Evolving Multi-label classification rules with GEP: a preliminary study

J.L. Avila. E.L. Gibaja. S. Ventura.

Departament of Computer Science and Numerical Analysis

KDIS Lab KNOWLEDGE DISCOVERY AND INTELLIGENT SYSTEMS

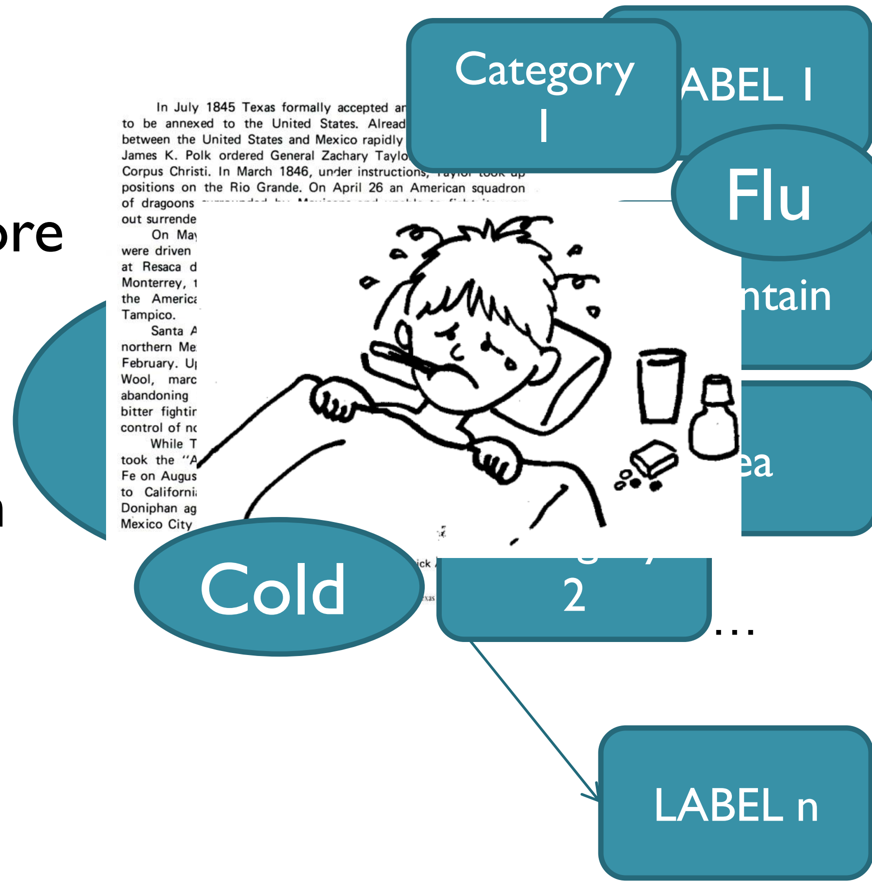
UNIVERSITY OF CORDOBA

Table of contents

- Classification and Multi-label Clasification
  - Multi -label classification
  - Techniques
- Proposed Algorithm
  - Main features
  - Individual representation
  - Genetic operators
  - Individual evaluation
  - Token competition
- Result and discussion
  - Experiments
  - Results
  - Discussion
- Conclusions

Classification and MLC

- Multilabel classification:
  - Each pattern associated with more than one label
- Many problems:
  - Text categorization
  - Image classification
  - Medical diagnosis
  - …



MLC Techniques

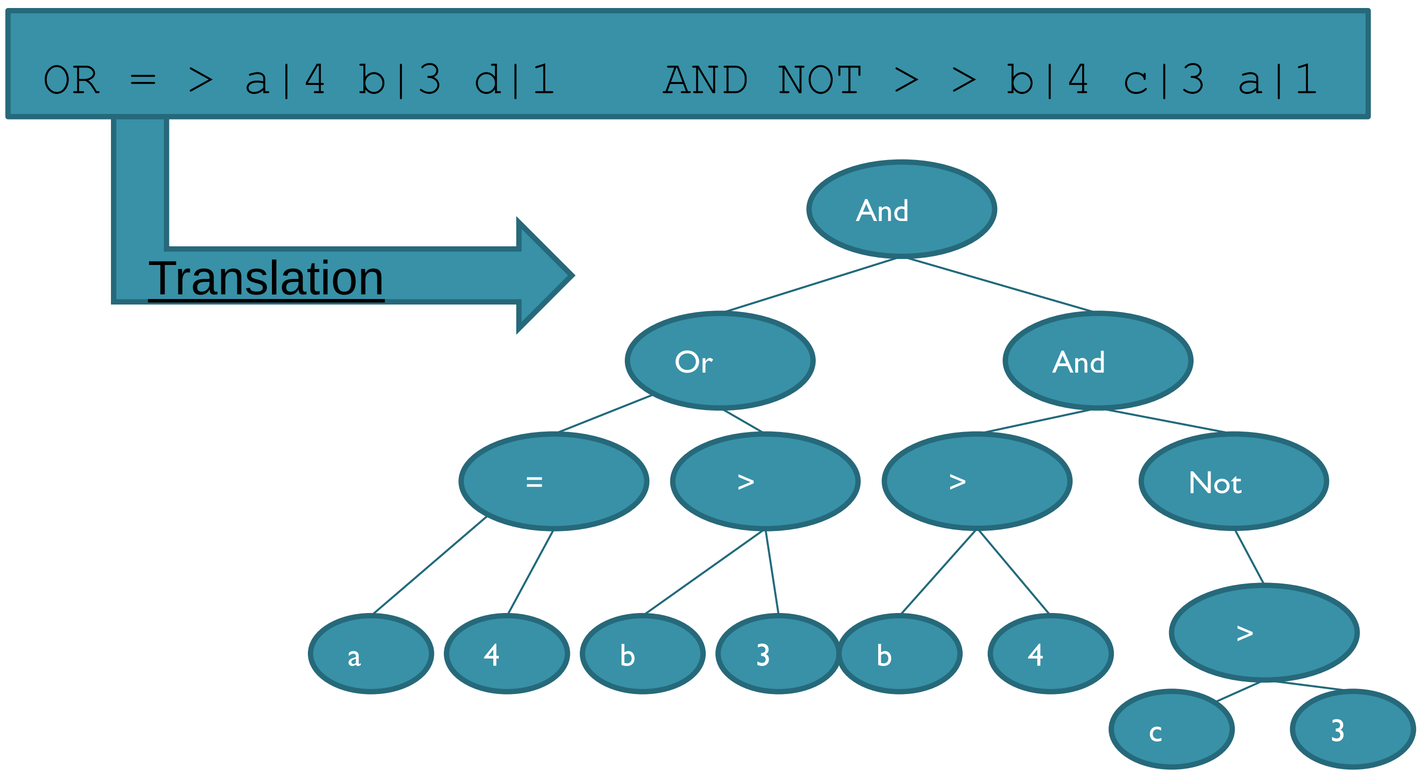
- Pre-processing techniques:
  - Transform a ML problem into several single label problems
    - Binary Relevance
    - Label Powerset
- Multi-label specific techniques:
  - Support vector machines
  - ML-KNN
  - Ensemble methods.
  - …

GC-ML: MAIN features

- Gene Expression Programming: succesfully used in classification
- Each Individual encodes a rule
  - *IF(CONDITION) THEN LABEL*
    - *Condition has both logical and relational operators*
- Niching algorithm to improve genetic diversity
- Final classifier is built using a set of rules

Individual representation

- Dual encoding: Genotype and phenotype
  - Genotype: Lineal String
- Phenotype: Syntax tree and codifies a rule



Genetic Operators

- Recombination operators
  - One point recombination
  - Two points recombination
  - Gene recombination
- Mutation operator
- Transposition operators
  - IS transposition
  - RIS transposition
  - Gene transposition

Individual evaluation

- Fitness function: F-score:

$$raw\_fitness = \frac{2 \times precission \times recall}{precission + recall}$$

- Calculated for each label
  - N raw fitness for individual
- Fitness is obtained after Token Competition

Token competition

- Niching effect
- One Token for each pattern and class
- Corrects the fitness

$$new\_fitness = \frac{raw\_fitness \times tokens\_won}{Total\_tokens}$$

- Penalizes individual which does not contribute to the classifier

Experiments

- GC ML has been compared with
  - Binary Relevance
  - Label Powerset
  - ML-KNN
- Measures: Accuracy, precision and recall
- Datasets

	Scene	Yeast	Genbase	Medical
Number of labels	6	14	27	45
Label cardinality	1,06	4,23	1,25	1,24
Label density	0,18	0,30	0,04	0,028
Number of patterns	2407	2417	662	978

Results

	Binary rel.			Label pow.			ML-KNN			GC-ML		
	Acc	Prec	Rec	Acc	Prec	Rec	Acc	Prec	Rec	Acc	Prec	Rec
Scene	0,43	0,44	0,81	0,57	0,60	0,59	0,62	0,66	0,67	0,57	0,55	0,69
Genbase	0,27	0,28	0,27	0,68	0,67	0,65	0,63	0,67	0,63	0,77	0,75	0,68
Yeast	0,42	0,61	0,62	0,39	0,52	0,52	0,49	0,54	0,54	0,43	0,57	0,57
Medical	0,59	0,65	0,61	0,61	0,67	0,65	0,56	0,57	0,56	0,65	0,70	0,70

- GC ML shows better results that other
  - Better than trasformation methods
- Results are better with nominal datasets

Conclusions

- GC-ML
  - Evolutionary: GEP
  - Learn classification rules
  - Niching technique
- Similar performance
  - Better than transformation methods
  - Better with categorical datasets
- Future research
  - Compare with other implementations
  - Test in other domains
  - Improve efficiency

Evolving Multi label classification rules with GEP: a preliminary study

J.L. Avila. E.L. Gibaja. S. Ventura.

Departament of Computer Science and Numerical Analysis

Thank you for your attention

Any question?

KDIS Lab

KNOWLEDGE DISCOVERY AND INTELLIGENT SYSTEMS

UNIVERSITY OF CORDOBA