21st INTERNATIONAL CONFERENCE ON COMPUTATIONAL STATISTICS
AUGUST 19-22, 2014 – Geneva, Switzerland

An association rule miner for unbalanced data based on artificial bee colony optimization

Goals and Methodology

Emmanuel Rousseaux

Institute of Service Science

Emmanuel.Rousseaux@unige.ch

Gilbert Ritschard

Institute for Demographic and Life Course Studies

Gilbert.Ritschard@unige.ch

University of Geneva
Switzerland

LIVES
Swiss National Centre of Competence in Research

Outline

Introduction

Association rule mining

Artificial bee colony optimization

AR mining with ABCO

What's next

Motivation

NCCR LIVES

"Overcoming vulnerability: life course perspectives"

Methodological team

"Measuring life sequences and the disorder of lives" ruled by Gilbert Ritschard

My contribution

Data mining approaches for the discovery of critical events in life courses

Temporal association rules

We want to discover rules: A $\Rightarrow$ B

- If I experience A, then I often experience B
- If I'm older than 50 with a low educational level and I lose my job, then I may fall in long-term unemployment.
- $A^{t_1} \Rightarrow B^{t_2}$

Temporal association rules

We want to discover rules: A $\Rightarrow$ B

- If I experience A, then I often experience B
- If I'm older than 50 with a low educational level and I lose my job, then I may fall in long-term unemployment.
- $A^{t_1} \Rightarrow B^{t_2}$

Exclusion rules

Then we may want to find exclusions to these rules

- $A^{t_1} \wedge Z^{t_2} \Rightarrow \bar{B}^{t_3}$
- If I experience A but I experience Z too; I won't experience B

And look on the whole life course

- $A^{t_1}_{work\ traj.} \wedge B^{t_2}_{family\ traj.} \Rightarrow C^{t_3}_{health\ traj.}$

LIVES
Swiss National Centre of Competence in Research

Exclusion rules

Then we may want to find exclusions to these rules

- $A^{t_1} \wedge Z^{t_2} \Rightarrow \bar{B}^{t_3}$
- If I experience A but I experience Z too; I won't experience B

And look on the whole life course

- $A^{t_1}_{work\ \ traj.} \wedge B^{t_2}_{family\ \ traj.} \Rightarrow C^{t_3}_{health\ \ traj.}$

Exclusion rules

Then we may want to find exclusions to these rules

- $A^{t_1} \wedge Z^{t_2} \Rightarrow \bar{B}^{t_3}$
- If I experience A but I experience Z too; I won't experience B

And look on the whole life course

- $A^{t_1}_{work\ \ traj.} \wedge B^{t_2}_{family\ \ traj.} \Rightarrow C^{t_3}_{health\ \ traj.}$

Outline

Introduction

Association rule mining

Artificial bee colony optimization

AR mining with ABCO

What's next

Association rule: Definition

Let:

- $I = \{I_1, I_2, \ldots, I_p\}$ a set of binary attributes
- T a database of n observations on I.

An association rule is given by (Agrawal, Imieliński, and Swami, 1993)

- $X \subset I,\ Y \subset I,\ X \cap Y = \varnothing$
- A direction: $X \Rightarrow Y$

Association rule: Quality assessment

More than 50 criterias...

- Support: $\text{supp}(A \Rightarrow B) = n_{A \cup B}/n$
- Confidence: $\text{conf}(A \Rightarrow B) = n_{A \cup B}/n_A \approx P(B|A)$
- Lift: $\text{lift}(A \Rightarrow B) = \text{supp}(A \Rightarrow B)/(\text{supp}(A) * \text{supp}(B)) \approx P(B|A)/P(B)$
- Statistical criterias: chi-squared, implicative intensity, etc.
- ...

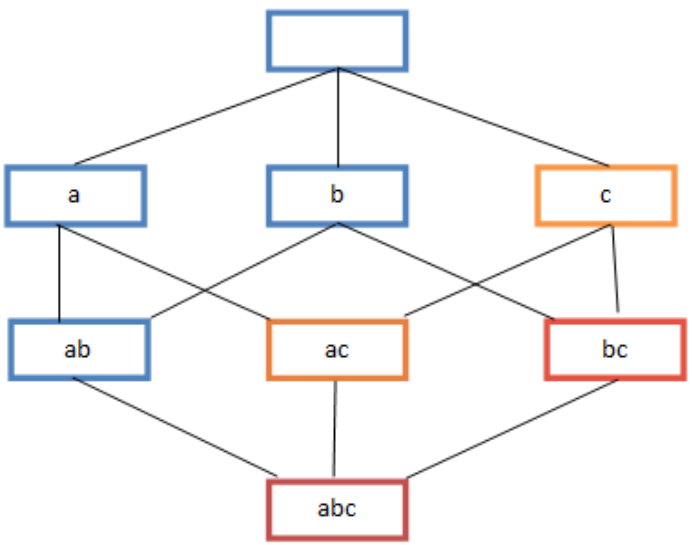
Application to life-course mining

Specific features

- Temporality (Srikant and Agrawal, 1996; Harms and Deogun, 2004)
  - Gap threshold between events
  - Ordering
- Positive and negative rules (Swesi, Bakar, and Kadir, 2012)

Frequent pattern mining

Exp(p) search space



- Limit exploration: support threshold
- Classical algorithms:
  - Apriori (Agrawal, Imieliński, and Swami, 1993)
  - FP-growth (Han, Pei, and Yin, 2000)
- Extract too many rules, most of them are useless
- Impossible to discover rare rules

Mining rare classes

Recent approaches

- Multiple support thresholds (Liu, Hsu, and Ma, 1999)
- Particle swarm optimization (Sarath and Ravi, 2013)
- Genetic algorithm (Ghosh and Nath, 2004; Salleb-Aouissi, Vrain, and Nortet, 2007)
- **Proposal:** Use an artificial bee colony algorithm

Mining rare classes

Recent approaches

- Multiple support thresholds (Liu, Hsu, and Ma, 1999)
- Particle swarm optimization (Sarath and Ravi, 2013)
- Genetic algorithm (Ghosh and Nath, 2004; Salleb-Aouissi, Vrain, and Nortet, 2007)
- **Proposal:** Use an artificial bee colony algorithm

Outline

Introduction

Association rule mining

Artificial bee colony optimization

AR mining with ABCO

What's next

Artificial bee colony algorithm

Recent Evolutionary Population-based stochastic optimization algorithm

(Karaboga and Basturk, 2007; Karaboga and Basturk, 2008)

- Based on the foraging ability of bees
- Originally design for optimization in $\mathbb{R}^d$
- Adapted for different problems (scheduling, feature selection, non-linear equation, clustering, classification)

(Karaboga, Gorkemli, et al., 2014)

Artificial bee colony algorithm

Recent Evolutionary Population-based stochastic optimization algorithm

(Karaboga and Basturk, 2007; Karaboga and Basturk, 2008)

- Based on the foraging ability of bees
- Originally design for optimization in $\mathbb{R}^d$
- Adapted for different problems (scheduling, feature selection, non-linear equation, clustering, classification)

  (Karaboga, Gorkemli, et al., 2014)

Example

Cost function: $f(x) = x^2$

Search space: $\mathbb{R}$

$$\text{fitness}(x) = \frac{1}{1 + f(x)}$$

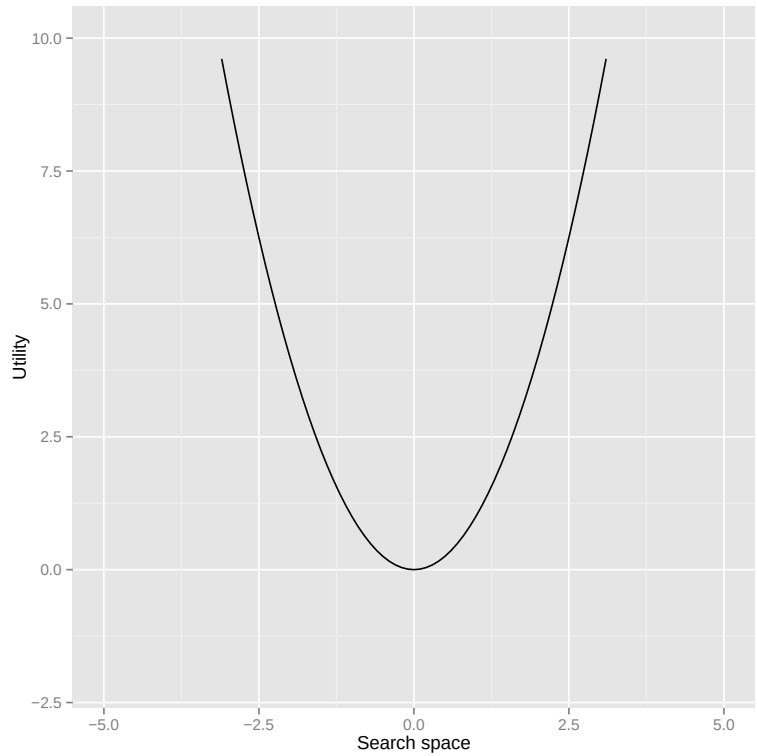
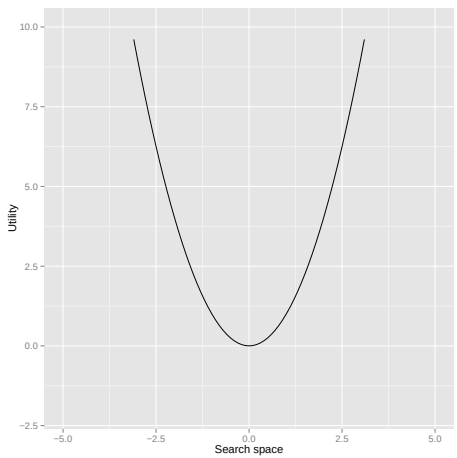


Figure : *A (too much) easy function to minimize*

Methodology

Parameters

- Let $S = (s_1, c_1; \ldots; s_N, c_N)$ N food sources
- N employed bees and N onlooker bees
- L limit to the number of try per source: $c_i < L$



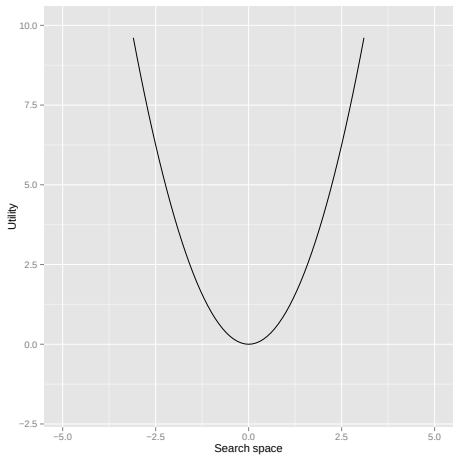
Initialization

- Each source is randomly initialized
- Counter c_i are all equal to 0

Methodology

Employed bee phase

- Update solutions using Eq. 1
- Calculate fitness values of new solutions
- Keep a new solution when better, increment its counter otherwise

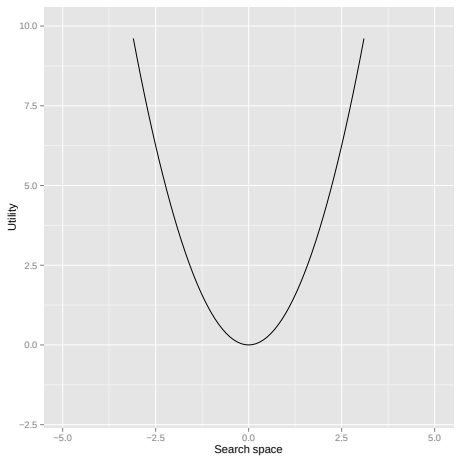


$$v_i = x_i + r_i(x_i - x_k), i \neq k \qquad (1)$$

Methodology

Onlooker bee phase

- Calculate selection probability by using Eq. 2
- Select an employed bee and update its solution by using Eq. 1
- Keep new solutions when better, increment counter otherwise

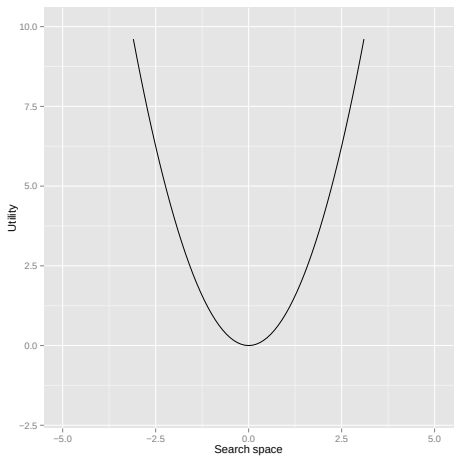


$$p_i = \frac{\text{fitness}(x_i)}{\sum_{k=1}^{N} \text{fitness}(x_k)} \qquad (2)$$

Methodology

Scout bee phase

- Select an abandonment counter having the highest value
- If higher than L, generate a new source for the employed bee by using Eq. 1



Outline

Introduction

Association rule mining

Artificial bee colony optimization

AR mining with ABCO

What's next

Data transformation

Let T be a database with p attributes

- Binary variable: do nothing
- Categorical variable: one binary variable per class
- Quantitative variable: discretization, then binarization

Result: each individual is a pattern in $\{0,1\}^{N_p}$.

Coding of a rule

Let: $A = 0010$ and $B = 0101$, then

- $A \cup B$ =0111

Then a rule can be coded:

- $A \Rightarrow B = A \cup B +$ is.conclusion $= 01110101$

This ensure $A \cap B = \varnothing$

Binary optimization with ABC

Initialization: For N sources

- We have to generate N patterns in $\{0,1\}^{2N_p}$.
- $S_i = S_{i1}S_{i1}S_{2N_p}$
- We can use a Bernouilli process:

$$S_{ij} \textbf{ realization of } X_{ij} \sim \textbf{Bernouilli}(p_0)$$

Binary optimization with ABCO

Very long sequences $\Rightarrow$ Need a low complexity candidate generation method

For each dimension (Kiran and Gunduz, 2013)

$v_i = x_i + r_i(x_i - x_k)$ becomes $V_i = X_i \oplus [R_i \otimes (X_i \oplus X_k)], i \neq k$

with

- $\otimes$ = AND
- $\oplus$ = XOR
- R_i a realization of the logic NOT gate with 50% probability

Fitness function

The fitness function has to be fast to compute

We use:

- A support filter SF_{θ_0}: 1 if support$(A \Rightarrow B)$ >= θ_0, 0 otherwise
- lift$(A \Rightarrow B)$
- conviction$(A \Rightarrow B) = 1/\text{lift}(A \Rightarrow \bar{B})$

$$\text{fitness}(A \Rightarrow B) = (SF_{\theta_0} \ . \ \text{lift} \ . \ \text{conviction})(A \Rightarrow B)$$

Rule generation

Generation of a set of M rules

- Each run generate a single rule: the best solution of the run
- We iterate enough to get M different rules
- Ensemble scheme: repeat the generation ℓ times and keep the M most frequents

Outline

Introduction

Association rule mining

Artificial bee colony optimization

AR mining with ABCO

What's next

What's next

- Better handling of
  - Time
  - Negative items
  - Quantitative covariates
- Development in process
- Available in R, surely bundled in a package
- `http://emmanuel.rousseaux.me/`
- Experimental assessment
- Visualization

LIVES
Swiss National Centre of Competence in Research

Selected references

Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami (1993). "Mining association rules between sets of items in large databases". In: *ACM SIGMOD Record*. Vol. 22. 2. ACM, pp. 207–216.

Ghosh, Ashish and Bhabesh Nath (2004). "Multi-objective rule mining using genetic algorithms". In: *Information Sciences* 163.1, pp. 123–133.

Han, Jiawei, Jian Pei, and Yiwen Yin (2000). "Mining frequent patterns without candidate generation". In: *ACM SIGMOD Record*. Vol. 29. 2. ACM, pp. 1–12.

Harms, Sherri K and Jitender S Deogun (2004). "Sequential association rule mining with time lags". In: *Journal of Intelligent Information Systems* 22.1, pp. 7–22.

Karaboga, Dervis and Bahriye Basturk (2007). "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm". In: *Journal of global optimization* 39.3, pp. 459–471.

— (2008). "On the performance of artificial bee colony (ABC) algorithm". In: *Applied soft computing* 8.1, pp. 687–697.

Karaboga, Dervis, Beyza Gorkemli, et al. (2014). "A comprehensive survey: artificial bee colony (ABC) algorithm and applications". In: *Artificial Intelligence Review* 42.1, pp. 21–57.

LIVES
Swiss National Centre of Competence in Research

Kiran, Mustafa Servet and Mesut Gunduz (2013). "XOR-based artificial bee colony algorithm for binary optimization". In: *Turkish Journal of Electrical Engineering & Computer Sciences* 21.Sup. 2, pp. 2307–2328.

Liu, Bing, Wynne Hsu, and Yiming Ma (1999). "Mining association rules with multiple minimum supports". In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 337–341.

Salleb-Aouissi, Ansaf, Christel Vrain, and Cyril Nortet (2007). "QuantMiner: A Genetic Algorithm for Mining Quantitative Association Rules." In: *IJCAI*. Vol. 7.

Sarath, KNVD and Vadlamani Ravi (2013). "Association rule mining using binary particle swarm optimization". In: *Engineering Applications of Artificial Intelligence* 26.8, pp. 1832–1840.

Srikant, Ramakrishnan and Rakesh Agrawal (1996). *Mining sequential patterns: Generalizations and performance improvements*. Springer.

Swesi, Idheba Mohamad Ali O, Azuraliza Abu Bakar, and Anis Suhailis Abdul Kadir (2012). "Mining positive and Negative Association Rules from interesting frequent and infrequent itemsets". In: *Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on*. IEEE, pp. 650–655.

Thank you for your attention

Questions/remarks?