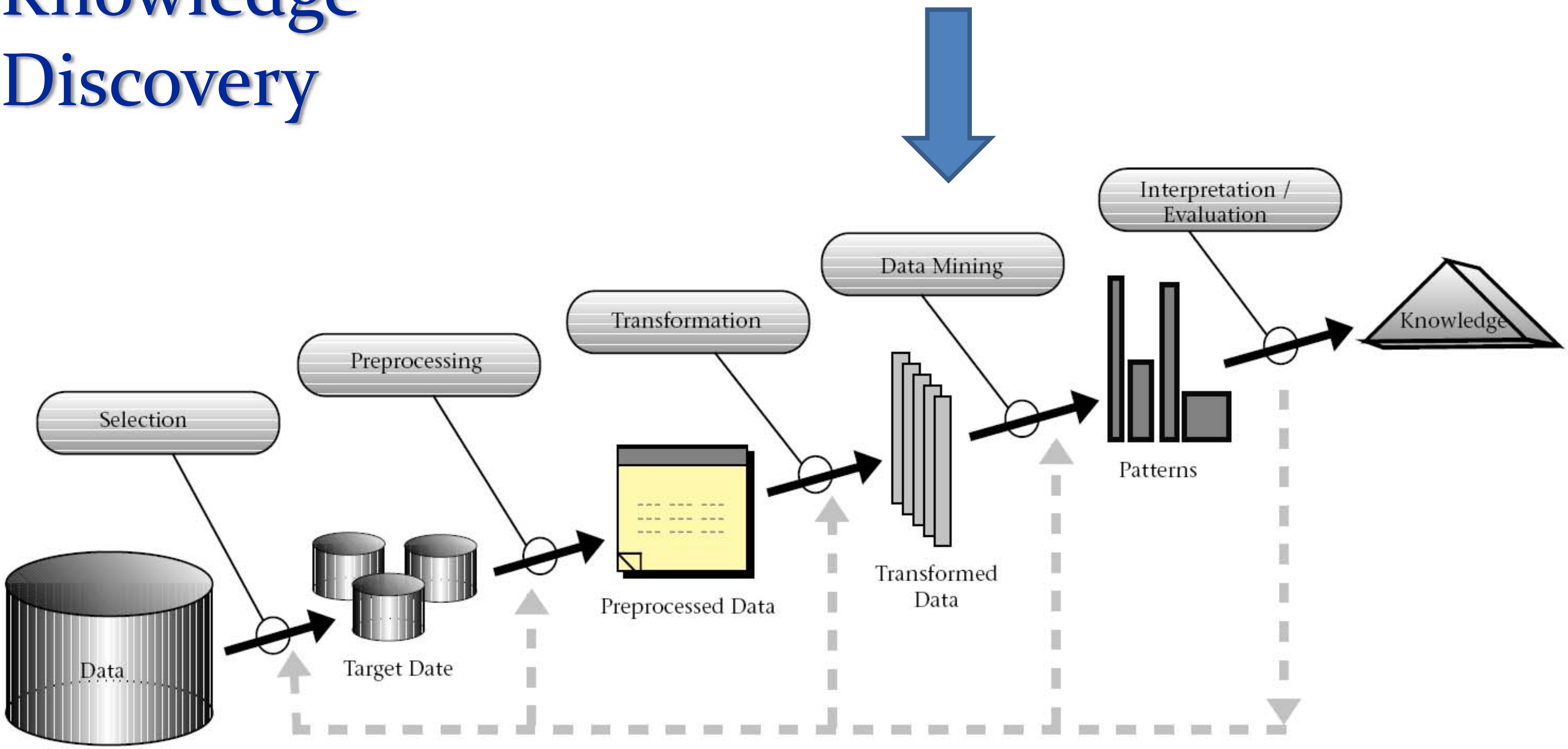
PGN: асоциативен класификатор, генериращ правила с висока степен на доверие.

Програмна реализация и експерименти.

Илия Митов, Красимира Иванова

Годишна научна сесия на ИМИ-БАН;   11.12.2013

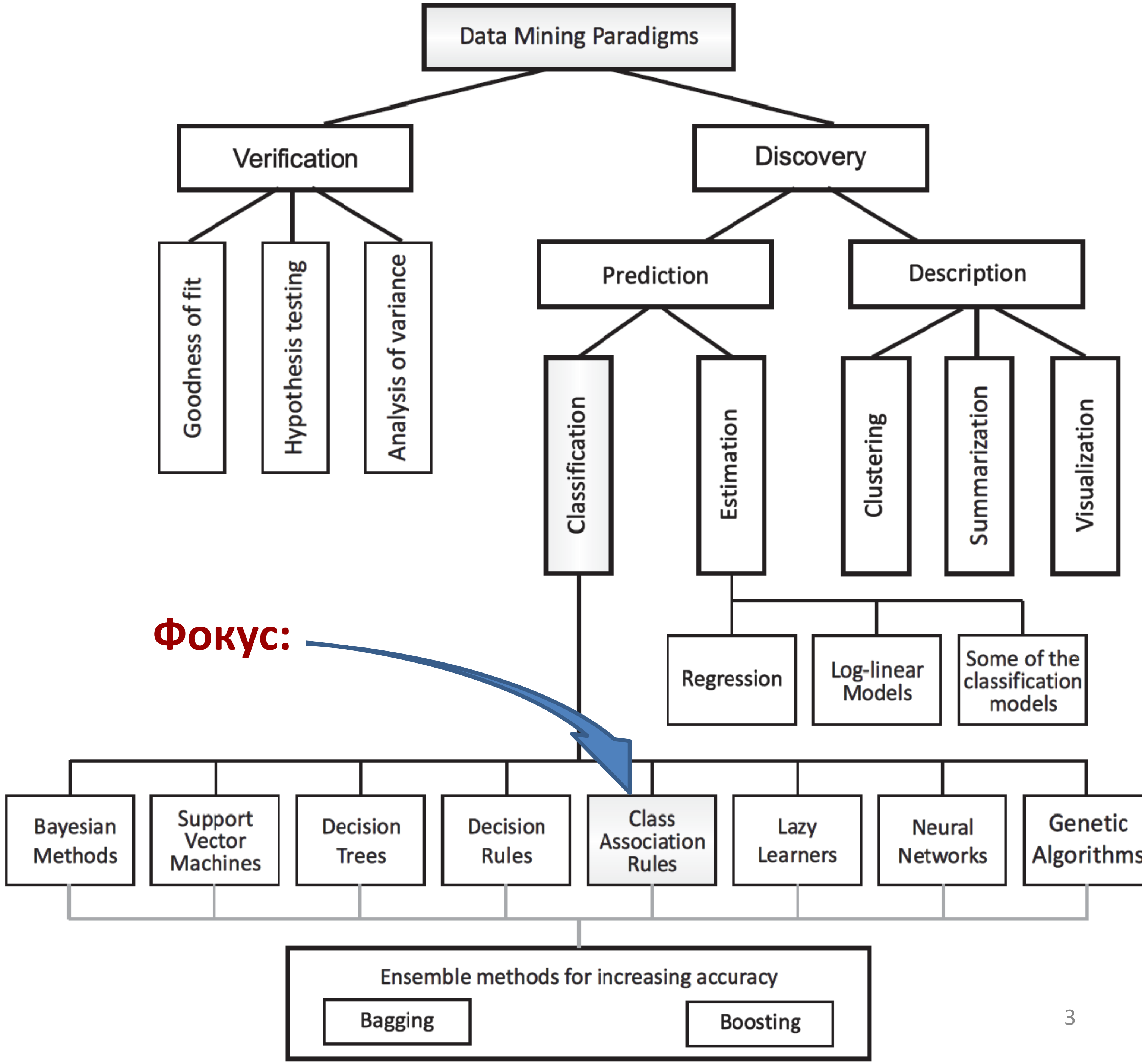
Knowledge Discovery



Основни стъпки на процеса "Извличане на знания"

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery: an overview. In Advances in Knowledge Discovery and Data Mining. American Association for AI, Menlo Park, CA, USA, 1996, pp.1-34.

Data Mining



Data Mining Paradigms
Verification
Discovery
Goodness of fit
Hypothesis testing
Analysis of variance
Prediction
Description
Classification
Estimation
Clustering
Summarization
Visualization
Regression
Log-linear Models
Some of the classification models
Фокус:
Bayesian Methods
Support Vector Machines
Decision Trees
Decision Rules
Class Association Rules
Lazy Learners
Neural Networks
Genetic Algorithms
Ensemble methods for increasing accuracy
Bagging
Boosting

Асоциативни класификатори

Плюсове:

- Ефективно обучение независимо от размера на обучаващото множество;
- Не се влияят от това дали има зависимости между атрибутите;
- Много бързо разпознаване;
- Висока точност на разпознаване;
- Класификационният модел се представя чрез множество от правила, които са интерпретируеми от човека.

Zaiane, O., Antonie, M.-L.: On pruning and tuning rules for associative classifiers. In Proc. of Int. Conf. on Knowledge-Based Intelligence Information & Engineering Systems, LNCS, Vol. 3683, 2005, pp.966-973.

Асоциативни класификатори

Структура:

1. Извличане на асоц. правила (Association rule mining)
2. Съкращаване (Pruning) - опционна
3. Разпознаване (Recognition)

Примери:

- CBA [Liu et al, 1998]
- CMAR [Li et al, 2001]
- ARC-AC and ARC-BC [Zaïane and Antonie, 2002]
- CPAR [Yin and Han, 2003]
- CorClass [Zimmermann and De Raedt, 2004]
- ACRI [Rak et al, 2005]
- TFPC [Coenen and Leng, 2005]
- HARMONY [Wang and Karypis, 2005]
- MCAR [Thabtah et al, 2005]
- CACA [Tang and Liao, 2007]
- ARUBAS [Depaire et al, 2008]

Нотация

- асоциативни правила $\leftrightarrow$ транзакционни множества
- Класификатори $\leftrightarrow$ таблични множества от данни
- $X_1=\{a,b,d,e\}$ $\leftrightarrow$ $X_1=\{\langle a,1\rangle, \langle b,1\rangle, \langle c,0\rangle, \langle d,1\rangle, \langle e,1\rangle\}$

***

- Записите: $D=\{X_i^j\},\ \ X_i=\{a_1^i,\ldots,a_j^i,\ldots,a_{J-1}^i,a_C^i\}$

$$a_j^i=\left\langle a_j, x_j^i\right\rangle,\ \ a_C^i=\left\langle a_C, c^i\right\rangle,\ \ x_j^i\in\{-,1,2,\ldots K^j\}$$

"-": липсваща стойност на атрибут

- Асоциативните правила – същата нотация

$$R_l:\{x_1^l,\ldots,x_{J-1}^l\}\Rightarrow\{c^l\}\qquad\leftrightarrow\qquad R_l=\{x_1^l,\ldots,x_{J-1}^l,c^l\}$$

"-": атрибут, който не участва в правилото,
т.е. произволна стойност на атрибута

Дефиниции

- Def.1: **Covering relation** "$\subset$"
  A rule R_l **covers** a record X_i (rule R_i)
  if the rule's antecedent corresponds with the record (rule):

$$R_l \subset X_i \Leftrightarrow \forall x_j^l \,|\, 1 \le j \le J-1, x_j^l \ne -\} : x_j^l = x_j^i$$

- Def.2: **Matching relation** "$\subseteq$"
  A rule R_l **matches** a record X_i (rule R_i)
  if both the rule's antecedent and consequent corresponds:

$$R_l \subseteq X_i \Leftrightarrow R_l \subset X_i \text{ and } c^l = c^i$$

- Def. 3: **Support** — **Поддръжка**

$$support(R_l, D) = \left| \{ X_i \in D \,|\, R_l \subseteq X_i \} \right|$$

- Def. 4: **Confidence** — **Доверие**

$$confidence(R_l, D) = \frac{\left| \{ X_i \in D \,|\, R_l \subseteq X_i \} \right|}{\left| \{ X_i \in D \,|\, R_l \subset X_i \} \right|}$$

Класификатор PGN

- Типичното за асоциативните класификатори:
  първо се взема пред вид степента на поддръжка
  на асоциативното правило,
  и след това на доверието.
- PGN обръща приоритета и се фокусира
  първо върху доверието
  като оставя само правилата със 100% доверие.
- Основна цел:
  да се изследва качеството на тази концепция.

Примерно обучаващо множество

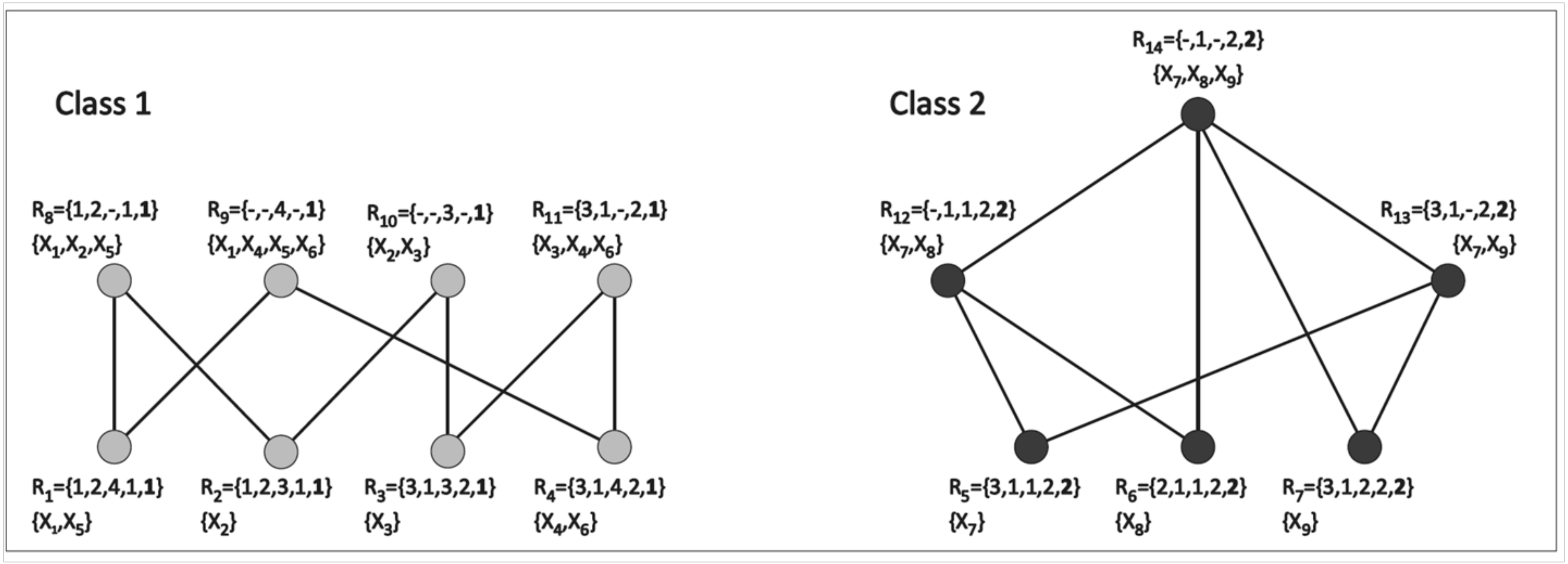
Записи

X_1 = {1, 2, 4, 1, **1**}
X_2 = {1, 2, 3, 1, **1**}
X_3 = {3, 1, 3, 2, **1**}
X_4 = {3, 1, 4, 2, **1**}
X_5 = {1, 2, 4, 1, **1**}   Equal to X_1
X_6 = {3, 1, 4, 2, **1**}   Equal to X_4
X_7 = {3, 1, 1, 2, **2**}
X_8 = {2, 1, 1, 2, **2**}
X_9 = {3, 1, 2, 2, **2**}

PGN обучение: генериране

- Във всеки клас поотделно:
  - От най-дългите правила (записите) към по-късите – докато има нови правила, получени като пресичане на предишните
- Def. 5: **Intersecton**

$$R_1 \cap R_2 = R_3 : \quad \forall x_j^3 \in R_3 \quad x_j^3 = \begin{cases} x_j^1 & if \quad x_j^1 = x_j^2 \\ - & if \quad x_j^1 \neq x_j^2 \end{cases}$$



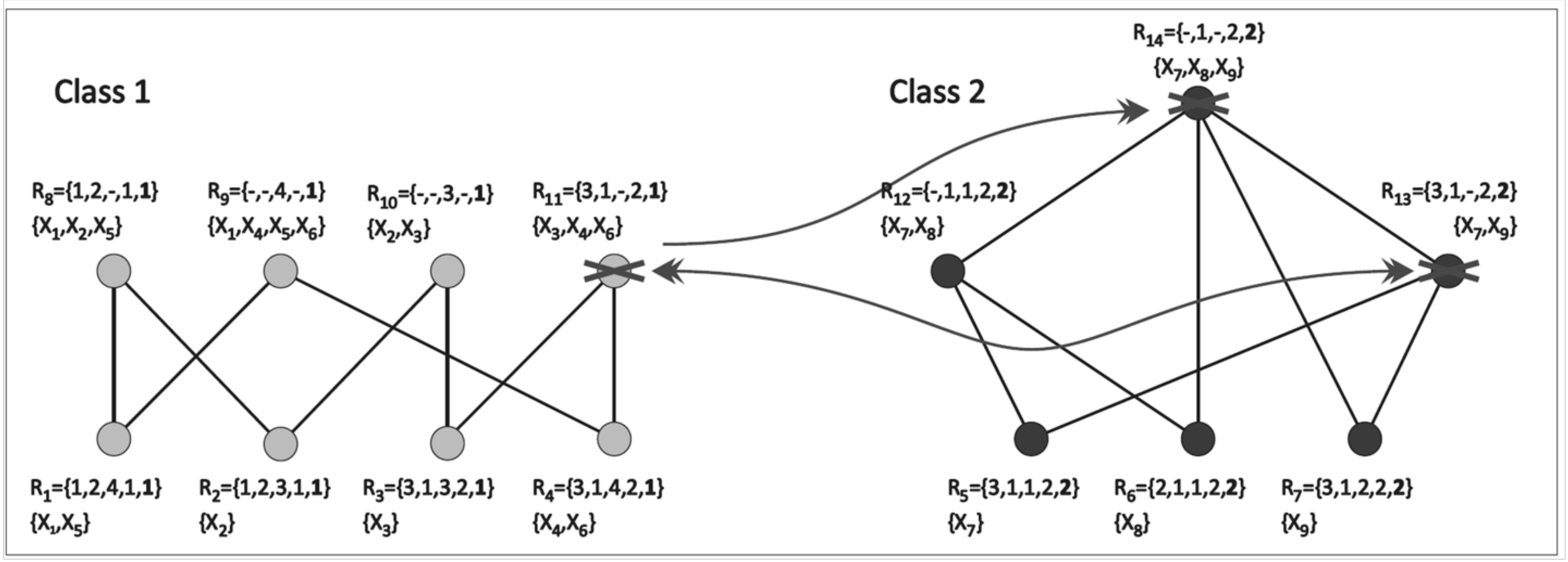
PGN обучение: съкращаване (1)

- Първо – между класовете:
  - При наличие на противоречия се премахват по-общите правила

Def. 6: **Pruning for confidence**

$$R_1 \subset R_2 \wedge c^1 \neq c^2 \Rightarrow \ mark \ R_1 \ for \ removal$$

Proof: ***Pruning for confidence retains only rules with confidence =100%***.

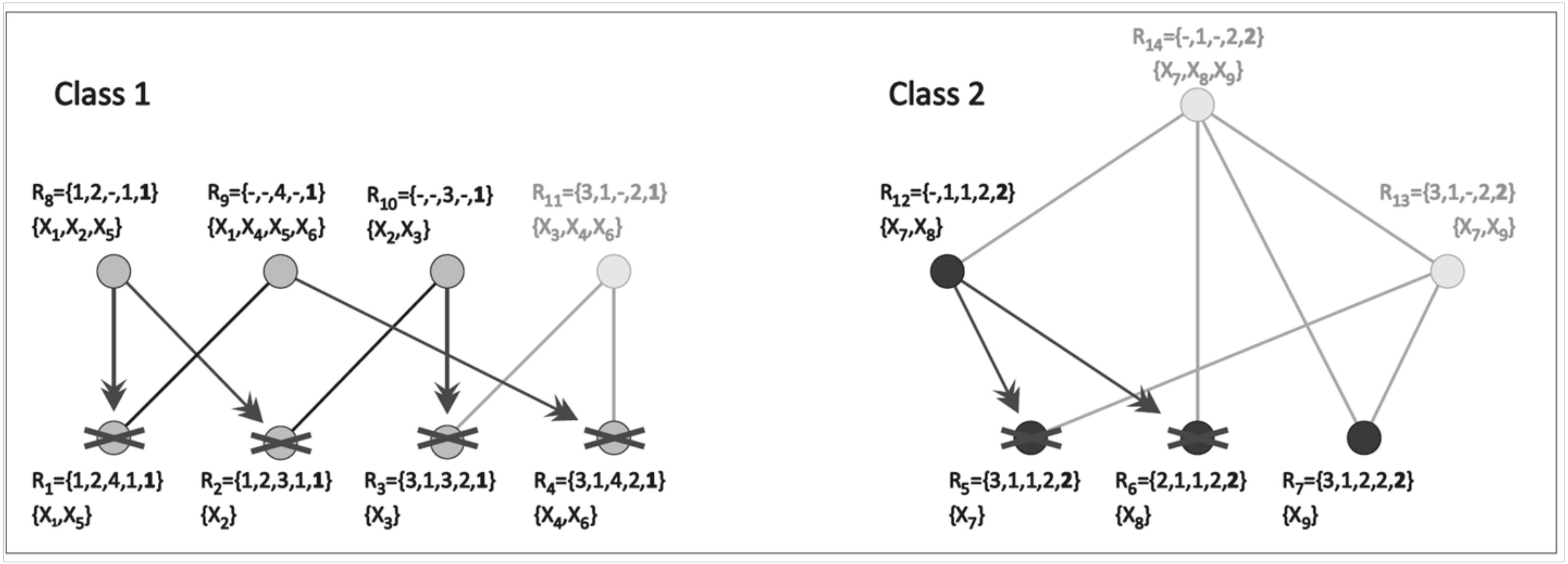


PGN обучение: съкращаване (2)

- Второ – в класовете:
  - Олекотяване на множеството от правила

Def. 7: **Pruning for general rules**

$$R_1 \subset R_2 \wedge c^1 = c^2 \Rightarrow \ mark\ R_2\ for\ removal$$



PGN - разпознаване

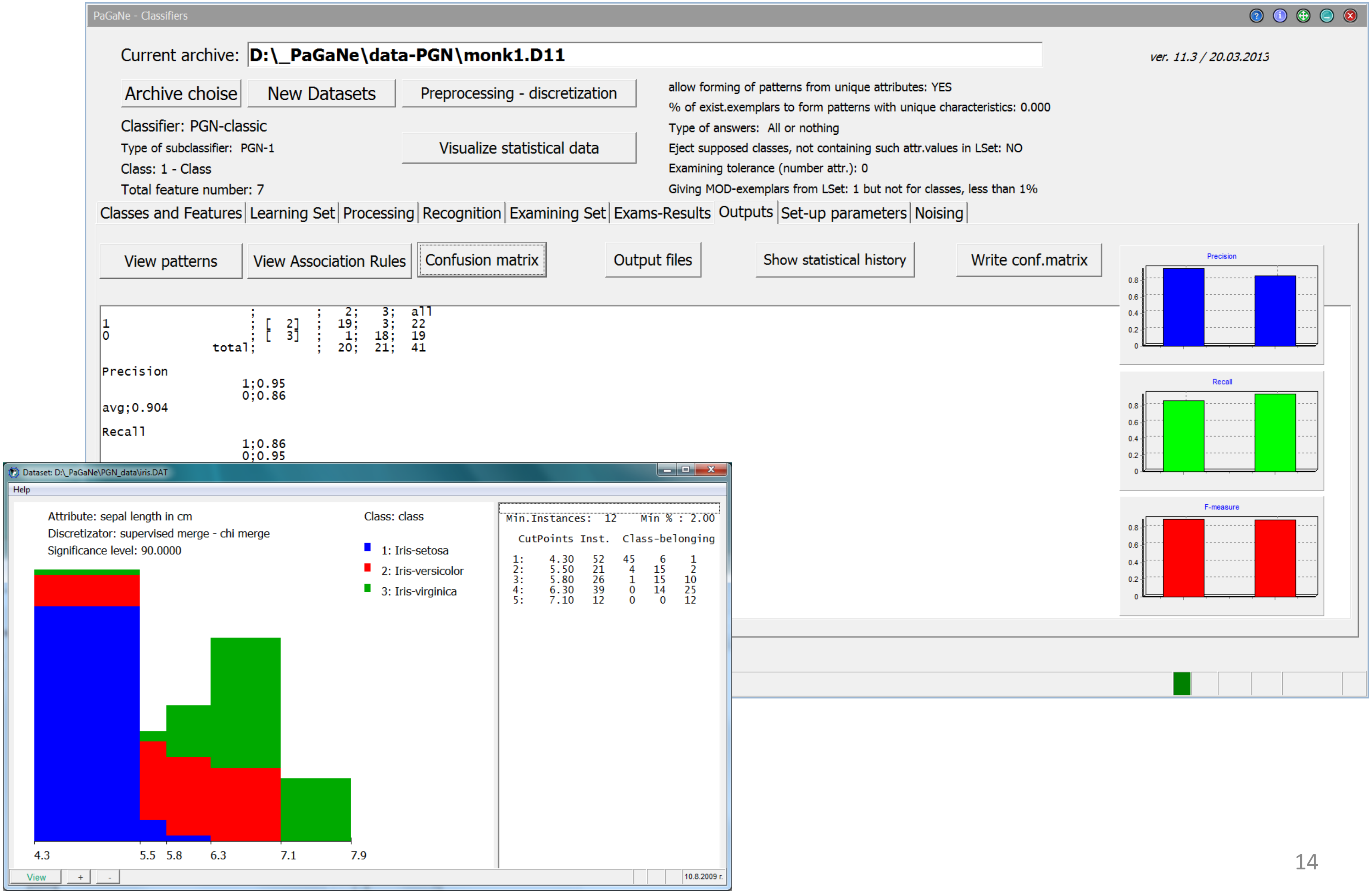
- Def. 8: **Association Rule Size** $\left|R_l\right| = \left|\left\{x_j^l \mid 1 \le j \le J-1, x_j^l \ne -\right\}\right|$
- Def. 9: **Intersection Percentage** $IP(X_i, R_l) = 100 * \frac{\left|X_i \cap R_l\right|}{\left|R_l\right|}$

- **Класификация:**

Classification of $X_i = \{1, 2, 1, 2, ?\}$

R_l	$X_i \cap R_l$	$IP(X_i, R_l)$	Support
R_1 = {1, 2, –, 1, **1**}	{1, 2, –, –, ?}	0.667	3
R_2 = {–, –, 4, –, **1**}	{–, –, –, –, ?}	0	4
R_3 = {–, –, 3, –, **1**}	{–, –, –, –, ?}	0	2
R_4 = {3, 1, 2, 2, **2**}	{–, –, –, 2, ?}	0.250	1
R_5 = {–, 1, 1, 2, **2**}	{–, –, 1, 2, ?}	0.667	2

Програмна реализация - PaGaNe



Експерименти

- 25 data sets from the UCI Machine Learning Repository
  (continuous attributes were discretized first by means of the Chi-merge method with 95% Chi-square threshold)
- 21 classifiers:
  - Associative classifiers: PGN and CMAR (supp. 1%; conf. 50%);
  - Decision Rules: One R, JRip (pruned and unpruned), Decision Table; NNge;
  - Decision Trees: REP Tree, J48 (pruned and unpruned), LAD Tree;
  - Nearest Neighbor learners: IBk, KStar;
  - Bayes: Naïve Bayes, Bayes Net, HNB, WAODE, LBR;
  - Ensemble methods (Bagging): Random Forest;
  - Support Vector Machines: SMO;
  - Neural Networks: Multilayer Perceptron.
- Used programs:
  - PGN - data mining environment PaGaNe;
  - CMAR - LUCS-KDD Repository;
  - for all other classifiers their Weka implementation are used.

Сравнение на класификатори: методика Demšar

n=25 множества от данни
k=21 класификатора
Five-fold cross validation -> точностите на разпознаване на класификаторите

1: Friedman test ->

- Класиране на алгоритмите за всяко от множествата (ranking)
- Изчисляне на среден ранг на класификатор R_j
- Нулева хипотеза (за статистическа неразличимост на класификаторите)

$$\chi_F^2 = \frac{12n}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right]$$

2: Nemenyi test – за сравняване на конкретен класификатор спрямо останалите

$$z = (R_i - R_j)$$

$$\text{Critical Difference } CD = q_\alpha \sqrt{\frac{k(k+1)}{6n}}$$

q_α : based on the Studentized range statistic divided by $\sqrt{2}$

Demsar, J.: Statistical comparisons of classifiers over multiple data sets.
J. Mach. Learn. Res., 7, 2006, pp.1-30.

Точност на разпознаване (в проценти) на класификаторите

classifier / dataset	PGN	CMAR	One R	JRip-unpr.	Dec. Table	JRip-pruned	NNge	REP Tree	J48-pruned	J48-unpr.	LAD Tree	IBk	KStar	Naïve Bayes	Bayes Net	HNB	WAODE	LBR	Rand. Forest	SMO	Mult. perc.
annealing	96.24	95.99	83.71	99.00	98.37	98.12	97.24	97.62	98.12	98.75	98.12	98.25	98.75	91.61	91.11	97.62	96.99	96.87	97.62	99.12	99.12
audiology	75.50	59.18	47.00	68.50	61.00	69.50	67.00	62.50	72.00	72.00	71.50	76.50	76.00	64.50	71.00	68.00	71.50	65.00	73.50	76.50	78.50
balance_scale	77.89	86.70	60.10	72.76	66.83	71.95	70.68	67.15	66.18	69.87	82.69	85.26	86.70	90.54	90.54	87.02	88.14	90.54	75.48	89.42	98.72
breast_cancer_wo	96.43	93.85	91.85	92.85	92.42	93.28	94.99	93.99	94.28	94.71	94.85	95.85	95.28	97.14	97.14	95.13	95.85	97.14	95.42	95.99	96.28
car	92.59	81.77	70.03	87.44	91.43	86.75	94.33	88.2	90.8	93.17	90.45	92.94	86.81	85.19	85.30	92.24	90.11	91.95	93.52	92.59	99.83
cmc	49.90	53.16	47.25	45.55	49.42	50.38	44.81	50.17	51.60	48.07	54.86	47.12	50.31	50.45	50.31	52.96	52.68	52.55	48.68	53.50	47.73
credit	87.54	87.10	85.51	81.45	85.8	85.07	80.14	85.07	85.36	83.91	86.67	82.90	84.78	86.38	86.38	84.93	85.94	86.23	85.51	85.94	86.09
ecoli	79.76	81.26	60.42	75.91	76.50	80.07	77.70	79.17	77.09	78.28	81.27	79.76	80.36	84.84	84.54	79.77	82.75	84.84	80.36	84.24	80.67
forestfires	57.63	58.80	53.38	55.31	52.03	54.76	54.36	53.95	53.96	52.41	57.26	56.69	56.68	58.02	58.21	56.29	60.73	58.02	58.42	61.11	58.01
glass	78.51	78.04	54.67	71.98	61.23	66.40	71.53	67.29	73.38	74.76	71.95	78.98	78.99	74.33	74.34	75.70	77.13	74.33	76.19	77.12	74.32
hayes-roth	81.94	83.42	50.77	78.86	51.51	78.12	75.10	73.53	68.23	69.00	87.24	63.67	61.40	85.67	85.67	72.82	76.61	85.67	76.61	83.39	83.45
hepatitis	80.65	84.52	81.94	76.78	82.58	77.42	81.29	79.36	79.36	77.42	77.42	81.29	80.65	86.45	85.16	85.81	83.87	87.10	83.23	77.42	81.29
iris	92.67	92.67	94.67	93.33	92.67	92.67	94.67	93.33	94.67	93.33	93.33	93.33	93.33	92.67	92.67	92.00	93.33	92.67	94.67	93.33	94.67
lenses	74.00	88.00	62.00	87.00	92.00	83.00	70.00	80.00	83.00	75.00	83.00	78.00	78.00	70.00	70.00	54.00	70.00	70.00	74.00	70.00	74.00
mammographic	80.75	82.11	82.00	78.98	82.73	81.69	76.28	81.69	81.69	83.46	80.96	80.44	81.17	82.62	82.42	82.42	82.94	82.42	81.90	81.48	80.44
monks1	100.00	100.00	74.98	99.31	100.00	87.53	96.05	88.91	94.68	93.28	80.08	97.92	97.92	74.98	74.98	100.00	74.29	100.00	96.30	74.98	100.00
monks2	73.06	59.74	65.73	58.74	64.40	58.73	73.87	63.90	59.90	60.91	68.39	71.55	76.88	61.41	61.24	67.90	63.73	66.57	65.39	65.73	100.00
monks3	98.56	98.92	79.97	98.56	98.92	98.92	98.20	98.92	98.92	98.92	98.92	97.66	97.84	96.39	96.39	98.38	98.56	98.74	98.02	96.75	98.92
soybean	93.15	78.48	37.44	87.28	75.24	85.35	89.24	78.18	87.64	87.95	77.85	90.87	91.85	82.76	86.33	91.85	90.87	86.99	89.91	90.87	92.18
tae	52.94	35.74	45.76	33.72	47.70	34.43	50.88	40.43	46.97	47.61	45.64	57.53	55.57	46.99	46.34	52.93	53.59	50.30	48.92	51.61	54.92
tic_tac_toe	88.93	98.75	69.93	97.29	73.70	98.02	86.53	80.37	84.23	84.23	73.70	97.39	95.30	71.29	71.40	77.03	73.27	84.97	91.13	98.33	97.81
votes	95.86	94.02	95.63	94.25	93.79	94.71	94.71	95.40	95.17	94.25	95.63	93.79	93.56	89.89	89.89	94.48	95.40	94.02	95.63	95.86	95.40
wine	96.09	91.70	78.63	89.33	80.90	90.45	92.18	88.16	87.03	88.19	90.98	96.11	96.11	98.89	99.44	98.33	97.20	98.89	94.40	98.33	97.76
winequality-red	64.98	56.29	55.54	48.65	55.97	53.72	60.79	57.03	58.22	59.16	56.41	64.29	64.67	58.60	58.47	62.29	61.91	59.28	64.35	59.04	64.04
zoo	98.10	94.19	73.29	90.14	88.19	88.19	95.14	82.19	94.14	95.14	98.10	96.14	96.14	94.10	96.10	97.10	98.05	94.10	96.10	98.05	96.14

Рангове

dataset \ classifier	PGN	CMAR	One R	JRip-unpruned	Dec.Table	JRip-pruned	NNge	REPTree	J48-pruned	J48-unpruned	LADTree	IB k	K Star	NaiveBayes	BayesNet	HNB	WAODE	LBR	RandForest	SMO	Mult.perceptron
annealing	17	18	21	3	6	9	14	12	9	4.5	9	7	4.5	19	20	12	15	16	12	**1.5**	**1.5**
audiology	5	20	21	13	19	12	15	18	7.5	7.5	9.5	2.5	4	17	11	14	9.5	16	6	2.5	**1**
balance_scale	12	8.5	21	14	19	15	16	18	20	17	11	10	8.5	3	3	7	6	3	13	5	**1**
breast_canc.	4	17	21	19	20	18	12	16	15	14	13	7.5	10	**2**	**2**	11	7.5	**2**	9	6	5
car	6.5	20	21	15	10	17	2	14	11	4	12	5	16	19	18	8	13	9	3	6.5	**1**
cmc	13	3	18	20	14	9	21	12	7	16	**1**	19	10.5	8	10.5	4	5	6	15	2	17
credit	**1**	2	11.5	20	10	14.5	21	14.5	13	18	3	19	17	4.5	4.5	16	8.5	6	11.5	8.5	7
ecoli	13.5	7	21	20	19	11	17	15	18	16	6	13.5	9.5	**1.5**	3	12	5	**1.5**	9.5	4	8
forestfires	9	3	19	14	21	15	16	18	17	20	10	11	12	6.5	5	13	2	6.5	4	**1**	8
glass	3	4	21	15	20	19	17	18	14	9	16	2	**1**	11.5	10	8	5	11.5	7	6	13
hayes-roth	8	6	21	9	20	10	13	14	17	16	**1**	18	19	3	3	15	11.5	3	11.5	7	5
hepatitis	13.5	5	9	21	8	18.5	11	15.5	15.5	18.5	18.5	11	13.5	2	4	3	6	**1**	7	18.5	11
iris	17	17	**3**	9.5	17	17	**3**	9.5	**3**	9.5	9.5	9.5	9.5	17	17	21	9.5	17	**3**	9.5	**3**
lenses	12	2	20	3	**1**	5	16.5	7	5	10	5	8.5	8.5	16.5	16.5	21	16.5	16.5	12	16.5	12
mammogr.	17	8	9	20	3	12	21	12	12	**1**	16	18.5	15	4	6	6	2	6	10	14	18.5
monks1	**3.5**	**3.5**	18.5	7	**3.5**	15	11	14	12	13	16	8.5	8.5	18.5	18.5	**3.5**	21	**3.5**	10	18.5	**3.5**
monks2	4	19	9.5	20	12	21	3	13	18	17	6	5	2	15	16	7	14	8	11	9.5	**1**
monks3	11	**4.5**	21	11	**4.5**	**4.5**	14	**4.5**	**4.5**	**4.5**	**4.5**	17	16	19.5	19.5	13	11	9	15	18	**4.5**
soybean	**1**	17	21	12	20	15	9	18	11	10	19	6	3.5	16	14	3.5	6	13	8	6	2
tae	5	19	16	21	11	20	8	18	14	12	17	**1**	2	13	15	6	4	9	10	7	3
tic_tac_toe	9	**1**	21	6	16.5	3	10	14	12.5	12.5	16.5	5	7	20	19	15	18	11	8	2	4
votes	**1.5**	15.5	4	13.5	17.5	10.5	10.5	7	9	13.5	4	17.5	19	20.5	20.5	12	7	15.5	4	**1.5**	7
wine	10	13	21	16	20	15	12	18	19	17	14	8.5	8.5	2.5	**1**	4.5	7	2.5	11	4.5	6
wineq.-red	**1**	17	19	21	18	20	8	15	14	10	16	4	2	12	13	6	7	9	3	11	5
zoo	**1.5**	13	21	17	18.5	18.5	11.5	20	14	11.5	**1.5**	7	7	15.5	9.5	5	3.5	15.5	9.5	3.5	7

Сравнение – Friedman test

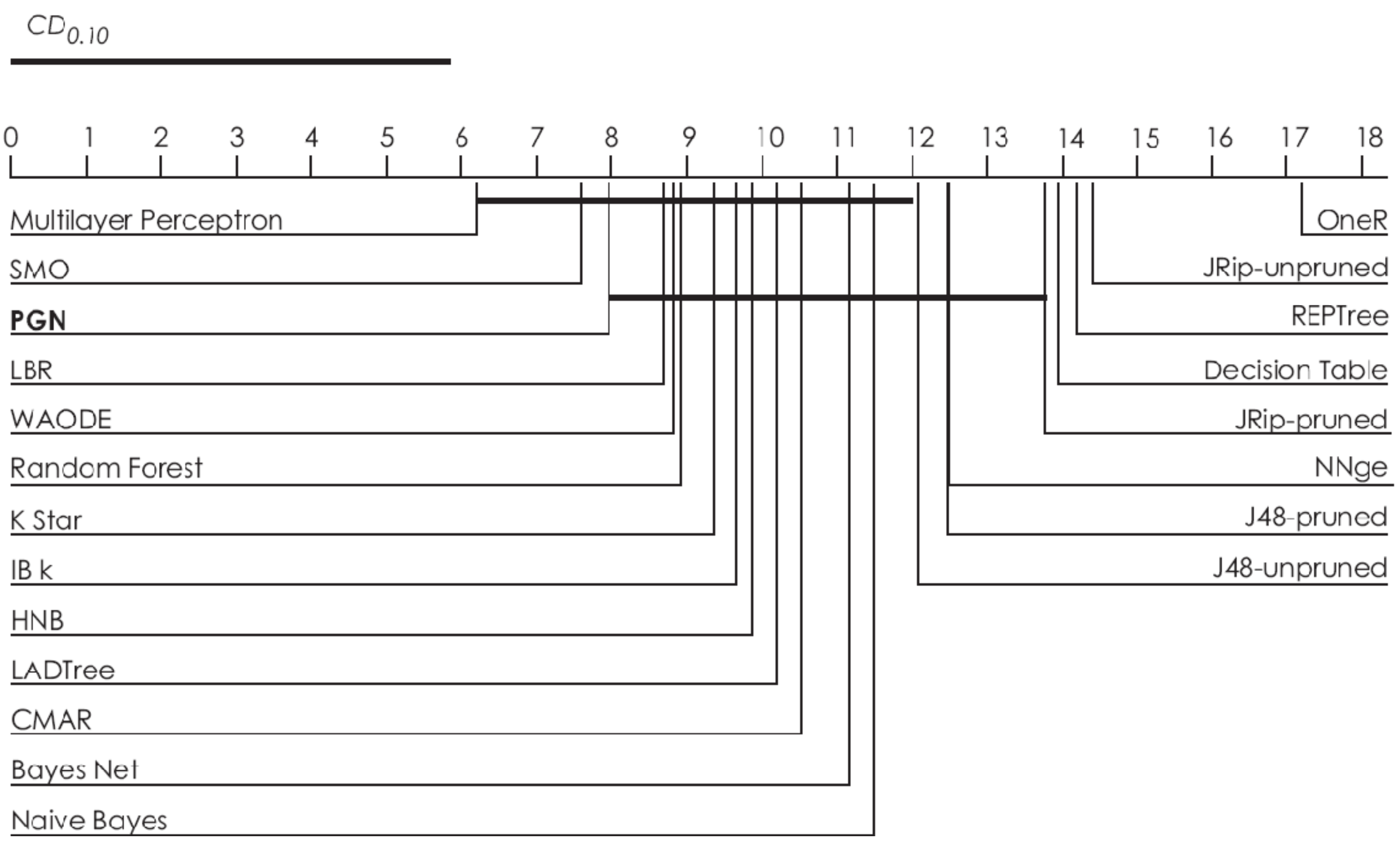
	PGN	CMAR	One R	JRip-unpr.	Dec. Table	JRip-pruned	NNge	REP Tree	J48-pruned	J48-unpr.	LAD Tree	IB k	K Star	Naïve Bayes	Bayes Net	HNB	WAO DE	LBR	Rand. Forest	SMO	Mult. perc.
Avg.Rank	7.96	10.52	17.18	14.40	13.94	13.78	12.50	14.20	12.48	12.08	10.20	9.66	9.36	11.48	11.18	9.86	8.82	8.68	8.92	7.60	6.20

20 степени на свобода,

$\chi^2 = 95.579 > \alpha_{0.10} = 28.412$

→ Класификаторите са статистически различими

Сравнение - Nemenyi test



Благодарим за вниманието!

Илия Митов
imitov@math.bas.bg

Красимира Иванова
kivanova@math.bas.bg