ANALÝZA A KLASIFIKACE DAT

prof. Ing. Jiří Holčík, CSc.

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

KDY A KDE SE BUDEME VÍDAT?

LITERATURA

- Holčík,J.: Analýza a klasifikace dat. Brno, CERM 2012, 112s.

  http://www.iba.muni.cz/res/file/ucebnice/holcik-analyza-klasifikace-dat.pdf

  http://www.iba.muni.cz/index.php?pg=vyuka--ucebnice

- Holčík, J.: přednáškové prezentace
- Holčík, J.: Analýza a klasifikace signálů. [Učební texty VŠ], Brno, FE VUT 1992.

LITERATURA

- Duda,R.O., Hart,P., Stork,D.G. Pattern Classification. New York, John Wiley & Sons 2001
- Theodoridis S., Koutroumbas K., Pattern Recognition. Amsterdam, Elsevier 2009
- McLachlan,G.J.: Discriminant Analysis and Statistical Pattern Recognition. J.Wiley&Sons, Hoboken 2004
- Webb,A.: Statistical Pattern Recognition. J.Wiley&Sons, Chichester 2002
- Meloun, M., Militký,J.: Statistická analýza experimentálních dat. Praha, Academia 2004.

UKONČENÍ PŘEDMĚTU

Požadavky:

- ☑ ústní zkouška
  - ➔ dvě části:
    - učená rozprava o některém z témat, která budou náplní předmětu;
    - diskuze nad individuálním vyřešeným problémem týkajícím se problematiky klasifikace dat a používajícím některé z technik, které budou náplní předmětu;

I. ZAČÍNÁME

CÍL ZPRACOVÁNÍ DAT

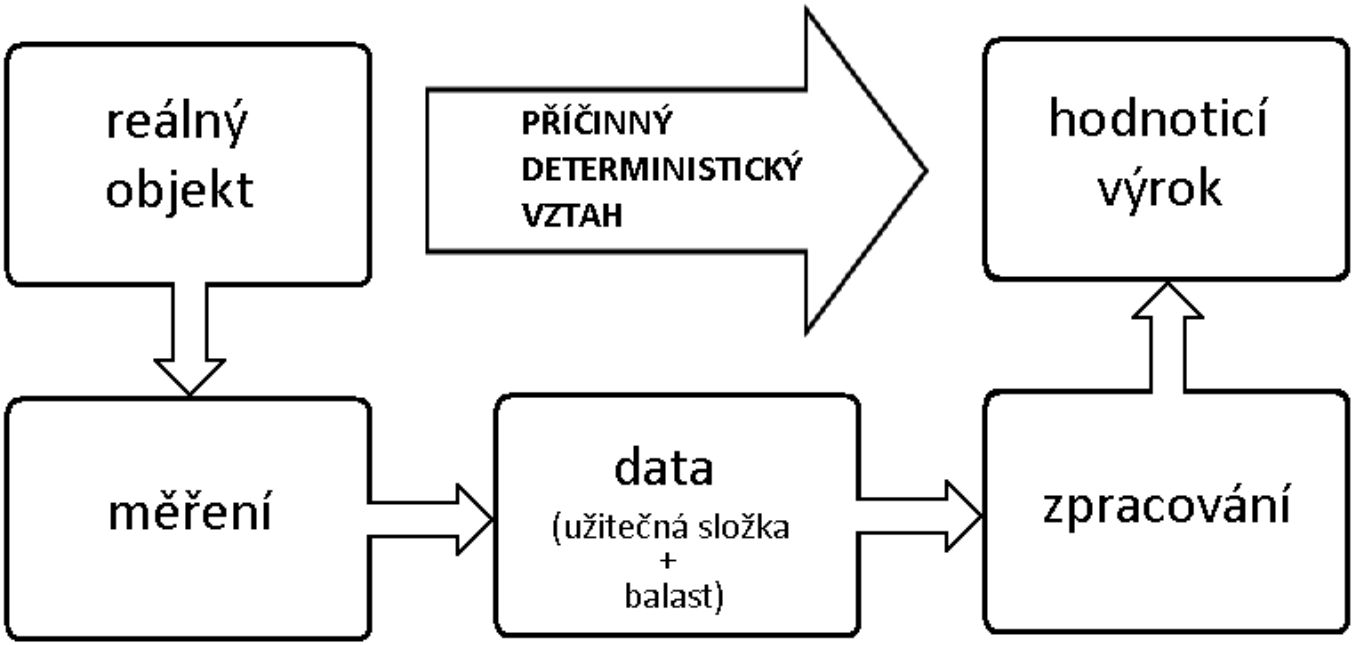
CÍL ZPRACOVÁNÍ DAT

Cílem jakéhokoliv zpracování (analýzy) dat je zpravidla posouzení zkoumaného reálného objektu (živého či neživého), který je zdrojem analyzovaných dat, příp. jeho stavu.

Toto posouzení může nejčastěji vyústit:

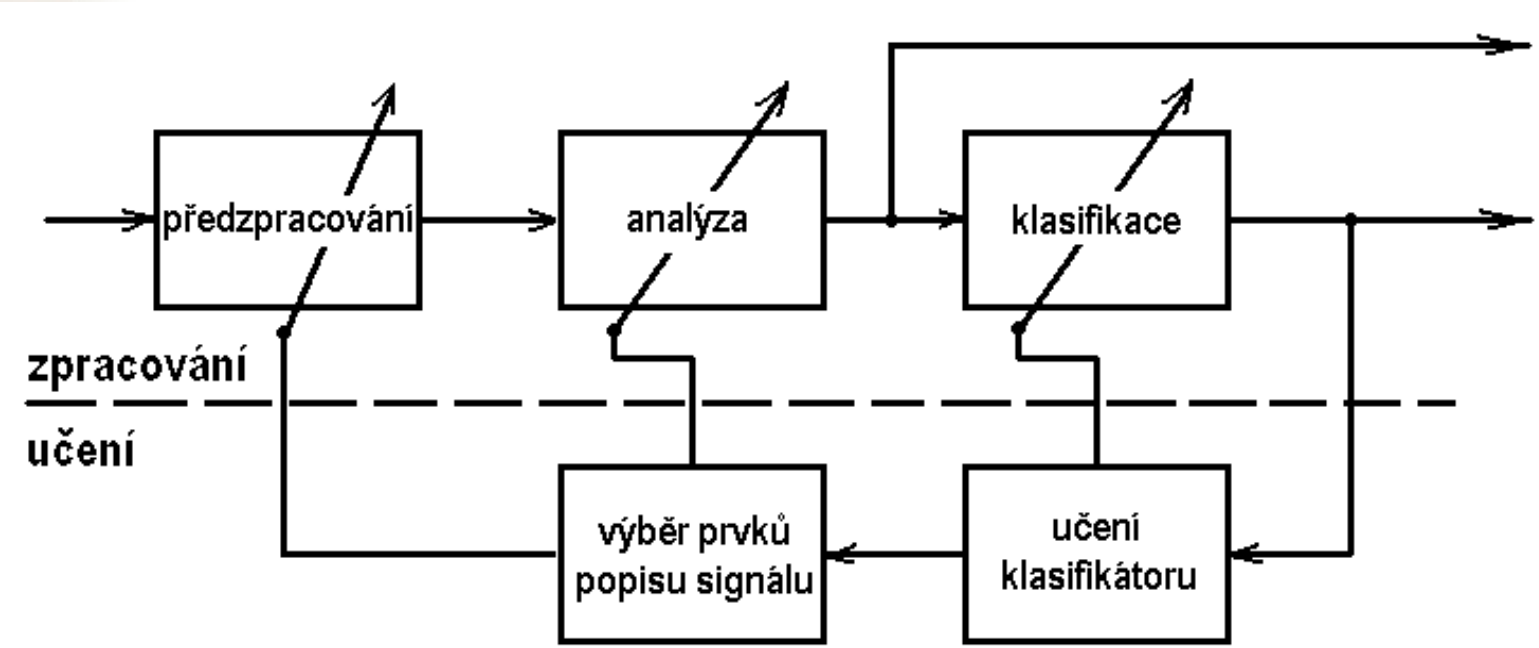
- ☑ v rozhodnutí o typu či charakteru objektu – např. že daná rostlina je pomněnka lesní (*Myosotis sylvatica*), zvíře že je medvěd hnědý (*Ursus arctos*), nebo že daná budova je vystavěna v renesančním slohu – ***klasifikační úloha***, resp. ***rozpoznávací***;
- ☑ v posouzení kvality stavu analyzovaného objektu, např. zda je pacient v pořádku, nebo má infarkt myokardu, cirhózu jater, apod. – opět ***klasifikační***, resp. ***rozpoznávací úloha***;
- ☑ v rozhodnutí o budoucnosti objektu – např. zda lze pacienta léčit a vyléčit, zda les po 20 letech odumře, jaké bude sociální složení obyvatelstva na daném území a v daném čase – ***klasifikační*** nebo také ***predikční úloha***

CÍL ZPRACOVÁNÍ DAT

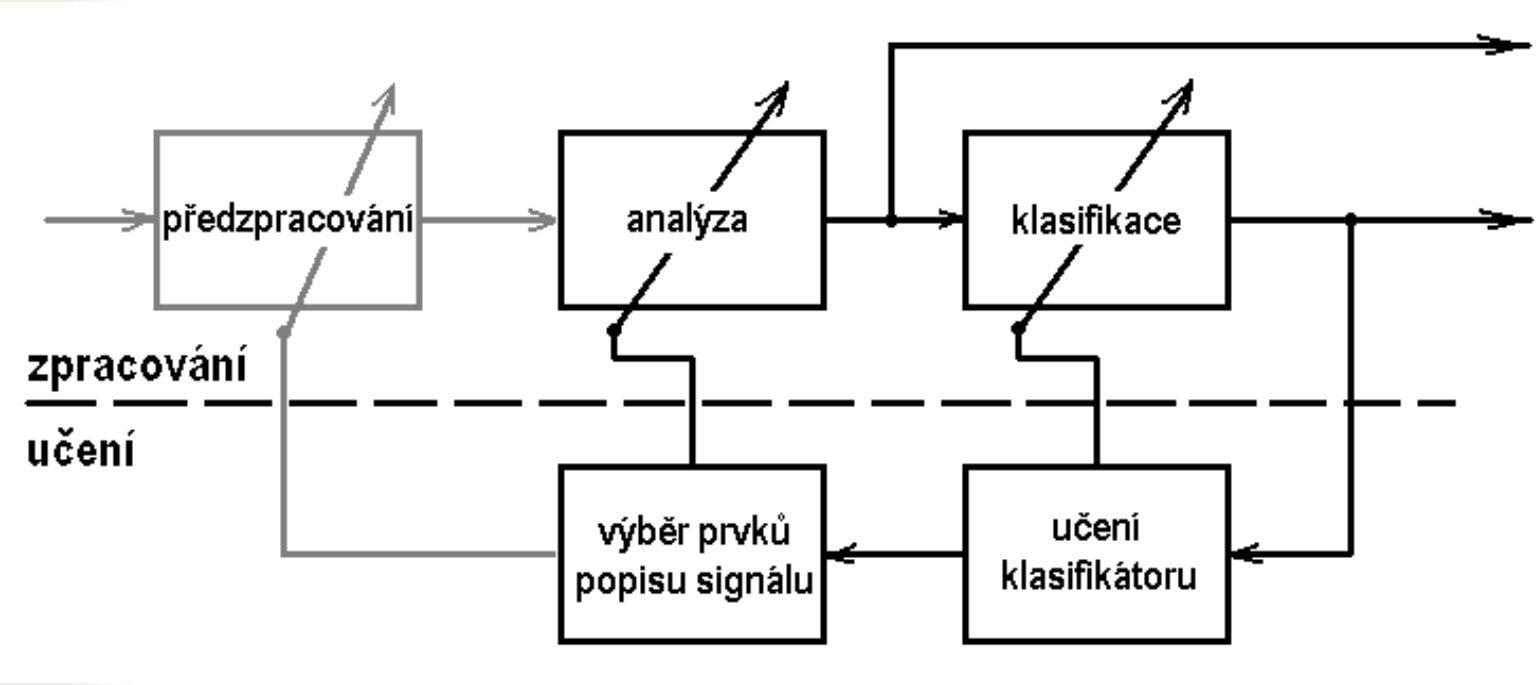


Chceme-li upřesnit dříve definovanýcíl zpracování (analýzy) dat, pak je to právě odhalení toho příčinného deterministického vztahu, navzdory všemu tomu, co to odhalení kazí.

OBECNÉ SCHÉMA ZPRACOVÁNÍ DAT



OBECNÉ SCHÉMA ZPRACOVÁNÍ DAT



OBECNÉ SCHÉMA ZPRACOVÁNÍ DAT

ZPRACOVÁNÍ

- **předzpracování**
  - filtrace rušivých složek x zvýraznění užitečných složek dat;
  - rekonstrukce a doplnění chybějících údajů;
  - konverze typu dat (A/Č převod);
  - redukce dat;
- **analýza dat**
  - určení hodnot příznaků (reprezentativních parametrů) – pro příznakové klasifikátory;
  - nalezení primitiv (charakteristických tvarových segmentů) – strukturální klasifikátory
- **klasifikátor** –
  - zatřídění do diagnostických kategorií

OBECNÉ SCHÉMA ZPRACOVÁNÍ DAT

ZPRACOVÁNÍ

- **předzpracování**
  - filtrace rušivých složek x zvýraznění užitečných složek dat;
  - rekonstrukce a doplnění chybějících údajů;
  - konverze typu dat (A/Č převod);
  - redukce dat;
- **analýza dat**
  - určení hodnot příznaků (reprezentativních parametrů) – pro příznakové klasifikátory;
  - nalezení primitiv (charakteristických tvarových segmentů) – strukturální klasifikátory
- **klasifikátor –**
  - zatřídění do diagnostických kategorií

OBECNÉ SCHÉMA ZPRACOVÁNÍ DAT

- **Analýza** (z řečtiny – *rozbor, rozčlenění*) je vědecká metoda založená na dekompozici celku na elementární části. Cílem analýzy je identifikovat podstatné a nutné vlastnosti elementárních částí celku, poznat jejich podstatu a zákonitosti.

OBECNÉ SCHÉMA ZPRACOVÁNÍ DAT

- **Analýza** (z řečtiny – *rozbor, rozčlenění*) je vědecká metoda založená na dekompozici celku na elementární části. Cílem analýzy je identifikovat podstatné a nutné vlastnosti elementárních částí celku, poznat jejich podstatu a zákonitosti.
- **Syntéza** je obecné označení pro proces spojení dvou nebo více částí do jednoho celku. S tímto pojmem se lze setkat v různých spojeních: syntéza obrazu, syntéza řeči, syntéza zvuku, chemická syntéza, jaderná syntéza, termonukleární syntéza, syntéza látek, fotosyntéza, proteosyntéza, biosyntéza, evoluční syntéza.

ANALÝZA

V bloku analýzy se vytváří formální (abstraktní) popis zpracovávaných dat, který nese podstatnou informaci z hlediska kvality rozhodování při klasifikaci. Abstraktní popis se často nazývá obrazem (pattern)⇒ rozpoznávání obrazů (pattern recognition). V datech je vybrána určitá množina elementárních vlastností, příp. jejich elementárních částí a jejich vazeb, jejichž způsob popisu je apriori znám.

KLASIFIKACE

- rozumí se rozdělení (konkrétní či teoretické) dané skupiny (množiny) předmětů či jevů na konečný počet dílčích skupin (podmnožin), v nichž všechny předměty či jevy mají dostatečně podobné společné vlastnosti. Vlastnosti podle nichž lze klasifikaci zadat či provádět, určují klasifikační kritéria. Předměty (jevy), které mají podobnou uvažovanou vlastnost tvoří třídu.

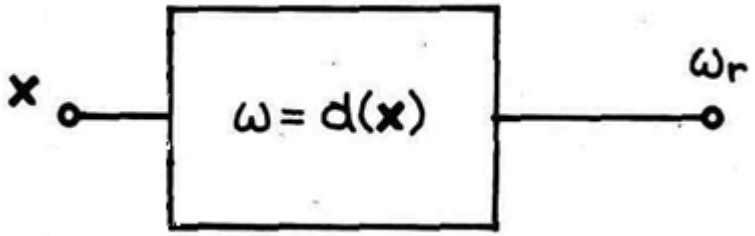
KLASIFIKÁTOR

- **Klasifikátor** je stroj (algoritmus,…) s jedním diskrétním výstupem, který udává třídu, do které klasifikátor zařadil vstupní reprezentaci dat

$$\omega_r = d(\mathbf{x})$$

$d(\mathbf{x})$ je funkce argumentu $\mathbf{x}$ představujícího reprezentaci vstupních dat, kterou nazýváme **rozhodovací pravidlo klasifikátoru**;

ω_r je **identifikátor klasifikační třídy;** $\omega_r|_{r=1,\dots,R} \in \Omega$

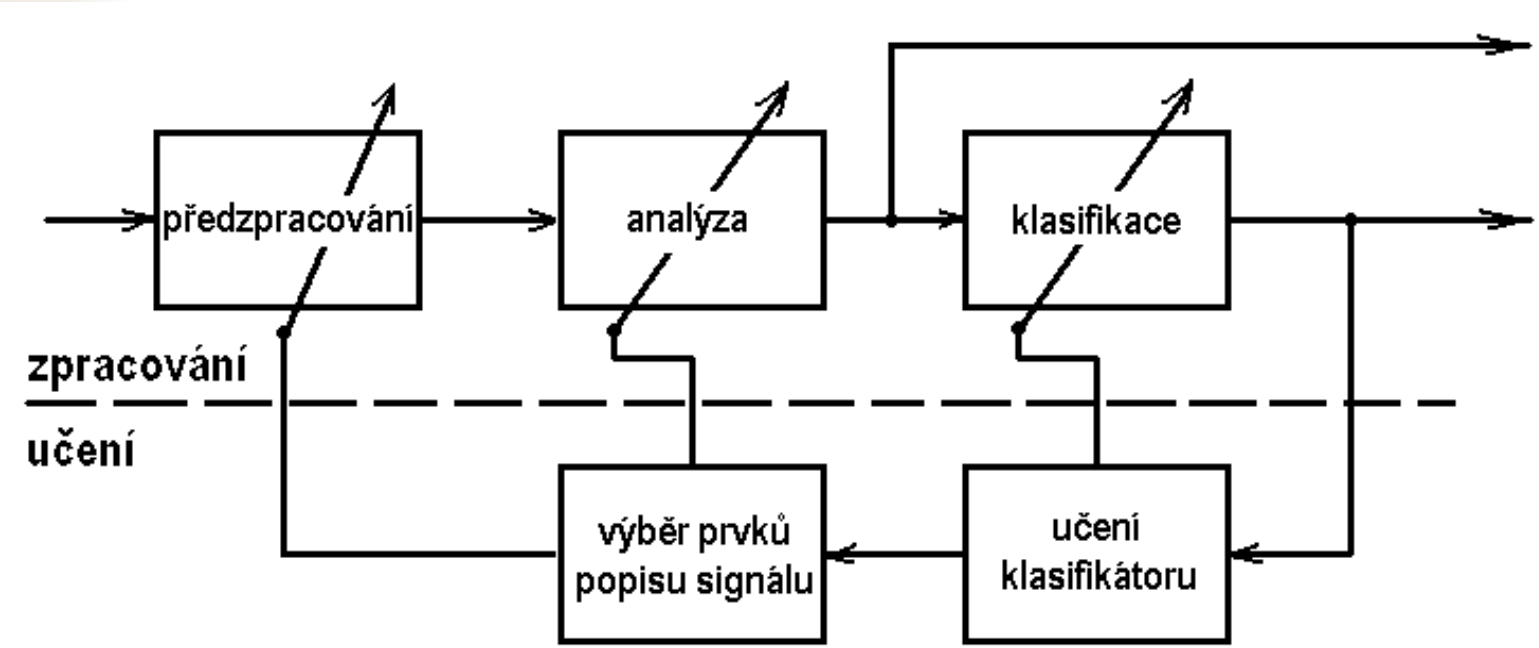


PRINCIPY KLASIFIKACE

PRINCIPY KLASIFIKACE

- pomocí diskriminačních funkcí – funkcí, které určují míru příslušnosti k dané klasifikační třídě;
- pomocí definice hranic mezi jednotlivými třídami a logických pravidel;
- pomocí vzdálenosti od reprezentativních obrazů (etalonů) klasifikačních tříd;
- pomocí ztotožnění s etalony;

OBECNÉ SCHÉMA ZPRACOVÁNÍ DAT



OBECNÉ SCHÉMA ZPRACOVÁNÍ DAT

UČENÍ

- **učení klasifikátoru**
  - nastavení klasifikačních kritérií;
    - s učitelem
      - dokonalým
      - nedokonalým
    - bez učitele – typicky shlukování
- **výběr prvků popisu dat**
  - stanovení reprezentativních charakteristických rysů zpracovávaného dat;

TYPY KLASIFIKÁTORŮ

Základní členění vychází z reprezentace vstupních dat

- příznakové – každý vstupní data jsou vyjádřena vektorem hodnot (příznaků);
  - paralelní (např. Bayesův klasifikátor, …)
  - sekvenční (např. klasifikační stromy, …)
- strukturální (syntaktické) – vstupní data jsou popsána relačními strukturami;
- kombinované – jednotlivá primitiva jsou doplněna příznakovým popisem

TYPY KLASIFIKÁTORŮ

Deterministický klasifikátor

- každá deterministická klasifikace musí být jednoznačná a úplná, tzn., že každý obraz (předmět, jev) musí patřit do nějaké třídy a nemůže být současně ve dvou či více třídách.

Pravděpodobnostní klasifikátor

- pravděpodobnostní klasifikátor stanoví pravděpodobnost zařazení obrazů do daných klasifikačních tříd

TYPY KLASIFIKÁTORŮ

Na základě typů klasifikačních a učících algoritmů:

- ☑ parametrické;
- ☑ neparametrické

KLASIFIKACE x PREDIKCE

predikce (z lat. *prae*-, před, a *dicere*, říkat) zjevně nese časové hledisko, když jej používáme ve významu předpověď či prognózu, jako soud o tom, co se stane nebo nestane v budoucnosti. V tomto významu je používán např. v analýze či zpracování časových řad.

(prediction x forecasting)

KLASIFIKACE x PREDIKCE

pojem klasifikace je používán, použije-li se klasifikačního algoritmu pro známá data. Pokud jsou data nová, pro která apriori neznáme klasifikační třídu, pak hovoříme o predikci klasifikační třídy.

http://www.kdnuggets.com/faq/classification-vs-prediction.html (23.8.2010)

KLASIFIKACE x PREDIKCE

pojem klasifikace používáme, pokud vybíráme identifikátor klasifikační třídy z určitého diskrétního konečného počtu možných identifikátorů. Pokud určujeme (predikujeme) spojitou hodnotu, např. pomocí regrese, pak hovoříme o predikci, i když tento pojem nemá časovou dimenzi.

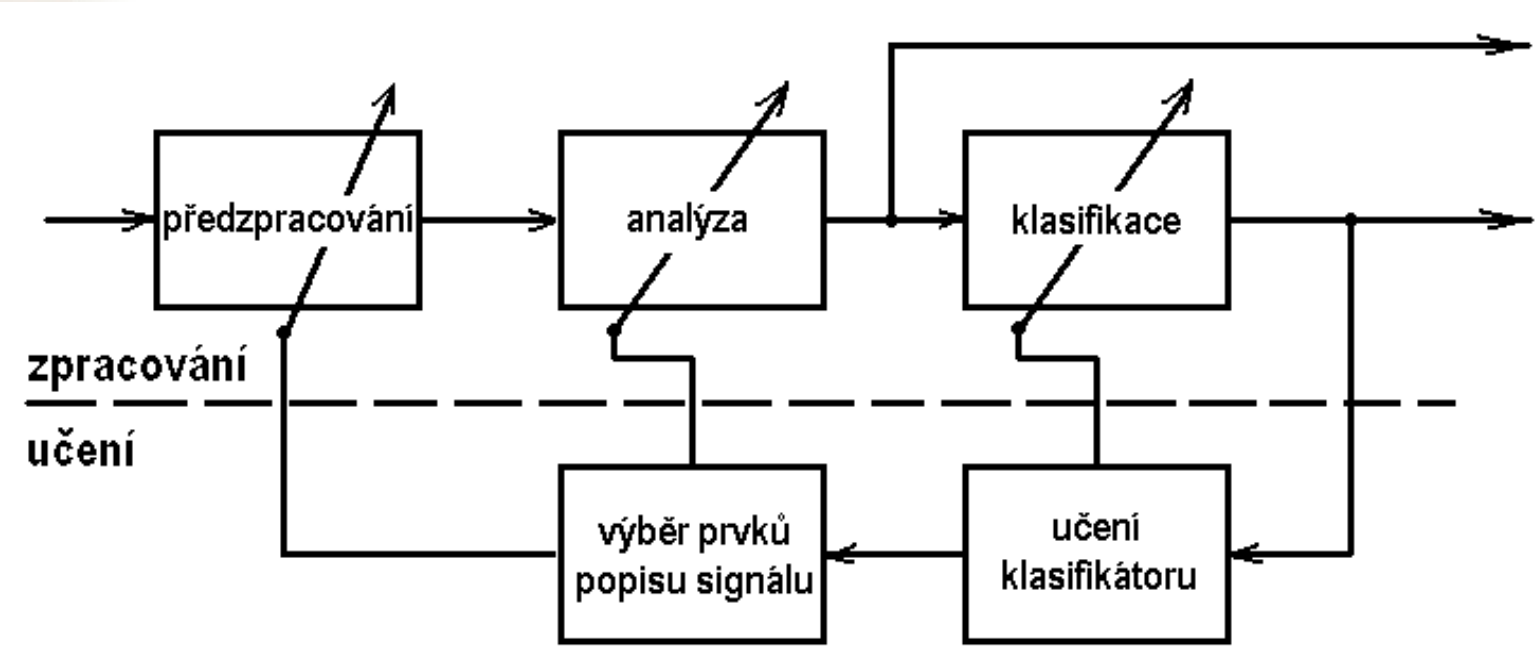
Han, J., Kamber, M.: Data Mining Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems. 2nd edition, Elsevier; Amsterdam(2005), 800 s.

DISKRIMINAČNÍ ANALÝZA

týká se obecně vztahu mezi kategoriální proměnnou a množinou vzájemně vázaných příznakových proměnných.

Konkrétně, předpokládejme že existuje konečný počet, řekněme R, různých a priori známých populací, kategorií, tříd nebo skupin, které označujeme ω_r, r=1,…,R a úkolem diskriminační analýzy je nalézt vztah, na základě kterého pro daný vektor příznaků popisujících konkrétní objekt tomuto vektoru přiřadíme hodnotu ω_r.

OBECNÉ SCHÉMA ZPRACOVÁNÍ DAT



ZÁVĚREM SHRNUTÍ

- ☑ co je to klasifikace?
- ☑ klasifikace vs. predikce vs. diskriminační analýza
- ☑ základní principy klasifikace
- ☑ parametrická vs. neparametrická klasifikace

Příprava nových učebních materiálů pro obor Matematická biologie

byla podporována projektem ESF

č. CZ.1.07/2.2.00/07.0318

„VÍCEOBOROVÁ INOVACE STUDIA MATEMATICKÉ BIOLOGIE"

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ