



Obrázek 13: plotsomhits

Váhy se taky dají vizualizovat v následujícím obrázku plotsomnd. Na obrázku lze vidět 6 vstupních vektorů, které zobrazují neurony, hrany spojující neurony a ilustruje různými barvami vzdálenosti mezi váhami vektorů odpovídající sousedních neuronů. Tmavší barvy představují velkou vzdálenost. Jako například modré jsou právě ty nejvíce negativní. Nulové připojení je černá barva. Světlé barvy naznačují malé vzdálenosti. Lze vidět, že vstupy mají odlišné spojení.



Obrázek 14: plotsomnd

LVQ

LVQ je metoda pro trénování kompetitivní vrstvy hlídaným způsobem (s cílovými výstupy). Kompetitivní vrstva automaticky učí klasifikovat vstupní vektory.

Třídy které tato vrstva našla, jsou závislé na vzdálenosti mezi vstupními vektory. Jestliže dva vstupní vektory jsou velmi podobné, budou pravděpodobně zařazeny do stejné třídy.(MathWorks, 2012)

Jako vstup byl použit výstup ze SOFM a cílové hodnoty byly zvolené data z dotazníku na které respondenti odpovídali. Bylo tedy použito 6 cílových hodnot. Legenda, která je uvedena v příloze A(obrázek 22) tyto target hodnoty popisuje. Protože předtím byl výstup zkonvertován do tříd, zde bude opět překonvertován na 0 a 1. To stejné je potřeba udělat i s target. Protože struktura dat je ve sloupci, musíme napřed použít konverzi a opět zkonvertovat na 1 a 0.

```
Tc = ind2vec(a1);
a2= full(Tc);

target=nizka_cena';
Tc = ind2vec(target);
t= full(Tc);
```

Nejdříve byla použita process.fce lvqoutputs (a2), která vrací argumenty beze změny, ale ukládá poměr cílových tříd v nastavení (PS) pro použití inicializace vah(initlvq). Initlvq ('configure', a2) má vstupní data, které vrací inicializaci nastavení pro váhy LVQ spojené s tímto vstupem.

```
[a,PS] = lvqoutputs(a2)

initlvq('configure',a2);
```

Byla potřeba vytvořit síť. K tomuto účelu byla použita funkce lvqnet, která se skládá ze třech parametrů. Jako první je počet neuronů ve skryté vrstvě. Je potřeba, aby počet neuronů v první vrstvě byl vždy větší než ve druhé vrstvě.

Druhým parametrem je rychlost učení, jehož hodnota byla zvolena na 0,001 a poslední proměnnou je funkce učení. Zvolená funkce učení byla learnlv1. Počet neuronů ve skryté vrstvě byly nastaveny na 170. Tato hodnota se zdála nejoptimálnější, protože do této sítě vstupuje 81 neuronů a je potřeba, aby neurony ve skryté vrstvě byly alespoň 1 x větší.

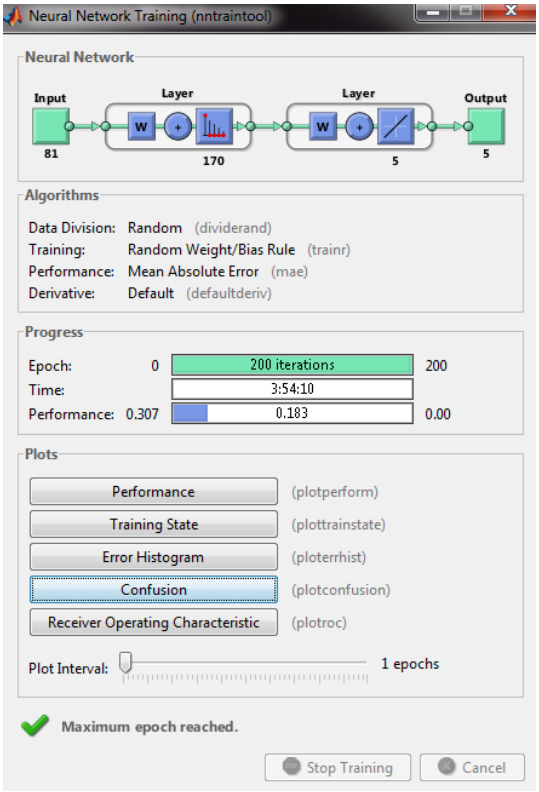
```
hiddenSize=170;
lvqLR=0.001;
lvqLF = 'learnlv1';
net = lvqnet(hiddenSize,lvqLR,lvqLF);
```

Síť byla nakonfigurována. Konfigurace je proces nastavení sítě, její vstupní a výstupní velikosti a rozsah.

```
net = configure(net,a2,t);
```

Z dat byly náhodně vybrány vzorky pro trénování, testování a validaci. Pro náhodný výběr byla použita funkce dividerand:

```
net.divideFcn = 'dividerand';
net.divideMode = 'sample';
net.divideParam.trainRatio = 70/100;
net.divideParam.valRatio = 15/100;
net.divideParam.testRatio = 15/100;
```



Obrázek 15: trénování

Trénování

Dále byla potřeba síť natrénovat k získání první vrstvy vah, které vedou ke správnému zařazení vstupních vektorů. Epochy byly nastaveny různě. Avšak nejvíce bylo nastaveno 200 epoch. Při použití 653 epoch bylo trénování zbytečné, protože chyba už neklesala. Výpočetně je LVQ velice náročné a tento počet epoch trval 10 hodin. Jako trénovací funkce byla zvolena trainr (trénink s náhodnými přírůstky)

```
net.trainFcn = 'trainr'
net.trainParam.epochs=200;
[net,tr2] = train(net,t,a2);

net.iw{1,1}
net.lw{2,1}
```

Testování a výpis chyb

Testuje se, zda tyto váhy skutečně vedou ke správnému zařazení. Vezme se výstup ze SOFM a jako vstup a simuluje se síť. Pomocí funkce vec2ind se výstup konvertuje do tříd, které jsou označená. Funkce mae vrátí výpočet průměrné absolutní chyby, kde se srovnává skutečná hodnota s odhadovanou.

```
outputs= sim(net,a2);

vystup = vec2ind(outputs)

E = t - outputs
per = mae (E)
```

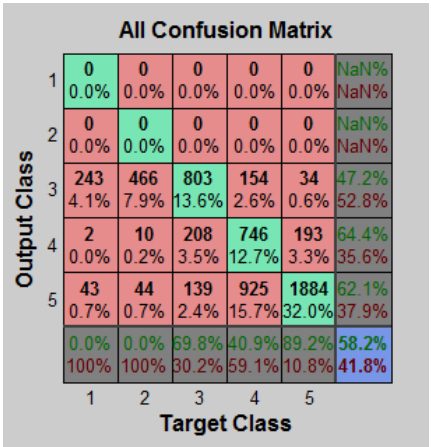
12.3 Výsledky LVQ

Zde je použito 6 souborů s cílovými hodnotami. Výsledky budou zvlášť rozdělena dle target.

Jako první byla použita cílová data:

Výrobek je v akční nabídce

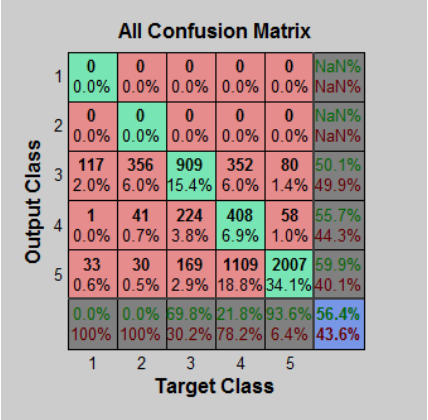
Pro trénování stačilo pouze 65 epoch. Na obrázku matice záměn lze vidět, že bylo dosaženo celkově 58,2% úspěšnosti. Nejlépe byla zařazena data s 64,4% do třídy, která preferuje výrobky v akční nabídce. 746 osob bylo zařazeno dobře a 621 špatně. Dobře byly také klasifikovány instance do třídy, ve které je brána velká důležitost pro výrobky v akční nabídce. Zde bylo klasifikováno správně 1884 respondentů a špatně 1151.



Obrázek 16: matice záměn - akční nabídka

Nízká cena

67 epoch pro trénování s 56% úspěšností. Data byla rozdělena do třech tříd. Třetí třída s 59,9% úspěšností říká, že pro 3348 respondentů je při výběru prodejny potravin nejdůležitější nízká cena.

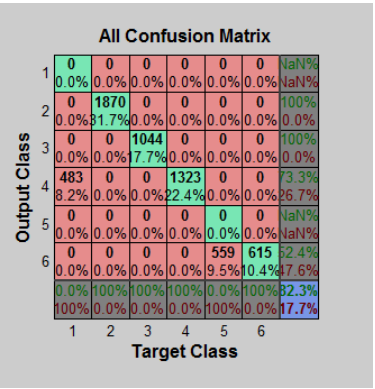


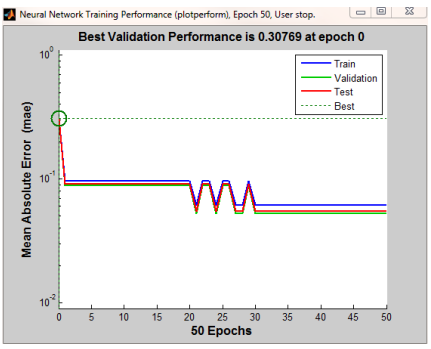
Obrázek 17: matice záměn - nízká cena

Věk respondenta

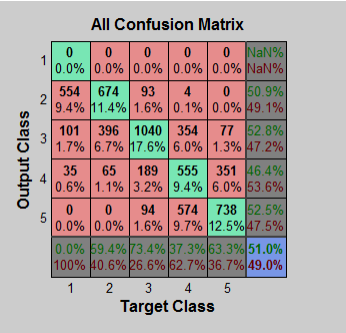
Zde bylo správně klasifikováno 82,3%. 100% proběhlo zařazení u druhé třídy s 1870 osob ve věku 18 až 24 let a třetí třídy s počtem 1044 ve věku 25 až 34. Do čtvrté třídy bylo správně zařazeno 1323 respondentů ve věku 35 až 54 a špatně 483. Pátá třída obsahovala správně klasifikované 615 osob ve věku 65 let a více a 599 špatně zařazených instancí.

Na obrázku perform je možno vidět, že chyba zde zpočátku klesla a u třicáté epochy opět klesla. Bylo použito k trénování 50 epoch.





Obrázek 18: matice záměn, perform - věk



Obrázek 19: matice záměn - zásoby

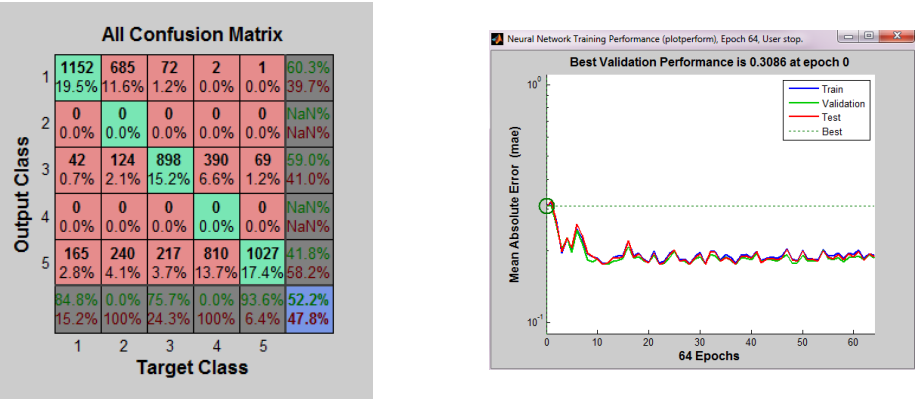
Nakupuji potraviny raději do zásoby, než když je to nutné, protože došly

Zde bylo provedeno 178 epoch s 51% úspěšností. Data byla rozdělena do 4 tříd. Největší úspěšnost měla druhá třída s 52,8%, kde bylo správně klasifikováno 1040 instancí a špatně 928 vzorků a čtvrtá kategorie s 52,5%

Před nákupem se nechám nejprve inspirovat nabídkou v letácích

U trénování stačilo pouze 64 epoch s 52% úspěšností. Hodnoty byly klasifikovány do třech tříd. Na obrázku matice záměn lze vidět klasifikaci s 60% úspěšnosti, kde je zařazeno 1912 respondentů pro které je inspirace nabídky v letáku velmi důležitá. 1523 osob bylo klasifikováno do třídy, která zastupuje středně důležitou inspiraci v letáku s 59% úspěšnosti. Nejméně procent úspěšnosti měla třída, která je zastoupena respondentama, pro které je tato inspirace nedůležitá. V této třídě bylo klasifikováno 2459 osob.

Na obrázku perform lze vidět, že chyba klesala zprvu rychle dolů. U dvacáté epochy začala kmitat nahoru a dolů. Po 60 iterací se začala dostávat do roviny.

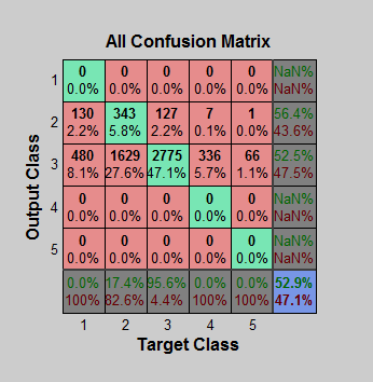


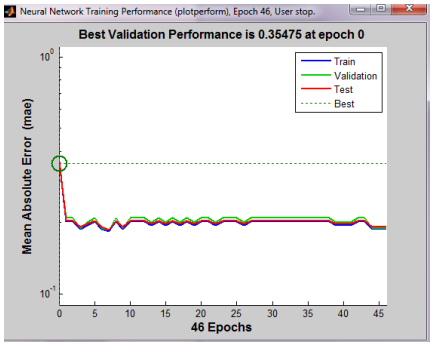
Obrázek 20: matice záměn, perform - letáky

Jak hodnotíte příjem vaší domácnosti z pohledu pokrytí potřeb a kvality života

Celková úspěšnost kvalifikace měla 52,9%. Klasifikace proběhla do dvou tříd. S 56,4% bylo klasifikováno 68 respondentů do třídy, která značí že jejich domácnost z pohledu pokrytí potřeb a kvality života je vyhovující (domácnost pokryje veškeré potřeby v přiměřeném rozsahu). 5286 osob bylo klasifikováno do třídy, kde jejich pokrytí je vysoké(možnost větších investic a nákupu luxusního zboží) s 52% úspěšností zařazení.

Celkově proběhlo 46 iterací. Chyba velmi rychle klesla, poté oscilovala a u 43 epochy ještě poklesla.





Obrázek 21: matice záměn, perform - příjem v domácnosti

Porovnání výsledků - tabulky

V níže uvedené tabulce jsou vypsány výsledky LVQ, kde největší úspěšnost klasifikace měl věk respondenta a naopak nejmenší měl atribut "Nakupuji potraviny raději do zásoby, než až když je to nutné, protože došly"s 51 %. U všech atributů byla úspěšnost na 50 procent.

V tabulce souhrn výsledků LVQ jsou sumarizované hodnoty počet epoch, čas výpočtu a velikost chyby mea s počáteční hodnotou 0,380. Nejvíce bylo provedeno 178 iterací, kde byl čas na výpočet velmi náročný, trvalo přes čtyři hodiny a chyba byla snížená na 0,2193. U cílové třídy věk respondenta stačilo pouze 50 epoch k dosažení velmi dobrého výsledku s hodnotou chyby 0,0589. U target "Nakupuji potraviny raději do zásoby,..."bylo provedeno pouze 46 epoch, kde byl dosažen výpočet chyby 0,2193 s časem jedna hodina a třicet minut.

Porovnání výsledků

V následující tabulce bylo provedeno porovnání výsledků algoritmu LVQ vypočítané v MATLABU a také metody MLP se všemi a s protříděnými atributy vypočítané

Tabulka 3: Výsledky LVQ

název	**úspěšnost v %**
Před nákupem se nechám nejprve inspirovat nabídkou v letácích	52
Nakupuji potraviny raději do zásoby, než když je to nutné, protože došly	51
Nízká cena	56
Jak hodnotíte příjem vaší domácnosti z pohledu pokrytí potřeb a kvality života	52,9
Věk respondenta	82,3
Výrobek je v akční nabídce	58,2

Tabulka 4: Souhrn výsledků LVQ

target	**počet epoch**	**chyba mae**	**čas výpočtu**
Před nákupem se nechám nejprve inspirovat nabídkou v letácích	64	0,192	1:07:35
Nakupuji potraviny raději do zásoby, než až když je to nutné, protože došly	178	0,2193	4:12:02
Nízká cena	67	0,2183	1:37:31
Jak hodnotíte příjem vaší domácnosti z pohledu pokrytí potřeb a kvality života	46	0,1884	1:28:01
Věk respondenta	50	0,0589	1:02:31
Výrobek je v akční nabídce	64	0,1670	1:46:20

ve Wece. Lze vidět, že nejlepších výsledků bylo dosaženo u algoritmu LVQ, kde se použilo 6 atributů pro vstup. Při porovnání s MLP, kde bylo také na vstupu jen 6 atributu, byla horší úspěšnost, až na nízkou cenu, kde bylo o 0,40% dosaženo lepšího výsledku.

Lze si však všimnout, že srovnání MLP s výběrem atributů a MLP použití všech proměnných na vstupu, je úspěšnější první varianta. Avšak lze si povšimnout, že u třídy "věk"byla úspěšnost s omezeným počtem proměnných u výpočtu MLP velmi mizivá. Výpočet zde byl o více než 30% horší než za použití všech parametrů. Metoda LVQ zde dosáhla velmi dobrých výsledků. Mimoto lze vidět, že u ostatních tříd proběhla klasifikace lépe za použití menšího počtu atributů.

U vybraných atributů byla úspěšnost u metody j48 horší ve dvouch případech a to zejména u třídy věk. Zde byl rozdíl téměř o 40%.

Tabulka 5: Porovnání výsledků LVQ a MLP úspěšnosti

název	**LVQ**	**MLP(výběr)**	**j48(výběr)**	**MLP(všechny)**	**j48(všechny)**
Před nákupem se nechám nejprve inspirovat nabídkou v letácích	52%	41,07%	36,87%	40%	55,92%
Nakupuji potraviny raději do zásoby, než až když je to nutné, protože došly	51%	33,72%	35%	31,25%	34,78%
Nízká cena	56%	56,40%	56,70%	49,85%	54,76%
Jak hodnotíte příjem vaší domácnosti z pohledu pokrytí potřeb a kvality života	52,9%	46,06%	51,80%	43,60%	51,24%
Věk respondenta	82,3%	32,37%	36%	67,75%	74,14%
Výrobek je v akční nabídce	58,2%	56,42%	56,50%	46,32%	55,87%

13 Závěr

Úkolem diplomové práce byla analýza daných ekonomických dat pomocí metod strojového učení. Byl proveden výběr vhodných kombinací tříd ze vstupních dat a jejich shlukování pomoci metody SOM a LVQ v programu Matlab a klasifikace v programu Weka.

Aby bylo možno cíl splnit, bylo potřeba upravit vstupní data. Nejprve byly použity data z prvního dotazníku, kde se tohoto výzkumu zúčastnilo 2335 respondentů a bylo jim položeno 29 otázek. Dohromady 105 atributů, protože některé otázky obsahovaly podotázky. I po provedení různých úprav byla stále vysoká chybovost, kterou lze vysvětlit velkou otevřenosti otázek. Úspěšnost se pohybovala pouze v rozmezí 30 - 40%. Byl proto použit jiný datový soubor s podobným typem otázek. Velkým rozdílem byl počet atributů a počet instancí. Celkový počet atributů byl 29 a počet respondentů činil 5984. Z tohoto důvodů bylo dosaženo lepších výsledků.

Po výběru datasetu byla potřeba také jejich úprava. Mnoho respondentů nezodpovědělo velkou míru otázek, z tohoto důvodu bylo nutné promazat některé instance. Data byla připravena ve formátu xls v číselné a nominální podobě. Pro program Weka bylo potřeba změnit číselné odpovědi na nominální hodnoty a přeuložit do souboru csv. Naopak program Matlab pracuje s číselnými hodnotami. Vybrané atributy se musely změnit z nominálních hodnot na číselné.

Nejprve byl použit program Weka, kde byly klasifikovány všechny atributy a poté byla použita filtrace. K tomuto účelu se zvolila metoda hodnocení příznaků ChiSquaredAttributeEval, která dle vstupních hodnot volí příznaky na základě výpočtu chí statistiky a metodu vyhledávání Ranker, která tyto příznaky seřadí dle individuálního ohodnocení. Po tomto seřazení bylo vybráno šest atributů, na kterých se dále prováděla klasifikace pomocí metody MLP, rozhodovacího stromu j48 a bylo také použito rozhodovací pravidlo PART. Tento výběr atributů byl použit také pro metody SOM a LVQ v programu Matlab. Vzhledem tomu, že závislost atributů byl vybrán v programu Weka, nebyl vytvořen skript pro nalezení závislosti mezi atributy.

Při použití všech atributů bylo dosaženo velmi pozitivních výsledků u tříd, které charakterizují respondenty, avšak pro nás ne moc zajímavé. Nejprve bylo použito rozhodovací pravidlo zeroR. Tato metoda je nejjednodušší pro klasifikaci. Úspěšnost u zeroR je jen informativní a s porovnáním s ní by ostatní metody neměli mít horší výsledky. U tříd, které nám poskytly informace, byla úspěšnost klasifikace okolo 55% a následně byly použity pro metodu PART, kde se objevily zajímavé závislosti a výsledky. U třídy nízká cena bylo správně zařazených 55,39%, kde bylo zjištěno, že lidé, kteří preferují nízkou cenu, nakupují potraviny v akční nabídce, a dávají přednost trvanlivým potravinářským výrobkům. Příjem v těchto domácnostech je nízký. Dále že důchodci z Jihomoravského kraje nakupují výrobky s nízkou cenou v akční nabídce.

U třídy "datum spotřeby"byla úspěšnost 57,63 %, kde jsme se dozvěděli, že ženy upřednostňují prodejny s čerstvými potravinami českého původu. Pro ženy z kraje

Vysočina je velmi důležité kupovat čerstvé potraviny českého původu a v nejvyšší kvalitě. U třídy "výrobek je v akční nabídce"bylo klasifikováno správně 57,67%. Respondenti nakupující v akční nabídce upřednostňující nízkou cenu a před nákupem se nechají inspirovat v letácích.

Po použití filtrace, bylo vybráno 6 atributů, kde bylo dosaženo lepších výsledků, než u použití všech atributů. Nejlepší výsledky byly dosažené u třídy "nízká cena", "výrobek je v akční nabídce"a "věk respondenta". U rozhodovacího pravidla PART bylo zjištěno, že lidé, kteří nakupují levné výrobky, mají nízký příjem domácnosti (základní potřeby domácnost pokryje, ale musí v nich šetřit eventuelně se omezovat) a zaměřují se na výrobky v akční nabídce a také se právě pro to nechají inspirovat nabídkou v letácích.

Metoda SOM rozdělila instance do 5 shluků a následně byla použita pro metodu LVQ, která 100% klasifikovala respondenty ve věku 18 až 24 let. Celková úspěšnost třídy "věk"byla 82,3%. A u výrobku v akční nabídce byla celková úspěšnost 58,2%, kde nejlépe byly zařazeny instance s 64,4% do třídy preferující výrobky v akční nabídce. Metoda LVQ byla velmi náročná na výpočet, kde při 178 iterací trvala klasifikace přes čtyři hodiny.

Po porovnání MLP s použitím všech atributů a poté jen s šesti, byla lepší úspěšnost ve vyfiltrovaných proměnných, avšak u třídy věk byl horší výsledek. To stejné bylo také u metody J48. MLP je na výpočet také náročný. Výpočet jedné třídy se pohyboval okolo dvou hodin.

Celkově lze říci, že byly zjištěny zajímavé závislosti. Některé třídy nám poskytly zajímavé informace a naopak jiné nám vypočítaly výsledky, které jsou nepoužitelné. Některé třídy byly klasifikovány lepé při použití všech atributů, a naopak některé při využití šesti atributů.

14 Literatura

HŘEBÍČEK Jiří, ŽIŽKA Jan. *Vědecký výpočty v biologii a biomedicíně. Brno, 2007. č. 2588/2007.*. Učební text. Masarykova univerzita..

CHURÝ, Lukáš. Umělá inteligence, díl 2. - neuronové sítě. Programujte.com [online]. 2005, červen [cit. 2012-12-14]. Dostupné z: http://programujte.com/clanek/2005061201-umela-inteligence-dil-2-neuronove-site/..

JANDOVÁ, Michaela, Marie PŘIBOVÁ. Destinacní management a vytvárení produktu v cestovním ruchu: Výzkumy trhu v destinaci. Mmr.cz [online]. Praha, 2006 [cit. 2012-12-25]. Dostupné z: http://mmr.cz/getmedia/657aafef-2893-48d7-baa7-2fb459b20625/GetFile1$_2$*.pdf*.

KARBAN, Pavel. *Výpočty a simulace v programech Matlab a Simulink. 1. vyd.*. Brno: Computer Press, 2006. 220 s. ISBN 80-251-1301-9..

KOHONEN, T. *Self-organizing maps.*.3. vyd. New York: Springer Verlag, 2001. 501 s. ISBN 3-540-67921-9..

KUBIK, Aleš. *Inteligentní agenty. Vyd. 1.* . Brno: Computer Press, 2004, 280 s. ISBN 80-251-0323-4. .

MathWorks [online]. © 1994-2012 [cit. 2012-01-15]. Dostupné z: http://www.mathworks.com/.

MIHAESCU, Cristi. Laboratory Module 1 Description of WEKA (Java-implemented machine learning tool). In: Software.ucv.ro [online]. 2011 [cit. 2012-12-14]. Dostupné z: http://software.ucv.ro/ cmihaescu/ro/teaching/AIR/docs/Lab2-DescriptionOfWEKA.pdf.

PEJČOCH, David. Metody řešení problematiky neúplných dat [online]. Praha, 2011 [cit. 2012-12-22]. Dostupné z: http://www.dataquality.cz/tutorial/tutorial_04.pdf. Tutoriál. Vysoká škola ekonomická..

SCUSE, David. WEKA Manual for Version 3-6-8 [online]. 2012 [cit. 2012-12-23]. Dostupné z: http://freefr.dl.sourceforge.net/project/weka/documentation/3.6.x/WekaManual-3-6-8.pdf..

SIMOVÁ, Jozefína. *Marketingový výzkum. Vyd. 1.* . Liberec: Technická univerzita v Liberci, 2005. 121 s. ISBN 80-737-2014-0..

SMITH, K. A. – GUPTA, J. N. D. *Neural networks in business : techniques and applications.*. Hershey, PA: IRM Press, 2003. 258 s. ISBN 1-931777-79-9..

ŠNOREK, Miroslav. *Neuronové sítě a neuropočítače. 1. vyd.*. Praha: ČVUT Praha, 2004, 156 s. ISBN 80-010-2549-7..

ŠTENCL, M. – ŠŤASTNÝ, J. *Artificial Neural Networks Numerical Forecasting of Economic Time Series.*. In: Artificial Neural Networks - Application. Artificial Neural Network. Riejka, Croatia: InTech, 2011. s. 13–28. ISBN 978-953-307-188-6..

TUČKOVÁ, Jana. *Úvod do teorie a aplikací umělých neuronových sítí: Neuronové sítě a genetické algoritmy. 1. vyd.*. Praha: Vydavatelství ČVUT, 2003, 103 s. ISBN 80-010-2800-3. .

TUČKOVÁ, Jana. *Vybrané aplikace umělých neuronových sítí při zpracování signálů. Vyd. 1.*. Praha: České vysoké učení technické v Praze, 2009, 224 s. ISBN 978-80-01-04229-8..

VONDRÁK, Ivo. *Umělá inteligence a neuronové sítě. 2. vyd.*. Ostrava: Vysoká škola báňská - Technická univerzita, 2000. 140 s. ISBN 80-707-8949-2..

ZAPLATÍLEK, Karel. *MATLAB:průvodce začínajícího uživatele.*. Brno: Tribun EU s.r.o, 2011. ISBN 978-80-263-0014-4..

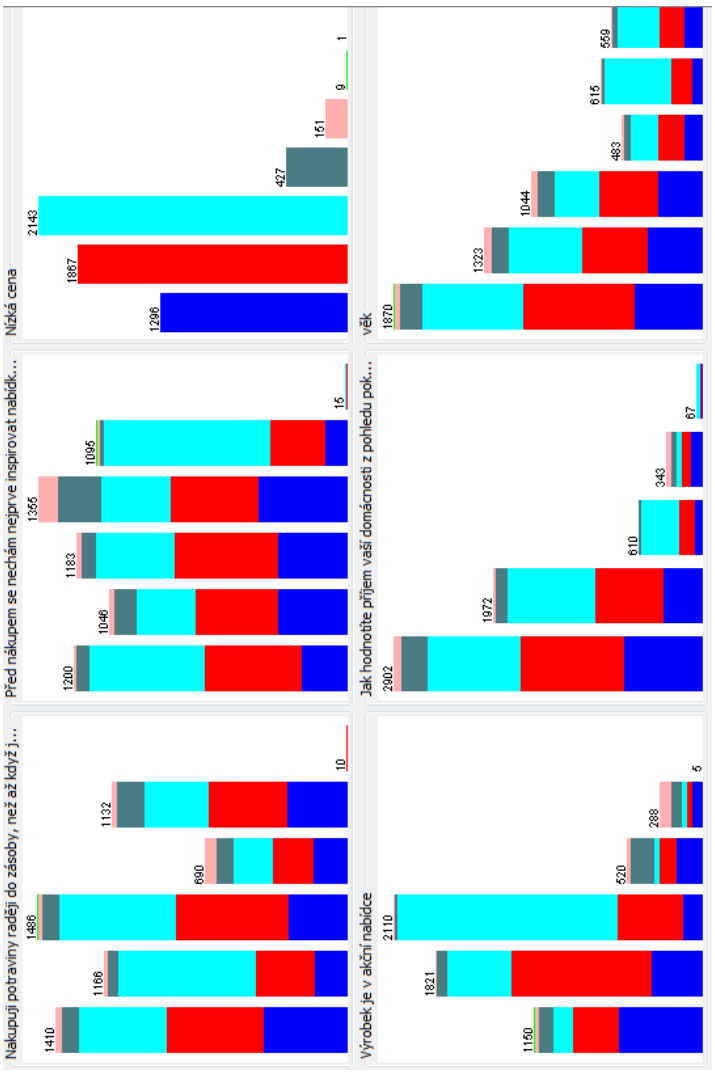
ZELINKA, Ivan. *Umělá inteligence I: Neuronové sítě a genetické algoritmy. 1. vyd.*. Brno: VUT v Brně, 1998, 126 s. ISBN 80-214-1163-5..

Přílohy

legenda	
Nakupuji potraviny raději do zásoby, než až když je to nutné, protože došly.	10
I	1
II	2
III	3
IV	4
V	5
Před nákupem se nechám nejprve inspirovat nabídkou v letácích.	20
I	1
II	2
III	3
IV	4
V	5
Nízká cena	30
I	1
II	2
III	3
IV	4
V	5
Výrobek je v akční nabídce	40
I	1
II	2
III	3
IV	4
V	5
Jak hodnotíte příjem vaší domácnosti z pohledu pokrytí potřeb a kvality života	50
A	1
B	2
C	3
D	4
E	5
věk	60
12 až 17	17
18 až 24	24
25 až 34	34
35 až 54	54
55 až 64	60
65 a více	65

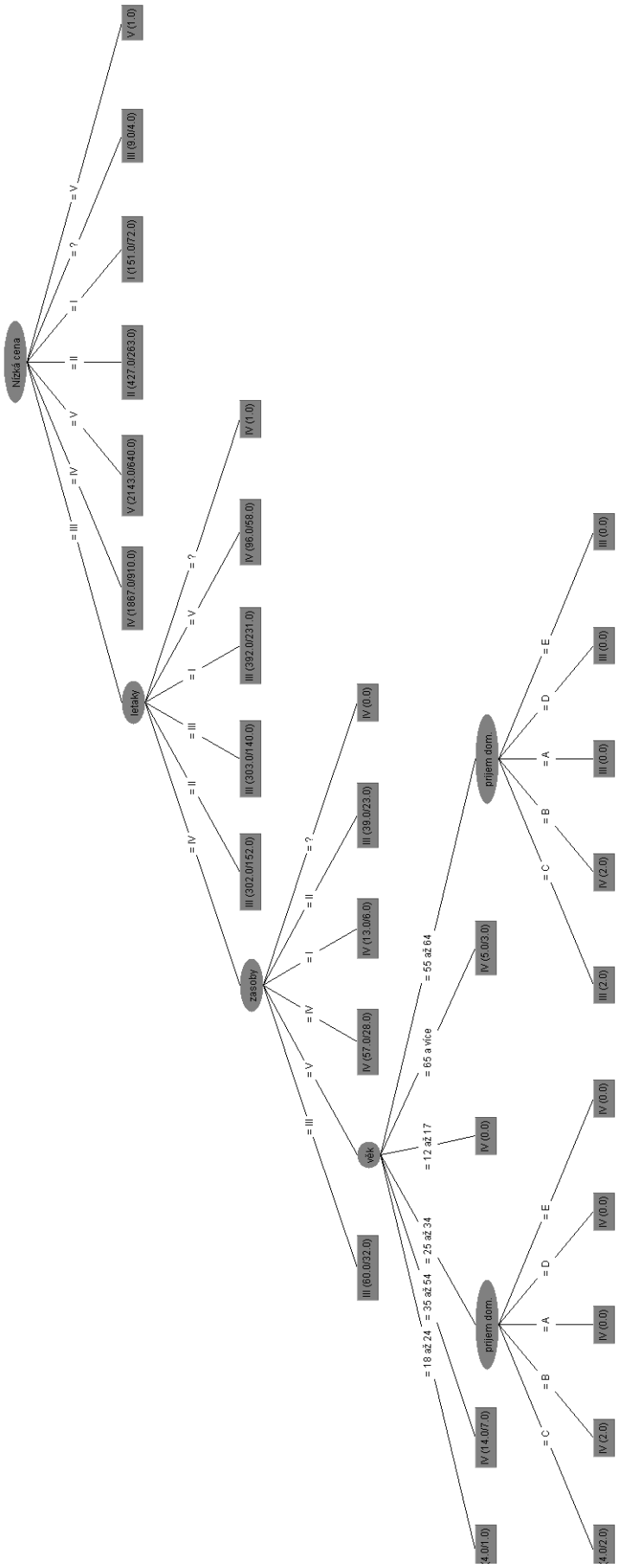
Obrázek 22: lagenda dat pro matlab

A legenda



Obrázek 23: rozložení hodnot

B výstupy - weka



Obrázek 24: rozhodovací strom J48 - výrobky v akční nabídce