8. Systémy pro dobývání znalostí z databází

Jako v jiných oblastech umělé inteligence, tak i v oblasti strojového učení se první programové systémy objevily v akademické sféře. Obvykle se jednalo o systémy, ve kterých se kladl důraz na implementování vlastního algoritmu; uživatelská přátelskost stála na pokraji zájmu. Přesto tyto systémy výrazně ovlivnily vývoj celé disciplíny. Připomeňme v této sovislosti alespoň Quinlanův *C4.5* [Quinlan, 1993] nebo *CN2* Clarka a Nibbleta [Clark, Nibblet, 1989]. Systémy pro dobývání znalostí z databází navazují tedy jednak na tuto linii, často přebírají úspěšné algoritmy. Druhou oblastí inspirace se staly velké balíky statistického software obsahující desítky metod pro analýzu dat i moduly pro transformaci dat. Aby systémy prorazily ke koncovému uživateli, dostaly (ve srovnání s programy pro strojové učení) přívětivější podobu.

Systémy pro dobývání znalostí z databází tedy

- pokrývají celý proces dobývání znalostí (od předzpracování po interpretaci),
- nabízejí více algoritmů pro analýzu (než „jednoúčelové" systémy strojového učení),
- kladou důraz na vizualizaci (ve způsobu práce se systémem i při interpretaci výsledků).

Tyto systémy lze rozdělit zhruba na výzkumné a komerční ([Siebes, 2000]). V každé této skupině můžeme opět nalézt dva typy systémů: zaměřené na dobývání znalostí obecně (tzv. horizontální) a zaměřené na nějakou konkretní aplikační oblast (tzv. vertikální). Zde uváděný přehled je zaměřen na obecné systémy pro dobývání znalostí komerční i výzkumné.

Systémy pro dobývání znalostí nabízejí jak malé firmy vzešlé z akademického prostředí (RuleQuest nebo Dialogis), tak význační producenti statistického software (SAS, SPSS nebo StatSoft). O rostoucím zájmu o dobývání znalostí z databází svědčí fakt, že mezi výrobce softwaru se zařadily i firmy jako IBM nebo Silicon Graphics. Na druhé straně existují akademické, volně šířené systémy; k nejznámějším z nich patří systém Weka. Tab. 1 uvádí některé systémy pro dobývání znalostí z dat. Podrobný seznam systémů je možno nalézt např. na http://www.kdnuggets.com.

Systém	***Výrobce***	***URL***
CART	Salford Systems	http://www.salford-systems.com
Clementine	SPSS	http://www.spss.com
Enterprise Miner	SAS Institute	http://www.sas.com/software/components/miner.html
Intelligent Miner	IBM	http://www-4.ibm.com/software/data/iminer
Kepler	Dialogis	http://www.dialogis.de
KnowledgeStudio	Angoss	http://www.angoss.com
LISp Miner	VŠE	http://lispminer.vse.cz
MineSet	Silicon Graphics	http://www-europe.sgi.com/software/mineset
See5	RuleQuest Research	http://www.rulequest.com/see5-info.html
Statistica Data Miner	StatSoft	http://www.statsoft.com
Weka	University of Waikato	http://www.cs.waikato.ac.nz/~ml/weka
WizWhy	WizSoft	http://www.wizsoft.com/why.html

Tab. 1 Systémy pro dobývání znalostí z databází

CART je implementace známého algoritmu „Classification And Regression Trees" popsaného v pol. 80.let [Bierman a kol, 1984] *See5* je uživatelsky přátelštější verze Quinlanova *C4.5* [Quinlan, 1993] pro tvorbu rozhodovacích stromů a pravidel. *WizWhy* nabízí sice jen algoritmus pro tvorbu pravidel, ale klade velký důraz na vizualizaci. Ostatní produkty jsou již plnohodnotné systémy pro dobývání znalostí z databází ve smyslu výše uvedených tří bodů a proto se na ně podíváme trochu podrobněji. V případě systémů *Enterprise Miner*, *Intelligent Miner* a *MineSet* se opírám o dostupné prameny, se systémy *Clementine*, *Kepler, KnowledgeStudio*, *Statistica Data Miner* a *Weka* máme praktické zkušenosti. *LISp-Miner* je náš vlastní příspěvek k plejádě systémů pro dobývání znalostí; tento systém nabízí algoritmy pro tvorbu asociačních a rozhodovacích pravidel.

8.1 Clementine

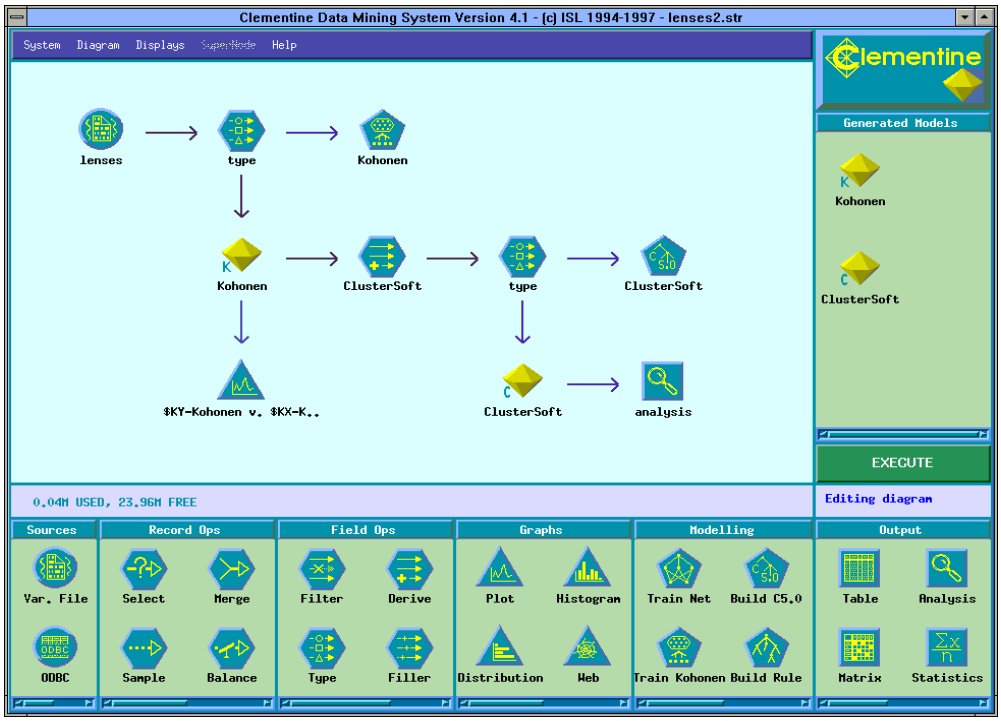
Systém *Clementine* vyvinula britská firma Integral Solutions Ltd. v polovině 90.let. K 1.lednu 1999 tuto firmu (a s ní i systém *Clementine*) převzal přední výrobce statistického software, firma SPSS. Clementine patří mezi přední komerční systémy pro dobývání znalostí. Systém důsledně vychází z metodologie *CRISP-DM* [1]. Systém nabízí řadu metod pro klasifikační (predikční) i deskriptivní úlohy, mimo jiné standardní algoritmy C4.5, apriori, vícevrstvý perceptron, nebo lineární regresi.

Clementine má velice propracovaný způsob ovládání, tzv. vizuální programování (vizual programming). Z nástrojů v jednotlivých paletách (tyto palety odpovídají jednotlivým krokům procesu dobývání znalostí; předzpracování, modelování, vizualizace a interpretace) se na pracovní ploše poskládá sekvence řešení úlohy (stream). Obr. 2 ukazuje úlohu, ve které je nejprve vytvořeno několik klasifikačních modelů pro stejná data (horní část sekvence kroků). Všechny modely (žluté "diamanty") jsou pak testovány na datech a výsledky testování jsou analyzovány (dolní část sekvence, viz též). Obr. 1 ukazuje úlohu, ve které jsou nejprve hledány zajímavé skupiny příkladů (deskriptivní úloha řešená pomocí Kohonenovy mapy), pro jednu ze skupin je pak vytvořen model umožňující klasifikaci (rozhodovací strom).

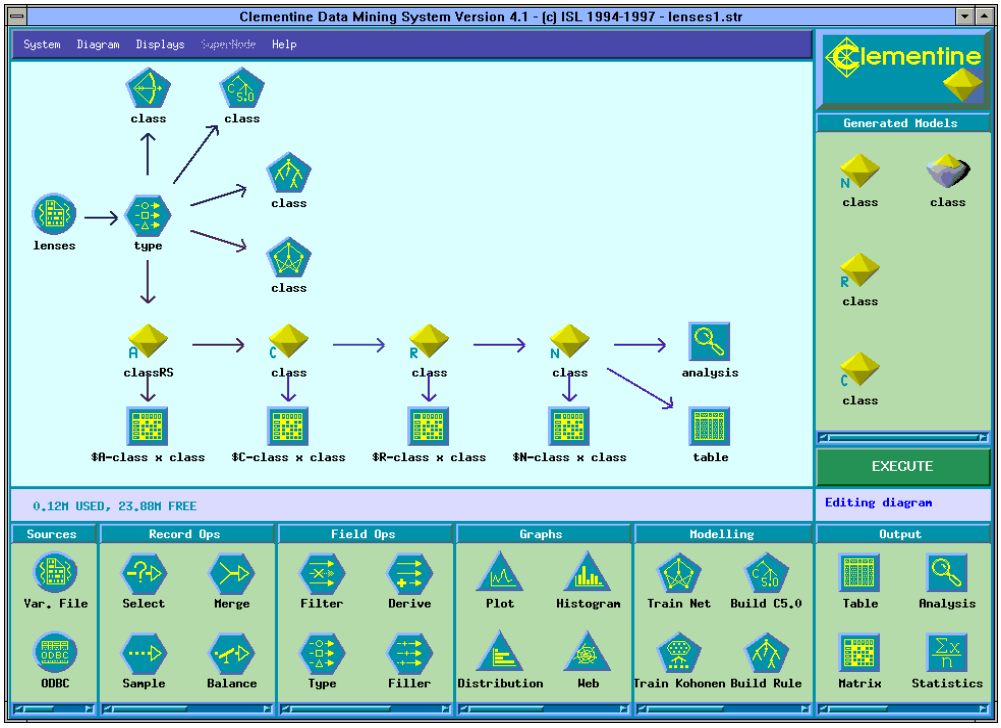
Systém *Clementine* nabízí analytikům tzv. Clementine External Module Interface pro přidávání vlastních algoritmů. Tak lze přidávat programy do jednotlivých palet nástrojů. Z implementačního hlediska je třeba zajistit správné začlenění nového programu do sekvence (streamu). Program tedy musí umět číst data a parametry ze sekvence a zapisovat do sekvence své výsledky.

Přenesení provedené analýzy ke koncovým uživatelům usnadňuje tzv. Clementine Solution Publisher, na jehož základě vznikne samostatná aplikace obsahující všechny provedené kroky v sekvenci (jako zdrojové programy v C, SQL příkazy apod.).

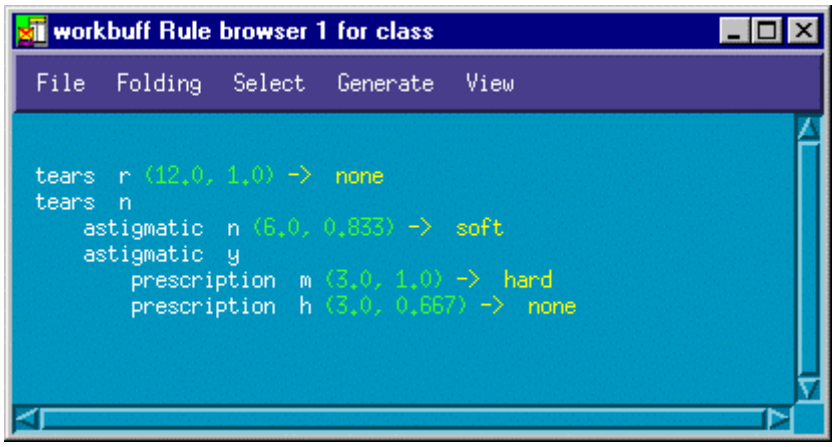
[1] Firma ISL se podílela na návrhu této metodologie.



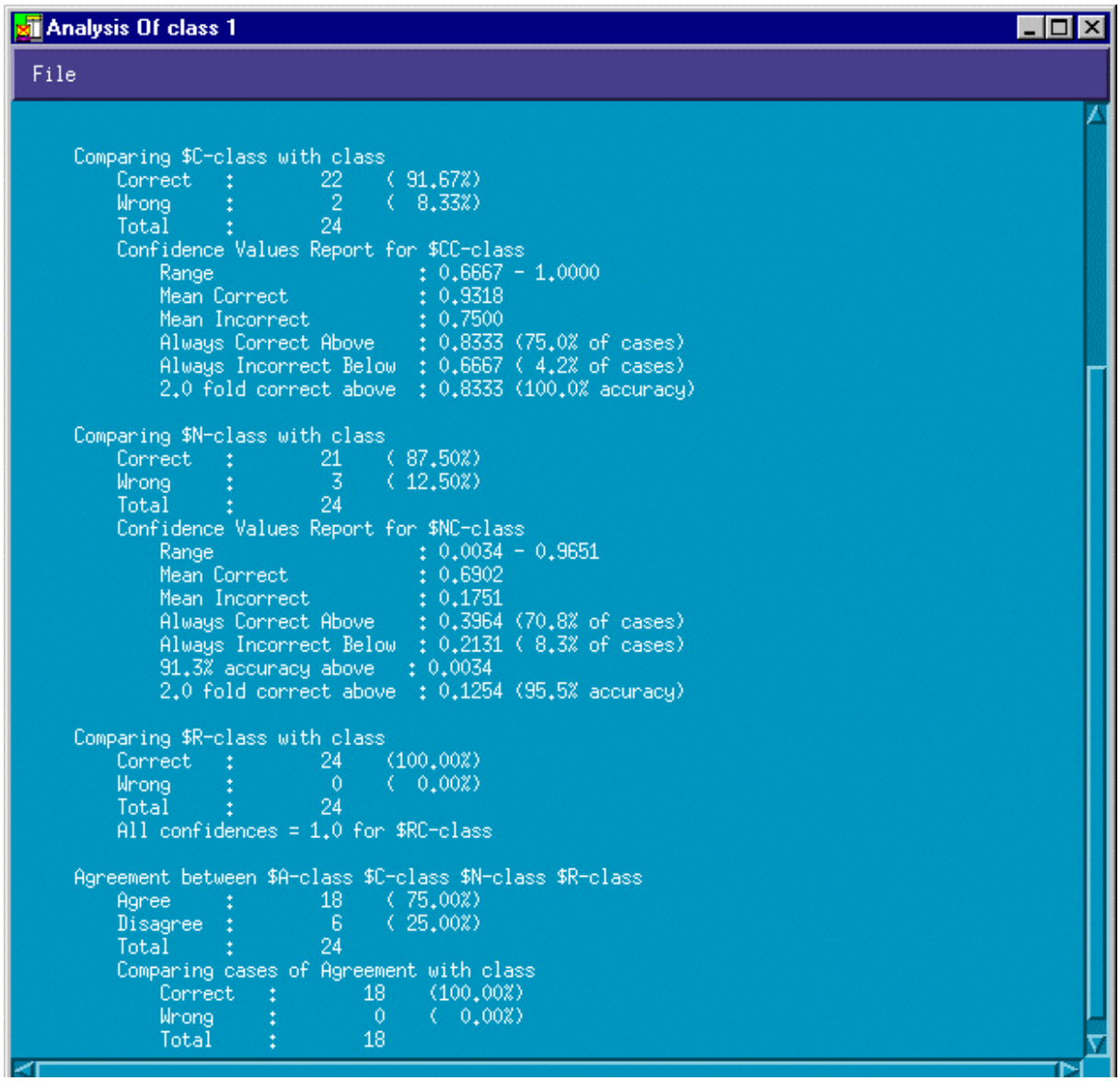
Obr. 1 Deskripce a klasifikace



Obr. 2 Více klasifikačních modelů



Obr. 3 Rozhodovací strom



Obr. 4 Porovnání klasifikátorů

8.2 Enterprise Miner

Enterprise Miner je produkt firmy SAS, jednoho z předních producentů statistického software. To je vidět i na algoritmech pro analýzu dat, které systém nabízí. Nejpropracovanější jsou statistické metody, které využívají již implementované procedury. *Enterprise Miner* použité metody integruje a nabízí uživatelsky příjemnější prostředí než je příkazový jazyk (SAS kód) jinak běžně používaný v „klasickém" SASu.

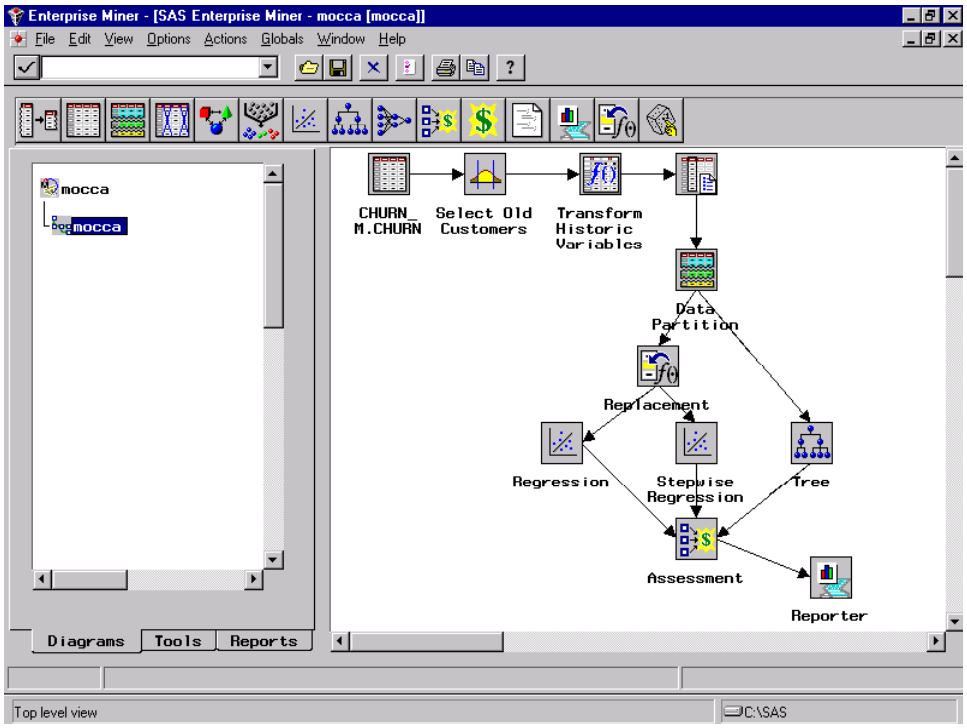
Enterprise Miner vychází z vlastní metodologie pro dobývání znalostí z databází. Název této metodologie SEMMA je akronym pro jednotlivé prováděné kroky:

- *Sample* (vybrání vhodných objektů),
- *Explore* (vizuální explorace a redukce dat),
- *Manipulate* (seskupování objektů a hodnot atributů, datové transformace),
- *Model* (analýza dat: neuronové sítě, rozhodovací stromy, statistické techniky, asociace a shlukování),
- *Assess* (porovnání modelů a interpretace).

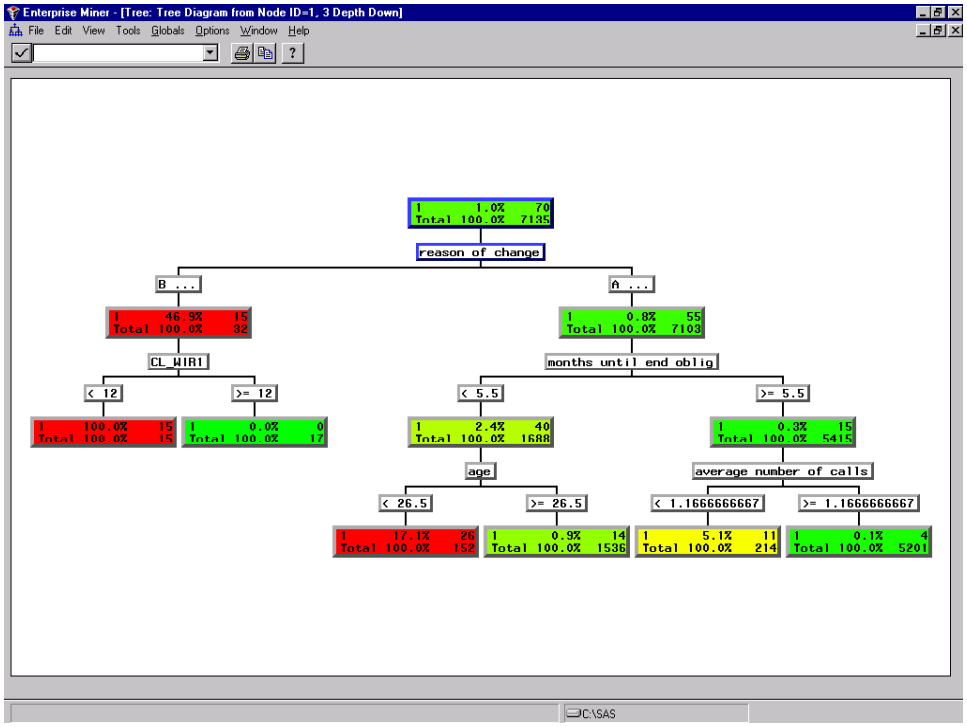
Důraz se klade na snadnou interpretaci výstupů ve formě srozumitelné obchodnímu uživateli.

Proces dobývání znalostí pro danou úlohu se definuje („programuje") pomocí procesních diagramů (Process Flow Diagrams). Jde vlastně o analogii k vizuálnímu programování použitém v systému *Clementine*.

..



Obr. 5 Procesní diagram



Obr. 6 Rozhodovací strom

8.3 Intelligent Miner

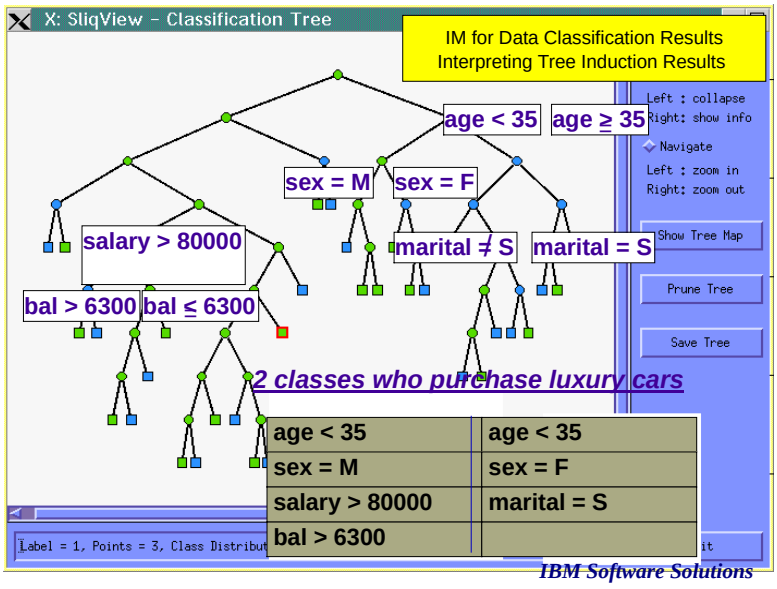
Intelligent Miner for Data a *Intelligent Miner for Text* jsou dva produkty firmy IBM, které pomáhají zákazníkům identifikovat a extrahovat cenné obchodní informace z databází a datových skladů

Jádrem produktu *Intelligent Miner* jsou algoritmy pro získávání dat vyvinuté ve výzkumných laboratořích společnosti IBM a prověřené v zákaznických instalacích po celém světě. Tyto algoritmy lze aplikovat na širokou škálu obchodních problémů. Mohou být uplatněny při rozhodování v takových oblastech, jako je plánování reklamních kampaní, správa vztahů se zákazníky, zpracování a revize procesů, plánování a sledování produkce nebo ochrana proti podvodům a zneužitím.

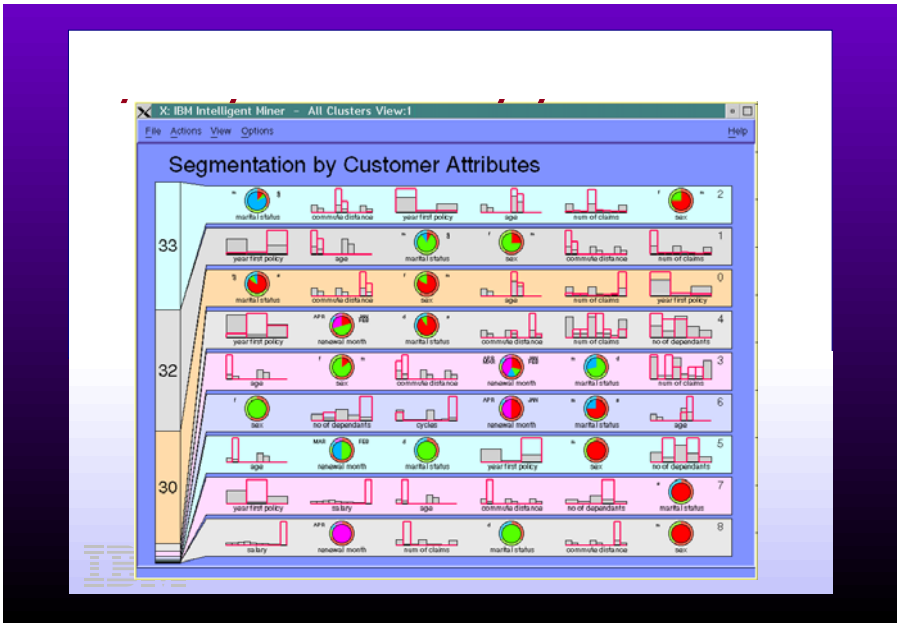
Intelligent Miner nabízí následující metody:

- vytváření klasifikačních a predikčních modelů,
- hledání vazeb a sekvenčních vzorů ve velkých databázích,
- automatická segmentace databází do skupin souvisejících záznamů.

Systém je „šit na míru" pro spolupráci s databází DB2, umožňuje ale samozřejmě používat i jiné databáze pro ukládání dat. Systém je implementován pro platformy UNIX, OS/400 a Windows.



Obr. 7 Rozhodovací strom

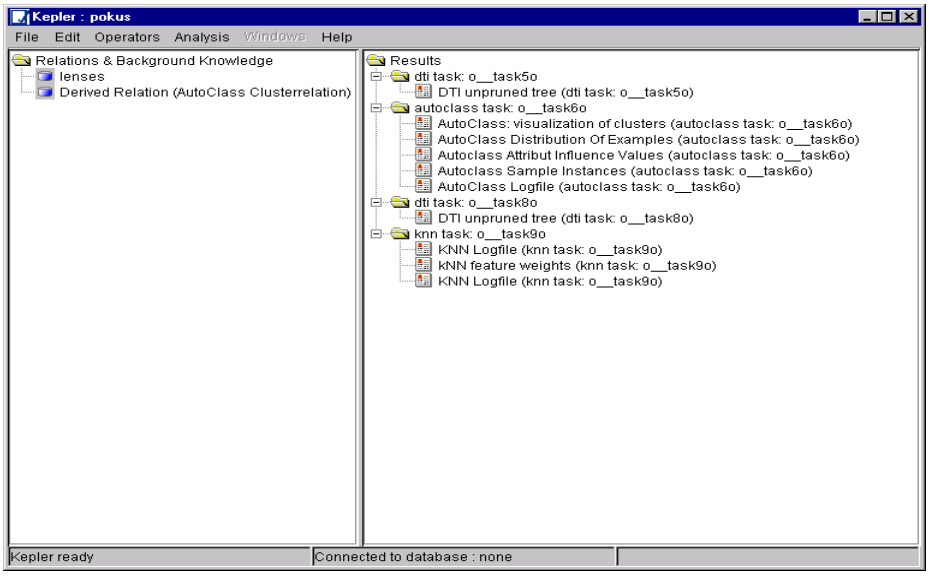


Obr. 8 Shlukování

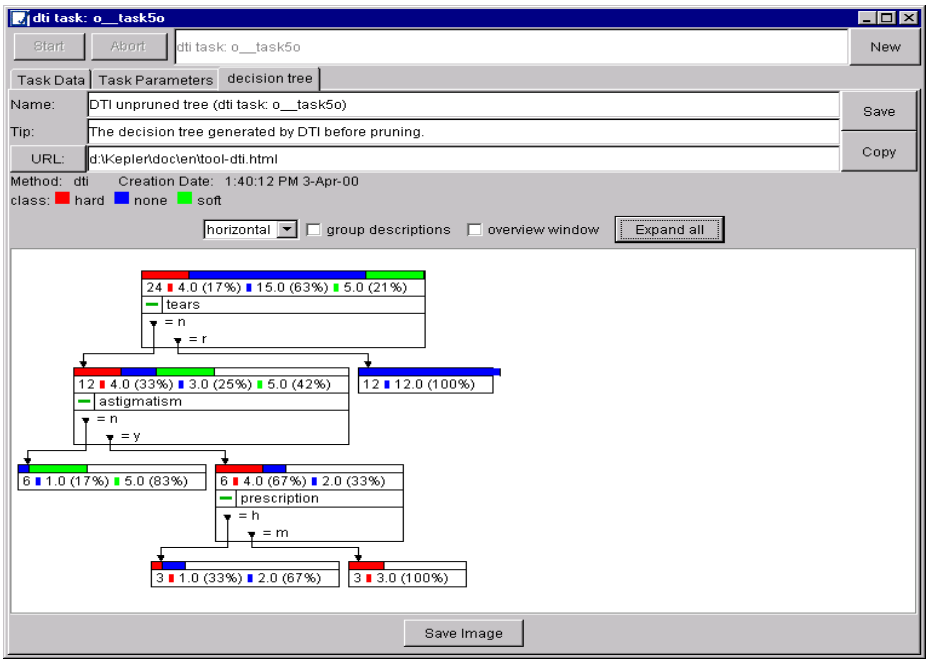
8.4 Kepler

Kepler je komerční systém německé firmy Dialogis. Vývoj tohoto systému byl zahájen na akademické půdě, v německém výzkumném ústavu informačních technologií (GMD). Proto *Kepler* obsahuje i méně běžné metody z oblasti induktivního logického programování (ILP), nabízí rovněž i některé algoritmy předzpracování (diskretizace) známé v akademickém světě. Systém je otevřený, to znamená, že ho lze rozšiřovat o další algoritmy.

Při práci se systémem se uchovává posloupnost provedených kroků (Obr. 9). Jednotlivé algoritmy učení nabízejí "Keplerovský", grafický výstup modelů (nově implementované algoritmy, viz Obr. 10), nebo původní, často textový výstup (algoritmy převzaté, např. C5.0).



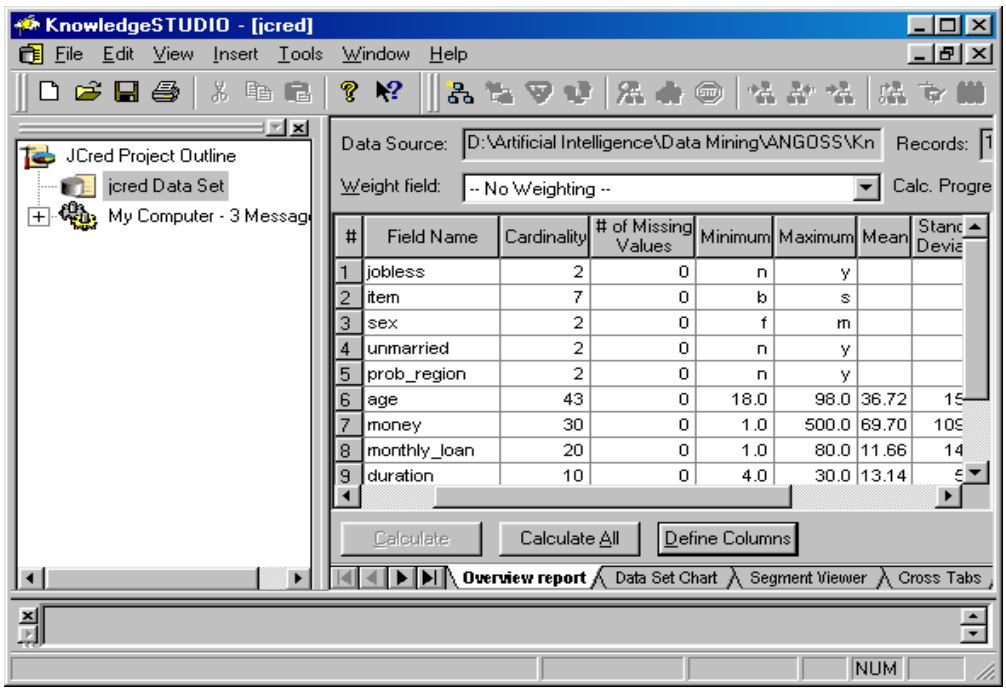
Obr. 9 Schéma kroků



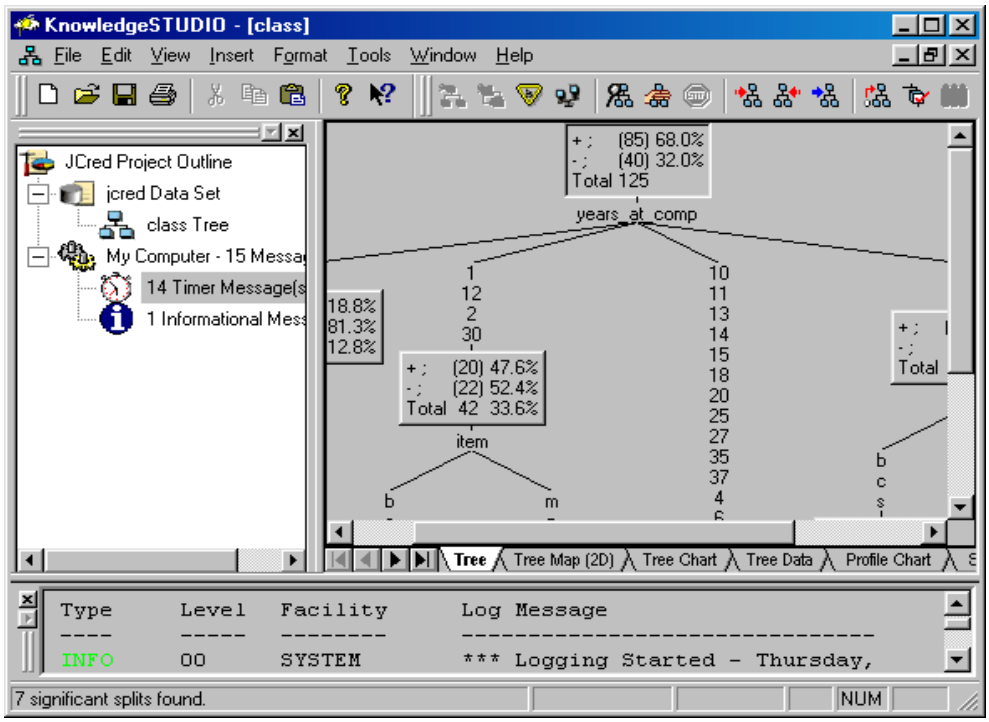
Obr. 10 Rozhodovací strom

8.5 KnowledgeSTUDIO

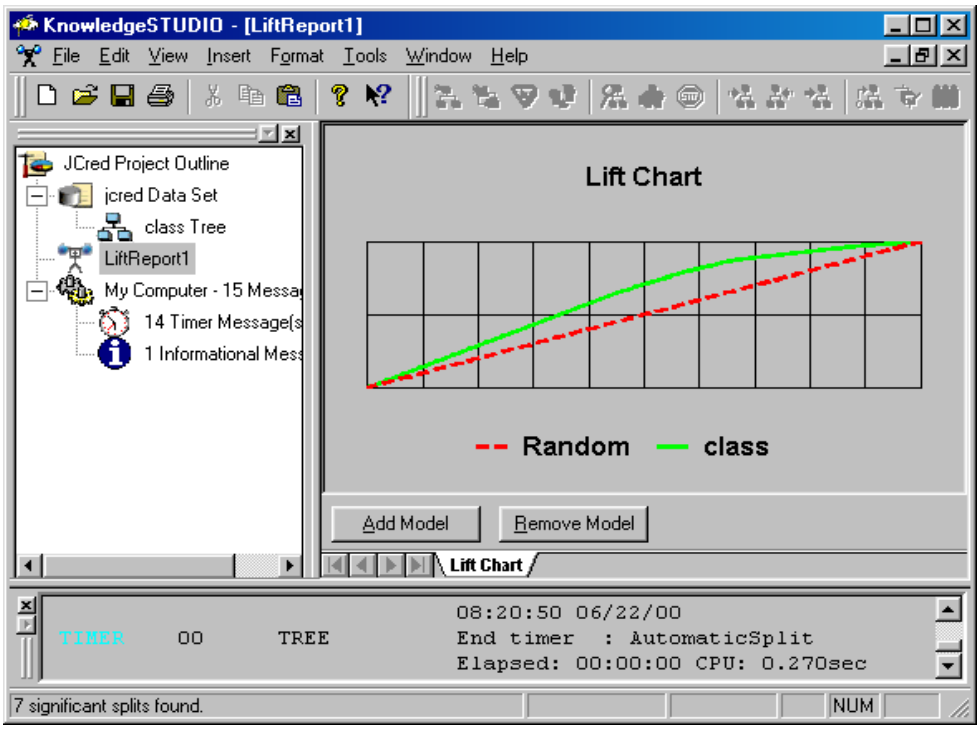
KnowledgeSTUDIO kanadské firmy Angoss je, ve srovnání s předcházejícími, relativně méně známým systémem. Základními metodami jsou rozhodovací stromy (zde firma navazuje na úspěšný systém *KnowledgeSeeker*), regrese, neuronové sítě a shlukování. Vytvořené rozhodovací stromy lze převést do spustitelného kódu (java, SAS). Propracovaná je i část interpretace (Obr. 13).



Obr. 11 Práce s atributy



Obr. 12 Rozhodovací strom



Obr. 13 Křivka návratnosti

8.6 LISP-Miner

LISp-Miner je systém vyvíjený na VŠE v Praze [Rauch, 2001]. Systém v současnosti nabízí dvě základní metody; tvorbu asociačních pravidel a tvorbu klasifikačních pravidel. Co se týká asociačních pravidel, systém navazuje na mnohaletý výzkum v této oblasti spojený s metodou *GUHA* [Hájek, Havránek, 1978], konkrétně jde o *GUHA* proceduru *4FT-Miner* [Šimůnek, Rauch, 1999]. Algoritmus pro tvorbu klasifikačních pravidel byl převzat ze systému *KEX*, rovněž vyvinutého na VŠE [Ivánek, Stejskal, 1988], [Berka, Ivánek, 1994]. Oba algoritmy již byly podrobně popsány v předcházejících kapitolách, zde jen připomeňme, že asociační pravidla mají podobu

$$\text{Ant} \sim \text{Suc} / \text{Cond},$$

a klasifikační pravidla mají podobu

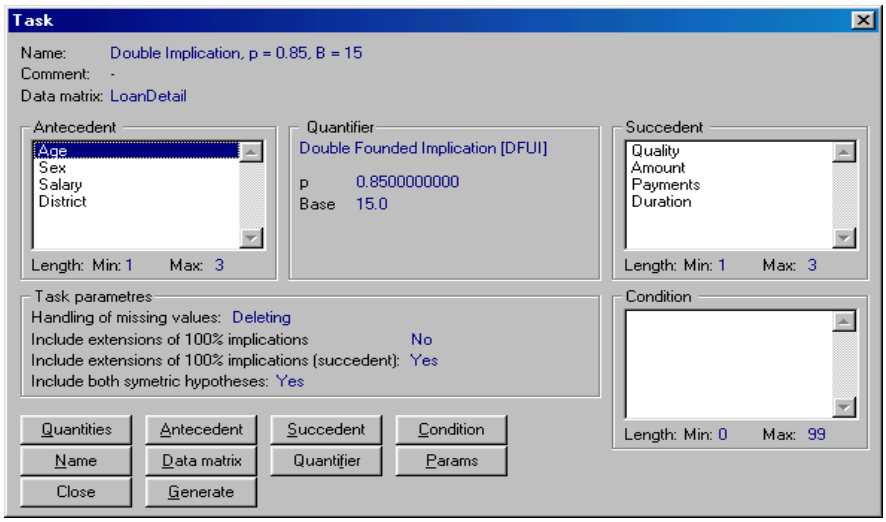
$$\text{Ant} \Rightarrow \text{Class (w)}.$$

Systém rovněž nabízí bohaté možnosti pro předzpracování založené na SQL. Systém je totiž úzce svázán s databázemi (současná implementace s databází MS Access), ze kterých načítá data i kam ukládá výsledky. Posledně zmiňovaný rys umožňuje uživatelům vytvářet vlastní interpretační procedury nad asociačními pravidly (tzv. open 4FT Miner).

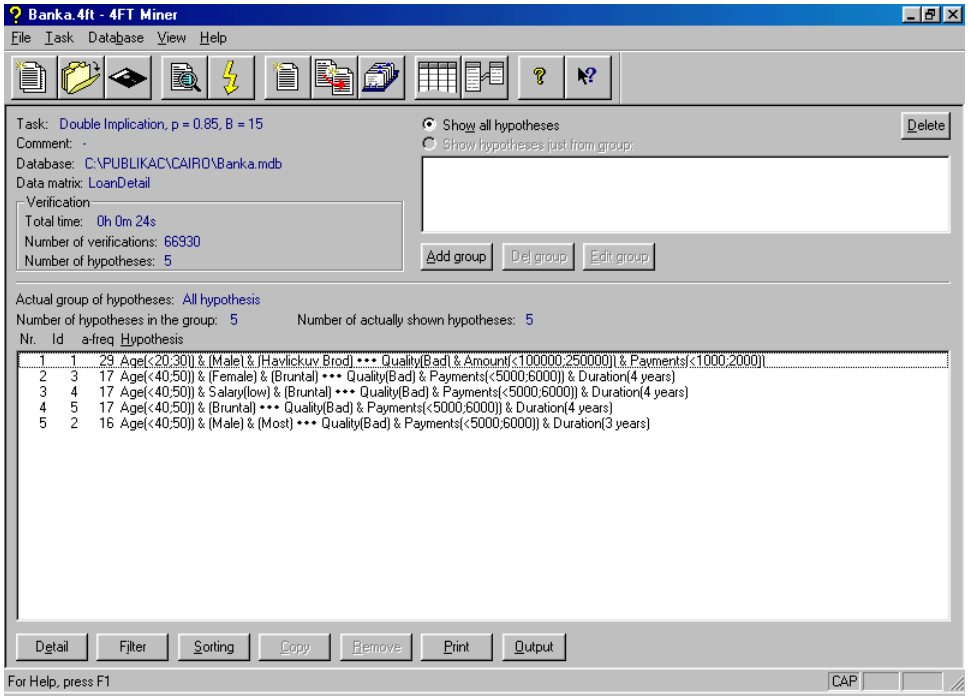
Následující obrázky ukazují způsob zadávání relevantních otázek pro *4FT-Miner* (Obr. 14), seznam nalezených hypotéz (Obr. 15), bázi znalostí vytvořenou pro *KEX* (Obr. 16)[2], detailní pohled na jedno pravidlo vytvořené KEXem (Obr. 17)[3] a příklad výsledku konzultace s bází znalostí (Obr. 18).

[2] Zadávání parametrů pro KEX je analogické k zadávání parametrů pro 4FT-Miner.

[3] Stejným způsobem se zobrazují i hypotézy nalezené 4FT-Minerem.



Obr. 14 Vstup parametrů pro 4FT Miner



Obr. 15 Nalezené hypotézy

LISp-Miner DB Data - LISp-Miner KEX Result module

Datasource Task description Rules Consultation

Task: KEX JCred 2-1-0
Comment: -
Group of tasks: Default task-group
Data matrix: JCRED_kat

Verification
Total time: 0h 0m 13s
Number of verifications: 562
Number of rules: 31

Total number of rules: 31 Number of actually shown rules: 31

Delete all r

Nr.	Id	r-freq	Wg.	Rule
1	1	125	0.680	(Default rule) ==> JCred class(+)
2	2	101	0.615	JCred jobless(n) & JCred prob_region(n) ==> JCred class(+)
3	3	61	0.272	JCred years_at_comp(a) ==> JCred class(+)
4	4	55	0.679	JCred jobless(n) & JCred unmarried(n) ==> JCred class(+)
5	5	47	0.329	JCred deposit(a) ==> JCred class(+)
6	7	44	0.679	JCred age (b) & JCred jobless(n) ==> JCred class(+)
7	6	44	0.679	JCred deposit(b) & JCred jobless(n) ==> JCred class(+)
8	8	37	0.751	JCred age (c) ==> JCred class(+)
9	9	36	0.702	JCred sex(m) & JCred unmarried(n) ==> JCred class(+)
10	10	34	0.883	JCred years_at_comp(d) ==> JCred class(+)
11	11	29	0.305	JCred age (a) ==> JCred class(+)
12	13	25	0.693	JCred age (a) & JCred years_at_comp(a) ==> JCred class(+)
13	12	25	0.775	JCred item(m) & JCred jobless(n) ==> JCred class(+)
14	14	24	0.767	JCred deposit(b) & JCred sex(m) ==> JCred class(+)
15	15	23	0.832	JCred deposit(b) & JCred month_payment(a) ==> JCred class(+)
16	16	20	0.899	JCred deposit_covers_loan(y) & JCred jobless(n) ==> JCred class(+)
17	17	19	0.800	JCred deposit(b) & JCred item(j) ==> JCred class(+)
18	18	15	0.868	JCred item(m) & JCred unmarried(n) ==> JCred class(+)
19	20	14	0.073	JCred jobless(y) ==> JCred class(+)
20	19	14	0.207	JCred prob_region(y) ==> JCred class(+)
21	21	14	0.860	JCred deposit_covers_loan(y) & JCred sex(m) ==> JCred class(+)
22	22	14	0.860	JCred item(m) & JCred month_payment(b) ==> JCred class(+)
23	23	11	0.908	JCred deposit_covers_loan(y) & JCred month_payment(a) ==> JCred class(+)
24	24	11	0.908	JCred deposit_covers_loan(y) & JCred no_months(b) ==> JCred class(+)

Detail Go to ID Filter Sorting Output Print

Ready NUM

Obr. 16 KEX pravidla

Hypothesis

*** Hypothesis number: 13

Antecedent	JCred age	a
	JCred years_	a
Succedent	JCred class	+

Values from frequency table:

a/(a+b) =	0.440000	(Prob, Validity)
a/(a+b+c) =	0.111111	
a/n =	0.088000	
Weight =	0.692562	

Frequency table

	Succedent	NOT Succedent	
Antecedent	11	14	25
NOT Antecedent	74	26	100
	85	40	125

Row percentage

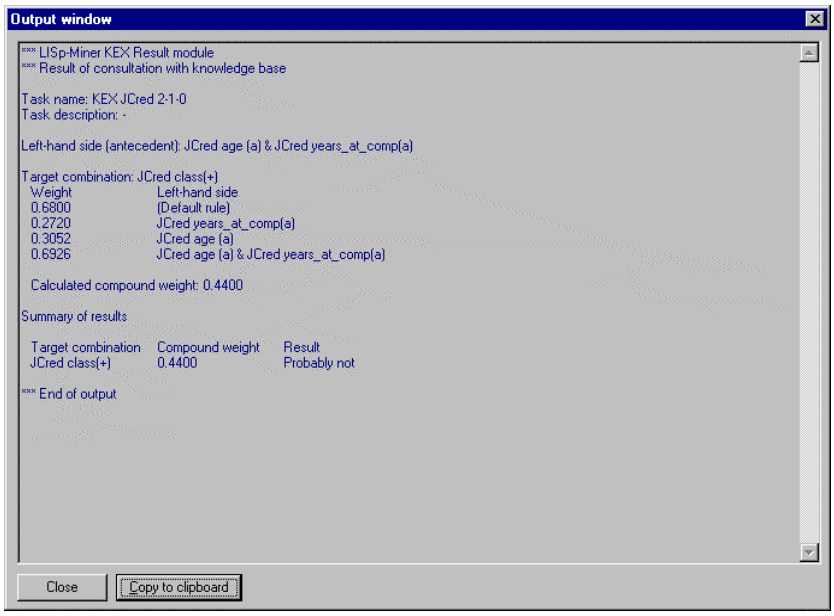
	Succedent	NOT Succedent	
Antecedent	44.00	56.00	100.00
NOT Antecedent	74.00	26.00	100.00
	68.00	32.00	100.00

Column percentage

	Succedent	NOT Succedent	
Antecedent	12.94	35.00	20.00
NOT Antecedent	87.06	65.00	80.00

Close Copy to clipboard

Obr. 17 KEX jedno pravidlo



Obr. 18 KEX konzultace

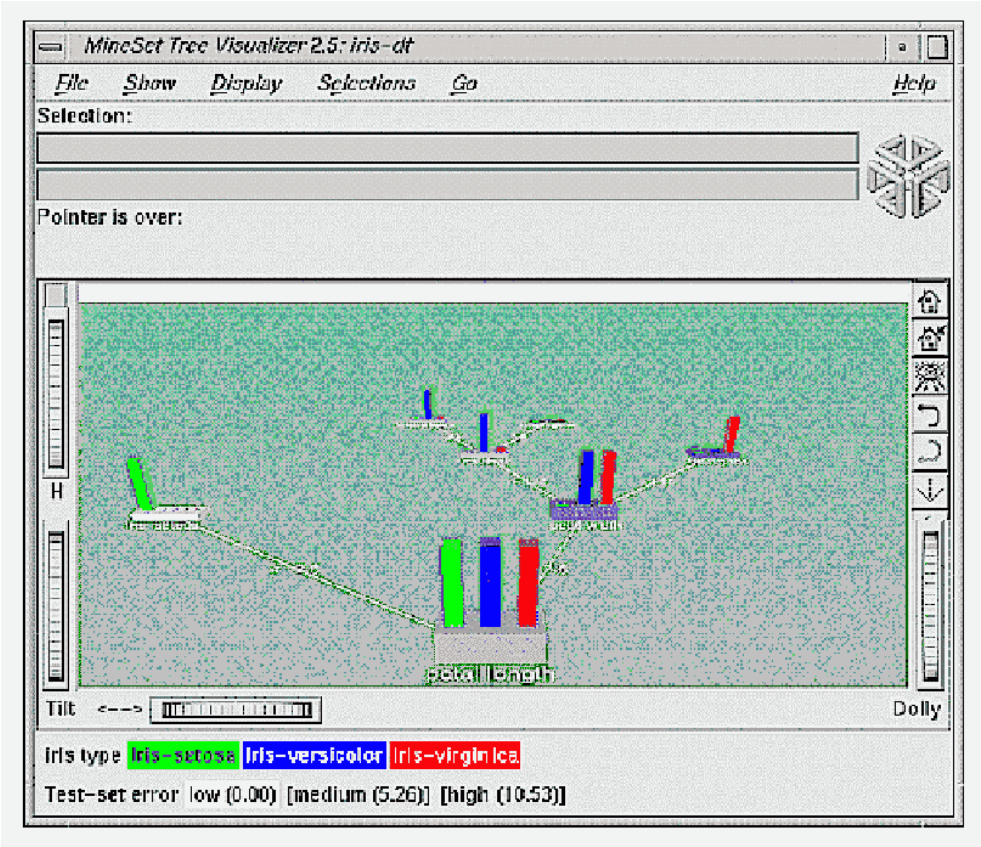
8.7 MineSet

Produkt *MineSet* uvedla firma Silicon Graphics (SGI) na trh v roce 1996 a od té doby je intenzivně vyvíjen. Systém navazuje na akademický projekt knihovny programů pro strojové učení *MLC++* řešený na universitě ve Stanfordu [Kohavi, 1994].

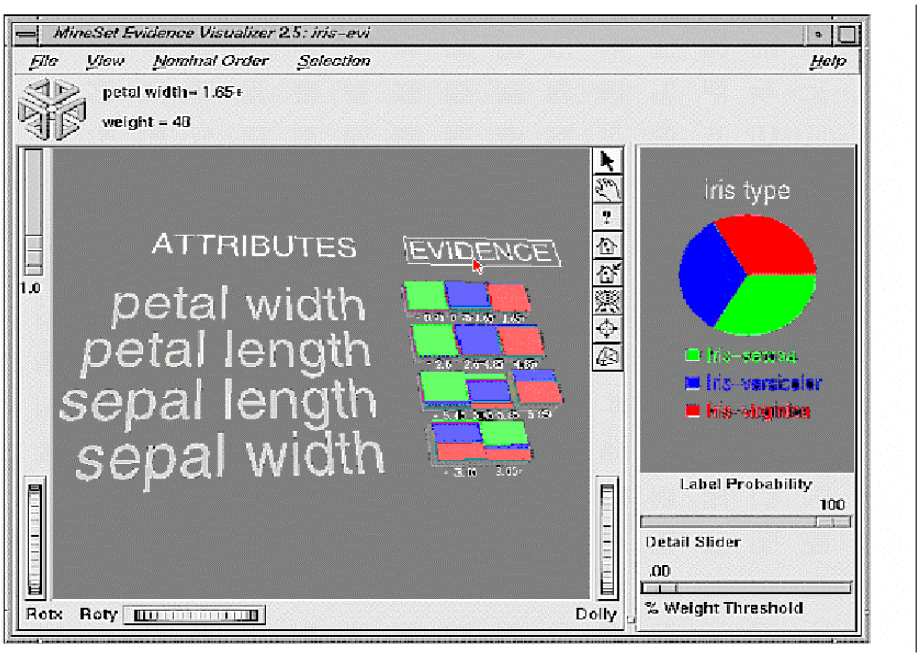
Programový systém *MineSet* je navržen pro splnění požadavků uživatelů na několika úrovních. Technickým uživatelům a analytikům nabízí sadu výkonných implementací metod dobývání znalostí založených na metodách strojového učení (rozhodovací stromy, asociační pravidla, regresní stromy, shluková analýza), prostředky pro spolupráci s databázemi a pro snadnou integraci do zákaznických řešení. Pro koncové uživatele a vlastníky dat přináší interaktivní vizualizační ástroje a prostředky pro sdílení výsledků v prostředí podnikových IS a intranetů. Právě vizualizace (včetně geografické vizualizace) je silnou stránkou tohoto produktu.

Kromě tradiční platformy IRIX bude *MineSet* dostupný i pro platformu Windows. Intenzivně se rovněž pracuje na společném projektu firem Silicon Graphics a MicroStrategy, jehož výsledkem bude první integrované řešení pro data mining a OLAP.

MineSet klade velký důraz na vizualizaci modelů. Je to ostatně vidět na obou ukázkách; Obr. 19 ukazuje rozhodovací strom (výška sloupců v uzlu odpovídá počtu objektů jednotlivých tříd), Obr. 20 ukazuje relativní četnosti hodnot atributů opět rozdělené podle tříd.



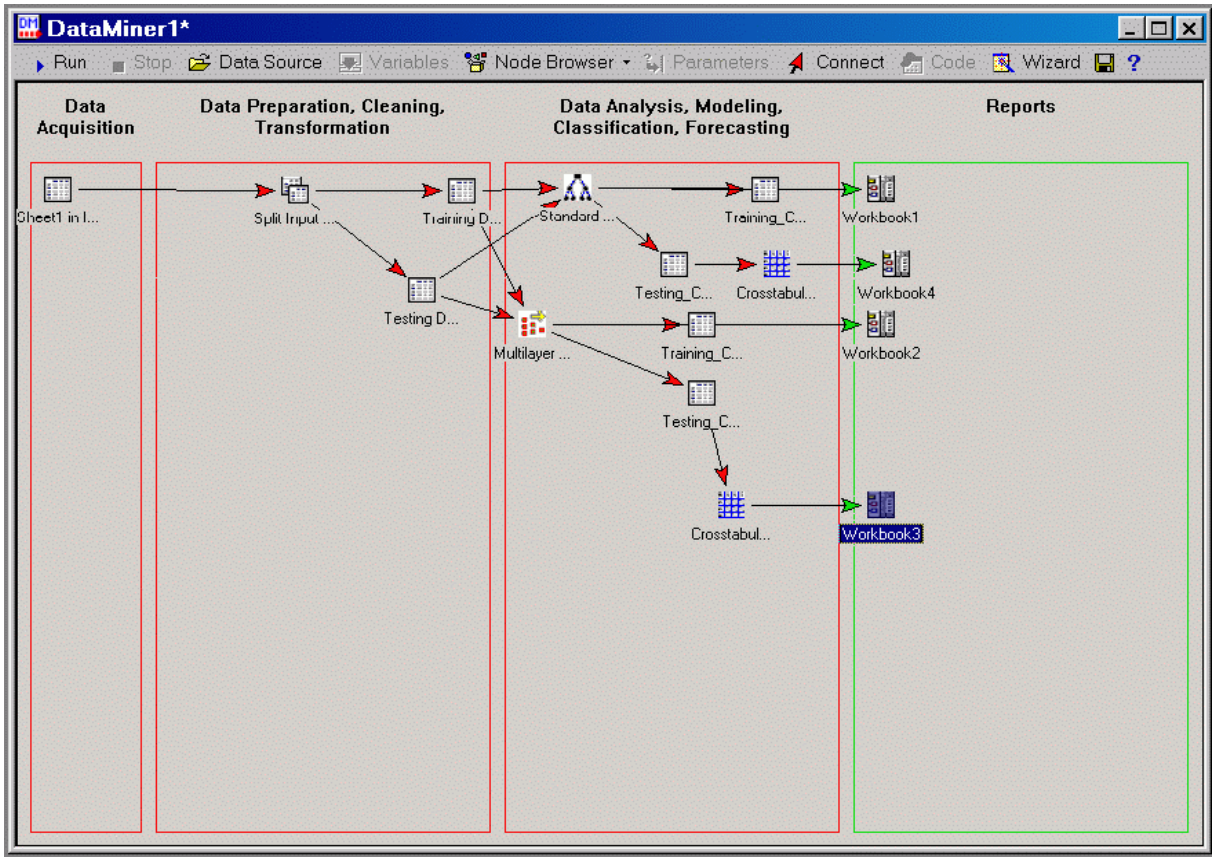
Obr. 19 Rozhodovací strom



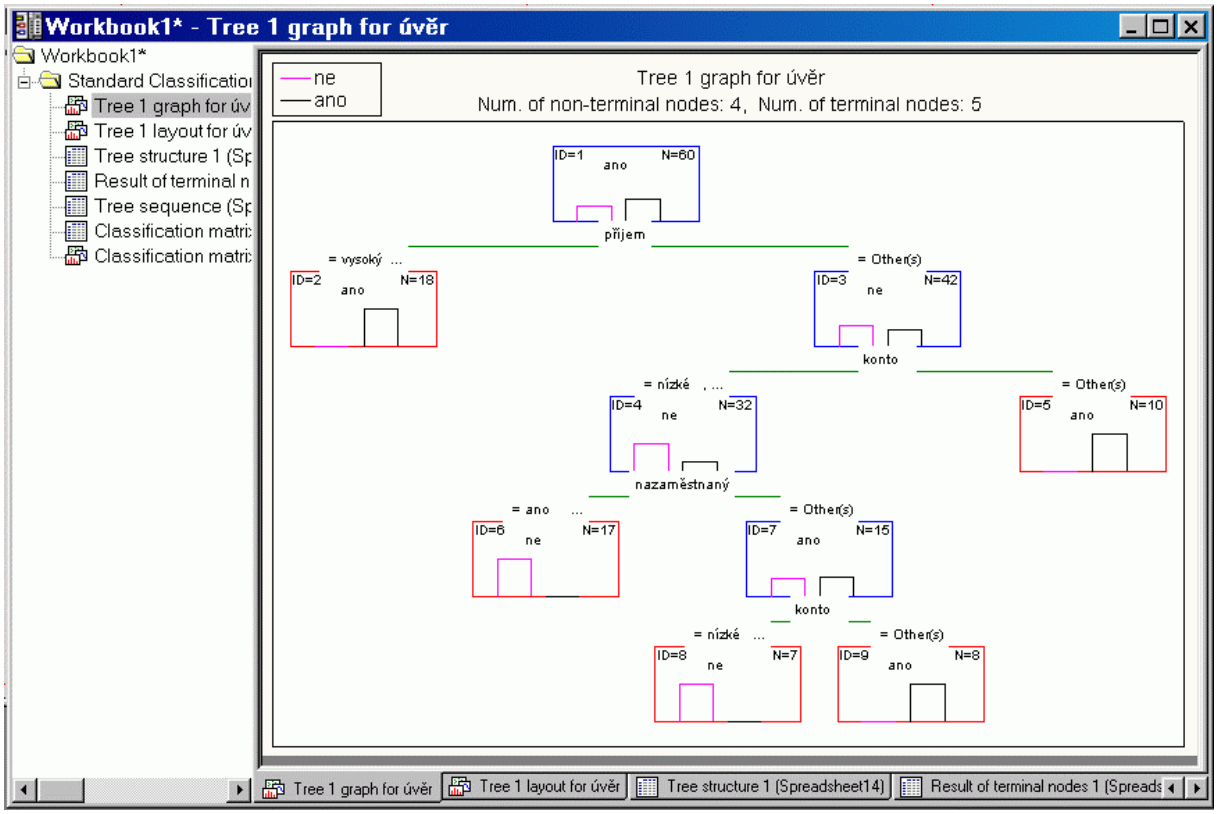
Obr. 20 Četnosti hodnot atributů

8.8 Statistica Data Miner

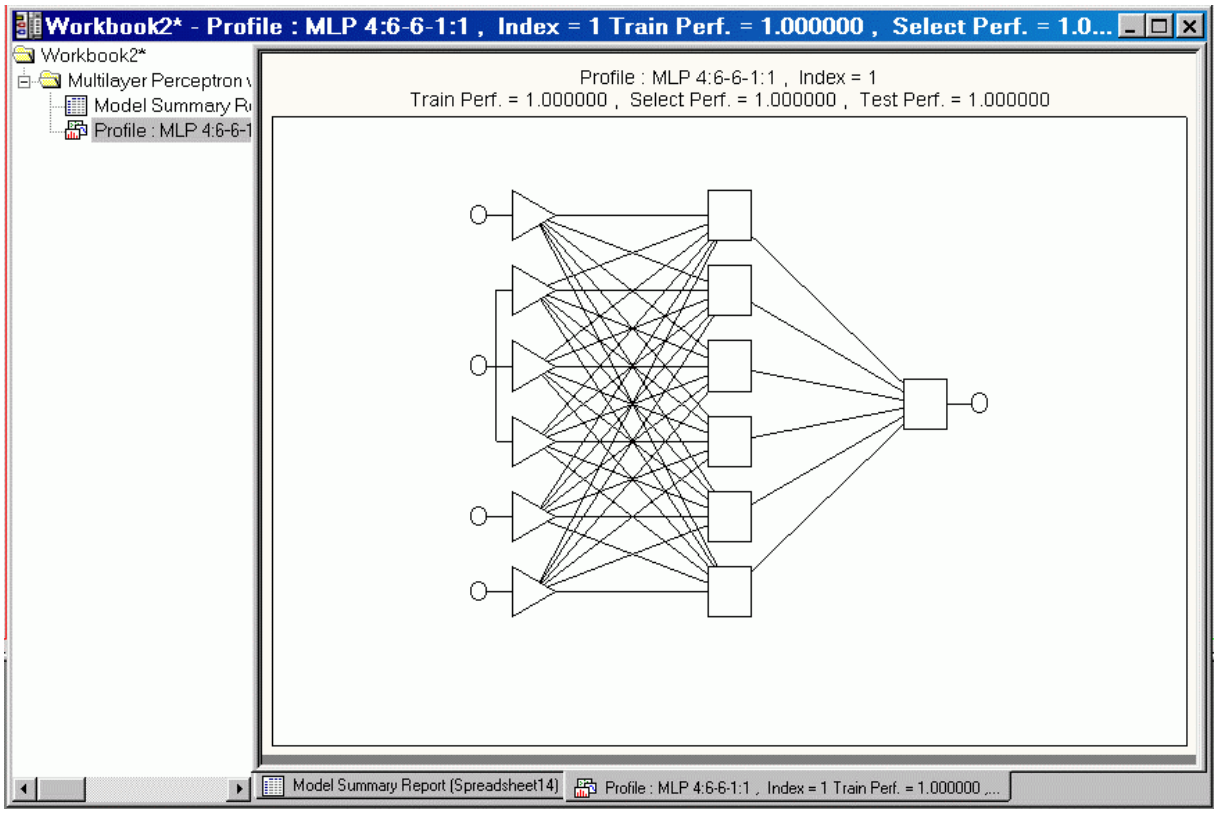
Statistica Data Miner je dalším příkladem systému pro dobývání znalostí, který vyvinula firma specializovaná na statistické programy. Systém opět pokrývá proces dobývání znalostí počínaje přípravou dat a využitím výsledků konče. Sekvence zpracování dat se vytvoří na pracovní ploše z jednotlivých nástrojů-uzlů (Obr. 21). Pro modelování nabízí *Data Miner* asociační pravidla, klasifikační a regresní stromy (Obr. 22), shlukování, neuronové sítě (mimo jiné vícevrstvý perceptron - Obr. 23 i Kohonenovu mapu), diskriminační analýzu a regresní metody. Z výsledných modelů (např. rozhodovacích stromů) lze vygenerovat spustitelný kód v C++, Visual Basicu nebo SQL. Visual Basic je ostatně přímo součástí systému. V systému najdeme i prvky OLAP a možnost kombinování modelů (bagging, boosting, metaučení). Otevřená architektura umožňuje uživatelům přidávat do systému vlastní nástroje (uzly).



Obr. 21Sekvence kroků



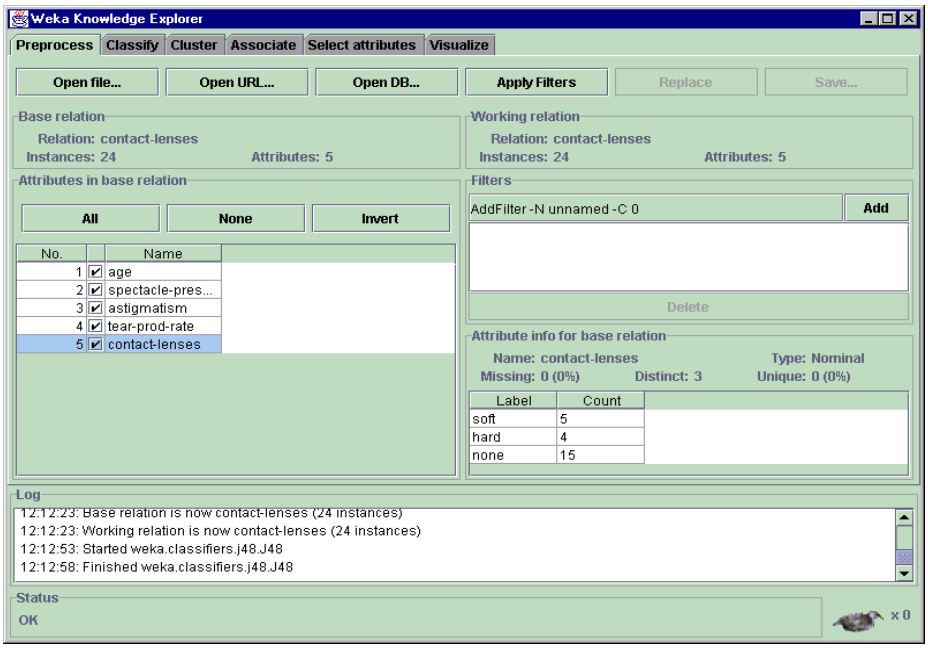
Obr. 22 Rozhodovací strom



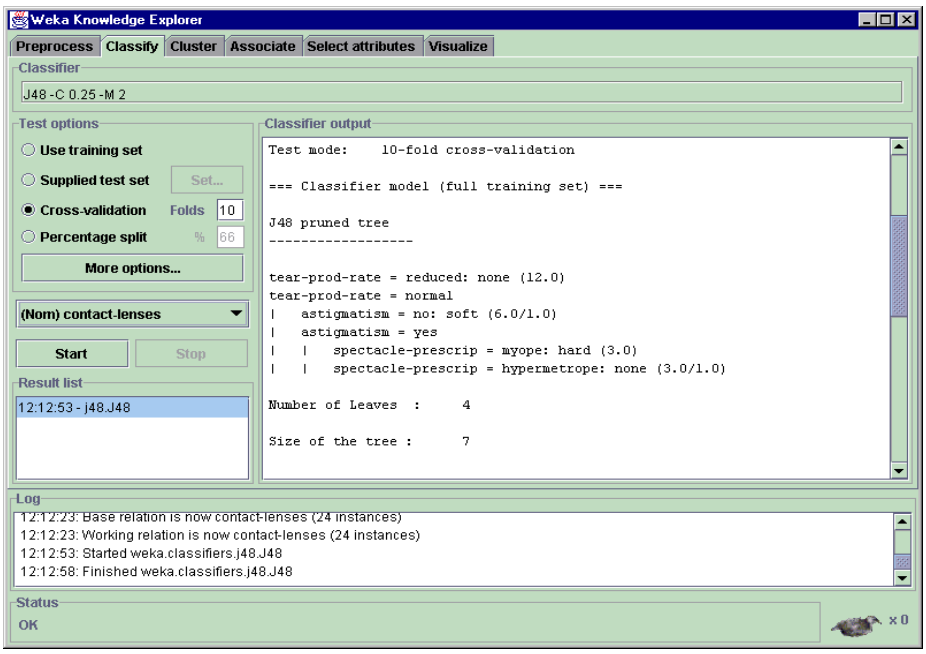
Obr. 23 Neuronová síť

8.9 Weka

Weka je systém vyvinutý na universitě Waikato na Novém Zélandě [Witten, Frank, 1999]. Přestože se jedné o freeware volně dostupný na Internetu[4], v ničem si nezadá s komerčními systémy. *Weka* nabízí celou řadu algoritmů pro učení i předzpracování, známých v akademickém světě. K dispozici jsou i možnosti vizualizace (Obr. 26) a kombinování modelů. Systém je řešen jako knihovna programů v Javě volaných z jednotného (grafického) rozhranní. Většina modelů si ale ponechává původní textový výstup (Obr. 25).

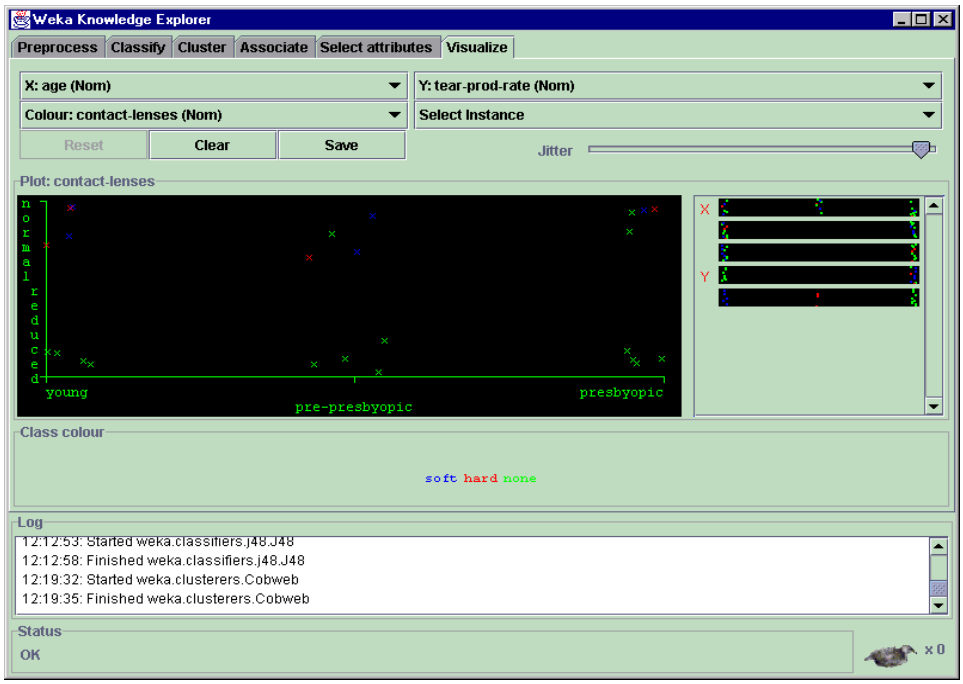


Obr. 24 Práce s atributy



Obr. 25 Rozhodovací strom

[4] Weka je šířena jako open source software spadající pod GNU licenci.



Obr. 26 Shlukování

8.10 Který systém zvolit?

Uvedený přehled jistě nezahrnuje všechny systémy pro dobývání znalostí z databází. Který systém z tak bohaté nabídky tedy zvolit? V současné době totiž nelze hovořit o nějakém standardním, všeobecně používaném systému. Dobrým vodítkem může být to, jak posuzovaný produkt podporuje jednotlivé kroky celého procesu dobývání znalostí. Je tedy dobré:

- posoudit zda systém umožňuje snadný přístup k externím datovým zdrojům, zda umožňuje číst standardní formáty dat (databáze, tabulkové kalkulátory, ASCII) a zda umožňuje pracovat s rozsáhlými soubory,
- posoudit jaké metody (a jak propracované algoritmy těchto metod) systém nabízí, zda nabízí vizualizační možnosti (kancelářská grafika, statistické grafy), zda nabízí průvodce (nebo alsepoň rozumné default nastavení parametrů) pro usnadnění práce s algoritmy, zda nabízí nástroje pro datové transformace (předzpracování), manipulaci se soubory, zda nabízí pracovní prostředí umožňující automatizovat opakovaně prováděné kroky (např. skripty, ukládání sekvence kroků),
- posoudit, zda systém nabízí snadné začlenění výstupů do již používaných aplikací (např. OLE, šablony pro přeformátování výstupu, podpora vytváření reportů),
- posoudit např. to zda lze vytvářet samostatně běžící aplikace (generování spustitelného kódu).

Důležitým kritériem bude jistě i cena. Zde je nutno říci, že ceny za komerční systémy dobývání znalostí mnohonásobně převyšují ceny např. běžného kancelářského software. Podstatná je i snadnost ovládání, dokumentace a konzultace k systému.

Svou roli hraje i to, zda hledáme univerzální systém použitelný pro různé aplikační oblasti nebo zda hledáme systém vyvinutý pro jeden typ aplikací.

Literatura:

[Berka, Ivánek, 1994] Berka, P. - Ivánek,J.: Automated Knowledge Acquisition for PROSPECTOR-like Expert Systems. In: (Bergadano, deRaedt, eds.) Proc. ECML'94? Springer, 1994, 339-342.

[Bierman a kol., 1984] Breiman,L. – Friedman,J. – Olshen,R. – Stone,C.: Classification and Regression Trees, Wadsworth, 1984.

[Clark, Nibblet, 1989] Clark,P. – Nibblet,T.: The CN2 Induction Algorithm. Machine Learning, 3(4), 1989, 261-284.

[Elder, Abbott, 1998] Elder,J.F. – Abbott,D.W.: A Comparison of Leading Data Mining Tools. 4th. Int. Conf. On Knowledge Discovery and Data Mining, New York, 1998.

[Hájek, Havránek, 1978] Hájek,P. – Havránek,T.: Mechanising Hypothesis Formation – Mathematical Foundations for a General Theory. Springer, 1978.

[Ivánek, Stejskal, 1988] Ivánek,J. – Stejskal,B.: Automatic Acquisition of Knowledge Base from Data without Expert: ESOD (Expert System from Observational Data). In: Proc. COMPSTAT"88, Physica-Verlag, 1988.

[Kohavi, 1994] Kohavi,R.: MLC++. A Machine Learning Library in C++, Tech.Rep. CS229B, Stanford Univ. 1994.

[Quinlan, 1993] Quinlan,J.R.: C4.5: Programs for Machine Learning. Morgan Kaufman, 1993, ISBN 1-55860-238-0.

[Rauch, 2001] Rauch,J.: Systém LISp-Miner příklad aplikace. Acta Oeconomica Pragensia Vol 9, No 1. 2001, 125-154.

[Siebes, 2000] Siebes,A.: Developing KDD systems. Zvaná přednáška na 4th. European. Conf. On Principles of Data Mining and Knowledge Discovery PKDD2000, Lyon, 2000.

[Šimůnek, Rauch, 1999] Šimůnek,M. - Rauch,J.: Procedura 4FT-Miner pro KDD. In: (Rychta, ed.) Proc. DATASEM'99, 1999, 281-290.

[Witten, Frank, 1999] Witten,I.H. – Frank,E.: Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufman, 1999, ISBN 1-55860-552-5.