algoritmus zvolit pro danou úlohu. Odpověď můžeme hledat na základě znalosti silných a slabých stránek jednotlivých algoritmů nebo experimentálně. Mezi známé charakteristiky algoritmů, které můžeme brát do úvahy, patří například:

- rozdíl mezi způsobem reprezentace příkladů (hodnoty atributů nebo relace),
- rozdíl mezi vyjadřovací silou jednotlivých algoritmů (rozhodovací stromy a pravidla rozdělují prostor atributů rovnoběžně s osami, neuronové sítě nebo diskriminační funkce naleznou i diagonální hranici mezi třídami),
- schopnost práce s numerickými atributy (některé algoritmy vyžadují pouze kategoriální atributy),
- schopnost práce se zašuměnými a chybějícími daty (různé způsoby náhrady nepoužitelné hodnoty),
- schopnost práce s maticí cen (ve fázi učení, ve fázi testování, vůbec ne),
- předpoklad nezávislosti mezi atributy (např. naivní bayesovský klasifikátor),
- ostrá nebo neostrá klasifikace (tj. zda odvozujeme jen indikátor třídy nebo i pravděpodobnost či váhu klasifikace).

K empirickým studiím vhodnosti jednotlivých algoritmů na různé typy dat patří dva rozsáhlé výzkumné projekty celoevropského rozsahu, *STATLOG* a *METAL*.

6.4.1 STATLOG

V letech 1991–1994 se v rámci výzkumného programu Evropského společenství řešil projekt[202], jehož cílem bylo komparativní testování a vyhodnocení různých učících se algoritmů na rozsáhlých aplikacích v oblasti klasifikace a predikce (Michie a kol., 1994). V rámci projektu bylo porovnáváno asi 20 různých algoritmů na zhruba 20 různých datových souborech. Cílem bylo zjistit, na jaké typy dat se hodí ten který algoritmus. Jednotlivé datové soubory byly popsány souborem[203] jednoduchých charakteristik, statistických charakteristik a charakteristik z oblasti teorie informace. Vznikla tak jakási metadata (data o datech), která byla použita při následné analýze.

Pro každý datový soubor se (běžnými metodami testování) zjistilo, jaké dávají jednotlivé algoritmy výsledky (správnost, chyba). Byla tedy k dispozici informace, které algoritmy jsou na daná data vhodné (dávají malou chybu) a které jsou nevhodné. Tato informace byla (jakožto cílový atribut) přidána k metadatům; vznikla tak trénovací množina, na níž byl nasazen algoritmus

[202] ESPRIT projekt č. 5170, známý pod akronymem STATLOG.

[203] Mezi jednoduché charkteristiky patří počet příkladů, počet atributů, počet tříd, počet binárních atributů. Mezi statistické charakteristiky patří parametry rozdělení dat, jako je šikmost, špičatost nebo různé korelační charakterisitky (střední korelace atributů, první kanonická korelace). K charakteristikám z teorie informace patří entropie třídy, střední entropie atributů, střední vzájemná informace třídy a atributů nebo poměr signálu a šumu.

C4.5[204]. Výsledkem byla řada pravidel typu: „IF charakteristiky dat jsou CH THEN použij algoritmus A“. Příkladem může být pravidlo

```
IF počet příkladů ≤ 6435 AND šikmost > 0.57 THEN použij CART.
```

K souhrnným výsledkům patří to, že:

- pro rozsáhlá data se hodí diskriminační analýza (lineární, kvadratická),
- není velký rozdíl mezi „obyčejnou“ a logistickou diskriminační analýzou,
- na rozsáhlých datech je nejpomalejší metoda *k*-nejbližších sousedů,
- použité algoritmy na tvorbu rozhodovacích stromů se chovaly zhruba stejně; nezdá se tedy, že by nějak zvlášť záleželo na kritériu pro volbu větvení,
- neuronové sítě dávaly výborné výsledky u dat, kde se nepoužívala matice cen.

Řešitelé projektu STATLOG správně upozorňují na různé možné příčiny odlišností v správnosti jednotlivých algoritmů na daných datech, jimiž jsou:

- různá vhodnost algoritmů jako takových,
- různě vhodná implicitní nastavení parametrů jednotlivých modelů,
- různá zkušenost uživatelů s laděním parametrů,
- vliv předzpracování dat (např. diskretizace do různých intervalů).

6.4.2 METAL

Projekt METAL (metalearning) volně navazuje na projekt STATLOG. Tento projekt, řešený jako ESPRIT projekt č. 26357 v rámci 5. programu, byl zahájen v roce 2000. Cílem projektu je pomoci běžnému uživateli s volbou vhodné metody strojového učení i s volbou metod pro předzpracování dat.

V současné době se práce zaměřily na otázku řazení klasifikátorů. Zatím je navrženo několik metod využívajících správnosti a doby výpočtu pro hodnocení klasifikátorů testovaných na různých datech (Brazdil, Soares, 2000), a to:

- průměrné pořadí (na základě uspořádání všech klasifikátorů pro jednotlivé datové soubory),
- poměr úspěšnosti (na základě vyhodnocení dvojic klasifikátorů),
- významní vítězové (na základě vyhodnocení dvojic klasifikátorů).

6.5 Kombinování modelů

Možností jak zlepšit výsledky dosažené jednotlivými modely je jejich vzájemná kombinace. Nejběžnějším způsobem je kombinovat rozhodnutí z více modelů do

[204] Důvodem volby tohoto algoritmu byla snadná interpretovatelnost vytvořených rozhodovacích stromů i skutečnost, že C4.5 si v testech na vlastních datech vedl poměrně dobře.