CLIP4 Inductive Machine Learning Algorithm

Lukasz A. Kurgan
Advisor: Dr. Krzysztof J. Cios

Intelligent Systems Laboratory
Department of Bioengineering

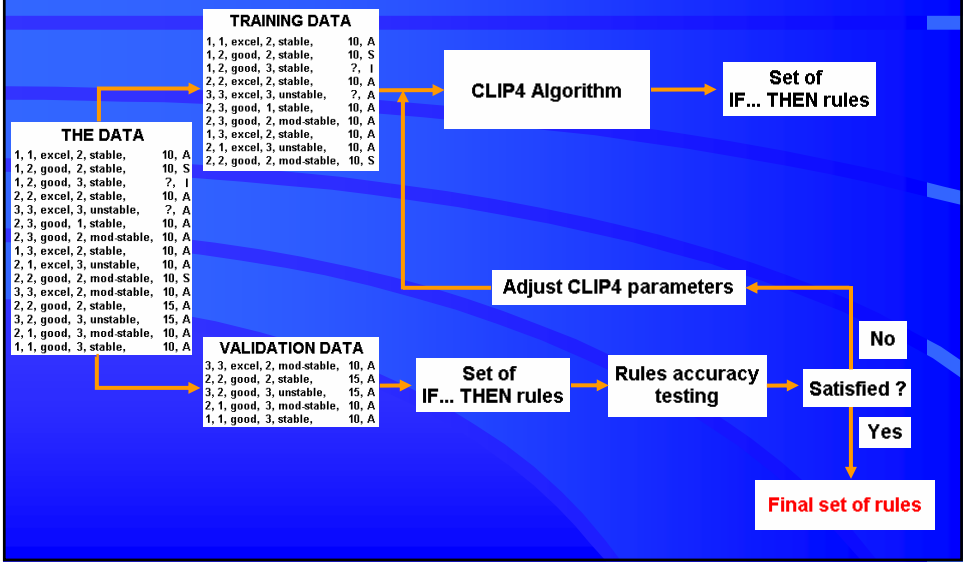
Presentation Outline

- Goal
- Introduction
- CLIP4 Algorithm
- Results and Summary

Goal

- Development of a new inductive machine learning algorithm CLIP4 (Cover Learning using Integer Programming)
- This algorithm is a descendent of CLIP3 algorithm (Cios, Wedding and Liu in 1997)

Introduction



Introduction

CLIP4 Algorithm:

- **First, it divides data into noise-free subsets by using integer-programming mode, and a pruning technique to partition the data**
- **Then, it generates rules from these subsets using a set covering technique.**

Introduction

Training data

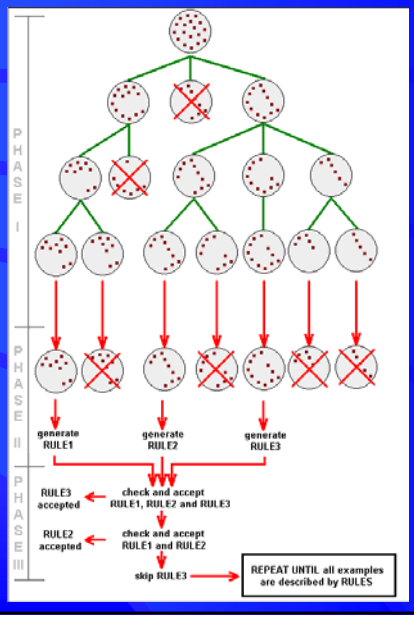
Division of the data into subsets

Tree-leaves represent the terminal subsets

Verification of the terminal subsets

Generation of rules from the chosen subsets

Verification and acceptance of rules



Introduction

CLIP4 algorithm can be used for:

- **classification problems**
  (generation of rules describing classification of the data)
- **data analysis problems**
  (finding significant attributes in the data)

Example applications:

- **generation of diagnostic rules for heart perfusion classification, cancer type classification**
- **generation of compact and highly accurate user satisfaction instruments**

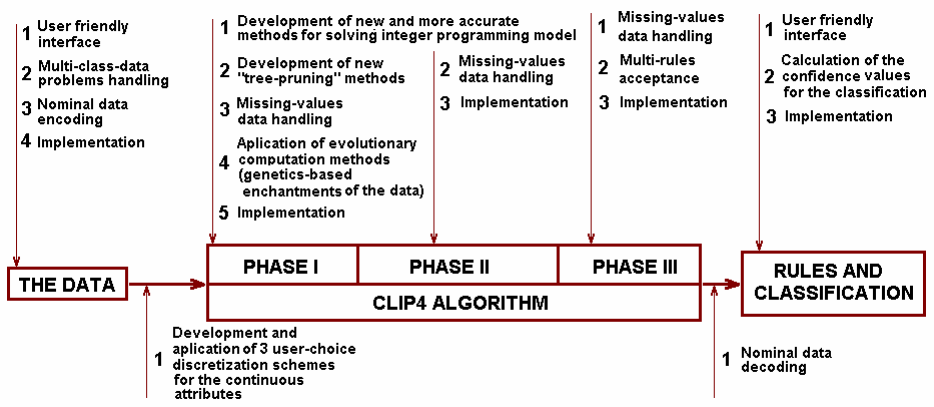
CLIP4 Inductive Machine Learning Algorithm, Lukasz Kurgan

CLIP4 Algorithm

- **Defining research goals**
- **Initial implementation of the algorithm**
- **Literature search**
- **Iterative process of goals realization and re-definition based on literature and own ideas**
- **Validation and adjustments of the algorithm based on its performance**
- **User-friendly implementation of the algorithm**
- **Publication and popularization of the algorithm**

CLIP4 Algorithm

Improvements and new features



Outcome

- Faster data processing that allows it to analyze large data sets
- Generation of more accurate rules
- Ability to work with missing-value data, multi-class problems, to discretize the data
- Easy to use software

CLIP4 Algorithm

CLIP3 vs. CLIP4

- **MONKS data**

Algorithm	MONKS 1		MONKS 2		MONKS 3	
	Nr rules	Accuracy	Nr rules	Accuracy	Nr rules	Accuracy
CLIP4	**4**	**100**	**15**	**88.66**	**3**	**88.89**
CLIP4	**4**	**100**	**8**	**81.94**	**8**	**95.14**
CLIP3 (threshold 1)	4	100	10	82.7	3	88.9
CLIP3 (threshold 2)	4	100	7	72.7	2	97.2
ID3 with windowing	28	98.6	110	67.9	29	94.4
CN2	10	100	58	69.0	24	89.1
C4.5 decision tree	----	75.7	----	65.0	----	97.2

- **Breast Cancer data**

Algorithm	Breast Cancer	
	Nr rules	Accuracy
CLIP4	**3**	**95.71**
CLIP4	**1**	**93.71**
CLIP3 (threshold 0)	----	89.6
CLIP3 (threshold 1)	----	86.8
CLIP3 (threshold 2)	----	92.4
C4.5	----	90.1

CLIP4 Inductive Machine Learning Algorithm, Lukasz Kurgan

Summary

The new algorithm was developed and implemented that is

- **more accurate**
- **more data compatible**
- **very user-friendly**

Feature goals

- **To do more data benchmarking on large datasets**
- **To compare its results with other algorithms**
- **To publish the findings in a machine learning journal**
- **To popularize CLIP4 by offering it as a freeware to the research community**

Major references

1. Cios, K.J., Liu, N., "An algorithm which learns multiple covers via integer linear programming, Part I - the CLILP2 algorithm", *Kybernetes*, 24(2): 29-50, 1995

2. Cios, K.J., Wedding, D.K., Liu, N., "CLIP3: cover learning using integer programming", *Kybernetes,* 26(4,5): 513-536, 1997

3. Cios, KJ, Pedrycz, W., Swiniarski, R., "Data Mining Methods for Knowledge Discovery", Kluwer, 1998

4. Kurgan, L.A. et al. "Knowledge Discovery Approach to Automated Cardiac SPECT Diagnosis", *Artificial Intelligence in Medicine*, 2000, submitted