HAIS'10

5th International Conference on HYBRID ARTIFICIAL INTELLIGENCE SYSTEMS

MULTIVARIATE DISCRETIZATION FOR ASSOCIATIVE CLASSIFICATION IN A SPARSE DATA APPLICATION DOMAIN

María N. Moreno García, Joel Pinho Lucas, Vivian F. López Batista and M. José Polo Martín

Dept. of Computing and Automatic

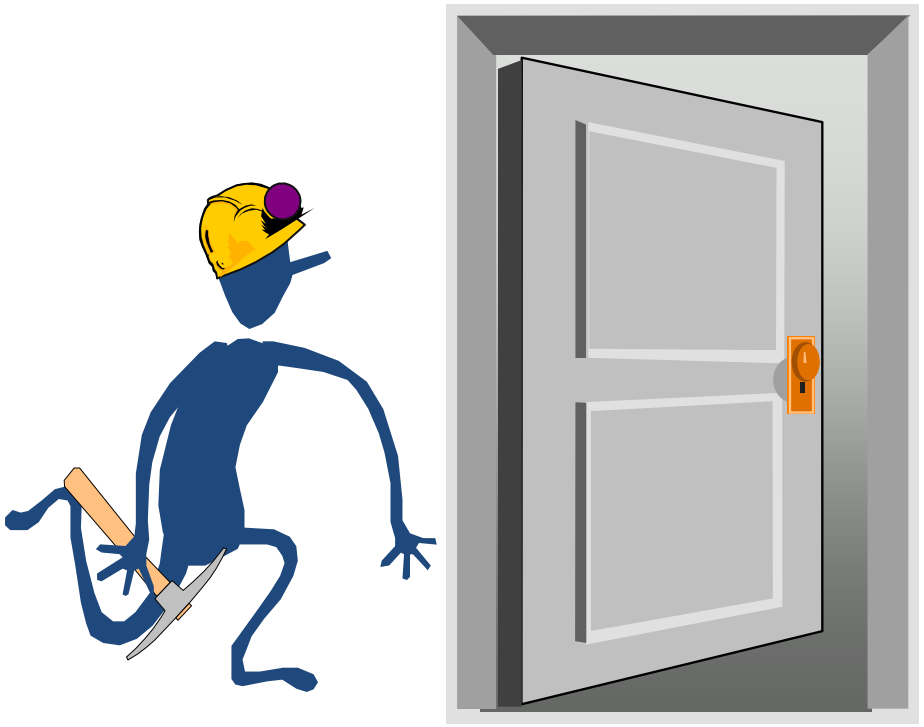
VNiVERSiDAD D SALAMANCA

Contents

- Introduction
- Proposed method
- Experimental study
- Results
- Conclusions

Contents

- **Introduction**
- Proposed method
- Experimental study
- Results
- Conclusions



Introduction

- **Objective**

To improve the precision of software estimations in the project management field

- Drawbacks of applying data mining techniques:
  - Data sparsity
    - Many attributes
    - Scarce number of available examples
  - Most of the involved attributes are continuous

Introduction

- **Proposal**
  - **Associative classification**
    - Machine learning technique that combines concepts from classification and association
    - Input: discrete attributes
  - **CBD** (Clustering Based Discretization) **algorithm**
    - Supervised, multivariate discretization process
    - Selection of the best attributes for classification
    - Based on supervised clustering

Introduction

Associative classification

- Set of discrete attributes $I = \{i_1, i_2, \ldots, i_m\}$
- Set of N transactions $D = \{T_1, T_2, \ldots, T_N\}$
- Atomic condition:

  $value_1 \leq attribute \leq value_2$ or $attribute = value$

  value, $value_1$ and $value_2$ in D

Association rule

$X \rightarrow A$

X is an itemset: the conjunction of atomic conditions

A can be an itemset or an atomic condition

Associative classification

A is the **class** attribute

CARs

Class Association Rules

Introduction

- **Associative classification methods**
  - Build a classifier from the associative model
  - Classification model is presented as an ordered list of rules obtained by a rule ordering mechanism
  - The most popular methods:
    - **CBA** (Classification Based in Association)
    - **MCAR** (Classification based on Predictive Association Rules)
    - **CMAR** (Classification based on Multiple class-Association Rules)
    - **CPAR** (Classification based on predictive association rules)

Introduction

- **Advantages of Associative Classification**
  - Associative classification methods are slightly sensitive to data sparsity
  - Association models are commonly more effective than classification models
  - Several works (Liu et. al) (Li et. al.) (Thabtah et. al) (Yin y Han) verified that classification based on association methods presents higher accuracy than traditional classification methods

Introduction

- Types of association rules
  - **Boolean:** binary attributes
  - **Nominal:** discrete attributes
  - **Quantitative:** continuous numerical attributes

Cost = 5.25 → precision = 85.3

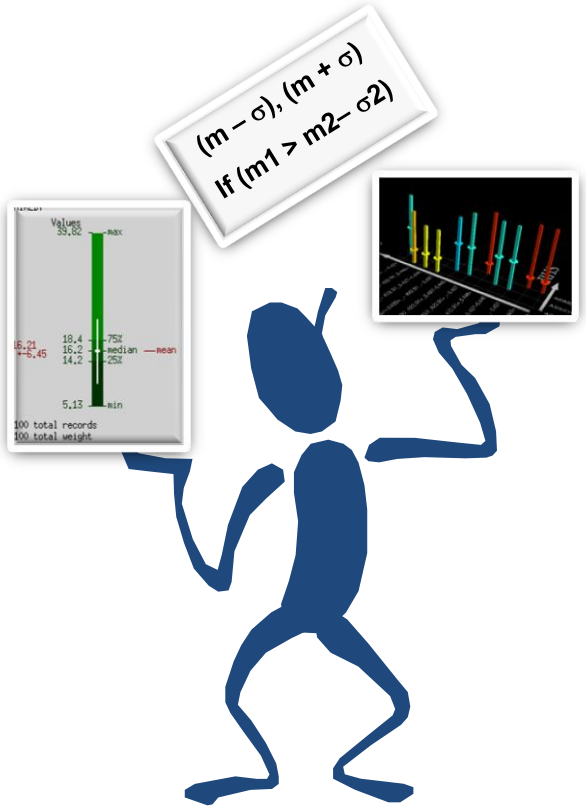
Quantiative association rules

↓

Discretization process

Contents

- Introduction
- **Proposed method**
- Experimental study
- Results
- Conclusions



Proposed method

- **Types of discretization**

  - **Univariate**: quantifies one continuous attribute at a time
  - **Multivariate**: considers simultaneously multiple attributes

  - **Supervised**: considers class (or other attribute) information for generating the intervals
  - **Unsupervised**: does not considers class (or other attribute) information for generating the intervals

Proposed method

- **Types of discretization**
  - **Univariate**: quantifies one continuous attribute at a time
  - **Multivariate**: considers simultaneously multiple attributes
  - **Supervised**: considers class (or other attribute) information for generating the intervals
  - **Unsupervised**: does not considers class (or other attribute) information for generating the intervals

Proposed method

- **CBD discretization method**

  - **Multivariate**

    Clustering based method

  - **Supervised**

    Considers consequent part of the rule, the class

Proposed method

- **Attributes' selection**
  - CARs have the consequent part formed only by the class attribute
  - For the antecedent part the selected attributes are the most influential in the prediction of the classes
  - The selection is based on the purity measure. It informs about how well the attributes discriminate the classes. It is based on the amount of information (entropy) that the attribute provides:

$$\mathrm{I}\,(P(c_1), \ldots, P(c_n)) = \sum_{i=1}^{n} - P(c_i) \log_n P(c_i)$$

where $P(c_i)$ is the probability of the class i and n is the number of classes

Proposed method

CBD discretization algorithm

Clusters of similar records are built giving more weight to the class attribute. This is a supervised way to obtain the best intervals for classification, according the following procedure:

- # intervals = # clusters
- Initial interval boundaries:
  - $(m - \sigma)$, $(m + \sigma)$
- For adjacent intervals 1 and 2:
  - If $(m_1 > m_2 - \sigma_2)$ or $(m_2 < m_1 + \sigma_1)$
    - Two intervals are merged into one: $(m_1 - \sigma_1)$, $(m_2 + \sigma_2)$
  - else
    - Cut point between intervals 1 and 2: $(m_2 - \sigma_2 + m_1 + \sigma_1)/2$

Contents

- Introduction
- Related work
- Proposed method
- **Experimental study**
- Results
- Conclusions

Experimental study

- **Objective**

  To estimate the final software size from some project attributes that can be obtained early in the life cycle

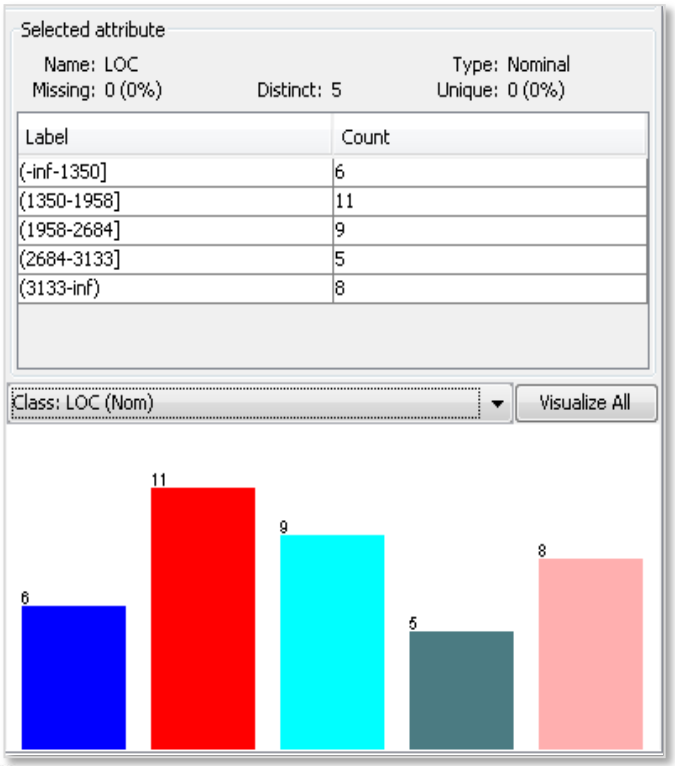
- **Proposed method**
  - Search for the best attributes for classification by calculating their cumulative purity
  - Discretization of continuous attributes by the **CBD algorithm**
  - Application of an associative classification method

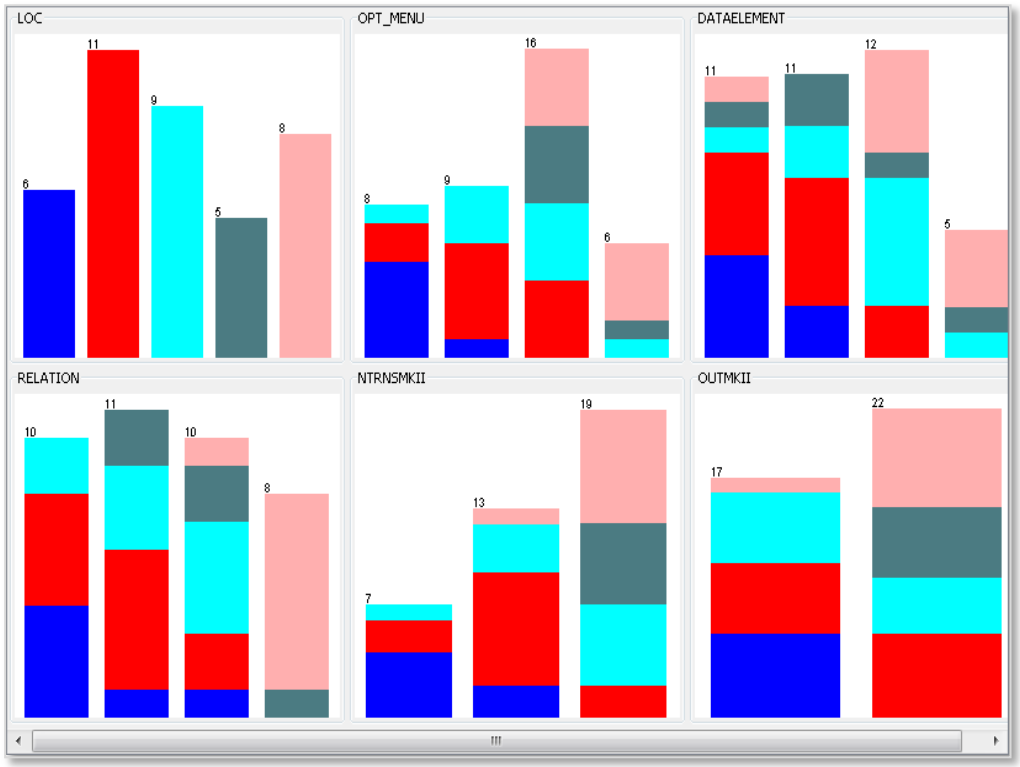
Experimental study

- **Dataset**
  - The data comes from 47 academic projects in which students developed accounting information systems
  - Class attribute
    - ***LOC*** : *Lines of Code*
  - Descriptive attributes
    - ***NOC-MENU***: total number of menu components
    - ***NOC-INPUT*** : total number of input components
    - ***NOC-RQ***: total number of report/query components
    - ***OPT-MENU*** : total number of menu choices
    - ***DATAELEMENT*** : total number of data elements
    - ***RELATION*** : total number of relations

Experimental study

- **Attribute discretization by means of the CBD algorithm**





Experimental study

- **Associative classification**
  - Application of CMAR with data discretized by means of four different algorithms
    - Equal width
    - Equal frequency
    - Fayyad and Irani method
    - CBD algorithm
- **Classical classification**
  - Applied methods
    - Bayes Net
    - Decision tree J4.8
    - Two multiclassifiers: Bagging with RepTree and Staking with CeroR

Contents

- Introduction
- Related work
- Proposed method
- Experimental study
- **Results**
- Conclusions

Results

- **Classification methods**

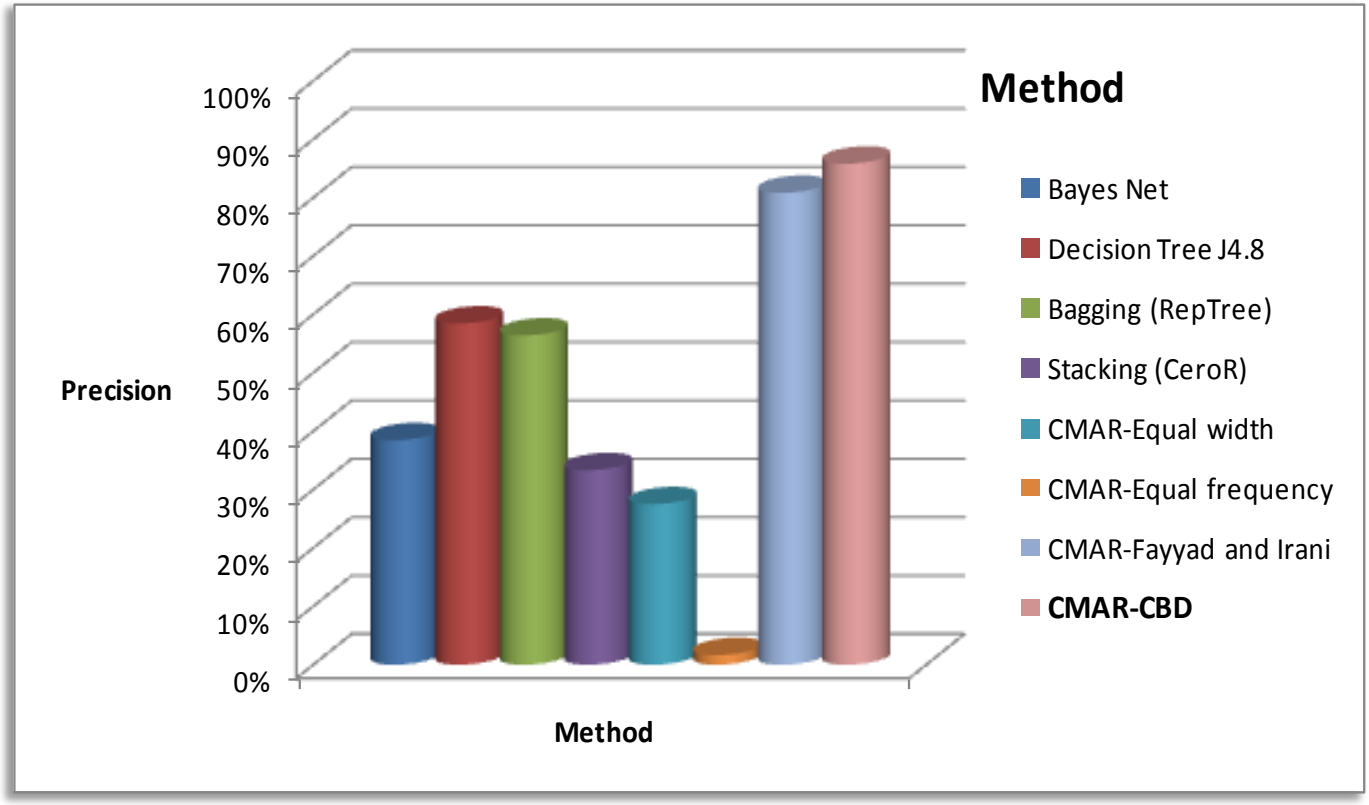
CLASSIFICATION METHOD	PRECISION
Bayes Net	38.46%
Decision Tree J4.8	58.97%
Bagging (RepTree)	56.41%
Stacking (CeroR)	33.33%

- **CMAR: Associative classification method**

DISCRETIZATION METHOD	PRECISION
Equal width	27.50%
Equal frequency	1.67%
Fayyad and Irani	80.83%
CBD	**85.83%**

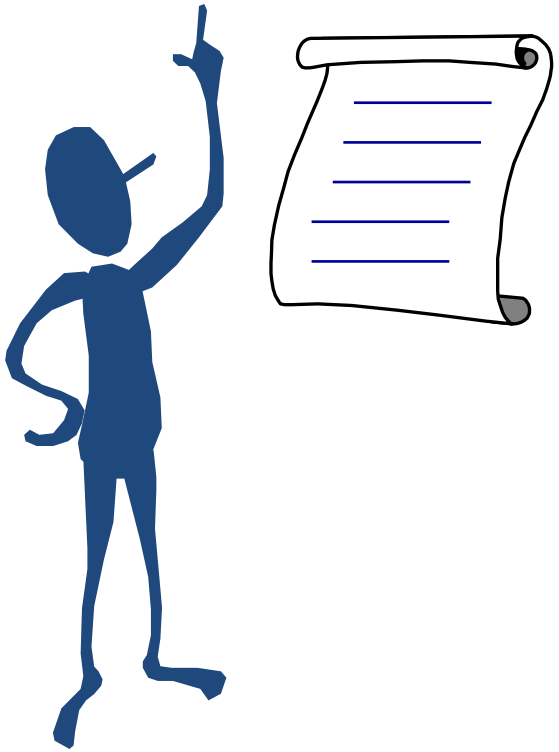
Results

- **Graphical representation**



Contents

- Introduction
- Related work
- Proposed method
- Experimental study
- Analysis of results
- **Conclusions**



Conclusions

- Data sparsity is one of the factors that produce the worst negative effects on the precision of machine learning methods
- Associative classification methods are less susceptible to sparsity but they have de drawback of working with discrete attributes
- In this work the CBD supervised multivariate discretization procedure is presented
- We have demostrated that the combination of the CMAR associative classification method with the CBD algorithm yields significantly better precision values than other classification methods in the project management field

THANKS FOR YOUR ATTENTION !

Multivariate discretization for associative classification in a sparse data application domain

María N. Moreno*, Joel Pinho Lucas, Vivian F. López and M. José Polo

***mmg@usal.es**

Dept. of Computing and Automatic

VNiVERSiDAD D SALAMANCA