Detecting unexpected patterns by Clustering

Danh Bui, UAntwerpen

biomina

ADReM

University of Antwerp

Outline

- Preliminaries
- What is unexpected pattern?
- How to mine it?
- Experimental result

Preliminaries

- Patterns: frequent item-sets, association rules.
- Frequent item-set: a set of items (variables, events, properties…) that frequently co-occur.
  - Breast cancer:
    - $age = 50 - 59, node - caps = no, irradiat = no$ (22.08%)
  - TCR data
    - ASS, CAS, YEQ (34.23%)
- Association rule: $X \rightarrow Y$ (If X occurs then Y occurs)
  - Breast cancer:
    - $age = 50 - 59, node - caps = no \rightarrow irradiat = no$ (100%)
    - $age = 50 - 59, node - caps = no, irradiat = no \rightarrow reccurence = no$ (83.01%)

Preliminaries

- Given a transaction dataset D; item-sets X, Y; a rule $X \rightarrow Y$
  - $Support(X)$: the percentage of transactions in D which contain X
  - $Confidence(X \rightarrow Y)$: the percentage of transactions containing X which also contain Y

- Given a database D, a minimum support and a minimum confidence threshold
  - Association rule mining (ARM) finds all rules whose supports and confidences are larger than given thresholds
  - Ranking the rules based on interestingness measures.

Preliminaries

- Different interestingness measures based on different criteria:
  - Conciseness
  - Reliability
  - Novelty
  - **Unexpectedness**
  - Utility
  - …

Preliminaries

- Interesting patterns are used in:
  - Recommendation system
  - Intrusion detection
  - Bioinformatics
  - Web usage mining
  - …

What is unexpected pattern?

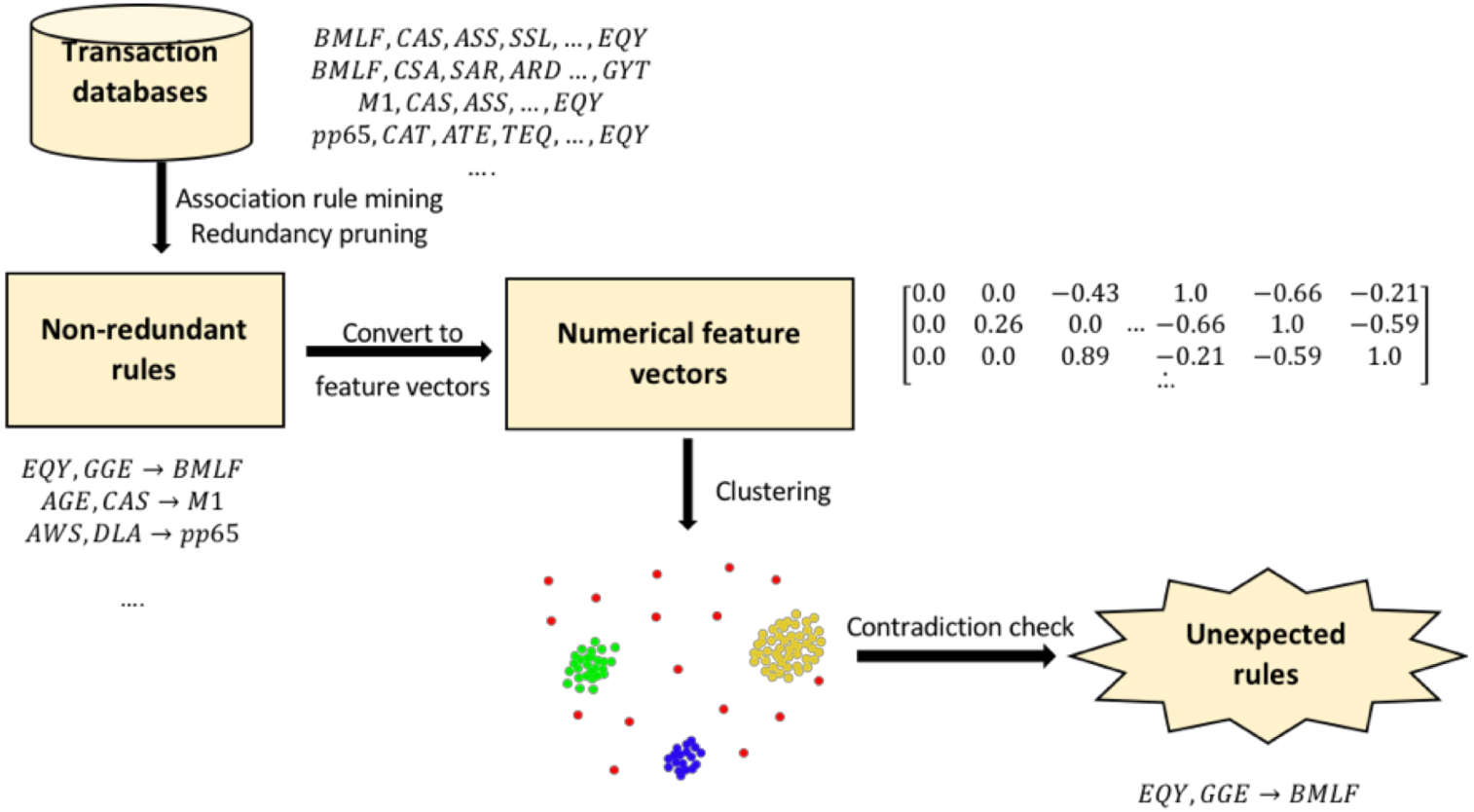
- **Unexpected patterns** are the patterns that allow us to identify a failing in prior knowledge (beliefs).
  - It may suggest an aspect of the data that needs further investigation
- Example: Breast cancer data
  - Belief:
    - $menopause = ge40, node\_caps = no, irradiat = no \rightarrow recurrence = no$
  - Unexpected pattern
    - $menopause = ge40, inv\_nodes = 3 - 5, node\_caps = no, irradiat = no \rightarrow recurrence = yes$

How to mine it?

- Belief – driven methods
  - Build a belief system (prior knowledge)
  - Using a comparison function to see whether a pattern violates the belief system.

How to mine it?

- A belief – driven method based on pattern clustering



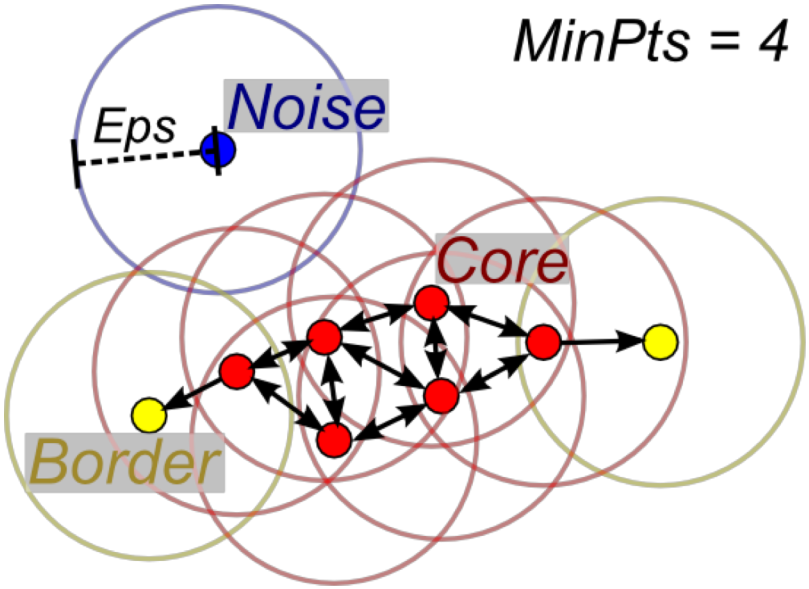
Contradiction Check

- Given a belief $X' \rightarrow Y'$, a pattern $X \rightarrow Y$ is unexpected if
  - Y and Y' logically contradict each other
  - X and X' co-occur in a substantial subset of transactions in D
  - X and X' are similar to each other

- We used correlation coefficient to judge contradiction between items.

Clustering

- DBSCAN algorithm for pattern clustering
- It depends on 2 parameters: Eps and $MinPts$
- How to choose parameters?
  - Fix $MinPts$
  - Reduce Eps until there's no contradiction between clusters
  - Relaxation: there's no contradiction between large clusters.



Experimental result

- Run experiments on imbalanced datasets

Dataset	train	test	#classes	#attributes	Minority class
Adult (UCI)	32561	16281	2	14	0.241
Breast cancer (UCI)	240	46	2	9	0.296
Credit approval (UCI)	598	92	2	15	0.446
Smtp (KDDCup99)	96554	8268	2	9	0.0122
Pima (UCI)	630	138	2	6	0.346
TCR (Dash et al.)	333	79	3	822	0.156

Experimental result

- Evaluate the quality of unexpected patterns based on their contribution to performance of the classifiers, SVM and Random Forest.

Dataset	SVM	SVM+UnexpRules	RF	RF+UnexpRules
Adult	0.251	**0.333**	0.626	**0.630**
Breast cancer	0.0	**0.696**	0.636	**0.72**
Credit	0.905	**0.916**	0.886	0.886
Pima	0.533	**0.575**	0.659	**0.694**
Smtp	0.021	**0.625**	0.021	**0.625**
TCR	0.0	**0.133**	0.72	**0.769**

Table F1 score on minority class of the classifiers, SVM and Random Forest (RF) with and without using UnexpRules on testing data.

Some unexpected patterns

- Breast cancer data

	Rules	**conf**	**sup**	**Rank**
Belief	$menopause = ge40$, $node-caps = no$, $irradiat = no \rightarrow recurrence = no$	0.823	0.233	
UnexpRule	$menopause = ge40$, $\mathbf{inv-nodes = 3..5}$, $node-caps = no$, $irradiat = no \rightarrow recurrence = yes$	1.0	0.0167	0.878
Reference	$menopause = ge40, inv-nodes = 3-5, node-caps = no, irradiat = no$		0.0167	
Belief	$menopause = ge40$, $node-caps = no$, $\mathbf{breast{=}left} \rightarrow recurrence = no$	0.833	0.146	
UnexpRule	$menopause = ge40$, $\mathbf{inv-nodes = 3..5}$, $node-caps = no$, $\mathbf{breast-quad = left-low} \rightarrow recurrence = yes$	1.0	0.0167	0.768
Reference	$menopause = ge40$, $inv-nodes = 3..5$, $node-caps = no$, $breast = left$, $breast-quad = left-low$		0.0125	

Table Unexpected rule examples of the Breast cancer dataset. *Reference* is the LHS co-occurrence of belief and unexpected rule in data.

Some unexpected patterns

- TCR dataset

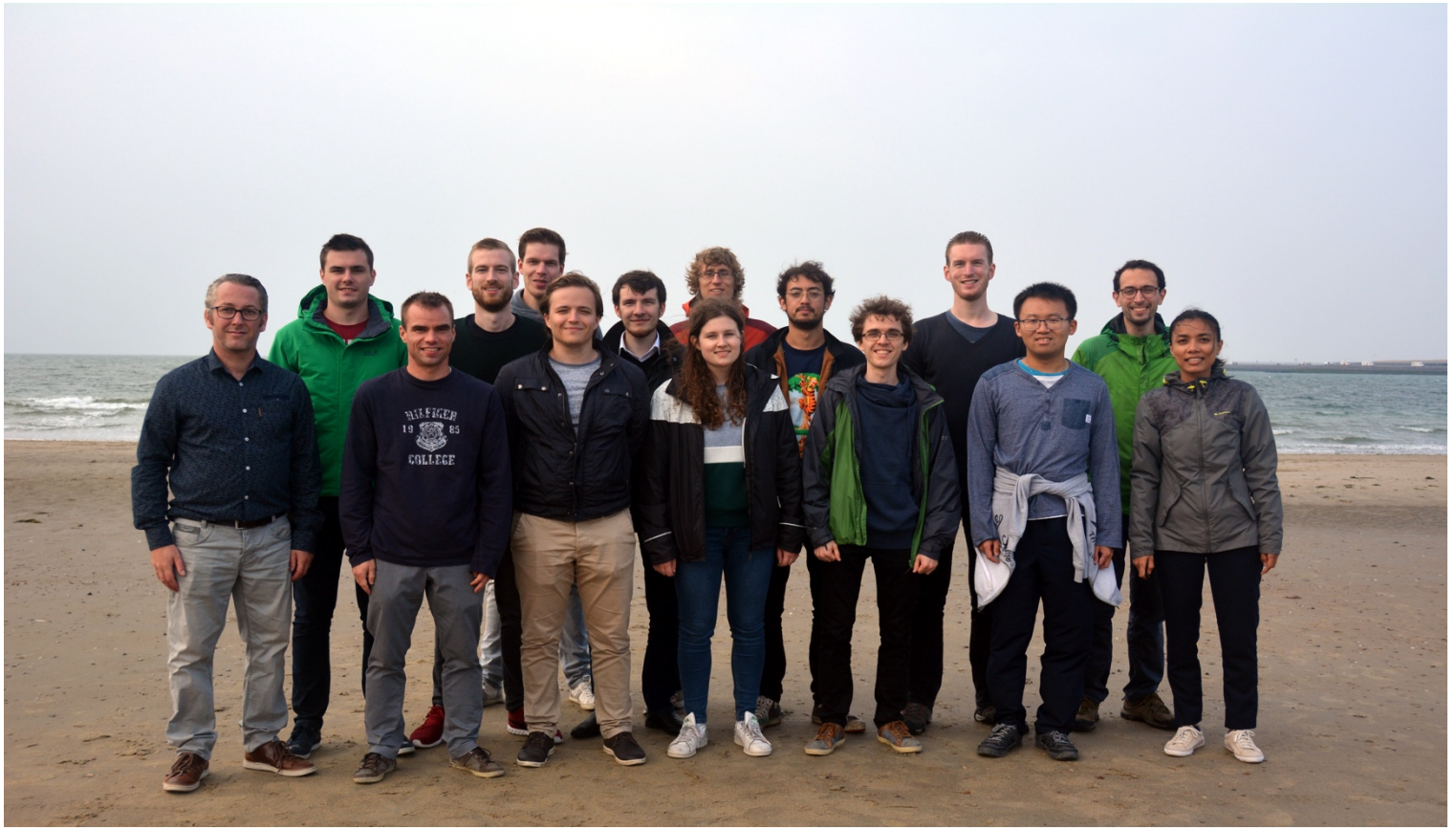
	Rules	conf(%)	supp(%)	Rank
Belief	$EQY \rightarrow antigen = M1$	0.894	0.354	
UnexpRule	EQY, **GGE** $\rightarrow antigen = BMLF$	1.0	0.012	0.812
Reference	EQY, GGE		0.012	
Belief	ASS ,**EQY** $\rightarrow antigen = M1$	0.911	0.339	
UnexpRule	ASS, **GGE** $\rightarrow antigen = BMLF$	1.0	0.012	0.735
Reference	ASS ,EQY, GGE		0.012	
Belief	**EQY** $\rightarrow antigen = M1$	0.894	0.354	
UnexpRule	**GGE** $\rightarrow antigen = BMLF$	0.8	0.012	0.0264
Reference	EQY, GGE		0.012	

Table Unexpected rule examples of the TCR dataset. *Reference* is the LHS co-occurrence of belief and unexpected rule in data.

Conclusion

- Mine unexpected patterns by a belief-driven method
  - Belief system are derived automatically from data by clustering
  - Outliers are candidates for unexpected patterns
- Unexpected patterns have great potential for imbalance datasets
  - Appear in many real-world applications: fault detection, fraud detection, medical diagnosis

Biomina team



http://www.biomina.be/

References

- B. Padmanabhan and A. Tuzhilin, *A Belief-Driven Method for Discovering Unexpected Patterns*, KDD (1998), pp. 94-100.
- M. Y. Chang, R. D. Chiang, S. J. Wu and C. H. Chan, *Mining unexpected patterns using decision trees and interestingness measures: a case study of endometriosis*, Soft Computing, vol.20, issue.10 (2016), pp. 3991-4003.
- P. Dash, A. J. Fiore-Gartland, T. Hertz, G. C. Wang, S. Sharma, A. Souquette, J. C Crawford, E. B Clemens, T. H. O. Nguyen, K. Kedzierska, N. L. La Gruta, P. Bradley and P. G. Thomas, *Quantiable predictive features dene epitope-specic T cell receptor repertoires*, Nature 547 (2017), pp. 89-93.
- R. Agrawal and R. Srikant R, *Fast algorithms for mining association rules*, VLDB, vol.1215 (1994), pp. 487-499.
- UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/index.php