UNIVERSIDAD
D
CÓRDOBA

KDIS Lab
KNOWLEDGE DISCOVERY AND INTELLIGENT SYSTEMS

Analysis of the Effectiveness of G3PARM Algorithm

J.M. Luna, J.R. Romero and S. Ventura

Knowledge Discovery and Intelligent Systems Research Group
University of Córdoba, Spain

HAIS 2010. San Sebastián, Spain. June 2010.

Outline

- Introduction
- Model description
- Experiments
- Results
- Conclusions
- Future research lines

Introduction

- Association rules mining (ARM): a Data Mining technique
- Classical algorithms for mining association rules:
  - Apriori
  - FP-Growth
- Genetic Programming: computer programs
- Grammar Guided Genetic Programming (G3P)
- Our proposal is a G3P-based approach for ARM

Model description

- G3PARM = G3P Association Rules Mining
- Designed for the extraction of association rules
- Based on the use of a context-free grammar
- Each individual represents a rule
- An auxiliary population:
  - Exceed a minimum threshold for two different measures
  - Maximum auxiliary population size

Model description

Context-free grammar

$G = (\Sigma_N, \Sigma_T, P, Rule)$ with:

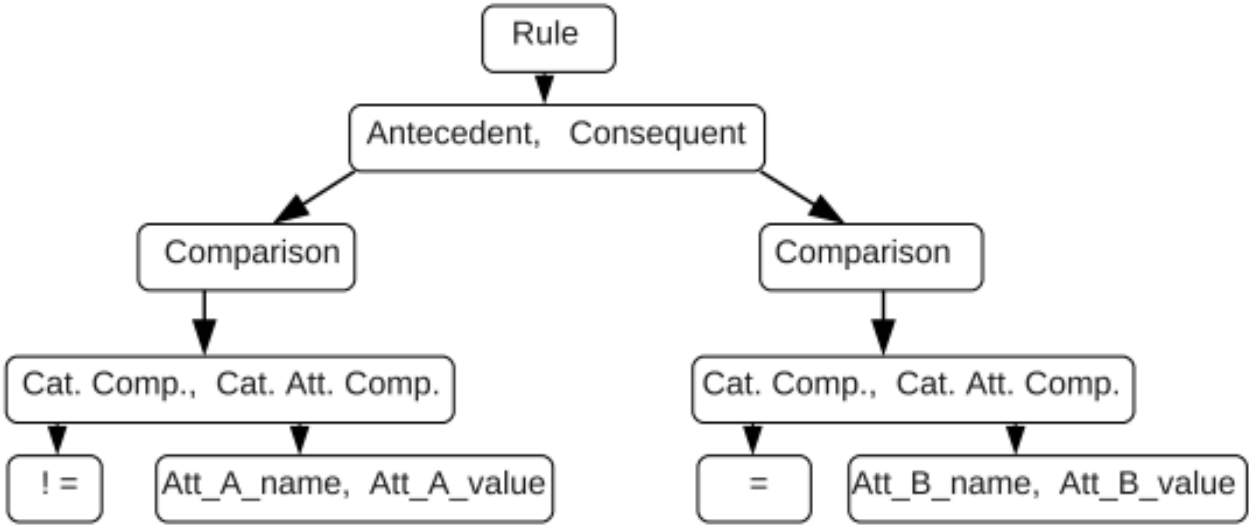
Σ_N = {Rule, Antecedent, Consequent, Comparison, Categorical Comparator, Categorical Attribute Comparison }

Σ_T = {AND, "! =", "=", "name", "value"}

P = {Rule = Antecedent, Consequent ;
Antecedent = Comparison | AND, Comparison, Antecedent ;
Consequent = Comparison ;
Comparison = Categorical Comparator, Categorical Attribute Comparison ;
Categorical Comparator = "! =" | "=" ;
Categorical Attribute Comparison = "name", "value" ;}

Model description

Syntax-tree structure

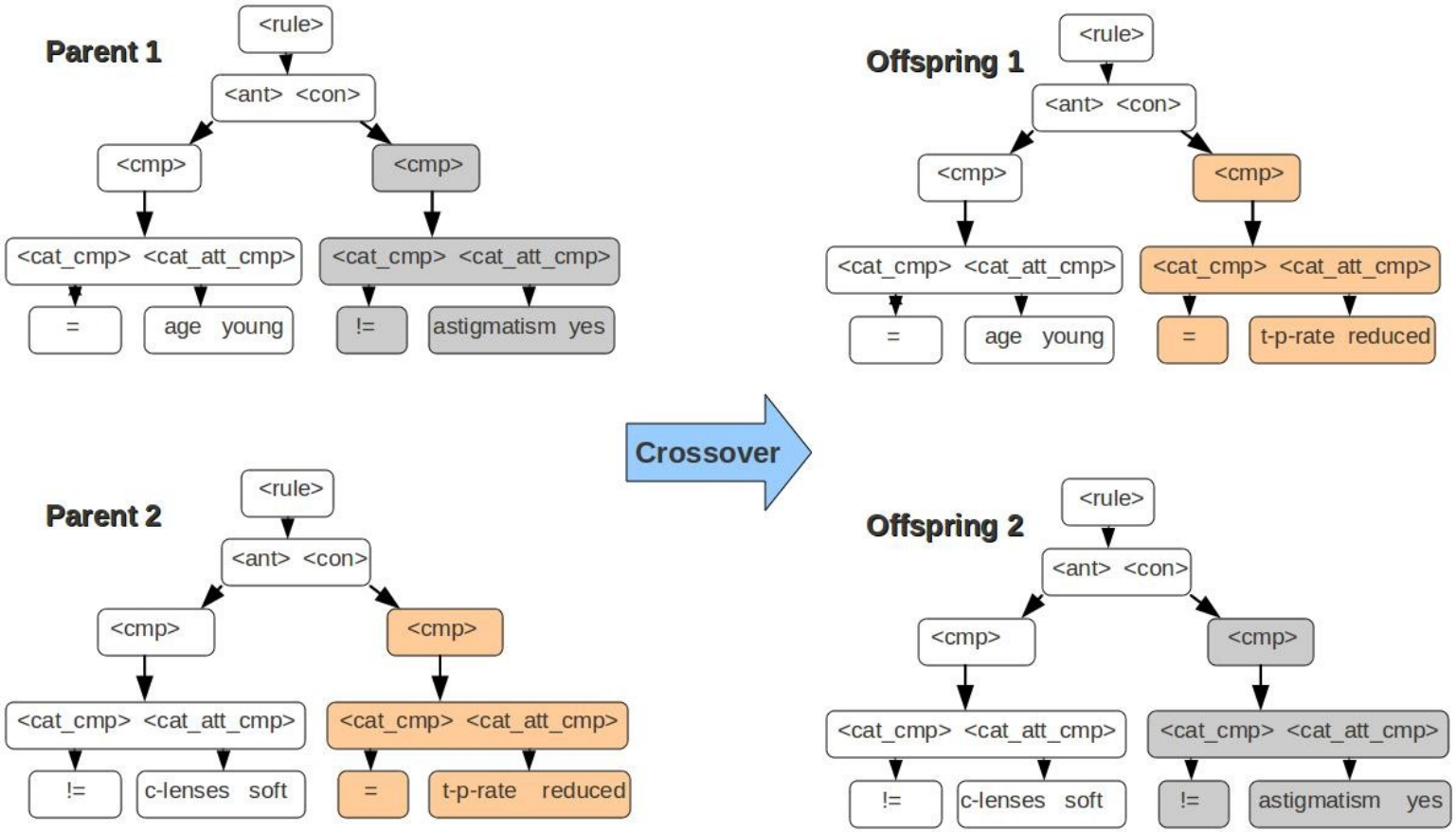


Phenotype:

(!= Att_A_name Att_A_value) → (= Att_B_name Att_B_value)

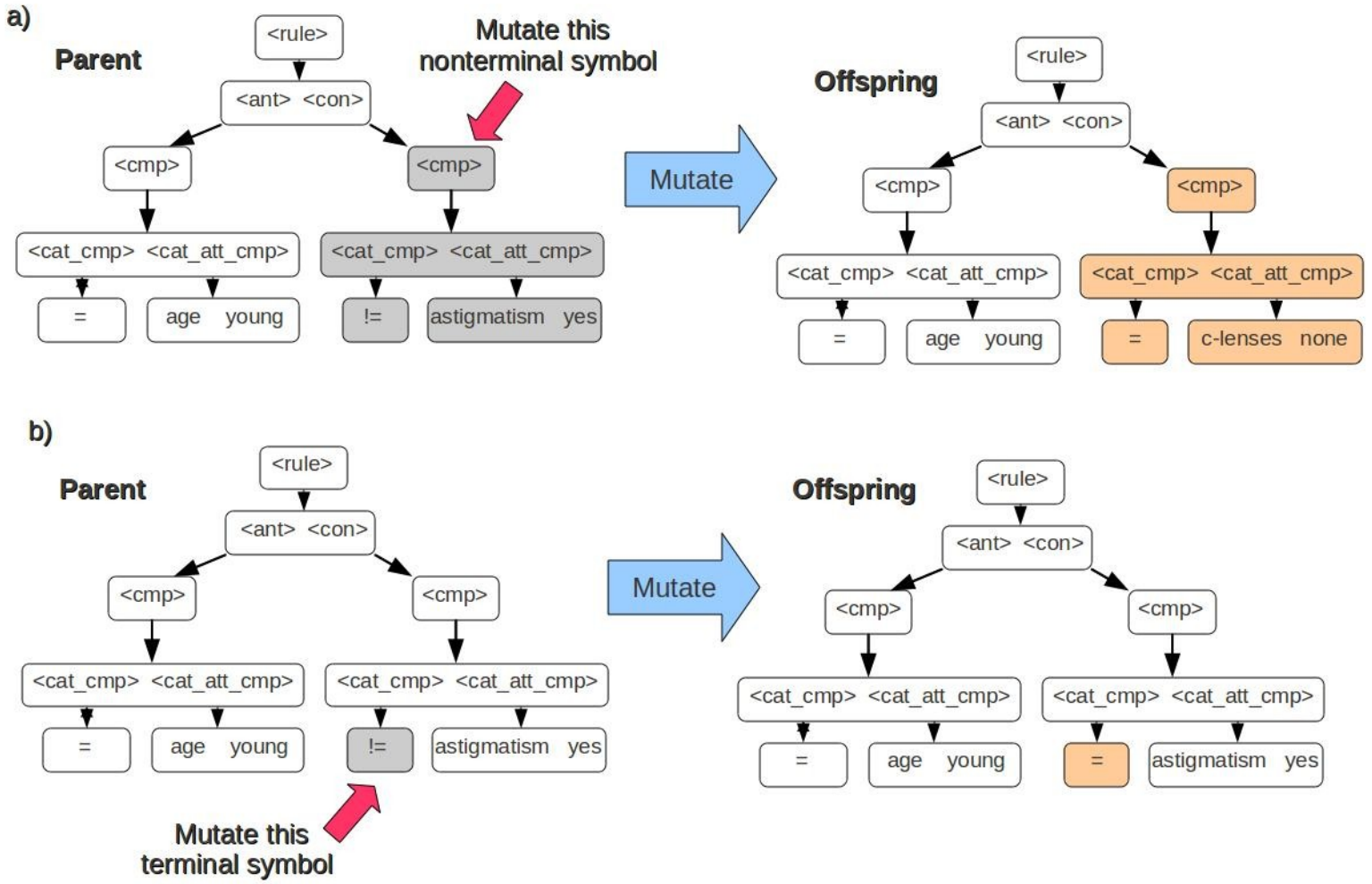
Model description

Crossover



Model description

Mutation



Model description

Measures and fitness function

- G3PARM uses two measures:
  - Support measure: proportion of the number of transactions *T* including the antecedent *A* and the consequent *C* in a dataset *D*.

$$supp(A \rightarrow C) = \frac{|\{A \cup C \subseteq T, T \in D\}|}{|D|}$$

  - Confidence measure: proportion of the number of transactions which include *A* and *C* in transaction which include *A*.

$$conf(A \rightarrow C) = \frac{|\{A \cup C \subseteq T, T \in D\}|}{|\{A \subseteq T, T \in D\}|}$$

- Fitness function = support measure

Model description

G3PARM algorithm

Algorithm 1 G3PARM algorithm

Require: $max\_generations, N$
Ensure: A
1: $P \leftarrow random(N)$
2: $A \leftarrow \emptyset$
3: $num\_generations \leftarrow 0$
4: **while** $num\_generations < max\_generations$ **do**
5: $Parents \leftarrow$Select parents $(P \cup A)$
6: Crossover $(Parents)$
7: Mutation $(Parents)$
8: Update auxiliary population (A, P)
9: $num\_generations++$
10: **end while**
11: **return** A

Algorithm 2 Update auxiliary population

Require: A, P
Ensure: A
1: $A' \leftarrow P \cup A$
2: Order (A')
3: Eliminate duplicate (A')
4: $A \leftarrow$ Threshold(A')
5: **return** A

Model description

G3PARM algorithm

Algorithm 1 G3PARM algorithm

Require: $max\_generations, N$
Ensure: A
1: $P \leftarrow random(N)$
2: $A \leftarrow \emptyset$
3: $num\_generations \leftarrow 0$
4: **while** $num\_generations < max\_generations$ **do**
5: $Parents \leftarrow$Select parents $(P \cup A)$
6: Crossover $(Parents)$
7: Mutation $(Parents)$
8: Update auxiliary population (A, P)
9: $num\_generations++$
10: **end while**
11: **return** A

Algorithm 2 Update auxiliary population

Require: A, P
Ensure: A
1: $A' \leftarrow P \cup A$
2: Order (A')
3: Eliminate duplicate (A')
4: $A \leftarrow$ Threshold(A')
5: **return** A

Rules comprised by the same attributes are considered equals:

1 AND 2 -> 3
2 AND 1 -> 3

Model description

G3PARM algorithm

Algorithm 1 G3PARM algorithm

Require: $max\_generations, N$
Ensure: A
1: $P \leftarrow random(N)$
2: $A \leftarrow \emptyset$
3: $num\_generations \leftarrow 0$
4: **while** $num\_generations < max\_generations$ **do**
5: $Parents \leftarrow$Select parents $(P \cup A)$
6: Crossover $(Parents)$
7: Mutation $(Parents)$
8: Update auxiliary population (A, P)
9: $num\_generations++$
10: **end while**
11: **return** A

Algorithm 2 Update auxiliary population

Require: A, P
Ensure: A
1: $A' \leftarrow P \cup A$
2: Order (A')
3: Eliminate duplicate (A')
4: $A \leftarrow$ Threshold(A')
5: **return** A

Minimum threshold for support and confidence measures

Experiments

Datasets

- To evaluate the usefulness of G3PARM, several experiments have been carried out on differents datasets:
  - Credit-g: 1000 instances and 21 attributes
  - HH: 22784 instances and 17 attributes
  - Mushroom: 8124 instances and 23 attributes
  - Segment: 1500 instances and 20 attributes
  - Sonar: 208 instances and 36 attributes
  - Soybean: 683 instances and 36 attributes
  - Wisconsin Breast Cancer: 683 instances and 11 attributes
- Discretization of datasets with numerical attributes.

Experiments

Experimental set-up

- G3PARM:
  - Population size: 50
  - Crossover probability: 70%
  - Mutation probability: 10%
  - Maximum derivation numbers: 24
  - External population size: 20
  - Support threshold: 70%
  - Confidence threshold: 90%
- Apriori and FP-Growth:
  - Support threshold: 70%
  - Confidence threshold: 90%

Results

Results obtained

Dataset	*Average_support* (1)	(2)	(3)	*Average_confidence* (1)	(2)	(3)	*%Instances* (1)	(2)	(3)
*CreditEqFre*10	0.780	0.709	**0.850**	**0.941**	0.855	0.939	0.987	0.987	**1.000**
*CreditEqFre*5	0.780	0.709	**0.850**	0.941	0.855	**0.953**	0.987	0.987	**1.000**
*CreditEqWid*10	0.780	0.709	**0.892**	0.941	0.855	**0.965**	0.987	0.987	**1.000**
*CreditEqWid*5	0.773	0.709	**0.858**	0.942	0.863	**0.961**	0.989	0.989	**1.000**
*HHEqFre*10	None	None	**0.803**	None	None	**0.913**	None	None	**1.000**
*HHEqFreq*5	None	None	**0.740**	None	None	**0.909**	None	None	**0.997**
*HHEqWid*10	0.761	0.761	**0.922**	0.950	0.950	**0.986**	**1.000**	**1.000**	**1.000**
*HHEqWid*5	0.765	0.765	**0.902**	0.955	0.955	**0.979**	**1.000**	**1.000**	**1.000**
Mushroom	0.824	0.817	**0.890**	0.968	0.960	**0.978**	**1.000**	**1.000**	**1.000**
*SegmentEqFre*10	**0.876**	**0.876**	0.813	**0.975**	**0.975**	0.926	0.996	0.996	**1.000**
*SegmentEqFre*5	**0.876**	**0.876**	0.817	**0.975**	**0.975**	0.974	0.996	0.996	**1.000**
*SegmentEqWid*10	0.815	0.815	**0.884**	0.968	0.968	**0.979**	**1.000**	**1.000**	**1.000**
*SegmentEqWid*5	0.860	0.860	**0.882**	0.964	0.964	**0.969**	**1.000**	**1.000**	**1.000**
*SonarEqFre*10	None	None	**0.782**	None	None	**0.909**	None	None	**1.000**
*SonarEqFre*5	None	None	**0.583**	None	None	**0.731**	None	None	**0.626**
*SonarEqWid*10	None	None	**0.958**	None	None	**0.887**	None	None	**1.000**
*SonarEqWid*5	0.747	None	**0.835**	0.942	None	**0.947**	0.846	None	**1.000**
Soybean	0.778	0.722	**0.822**	0.950	0.953	**0.957**	**1.000**	**1.000**	**1.000**
*WBCEqFre*10	None	None	**0.875**	None	None	**0.958**	None	None	**1.000**
*WBCEqFre*5	None	None	**0.806**	None	None	**0.928**	None	None	**1.000**
*WBCEqWid*10	0.821	0.821	**0.900**	**0.996**	**0.996**	0.971	0.821	0.821	**1.000**
*WBCEqWid*5	**0.872**	**0.872**	0.864	**0.996**	**0.996**	0.956	0.872	0.872	**1.000**
Ranking	2.204	2.522	**1.272**	2.159	2.431	**1.409**	2.340	2.386	**1.272**

- (1) Apriori
- (2) FP-Growth
- (3) G3PARM

Results

Results obtained

Dataset	*Average_support* (1)	(2)	(3)	*Average_confidence* (1)	(2)	(3)	*%Instances* (1)	(2)	(3)
CreditEqFre10	0.780	0.709	**0.850**	**0.941**	0.855	0.939	0.987	0.987	**1.000**
CreditEqFre5	0.780	0.709	**0.850**	0.941	0.855	**0.953**	0.987	0.987	**1.000**
CreditEqWid10	0.780	0.709	**0.892**	0.941	0.855	**0.965**	0.987	0.987	**1.000**
CreditEqWid5	0.773	0.709	**0.858**	0.942	0.863	**0.961**	0.989	0.989	**1.000**
HHEqFre10	None	None	**0.803**	None	None	**0.913**	None	None	**1.000**
HHEqFreq5	None	None	**0.740**	None	None	**0.909**	None	None	**0.997**
HHEqWid10	0.761	0.761	**0.922**	0.950	0.950	**0.986**	**1.000**	**1.000**	**1.000**
HHEqWid5	0.765	0.765	**0.902**	0.955	0.955	**0.979**	**1.000**	**1.000**	**1.000**
Mushroom	0.824	0.817	**0.890**	0.968	0.960	**0.978**	**1.000**	**1.000**	**1.000**
SegmentEqFre10	**0.876**	**0.876**	0.813	**0.975**	**0.975**	0.926	0.996	0.996	**1.000**
SegmentEqFre5	**0.876**	**0.876**	0.817	**0.975**	**0.975**	0.974	0.996	0.996	**1.000**
SegmentEqWid10	0.815	0.815	**0.884**	0.968	0.968	**0.979**	**1.000**	**1.000**	**1.000**
SegmentEqWid5	0.860	0.860	**0.882**	0.964	0.964	**0.969**	**1.000**	**1.000**	**1.000**
SonarEqFre10	None	None	**0.782**	None	None	**0.909**	None	None	**1.000**
SonarEqFre5	None	None	**0.583**	None	None	**0.731**	None	None	**0.626**
SonarEqWid10	None	None	**0.958**	None	None	**0.887**	None	None	**1.000**
SonarEqWid5	0.747	None	**0.835**	0.942	None	**0.947**	0.846	None	**1.000**
Soybean	0.778	0.722	**0.822**	0.950	0.953	**0.957**	**1.000**	**1.000**	**1.000**
WBCEqFre10	None	None	**0.875**	None	None	**0.958**	None	None	**1.000**
WBCEqFre5	None	None	**0.806**	None	None	**0.928**	None	None	**1.000**
WBCEqWid10	0.821	0.821	**0.900**	**0.996**	**0.996**	0.971	0.821	0.821	**1.000**
WBCEqWid5	**0.872**	**0.872**	0.864	**0.996**	**0.996**	0.956	0.872	0.872	**1.000**
Ranking	2.204	2.522	**1.272**	2.159	2.431	**1.409**	2.340	2.386	**1.272**

- (1) Apriori
- (2) FP-Growth
- (3) G3PARM

Results

Results obtained

Dataset	*Average_support* (1)	(2)	(3)	*Average_confidence* (1)	(2)	(3)	*%Instances* (1)	(2)	(3)
CreditEqFre10	0.780	0.709	**0.850**	**0.941**	0.855	0.939	0.987	0.987	**1.000**
CreditEqFre5	0.780	0.709	**0.850**	0.941	0.855	**0.953**	0.987	0.987	**1.000**
CreditEqWid10	0.780	0.709	**0.892**	0.941	0.855	**0.965**	0.987	0.987	**1.000**
CreditEqWid5	0.773	0.709	**0.858**	0.942	0.863	**0.961**	0.989	0.989	**1.000**
HHEqFre10	None	None	**0.803**	None	None	**0.913**	None	None	**1.000**
HHEqFreq5	None	None	**0.740**	None	None	**0.909**	None	None	**0.997**
HHEqWid10	0.761	0.761	**0.922**	0.950	0.950	**0.986**	**1.000**	**1.000**	**1.000**
HHEqWid5	0.765	0.765	**0.902**	0.955	0.955	**0.979**	**1.000**	**1.000**	**1.000**
Mushroom	0.824	0.817	**0.890**	0.968	0.960	**0.978**	**1.000**	**1.000**	**1.000**
SegmentEqFre10	**0.876**	**0.876**	0.813	**0.975**	**0.975**	0.926	0.996	0.996	**1.000**
SegmentEqFre5	**0.876**	**0.876**	0.817	**0.975**	**0.975**	0.974	0.996	0.996	**1.000**
SegmentEqWid10	0.815	0.815	**0.884**	0.968	0.968	**0.979**	**1.000**	**1.000**	**1.000**
SegmentEqWid5	0.860	0.860	**0.882**	0.964	0.964	**0.969**	**1.000**	**1.000**	**1.000**
SonarEqFre10	None	None	**0.782**	None	None	**0.909**	None	None	**1.000**
SonarEqFre5	None	None	**0.583**	None	None	**0.731**	None	None	**0.626**
SonarEqWid10	None	None	**0.958**	None	None	**0.887**	None	None	**1.000**
SonarEqWid5	0.747	None	**0.835**	0.942	None	**0.947**	0.846	None	**1.000**
Soybean	0.778	0.722	**0.822**	0.950	0.953	**0.957**	**1.000**	**1.000**	**1.000**
WBCEqFre10	None	None	**0.875**	None	None	**0.958**	None	None	**1.000**
WBCEqFre5	None	None	**0.806**	None	None	**0.928**	None	None	**1.000**
WBCEqWid10	0.821	0.821	**0.900**	**0.996**	**0.996**	0.971	0.821	0.821	**1.000**
WBCEqWid5	**0.872**	**0.872**	0.864	**0.996**	**0.996**	0.956	0.872	0.872	**1.000**
Ranking	2.204	2.522	**1.272**	2.159	2.431	**1.409**	2.340	2.386	**1.272**

- (1) Apriori
- (2) FP-Growth
- (3) G3PARM

Results

Results obtained

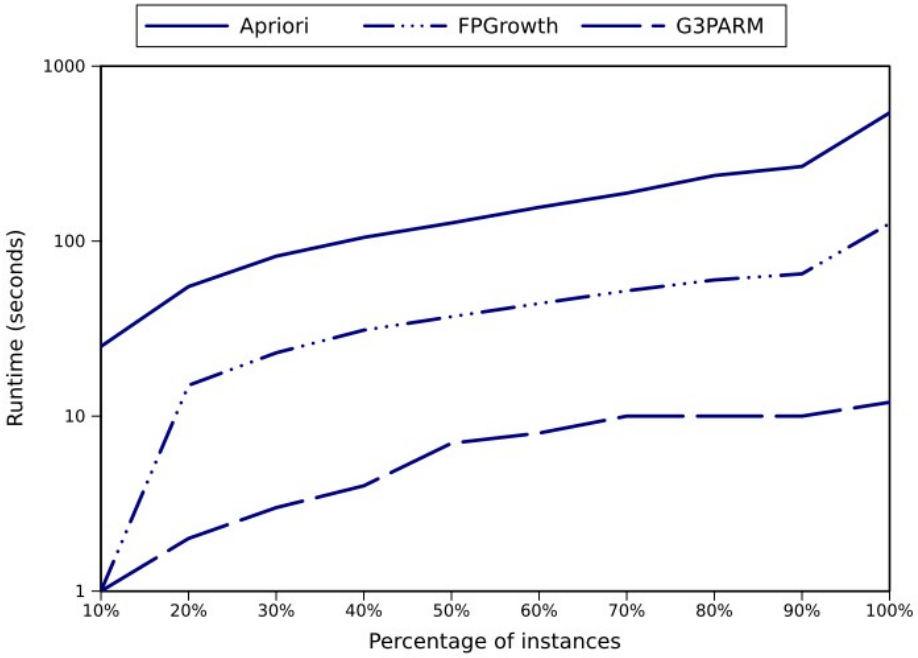
Dataset	*Average_support* (1)	(2)	(3)	*Average_confidence* (1)	(2)	(3)	*%Instances* (1)	(2)	(3)
*CreditEqFre*10	0.780	0.709	**0.850**	**0.941**	0.855	0.939	0.987	0.987	**1.000**
*CreditEqFre*5	0.780	0.709	**0.850**	0.941	0.855	**0.953**	0.987	0.987	**1.000**
*CreditEqWid*10	0.780	0.709	**0.892**	0.941	0.855	**0.965**	0.987	0.987	**1.000**
*CreditEqWid*5	0.773	0.709	**0.858**	0.942	0.863	**0.961**	0.989	0.989	**1.000**
*HHEqFre*10	None	None	**0.803**	None	None	**0.913**	None	None	**1.000**
*HHEqFreq*5	None	None	**0.740**	None	None	**0.909**	None	None	**0.997**
*HHEqWid*10	0.761	0.761	**0.922**	0.950	0.950	**0.986**	**1.000**	**1.000**	**1.000**
*HHEqWid*5	0.765	0.765	**0.902**	0.955	0.955	**0.979**	**1.000**	**1.000**	**1.000**
Mushroom	0.824	0.817	**0.890**	0.968	0.960	**0.978**	**1.000**	**1.000**	**1.000**
*SegmentEqFre*10	**0.876**	**0.876**	0.813	**0.975**	**0.975**	0.926	0.996	0.996	**1.000**
*SegmentEqFre*5	**0.876**	**0.876**	0.817	**0.975**	**0.975**	0.974	0.996	0.996	**1.000**
*SegmentEqWid*10	0.815	0.815	**0.884**	0.968	0.968	**0.979**	**1.000**	**1.000**	**1.000**
*SegmentEqWid*5	0.860	0.860	**0.882**	0.964	0.964	**0.969**	**1.000**	**1.000**	**1.000**
*SonarEqFre*10	None	None	**0.782**	None	None	**0.909**	None	None	**1.000**
*SonarEqFre*5	None	None	**0.583**	None	None	**0.731**	None	None	**0.626**
*SonarEqWid*10	None	None	**0.958**	None	None	**0.887**	None	None	**1.000**
*SonarEqWid*5	0.747	None	**0.835**	0.942	None	**0.947**	0.846	None	**1.000**
Soybean	0.778	0.722	**0.822**	0.950	0.953	**0.957**	**1.000**	**1.000**	**1.000**
*WBCEqFre*10	None	None	**0.875**	None	None	**0.958**	None	None	**1.000**
*WBCEqFre*5	None	None	**0.806**	None	None	**0.928**	None	None	**1.000**
*WBCEqWid*10	0.821	0.821	**0.900**	**0.996**	**0.996**	0.971	0.821	0.821	**1.000**
*WBCEqWid*5	**0.872**	**0.872**	0.864	**0.996**	**0.996**	0.956	0.872	0.872	**1.000**
Ranking	2.204	2.522	**1.272**	2.159	2.431	**1.409**	2.340	2.386	**1.272**

- (1) Apriori
- (2) FP-Growth
- (3) G3PARM

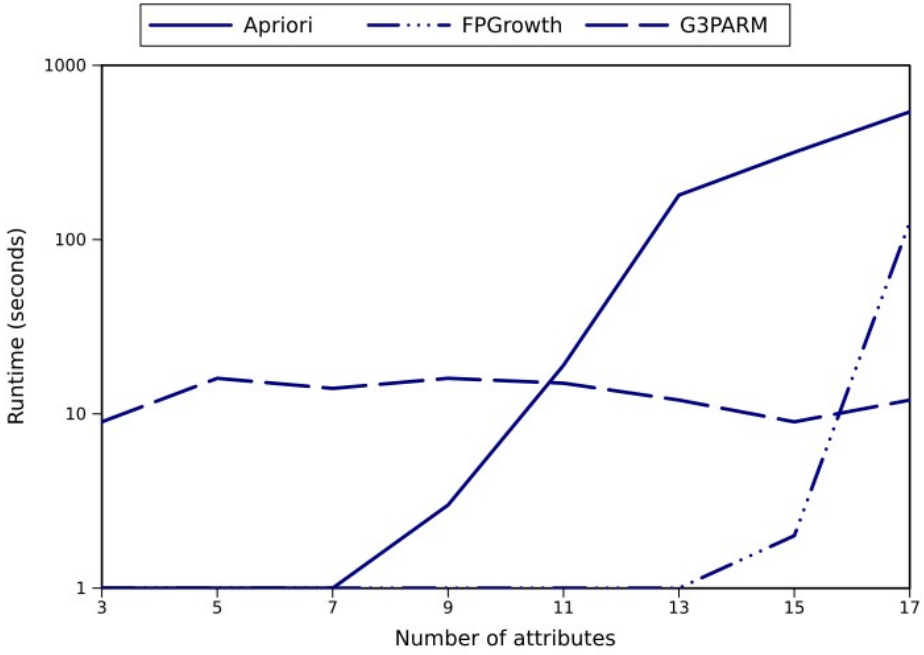
Results

Scalability

Different number of instances



Different number of attributes



Conclusions

- Novel G3P-based algorithm for mining association rules
- G3PARM obtains rules with:
  - High support
  - High confidence
  - High representative rules
- G3PARM scales quite linearly as we increase up the dataset size and the number of attributes

Future research lines

- Use of numerical attributes
- Modify our approach to work with infrequent pattern
- Multiobjective version using support and confidence as objectives to be optimized

UNIVERSIDAD
D
CÓRDOBA

Thanks!

KDIS Lab
KNOWLEDGE DISCOVERY AND INTELLIGENT SYSTEMS

Analysis of the Effectiveness of G3PARM Algorithm

J.M. Luna, J.R. Romero and S. Ventura

Knowledge Discovery and Intelligent Systems Research Group
University of Córdoba, Spain

HAIS 2010. San Sebastián, Spain. June 2010.