Mining Imprefect Data

Jerzy W. Grzymala-Busse

jerzy@ku.edu

Department of Electrical Engineering and Computer Science, University of Kansas,

Lawrence, KS 66045, USA

Outline of the talk

- Inconsistent Data
  - Blocks of Attribute-Value Pairs
  - Elementary Sets
  - Approximations
  - Experiments
- Incomplete Data
  - Sequential Methods
  - Parallel Methods
    - Characteristic Sets
    - Global Approximations
    - Local Approximations
    - Experiments
    - Conclusions

US Congressional Voting 1984, I

- handicapped-infants
- water-project-cost-sharing
- adoption-of-the-budget-resolution
- physician-fee-freeze
- el-salvador-aid
- religious-groups-in-schools
- anti-satellite-test-ban
- aid-to-nicaraguan-contras
- mx-missile
- immigration
- synfuels-corporation-cutback
- education-spending

US Congressional Voting 1984

- superfund-right-to-sue
- crime
- duty-free-exports
- export-administration-act-south-africa class

n y n y y y n n n y ? y y y n y republican

n y n y y y n n n n n y y y n ? republican

? y y ? y y n n n n y n y y n n democrat

? ? ? ? ? ? ? ? y ? ? ? ? ? ? ? democrat

? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? republican

An Inconsistent Data Set

	Attributes			Decision
Case	Temperature	Headache	Cough	Flu
1	high	yes	no	no
2	high	no	yes	no
3	normal	no	no	no
4	normal	no	no	no
5	high	yes	no	yes
6	high	yes	no	yes
7	high	no	yes	yes
8	high	no	no	maybe
9	high	no	no	maybe

Blocks of Attribute-Value Pairs

$a \in A$ and

v be a value of *a* for some case *x*,

denoted by $a(x) = v$,

for complete decision tables if

$t = (a, v)$ is an attribute-value pair then

a *block* of *t*, denoted [*t*],

is a set of all cases from *U* that

for attribute *a* have value *v*.

A Data Set

	Attributes			Decision
Case	Temperature	Headache	Cough	Flu
1	high	yes	no	no
2	high	no	yes	no
3	normal	no	no	no
4	normal	no	no	no
5	high	yes	no	yes
6	high	yes	no	yes
7	high	no	yes	yes
8	high	no	no	maybe
9	high	no	no	maybe

Blocks of Attribute-Value Pairs

[(Temperature, high)] = {1, 2, 5, 6, 7, 8, 9},

Blocks of Attribute-Value Pairs

[(Temperature, high)] = {1, 2, 5, 6, 7, 8, 9},

[(Temperature, normal)] = {3, 4},

Blocks of Attribute-Value Pairs

[(Temperature, high)] = {1, 2, 5, 6, 7, 8, 9},

[(Temperature, normal)] = {3, 4},

[(Headache, yes)] = {1, 5, 6},

[(Headache, no)] = {2, 3, 4, 7, 8, 9},

[(Cough, no)] = {1, 3, 4, 5, 6, 8, 9},

[(Cough, yes)] = {2, 7}.

Elementary Sets of B

Let B be a nonempty subset of the set A of all attributes.

$$[x]_B = \cap\{[(a, a(x))] | a \in B\}.$$

A union of B-elementary sets is called a B-definable set.

Elementary Sets of A

$[1]_A = [(Temperature, high)] \cap [(Headache, yes)] \cap$
$[(Cough, no)] = \{1, 5, 6\},$

Elementary Sets of *A*

$[1]_A = [5]_A = [6]_A = [(Temperature, high)] \cap [(Headache, yes)] \cap [(Cough, no)] = \{1, 5, 6\}$,

$[2]_A = [7]_A = [(Temperature, high)] \cap [(Headache, no)] \cap [(Cough, yes)] = \{2, 7\}$,

$[3]_A = [4]_A = [(Temperature, normal)] \cap [(Headache, no)] \cap [(Cough, no)] = \{3, 4\}$,

$[8]_A = [9]_A = [(Temperature, high)] \cap [(Headache, no] \cap [(Cough, no)] = \{8, 9\}$.

Indiscernibility Relation

The *indiscernibility relation* $IND(B)$ is a relation on U defined for $x, y \in U$ as follows

$$(x, y) \in IND(B) \;\; if\ and\ only\ if\ a(x) = a(y)\ for\ all\ a \in B.$$

$IND(A)$ = {(1, 1), (1, 5), (1, 6), (2, 2), (2, 7), (3, 3), (3, 4),

(5, 1), (5, 5), (5, 6), (6, 1), (6, 5), (6, 6), (7, 2), (7, 7), (8, 8),

(8, 9), (9, 8), (9, 9)}.

Lower and Upper Approximations

- First definition

$$\underline{B}X = \{x \in U \mid [x]_B \subseteq X\},$$

$$\overline{B}X = \{x \in U \mid [x]_B \cap X \neq \emptyset.$$

Lower and Upper Approximations

- First definition

$$\underline{B}X = \{x \in U \mid [x]_B \subseteq X\},$$

$$\overline{B}X = \{x \in U \mid [x]_B \cap X \neq \emptyset.$$

- Second definition

$$\underline{B}X = \cup\{[x]_B \mid x \in U, [x]_B \subseteq X\},$$

$$\overline{B}X = \cup\{[x]_B \mid x \in U, [x]_B \cap X \neq \emptyset).$$

Lower and Upper Approximations

The largest B-definable set contained in X is called the *B-lower approximation* of X, denoted by $\underline{appr}_B(X)$, and defined as follows

$$\cup\{[x]_B \mid [x]_B \subseteq X\}$$

the smallest B-definable set containing X, denoted by $\overline{appr}_B(X)$ is called the *B-upper approximation* of X, and is defined as follows

$$\cup\{[x]_B \mid [x]_B \cap X \neq \emptyset\}.$$

An Example

For the concept [(*Flu*, *no*)] = {1, 2, 3, 4},

$$\begin{aligned} \underline{appr}_A([(Flu, no)]) &= \{3, 4\}, \\ \overline{appr}_A([(Flu, no)]) &= \{1, 2, 3, 4, 5, 6, 7\}. \end{aligned}$$

Probabilistic Approximations

A *probabilistic (parameterized) approximation*, denoted by $appr_\alpha(X)$, is defined by

$$\cup\{[x] \mid x \in U,\ Pr(X|[x]) \geq \alpha\},$$

where α is called a *threshold* and $1 \geq \alpha > 0$.

We excluded the case of $\alpha = 0$ since then $appr_\alpha(X) = U$ for any X.

Since we consider all possible values of α, our definition of $appr_\alpha(X)$ covers both lower and upper probabilistic approximations.

Standard Approximations

If $\alpha = 1$, the probabilistic approximation becomes the standard lower approximation.

If α is small, close to 0, the same definition describes the standard upper approximation.

All Conditional Probabilities

For the fixed set X and all

equivalence classes $[x]$ from R^*

we may compute the set of all

distinct conditional probabilities $Pr(X|[x])$

and then sort these numbers in the ascending order.

The number of all probabilistic approximations of X is

smaller than or equal to

the number of elementary sets $[x]$.

Conditional Probabilities

$[x]$	{1, 5, 6}	{2, 7}	{3, 4}	{8, 9}
$P(\{1, 2, 3, 4\} \mid [x])$	0.333	0.5	1.0	0

Probabilistic Approximations

for the concept {1, 2, 3, 4} we may define only three distinct probabilistic approximations:

$$appr_{0.333}(\{1, 2, 3, 4\}) = \{1, 2, 3, 4, 5, 6, 7\},$$

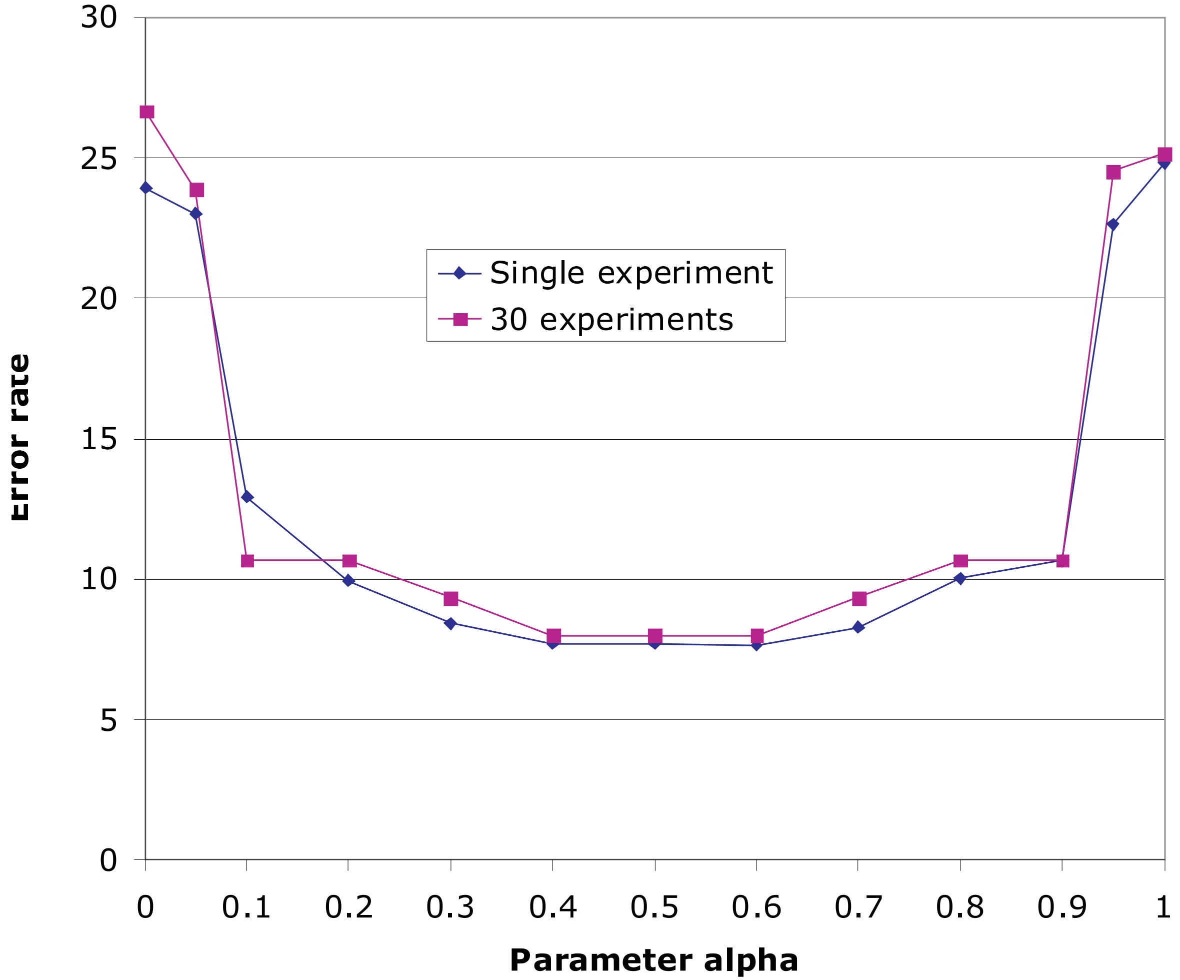
$$appr_{0.5}(\{1, 2, 3, 4\}) = \{2, 3, 4, 7\},$$

$$appr_{1.0}(\{1, 2, 3, 4\}) = \{3, 4\}.$$

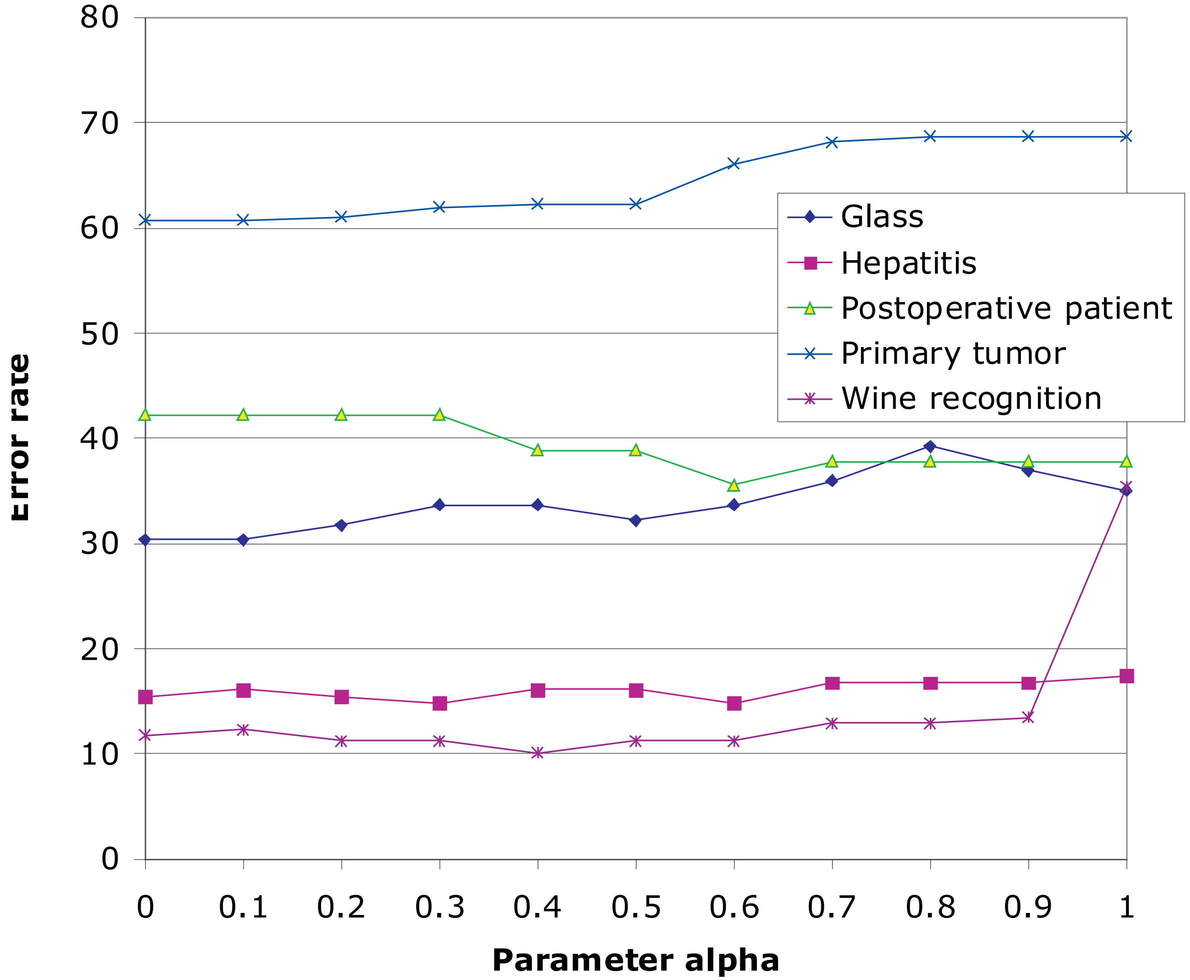
Experiments - Data

Data set	Number of			Consistency
	cases	attributes	concepts	
Glass	214	9	6	55.14
Hepatitis	155	19	2	65.81
Iris	150	4	3	56.0
Postoperative patient	90	8	3	84.44
Primary tumor	339	17	21	72.27
Wine recognition	178	13	3	61.80

Extensive Experiments - Iris



Experiments - Five Data Sets



An Incomplete Data Set

	Attributes			Decision
Case	Temperature	Headache	Nausea	Flu
1	high	–	no	yes
2	very_high	yes	yes	yes
3	?	no	no	no
4	high	yes	yes	yes
5	high	?	yes	no
6	normal	yes	no	no
7	normal	no	yes	no
8	–	yes	*	yes

Sequential Methods, I

- Deleting cases with missing attribute values (*listwise deletion, casewise deletion, complete case analysis*)
- The most common value of an attribute
- The most common value of an attribute restricted to a concept
- Assigning all possible attribute values to a missing attribute value
- Assigning all possible attribute values restricted to a concept

Sequential Methods, II

- Replacing missing attribute values by the attribute mean
- Replacing missing attribute values by the attribute mean restricted to a concept
- Global closest fit
- Concept global fit
- Imputation
  - ML method (*maximum likelihood*)
  - EM method (*expectation-maximization*)
  - Single random imputation
  - Multiple random imputation

Parallel Methods

- C4.5
- CART
- MLEM2
  - Characteristic Relations
  - Singleton, Subset, and Concept Approximations
  - Local Approximations
  - Rule Induction

Incomplete Data

- Missing attribute values:
  - Lost values are denoted by ?
  - "do not care" conditions are denoted by *
  - attribute-concept values are denoted by –
- All decision values are specified
- For each case at least one attribute value is specified

Blocks of Attribute-Value Pairs, I

- If for an attribute a there exists a case x such that $a(x) =?$ then the case x should not be included in any block $[(a, v)]$ for all specified values v of attribute a,

Blocks of Attribute-Value Pairs, I

- If for an attribute a there exists a case x such that $a(x) = ?$ then the case x should not be included in any block $[(a, v)]$ for all specified values v of attribute a,
- If for an attribute a there exists a case x such that $a(x) = *$, then the case x should be included in all blocks $[(a, v)]$ for all specified values v of attribute a.

Blocks of Attribute-Value Pairs, II

- If for an attribute a there exists a case x such that $a(x) = -$ then the corresponding case x should be included in blocks $[(a, v)]$ for all specified values $v \in V(x, a)$ of attribute a, where

$$V(x, a) = \{a(y) \mid a(y) \ is \ specified, y \in U, \ d(y) = d(x)\}.$$

An Incomplete Decision Table

	Attributes			Decision
Case	Temperature	Headache	Nausea	Flu
1	high	–	no	yes
2	very_high	yes	yes	yes
3	?	no	no	no
4	high	yes	yes	yes
5	high	?	yes	no
6	normal	yes	no	no
7	normal	no	yes	no
8	–	yes	*	yes

Blocks of Attribute-Value Pairs, III

[(Temperature, high)] = {1, 4, 5, 8},

An Incomplete Decision Table

	Attributes			Decision
Case	Temperature	Headache	Nausea	Flu
1	high	–	no	yes
2	very_high	yes	yes	yes
3	?	no	no	no
4	high	yes	yes	yes
5	high	?	yes	no
6	normal	yes	no	no
7	normal	no	yes	no
8	–	yes	*	yes

Blocks of Attribute-Value Pairs, III

[(Temperature, high)] = {1, 4, 5, 8},

[(Temperature, very_high)] = {2, 8},

Blocks of Attribute-Value Pairs, III

[(Temperature, high)] = {1, 4, 5, 8},

[(Temperature, very_high)] = {2, 8},

[(Temperature, normal)] = {6, 7},

[(Headache, yes)] = {1, 2, 4, 6, 8},

[(Headache, no)] = {3, 7},

[(Nausea, no)] = {1, 3, 6, 8},

[(Nausea, yes)] = {2, 4, 5, 7, 8}.

Characteristic sets $K_B(x)$, I

- Characteristic set $K_B(x)$ is the intersection of the sets $K(x, a)$, for all $a \in B$:
- If $a(x)$ is specified, then $K(x, a)$ is the block $[(a, a(x)]$,
- If $a(x) = *$ or $a(x) =?$ then the set $K(x, a) = U$,
- If $a(x) = -$, then $K(x, a)$ is equal to the union of all blocks of attribute-value pairs (a, v), where $v \in V(x, a)$.

Characteristic sets $K_A(x)$, II

$$K_A(1) = \{1,4,5,8\} \cap \{1,2,4,6,8\} \cap \{1,3,6,8\} = \{1,8\},$$

Characteristic sets $K_A(x)$, II

$K_A(1) = \{1, 4, 5, 8\} \cap \{1, 2, 4, 6, 8\} \cap \{1, 3, 6, 8\} = \{1, 8\}$,

$K_A(2) = \{2, 8\} \cap \{1, 2, 4, 6, 8\} \cap \{2, 4, 5, 7, 8\} = \{2, 8\}$,

$K_A(3) = U \cap \{3, 7\} \cap \{1, 3, 6, 8\} = \{3\}$,

$K_A(4) = \{1, 4, 5, 8\} \cap \{1, 2, 4, 6, 8\} \cap \{2, 4, 5, 7, 8\} = \{4, 8\}$,

$K_A(5) = \{1, 4, 5, 8\} \cap U \cap \{2, 4, 5, 7, 8\} = \{4, 5, 8\}$,

$K_A(6) = \{6, 7\} \cap \{1, 2, 4, 6, 8\} \cap \{1, 3, 6, 8\} = \{6\}$,

$K_A(7) = \{6, 7\} \cap \{3, 7\} \cap \{2, 4, 5, 7, 8\} = \{7\}$, and

$K_A(8) = (\{1, 4, 5, 8\} \cup \{2, 8\}) \cap \{1, 2, 4, 6, 8\} \cap U = \{1, 2, 4, 8\}$.

Definability of Sets

A union of some intersections

of attribute-value pair blocks,

in any such intersection all attributes

should be different and attributes are members of B,

will be called B-*locally definable* sets.

Definability of Sets

A union of some intersections
of attribute-value pair blocks,
in any such intersection all attributes
should be different and attributes are members of B,
will be called B-*locally definable* sets.

A union of characteristic sets $K_B(x)$,
where $x \in X \subseteq U$ will be called
a B-*globally definable* set.

Definability of Sets

A union of some intersections

of attribute-value pair blocks,

in any such intersection all attributes

should be different and attributes are members of B,

will be called B-*locally definable* sets.

A union of characteristic sets $K_B(x)$,

where $x \in X \subseteq U$ will be called

a B-*globally definable* set.

Any set X that is *B*-globally definable

is *B*-locally definable,

the converse is not true.

Singleton Approximations

$$\underline{B}X = \{x \in U \mid K_B(x) \subseteq X\},$$

$$\overline{B}X = \{x \in U \mid K_B(x) \cap X \neq \emptyset\}.$$

Singleton Approximations

$$\underline{B}X = \{x \in U \mid K_B(x) \subseteq X\},$$

$$\overline{B}X = \{x \in U \mid K_B(x) \cap X \neq \emptyset\}.$$

$$\underline{A}\{1,2,4,8\} = \{1,2,4,8\},$$

$$\underline{A}\{3,5,6,7\} = \{3,6,7\},$$

$$\overline{A}\{1,2,4,8\} = \{1,2,4,5,8\},$$

$$\overline{A}\{3,5,6,7\} = \{3,5,6,7\}.$$

Singleton Approximation and Definability

{3, 5, 6, 7} = $\overline{A}\{3, 5, 6, 7\}$ is not A-locally definable—

no way to separate cases: 5 from 4 and 8:

[(Temperature, high)] = {1, 4, 5, 8},

[(Temperature, very_high)] = {2, 8},

[(Temperature, normal)] = {6, 7},

[(Headache, yes)] = {1, 2, 4, 6, 8},

[(Headache, no)] = {3, 7},

[(Nausea, no)] = {1, 3, 6, 8},

[(Nausea, yes)] = {2, 4, 5, 7, 8}.

Subset Approximations

$$\underline{B}X = \cup\{K_B(x) \mid x \in U, K_B(x) \subseteq X\},$$

$$\overline{B}X = \cup\{K_B(x) \mid x \in U, K_B(x) \cap X \neq \emptyset\}.$$

Subset Approximations

$$\underline{B}X = \cup\{K_B(x) \mid x \in U, K_B(x) \subseteq X\},$$

$$\overline{B}X = \cup\{K_B(x) \mid x \in U, K_B(x) \cap X \neq \emptyset\}.$$

$$\underline{A}\{1,2,4,8\} = \{1,2,4,8\},$$

$$\underline{A}\{3,5,6,7\} = \{3,6,7\},$$

$$\overline{A}\{1,2,4,8\} = \{1,2,4,5,8\},$$

$$\overline{A}\{3,5,6,7\} = \{3,4,5,6,7,8\}.$$

Concept Approximations

$$\underline{B}X = \cup\{K_B(x) \mid x \in X, K_B(x) \subseteq X\},$$

$$\overline{B}X = \cup\{K_B(x) \mid x \in X, K_B(x) \cap X \neq \emptyset\} = \cup\{K_B(x) \mid x \in X\}.$$

Concept Approximations

$$\underline{B}X = \cup\{K_B(x) \mid x \in X, K_B(x) \subseteq X\},$$

$$\overline{B}X = \cup\{K_B(x) \mid x \in X, K_B(x) \cap X \neq \emptyset\} = \cup\{K_B(x) \mid x \in X\}.$$

$$\overline{A}\{1, 2, 4, 8\} = \{1, 2, 4, 8\},$$

$$\overline{A}\{3, 5, 6, 7\} = \{3, 4, 5, 6, 7, 8\}.$$

Local Approximations

A set T of attribute-value pairs, where all attributes belong to the set B and are distinct, will be called a *B-complex*. A

B-*local lower* approximation of the concept X is defined as follows

$$\cup\{[T] \mid T \ is \ a \ \textit{B-complex} \ of \ X, \ [T] \subseteq X\}.$$

A B-*local upper* approximation of the concept X is defined as the minimal set containing X and defined in the following way

$$\cup\{[T] \mid \exists \ a \ family \ \mathcal{T} \ of \ \textit{B-complexes} \ of \ X \\ with \ \forall \ T \in \mathcal{T}, \ [T] \cap X \neq \emptyset\}.$$

Data Sets

Data set	Number of		
	cases	attributes	concepts
Breast cancer (Slovenia)	277	9	2
Hepatitis	155	19	2
Image segmentation	210	19	7
Lymphography	148	18	4
Wine	178	13	3

Incomplete Data Sets

For every data set a set of templates was created.

Templates were formed by replacing incrementally (with 5% increment) existing specified attribute values by lost values.

We started each series of experiments with no lost values, then we added 5% of lost values, then we added additional 5% of lost values, etc., until at least one entire row of the data sets was full of lost values.

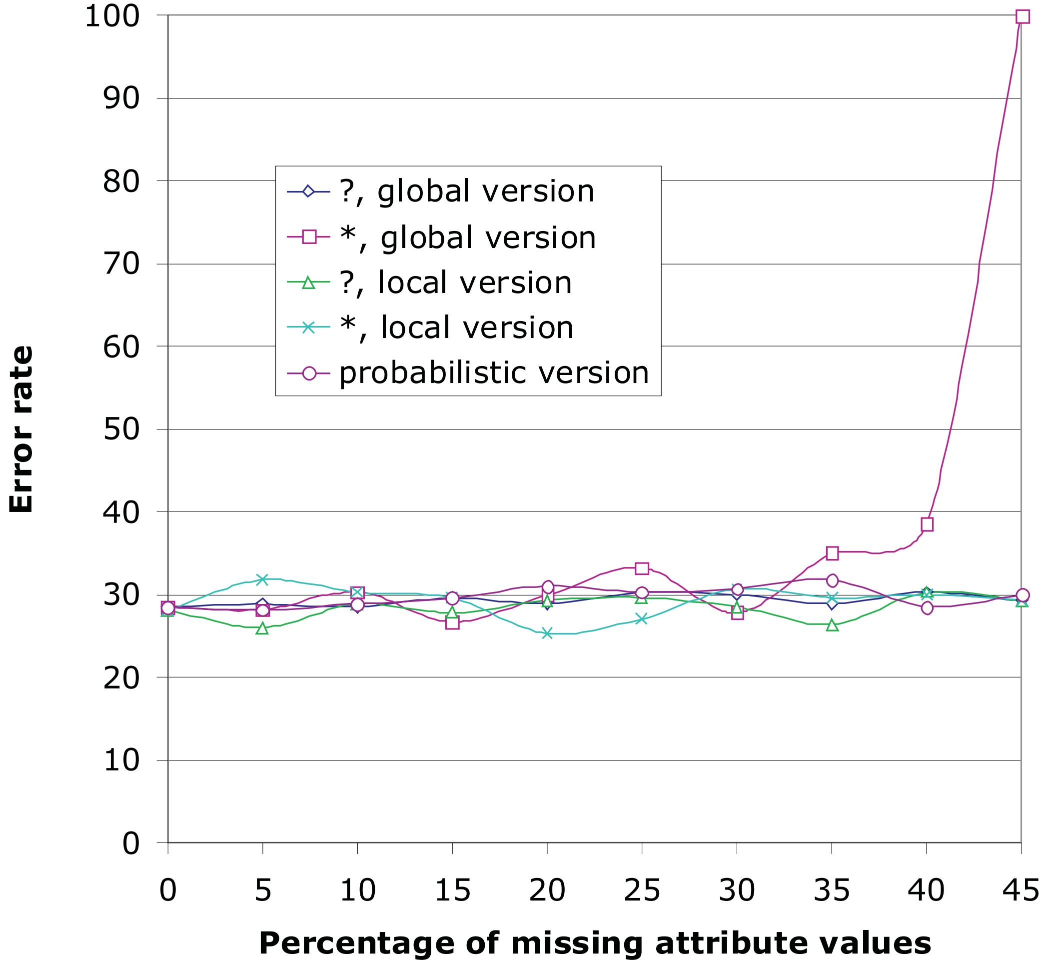
Then three attempts were made to change configuration of new lost values and either a new data set with extra 5% of lost values was created or the process was terminated.

For example, for the *breast cancer* data set that limit was 45% (in all three attempts with 50% of lost values, at least one row was full of lost values).

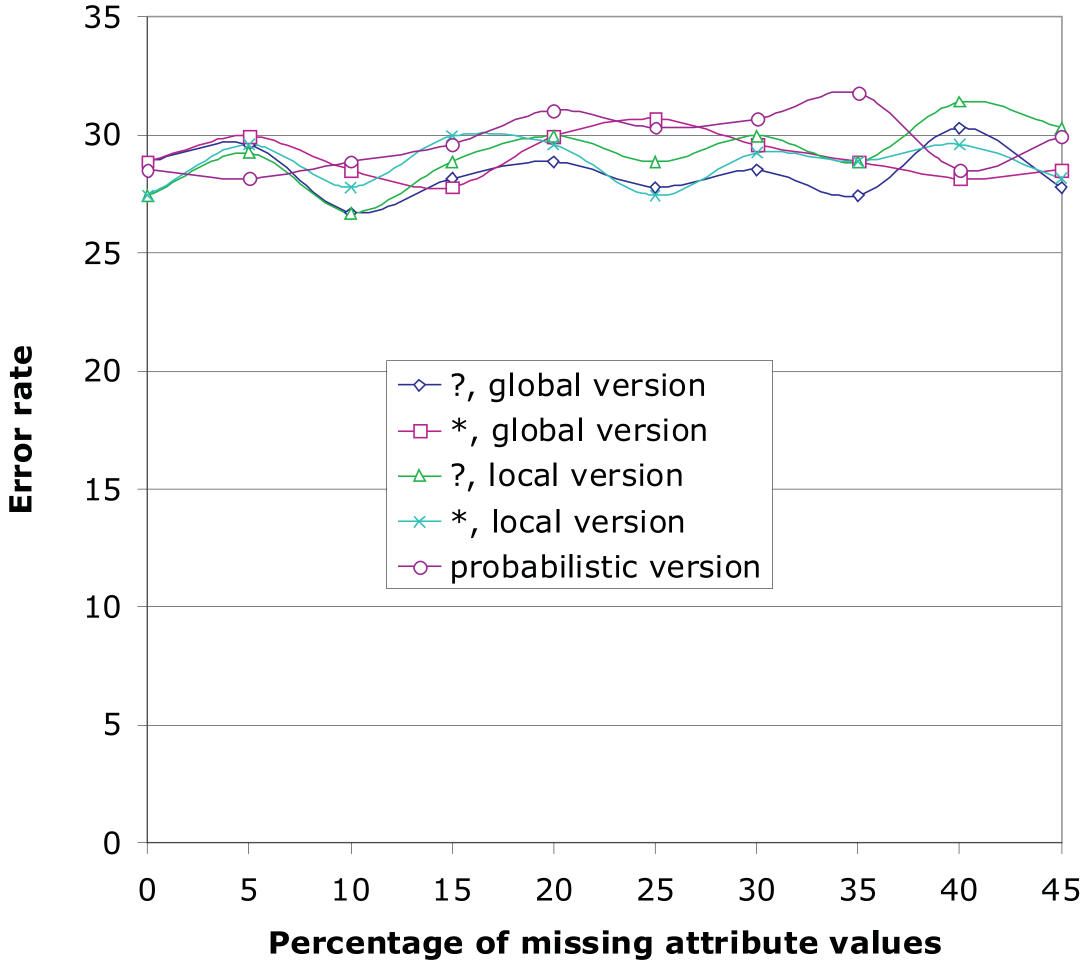
A pattern of 40% missing attribute values

Case	Temperature	Headache	Nausea	Flu
1	high	?	?	yes
2	very_high	?	yes	yes
3	?	no	no	no
4	high	?	yes	yes
5	?	?	yes	no
6	normal	yes	?	no
7	normal	no	yes	no
8	?	yes	?	yes
9	?	no	yes	yes
10	very_high	no	?	yes

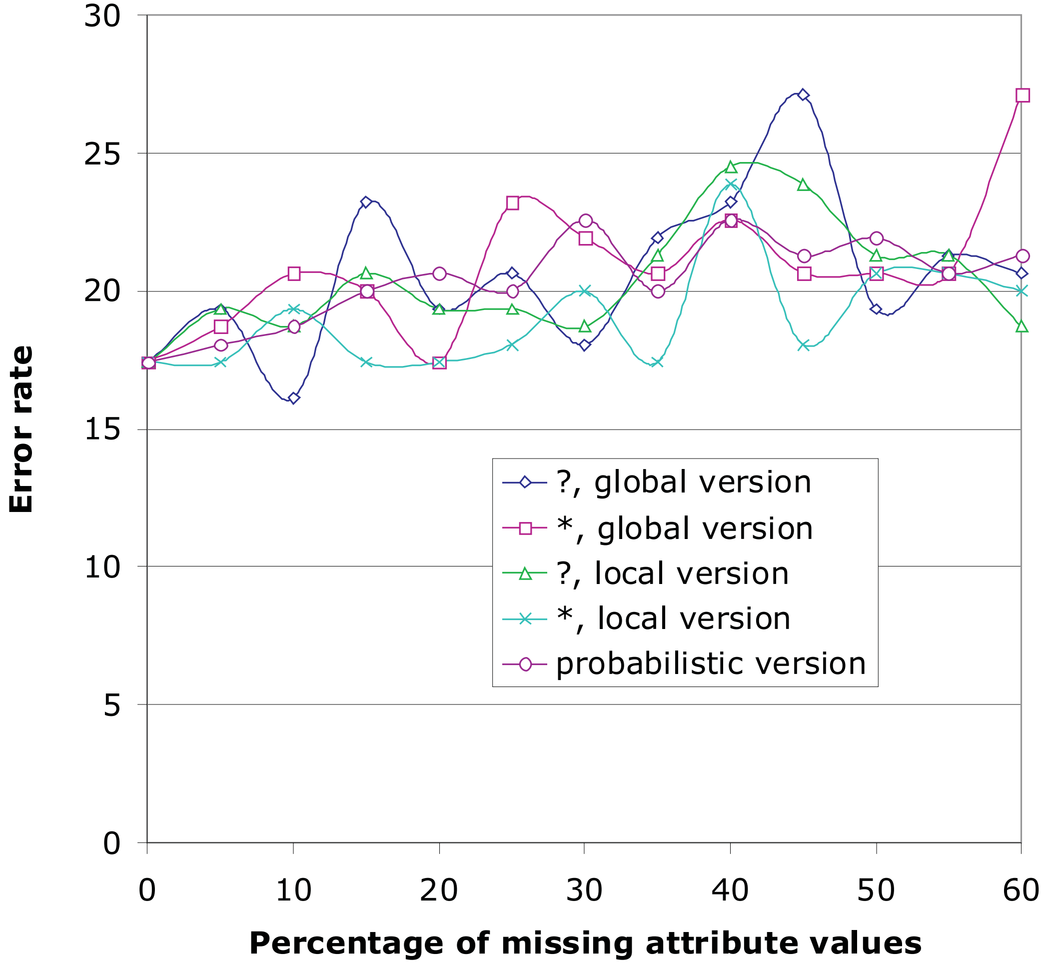
Breast Cancer, Certain Rules



Breast Cancer, Possible Rules



Hepatitis, Certain Rules



Hepatitis, Possible Rules

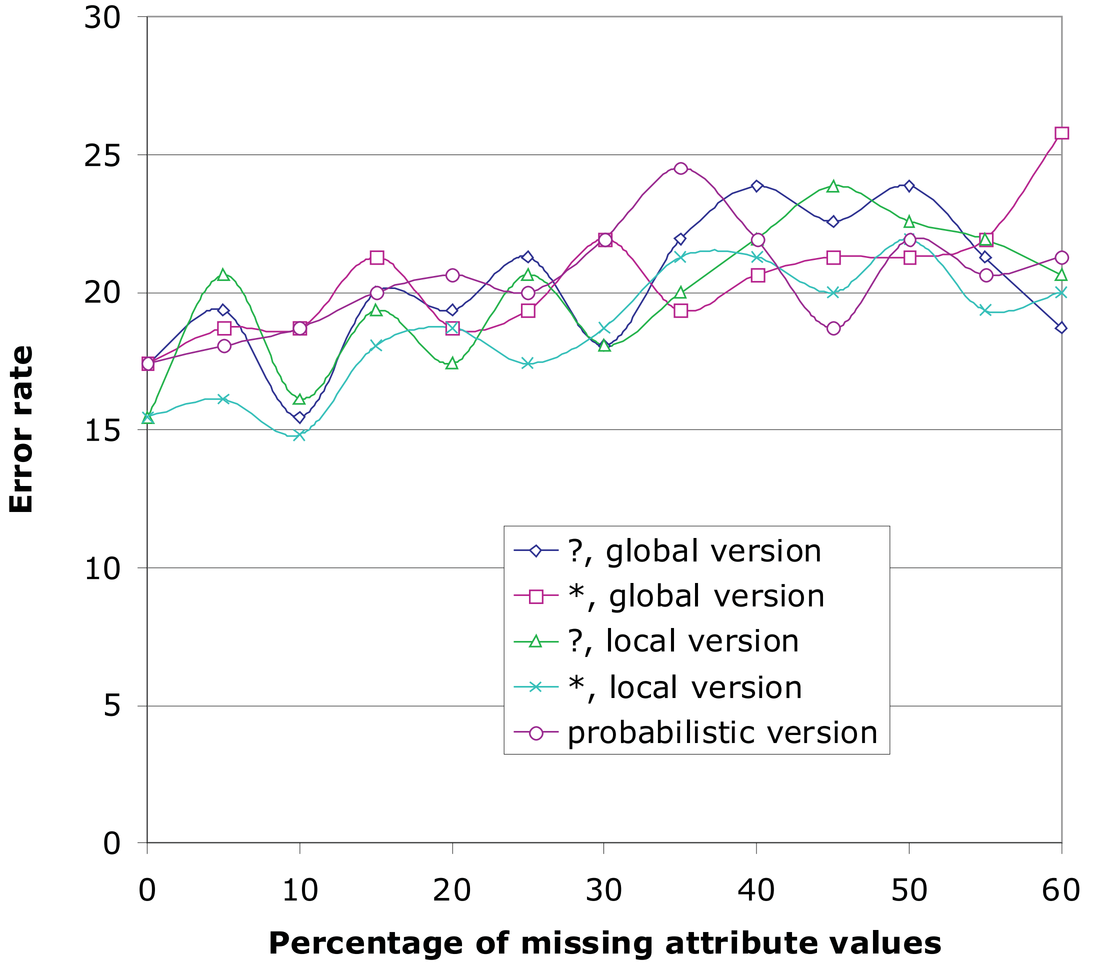


Image Segmentation, Certain Rules

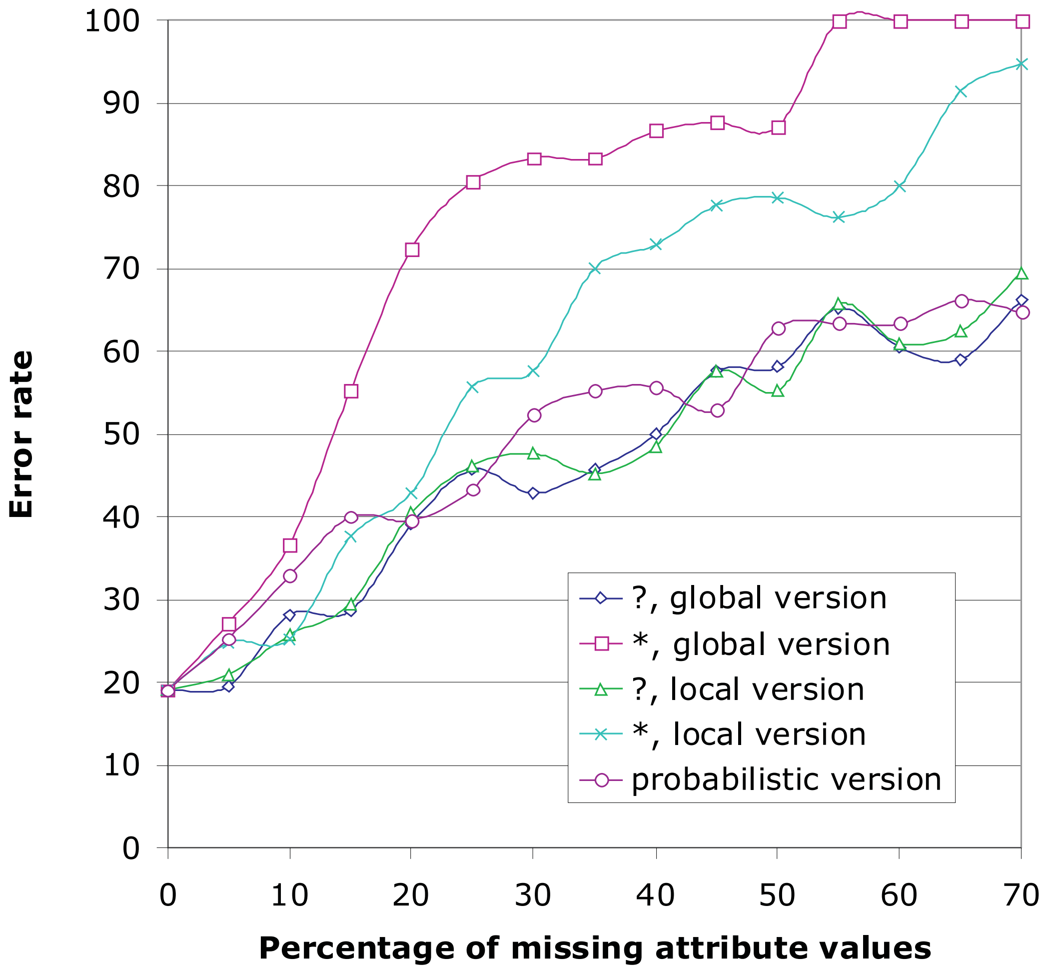
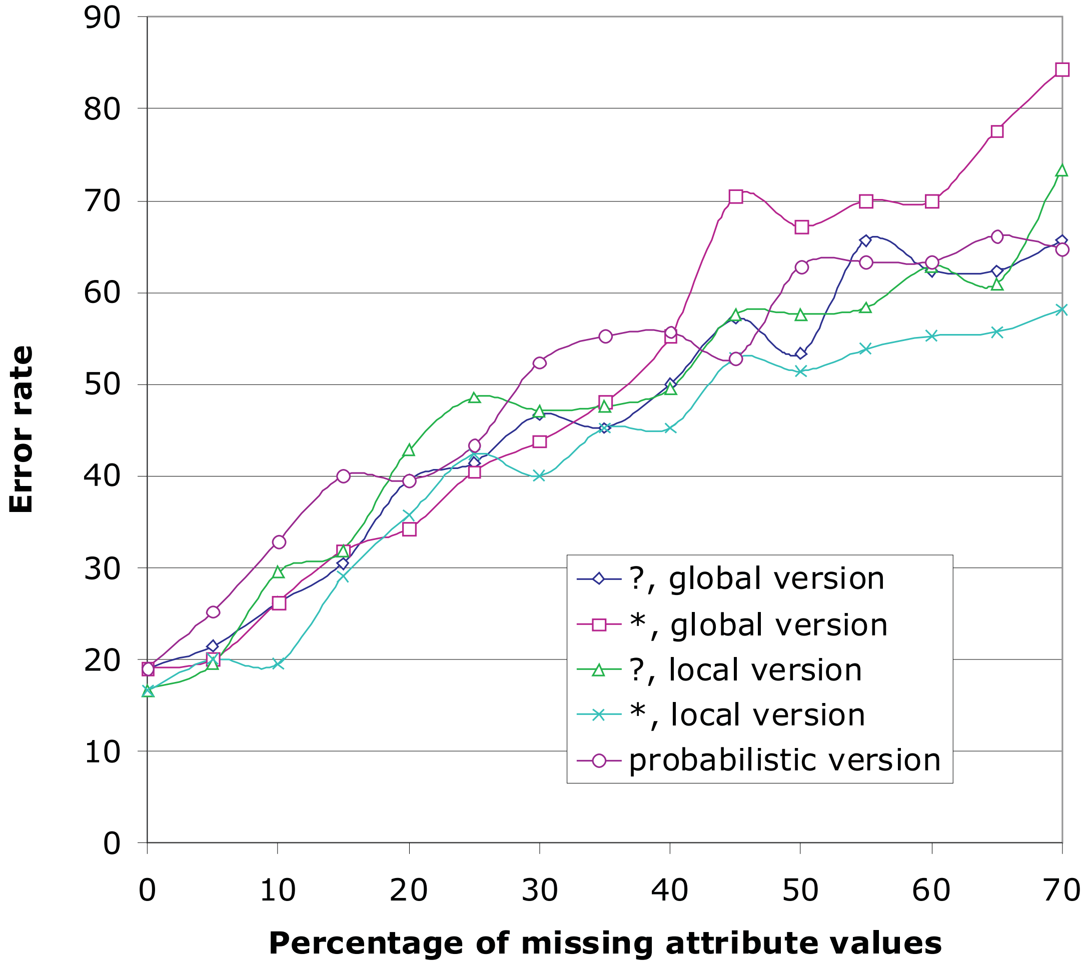
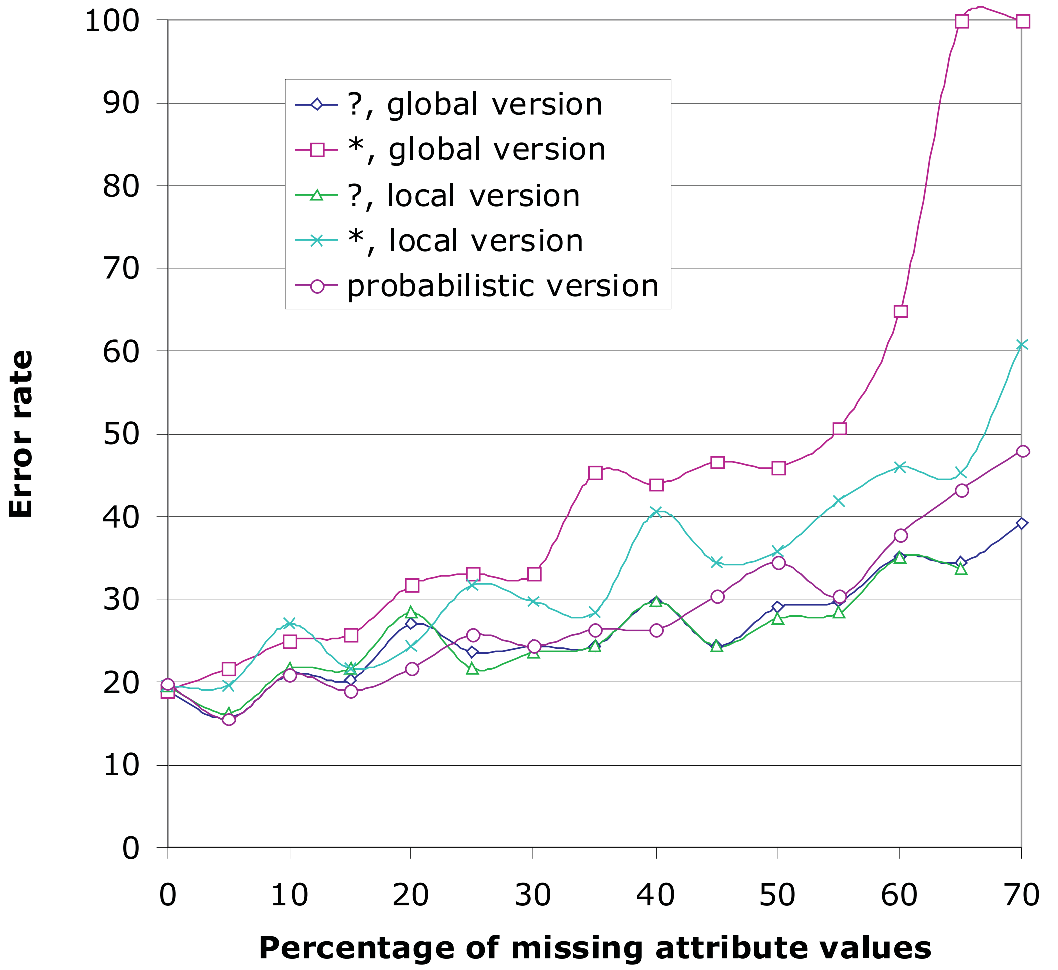


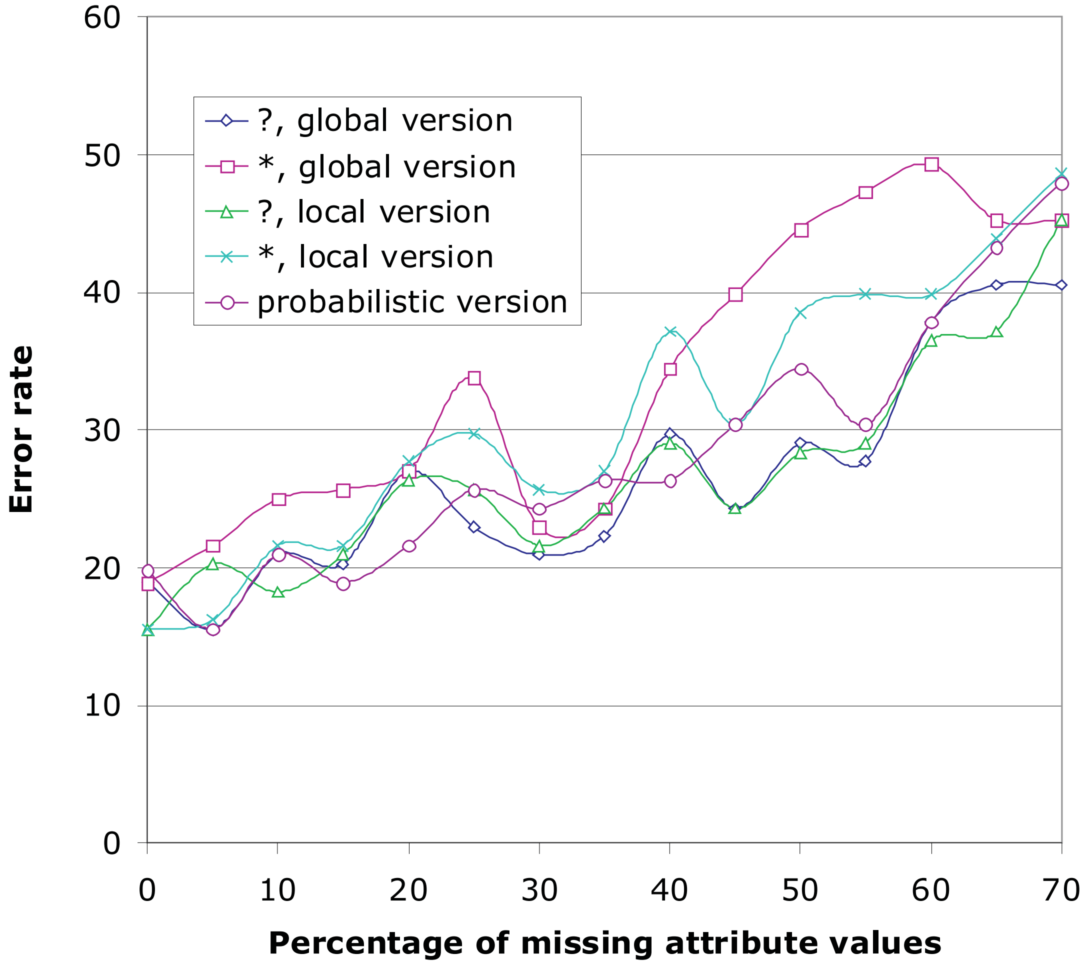
Image Segmentation, Possible Rules



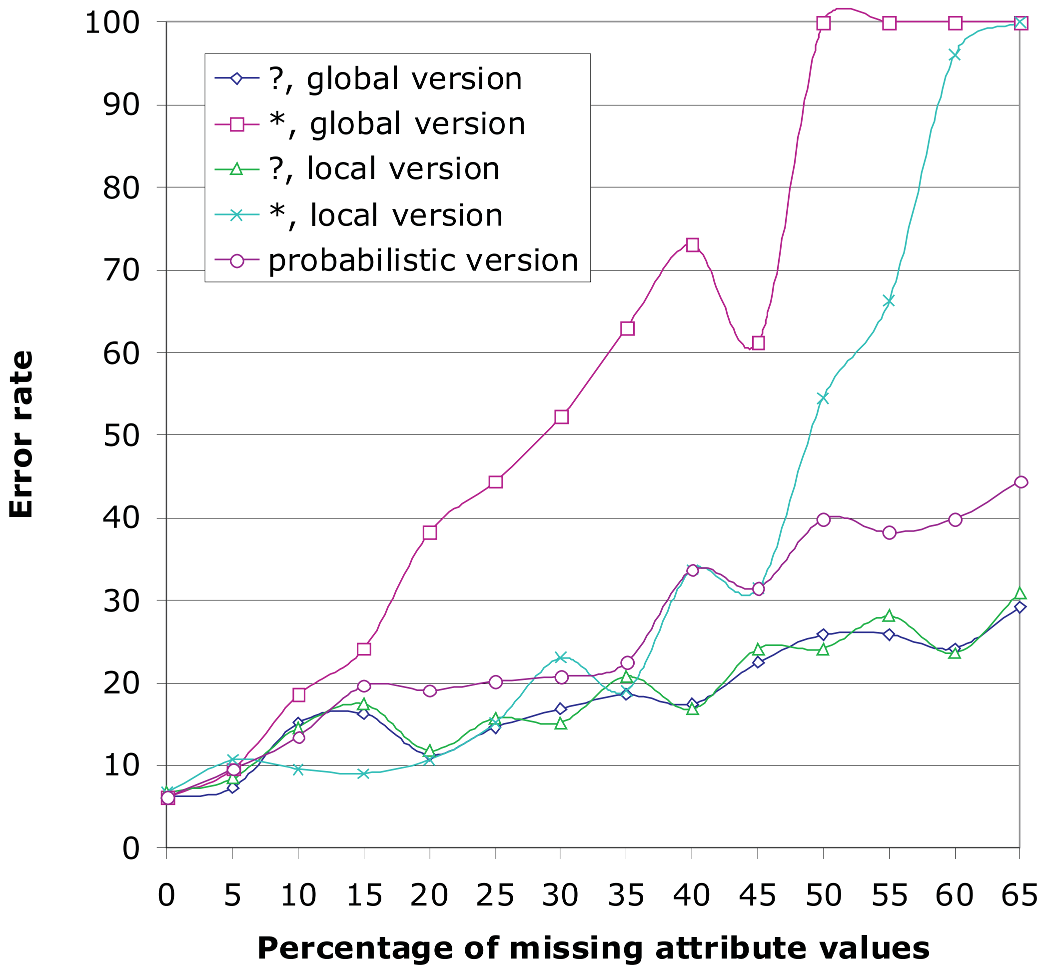
Lymphography, Certain Rules



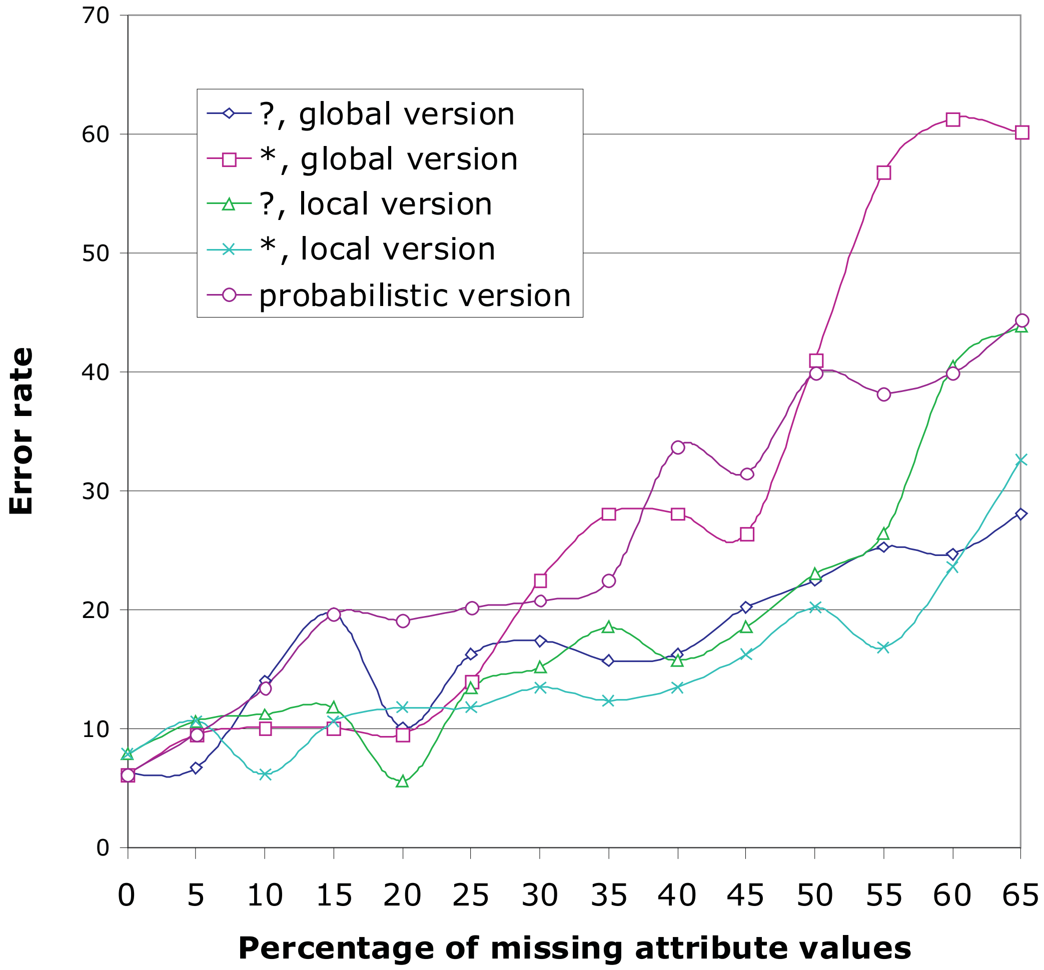
Lymphography, Possible Rules



Wine, Certain Rules



Wine, Possible Rules



Error rates

Data set	Error rate in percent	
	better of MCV-AV and CMCV-CAV	best rough set approach
Breast cancer	29.96	27.08
Echocardiogram	22.97	27.03
Hepatitis	19.35	17.42
Image segmentation	64.76	63.81
Lymphography	41.22	37.84
Wine recognition	31.46	26.97

Breast Cancer (Slovenia) Data Set

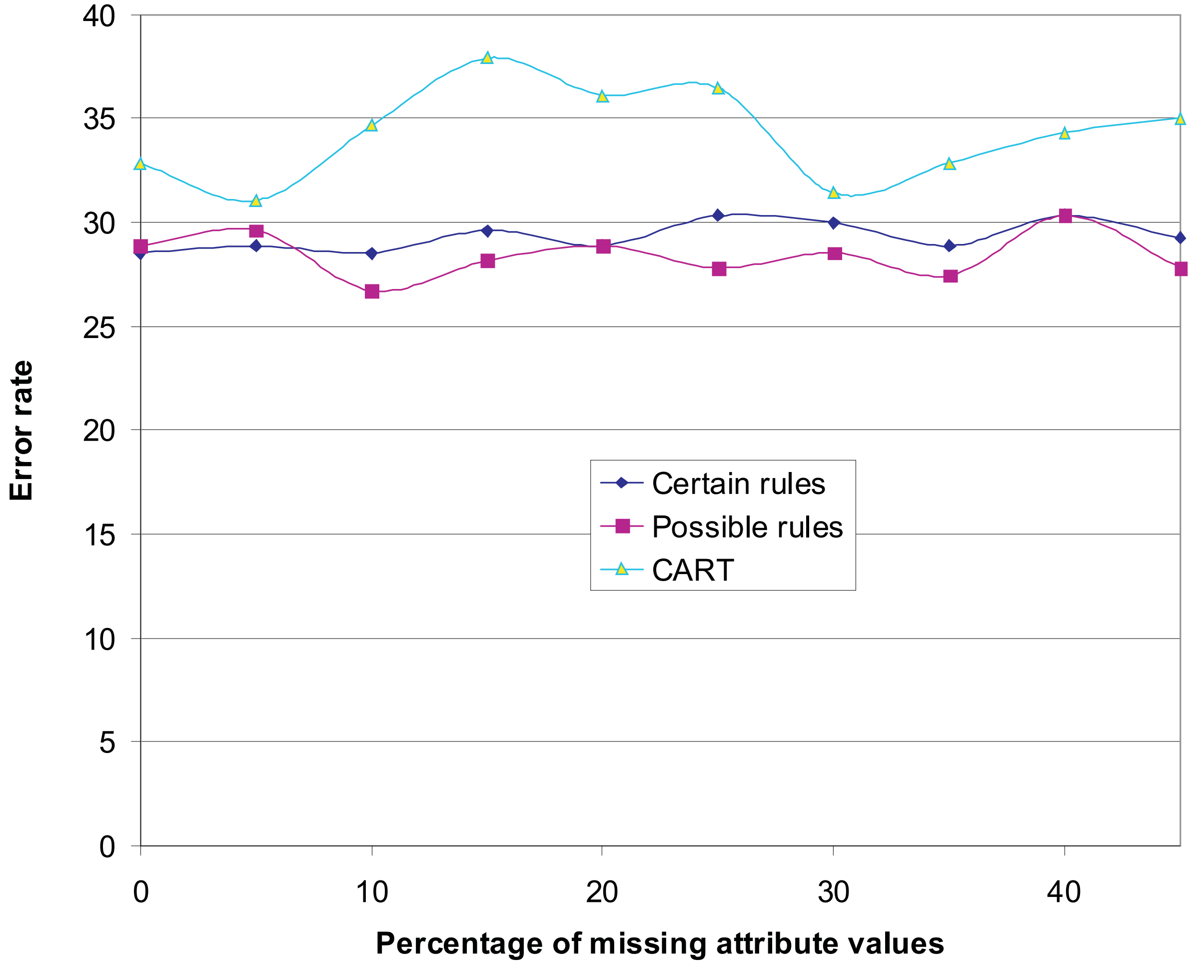
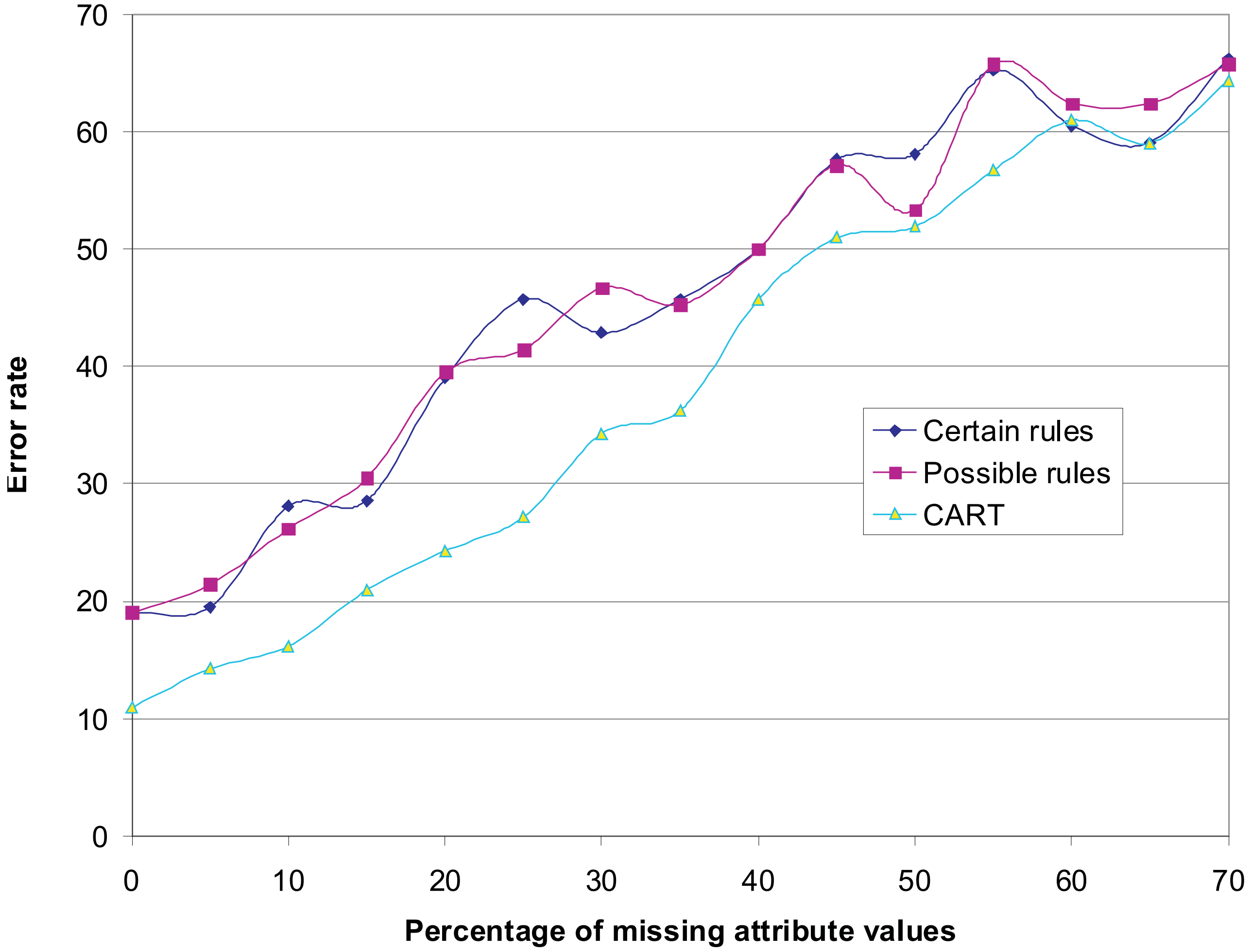
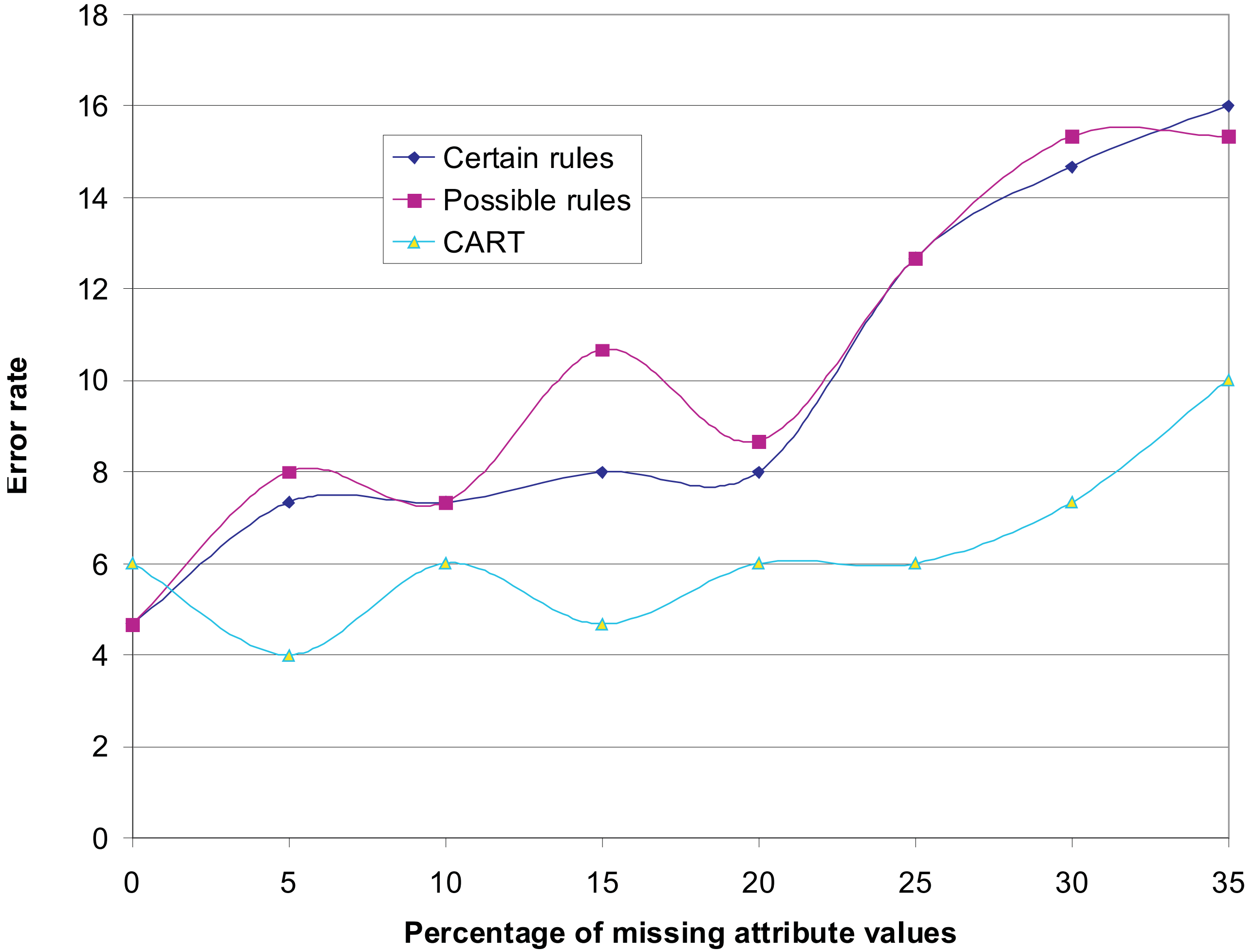


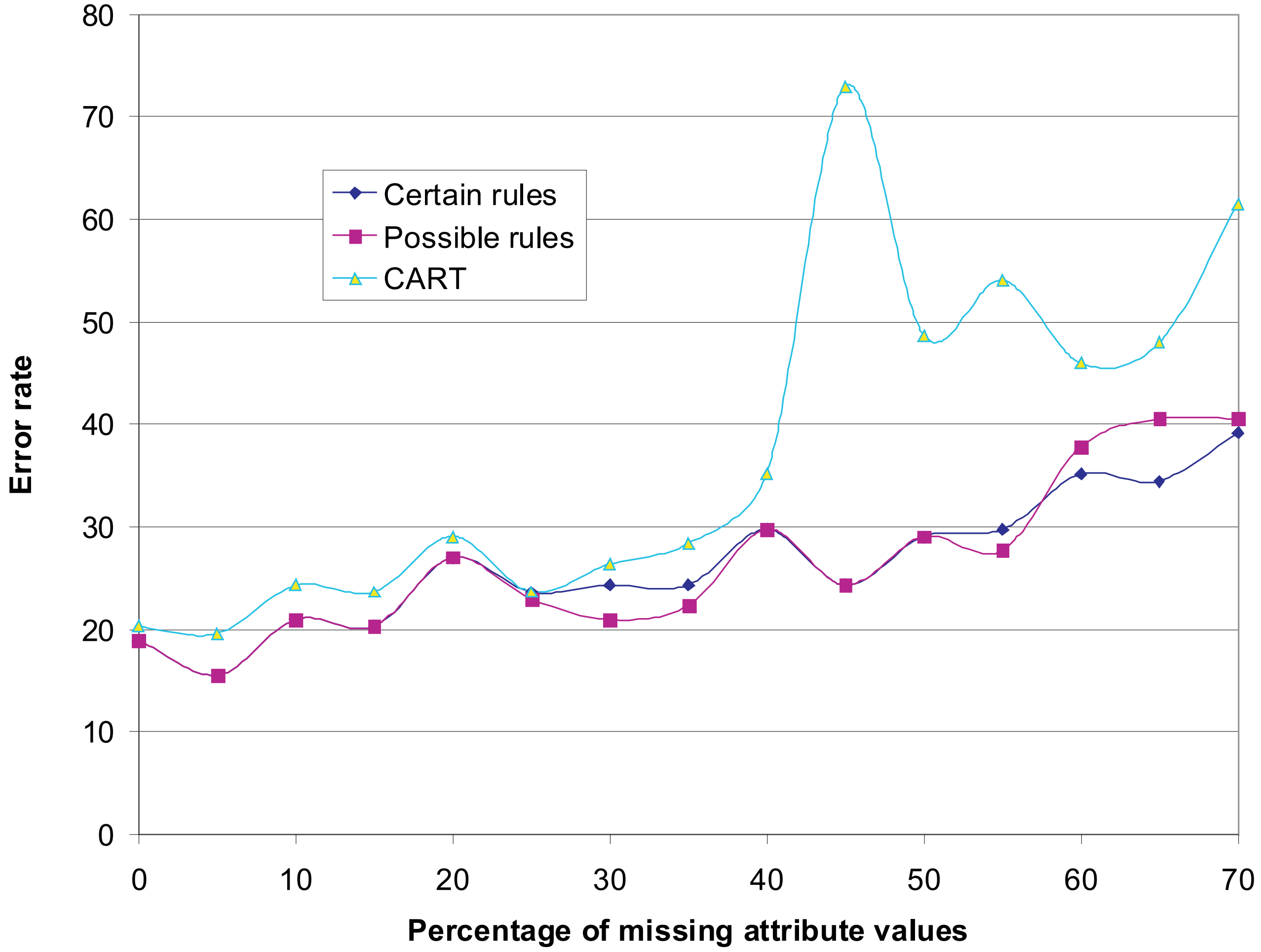
Image Segmentation Data Set



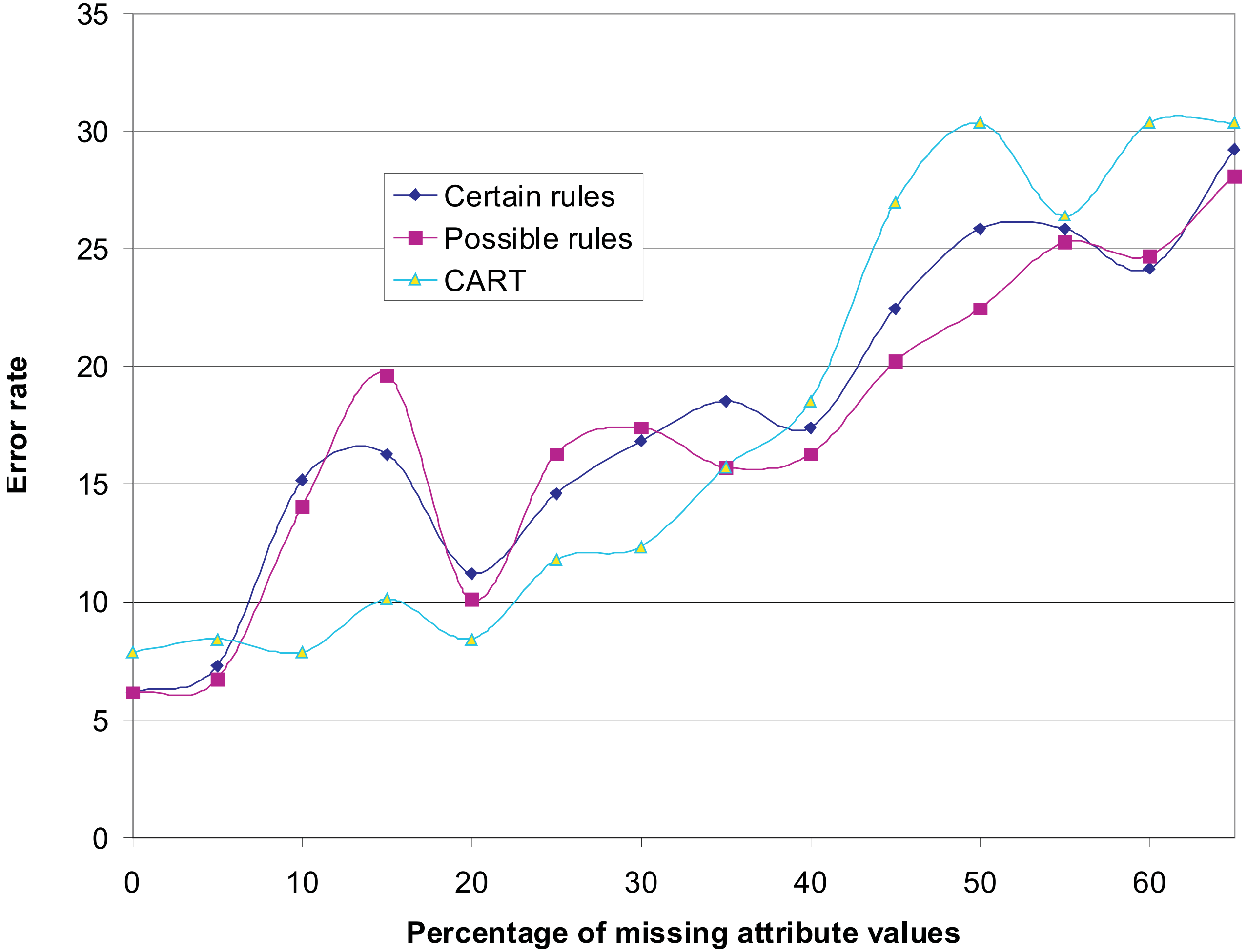
Iris Data Set



Lymphography Data Set



Wine Data Set



Wine data set, lost values, certain rule sets

Percentage of lost values	Average error rate	Standard deviation	Z score
0	7.66	1.32	
5	7.17	1.74	1.22
10	7.13	2.00	1.20
15	8.76	1.85	–2.66
20	7.06	1.38	1.72
25	7.27	1.55	1.06
30	6.20	1.39	4.17
35	6.55	1.16	3.43
40	6.8	1.28	2.56
45	7.73	1.48	–0.21

Wine data set, lost values, possible rule sets

Percentage of lost values	Average error rate	Standard deviation	Z score
0	7.66	1.32	
25	7.64	1.63	0.05
30	6.33	1.15	4.15
35	6.57	1.12	3.44
40	6.22	1.29	4.27
45	7.79	1.30	–0.39
50	7.12	0.68	2.00
55	7.68	0.98	–0.06
60	6.89	0.78	2.74
65	8.31	1.17	–2.04

Some conclusions

- An interpretation of the *lost values* seems to be the best approach to missing attribute values,
- An interpretation of the *"do not care" conditions* and certain rule sets is the worst approach,
- All three approaches: rough set, probabilistic and CART are comparable in terms of an error rate,
- For some data sets increasing incompleteness reduces the error rate.