Attractiveness measures
for decision rules induced from data
-critical survey

Izabela Brzezińska

Krytyczny przegląd miar atrakcyjności

reguł decyzyjnych wyindukowanych z danych

Plan

- Introduction
- Semantics of attractiveness measures
- Knowledge representation semantics
  - Brief attractiveness measures survey
  - Concept of mining the Pareto-optimal support/certainty border
  - Computational experiments
- Conclusions
- Further research plans
- References

Introduction

- *S*=⟨*U*, *A*⟩ – ***data table***, where *U* and *A* are finite, non-empty sets
  U – *universe*; *A* – set of *attributes*
- *S*=⟨*U*, *C*, *D*⟩ – ***decision table***, where C – set of *condition attributes*,
  D – set of *decision attributes*, $C \cap D = \emptyset$
- ***Decision rule*** or ***association rule*** induced from *S*
  is a ***consequence relation***: *F* ® *Y* ***read as*** ***if F, then Y***

  where *F* and *Y* are condition and decision formulas expressed as attribute-value pairs

Introduction

- $\|\Phi\|$ is the set of all objects from ***U***, **having property** F
- $\|\Psi\|$ is the set of all objects from ***U***, **having property** Y

- ***Support*** **of decision rule** $F \to Y$ **:**

$$\sup(f \to y) = card(\|f \wedge y\|)$$

- ***Certainty factor*** **for decision rule** $F \to Y$ **(Łukasiewicz, 1913): (called also *confidence*)**

$$cert(f \to y) = \frac{card(\|f \wedge y\|)}{card(\|f\|)}$$

Semantics of attractiveness measures

- In all practical applications, like ***medical practice***, ***market basket***, ***customer satisfaction or risk analysis***, it is crucial to know how good the rules are for:
  - knowledge representation
  - prediction
  - efficient intervention
- "How good" is a question about ***attractiveness measures*** of rules
- Review of literature shows that **there is no single measure which would be the best** for applications in all possible perspectives (e.g. Bayardo and Agrawal 1999, Greco, Pawlak & Slowinski 2004, Yao & Zhong 1999, Hilderman and Hamilton 2001)
- Claim 1: the adequacy of interestingness measures is dependent on their ***semantics***

Knowledge representation semantics

- Among commonly used attractiveness metrics are:
  - support
  - certainty (a.k.a. confidence)
  - conviction
  - lift
  - laplace
  - piatetsky-shapiro
  - gini
  - chi-squared
  - gray-orlowska
  - kamber-shinghal

- Several algorithms are known to efficiently find the best rules according to one of these metrics (e.g. *Webb'95, Fukada et al.'96, Rastoni and Shim '98).*

List of all mentioned attractiveness metrics

$$support(f \to y) = \mathrm{card}(\|f \wedge y\|)$$

$$\mathrm{certainty}(f \to y) = \mathrm{cert}(f \to y) = \frac{\mathrm{card}(\|f \wedge y\|)}{card(\|f\|)}$$

$$conviction(f \to y) = \frac{card(U) - card(\|y\|)}{card(U) * (1 - cert(f \to y))}$$

$$lift(f \to y) = \frac{card(U) * cert(f \to y)}{card(\|y\|)} = \frac{card(U) * card(\|f \wedge y\|)}{card(\|y\|) * card(\|f\|)}$$

$$laplace(f \to y) = \frac{card(\|f \wedge y\|) + 1}{card(\|f\|) + k}, k = number\_of\_classes$$

$$piatetsky - shapiro = card(\|f \wedge y\|) - \frac{card(\|f\|) * card(\|y\|)}{card(U)}$$

List of all mentioned attractiveness metrics

$$\mathrm{gini} = 1 - \left(\frac{\mathrm{card}(\|\psi\|)^2}{\mathrm{card}(U)^2} + \frac{\left(\mathrm{card}(U) - \mathrm{card}(\|\psi\|)\right)^2}{\mathrm{card}(U)^2} \right) -$$

$$- \frac{\mathrm{card}(\|\phi\|)}{\mathrm{card}(U)} * \left(1 - \left(\frac{\mathrm{card}(\|\phi \wedge \psi\|)^2}{\mathrm{card}(\|\phi\|)^2} + \frac{\left(\mathrm{card}(\|\phi\|) - \mathrm{card}(\|\phi \wedge \psi\|)\right)^2}{\mathrm{card}(\|\phi\|)^2} \right) \right) -$$

$$- \frac{\mathrm{card}(\|\sim \phi\|)}{\mathrm{card}(U)} * \left(1 - \left(\frac{\mathrm{card}(\|\sim \phi \wedge \psi\|)^2}{\mathrm{card}(\|\sim \phi\|)^2} + \frac{\left(\mathrm{card}(\|\sim \phi\|) - \mathrm{card}(\|\sim \phi \wedge \psi\|)\right)^2}{\mathrm{card}(\|\sim \phi\|)^2} \right) \right)$$

List of all mentioned attractiveness metrics

$$\mathrm{chi-squared} = \frac{\mathrm{card}(\|\phi\|) * \left(\frac{\mathrm{card}(\|\phi \wedge \psi\|)}{\mathrm{card}(\|\phi\|)} - \frac{\mathrm{card}(\|\psi\|)}{\mathrm{card}(U)} \right)^2 - \mathrm{card}(\|\sim \phi\|) * \left(\frac{\mathrm{card}(\|\sim \phi \wedge \psi\|)}{\mathrm{card}(\sim \|\phi\|)} - \frac{\mathrm{card}(\|\psi\|)}{\mathrm{card}(U)} \right)^2}{\frac{\mathrm{card}(\|\psi\|)}{\mathrm{card}(U)}} +$$

$$+ \frac{\mathrm{card}(\|\phi\|) * \left(\frac{\mathrm{card}(\|\phi \wedge \sim \psi\|)}{\mathrm{card}(\|\phi\|)} - \frac{\mathrm{card}(\|\sim \psi\|)}{\mathrm{card}(U)} \right)^2 - \mathrm{card}(\|\sim \phi\|) * \left(\frac{\mathrm{card}(\|\sim \phi \wedge \sim \psi\|)}{\mathrm{card}(\sim \|\phi\|)} - \frac{\mathrm{card}(\|\sim \psi\|)}{\mathrm{card}(U)} \right)^2}{\frac{\mathrm{card}(\|\sim \psi\|)}{\mathrm{card}(U)}} +$$

List of all mentioned attractiveness metrics

$$gray-orlowska = \left(\left(\frac{card(\|f \wedge y\|) * card(U)}{card(\|f\|) * card(\|y\|)} \right)^{l} - 1 \right) * \left(\frac{card(\|f\|) * card(\|y\|)}{card(U)^{2}} \right)^{m}$$

where *l*, *k* are parameters to weight the relative importance of the discrimination and support components.

$$kamber-shinghal = \frac{card(f \wedge y)}{card(f \wedge \sim y)} * \left(1 - \frac{card(\sim f \wedge y)}{card(\sim f \wedge \sim y)} \right)$$

The first component measures „how sufficient is Φ for Ψ".

The second component measures „how necessary is Φ for Ψ".

Conviction metric*

$$conviction(f \to y) = \frac{card(U) - card(\|y\|)}{card(U) * (1 - cert(f \to y))}$$

$$conviction(f \to y) = \frac{card(\|f\|) * card(\|\sim y\|)}{card(U) * card(\|f \wedge \sim y\|)}$$

$$conviction(f \to y) = \frac{card(\|\sim y\|)}{card(U) * cert(f \to \sim y)}$$

- **The metric takes into account occurrence of objects with ~Ψ in the decision table.**
- **Division by zero occurs when there are no (Φ and ~Ψ) objects.**

It is a strong disadvantage of the metric!

Bayardo, R.J.;Agrawal, R.;and Gunopulos, D. 1999. Constraint-Based Rule Mining in Large, Dense Databases. In *Proc. of the 15th Int'l Conf. on Data Engineering, 188-197.

Piatetsky-Shapiro's metric*

$$piatetsky - shapiro = card(\|f \wedge y\|) - \frac{card(\|f\|) * card(\|y\|)}{card(U)}$$

- This rule interest function is used to quantify the correlation between condition and decision attributes.
- When ps=0, then Φ and Ψ are statistically independent and the rule is not interesting.
- When ps>0 (ps<0), then Φ is positively (negatively) correlated to Ψ .
- The metric does not take into account occurrence of objects with ~Φ nor ~Ψ in the decision table.
- It can be transformed to be identical to *gain metric***.

*Piatetsky-Shapiro, G. 1991. Discovery , Analysis, and Presentation of Strong Rules. Chapter 13 of *Knowledge Discovery in Databases,* AAAI/MIT press, 1991.

**Fukada,T. et al.1996. Data Mining using Two-Dimensional Optimized Association Rules: Scheme, Algorithms, and Visualization. In *Proc. of the 1996 ACM-SIGMOD Int'l Conf. on the Management of Data, 13-23.*

Lift metric*

$$lift(f \rightarrow y) = \frac{card(U) * cert(f \rightarrow y)}{card(\|y\|)} = \frac{card(U) * card(\|f \wedge y\|)}{card(\|y\|) * card(\|f\|)}$$

- The metric is not straightforwardly influenced by number of objects with ~Ψ or ~Φ in the decision table.
- Measures the „independency" of Φ and Ψ.
- Looks very similar to Horwitch'82 confirmation measure r.

$$r(y \mid f) = \log\left(\frac{\Pr(y \wedge f)}{\Pr(y) * \Pr(f)}\right), where \Pr(X) = \frac{card(\|X\|)}{card(U)}$$

*International Business Machines, 1996. *IBM Intelligent Miner User's Guide*, Ver 1, Rel1.

Bayardo-Agrawal concept

- Bayardo and Agrawal ′99 introduced a concept involving a partial order on rules defined in terms of support and certainty.
- They demonstrated that the set of rules that are optimal according to this partial order includes all rules that are best according to any of these metrics.

Optimized rule mining – problem statement

- **The input to the problem of mining optimized rules: *<K, U, £, L, N>***
  - ***K*** **is a finite set of conditions;**
  - ***U*** **is a data set;**
  - **£ is a total order on rules;**
  - ***L*** **is a condition specifying the rule consequent;**
  - ***N*** **is a set of constraints on rules (e.g. minimum support, certainty).**
- **Optimized rule mining problem statement:**

  Find a rule r_1 such that:

  1. **r_1 satisfies the input constraints, and**
  2. **there exists no r_2 such that r_2 satisfies the input constraints and $r_1 < r_2$.**

Mining optimized rules under partial order

- With partial order, because some rules may be incomparable, there can be several equivalence classes containing optimal rules.
- The previous problem statement requires an algorithm to identify only a single rule from one of these equivalence classes.
- To mine at least one representative from each equivalence classes that contains an optimal rule, we need to rephrase the mining problem.
- Partial-order optimized rule mining problem statement:
  Find a set *R* of rules such that:
  1. *every rule* r_i in *R* is optimal as defined by the optimized rule mining problem
  2. for every equivalence class of rules if the equivalence class contains an optimal rule, then exactly one member of this equivalence class is in *R*.

Support-certainty optimality

- Consider the following partial order $\leq_{sc}$ on rules. Given rules r_1 and r_2, $r_1 \leq_{sc} r_2$ if and only if:

- $sup(r_1) \leq sup(r_2) \wedge cert(r_1) < cert(r_2)$, or

- $sup(r_1) < sup(r_2) \wedge cert(r_1) \leq cert(r_2)$.

Additionally, $r_1 =_{sc} r_2$ iff $sup(r_1) = sup(r_2)$ and $cert(r_1) = cert(r_2)$.

An optimal set of rules (optimized rule mining problem solutions) according to this partial order $\leq_{sc}$ is regarded as ***sc-upper border***. Intuitively, such a set of rules defines a *support-certainty border* above which no rule that satisfies the input constraints can fall.

Support-~ certainty optimality

- Consider the following partial order $\leq_{s\sim c}$ on rules. Given rules r_1 and r_2, $r_1 \leq_{s\sim c} r_2$ if and only if:

- sup(r_1) £ sup(r_2) Ù cert(r_1) > cert(r_2), or

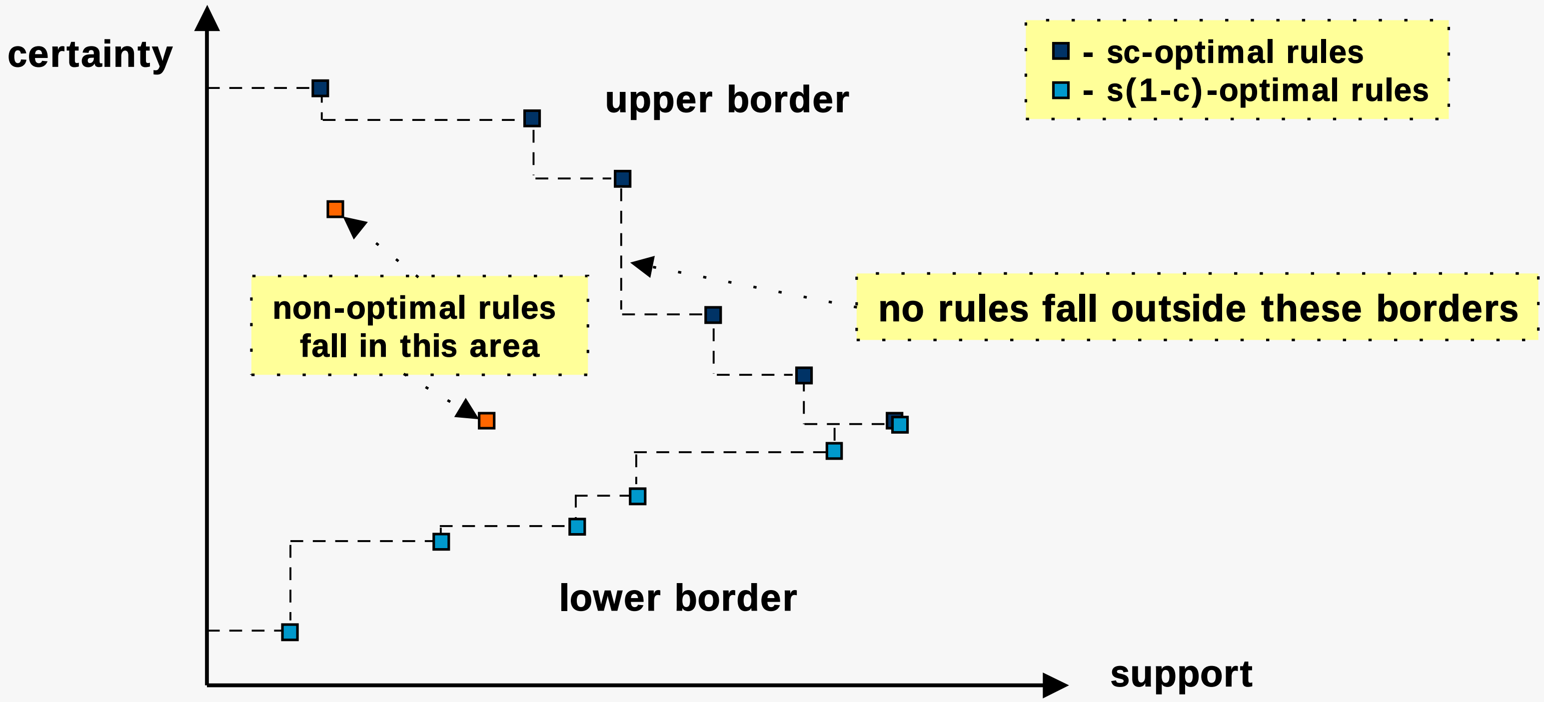
- sup(r_1) < sup(r_2) Ù cert(r_1) ³ cert(r_2).

Additionally, $r_1 =_{s\sim c} r_2$ iff sup(r_1) = sup(r_2) and cert(r_1) = cert(r_2).

An optimal set of rules according to this partial order $\leq_{s\sim c}$ forms a lower border.

Support vs certainty pareto-optimal borders

- **Pareto optimal borders:**
  - **support vs certainty => upper border**
  - **support vs ~certainty => lower border**



Border mining

- **Mining the upper support/certainty border identifies optimal rules according to such interestingness metrics:**
  - **support**
  - **certainty (a.k.a. confidence)**
  - **lift**
  - **conviction**
  - **laplace**
  - **piatetsky-shapiro's rule-interest function (p-s)**
- **If we also mine the lower border, such metrics will also be included:**
  - **gini**
  - **chi-squared**

Theoretical implications – lemma 1

Definition 1:

We say that an intended to rank rules in order of interestingness total order $\leq_t$ is implied by partial order $\leq_{sc}$ iff

$r_1 <_{sc} r_2$ Þ r_1 £$_t$ r_2, and $r_1 =_{sc} r_2$ Þ $r_1 =_t r_2$.

Lemma 1:

Given the problem instance *I* = <*K*, *U*, £$_t$, *L*, *N*>

such that $\leq_t$ is implied by $\leq_{sc}$,

an *I*-optimal rule is contained within any I_{sc}-optimal set

where I_{sc} = <*K*, *U*, £$_{sc}$, *L*, *N*>.

Theoretical implications – proof 1

Proof 1:

Consider any rule r_1 that is not I_{sc}-optimal. Because r_1 is non-optimal, there must exist some rule r_2 that is optimal such that $r_1 <_{sc} r_{2}$. But then we also have that r_1 £$_t$ r_2 since $\leq_t$ is implied by $\leq_{sc}$.

This implies that any non-I_{sc}-optimal rule is either non-I-optimal, or it is equivalent to some I-optimal rule which resides in an I_{sc}-optimal equivalence class.

At least one I_{sc}-optimal equivalence class must therefore contain an I-optimal rule. Further, because $=_t$ is implied by $=_{sc}$, every rule in this equivalence class must be I-optimal. By definition, an I_{sc}-optimal set will contain one of these rules, and the claim follows. •

Theoretical implications – lemma 2

To identify the interestingness metrics that are implied by $\pounds_{sc}$ ***we use:***

Lemma 2:

The following conditions are sufficient for establishing that a total order $\leq_t$ defined over a rule value function f(r) is implied by partial order $\leq_{sc}$:

1. **f(r) is monotone in support over rules with the same certainty, and**
2. **f(r) is monotone in certainty over rules with the same support.**

Proof 2:

Suppose **$r_1 <_{sc} r_2$, then consider a rule r where $sup(r_1)=sup(r)$ and $cert(r_2)=cert(r)$. By definition $r_1 \leq_{sc} r$ and $r \leq_{sc} r_2$.**

If the total order has the monotonicity properties 1,2,

then $r_1 \pounds_t r$ and $r \pounds_t r_2$. Since total orders are transitive, we then have that $r_1 \pounds_t r_2$, which establishes the claim. •

Example - Laplace function

Consider a Laplace function, which is commonly used to rank rules for classification purposes:

Definition 2:

$$laplace(f \rightarrow y) = \frac{card(\|f \wedge y\|) + 1}{card(\|f\|) + k}$$

where k **is a constant integer >1, usually set to the number of classes when building a classification model.**

Because:

$$card(\|f \wedge y\|) = \sup(\|f \wedge y\|), cert = \frac{card(\|f \wedge y\|)}{card(\|f\|)}$$

we have:

$$laplace(f \rightarrow y) = \frac{\sup(f \rightarrow y) + 1}{\frac{\sup(f \rightarrow y)}{cert(f \rightarrow y)} + k}$$

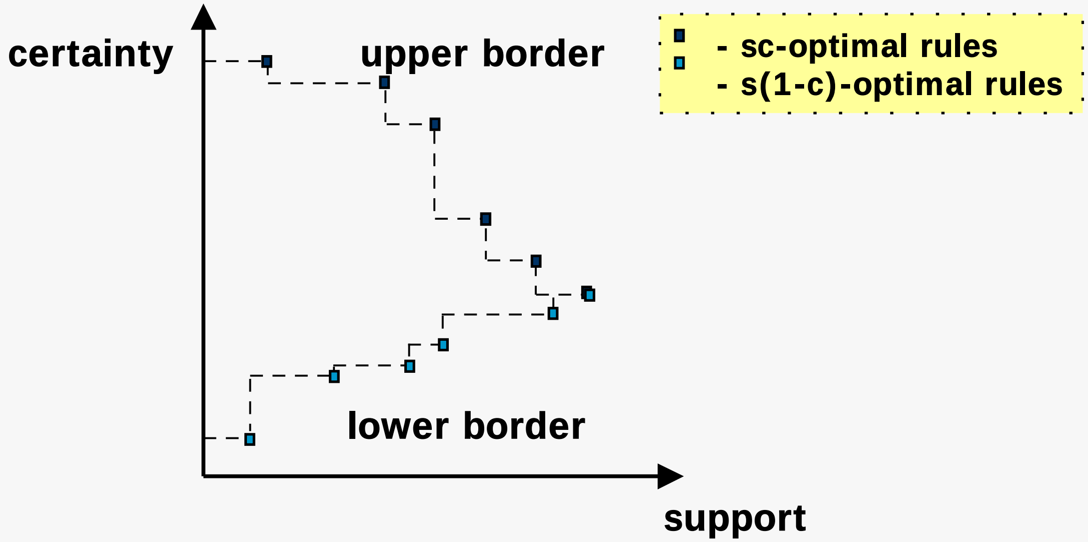
Example - Laplace function

$$laplace\ (f \rightarrow y\ ) = \frac{\sup(f \rightarrow y\ ) + 1}{\dfrac{\sup(f \rightarrow y\ )}{cert\ (f \rightarrow y\ )} + k}$$

- **This expression is monotone in rule support sine k>1 and cert≥0.**
- **It is also monotone in certainty among rules with equivalent support:**
  - **note that if support is held constant, in order to raise the function value, we need to decrease the value of the denonimator,**
  - **the decrease of the denominator can only be achieved by increasing certainty.**

Other attractiveness metrics included in the upper border

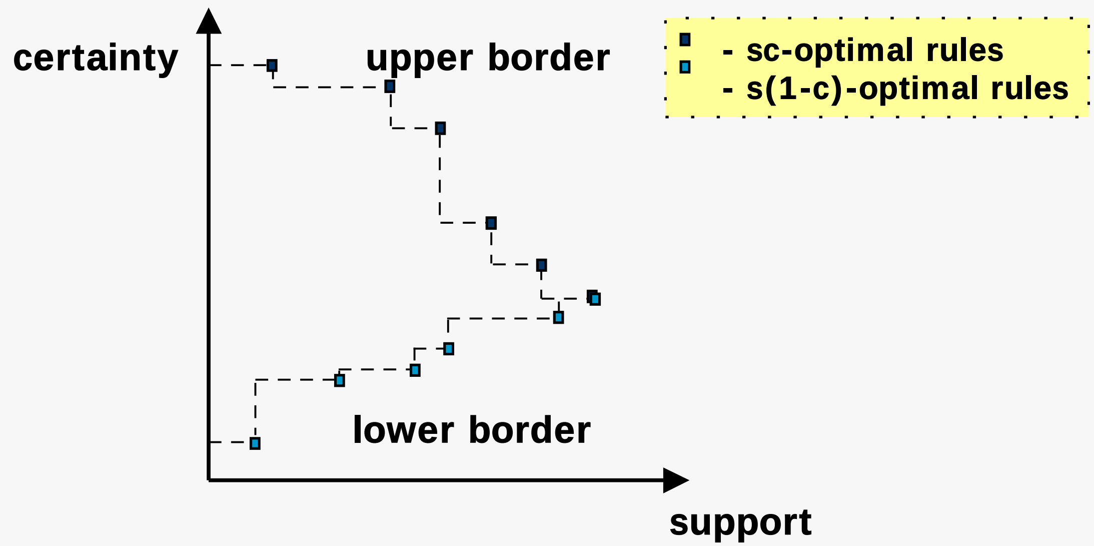
- Bayardo and Agrawal ′99 have also showed that the total orders listed below are also implied by partial order $\leq_{sc}$:
  - support
  - certainty
  - conviction
  - lift
  - laplace
  - piatetsky-shapiro



- **Thus, mining the upper support/certainty border identifies optimal rules according to these metrics.**

Mining the lower border

- The following metrics are not implied by $\leq_{sc.}$
  - gini
  - chi-squared



- However, Bayardo and Agrawal have shown that the optimal rules with respect to these metrics must reside on either the upper or lower support/certainty border.

Computational experiment

- **Decision rules were generated from lower approximations of preference-ordered decision classes defined according to Variable-consistency Dominance-based Rough Set Approach (VC-DRSA)**

  (Greco, Matarazzo, Slowinski, Stefanowski 2001)

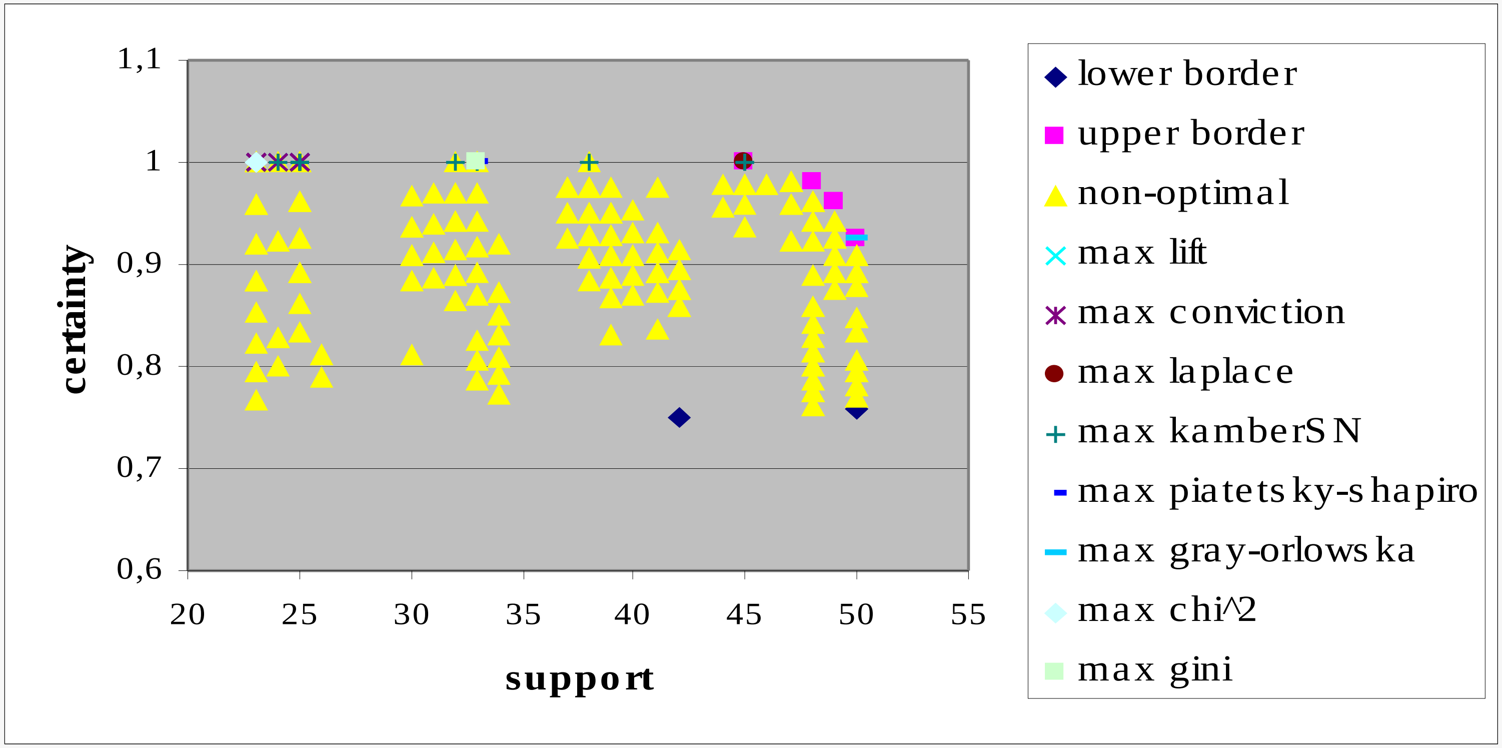
File	objects	atr+crit	classes	rules (alg)	consistency	length	coverage
Buses	76	0+8	3	266 (all)	$\geq$ 0.75	$\leq$ 3	$\geq$ 0.9
Nativity	342	0+33	2	64 (mc)	$\geq$ 0.75	no limit	no limit
Urology	500	18+9	3	186 (mc)	$\geq$ 0.96	no limit	no limit

Rule induction algorithms: „all" = all rules algorithm (DOMAPRIORI)

„mc" = minimal-cover algorithm (DOMLEM)

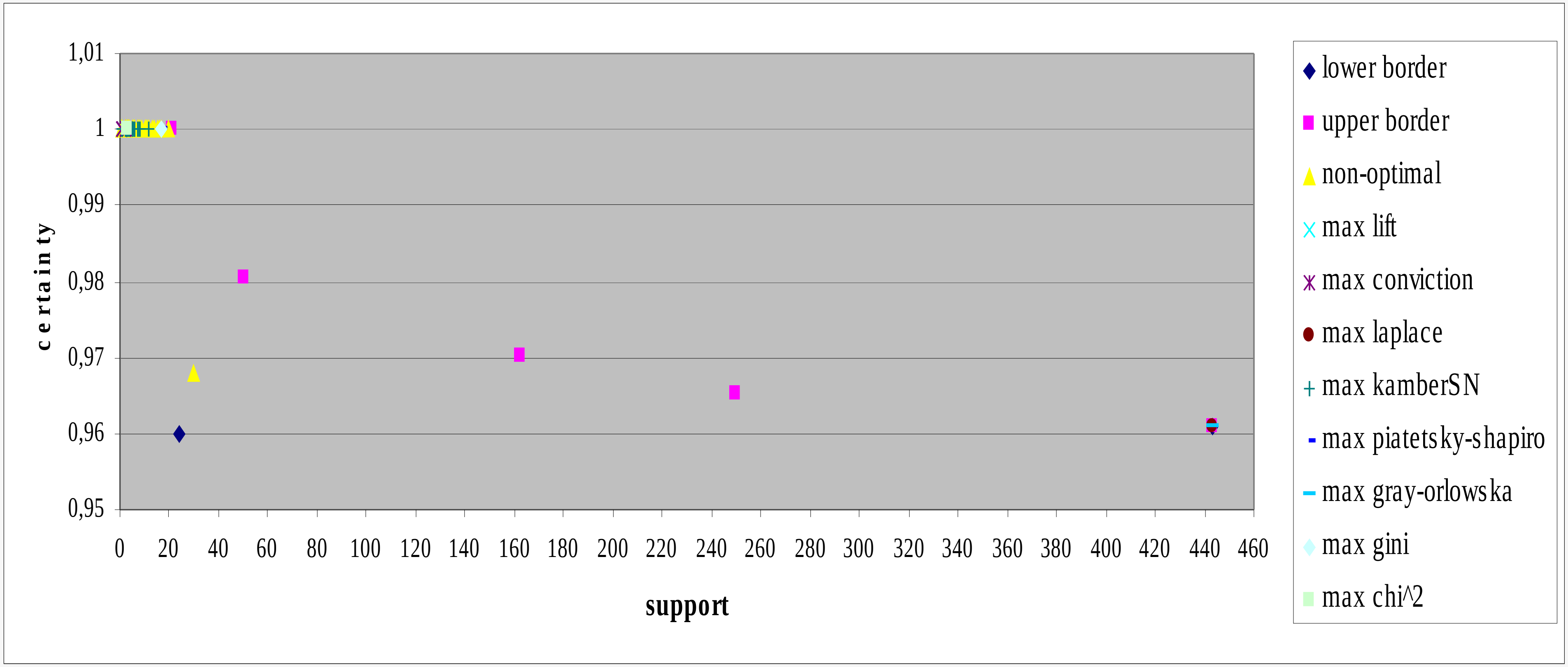
Computational experiment - Buses

- ***Buses***



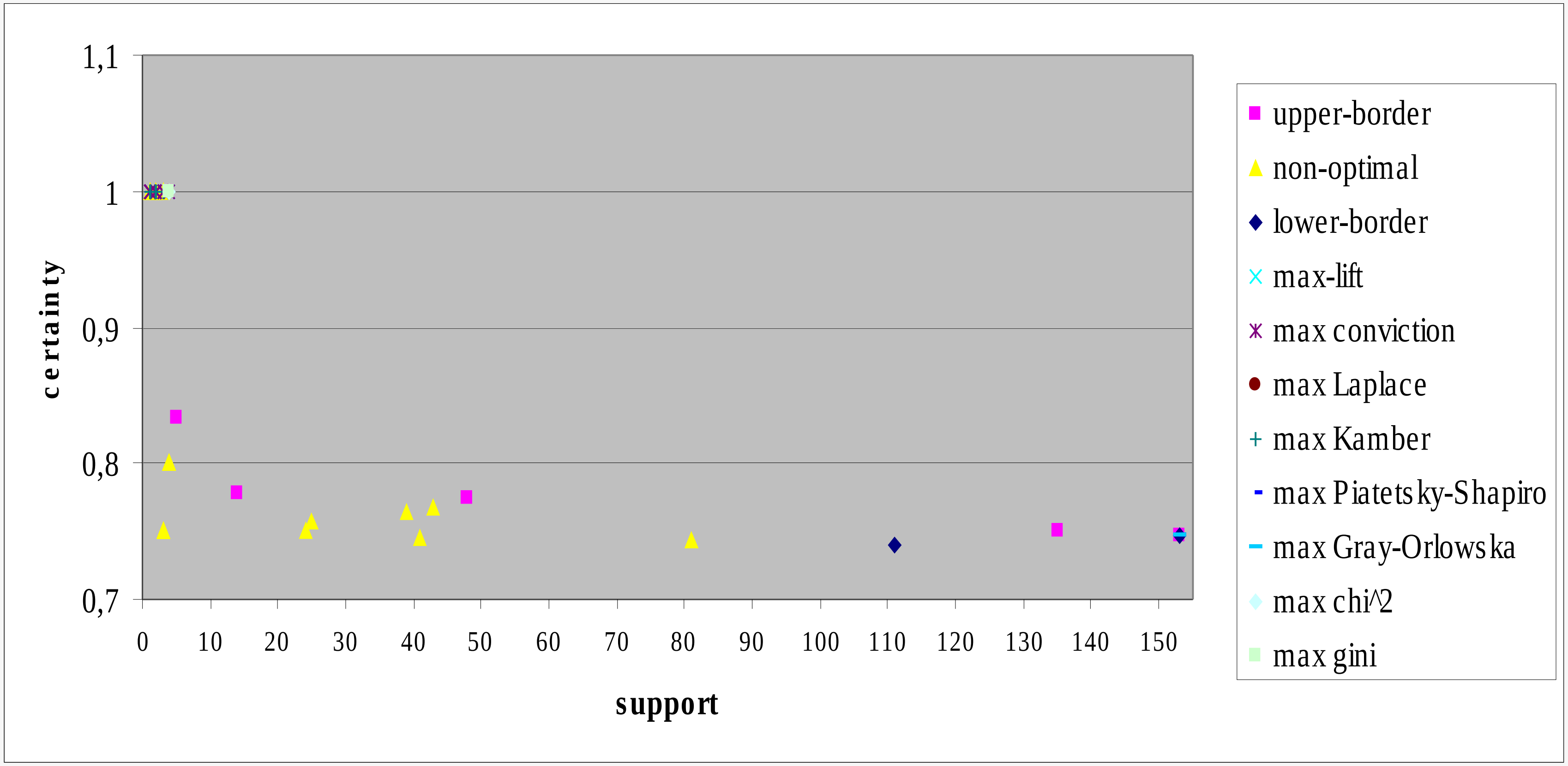
Computational experiment - Urology

- *Urology*



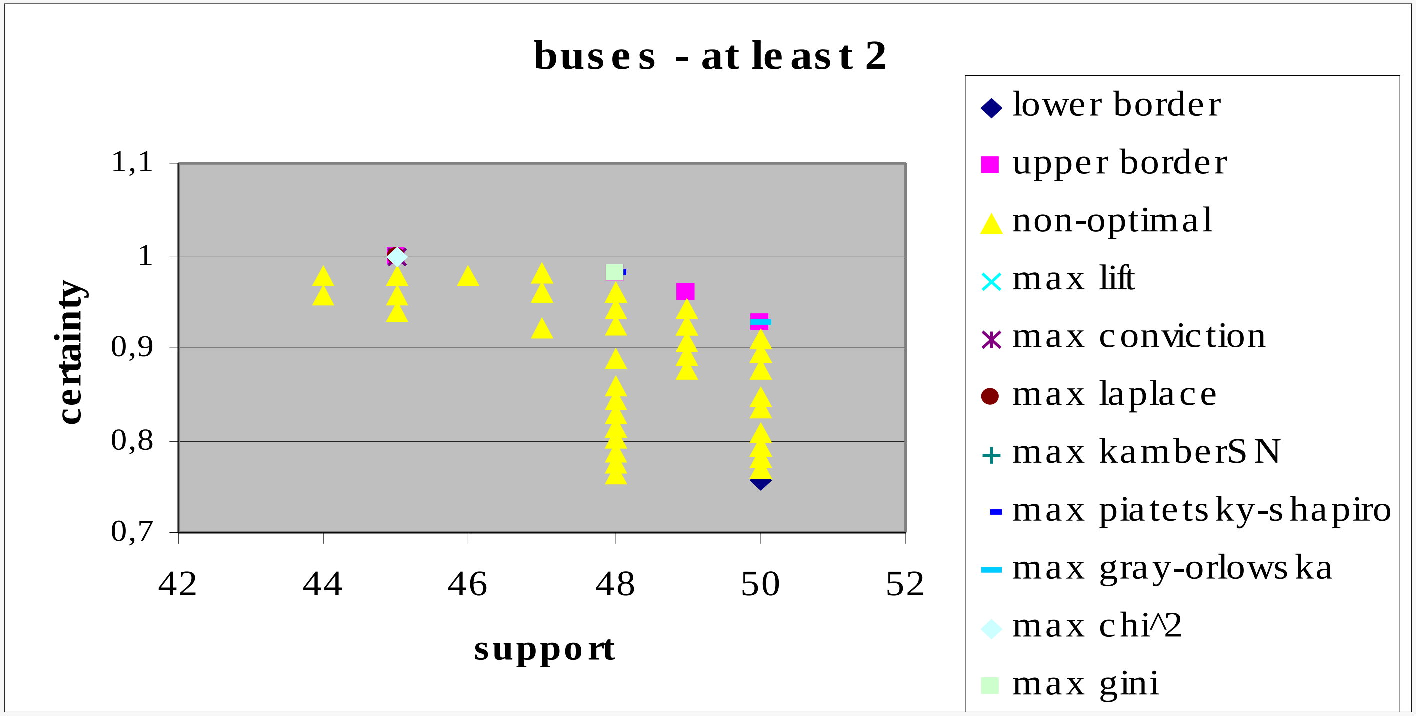
Computational experiment - Nativity

- ***Nativity***



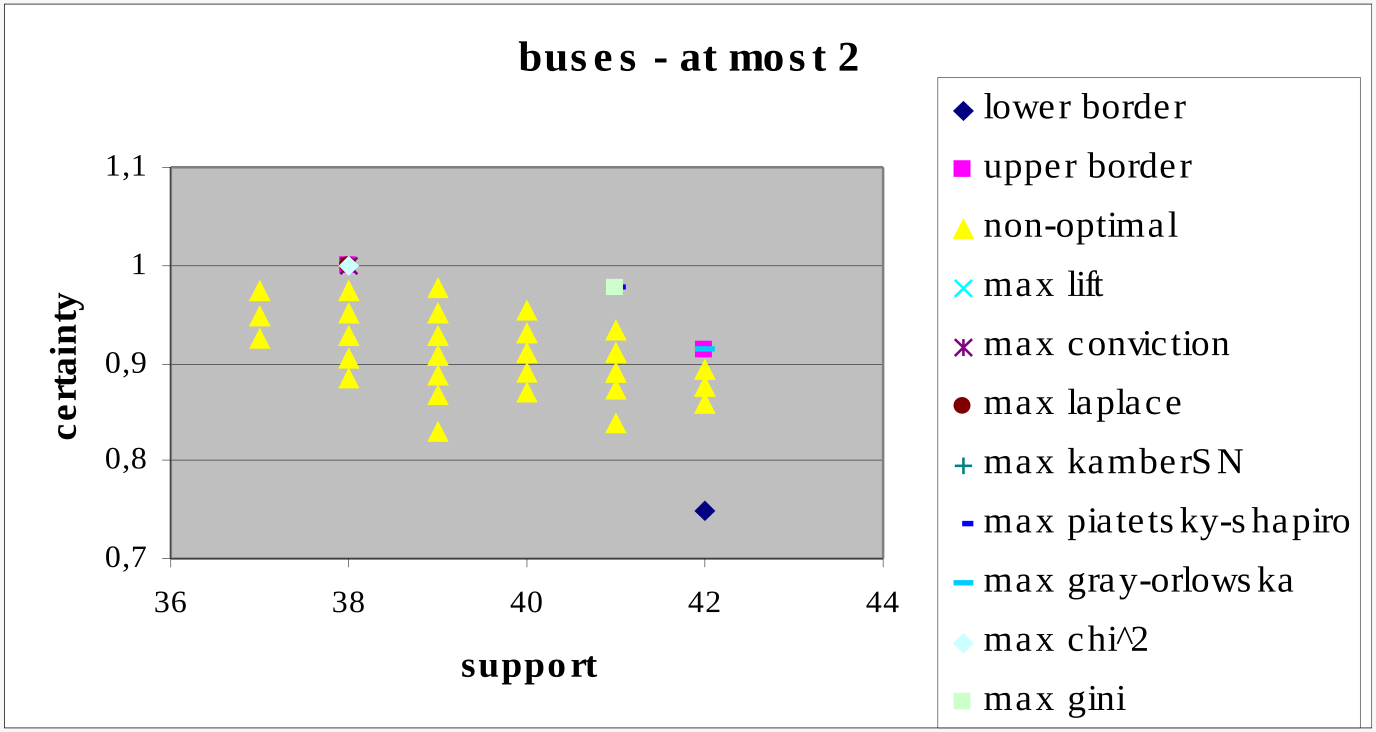
Computational experiment – Buses ″at least 2″

- *Buses – for union of classes 1 and 2*



Computational experiment – Buses ″at most 2″

- *Buses – for union of classes 2 and 3*



Conclusions – knowledge representation

- A survey on attractiveness measures in knowledge representation aspect has been done.
- By Bayardo and Agrawal a new optimized rule mining problem has been defined. It allows a partial order in place of the typical total order on rules.
- Solving this optimized rule mining problem with respect to a particular partial order $\leq_{sc}$ is guaranteed to identify a most-interesting rule according to several attractiveness metrics including: (support, certainty, laplace, conviction, piatetsky-shapiro, lift, gini, chi-squared).
- The computational experiment has expressed that indeed Pareto optimal support/certainty border contains rules optimal with respect to any of those metrics.
- Moreover, the computational experiment placed in upper support/certainty border also rules optimal according to kamber-shinghal and gray-orlowska metrics. However, an analytical proof is required.

Further research

- Computational experiments have placed rules optimal according to gray-orlowska metric and kamber-shinghal metric in the upper support/certainty border. Can it be analitically verified whether these total orders are implied by partial order $\leq_{sc}$?
- Is it possible to imply discussed total orders (like: support, certainty, laplace, etc.) by partial order other than support/certainty?
- Are some metrics (eg. lift) confirmation metrics? Analitical proof of posessing hypothesis symmetry and monotonicity properties.
- From decision to association rules…

References

- Bayardo, R.J., Agrawal, R.: Mining the most interesting rules. [In]: Proc. of the Fifth ACM SIGKDD Intl Conf. on Knowledge Discovery and Data Mining, (1999) 145-154
- Greco, S., Matarazzo, B., Pappalardo, N., Słowiński, R.: Measuring expected effects of interventions based on decision rules. Journal of Experimental and Applied Artificial Intelligence, 17 (1-2) (2005) 103-118
- Greco, S., Pawlak,. Z., Słowiński, R.: Can Bayesian confirmation measures be useful for rough set decision rules? Engineering Applications of Artificial Intelligence, 17 (4) (2004) 345-361
- Hilderman, R.J., Hamilton, H.J.: Knowledge Discovery and Measures of Interest. Kluwer Academic Publishers, Boston (2001)
- Yao, Y.Y., Zhong, N.: An analysis of quantitative measures associated with rules. [In]: Proceedings of the 3rd Pacific-Asia Conference on Knowledge Discovery and Data Mining, LNAI 1574, Springer-Verlag, Berlin (1999) 479-488