Mining Incomplete Data—A Rough Set Approach

Jerzy W. Grzymala-Busse

jerzy@ku.edu

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

Lawrence, KS 66045, USA

and

Institute of Computer Science, Polish Academy of Sciences, 01-237 Warsaw, Poland

Outline of the talk

- Complete Data
  - Blocks of Attribute-Value Pairs
  - Elementary Sets
  - Indiscernibility Relation
- Incomplete Data
  - Sequential Methods
  - Parallel Methods
    - Characteristic Relation
    - Singleton, Subset, and Concept Approximations
    - Local Approximations
    - Rule Induction
    - Experiments
    - Conclusions

A Complete Decision Table

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

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 Complete Decision Table

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

Blocks of Attribute-Value Pairs

[(Temperature, high)] = {1, 3, 4, 5},

Blocks of Attribute-Value Pairs

[(Temperature, high)] = {1, 3, 4, 5},

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

Blocks of Attribute-Value Pairs

[(Temperature, high)] = {1, 3, 4, 5},

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

[(Temperature, normal)] = {6, 7, 8},

[(Headache, yes)] = {1, 2, 4, 5, 6, 8},

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

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

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

Elementary Sets of *B*

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

Elementary Sets of *A*

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

Elementary Sets of *A*

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

$[2]_A = [(Temperature, very\_high)] \cap [(Headache, yes)] \cap [(Nausea, yes)] = \{2\}$,

$[3]_A = [(Temperature, high)] \cap [(Headache, no)] \cap [(Nausea, no)] = \{3\}$,

$[4]_A = [5]_A = [(Temperature, high)] \cap [(Headache, yes)] \cap [(Nausea, yes)] = \{4, 5\}$,

$[6]_A = [8]_A = [(Temperature, normal)] \cap [(Headache, yes)] \cap$

$[(Nausea, no] = \{6, 8\}$,

$[7]_A = [(Temperature, normal)] \cap [(Headache, no] \cap [(Nausea, yes)] = \{7\}$.

Indiscernibility Relation

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

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

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

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

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

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

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

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, 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

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

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

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

LERS

- A data mining system LERS (Learning from Examples based on Rough Sets) computes lower and upper approximations of all concepts,
- Rules induced from the lower approximation of the concept are called certain,
- Rules induced from the upper approximation of the are called possible,
- LEM2 explores the search space of attribute-value pairs,
- MLEM2, a modified version of LEM2.

Minimal Complex

Let X be a subset of U.

Let T be a set of attribute-value pairs $t = (a, v)$.
Set X *depends* on set T if and only if

$$\emptyset \neq \cap\{[t] \mid t \in T\} \subseteq X$$

Set T is a minimal complex of X if and only if

- X depends on T and
- No proper subset T' of T exists such that X depends on T'.

Local Coverings

Let $\mathcal{T}$ be a nonempty collection of nonempty sets of attribute-value pairs.

$\mathcal{T}$ is a local covering of X if and only if the following conditions are satisfied:

- Each member T of $\mathcal{T}$ is a minimal complex of B,
- $\cup\{T | T \in \mathcal{T}\} = X$, and
- $\mathcal{T}$ is minimal, i.e., $\mathcal{T}$ has the smallest possible number of members.

The kernel of LEM2

- Select a pair $t \in T(G)$ such that $|[t] \cap G|$ is maximum,

The kernel of LEM2

- Select a pair $t \in T(G)$ such that $|[t] \cap G|$ is maximum,
- If a tie occurs, select a pair $t \in T(G)$ with the smallest cardinality of $[t]$,

The kernel of LEM2

- Select a pair $t \in T(G)$ such that $|[t] \cap G|$ is maximum,
- If a tie occurs, select a pair $t \in T(G)$ with the smallest cardinality of $[t]$,
- If another tie occurs, select the first pair.

Classification of Unseen Cases, I

A classification system uses the rule set, induced from training data set, to classify new cases.

The classification system of LERS is a modification of the bucket brigade algorithm

- Strength is the total number of examples correctly classified by the rule during training,

Classification of Unseen Cases, I

A classification system uses the rule set, induced from training data set, to classify new cases.

The classification system of LERS is a modification of the bucket brigade algorithm

- Strength is the total number of examples correctly classified by the rule during training,
- Support is defined as the sum strengths for all rules matching the concept.

Classification of Unseen Cases, II

Support:

$$\sum_{matching\ rules\ R\ describing\ C} Strength(R)$$

The concept C for which the support is the largest is a winner and the case is classified as being a member of C.

Classification of Unseen Cases, III

If complete matching is impossible, all partially matching rules are identified.

Matching factor (R) is equal to the ratio of the number of matched attribute-value pairs of the rule R to the total number of attribute-value pairs of the rule R. Support:

$$\sum_{\substack{partially\ matching \\ rules\ R\ describing\ C}} Matching(R) * Strength(R)$$

The concept C for which the support is the largest is a winner and the case is classified as being a member of C.

Classification of Unseen Cases, IV

Classification of testing cases

- if for an unseen case *x*, $a(x) = ?$, any condition (a, v) of a rule should not be matched,
- if for an unseen case *x*, $a(x) = *$ or $a(x) = -$, any condition (a, v) of a rule should be matched.

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.

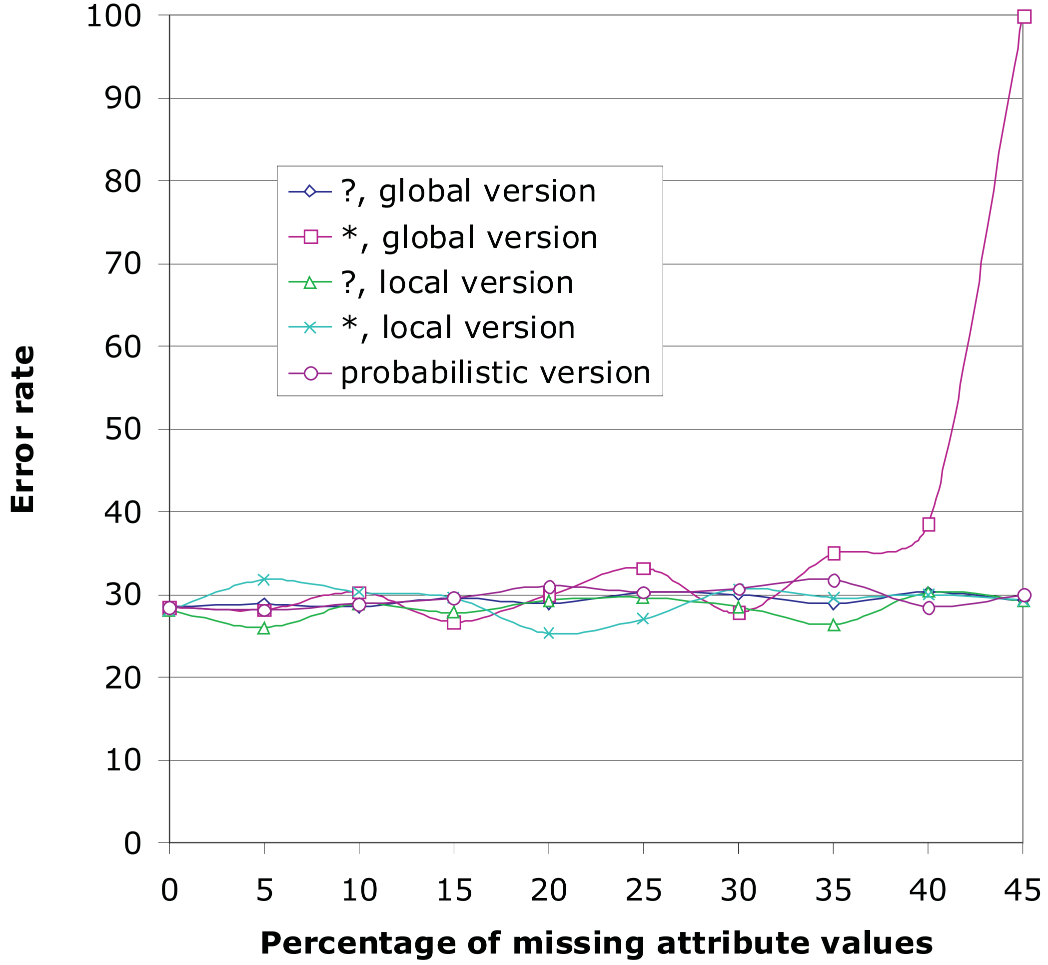
Then three attempts were made to change configuration of new lost values and either a new data set with extra 5% of lost values was created or the process was terminated.

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).

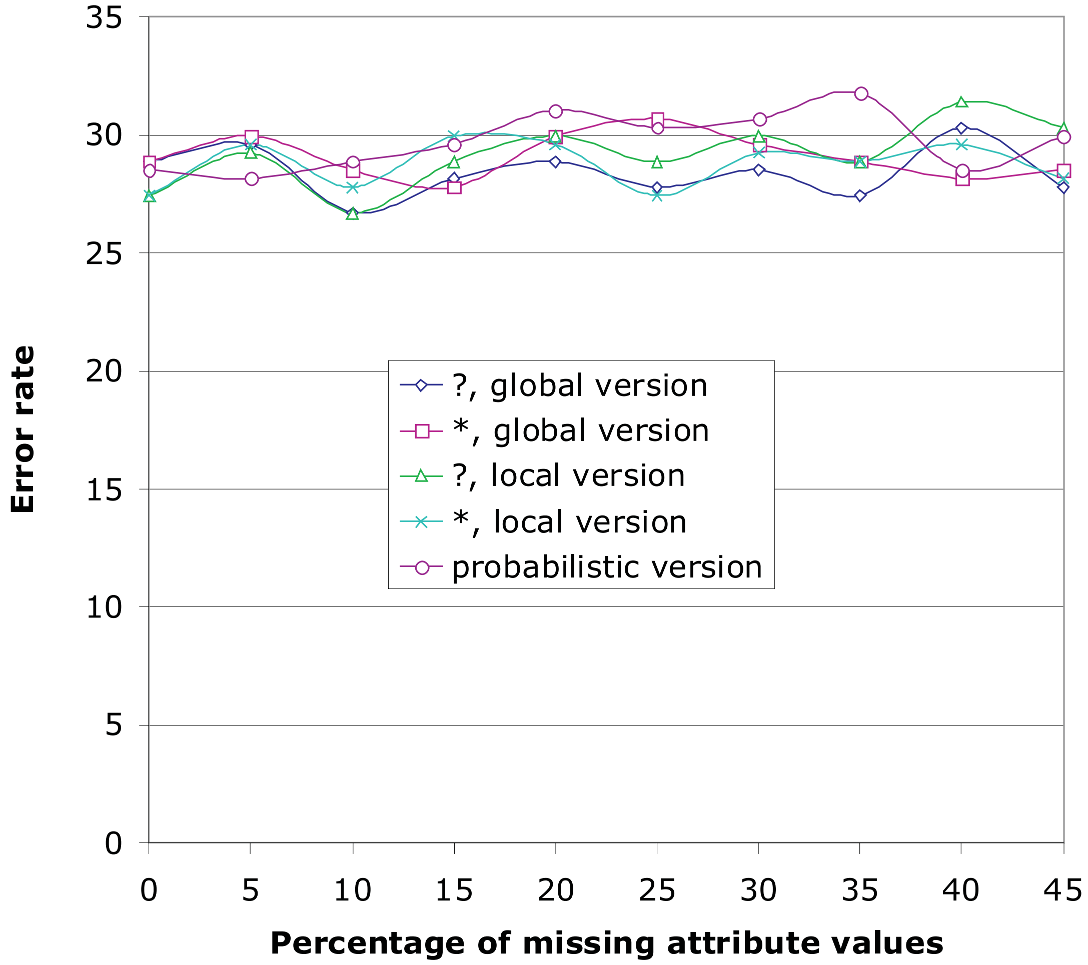
A pattern of 40% missing attribute values

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

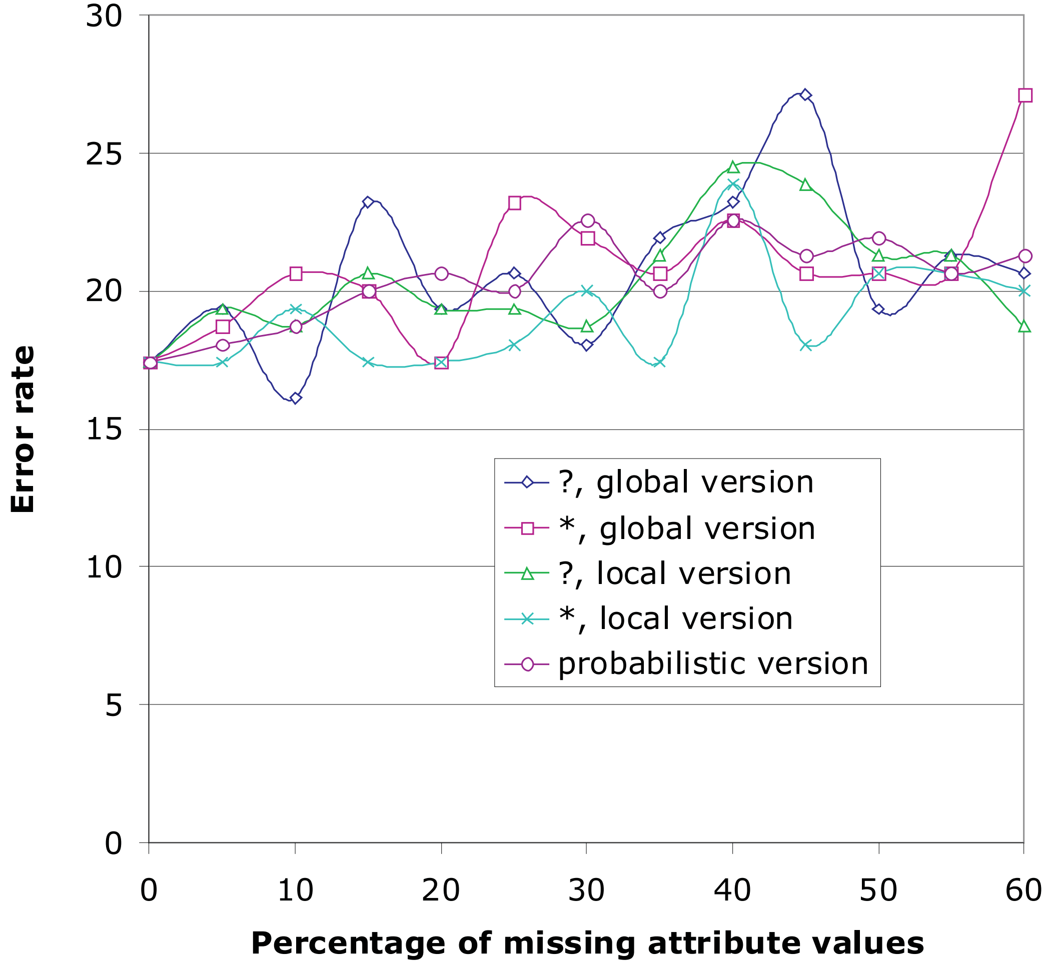
Breast Cancer, Certain Rules



Breast Cancer, Possible Rules



Hepatitis, Certain Rules



Hepatitis, Possible Rules

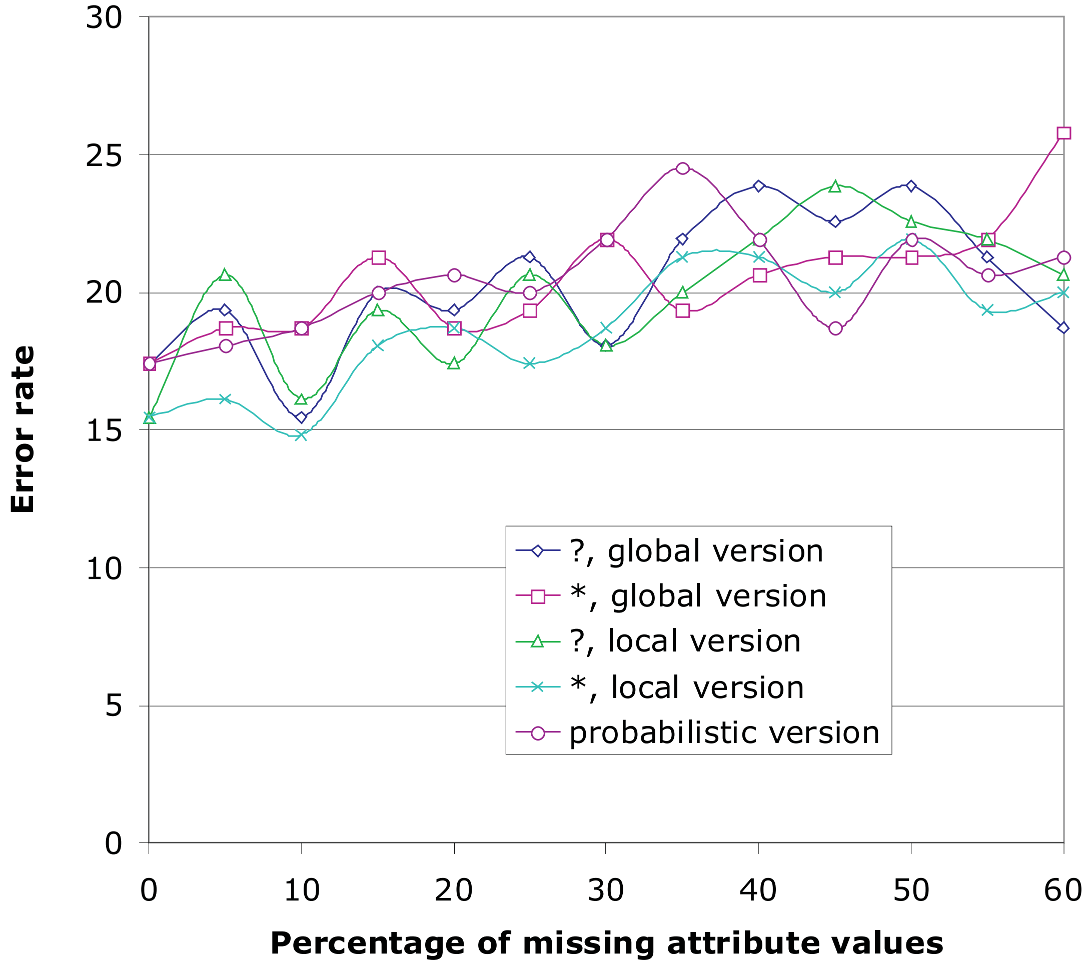


Image Segmentation, Certain Rules

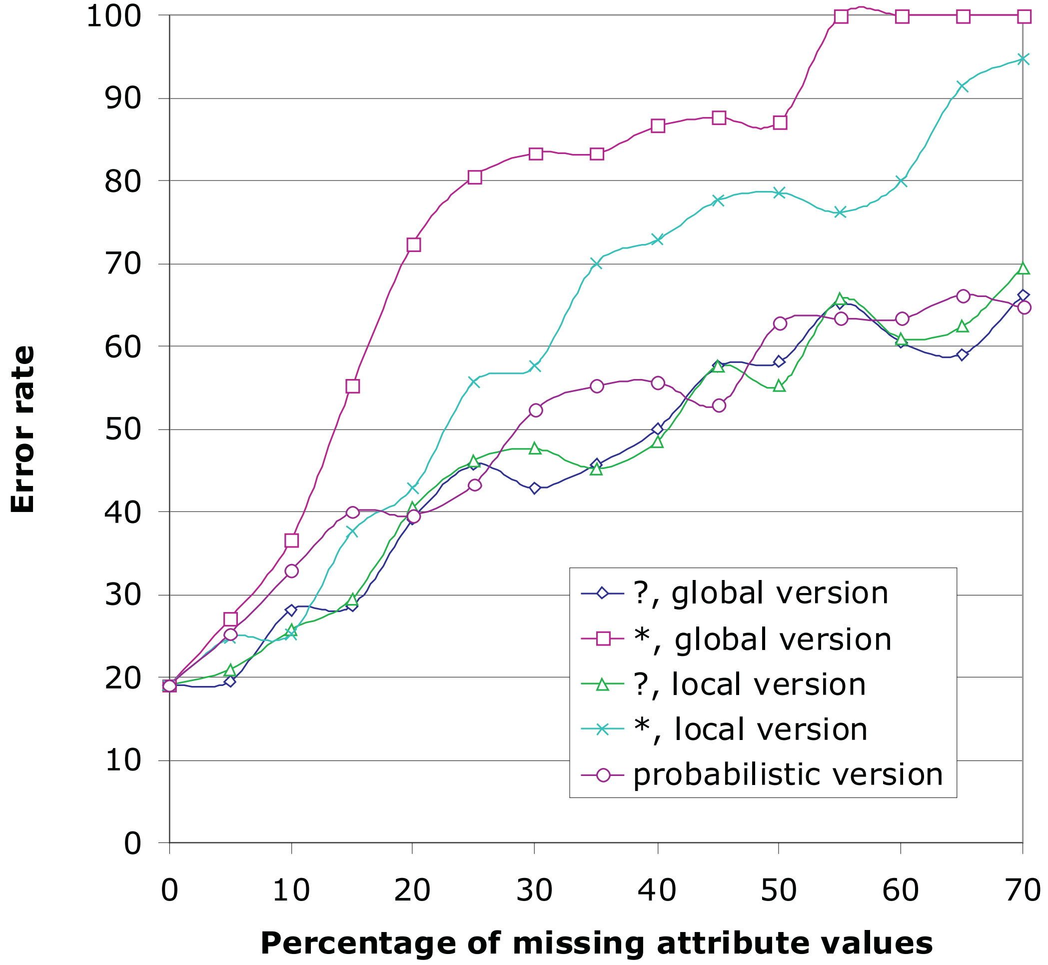
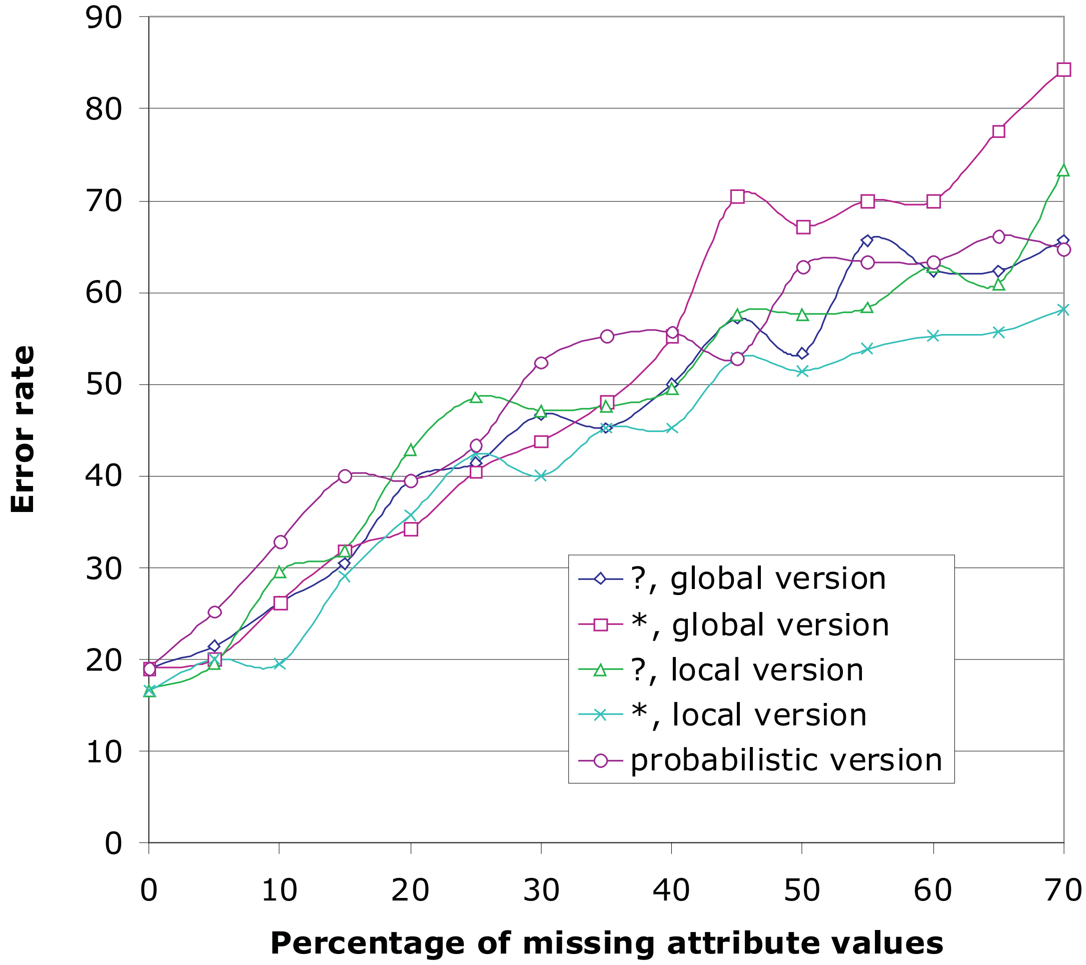
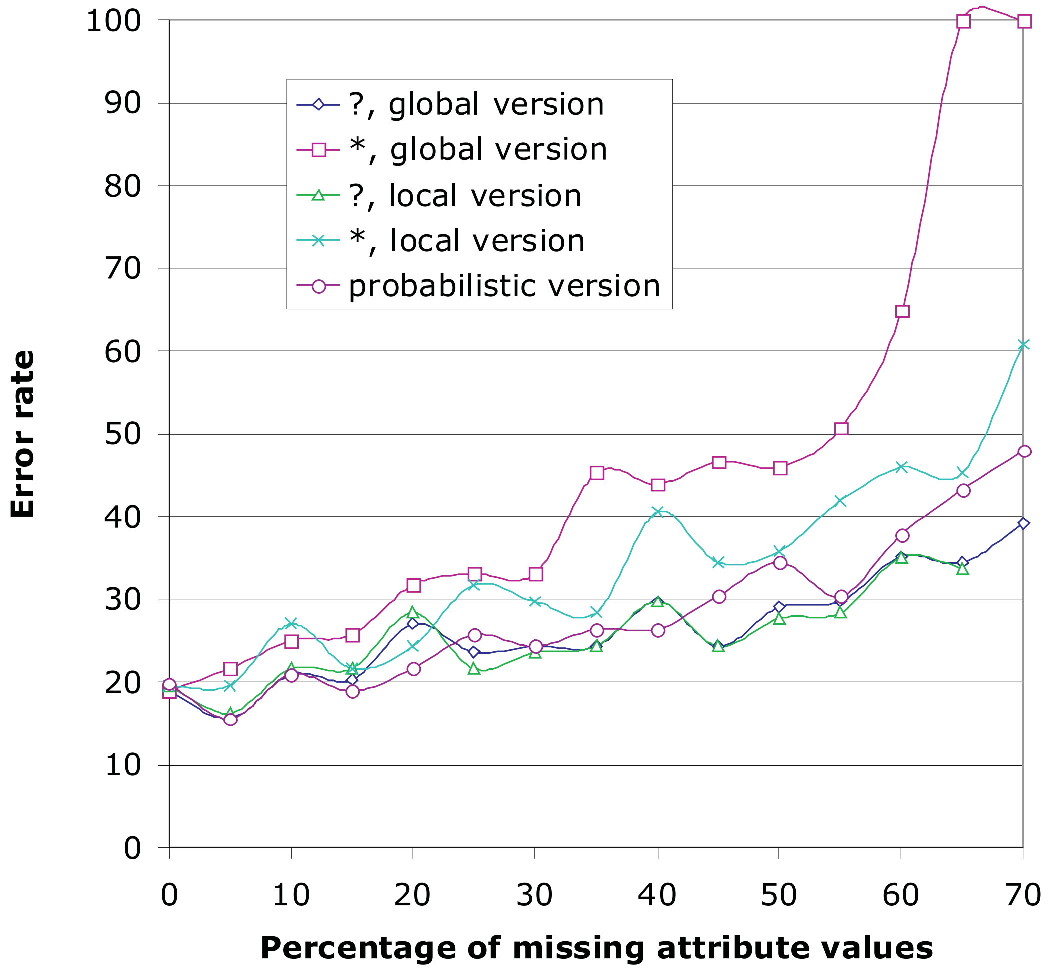


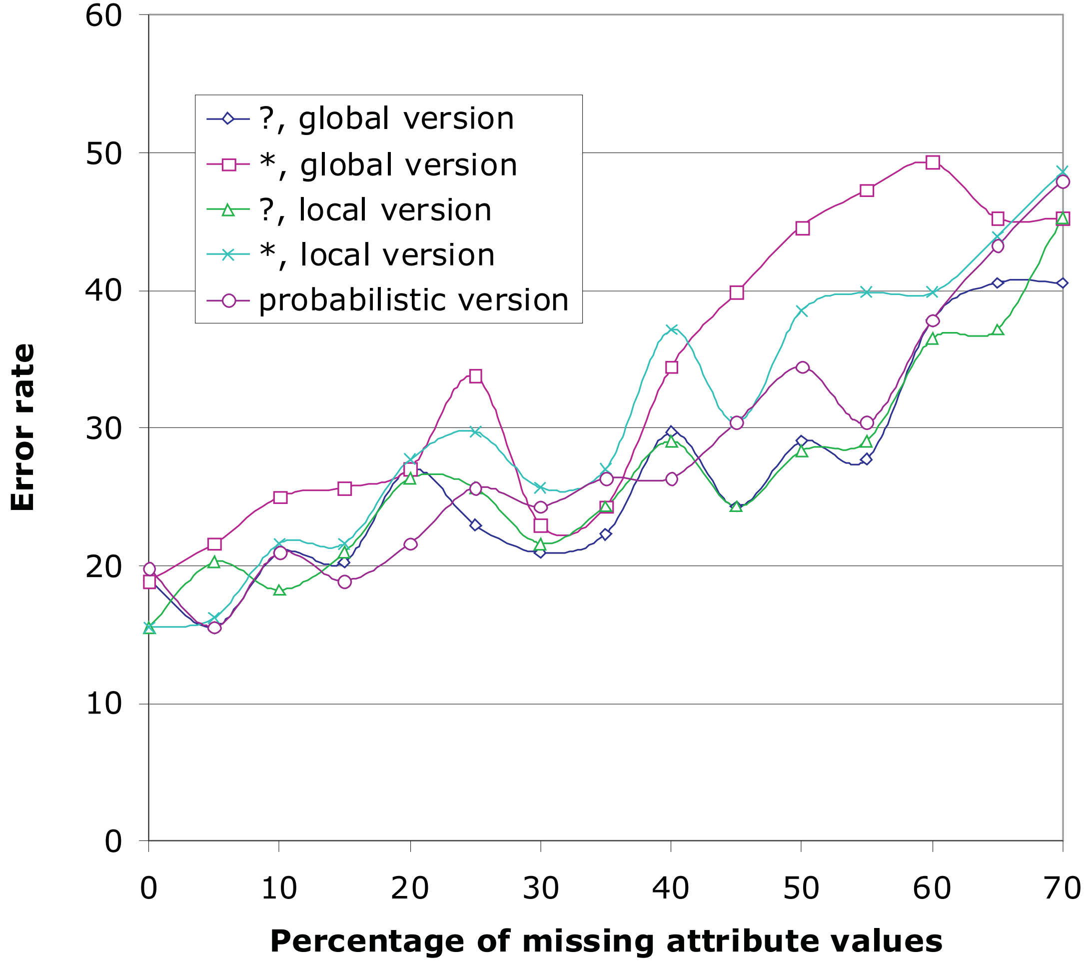
Image Segmentation, Possible Rules



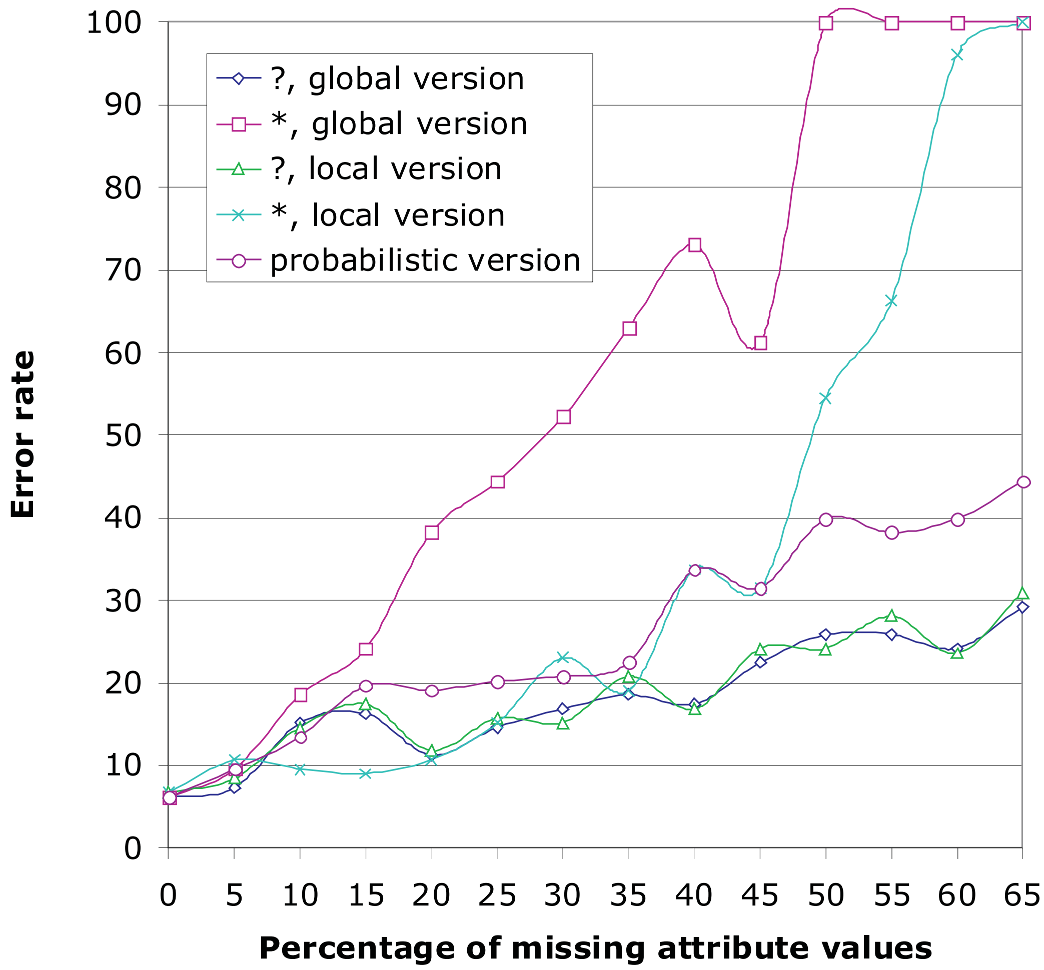
Lymphography, Certain Rules



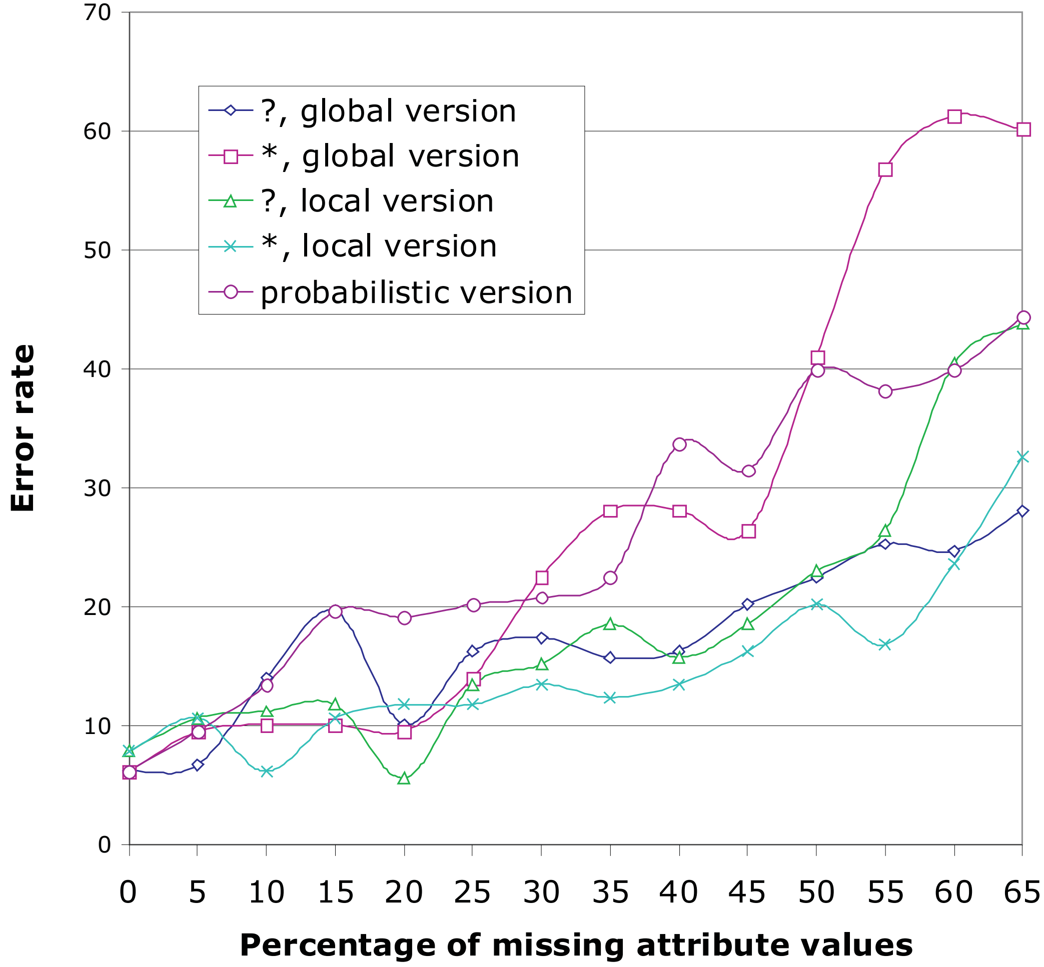
Lymphography, Possible Rules



Wine, Certain Rules



Wine, Possible Rules



Breast Cancer (Slovenia) Data Set

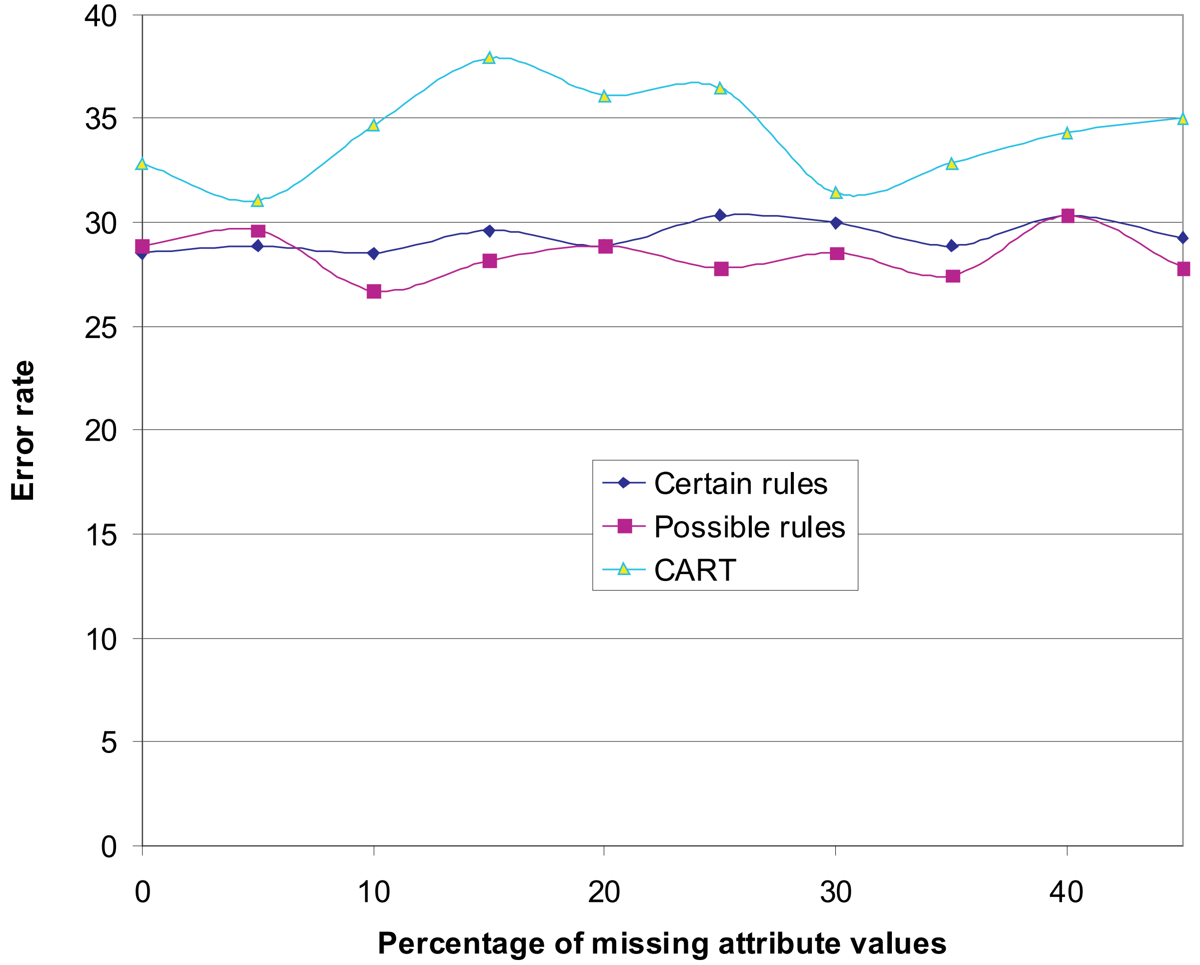
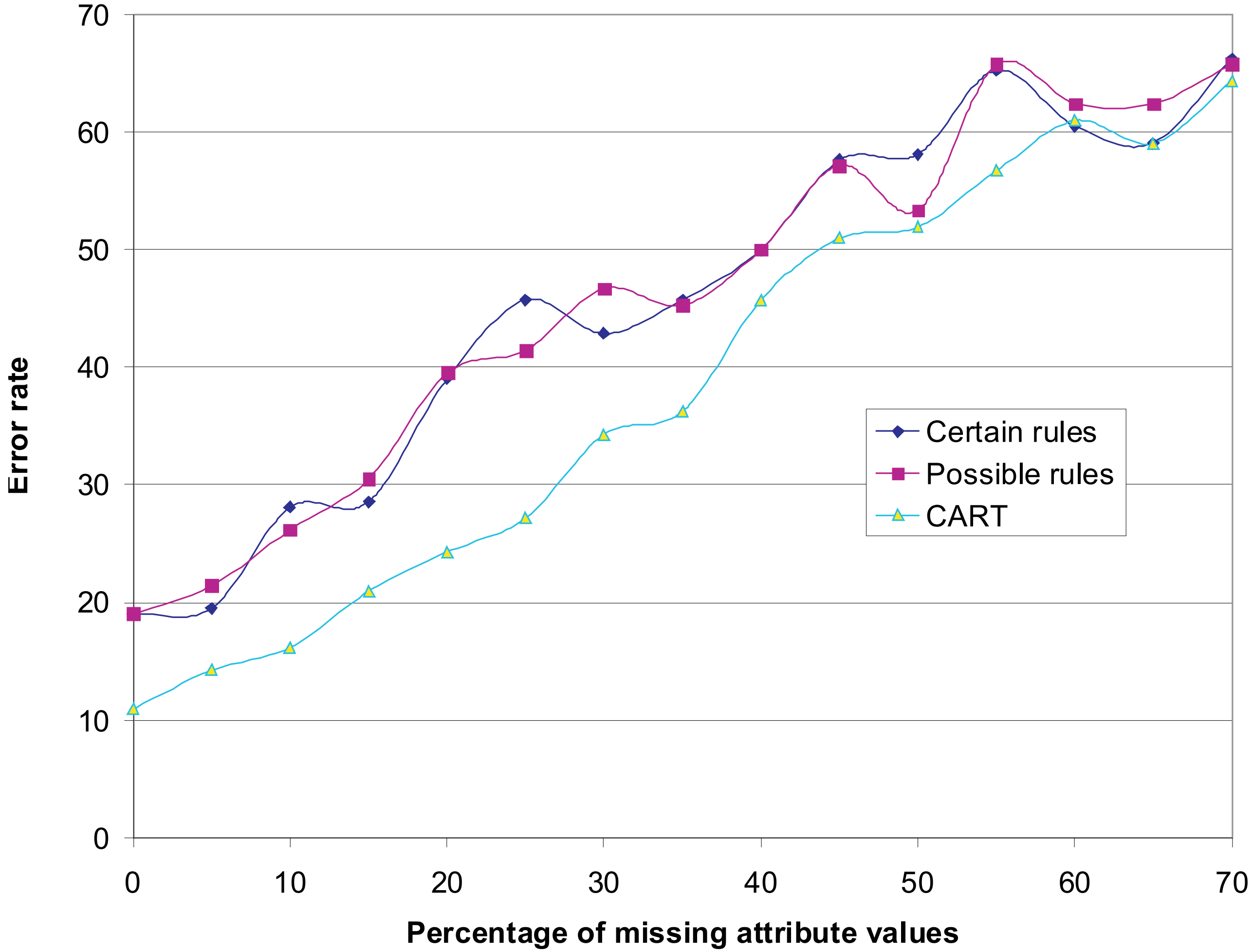
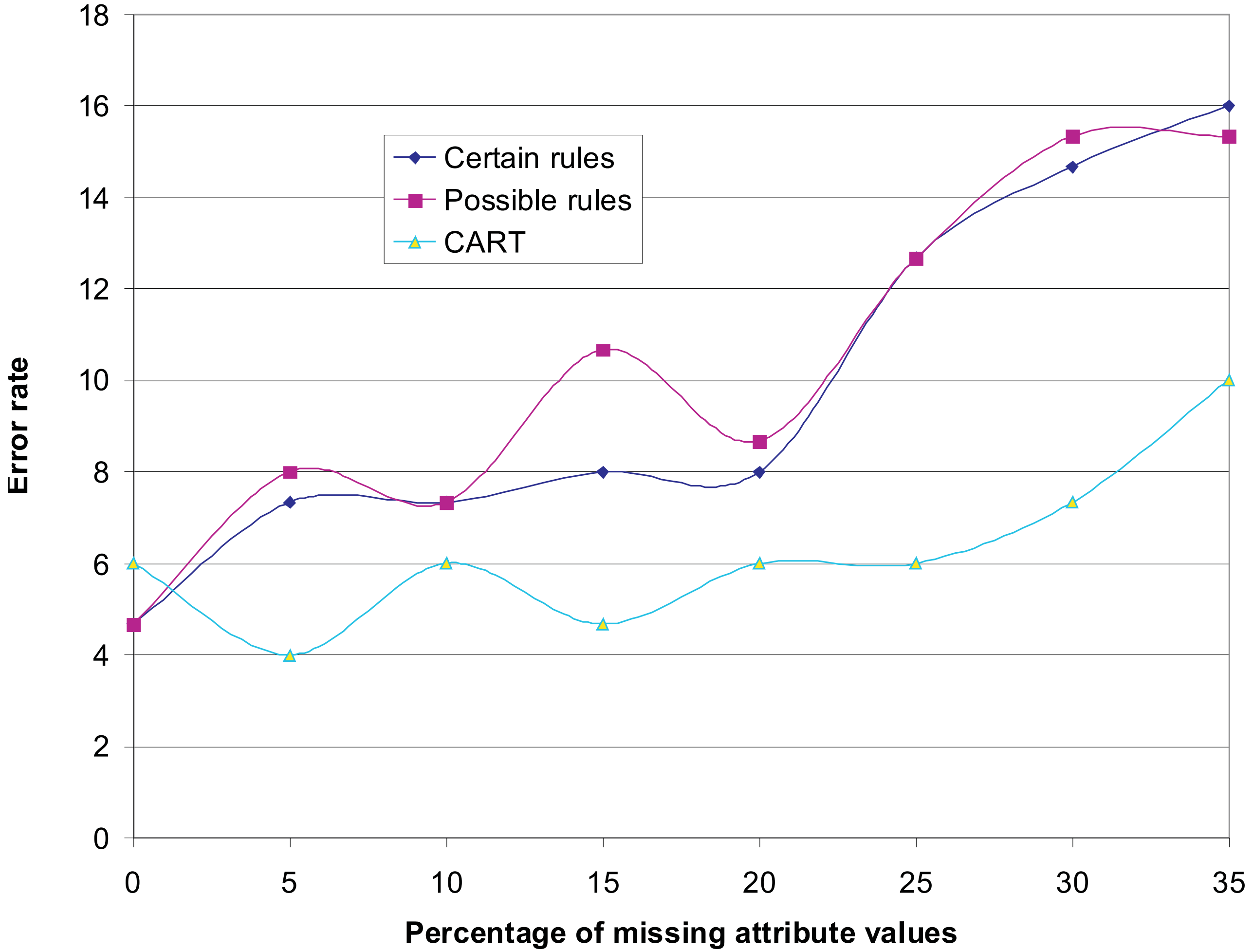


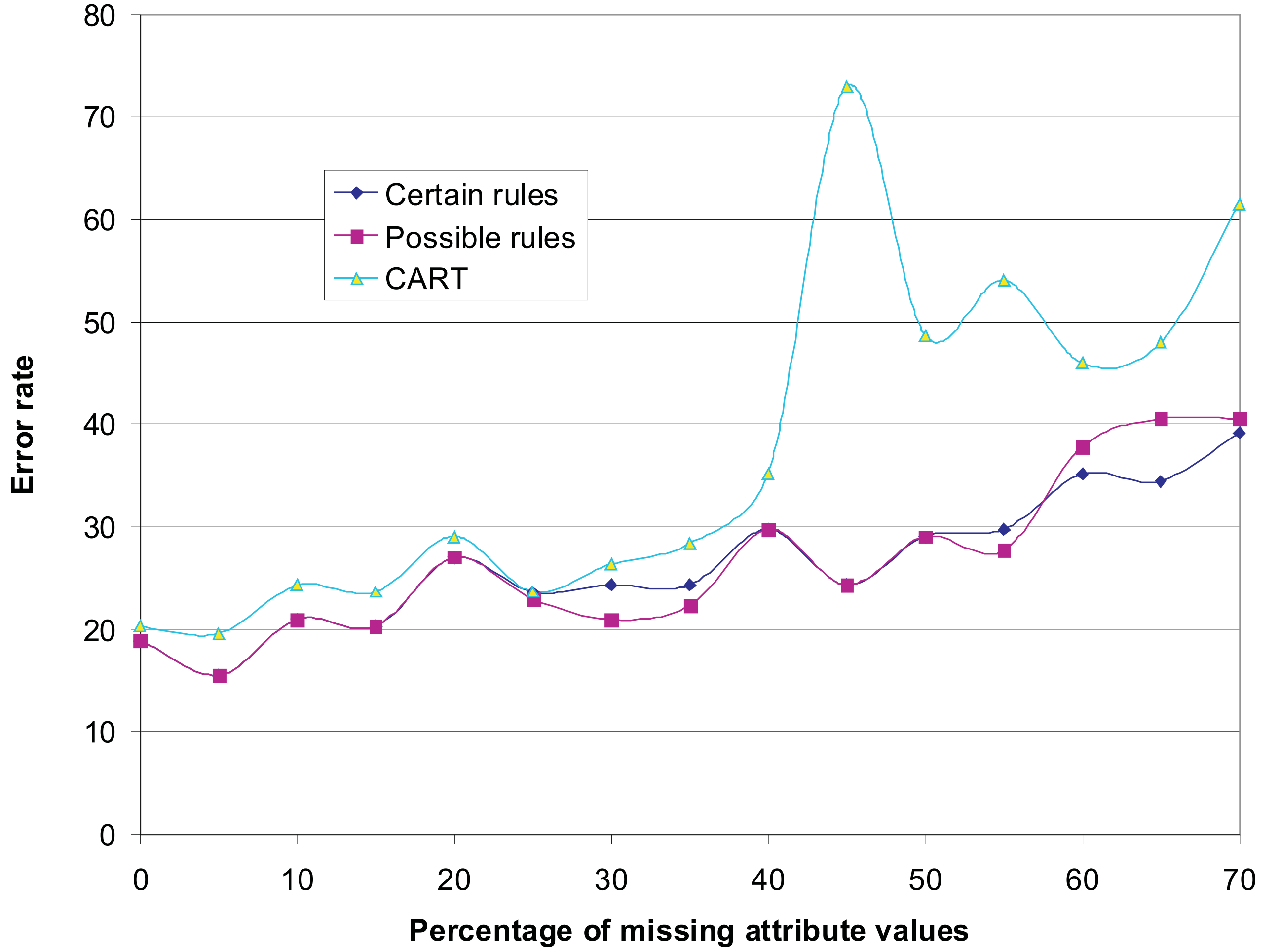
Image Segmentation Data Set



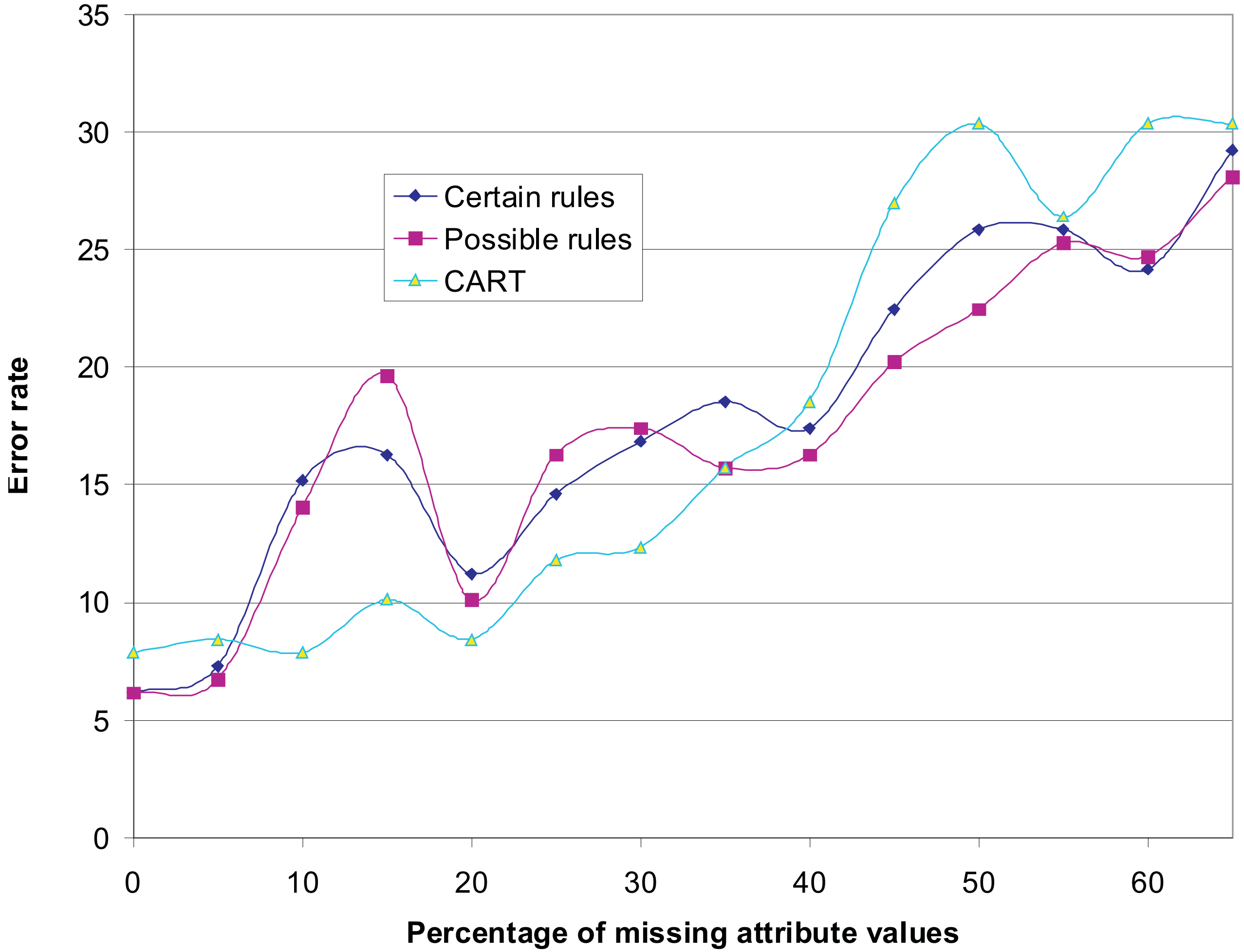
Iris Data Set



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

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.