Grant 2014/08996-0,
2015/21574-0

Post-Processing Association Rules: A Network based Label Propagation Approach

Renan de Padua
Veronica Oliveira de Carvalho
Solange Oliveira Rezende

Univ Estadual Paulista
Brazil, São Paulo

Univ de São Paulo
Brazil, São Paulo

Introduction

- Association rules are widely used to explore relations among items on a data set
- However, a great amount of rules is generated
  - Makes the manual exploration for interesting patterns infeasible
- Many researches try to direct the users on the exploration, helping them to find the interesting rules

Introduction

- Some works propose the use of networks
  - They are used as a mean to model and prune the rules that are not interesting to the user

- Problem: those works require previous information about what the user considers interesting (objective item), forcing him to have prior a knowledge on the data set

Objective

- To propose an approach that helps the user to find the most relevant rules in a way the user does not need to have a prior knowledge on the data set

- For that, the approach suggests some rules to be classified by him based on the rule's relevance in the network

Contribution

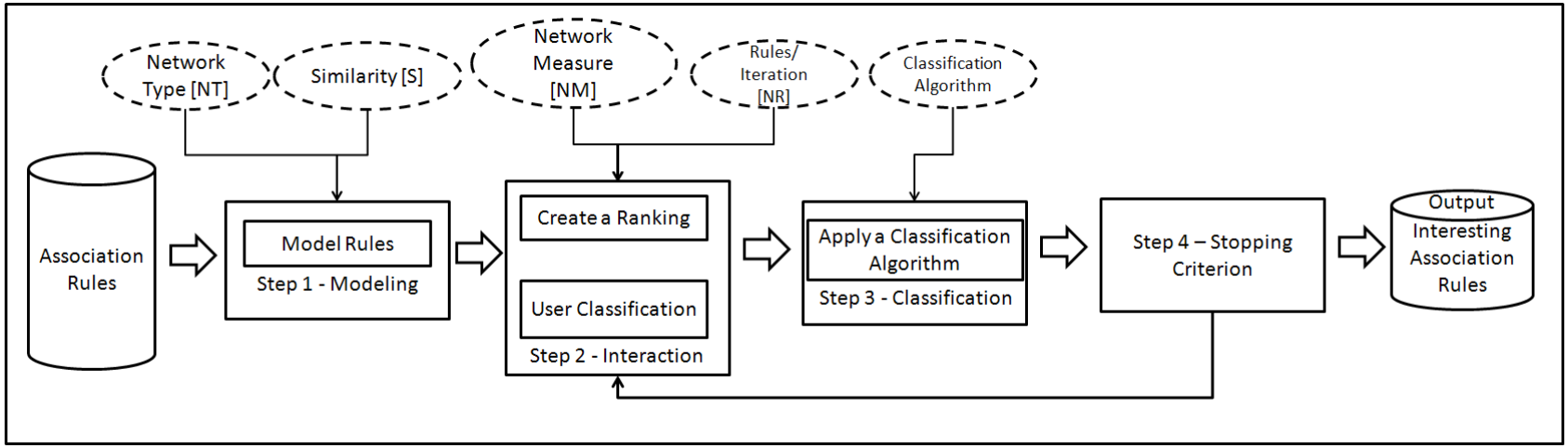
- It is not necessary to have a prior knowledge on the data set (objective item), making the classification (I, NI) easier

Post-Processing Association Rules using Label Propagation

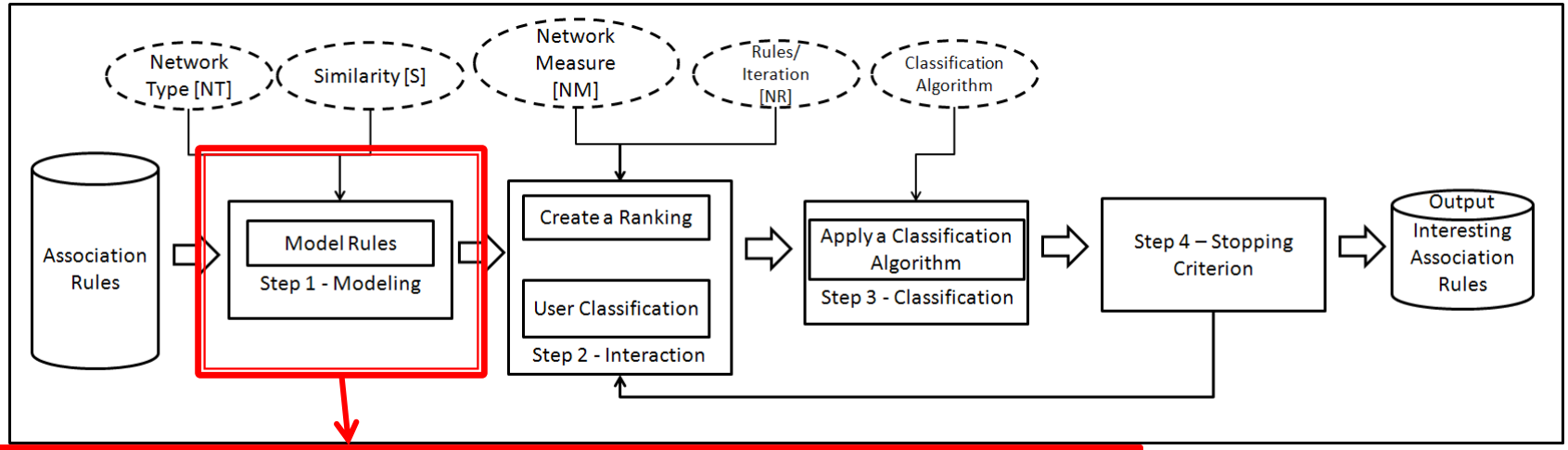
- The Post-Processing Association Rules using Label Propagation (PAR_{LP}) approach is based on the idea of using classification algorithms to post-processes association rules

- The classification algorithms allow the approach to learning from the previous iterations, reinforcing the user's knowledge on each new iteration

Post-Processing Association Rules using Label Propagation



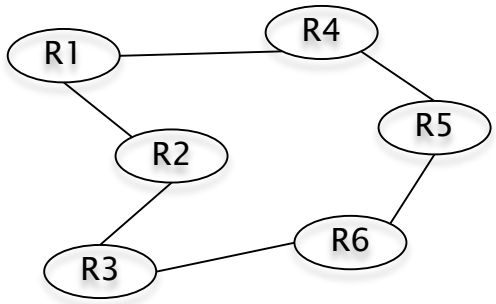
Post-Processing Association Rules using Label Propagation



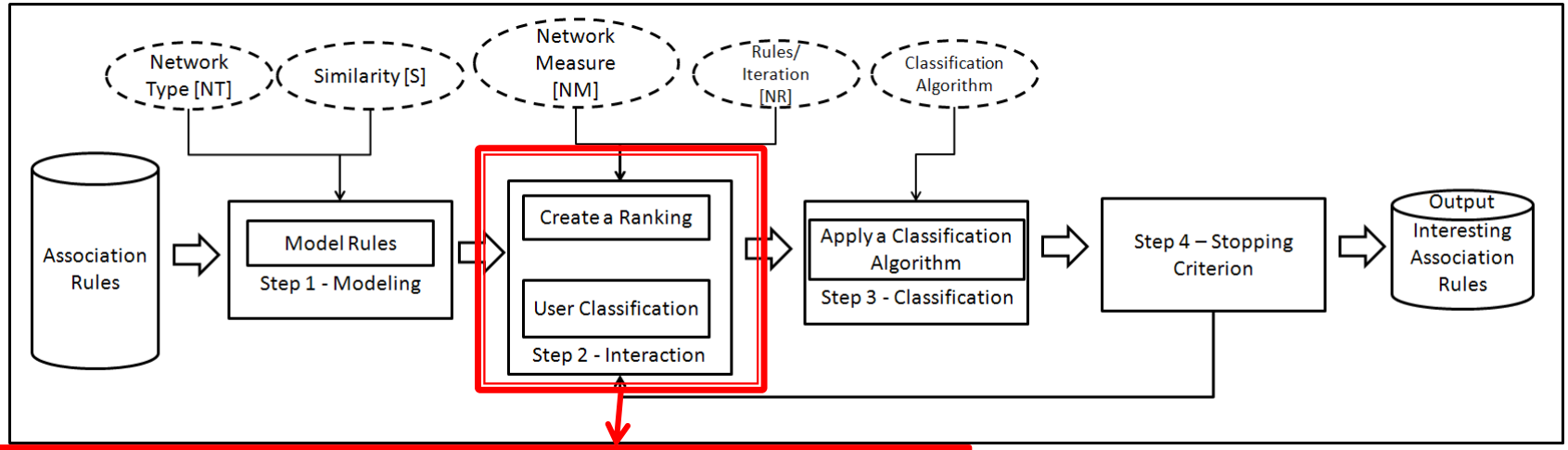
Models the rules on a network

Network Type: defines the network type (homogeneous, heterogeneous, etc.) and its configuration (Knn, Gaussian, etc.)

Similarity: measure that will be used as the weight between two vertex



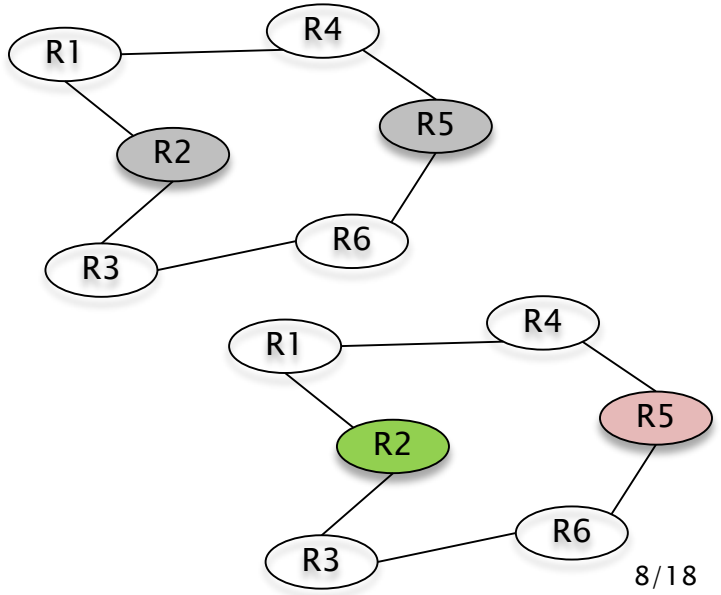
Post-Processing Association Rules using Label Propagation



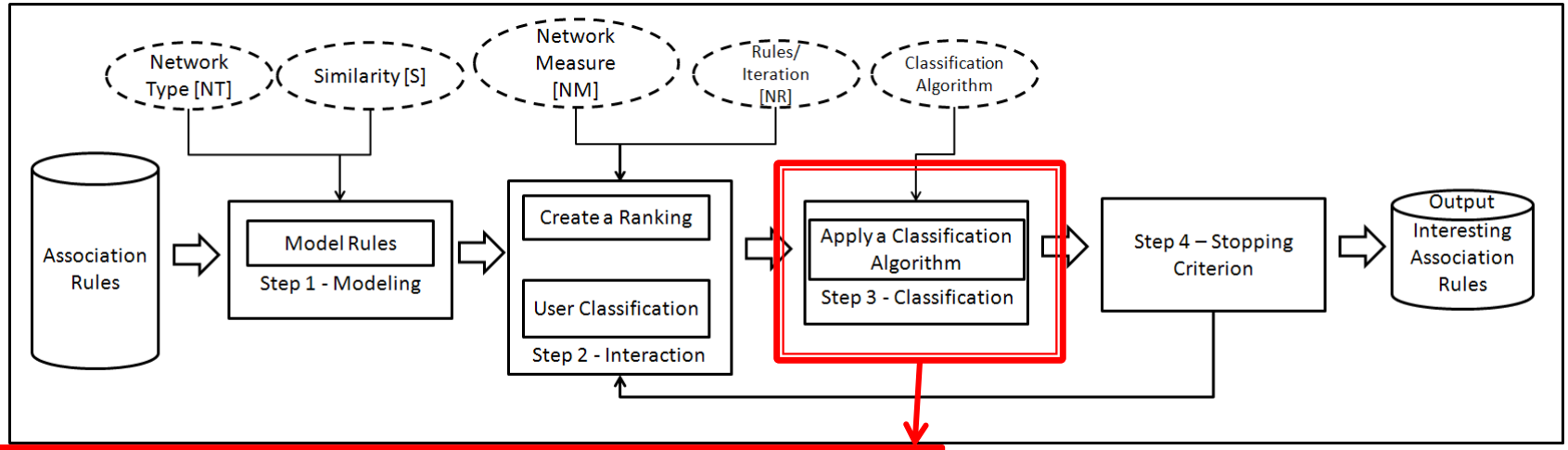
Approach interacts with the user to capture the user's knowledge, directing him to the rules he considers “Interesting”

Network Measure: measure used to create the rules’ ranking

Rules/Iteration: defines the number NR of rules to be analyzed

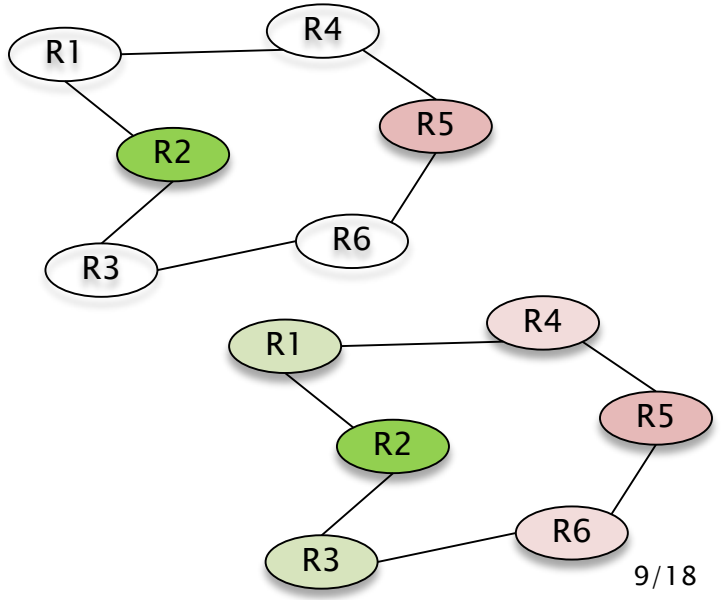


Post-Processing Association Rules using Label Propagation

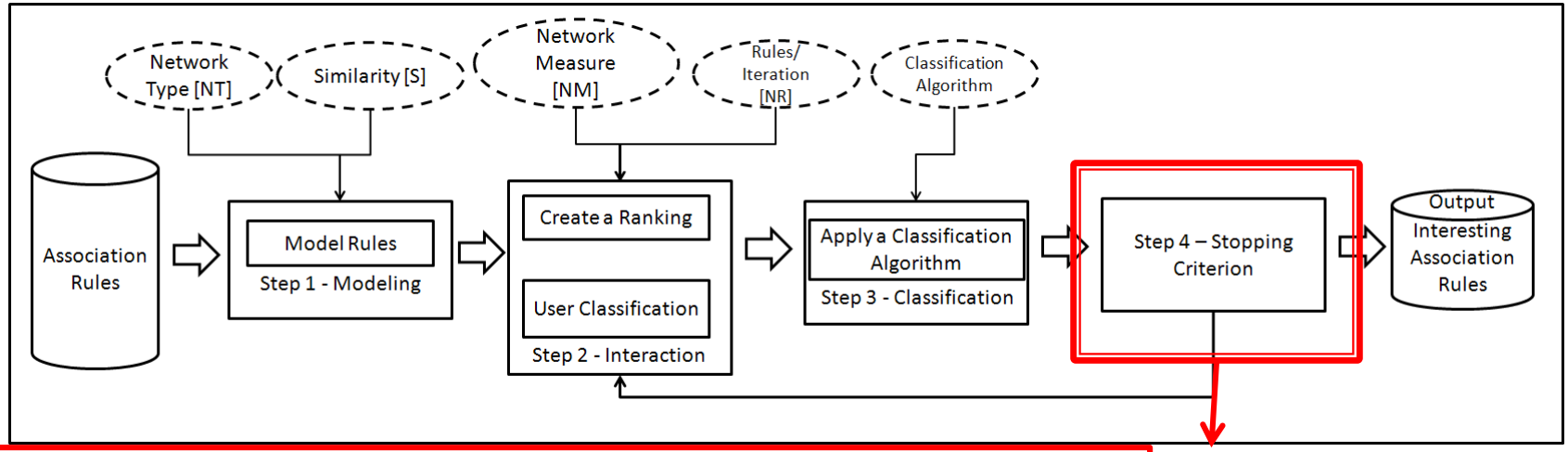


Applies a Label Propagation Algorithm considering the user's classification

Classification Algorithm: selects the algorithm to be used and its parameters in order to classify the entire network



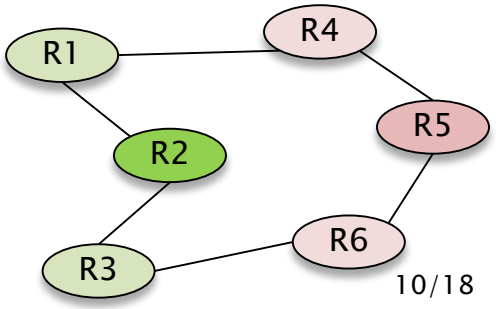
Post-Processing Association Rules using Label Propagation



Checks the obtained rules and tests if it is necessary to execute the process again

If so, the previous user's classifications are kept and considered

→ the approach will refine the nodes (classifications) considering the new knowledge that will be provided by the user



Experiments
PAR_{LP} Configurations

Network Type (NT)	NT Configuration	NR	Network Measure	Similarity	Classifier
Homogeneous	Knn (K = 10, 20, 30, 40, 50); Gaussian (α = 0.25, 0.50, 0.75); Conventional	10	Output Degree, PageRank	Jaccard, Confidence	GFHF; LLGC (α = 0.1, 0.3, 0.5, 0.7, 0.9)
Bipartite Heterogeneous	Conventional	10	Output Degree, PageRank	Jaccard, Confidence	LPBHN; GNetMine (α = 0.1, 0.3, 0.5, 0.7, 0.9)

Experiments
PAR_{LP} Configurations

- The PAR_{LP} was executed over 8 UCI data sets
- The user's interaction was simulated using a set of rules as an objective set – rules to be found on the rule set, simulating the user's interests

Experiments
PAR_{LP} Configurations

- Two different objective sets – to analyze how the approach would behave with different users
  - Random objective set: generated by randomly selecting rules until a total of 1% of the rule set size is reached
  - Similarity objective set: generated by randomly selecting one rule in the rule set and creating a similarity ranking among the selected rule and the entire rule set – 1% of the most similar are considered
- Due to the randomness, 30 objective sets were generated for each case

Experiments
PAR_{LP} Configurations

- Based on the objective sets, the user's classification is simulated considering a threshold

- In each iteration, the mean similarity among the rules to be classified and the objective set is calculated and compared to the threshold
  - if the mean similarity is greater than or equal to the threshold the rule is labeled as “Interesting”; otherwise, as “Non-Interesting”

Experiments
PAR_{LP} Configurations

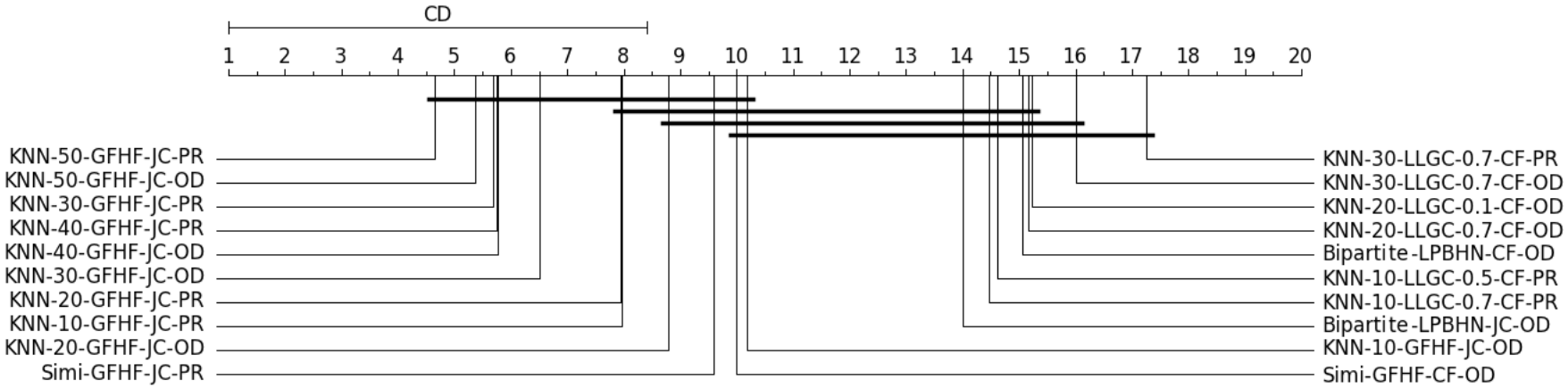
- Stopping criteria
  - The approach was executed until all the rules on an objective set were classified as “Interesting" – either by the user or by the classifier
- Validation measure
  - Number of rules the user does not need to explore to find all the interesting ones
    - To analyze the user’s effort

Results and Discussion

		Random Objective Set		Similarity Objective Set	
Data set	# Rules	Best ROS	Worst ROS	Best SOS	Worst SOS
Balance-scale	1746	40.66%	4.81%	63.69%	29.50%
Breast-cancer	1602	19.98%	5.37%	42.45%	4.56%
Car	1326	15.91%	4.68%	52.64%	22.17%
Ecoli	1685	28.66%	4.87%	51.57%	21.01%
Habermann	1006	46.12%	9.15%	58.45%	29.72%
Iris	967	51.71%	10.13%	66.49%	39.50%
Tic-tac-toe	1317	37.05%	4.02%	61.88%	16.02%
zoo	1658	30.88%	4.40%	46.38%	17.13%

It can be seen that an exploration guided by some "theme" or by some related topics will result in a higher reduction than an exploration where the user explores by selecting dissimilar rules as "Interesting"

Results and Discussion



- Friedman NxN with Nemenyi as post-test
- It is possible to see that the kNN network, together with the GFHF classifier, obtained the overall best results, being on 9 out of 10 best results

Conclusion

- The results indicate that the PAR_{LP} can become a very interesting way to post-process association rules
  - It helps the user to find the most relevant rules in an interactive way considering the user does not have a prior knowledge on the data set

Grants 2014/08996-0,
2015/21574-0

Post-Processing Association Rules: A Network based Label Propagation Approach

Renan de Padua
Veronica Oliveira de Carvalho
Solange Oliveira Rezende

Univ Estadual Paulista
Brazil, São Paulo

Univ de São Paulo
Brazil, São Paulo