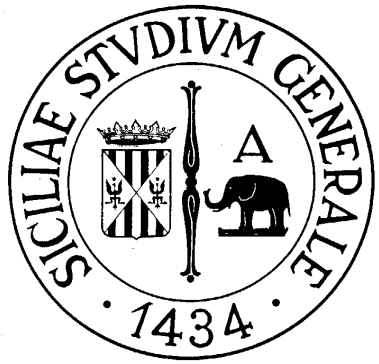


Mining Association Rules with respect to Support and Anti-support - Experimental results

Izabela Szczęch

Poznań University of Technology

Roman Słowiński

Poznań University of Technology,

Institute for Systems Research, PAS, Poland

Salvatore Greco

University of Catania, Italy

RSEISP 2007, Warszawa

Presentation plan

- Introduction
- Basic quantitative characteristics of rules

- Rule evaluation with respect to support and confidence
- Confirmation Perspective on the support-confidence evaluations

- Rule evaluation with respect to support and anti-support
- Confirmation Perspective on the support-confidence evaluations

Introduction – rule induction

- Patterns in form of rules are induced from a data table
- $S=\langle U, A\rangle$ – *data table*, where U and A are finite, non-empty sets
  U – *universe*; A – set of *attributes*
- $S=\langle U, C, D\rangle$ – *decision table*, where C – set of *condition attributes*,
  D – set of *decision attributes*, $C\cap D=\varnothing$

- *Decision rule* or *association rule* induced from S
  is a *consequence relation*: $\phi\rightarrow\psi$ *read as* ***if*** ϕ ***then*** ψ
  where ϕ and ψ are condition and conclusion formulas
  built from attribute-value pairs (q,v)

- **In this work we consider association rules with a fixed conclusion**

Introduction – attractiveness measures

- To measure the relevance and utility of rules, quantitative measures called attractiveness or interestingness measures, have been proposed

  (e.g. support, confidence, lift, gain, conviction, Piatetsky-Shapiro,…)

- Unfortunately, there is no evidence which measure(s) is (are) the best

- Notation:
  - $sup(\circ)$ is the number of all objects from *U*, having property $\circ$

    e.g. $sup(\phi)$, $sup(\psi)$

Basic quantitative characteristics of rules

- Basic quantitative characteristics of rules
  - *Support* of rule $\phi \rightarrow \psi$ in *S*:

$$sup(\phi \rightarrow \psi) = sup(\phi \wedge \psi)$$

  - *Confidence* (called also *certainty factor*) of rule $\phi \rightarrow \psi$ in *S*:

$$conf(\phi \rightarrow \psi) = \frac{sup(\phi \rightarrow \psi)}{sup(\phi)}$$

  - *Anti-support* of rule $\phi \rightarrow \psi$ in *S*:

$$anti\text{-}sup(\phi \rightarrow \psi) = sup(\phi \wedge \neg\psi)$$

Bayesian confirmation measures

- An attractiveness measure *c* has the property of confirmation if is satisfies the following condition:

$$c(\phi,\psi)\begin{cases} > 0 \;\; if \;\; Pr(\psi|\phi) > Pr(\psi) \\ = 0 \;\; if \;\; Pr(\psi|\phi) = Pr(\psi) \\ < 0 \;\; if \;\; Pr(\psi|\phi) < Pr(\psi) \end{cases}$$

- Measures of confirmation quantify the strength of confirmation that premise ϕ gives to conclusion ψ
- „ψ is verified more often, when ϕ is verified, rather than when ϕ is not verified”

Bayesian confirmation property - interpretation

- $c(\phi, \psi)>0$ means that property ψ is satisfied more frequently when ϕ is satisfied, rather than generically in S (where the frequency is $Pr(\psi)$)

- $c(\phi, \psi)=0$ means that property ψ is satisfied with the same frequency whether ϕ is satisfied or not

- $c(\phi, \psi)<0$ means that property ψ is satisfied less frequently when ϕ is satisfied, rather than generically

Confirmation measure *f*

- Measures of confirmation quantify the strength of confirmation that premise ϕ gives to conclusion ψ

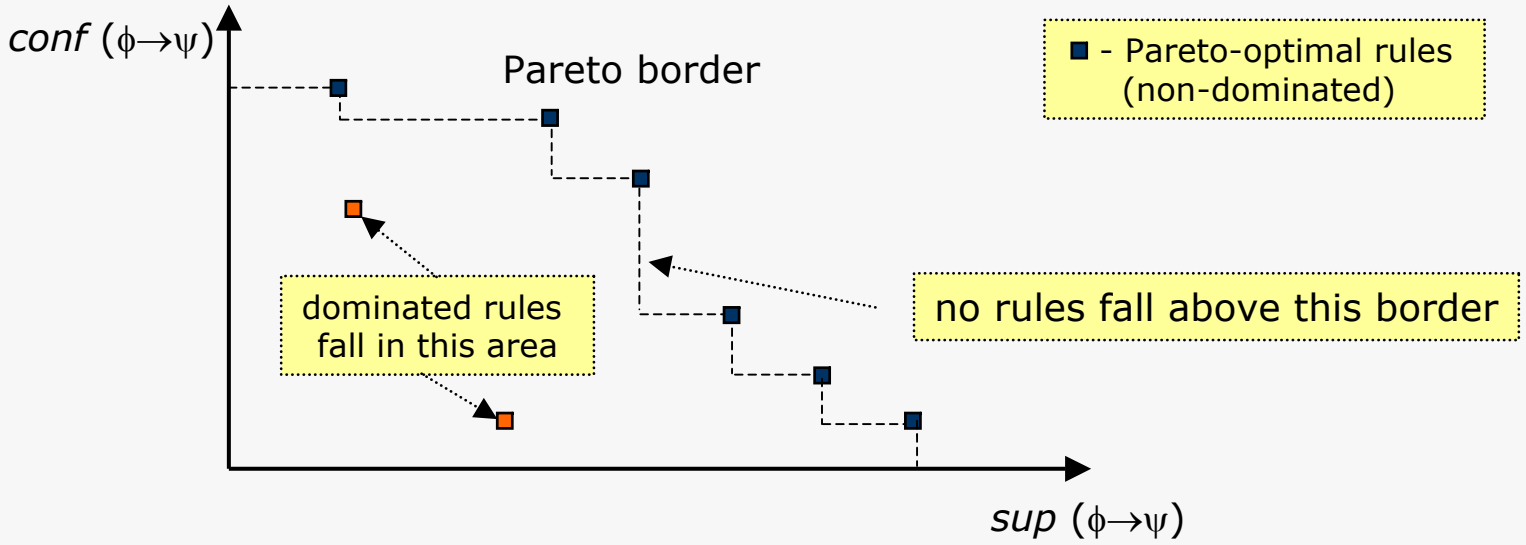
- Confirmation measure *f* (Good 1984, Heckerman 1988, Pearl 1988, Fitelson 2001)

$$f(\phi \rightarrow \psi) = \frac{conf(\psi \rightarrow \phi) - conf(\neg\psi \rightarrow \phi)}{conf(\psi \rightarrow \phi) + conf(\neg\psi \rightarrow \phi)}$$

Support-confidence evaluation

Support-confidence Pareto border

- Support-confidence Pareto border is the set of non-dominated, Pareto-optimal rules with respect to both rule support and confidence



- Mining the border identifies rules optimal with respect to measures such as: *lift*, *gain*, *conviction*, *Piatetsky-Shapiro*,…

Confirmation perspective on support-confidence space

- Is there a curve separating rules with negative value of any measure with the confirmation property in the support-confidence space?

- *Theorem:*
  Due to monotonicity of confidence in *c*,

$$c(\phi \rightarrow \psi) \geq 0 \quad \Leftrightarrow \quad conf\,(\phi \rightarrow \psi) \geq sup(\psi)/|U|$$

  Thus, rules lying below a constant:

$$sup(\psi)/|U|$$

  have a negative value of any confirmation measure.
  For those rules, the premise only disconfirms the conclusion!

- $sup(\psi)/|U|$ is a constant expressing what percentage of the whole data set is taken by considered class ψ

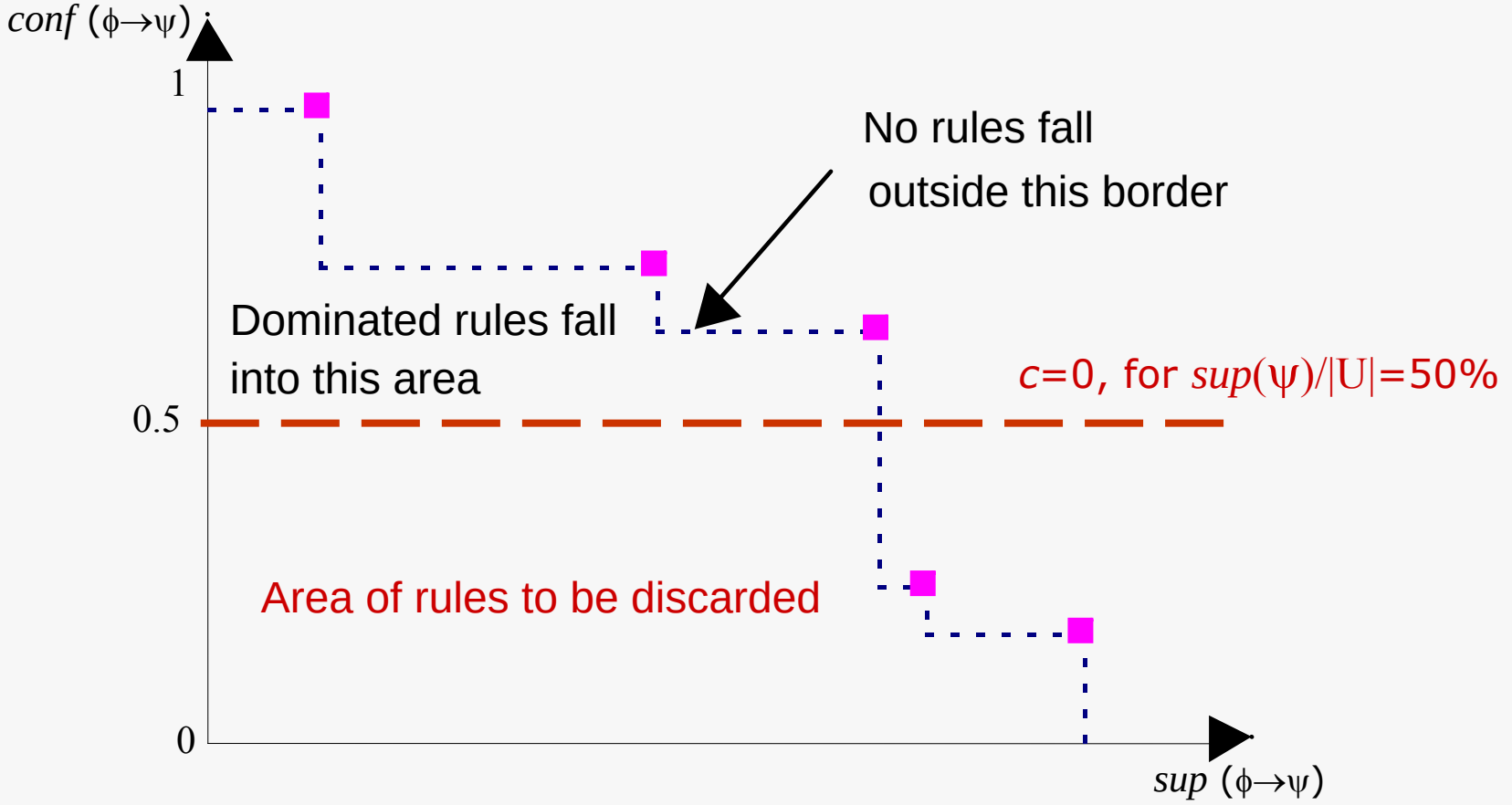
Confirmation perspective on support-confidence space

- A more general condition $c(\phi, \psi) \geq k$, $k \geq 0$ for some specific confirmation measure, $f(\phi, \psi)$, was also investigated.

- *Theorem:*

$$f(\phi \rightarrow \psi) \geq k \Leftrightarrow conf(\phi \rightarrow \psi) \geq sup(\psi)(k+1) \,/\, [|U| - k(|U| - 2sup(\psi))]$$

Confirmation perspective on support-confidence space

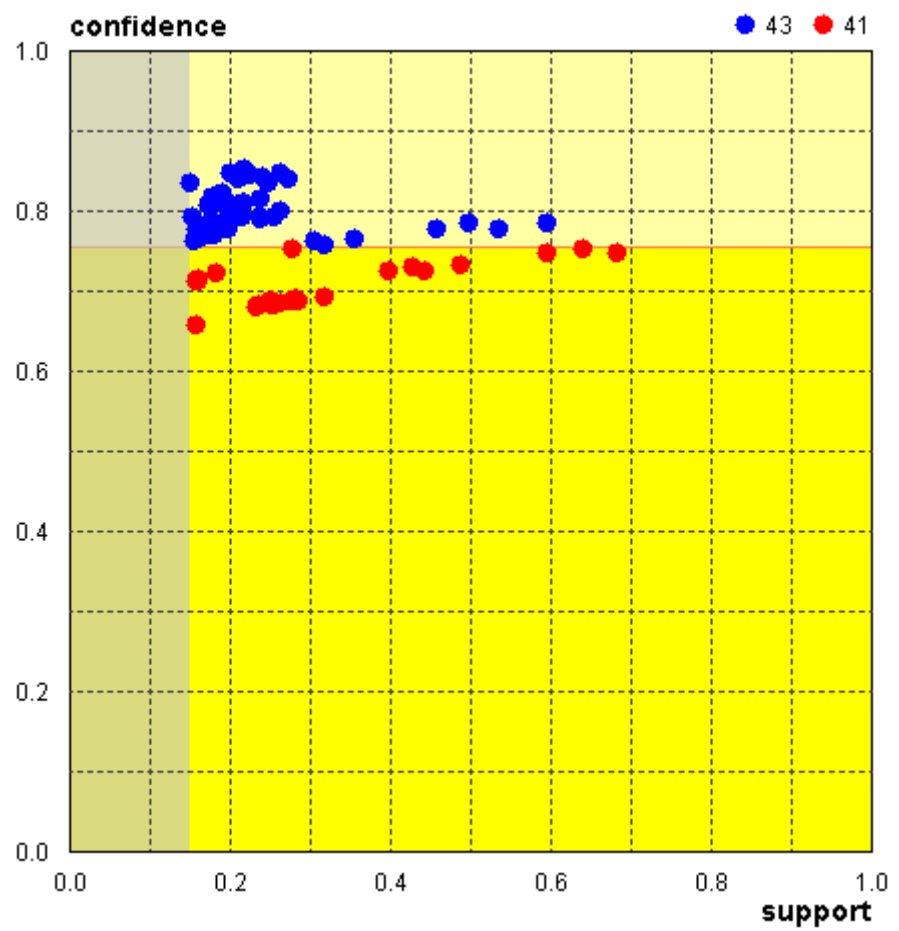


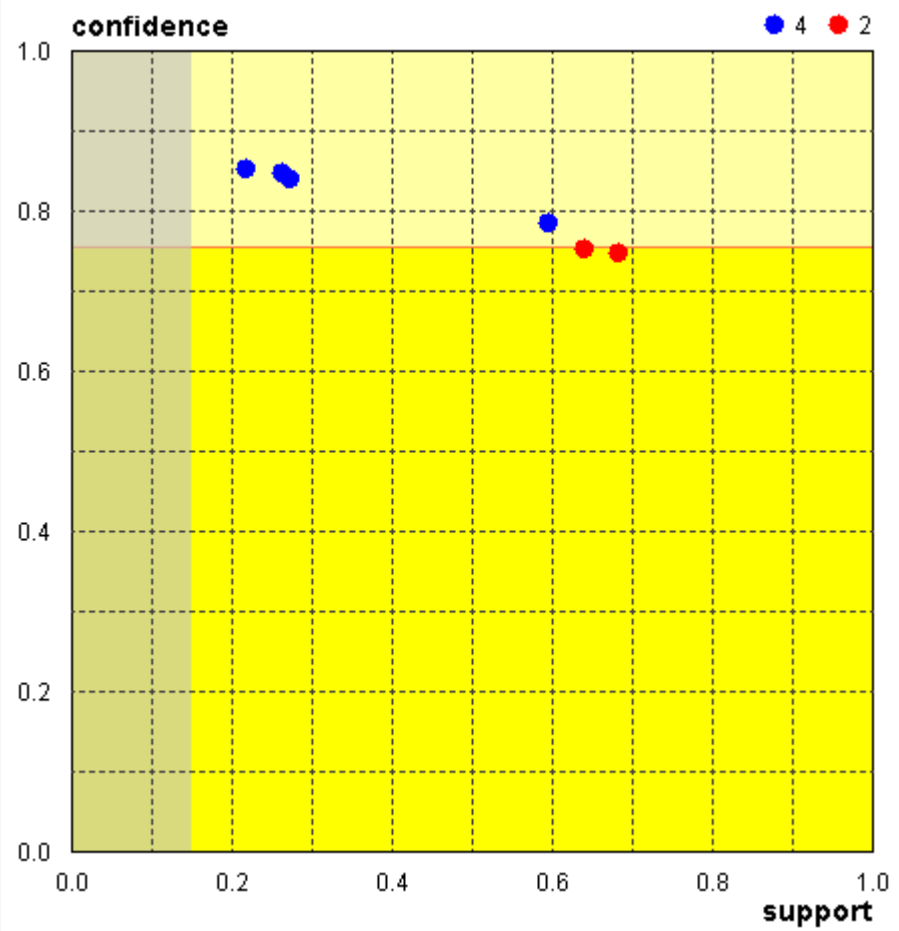
For rules lying below the curve for which $c=0$
the premise only disconfirms the conclusion

General info about the dataset

- Dataset ***adult***, created in ’96 by B. Becker/R. Kohavi from census database
- **32 561 instances**
- 9 nominal attributes
  - workclass: Private, Local-gov, etc.;
  - education: Bachelors, Some-college, etc.;
  - marital-status: Married, Divorced, Never-married, et.;
  - occupation: Tech-support, Craft-repair, etc.;
  - relationship: Wife, Own-child, Husband, etc.;
  - race: White, Asian-Pac-Islander, etc.;
  - sex: Female, Male;
  - native-country: United-States, Cambodia, England, etc.;
  - salary: >50K, <=50K
- **throughout the experiment, *sup*($\phi \rightarrow \psi$) is denoted as „support” and expressed as a relative rule support [0-1]**

Support-confidence (workclass=Private)





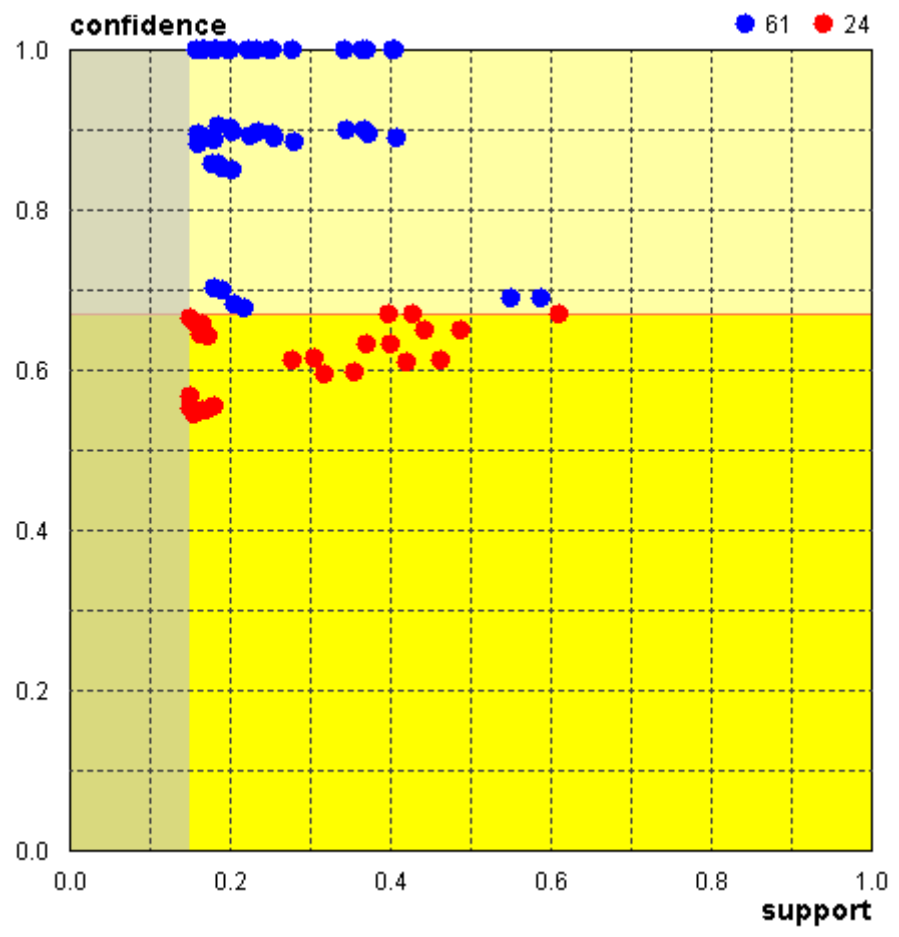
- ● indicates rules with negative confirmation
- the class constitutes over 70% of the whole dataset
- rules with high confidence can be disconfirming
- even some rules from the Pareto border need to be discarded

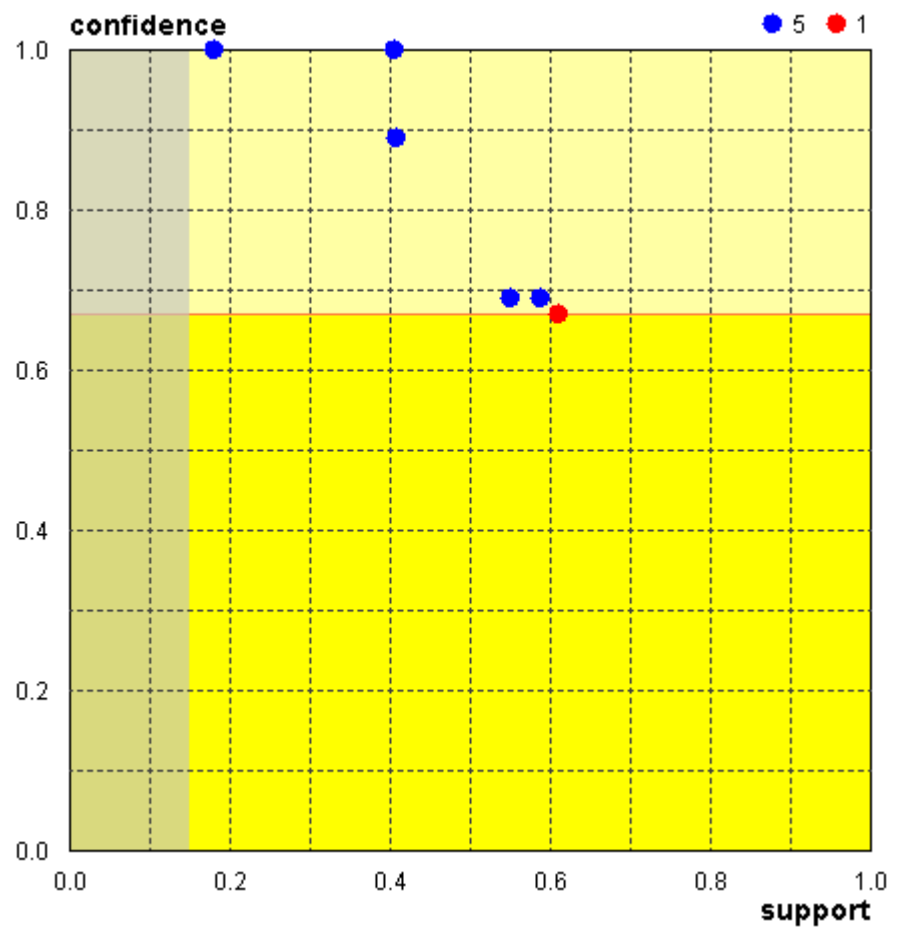
Few rules describing class: workclass=Private

premise	conclusion	supp	conf	s	f	a-supp
education is HS-grad **and race is** White **and native-country is** United-States	**workclass is** Private	0.20	0.79	0.05	0.10	0.06
education is HS-grad **and sex is** Male **and native-country is** United-States	**workclass is** Private	0.16	0.76	0.01	0.02	0.05
education is HS-grad **and native-country is** United-States	**workclass is** Private	0.24	0.79	0.05	0.10	0.06
education is Some-college **and native-country is** United-States	**workclass is** Private	0.16	0.77	0.02	0.03	0.05
marital-status is Married-civ-spouse **and relationship is** Husband **and race is** White	**workclass is** Private	0.25	0.69	-0.11	-0.17	0.12
relationship is Husband **and race is** White **and sex is** Male	**workclass is** Private	0.25	0.69	-0.11	-0.17	0.12
relationship is Husband **and sex is** Male **and native-country is** United-States	**workclass is** Private	0.25	0.68	-0.12	-0.18	0.12
race is White **and sex is** Male	**workclass is** Private	0.43	0.73	-0.06	-0.06	0.16
sex is Male **and native-country is** United-States	**workclass is** Private	0.44	0.72	-0.07	-0.07	0.17

- the table contains few examples of rules with the conclusion workclass=Private

Support-confidence (sex=Male)





- indicates rules with negative confirmation

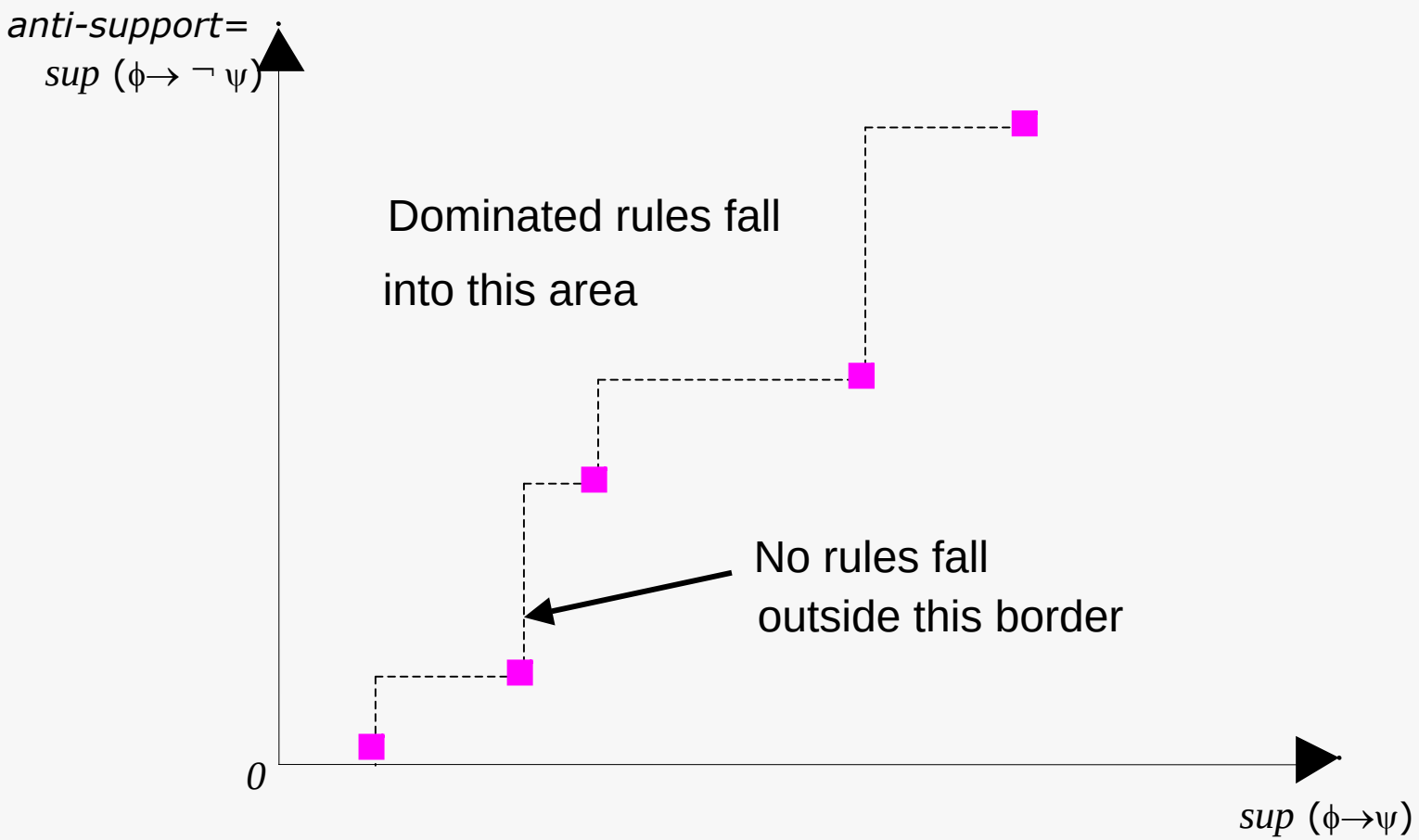
Support-confidence - summary

Considered conclusion	No. of all rules	No. of rules with non–positive confirm.	Reduction percentage
workclass='Private'	84	41	49%
sex=Male	85	24	28%
income<=50kUSD	87	43	49%

Considered conclusion	No. of all rules on Pareto border	No. of rules with non–positive confirm.	Reduction percentage
workclass='Private'	6	2	33%
sex=Male	6	1	17%
income<=50 kUSD	5	1	20%

Support - anti-support evaluation

Support - anti-support Pareto border



Theorem: The best rules according to any measure with the property M must reside on the support - anti-support Pareto border

Brzezińska I., Greco S., Słowiński R.: Mining Pareto-Optimal Rules with Respect to Support and Confirmation or Support and Anti-Support (EAAI Journal, 2007)

Confirmation perspective on support - anti-support space

- Is there a curve separating rules with negative value of any confirmation measure in the support-anti-support space?

- *Theorem:*

  Due to anti-monotonicity of anti-support in *c*,

$$c(\phi \rightarrow \psi) \geq 0 \quad \Leftrightarrow \quad anti\text{-}sup(\phi \rightarrow \psi) \leq sup(\phi \rightarrow \psi)[|U|/sup(\psi)\text{-}1]$$

  Thus, rules lying above a linear function:

$$sup(\phi \rightarrow \psi)[|U|/sup(\psi)\text{-}1]$$

  have a negative value of any confirmation measure.

  For those rules, the premise only disconfirms the conclusion!

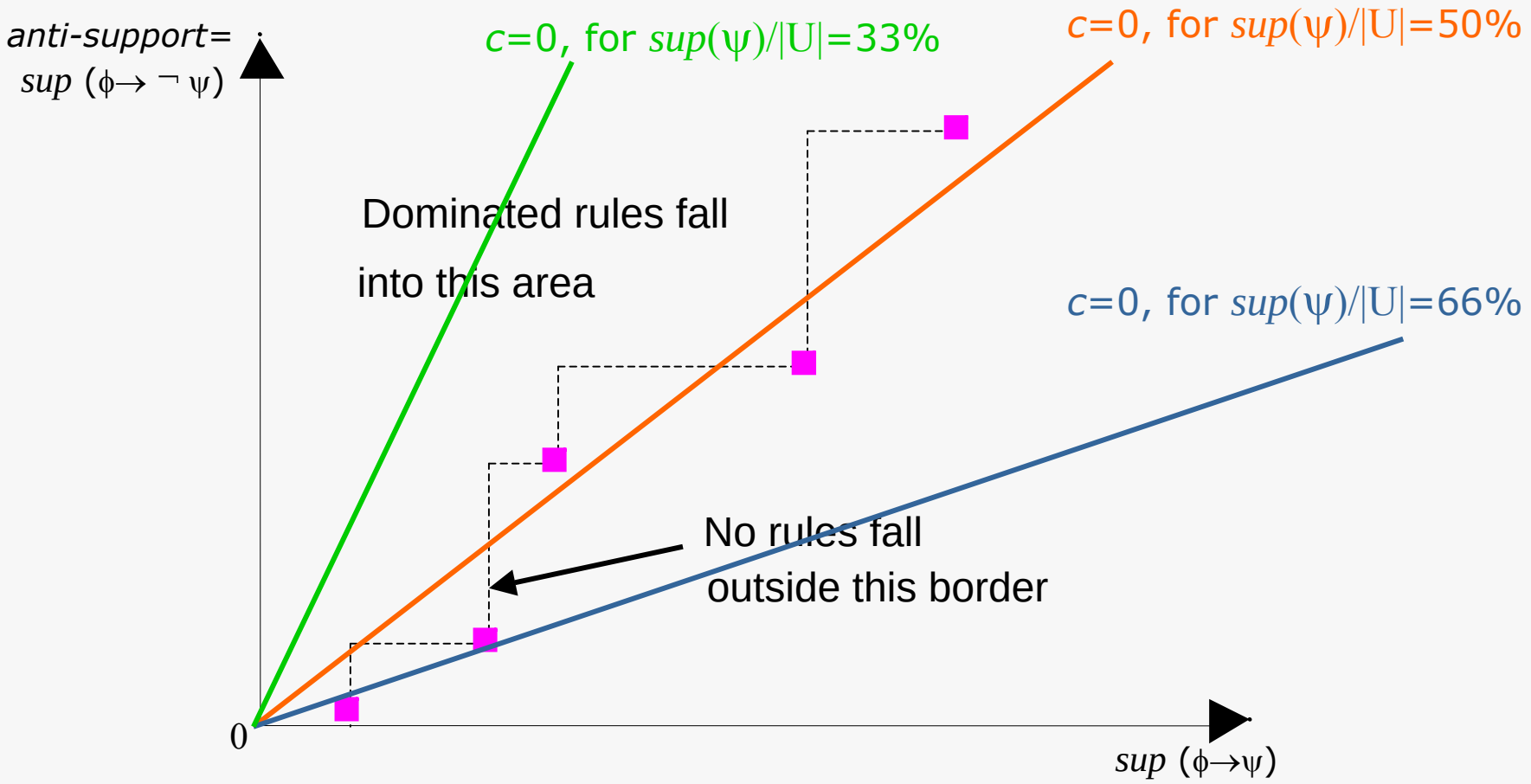
Confirmation perspective on support - anti-support space

- A more general condition $c(\phi, \psi) \geq k$, $k \geq 0$ for some specific confirmation measure, $f(\phi, \psi)$, was also investigated.

- *Theorem:*

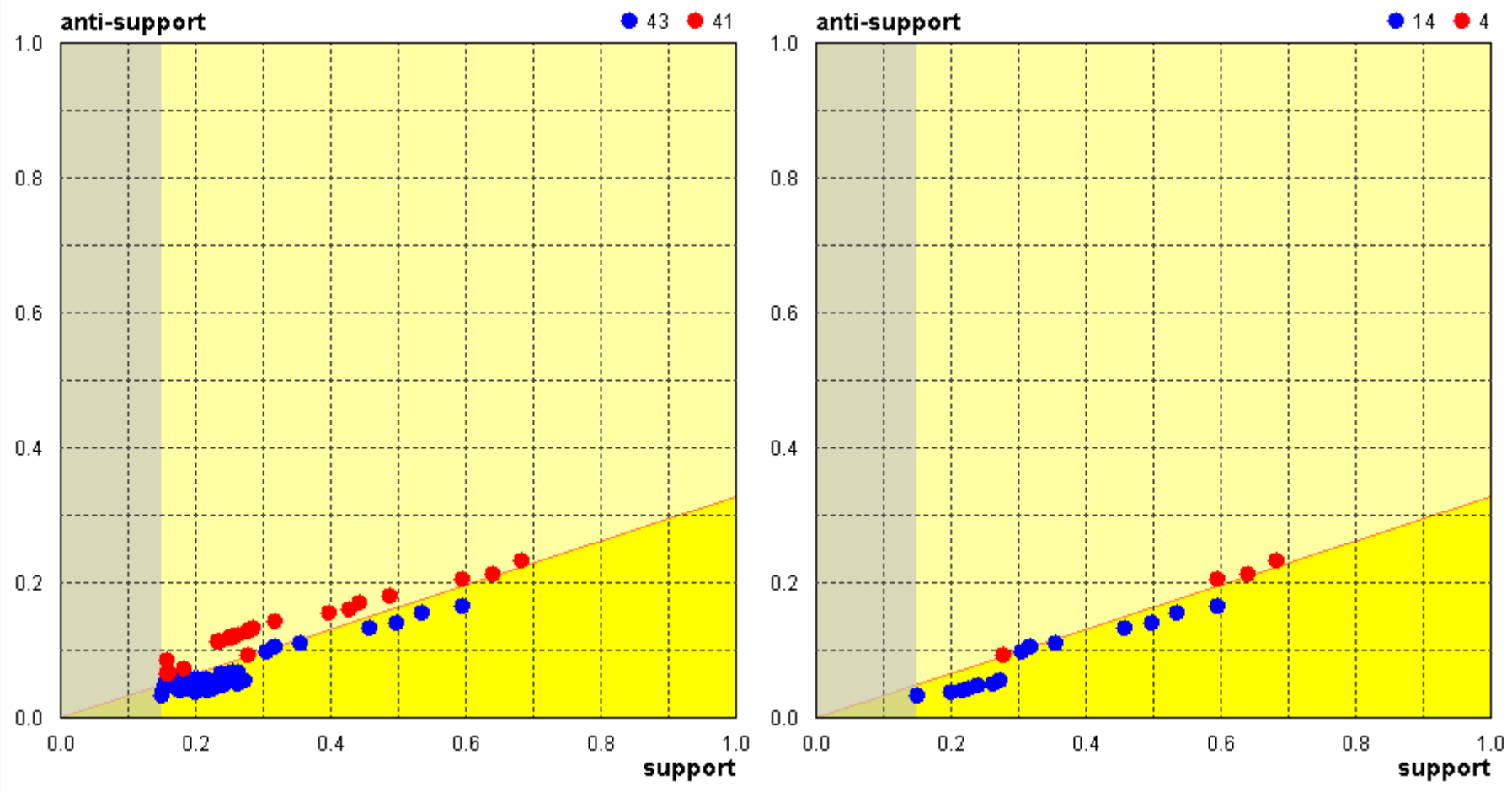
$$f(\phi \rightarrow \psi) \geq k \Leftrightarrow \textit{anti-sup}(\phi \rightarrow \psi) \leq \textit{sup}(\phi \rightarrow \psi)[|U|/\textit{sup}(\psi)\text{-}1]$$

Confirmation perspective on support - anti-support border



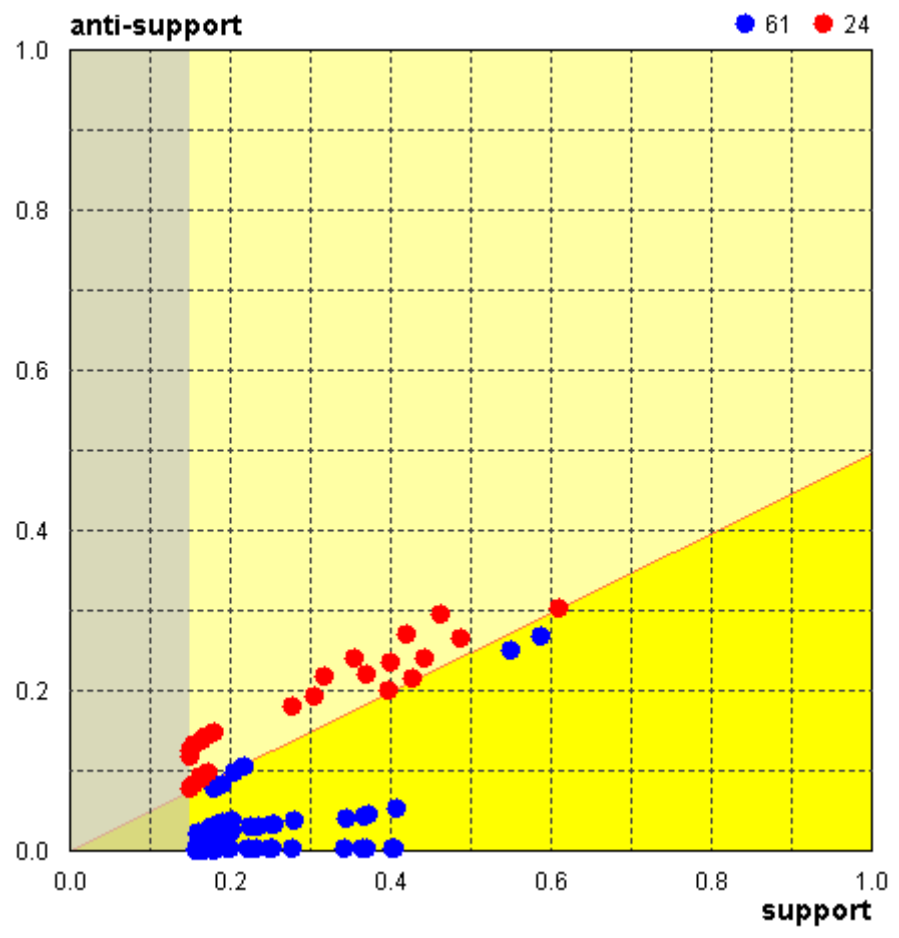
For rules lying above the curve for which $c=0$
the premise only disconfirms the conclusion

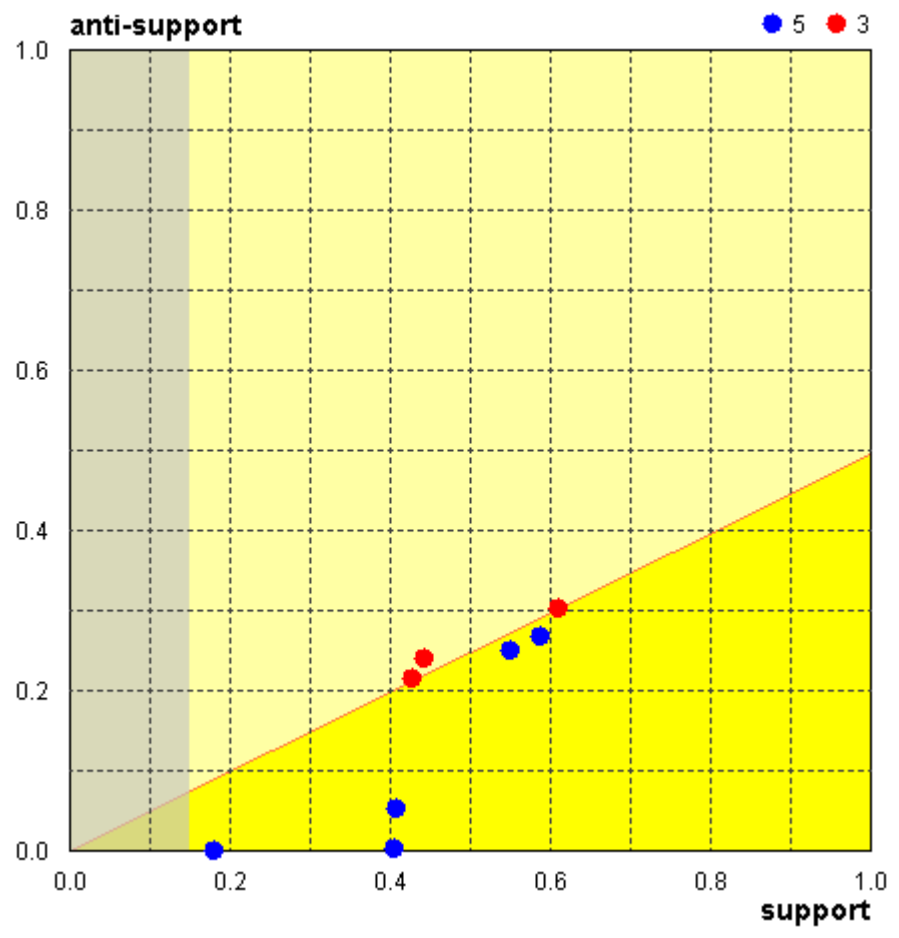
Support - anti-support (workclass=Private)



- ● indicates rules with negative confirmation
- even some rules from the Pareto border need to be discarded

Support-anti-support (sex=Male)





- ● indicates rules with negative confirmation

Support – anti-support - summary

Considered conclusion	No. of all rules	No. of rules with non–positive confirm.	Reduction percentage
workclass='Private'	84	41	49%
sex=Male	85	24	28%
income<=50kUSD	87	43	49%

Considered conclusion	No. of all rules on Pareto border	No. of rules with non–positive confirm.	Reduction percentage
workclass='Private'	18	4	22%
sex=Male	8	3	38%
income<=50 kUSD	15	4	27%

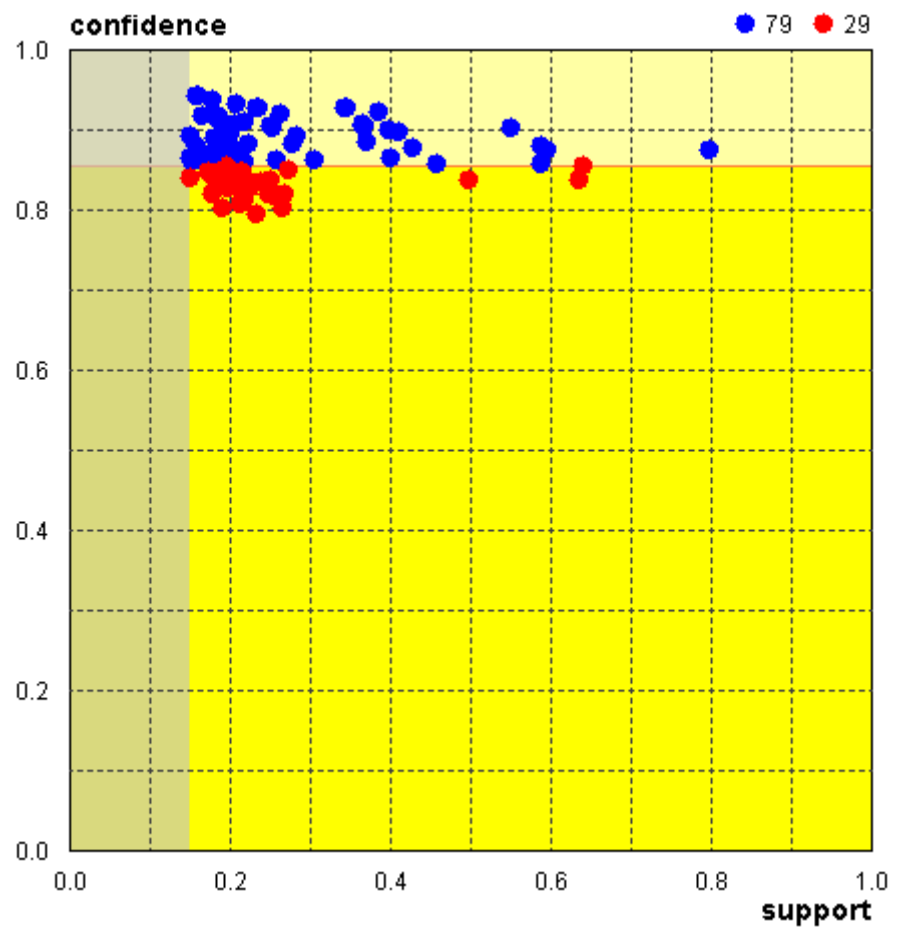
Summary

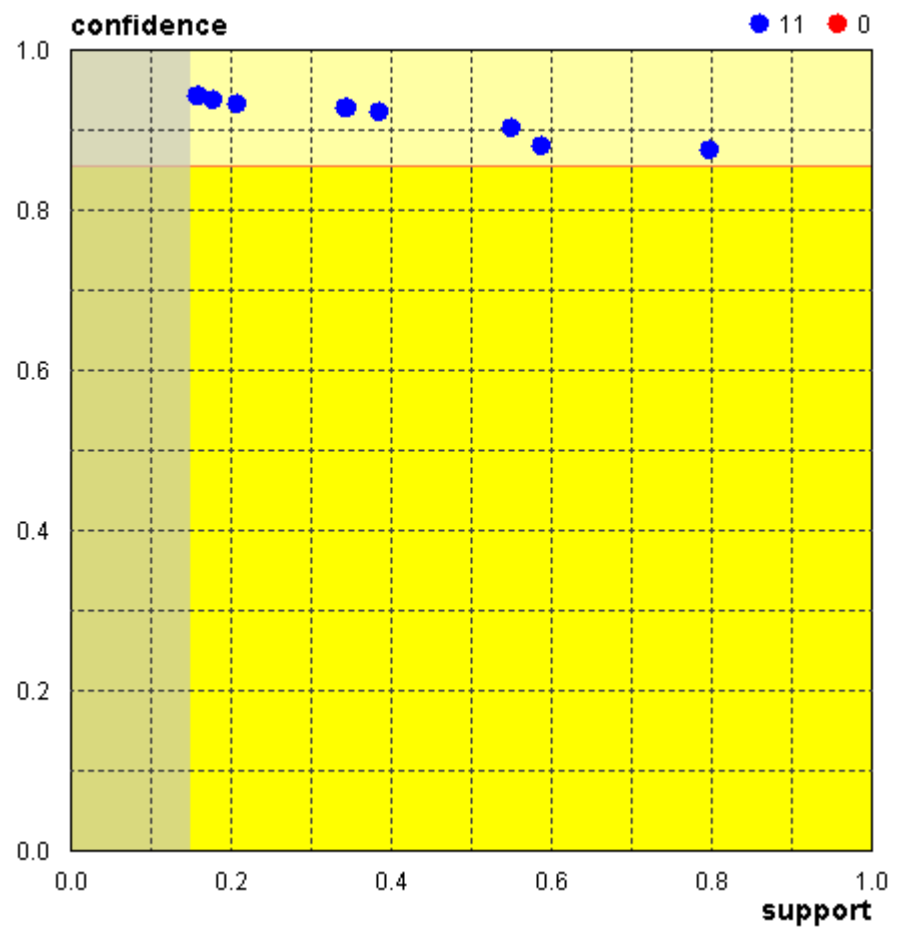
Summary

- The support-confidence and support – anti-support Pareto-optimal borders are characterized by valuable features.
- Inspired by the strength of the semantics of the family of confirmation measures, we have shown that it is reasonable to eliminate rules with non-positive or small values of confirmation.
- We have shown analytically that simple linear functions imposed on the two-dimensional spaces limit the set of induced rules to rules for which the premise confirms the conclusion.
- Experimental results show how big the reduction of a rule set can be.

Thank you!

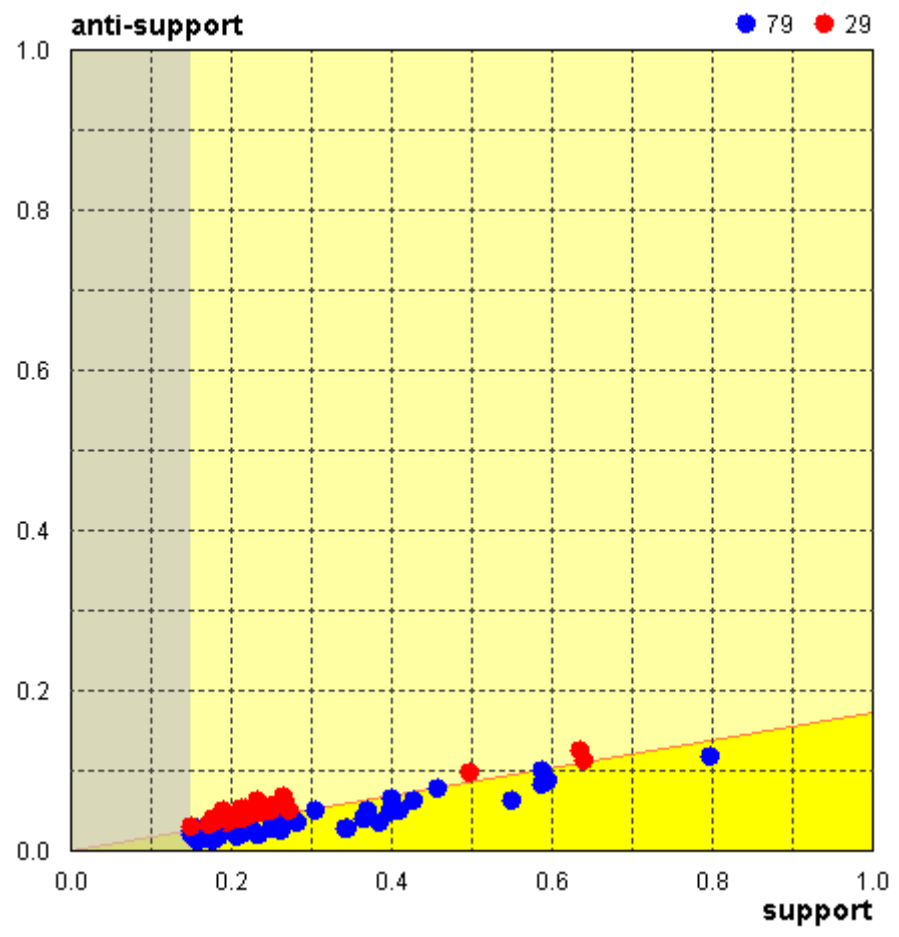
Support-confidence (race=White)

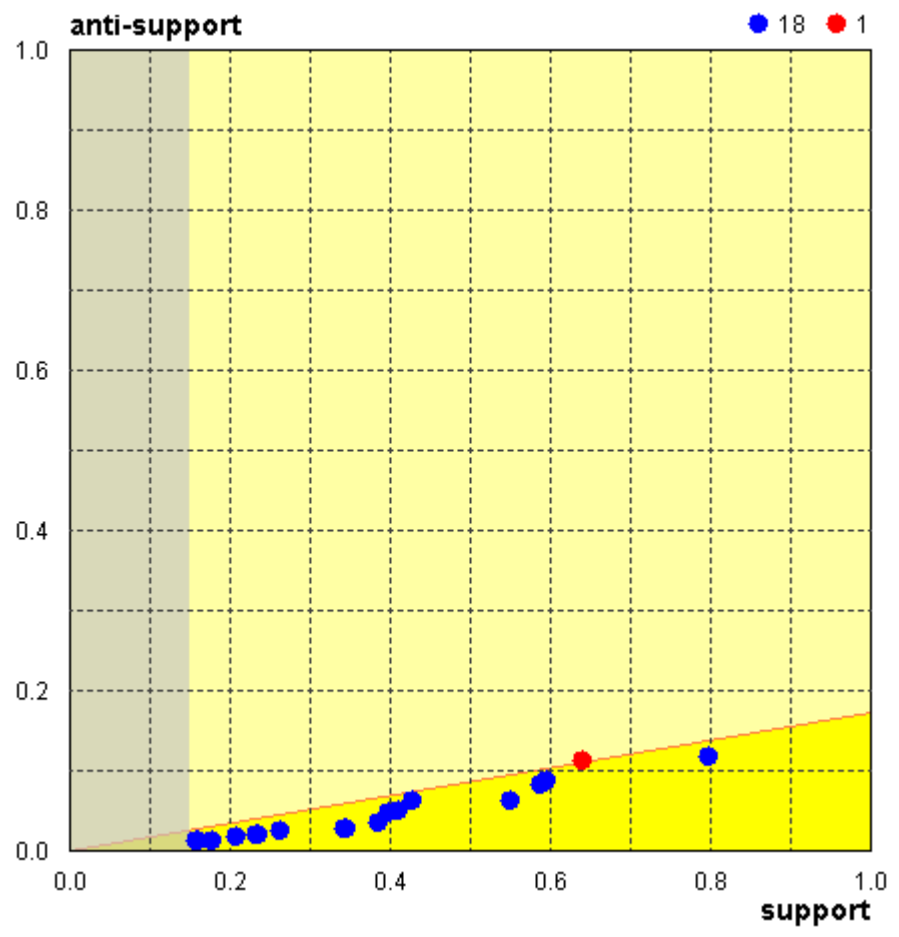




- indicates rules with negative confirmation

Support-anti-support (race=White)





- ● indicates rules with negative confirmation