Mining non-dominated rules with respect to support and anti-support

Izabela Szczęch
Poznań University of Technology

Poszukiwanie reguł niezdominowanych ze względu na wsparcie i anty-wsparcie

IDSS, 05.12.2006

Plan

Introduction

- Discovering rules from data is the domain of inductive reasoning (IR)
- IR uses data about a sample of larger reality to start inference
- $S=\langle U, A\rangle$ – *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*, $C\cap D=\varnothing$

e.g.

Characterization of nationalities

U	Height	Hair	Eyes	Nationality	Support
1	tall	blond	blue	Swede	270
2	medium	dark	hazel	German	90
3	medium	blond	blue	Swede	90
4	tall	blond	blue	German	360
5	short	red	blue	German	45
6	medium	dark	hazel	Swede	45

C (Height, Hair, Eyes) *D* (Nationality)

Introduction

- With every subset of attributes $B\subseteq A$, one can associate a formal language of formulas L, called *decision language*
- Formulas are built from attribute-value pairs (q,v), where $q\in B$ and $v\in V_a$ (domain of a), using logical connectives $\wedge$, $\vee$, $\neg$
- All formulas in L are partitioned into *condition* and *decision formulas* (called *premise* and *conclusion*, resp.)
- *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 decision formulas expressed in L

Introduction

- The number of rules generated from massive datasets can be very large and only a few of them are likely to be useful
- In all practical applications, like medical practice, market basket, it is crucial to know how good the rules are
- 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,...)
- There is no evidence which measure(s) is (are) the best

Basic quantitative characteristics of rules

- Notation:
  - $sup(\circ)$ is the number of all objects from U, having property ° in S
    e.g. $sup(\phi)$, $sup(\psi)$
- Basic quantitative characteristics of rules
  - *Support* of decision rule $\phi\rightarrow\psi$ in S:
    $$sup(\phi\rightarrow\psi)=sup(\phi\wedge\psi)$$
  - *Confidence* (called also *certainty factor*) of decision rule $\phi\rightarrow\psi$ in S (Łukasiewicz, 1913):
    $$conf(\phi\rightarrow\psi)=\frac{sup(\phi\rightarrow\psi)}{sup(\phi)}$$

Bayesian confirmation measures

- Among widely studied interestingness measures, there is a group of ***Bayesian confirmation measures***
- 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"

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

- Its meaning is different from a simple statistics of co-occurrence of properties ϕ and ψ in universe *U*

Bayesian confirmation measures

- Assuming $Fr(\psi)=\frac{sup(\psi)}{\text{card(U)}}$:

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

$$\Downarrow$$

$$c(\phi,\psi)\begin{cases} >0 \ \ if \ \ conf(\phi\rightarrow\psi) > Fr(\psi) \\ =0 \ \ if \ \ conf(\phi\rightarrow\psi) = Fr(\psi) \\ <0 \ \ if \ \ conf(\phi\rightarrow\psi) < Fr(\psi) \end{cases}$$

Desirable properties of confirmation measures

- Desirable properties of *c*(ϕ,ψ):
  - ***monotonicity* (M)** (Greco, Pawlak, Słowiński 2004):

    a=$sup(\phi\rightarrow\psi)$, ***b***=$sup(\neg\phi\rightarrow\psi)$, ***c***=$sup(\phi\rightarrow\neg\psi)$, ***d***=$sup(\neg\phi\rightarrow\neg\psi)$

    c(ϕ,ψ) = ***F***(***a***, ***b***, ***c***, ***d***), where ***F*** is a function non-decreasing with respect to ***a*** and ***d*** and non-increasing with respect to ***b*** and ***c***

  - ***hypothesis symmetry*** (Eells, Fitelson 2002): ***c***(ϕ,ψ)= - ***c***(ϕ, $\neg\psi$)

Properties of monotonicity (M)

- The property of monotonicity (M) takes into account four evidences in assessment of the impact of property ϕ on $\phi\rightarrow\psi$
- E.g. (Hempel) consider rule $\phi\rightarrow\psi$: ***if x is a raven, then x is black***
- ϕ is the property ***to be a raven*** and ψ is the property ***to be black***
  - ***a*** – the number of objects in ***S*** which are black ravens
  - ***b*** – the number of objects in ***S*** which are black non-ravens
  - ***c*** – the number of objects in ***S*** which are non-black ravens
  - ***d*** – the number of objects in ***S*** which are non-black non-ravens

Confirmation measure *f* and *s*

- As shown by (Greco, Pawlak, Słowiński 2004), confirmation measure *f* (Good 1984, Heckerman 1988, Pearl 1988, Fitelson 2001)

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

and confirmation measure *s* (Christensen 1999)

$$s(\phi\rightarrow\psi)=conf(\phi\rightarrow\psi)-conf(\neg\phi\rightarrow\psi)$$

are the only ones that enjoy both property of monotonicity (M) and hypothesis symmetry (HS), among the most well known confirmation measures

Utility of confidence vs. utility of confirmation *measures* (1)

- Utility of scales:
  - $conf(\phi\rightarrow\psi)$ is the truth value of the knowledge pattern ***„if*** ϕ, ***then*** ψ",
  - $f(\phi\rightarrow\psi)$, $s(\phi\rightarrow\psi)$ say to what extend ψ is satisfied more frequently when ϕ is satisfied rather than when ϕ is not satisfied

Utility of confidence vs. utility of confirmation measures e.g. 1

- Consider the possible result of rolling a die: 1,2,3,4,5,6, and let the conclusion be ψ=„the result is 6"
  - ϕ_1 ="the result is divisible by 3" $conf(\phi_1\rightarrow\psi)$=1/2, $f(\phi_1\rightarrow\psi)$=2/3
  - ϕ_2 ="the result is divisible by 2" $conf(\phi_2\rightarrow\psi)$=1/3, $f(\phi_2\rightarrow\psi)$=3/7
  - ϕ_3 ="the result is divisible by 1" $conf(\phi_3\rightarrow\psi)$=1/6, $f(\phi_3\rightarrow\psi)$=0
- In particular, rule $\phi_3\rightarrow\psi$, can be read as „in any case, the result is 6"; indeed, the „any case" does not add any information which could confirm that the result is 6, and this fact is expressed by $f(\phi_3\rightarrow\psi)=0$
- This example clearly shows that the value of f has a more useful interpretation than *conf*

13

Utility of confidence vs. utility of confirmation measures e.g. 2

- Consider the possible result of rolling a die: 1,2,3,4,5,6, and let the premise be ϕ=„the result is divisible by 2"
  - ψ_1="the result is 6" $conf(\phi\rightarrow\psi_1)$=1/3, $f(\phi\rightarrow\psi_1)$=3/7
  - ψ_2="the result is ***not*** 6" $conf(\phi\rightarrow\psi_2)$=2/3, $f(\phi\rightarrow\psi_2)$=–3/7
- In this example, rule $\phi\rightarrow\psi_2$ has greater confidence than rule $\phi\rightarrow\psi_1$
- However, rule $\phi\rightarrow\psi_2$ is less interesting than rule $\phi\rightarrow\psi_1$ because premise ϕ reduces the probability of conclusion ψ_2 from 5/6=$sup(\psi_2)$ to 2/3= $conf(\phi\rightarrow\psi_2)$, while it augments the probability of conclusion ψ_1 from 1/6=$sup(\psi_1)$ to 1/3= $conf(\phi\rightarrow\psi_1)$
- In consequence, premise ϕ disconfirms conclusion ψ_2, which is expressed by a negative value of $f(\phi\rightarrow\psi_2)=-3/7$, and it confirms conclusion ψ_1, which is expressed by a positive value of $f(\phi\rightarrow\psi_1)$=3/7

14

Support-confidence Pareto border

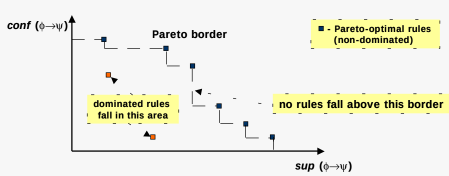
Support-confidence Pareto border

- In the set of rules induced from data, we look for rules that are optimal according to a chosen attractiveness measure
- This problem was addressed with respect to such measures as *lift*, *gain*, *conviction*, *Piatetsky-Shapiro*,...
- Bayardo and Agrawal (1999) proved, however, that given a fixed conclusion ψ, the *support-confidence Pareto border* (i.e. Pareto-optimal border w.r.t. rule support and confidence) includes optimal rules according to any of those attractiveness measures

16

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

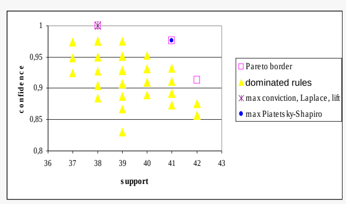


- Mining the border identifies rules optimal with respect to measures such as: *lift*, *gain*, *conviction*, *Piatetsky-Shapiro*,...

17

Support-confidence Pareto border

E.g. „Buses" data set, class of „good state"



- 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, Słowiński, Stefanowski 2001)

Rule induction algorithm: all rules algorithm (DOMAPRIORI)

18

Support-confidence Pareto border

- The following conditions are sufficient for verifying whether rules optimal according to a measure $g(x)$ are included on the support-confidence Pareto border:

  1. $g(x)$ is monotone in support over rules with the same confidence

  and

  2. $g(x)$ is monotone in confidence over rules with the same support

- A function $g(x)$ is understood to be *monotone* in x, if x_1 p x_2 implies that $g(x_1) \leq g(x_2)$

19

Support-f Pareto border

Monotonicty of *f* in support and confidence

- Is confirmation measure f included in the support-confidence Pareto border?
- Theorem 1:
  Confirmation measure f is independent of support, and, therefore, monotone in support, when the value of confidence is held fixed
- Theorem 2:
  Confirmation measure f is increasing, and, therefore, monotone in confidence
- Conclusion:

  Rules maximizing f lie on the support-confidence Pareto border
  (rules with fixed conclusion)

21

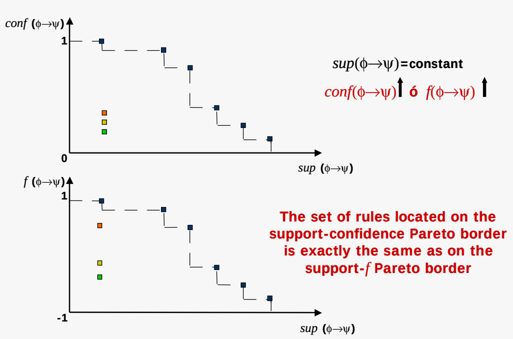
Monotonicty of confidence in support and *f*

- The utility of confirmation measure f outranks utility of confidence
- Claim 1: Substitute the $conf(\phi\rightarrow\psi)$ dimension for $f(\phi\rightarrow\psi)$ in the support-confidence Pareto border
- Corollary 1:
  Confidence is independent of support, and, therefore, monotone in support, when the value of $f(\phi\rightarrow\psi)$ is held fixed
- Corollary 2:
  Confidence is increasing, and, therefore, monotone in $f(\phi\rightarrow\psi)$
- Conclusion:

  The set of rules located on the support-confidence Pareto border is exactly the same as on the support-f Pareto border

22

Support-confidence vs. support-*f* Pareto border



23

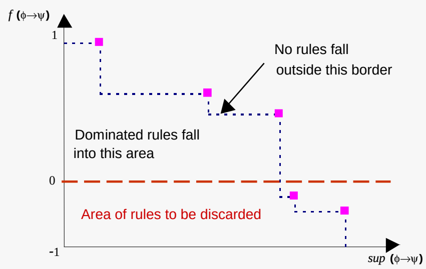
Support-confidence vs. support-*f* Pareto border

- All the other interestingness measures that were represented on the support-confidence Pareto border also reside on support-f Pareto border
- Any non-dominated rule with a negative value of $f(\phi\rightarrow\psi)$ must be discarded from further analysis as its premise only disconfirms the conclusion – such situation cannot be expressed by the scale of confidence
- Conclusion:

  The support-f Pareto border is more meaningful than the support-confidence Pareto border

24

Support-*f* Pareto border is more meaningful



Support-s Pareto border

Monotonicty of *s* in support and confidence

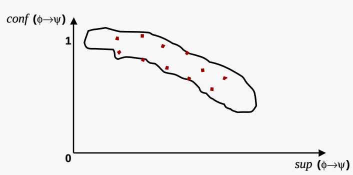
- Is confirmation measure s on rule support-confidence Pareto border?
- **Theorem 3:**
  Confirmation measure s is increasing, and, therefore, monotone in confidence when the value of support is held fixed
- **Theorem 4:**
  For a fixed value of confidence, confirmation measure s is:
  - increasing in $sup(\phi\rightarrow\psi) \Leftrightarrow s(\phi\rightarrow\psi)>0$
  - constant in $sup(\phi\rightarrow\psi) \Leftrightarrow s(\phi\rightarrow\psi)=0$
  - decreasing in $sup(\phi\rightarrow\psi) \Leftrightarrow s(\phi\rightarrow\psi)<0$
- Theorem 4 states the monotone relationship just in the non-negative range of the value of *s* (i.e. the only interesting)

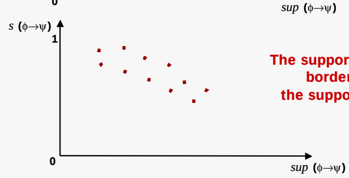
Support-confidence vs. support-*s* Pareto border

- **Theorem 5:**
  If a rule resides on the support-s Pareto border
  (in case of positive value of s),
  then it also resides on the support-confidence Pareto border,

  while one can have rules being on the support-confidence Pareto border which are not on the support-s Pareto border.
- **Conclusion:**
  The support-confidence Pareto border is, in general, larger than the support-s Pareto border

Support-confidence vs. support-*s* Pareto border





The support-confidence Pareto border is larger than the support-s Pareto border

Confirmation measures with the property of monotonicity (M)

- What are the necessary and sufficient conditions for rules maximizing a confirmation measure $c(\phi,\psi)$ with the property of monotonicity (M) to be included in the rule support-confidence Pareto border?
- Reminder of the property of monotonicity (M):
  $a=sup(\phi\rightarrow\psi)$, $b=sup(\neg\phi\rightarrow\psi)$, $c=sup(\phi\rightarrow\neg\psi)$, $d=sup(\neg\phi\rightarrow\neg\psi)$
  $c(\phi,\psi)=F(a,b,c,d)$, where F is a function non-decreasing with respect to a and d, and non-increasing with respect to b and c

Confirmation measures with the property of monotonicity (M)

- Let $F(a, b, c, d)$ be a confirmation measure with the property (M)
- Theorem 6:
  When the value of support is held fixed, then $F(a, b, c, d)$ is monotone in confidence.
- Theorem 7:
  When the value of confidence is held fixed, then $F(a, b, c, d)$ admitting derivative with respect to all its variables a, b, c and d, is monotone in support if:

$$\frac{\partial F}{\partial c}=\frac{\partial F}{\partial d}=0 \quad or \quad \frac{\frac{\partial F}{\partial a}-\frac{\partial F}{\partial b}}{\frac{\partial F}{\partial d}-\frac{\partial F}{\partial c}} \geq \frac{1}{conf(\phi \rightarrow \psi)}-1$$

31

Confirmation measures with the property of monotonicity (M)

- Conclusions:
  - *Theorem 6* states that for a set of rules with the same conclusion, any Bayesian confirmation measure satisfying the property of monotonicity (M) is always non-decreasing with respect to confidence when the value of support is kept fixed
  - Due to *Theorem 7*, all those confirmation measures that are independent of $c=sup(\phi \rightarrow \neg\psi)$ and $d=sup(\neg\phi \rightarrow \neg\psi)$ are always monotone in support when the value of confidence remains unchanged

32

Support-confidence vs. support-*i* Pareto border

- Theorem 8:
  Given an interestingness measure *i*, which is monotone with respect to support and confidence, if a rule resides on the support-*i* Pareto-optimal border, then it also resides on the support-confidence Pareto-optimal border

  while the opposite assertion is not necessarily true.

- Conclusion:

  The support-confidence Pareto border is, in general, larger than the support-*i* Pareto border

33

Pareto borders - summary

- Inclusion of Pareto-optimal borders:
  - Support-confidence = support-*f*
  - Support-confidence Ê support-*s*
  - Support-confidence Ê support-*i*
    - *i* is any interestingness measure monotone with respect to support and confidence

34

Area of interesting rules with respect to support and confidence

Which dominated rules (supp-conf) are definitely NOT interesting?

- Let us suppose that *F* is a confirmation measure with the property of monotonicity (M).
- We know that when $sup(\phi \rightarrow \psi)$=constant:
  - confidence is *monotone* (non-decreasing) w.r.t. *F*.
- Claim 2: Due to monotonicity of confidence in *F*, rules lying below the curve for which *F*=0 must be discarded.
  For those rules, the premise only disconfirms the conclusion!

36

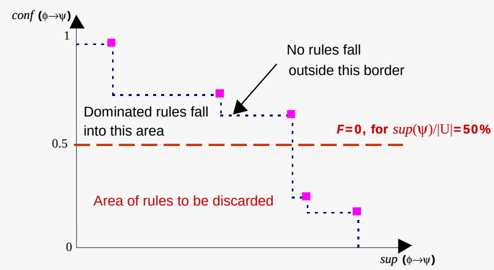
Which dominated rules (supp-conf) are definitely NOT interesting?

- Let us recall the definition of *F*:

$$c(\phi,\psi)\begin{cases} >0 & if\ \ conf(\phi\rightarrow\psi) > Fr(\psi) \\ =0 & if\ \ conf(\phi\rightarrow\psi) = Fr(\psi) \\ <0 & if\ \ conf(\phi\rightarrow\psi) < Fr(\psi) \end{cases}$$

- Let us assume that: $Fr(\psi)=\frac{sup(\psi)}{\mathrm{card}(U)}$
- Claim 3: *F*=0 ó $\boxed{conf(\phi\rightarrow\psi)=\frac{sup(\psi)}{\mathrm{card}(U)}}$
- $\frac{sup(\psi)}{\mathrm{card}(U)}$ is a constant expressing what percentage of the whole data set is taken by considered class ψ

Which dominated rules (supp-conf) are definitely NOT interesting?



For rules lying below the curve for which *F*=0 the premise only disconfirms the conclusion

Support-anti-support Pareto border

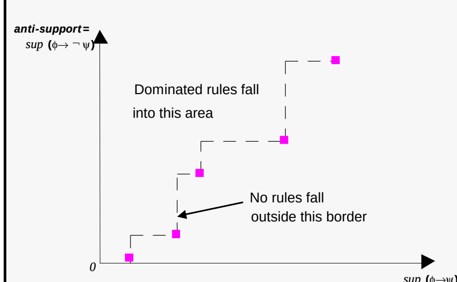
Support-anti-support Pareto border

- How to find rules optimal according to any confirmation measure with the property (M)?
- Anti-support is the number of examples which satisfy the premise of the rule but not its conclusion: $sup(\phi\rightarrow\neg\psi)$
- Theorem 9:
  When the value of support is held fixed, then *F*(*a*, *b*, *c*, *d*) is anti-monotone (non-increasing) in anti-support
- Theorem 10:
  When the value of anti-support is held fixed, then *F*(*a*, *b*, *c*, *d*) is monotone (non-decreasing) in support

Support-anti-support Pareto border

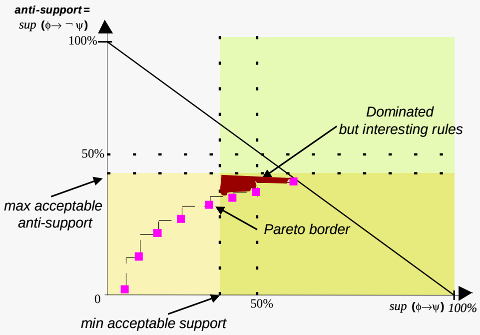
- Claim 4:
  - The best rules according to any of the confirmation measures with the property of monotonicity (M) must reside on the support-anti-support Pareto border
- The support-anti-support Pareto border is the set of rules such that there is no other rule having greater support and smaller anti-support

Support-anti-support Pareto border



The best rules according to any of the confirmation measures with the property of monotonicity (M) must reside on the support-anti-support Pareto border

Dominated rules can also be interesting



43

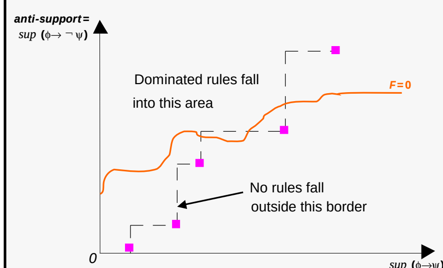
Area of interesting rules
with respect to support and anti-support

Which dominated rules are definitely NOT interesting?

- **Let us suppose that *F* is a confirmation measure with the property of monotonicity (M).**
- **We know that when $sup(\phi\rightarrow\psi)$=constant:**
  - **anti-support is *anti-monotone* (non-increasing) w.r.t. confindence,**
  - **anti-support is *anti-monotone* (non-increasing) w.r.t. *F*.**
- **Claim 5: Due to anti-monotonicity of anti-support in *F*, rules lying above the curve for which *F*=0 must be discarded.**
  For those rules, the premise only disconfirms the conclusion

45

Which dominated rules are definitely NOT interesting?



For rules lying above the curve for which *F*=0
the premise only disconfirms the conclusion

46

Which dominated rules are definitely NOT interesting?

- **Let us recall the definition of *F*:**

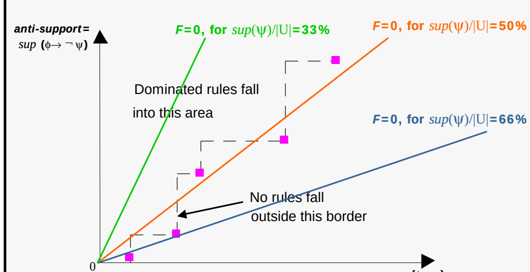
$$c(\phi,\psi)\begin{cases} >0 & if\ \ conf\,(\phi\rightarrow\psi) > Fr\,(\psi)\\ =0 & if\ \ conf\,(\phi\rightarrow\psi) = Fr\,(\psi)\\ <0 & if\ \ conf\,(\phi\rightarrow\psi) < Fr\,(\psi)\end{cases}$$

- **Claim 6: *F*=0 ó** $\boxed{anti\text{-}sup(\phi\rightarrow\psi)=sup(\phi\rightarrow\psi)\left[\frac{\mathrm{card}(U)}{sup(\psi)}-1\right]}$

- $anti\text{-}sup(\phi\rightarrow\psi)=sup(\phi\rightarrow\psi)\left[\frac{\mathrm{card}(U)}{sup(\psi)}-1\right]$ **is a linear function**

47

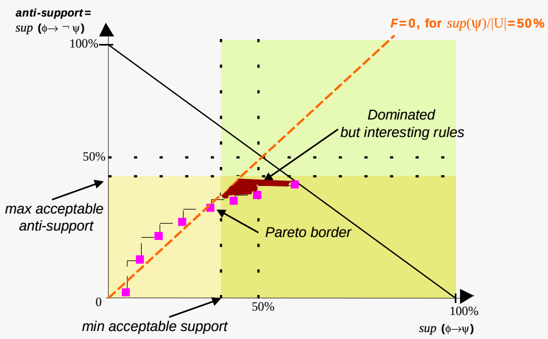
Which dominated rules are definitely NOT interesting?



For rules lying above the curve for which *F*=0
the premise only disconfirms the conclusion

48

Dominated rules can also be interesting



49

Summary

Summary

- Many attractiveness measures can be identified by mining the support-confidence Pareto border – very practical result
- The utility of confirmation measures outranks the utility of confidence
- Suggested new Pareto borders:
  - support-*f* Pareto border
  - support-*s* Pareto border
- Pareto border w.r.t. support and anti-support includes rules maximizing all confirmation measures with the property (M)

51

Summary

- Dominated rules can also be interesting
- We have shown that for
  - support-confidence Pareto border
  - support-anti-support Pareto border

  simple linear functions narrow the area of dominated rules only to rules for which the premise confirms the conclusion

52

Experimental results

Thanks to Mirek Urbanowicz

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*(f ® y) is denoted as „support” and expressed as a relative rule support [0-1]

54

The gist of the algorithm for support-anti-support rules

- **Traditional Apriori approach to generation of association rules (Agrawal et al) proceeds in a two step framework:**
  - **find frequent itemsets (i.e. sets of items which occur more frequently than the minimum support threshold),**
  - **generate rules from frequent itemsets and filter out those that do not exceed the minimum confidence threshold**
- **Generation of association rules regarding support and anti-support, in general, requires only the substitution of the parameter calculated in step 2. Confidence -> anti-support**

The gist of the algorithm for support-anti-support rules

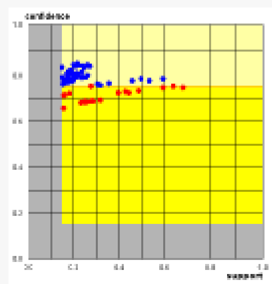
- **Since** $conf(\phi\rightarrow\psi)=sup(\phi\rightarrow\psi)/sup(\phi)$ **all the data needed to calculate it are already gathered in step 1 of Apriori**
- **Claim 7: calculation of anti-support (instead of confidence) does not introduce any more computational overhead to the algorithm**
- **Let us observe that:** $anti\text{-}sup(\phi\rightarrow\psi) = sup(\phi\rightarrow\neg\psi) = sup(\phi)\text{-}sup(\phi\rightarrow\psi)$.
- **All the data required to calculate anti-support are also gathered in step 1 of Apriori**
- **The data needed to calculate anti-support is the same as to calculate confidence, and moreover subtraction is easier than division**

The gist of the algorithm for support-anti-support rules

- **Claim 8: When generating association rules from a frequent set it is advisable to first generate rules with few conclusion elements (for optimisation reasons)**
- **Let us observe three different rules constructed from the same frequent itemset** $\{a, b, c, d\}$**:**
  - r_1: $a\text{->}bcd$ $\quad anti\text{-}sup(r_1) = sup(a)\text{-}sup(abcd)$
  - r_2: $ab\text{->}cd$ $\quad anti\text{-}sup(r_2) = sup(ab)\text{-}sup(abcd)$
  - r_3: $abc\text{->}d$ $\quad anti\text{-}sup(r_3) = sup(abc)\text{-}sup(abcd)$
- $anti\text{-}sup(r_1) \geq anti\text{-}sup(r_2) \geq anti\text{-}sup(r_3)$
- **Conclusion:** $anti\text{-}sup(r_3)$ **> max_acceptable anti-support =>** $anti\text{-}sup(r_2)$ **> max_acceptable anti-support**

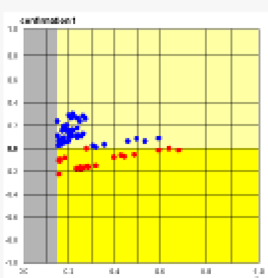
Generate and verify r_3 first!

Support-confidence (workclass=Private)



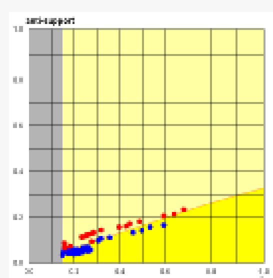
- ● indicates rules with negative confirmation
- the class constitutes over 70% of the whole dataset
- rules with high confidence can be disconfirming
- even some rules from the Pareto border need to be discarded

Support-f (workclass=Private)



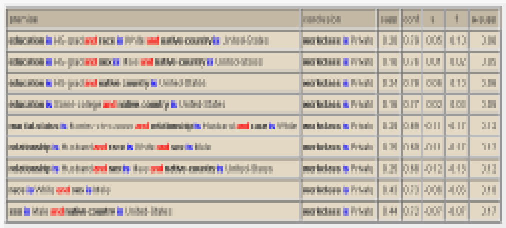
- ● indicates rules with negative confirmation
- this diagram does not (explicitly) show the ratio of the class cardinality to the whole dataset
- even some rules from the Pareto border need to be discarded

Support-anti-support (workclass=Private)



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

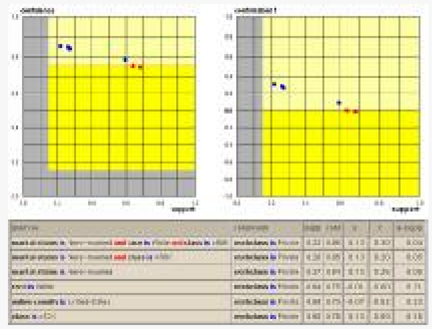
Few rules describing class: workclass=Private



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

61

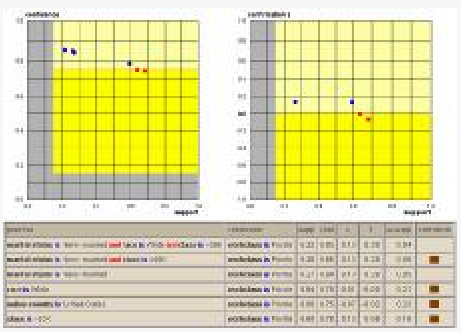
Support-confidence Pareto border vs. support-f



- both Pareto borders contain the same rules

62

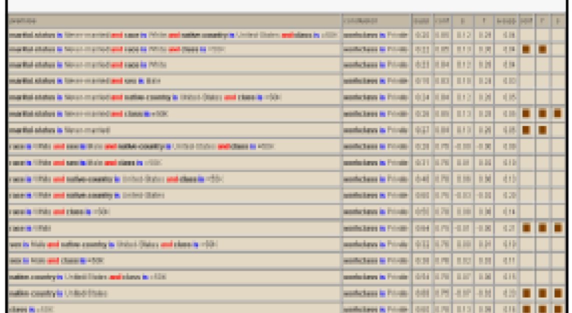
Support-confidence Pareto border vs. support-s



■ indicates rules that appeared on both Pareto borders

63

Comparison of all Pareto borders



■ indicates rules that appeared on a particular Pareto border

64

Final remarks

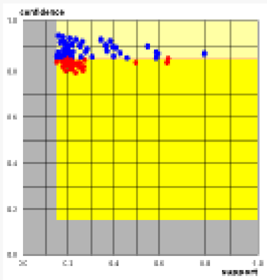
- The experiment is an illustration of all the studied and proved features of different Pareto-borders on a real dataset
- Further research will include
  - conducting of such an experiment for decision rules, at-least/at-most rules
  - searching for optimisation tricks (mostly structural) to improve the efficiency of the algorithm for rule generation

65

Thank you!

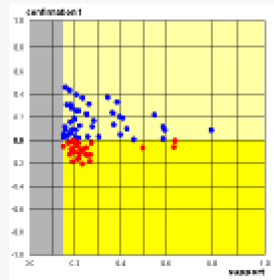
66

Support-confidence (race=White)



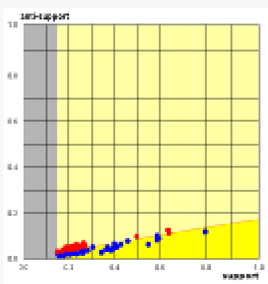
indicates rules with negative confirmation

Support-f (race=White)



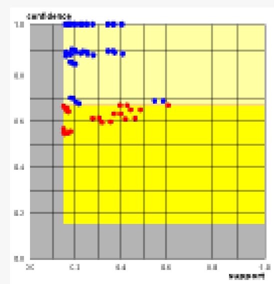
indicates rules with negative confirmation

Support-anti-support (race=White)



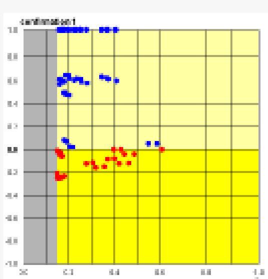
- indicates rules with negative confirmation

Support-confidence (sex=Male)



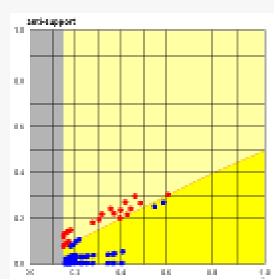
indicates rules with negative confirmation

Support-f (sex=Male)



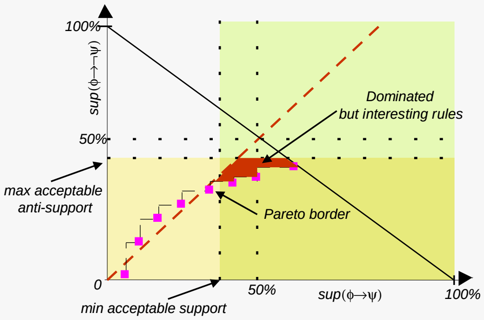
indicates rules with negative confirmation

Support-anti-support (sex=Male)



- indicates rules with negative confirmation
- the (sex=Male) class is smaller than workclass=Private

Support-anti-support Pareto border



More detailed info about the dataset

- **the dataset was downloaded from the repository of Univ. of California, Irvine**
- **32 561 instances**
- **9 nominal attributes**
  - **workclass:** Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked;
  - **education:** Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool;
  - **marital-status:** Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse;
  - **occupation:** Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces;
  - **relationship:** Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried;
  - **race:** White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.sex: Female, Male;
  - **sex :** Male, Female;
  - **native-country:** United-States, Cambodia, England, etc.;
  - **salary:** >50K, <=50K