Pontifícia Universidade Católica de Campinas
Centro Universitário das Faculdades Associadas de Ensino
Instituto de Ciências Matemáticas e de Computação

A Comparison of Methods for Rule Subset Selection Applied to Associative Classification

Gustavo Batista, Claudia Milaré,
Ronaldo Prati and Maria Carolina Monard

ASAI 2006

Content

- Introduction
- Rule Subset Selection
- GARSS
- Experimental Evaluation
- Conclusion and Future Work

Introduction

- Rules induction is an important approach to extract knowledge in KDD;
- However, some situations might lead to the induction of excessively large and complex rule sets;
- One instance is the generation of association rules.

Introduction

Age	Income	Gender	Risk
35	$10.000	Male	High
42	$21.000	Female	Low
21	$15.000	Male	High
69	$40.000	Female	Low
29	$35.000	Female	Low

- Gender = Female => Income > 20.000
- Income < 16.000 => Risk = High

Rule Subset Selection — RSS

- A very direct approach to reduce the number of rules is to post process the ruleset, selecting a subset of the induced rules, a method known as *rule subset selection*;
- Rule subsets can be constructed aiming to maximize several important ruleset quality criteria, such as comprehensibility, interestingness, and classification performance.

GARSS

- In this work, we present GARSS – Genetic Algorithm for Rule Subset Selection;
- GARSS searches for a rule set that maximizes the Area under the ROC Curve – AUC;
- GARSS results are compared with a recently proposed algorithm, ROCCER,which looks for a rule subset that creates a convex hull in the ROC space.

Associative Classification

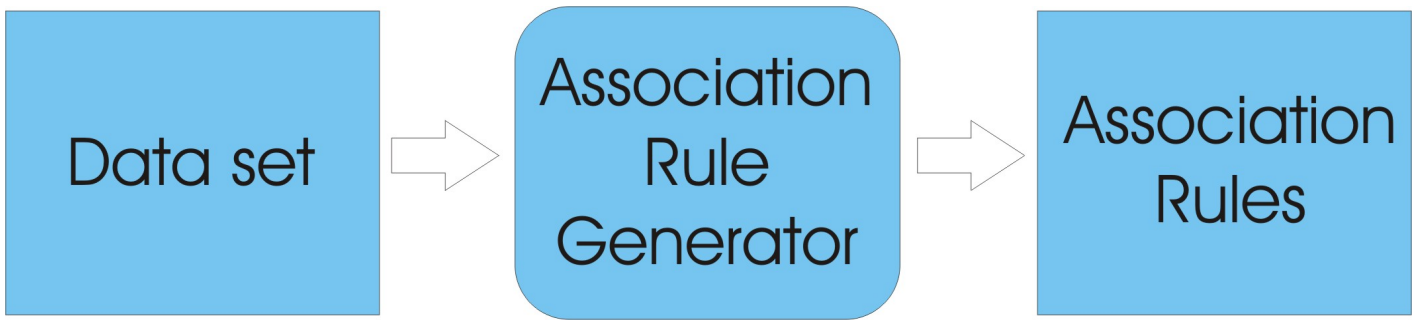
- Our experiments are performed using an *associative classifier*;
- An associative classifier is composed by all rules where the consequent is the class-attribute, known as *class association rules – CARs*;
- The number of CARs frequently outnumbers the number of examples, implying serious restrictions for knowledge deployment.

Introduction

Age	Income	Gender	Risk
35	$10.000	Male	High
42	$21.000	Female	Low
21	$15.000	Male	High
69	$40.000	Female	Low
29	$35.000	Female	Low

- Gender = Female => Income > 20.000
- Income < 16.000 => Risk = High

Associative Classification



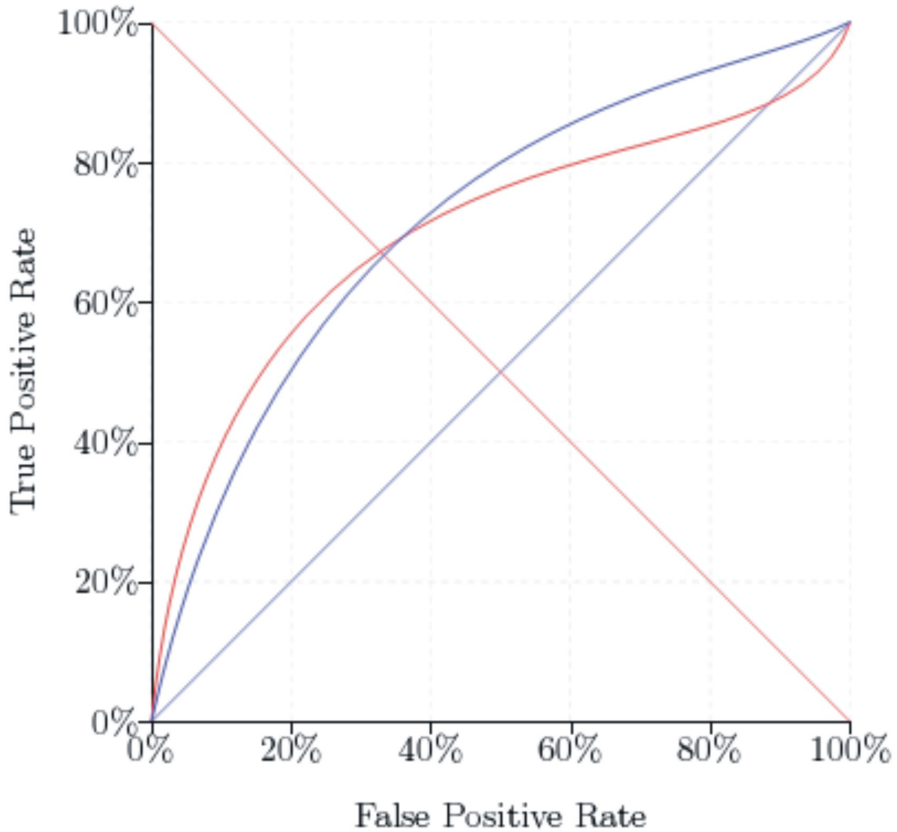
Associative Classification

Associative Classification

GARSS's Objective for Associative Classifiers

To considerably reduce the number of rules of associative classifiers, and still be able to achieve similar classification performance in terms of AUC.

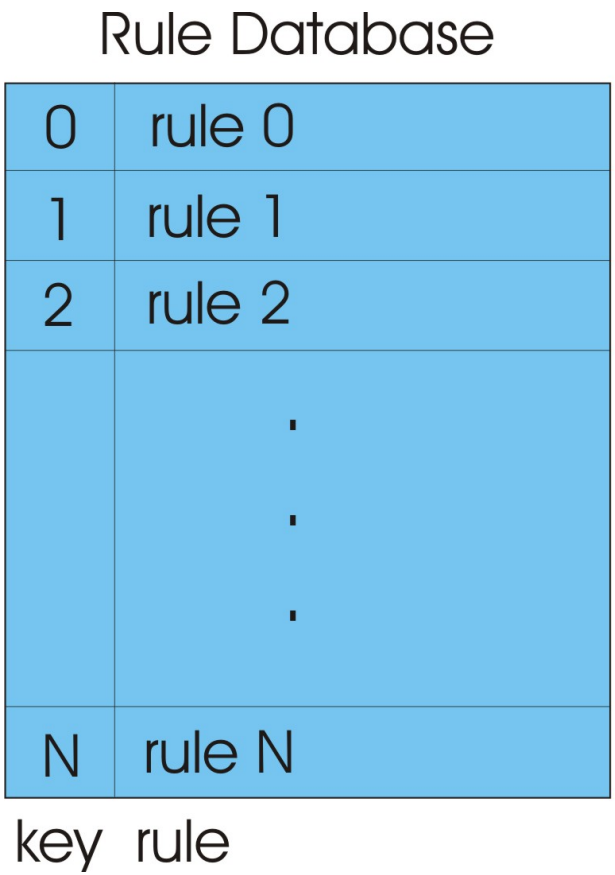
ROC Analysis

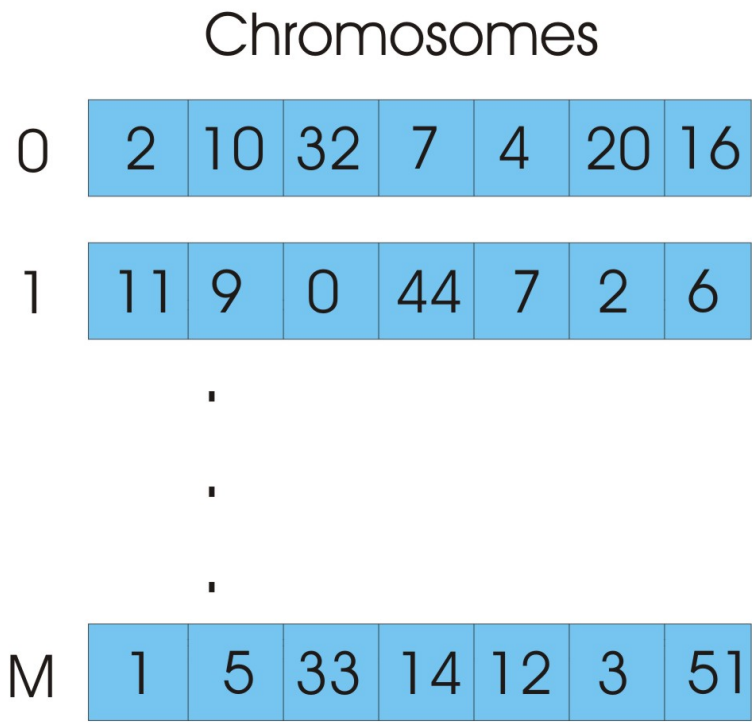


GARSS Algorithm

- GARSS is a rule selection algorithm that uses genetic algorithms to select rules aiming to maximize the AUC metric;
- We use a rule database composed by all CARs generated by the Apriori algorithm;
- A primary key composed by a positive integer is associated to each rule in the database. Therefore, each rule can be accessed independently by its key;
- An array of keys (integers) is used to represent a chromosome.

GARSS Algorithm

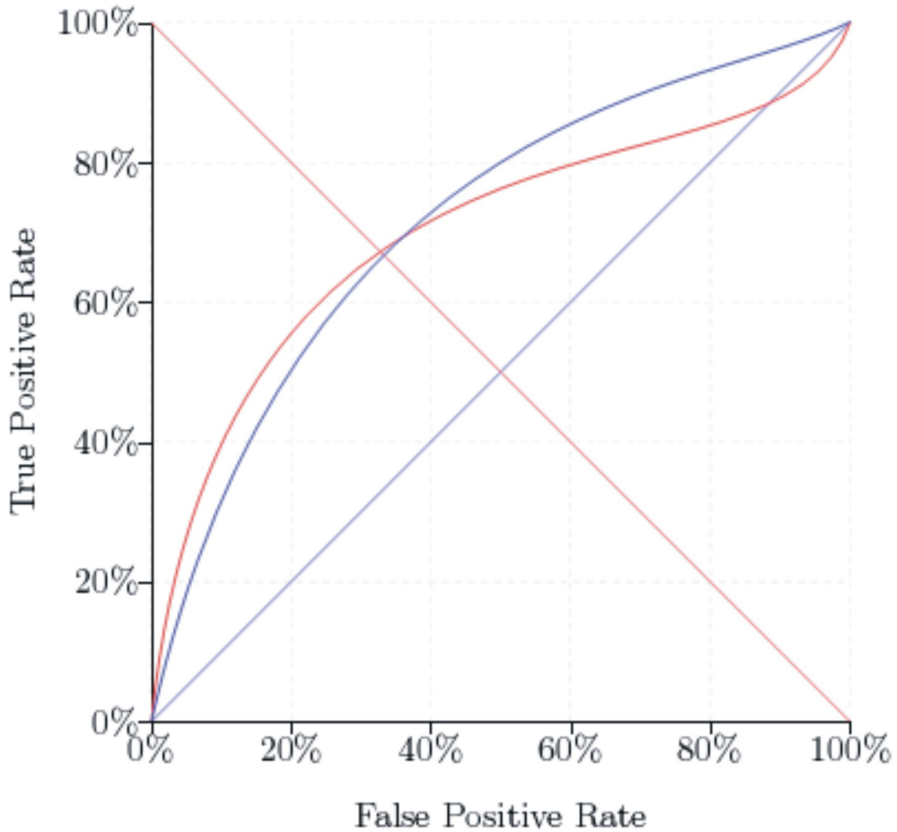




ROCCER Algorithm

- Uses ROC graphs to build rule sets;
- The algorithm works by selecting rules from a larger rule set;
- Rules are only kept if they improve the ROC convex hull.

ROC Analysis



Experimental Setup

- Four UCI datasets were used in the experiments;
- All of them have no missing data, since Apriori cannot handle them;
- Only two-class problems were used, in order to calculate AUC values.

Experimental Setup

Data set	#Attrs	#Examples	Maj. Class %
Breast	10	683	65.00
Bupa	7	345	57.98
German	21	1000	70.00
Heart	14	270	55.55

Experimental Setup

- AUC values were estimated using stratified 10-fold cross-validation;
- The experiments were paired, *i.e.*, all inducers were given the same training and test sets;
- Four algorithms were compared: GARSS and ROCCER; C4.5 (a well-known divide-and-conquer learning system); and All that is a classifier composed by all CARs generated by Apriori.

Experimental Results – AUC

Data sets	GARSS	ROCCER	C4.5	All
Breast	99.06(0.46)	98.63(1.88)	97.76(1.51)	99.07(0.87)
Bupa	64.65(3.96)	65.30(7.93)	62.14(9.91)	65.38(10.63)
German	74.16(1.60)	72.08(6.02)	71.43(5.89)	73.37(4.84)
Heart	82.86(3.36)	85.78(8.43)	84.81(6.57)	90.72(6.28)
Avg	80.18	80.45	79.04	82.14

Experimental Results – Number of Rules

Data sets	GARSS	ROCCER	C4.5	All
Breast	46.00(2.56)	48.4(2.32)	37.8(12.62)	502.1(8.96)
Bupa	61.10(1.55)	3.9(0.99)	15(10.53)	292.8(21.57)
German	34.90(5.88)	23.7(6.75)	78.2(18.5)	2886.1(577.3)
Heart	43.90(1.89)	68.2(4.42)	13.2(4.49)	1875.6(91.9)
Avg	46.48	36.05	36.05	1389.15

Results

- Classifiers generated by All approach have a large number of rules. For German and Heart data sets, the average number of rules is even greater than the number of examples;
- For GARSS, the average number of rules for Breast, Bupa, German and Heart data sets represents 9.16%, 20.87%, 1.21% and 2.34% of all generated CARs, respectively;
- For ROCCER, they represent 9.64%, 1.33%, 0.82% and 3.64% of all generated CARs, respectively.

Results

- In general, GARSS and ROCCER provided average AUC values that are similar or better than C4.5;
- Comparing with all class association rules, the average AUC results are similar for the data sets Breast, Bupa and German, although lower for Heart;
- On the other hand, GARSS and ROCCER provided a significant reduction in the average number of rules, thus improving the comprehensibility of the final rule sets.

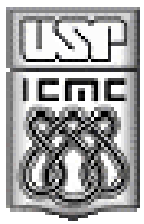
Conclusion

- GARSS and ROCCER showed a classification performance (in AUC) similar or better than C4.5;
- In addition, GARSS and ROCCER were able to create rule sets considerably smaller than the associative classifier with all CARs;
- GARSS and ROCCER were able to maintain similar classification performance in three of the four data sets used in the experiments.

Future Work

- As GARSS and ROCCER are post-processing algorithms, they are not limited to associative classifiers;
- An interesting approach for future research is to use GARSS and ROCCER as a pruning method, where instead of feeding them with class association rules, we use as input other symbolic classifiers (such as the tree induced by C4.5);
- Another interesting research direction is to evaluate the performance of GARSS and ROCCER as methods for combining rules from different classifiers.

gbatista@puc-campinas.edu.br
cmilare@fae.br
{mcmonard,prati}@icmc.usp.br



PUC
CAMPINAS
PONTIFÍCIA UNIVERSIDADE CATÓLICA