STARK:
Self-Tuning Association Rules for KNIME
A Pascal-2 Harvest Project

José L Balcázar
Departament de Llenguatges i Sistemes Informàtics UPC,
Barcelona, Spain
jose.luis.balcazar@upc.edu

Pascal-2 Steering Committee, Cumberland Lodge, march 2012

Implications

As a data analysis tool

Examples

From abstracts in reports of the Pascal repository:

descent $\Longrightarrow$ gradient

hilbert $\Longrightarrow$ space

margin support $\Longrightarrow$ vector

Implications

As a data analysis tool

Examples

From abstracts in reports of the Pascal repository:

descent $\Longrightarrow$ gradient

hilbert $\Longrightarrow$ space

margin support $\Longrightarrow$ vector

carlo $\Longrightarrow$ monte

monte $\Longrightarrow$ carlo

Implications

As a data analysis tool

Examples

From abstracts in reports of the Pascal repository:

descent $\Longrightarrow$ gradient

hilbert $\Longrightarrow$ space

margin support $\Longrightarrow$ vector

carlo $\Longrightarrow$ monte

monte $\Longrightarrow$ carlo

Example from a "census" dataset:

Exec-managerial Husband $\Longrightarrow$ Married-civ-spouse

Husband $\Longrightarrow$ Male

Implications

As a data analysis tool

Examples

From abstracts in reports of the Pascal repository:

descent $\Longrightarrow$ gradient

hilbert $\Longrightarrow$ space

margin support $\Longrightarrow$ vector

carlo $\Longrightarrow$ monte

monte $\Longrightarrow$ carlo

Example from a "census" dataset:

Exec-managerial Husband $\Longrightarrow$ Married-civ-spouse

Husband $\Longrightarrow$ Male... does not hold!

Implications
As a data analysis tool

Examples

From abstracts in reports of the Pascal repository:

descent $\Longrightarrow$ gradient

hilbert $\Longrightarrow$ space

margin support $\Longrightarrow$ vector

carlo $\Longrightarrow$ monte

monte $\Longrightarrow$ carlo

Example from a "census" dataset:

Exec-managerial Husband $\Longrightarrow$ Married-civ-spouse

Husband $\Longrightarrow$ Male. . . does not hold!

Similarly, Wife $\Longrightarrow$ Female does not hold either:

there are two tuples declaring Male and Wife.

We need to accept exceptions.

The Confidence-and-Support Framework, I

The confidence bound

The very notion of association rules is not fully defined until we define how exceptions are handled: the measure of "intensity of the implication".

Most popular measures:

Confidence (that is, frequentist conditional probability); lift, leverage, weighted relative accuracy. . . ;

The Confidence-and-Support Framework, I

The confidence bound

The very notion of association rules is not fully defined until we define how exceptions are handled: the measure of "intensity of the implication".

Most popular measures:

Confidence (that is, frequentist conditional probability); lift, leverage, weighted relative accuracy...; but:

- How to set the threshold?
- Many measures take values in unpredictable intervals.
- The basic properties of association rules depend of the implication intensity notion.

The Confidence-and-Support Framework, II

The support bound

Notion of support: amount of observations where all the items in the implication are present.

Lower bound on the support

Two reasons:

- Avoid potential spurious statistical artifacts.
- Exponential powerset size may blow up memory:
  - Slow-down due to virtual memory leads to stalling.
  - Even the hard drive availability can be exhausted.
  - Huge lattices take looooooong to explore.

The Confidence-and-Support Framework, II

The support bound

Notion of support: amount of observations where all the items in the implication are present.

Lower bound on the support

Two reasons:

- Avoid potential spurious statistical artifacts.
- Exponential powerset size may blow up memory:
  - Slow-down due to virtual memory leads to stalling.
  - Even the hard drive availability can be exhausted.
  - Huge lattices take looooooong to explore.
- How to set the support threshold? There are examples of datasets leading to both extreme behaviors:
  - algorithms choke if you ask them to go below 98%,
  - algorithms find nothing until reaching down to 0.1%.

As a Data Mining Tool

Associations are not the most successful technology so far

End-user point of view

The rumor: association rules don't actually work.

- Most association miners yield very redundant rules.
- Hardly any sensible rule is found:

  "Most sessions start at the front page"

  "Most sessions visiting assignments start at the front page"

  "Most sessions visiting grades start at the front page"

  "Most sessions visiting contents start at the front page"

  ...

Example

Dataset on Abstracts of PASCAL Reports

Standard miner output

At 70% confidence and 5% support, among others,

support $\longrightarrow$ vector (12.6, 81.3)

vector $\longrightarrow$ support (13.3, 77.1)

Example

Dataset on Abstracts of PASCAL Reports

Standard miner output

At 70% confidence and 5% support, among others,

support $\longrightarrow$ vector (12.6, 81.3)

vector $\longrightarrow$ support (13.3, 77.1)

machines support $\longrightarrow$ vector (6.4, 100.0)

machines vector $\longrightarrow$ support (6.5, 97.9)

support using $\longrightarrow$ vector (6.0, 88.4)

vector using $\longrightarrow$ support (6.2, 84.4)

support data $\longrightarrow$ vector (5.4, 82.1)

vector data $\longrightarrow$ support (5.8, 76.2)

support paper $\longrightarrow$ vector (5.4, 82.1)

vector paper $\longrightarrow$ support (5.3, 84.2)

...

Yet Another Example

Dataset on Contraceptive Method Choice

Standard miner output

At 90% confidence and 10% support:

1. wife-education=4 contraception=2

→

media-exposure=0 (conf:1)

2. husband-education=4 no-working-now=1 standard-of-living=3

→

media-exposure=0 (conf:1)

...

64. husband-occupation=1 contraception=2

→

media-exposure=0 (conf:0.98)

Redundancy in Association Rules

A Logic-based view

Standard Association Mining Process

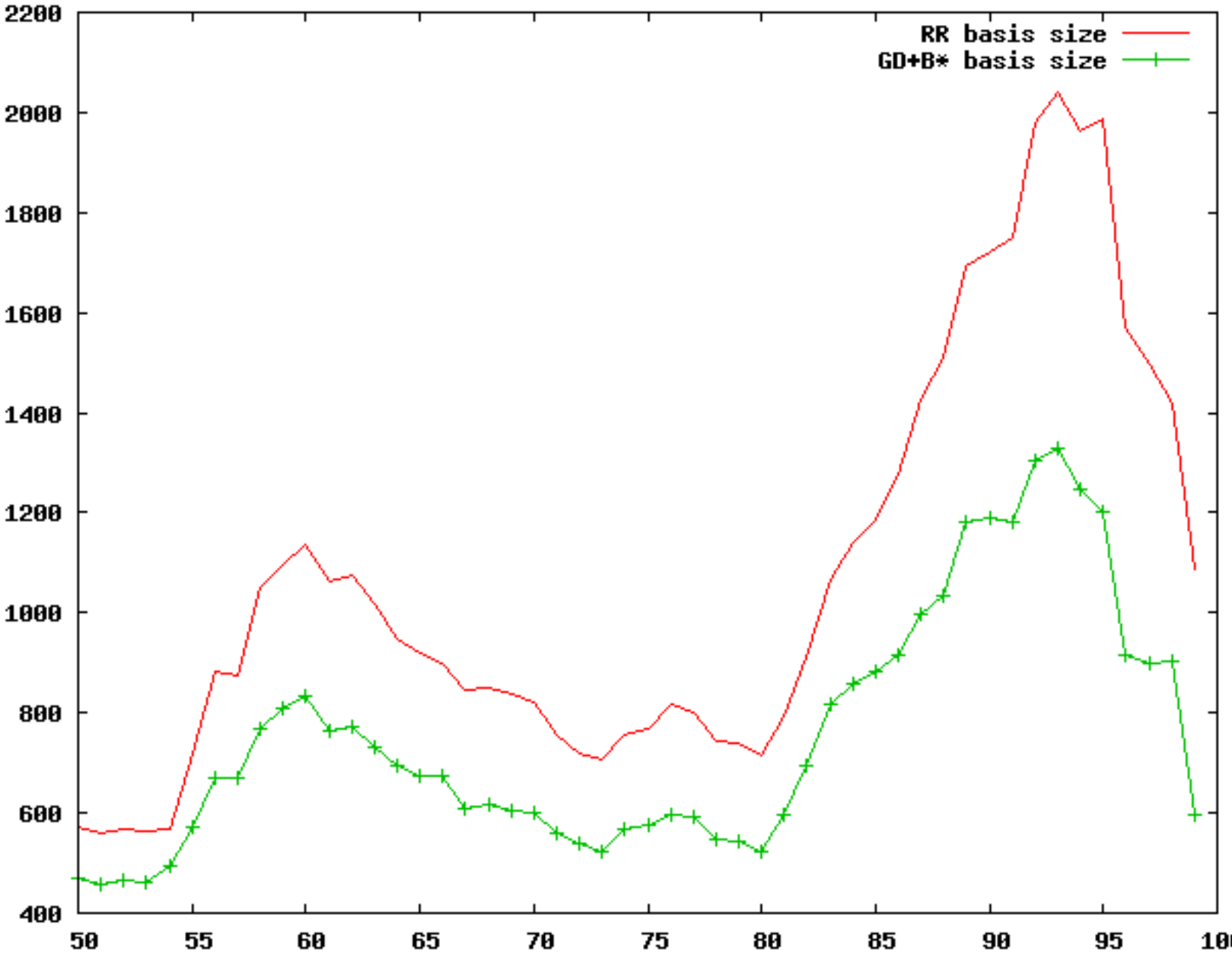
User provides dataset and thresholds for support and confidence, and gets all rules that hold in the dataset at those levels or higher.

Huge set of rules, growing further for lower thresholds. How to offer the user a smallish set of output rules?

- Several natural notions of "redundancy" and "minimum-size bases" that cover all rules.
- Essentially, two variants, according to whether full-confidence implications are treated separately or not.

Irredundant Rules for Dataset FIMI `pumsb-star`

Inspires a notion of "novelty"



Closure-Based Confidence Boost

One way through

For the usual confidence-and-support scheme:

High thresholds give nothing of interest, but lowering them (specially confidence) leads to too many rules to browse manually. How to discard rules?

- "Logical" redundancy approach (not really useful yet);
- "logical" novelty: confidence width (promising, but still somewhat unsatisfactory);
- (closure-based) confidence boost:
  - an "intuitive" variant of redundancy,
  - corresponds to "relative confidence",
  - advantageous not only in that it sets apart few rules, but also in the intuitive quality of the rules.

Parameter-Less, "Self-Tuning" Option

A somewhat bold attempt

Ask nothing from the user

(except of course the file name of the dataset).

- Confidence?

Parameter-Less, "Self-Tuning" Option

A somewhat bold attempt

Ask nothing from the user

(except of course the file name of the dataset).

- Confidence? Set it somewhat low, keep it constant!

Parameter-Less, "Self-Tuning" Option

A somewhat bold attempt

Ask nothing from the user

(except of course the file name of the dataset).

- Confidence? Set it somewhat low, keep it constant!
- Support?

Parameter-Less, "Self-Tuning" Option

A somewhat bold attempt

Ask nothing from the user

(except of course the file name of the dataset).

- Confidence? Set it somewhat low, keep it constant!
- Support? Set it low and keep it constant as far as you can afford it, then increase it if needed.

Parameter-Less, "Self-Tuning" Option

A somewhat bold attempt

Ask nothing from the user

(except of course the file name of the dataset).

- Confidence? Set it somewhat low, keep it constant!
- Support? Set it low and keep it constant as far as you can afford it, then increase it if needed.
- Confidence boost?

Parameter-Less, "Self-Tuning" Option

A somewhat bold attempt

Ask nothing from the user

(except of course the file name of the dataset).

- Confidence? Set it somewhat low, keep it constant!
- Support? Set it low and keep it constant as far as you can afford it, then increase it if needed.
- Confidence boost? Set it high and keep it constant as long as you keep finding rules, then decrease it if needed.
  - Connection to lift for certain syntactic form of (usually quite abundant) rules allows one to monitor the rules and trigger the weakening of the confidence boost threshold.
- Python open source implementation at `yacaree.sf.net`

KNIME

A tool, a community, and a company

KNIME.com AG, Zürich

A company centered around the open source Data Mining tool KNIME (very brief demo).

- Spinoff of University of Konstanz (strong relation kept),
- tool evolved from earlier "KoNstanz Information MinEr",
- but fully redesigned by professional software engineers,
- written in Java (Eclipse plugin),
- fast-growing user (and contributor) base, (KNIMEtech Lab, `tech.knime.org`)
- revenues from
  - license with tech support and scheduled releases,
  - training,
  - consulting,
  - development of customized solutions.

Key moments

- The CEO of KNIME participated in ECML PKDD 2010 and presented KNIME in the Industrial Day: exciting evening conversation.
- A conversation right here in this room with Nicola.
- Harvest submission...
- Harvest acceptance!...
- Javier de la Dehesa stays at KNIME for the summer...
- (manages to spend a fraction of the budgeted amount...)
- The *yacaree* node works...!
- Harvest presentations at WAPA 2011...
- but...

But:

- The new node was more difficult to program in Java than I anticipated,
- has been evaluated as "difficult to use" by our colleagues,
- it is not particularly slow, but definitely slower than we anticipated,
- occassionally the results are debatable.
- I have chosen not to make it available yet to the KNIMEtech Lab.
- In the meantime, a new algorithm has been designed, able to traverse the closure lattice (key slow operation) considerably faster

  (submitted to a major conference).

STARK

Will keep working with KNIME in 2012... and hopefully beyond!

Continuation

As we had some left-over funds,

- the Harvest Programme manager has granted a temporal extension with no additional funding,
- Diego García-Sáiz (UC) will visit KNIME along 2012
  - to replace the closure lattice traversal algorithm and
  - to work out an usability improvement;

STARK

Will keep working with KNIME in 2012... and hopefully beyond!

Continuation

As we had some left-over funds,

- the Harvest Programme manager has granted a temporal extension with no additional funding,
- Diego García-Sáiz (UC) will visit KNIME along 2012
  - to replace the closure lattice traversal algorithm and
  - to work out an usability improvement;

In the meantime, KNIME is applying, as coordinators, to FET Open, SME-high-tech track, for a STREP with two universities.

One of the teams will be our group at UPC.

The short proposal already passed; we are working on the full proposal due very soon.