APRENDIZAJE DE REGLAS SUPERVISADO DESCRIPTIVO

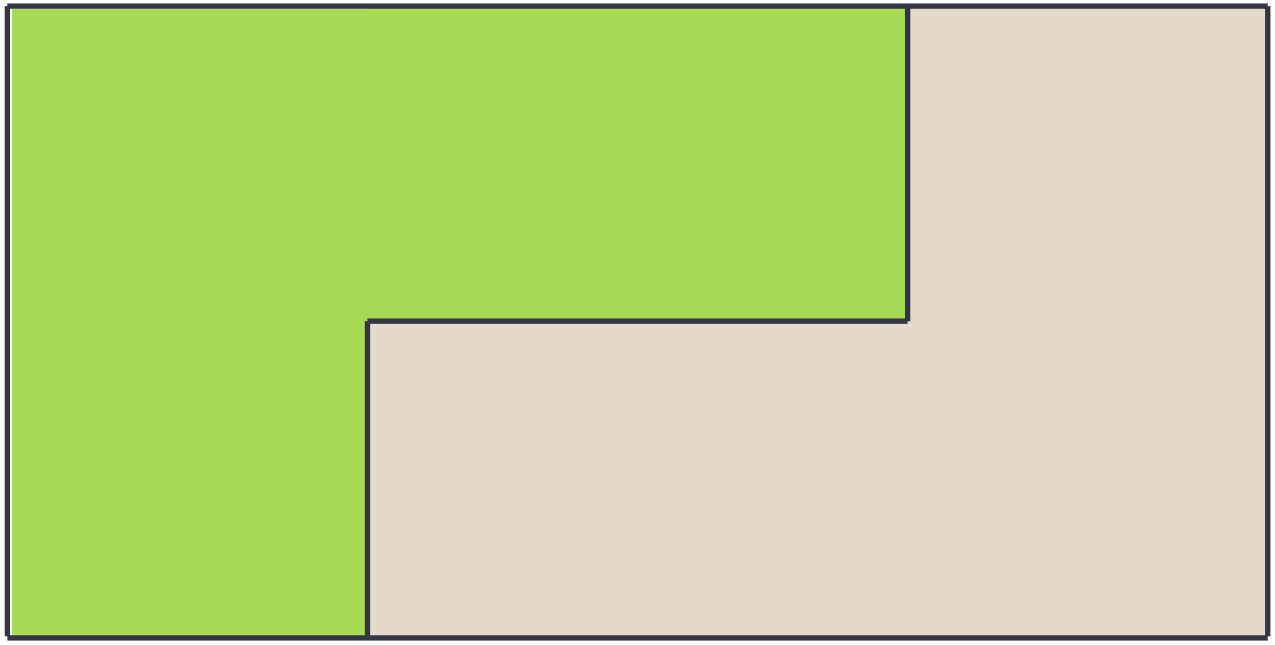
Minería de Datos: Aspectos Avanzados

Salvador García
salvagl@decsai.ugr.es

1. Data Mining

Machine learning task where data are labelled

Supervised learning



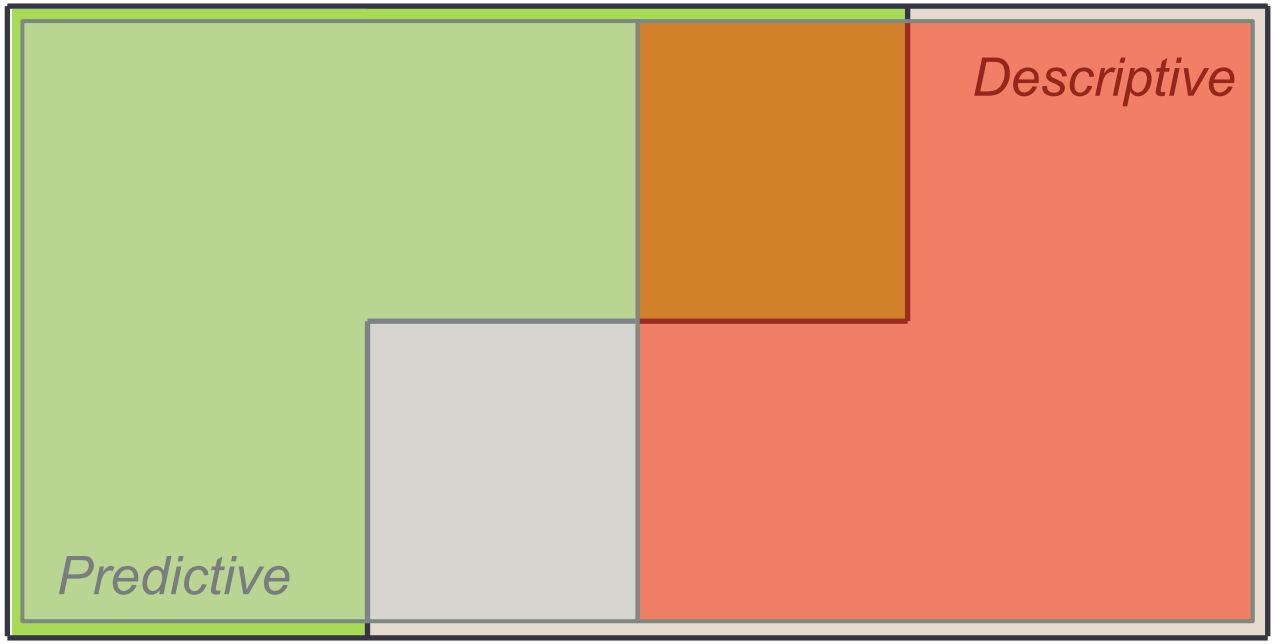
In this way, data are unlabelled and we can't evaluate the OUTPUT

Unsupervised learning

1. Data Mining

Supervised learning

Analysis relations between unlabeled data in order to obtain knowledge.

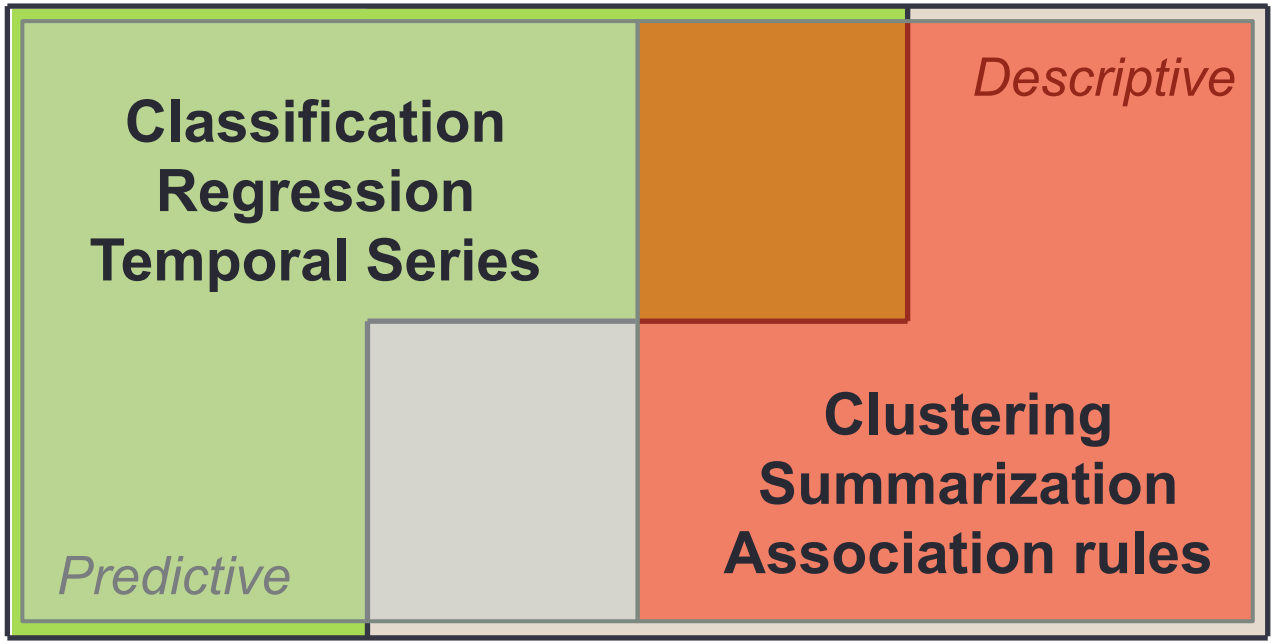


Search for relations between the OUTPUT in a sample and one or more known INPUT attributes or features of the unit.

Unsupervised learning

1. Data Mining

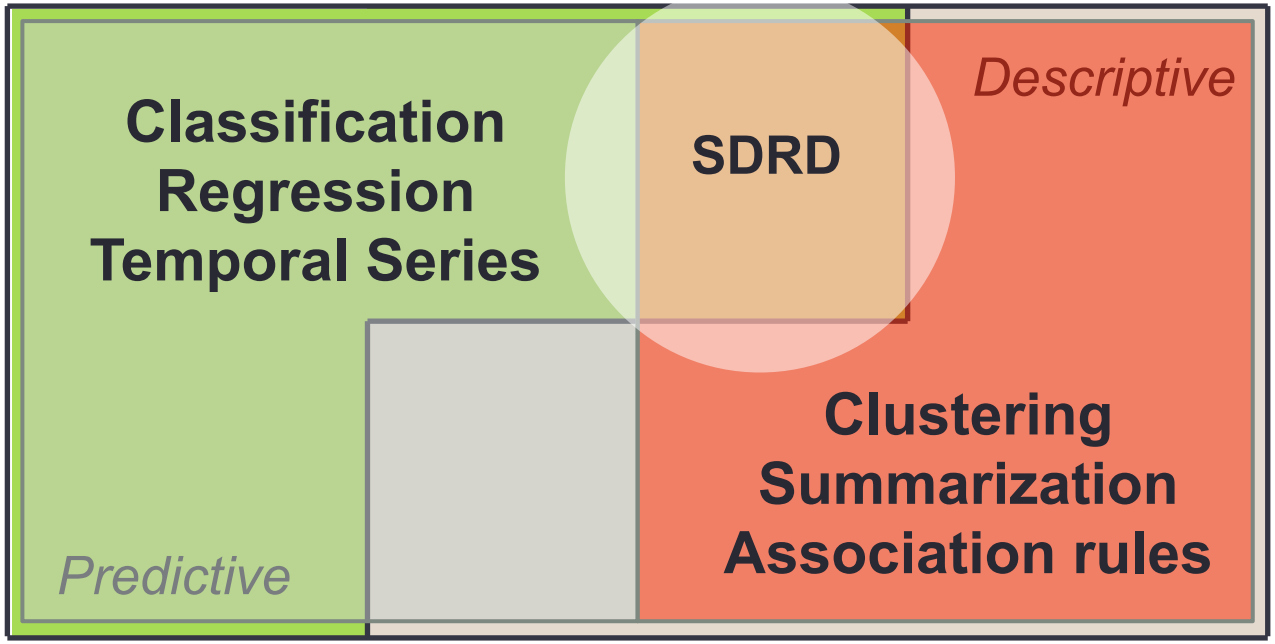
Supervised learning



Unsupervised learning

1. Data Mining

Supervised learning



Unsupervised learning

2. Supervised Descriptive Rule Learning

1. Definition
2. Contextualization
3. Main elements
4. Quality measures
5. Models and taxonomy
6. Applications

2.1. Definition

- **W. Klösgen, 1996:**

  "Given a population of individuals and a property of those individuals we are interested in, find population subgroups that are statistically 'most interesting', e.g., are as large as possible and have the most unusual statistical characteristics with respect to the property of interest".

W. Klösgen, Explora: A multipattern and multistrategy discovery assistant, Advance in Knowledge Discovery and Data Mining, MIT Press, 249-271, 1996.

S. Wrobel, An algorithm for multi-relational discovery of subgroups, Proc. 1st European Symposium Principles Data Mining Discovery (PKDD'97), Berlin (Germany), 1997, pp.78-87

2.1. Definition

- The representation of an induced subgroup (*R*) description has the form of an implication:

$$R : Cond \rightarrow Target_{value}$$

  - In the consequent part of the rule appears the value of the property of interest for the subgroup discovery task
  - The antecedent part of the rule is expressed by a conjunction of features selected from the features describing the training instances

2.1. Definition

- SD attempts to discover interesting relations between different properties of a set with respect to a target variable:
  - It is not necessary to obtain complete relations but rather partial ones.
- **So, the objective is to identify subgroups within the dataset whose behaviour is statistically different from that of the complete dataset.**
  - Smoker = true AND Family = positive → Heart_disease = true
    - *The percentage of smokers with positive family history, suffering coronary diseases is significantly different (in this case upper) to that of the complete set of individuals.*

2. Supervised Descriptive Rule Learning

1. Definition
2. Contextualization
3. Main elements
4. Quality measures
5. Models and taxonomy
6. Applications

2.2. Contextualization

How can we situate subgroup discovery in data mining?

- Classification

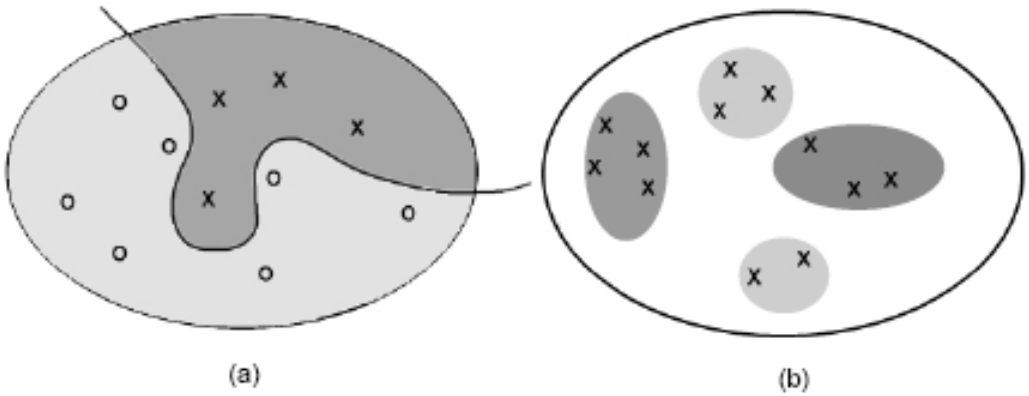
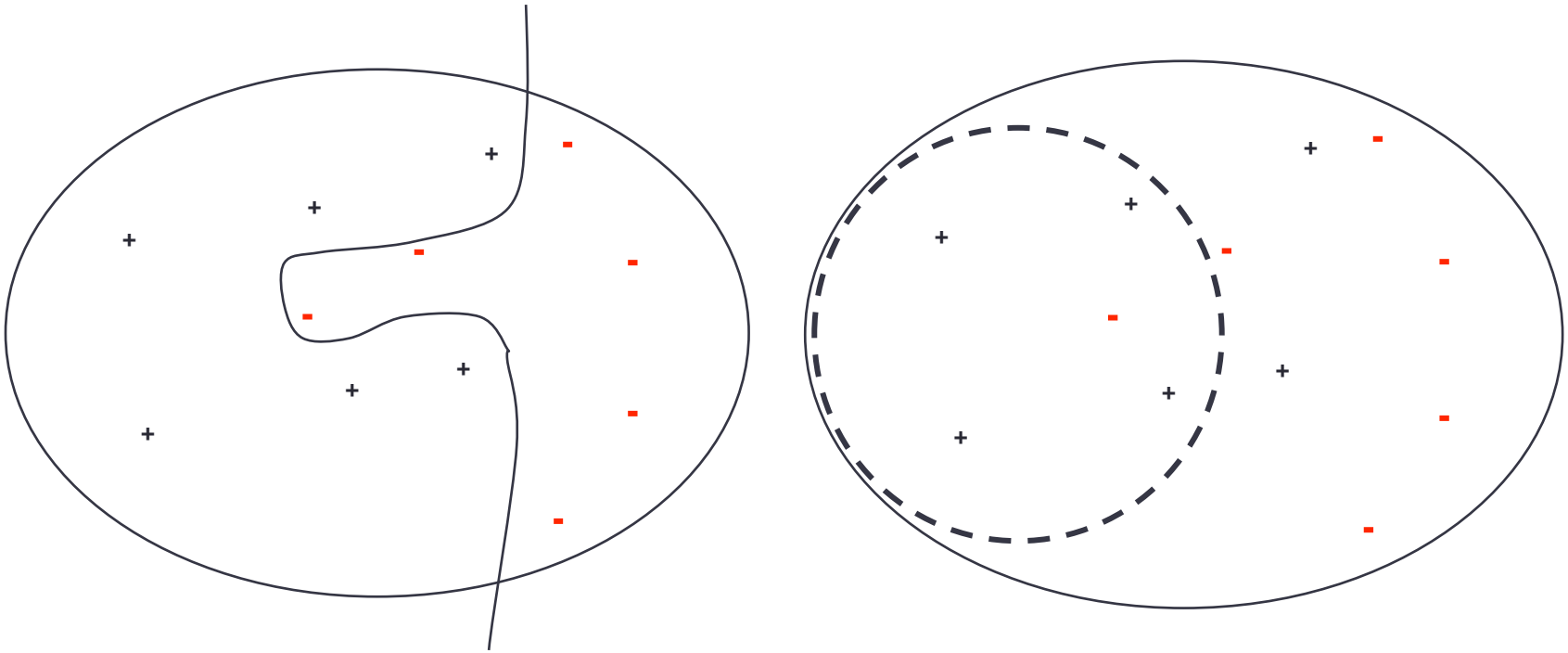


Fig. 1 Models of different techniques of knowledge discovery of databases

- Descriptive induction

2.2. Contextualization

Classification vs. Subgroup discovery



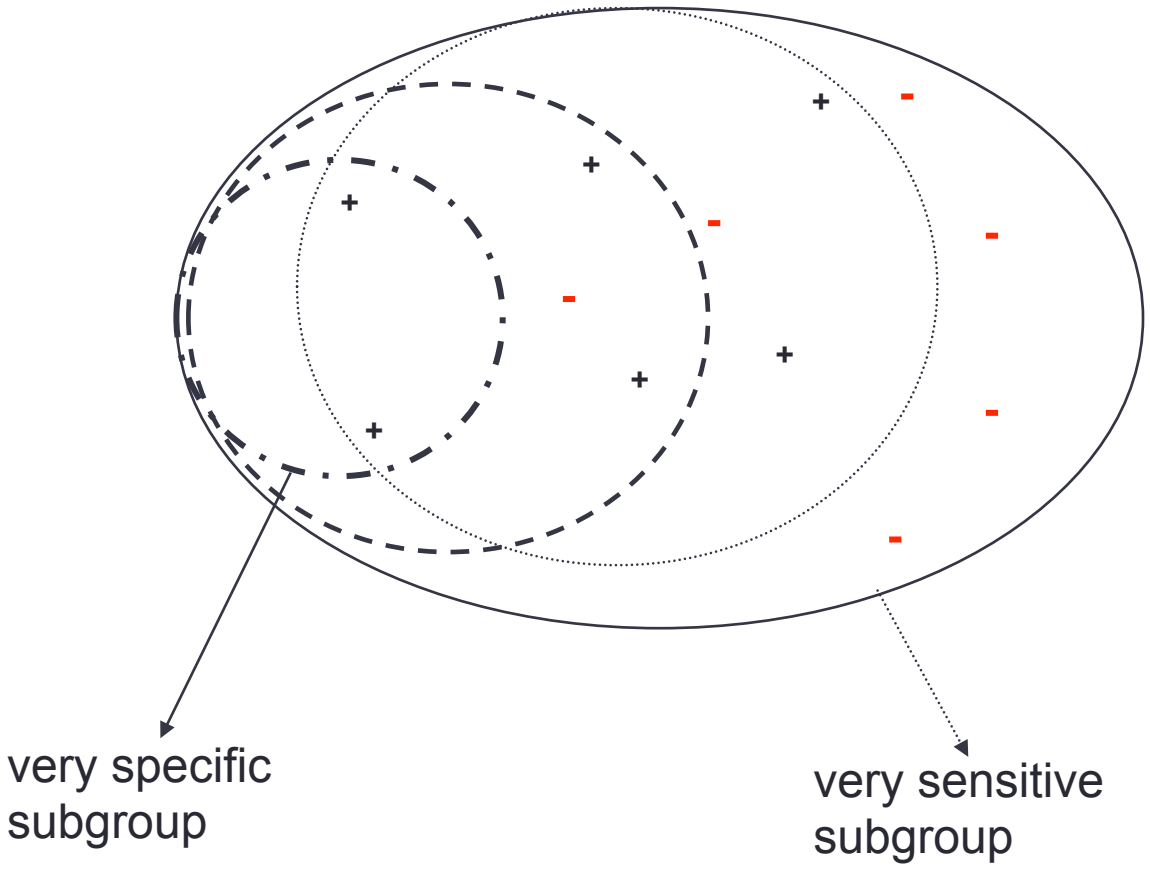
2.2. Contextualization

Classification vs. Subgroup discovery

- **Classification**
  - predictive induction
  - constructing sets of classification rules
  - aimed at learning a model for classification or prediction
  - rules are dependent

- **Subgroup Discovery**
  - descriptive induction
  - constructing individual subgroup-describing rules
  - aimed at finding interesting patterns in target class examples

2.2. Contextualization



generality – the main parameter of the subgroup induction process

2.2. Contextualization

- Interest for SDRD:
  - As large as possible
  - Significant
  - Non-redundant
  - Surprising to the user
  - **Simple**
  - Useful – actionable
  - **Good tradeoff generality and precision**

Overview on evolutionary subgroup discovery: analysis of the suitability and potential of the search performed by evolutionary algorithms, Carmona, C.J., González P., del Jesus M.J., and Herrera F. , WIREs Data Mining and Knowledge Discovery, Volume 4, Number 2, p.87-103, (2014)

2. Supervised Descriptive Rule Learning

1. Definition
2. Contextualization
3. Main elements
4. Quality measures
5. Models and taxonomy
6. Applications

2.3. Main Elements

- Type of the target variable: binary, nominal or numerical
- Description language
  - The representation of the rules must be suitable for obtaining interesting rules
  - The rules must be simple → usually represented as attribute-value pairs in conjunctive or disjunctive normal form
  - The values of the variables can be represented as positive or negative, through fuzzy logic, through the use of inequality or equality, etc.

2.3. Main Elements

- Quality measures, used in the data mining process and in the evaluation of the knowledge extracted
- Search strategy: beam search, evolutionary algorithms, search in multirelational spaces…
  - The dimension of the search space has an exponential relation to the number of features and values considered

2. Supervised Descriptive Rule Learning

1. Definition
2. Contextualization
3. Main elements
4. Quality measures
5. Models and taxonomy
6. Applications

2.4. Quality measures

- There is no current consensus in the choice of the quality measures more suitable to extract and evaluate knowledge in subgroup discovery
- There are a wide number of measures presented throughout the bibliography, classified on:

  - Measures of complexity
  - Measures of generality
  - Measures of precision
  - Measures of interest
  - Hybrid measures

An overview on Subgroup Discovery: Foundations and Applications, Herrera, F., Carmona C.J., González P., and del Jesus M.J. , Knowledge and Information Systems, Volume 29, Number 3, p. 495-525, (2011)

2. Supervised Descriptive Rule Learning

1. Definition
2. Contextualization
3. Main elements
4. Quality measures
5. Models and taxonomy
6. Applications

2.5. Models and taxonomy

Extensions of classification algorithms
EXPLORA [70]
MIDOS [108]
SubgroupMiner [72]
SD [43]
CN2-SD [85]
RSD [83, 112]
Extensions of association algorithms
APRIORI-SD [66, 68]
SD4TS [92]
SD-MAP [10]
DpSubgroup [55]
Merge-SD [54]
IMR [20]
Evolutionary algorithms
SDIGA [61]
MESDIF [18, 60]
NMEEF-SD [27, 28]

An overview on Subgroup Discovery: Foundations and Applications, Herrera, F., Carmona C.J., González P., and del Jesus M.J. , Knowledge and Information Systems, Volume 29, Number 3, p. 495-525, (2011)

2. Supervised Descriptive Rule Learning

1. Definition
2. Contextualization
3. Main elements
4. Quality measures
5. Models and taxonomy
6. Applications

2.6. Applications

Field	*Application*	*References*
Medical domain	Detection of risk groups with coronary heart disease	[43–45, 48, 49, 80, 84]
	Brain ischaemia	[47, 52, 76]
	Cervical cancer	[103]
	Psychiatric emergency	[29]
Bioinformatic domain	Leukemia cancer	[50, 104, 105, 115]
	Leukemia$_{improved}$ cancer	[106]
	Subtypes of leukemia ALL cancer	[106]
	Cancer diagnosis	[46, 50, 80, 104–106, 115]
	Interpreting positron emission tomography (PET) scans	[99]
Marketing	Financial	[71]
	Comercial	[44, 84]
	Planning trade fairs	[18, 61]
Learning	e-learning	[30, 96]
Spatial subgroup mining	Demographic data	[6]
	Census data	[72, 73, 75]
	Vegetation data	[90]
Others	Traffic accidents	[64, 65, 67]
	Production control	[71]
	Mutagenesis	[86, 113, 114]
	Social data	[79]
	Voltage sag source location	[16]

An overview on Subgroup Discovery: Foundations and Applications, Herrera, F., Carmona C.J., González P., and del Jesus M.J. , Knowledge and Information Systems, Volume 29, Number 3, p.495-525, (2011)