Feature Extraction from Top Association Rules:

Effect on Average Predictive Accuracy

José L. Balcázar
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
Barcelona, Spain

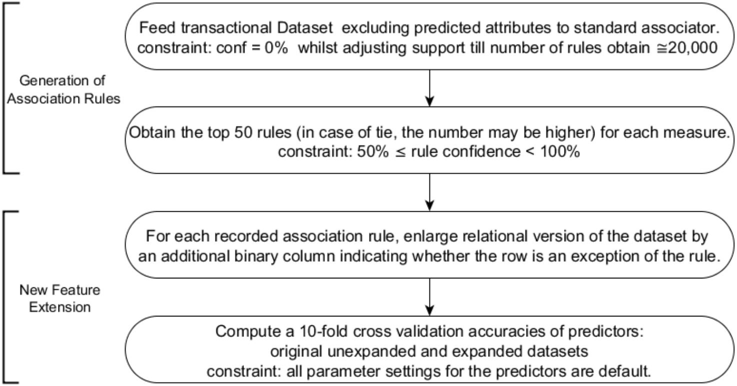
Francis Dogbey
Advanced Information Technology Institute
Ghana-India Kofi Annan Centre of Excellence in ICT
Accra, Ghana

EUCOG III Conference, Palma de Mallorca,2013

Introduction/Motivation

- Notation: Association rule, expressed, $X \rightarrow A$, where X is a set of items and A is an item.
  - Semantic Interpretations:
    - relaxed implication: A tends to appear when X appears
    - Standard implication: Every observation that includes X must include A
  - Issue/Challenge:
    - Different formalizations of a relaxed implication tend measures differently the intensity of the implications. Which measure leads to the most useful results for a given data mining application?
- Informal Hypothesis: Given that different rule quality measures select different rules, we could objectively compare the "top" set of rules by using them as help in other Data Mining decision processes, and to evaluate their contribution.
- Rule Quality Measures Evaluated: support(s), confidence(c), relative confidence(r), lift(l), leverage(v), improvement (i) and multiplicative improvement(m)

Process



Generation of Association Rules

Feed transactional Dataset excluding predicted attributes to standard associator.
constraint: conf = 0% whilst adjusting support till number of rules obtain ≅20,000

Obtain the top 50 rules (in case of tie, the number may be higher) for each measure.
constraint: 50% ≤ rule confidence < 100%

New Feature Extension

For each recorded association rule, enlarge relational version of the dataset by an additional binary column indicating whether the row is an exception of the rule.

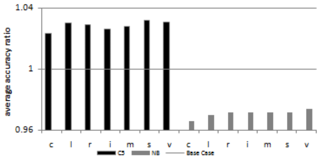
Compute a 10-fold cross validation accuracies of predictors:
original unexpanded and expanded datasets
constraint: all parameter settings for the predictors are default.

Discussion

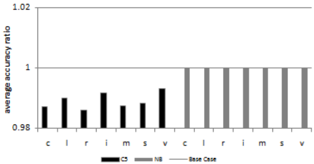
Dataset Information/Baseline Predictions on dataset

Dataset	Pred. Attr.	Size	Total Values	Supp.	Rules	Average Accuracy	
						C5.0	NB
Cmc	Contra. Meth.	1,473	74	1%	19,067	50.17 %	51.22%
Sweets	Product 27	384	431	6%	16,672	67.96%	69.78%
Votes	Party	435	50	22.5%	19,544	95.40%	61.42%

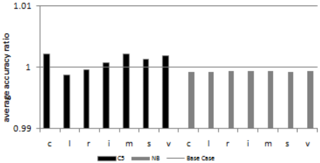
Effects of New Feature on the Expanded Datasets



Cmc dataset



sweets dataset



votes dataset

Observations & Conclusion

- CMC dataset
  - remarkable positive improvement in the average accuracy for C5.0 but negative improvement for NB for all measures.
    REASON: *NB is based on (conditional)independence of the attribute values, but the new attribute added is not independent at all of the original attributes.*
- VOTES dataset
  - marginal positive average accuracy improvement for measures except for l and r on C5.0 but NB predictor seems to ignore the extra feature as only minor negative improvements were noticed for expanded dataset.
- SWEETS dataset
  - predicting sweet 27 reports insiginficant but stable positive improvement for all measures on NB but C5.0 suffers a consistent deterioration of its accuracy.
- CONCLUSION
  - From the preliminary experiments, no measure fares particularly well, although some seem to be somewhat better. *A finer evaluation of more datasets and "good" rules among the top 50 rules of the measures whose corresponding new feature improves predictors will be furnished in a forthcoming paper.*