Highlight sub group intro

Highlight sub group definition

Highlight subgroup examples

Highlight subgroup measures

Highlight sub group algotithm

Highlight sub group pre process variables

Highlight references

**Notes in ‘An overview on subgroup discovery- foundations and applications’**

**Notes in Document**

**'An overview on subgroup discovery- foundations and applications':**

Highlight : Subgroup discovery is a data mining technique which extracts interesting rules with respect to a target variable. An important characteristic of this task is the combination of predictive and descriptive induction. *\cite{herrera2011overview}, p.1)*

Highlight : Subgroup discovery [70,108] is a broadly applicable data mining technique aimed at dis- covering interesting relationships between different objects in a set with respect to a specific property which is of interest to the user the target variable. The patterns extracted are normally represented in the form of rules and called subgroups *\cite{herrera2011overview}, p.1)*

Highlight : example, predictive techniques maximise accuracy in order to correctly classify new objects, and descriptive techniques simply search for relations between unlabelled objects. The need for obtaining simple models with a high level of interest led to statistical techniques which search for unusual relations [ *\cite{herrera2011overview}, p.2)*

Highlight : subgroup discovery is somewhere halfway between supervised and unsuper- vised learning [78]. It can be considered that subgroup discovery lies between the extraction of association rules and the obtaining of classification rules.  *\cite{herrera2011overview}, p.2)*

Highlight : 2.1 Definition of subgroup discovery  *\cite{herrera2011overview}, p.2)*

Highlight : The concept of subgroup discovery was initially introduced by Kloesgen [70] and Wrobel [108], and more formally defined by Siebes [101] but using the name Data Surveying for the discovery of interesting subgroups. It can be defined as [109]: In subgroup discovery, we assume we are given a so-called population of individuals (objects, customer, ...) and a property of those individuals we are interested in. The task of subgroup discovery is then to discover the subgroups of the population that are statistically “most interesting”, i.e. are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest. *\cite{herrera2011overview}, p.2)*

Highlight : Subgroup discovery attempts to search relations between different properties or variables of a set with respect to a target variable. Due to the fact that subgroup discovery is focused in the extraction of relations with interesting characteristics, it is not necessary to obtain complete but partial relations. These relations are described in the form of individual rules. Then, a rule (R), which consists of an induced subgroup description, can be formally defined as [43,84]: R : Cond → T argetvalue where T argetvalue is a value for the variable of interest (target variable) for the subgroup discovery task (which also appears as Class in the literature), and Cond is commonly a conjunction of features (attribute-value pairs) which is able to describe an unusual statistical distribution with respect to the T argetvalue.  *\cite{herrera2011overview}, p.2)*

Highlight : As an example, let D be a data set with three variables Age = {Less than 25, 25 to 60, More than 60}, Sex = {M, F} and Country = {Spain, USA, France, German}, and  *\cite{herrera2011overview}, p.2)*

Highlight : a variable of interest target variable Money = {Poor, Normal, Rich}. Some possible rules containing subgroup descriptions are: R1 : (Age = Less than 25 AND Country = German) → Money = Rich R2 : (Age = More than 60 AND Sex = F) → Money = Normal where rule R1 represents a subgroup of German people with less than 25 years old for which the probability of being rich is unusually high with respect to the rest of the population, and rule R2 represents that women with more than 60 years old are more likely to have a normal economy than the rest of the population.  *\cite{herrera2011overview}, p.3)*

Highlight : Data mining is a stage of the Knowledge Discovery in Databases defined as “the non-trivial extraction of implicit, unknown, and potentially useful information from data *\cite{herrera2011overview}, p.3)*

Highlight : Descriptive induction, whose main objective is the extraction of interesting knowledge from the data. Its features include association rules [2], summarisation [116] or subgroup discovery [70,108] can be mentioned.  *\cite{herrera2011overview}, p.3)*

Highlight : Subgroup discovery [70] is a technique for the extraction of patterns, with respect to a property of interest in the data, or target variable. This technique is somewhere halfway between predictive and descriptive induction, and its goal is to generate in a single and inter- pretable way subgroups to describe relations between independent variables and a certain value of the target variable. The algorithms for this task must generate subgroups for each value of the target variable *\cite{herrera2011overview}, p.3)*

Highlight : Currently, several techniques lie halfway between descriptive and predictive data mining. “Supervised Descriptive Rule Induction” [78] is a new recently proposed paradigm which includes techniques combining the features of both types of induction, and its main objective is to extract descriptive knowledge from the data of a property of interest. These techniques use supervised learning to solve descriptive tasks. Within this new paradigm, the following data mining techniques are included: – Subgroup Discovery [70,108], defined as the extraction of interesting subgroups for a target value. – Contrast Set Mining [17], defined as “a conjunction of attribute-value pairs defined on groups with no attribute occurring more than once”. – Emerging Pattern Mining [38] defined as “patterns whose frequencies in two classes differ by a large ratio”.  *\cite{herrera2011overview}, p.4)*

Highlight : Different elements can be considered the most important when a subgroup discovery approach must be applied. These elements are defined below [13] *\cite{herrera2011overview}, p.4)*

Highlight : Type of the target variable Different types for the variable can be found: binary, nominal or numeric. For each one, different analyses can be applied considering the target variable as a dimension of the reality to study. – Binary analysis. The variables have only two values (True or False), and the task is focused on providing interesting subgroups for each of the possible values. – Nominal analysis. The target variable can take an undetermined number of values, but the philosophy for the analysis is similar to the binary, to find subgroups for each value. – Numeric analysis. This type is the most complex because the variable can be studied different ways such as dividing the variable in two ranges with respect to the average, discretisising the target variable in a determined number of intervals [91], or searching for significant deviations of the mean among others.  *\cite{herrera2011overview}, p.5)*

Highlight : Description language The representation of the subgroups must be suitable for obtaining interesting rules. These rules must be simple and therefore are represented as attribute- value pairs in conjunctive or disjunctive normal form in general. *\cite{herrera2011overview}, p.5)*

Highlight : Quality measures These are a key factor for the extraction of knowledge because the interest obtained depends directly on them. *\cite{herrera2011overview}, p.5)*

Highlight : Search strategy This is very important, since the dimension of the search space has an exponential relation to the number of features and values considered. Different strate- gies have been used up to the moment, for example beam search, evolutionary algo- rithms, search in multi-relational spaces, etc. *\cite{herrera2011overview}, p.5)*

Highlight : One of the most important aspects in subgroup discovery is the choice of the quality measures employed to extract and evaluate the rules *\cite{herrera2011overview}, p.5)*

Highlight : The most common quality measures used in subgroup discovery are described here, classified by their main objective such as complexity, generality, precision, and interest.  *\cite{herrera2011overview}, p.5)*

Highlight : described. There are several proposals of algorithms for subgroup discovery. To classify these algorithms, it can be distinguished between extensions of classification algorithms, extensions of association *\cite{herrera2011overview}, p.8)*

Highlight : 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]  *\cite{herrera2011overview}, p.9)*

Highlight : algorithms and evolutionary fuzzy systems *\cite{herrera2011overview}, p.9)*

Highlight : The first algorithms developed for subgroup discovery—EXPLORA and MIDOS—are exten- sions of classification algorithms and use decision trees *\cite{herrera2011overview}, p.10)*

Highlight : EXPLORA [70] was the first approach developed for subgroup discovery. It uses decision trees for the extraction of rules. *\cite{herrera2011overview}, p.10)*

Highlight : MIDOS [108] applies the EXPLORA approach to multi-relational databases. The goal is to discover subgroups of the target variable (defined as first-order conjunctions) that have an unusual statistical distribution with respect to the complete population *\cite{herrera2011overview}, p.10)*

Highlight : SubgroupMiner [72] is an extension of EXPLORA and MIDOS. It is an advanced sub- group discovery system that uses decision rules and interactive search in the space of the solutions, *\cite{herrera2011overview}, p.11)*

Highlight : SD [43] is a rule induction system based on a variation of the beam search algorithms  *\cite{herrera2011overview}, p.11)*

Highlight : CN2-SD [84] is a subgroup discovery algorithm obtained by adapting a standard classi- fication rule learning approach CN2 [32,33] to subgroup discovery *\cite{herrera2011overview}, p.11)*

Highlight : RSD (Relational subgroup discovery) [83,112] has the objective of obtaining population subgroups which are as large as possible, with a statistical distribution as unusual as possible with respect to the property of interest, and different enough to cover most of the target population. It is an upgrade of the CN2-SD algorithm which enables relational subgroup discovery. *\cite{herrera2011overview}, p.12)*

Highlight : SubgroupMiner is the first algorithm which considers the use of numerical target variables, though it is necessary to perform a previous discretisation. *\cite{herrera2011overview}, p.12)*

Highlight : An association rule algorithm attempts to obtain relations between the variables of the data set. In this case, several variables can appear both in the antecedent and consequent of the rule. In contrast, in subgroup discovery the consequent of the rule, consisting of the property of interest is prefixed *\cite{herrera2011overview}, p.12)*

Highlight : APRIORI-SD [66,68] is developed by adapting to subgroup discovery the classifica- tion rule learning algorithm APRIORI-C [63], *\cite{herrera2011overview}, p.12)*

Highlight : SD4TS [92] is an algorithm based on APRIORI-SD but using the quality of the subgroup to prune the search space *\cite{herrera2011overview}, p.13)*

Highlight : SD-Map [10] is an exhaustive subgroup discovery algorithm that uses the well-known FP-growth method [56] for mining association rules with adaptations for the subgroup discovery task. *\cite{herrera2011overview}, p.13)*

Highlight : DpSubgroup [55] is a subgroup discovery algorithm that uses a frequent pattern tree to obtain the subgroups *\cite{herrera2011overview}, p.13)*

Highlight : Merge-SD [54] is a subgroup discovery algorithm that prunes large parts of the search space by exploiting bounds between related numerical subgroup descriptions. In this way, the algorithm can manage data sets with numeric attributes. *\cite{herrera2011overview}, p.13)*

Highlight : IMR [20] is an alternative algorithmic approach for the discovery of non-redundant sub- groups based on a breadth-first strategy. *\cite{herrera2011overview}, p.13)*

Highlight : Some of these algorithms like APRIORI-SD or SD4TS are obtained from the adaptation to subgroup discovery of the association rule learner algorithm APRIORI, but others like SD-MAP, DpSubgroup or Merge-SD are adaptations of FP-Growth. All of them use decision trees for representation. Only Merge-SD and SD-MAP can handle numeric or continuous variables, and for the rest a previous discretisation is necessary. *\cite{herrera2011overview}, p.13)*

Highlight : Evolutionary algorithms for extracting subgroups Subgroup discovery is a task which can be approached and solved as optimisation and search problems. Evolutionary algorithms [102] imitate the principles of natural evolution in order to form searching processes *\cite{herrera2011overview}, p.14)*

Highlight : One of the most widely used types of evolutionary algo- rithms are genetic algorithms, inspired by natural evolution processes *\cite{herrera2011overview}, p.14)*

Highlight : The heuristic used by this type of algorithm is defined by a fitness function, which determines which individuals (rules in this case) will be selected to form part of the new population in the competition process. This makes genetic algorithms very useful for the subgroup discovery task *\cite{herrera2011overview}, p.14)*

Highlight : SDIGA [61] is an evolutionary fuzzy rule induction system *\cite{herrera2011overview}, p.14)*

Highlight : The algorithm evaluates the quality of the rules by means of a weighted average of the mea- sures selected. *\cite{herrera2011overview}, p.14)*

Highlight : – MESDIF [18,60] is a multi-objective genetic algorithm for the extraction of fuzzy rules which describe subgroups. The algorithm extracts a variable number of different rules expressing information on a single value of the target variable *\cite{herrera2011overview}, p.15)*

Highlight : NMEEF-SD [27,28] is an evolutionary fuzzy system whose objective is to extract descrip- tive fuzzy and/or crisp rules for the subgroup discovery task, depending on the type of variables present in the problem *\cite{herrera2011overview}, p.15)*

Highlight : The evolutionary algorithms proposed so far for the subgroup discovery task are based on the hybridisation between fuzzy logic and evolutionary algorithms, known as evolution- ary fuzzy systems [34,57]. *\cite{herrera2011overview}, p.15)*

Highlight : 5.2 Preprocessing of the variables  *\cite{herrera2011overview}, p.17)*

Highlight : It is very common that some of the variables collected in the data sets used to apply subgroup discovery techniques are continuous variables. Most of the subgroup discovery algorithms are not able to handle continuous variables. In this case, a previous discretisation can be applied using different mechanisms [40,88] *\cite{herrera2011overview}, p.17)*

Highlight : Some approaches can manage continuous variables in the condition of the rule (explaining variables) without the need of a previous discretisation. This is performed in the evolu- tionary algorithms proposed in [27,59–61] using fuzzy logic [111]. *\cite{herrera2011overview}, p.17)*

Highlight : A recent approach presented in [91], TargetCluster, discretises a continuous target variable before applying a subgroup discovery algorithm based on a clustering approach *\cite{herrera2011overview}, p.17)*

Highlight : In [107], a methodology for grouping different variables as a target variable was pre- sented. This proposal is based on clustering to separately find clusters as values of the target variable. *\cite{herrera2011overview}, p.17)*

Highlight : Using domain knowledge in data mining methods can improve the quality of data mining results [94]. In subgroup discovery, it can help to focus the search on the interesting subgroups related to the target variable by restricting the search space. Different approaches to include domain knowledge in subgroup discovery have been presented recently:  *\cite{herrera2011overview}, p.17)*

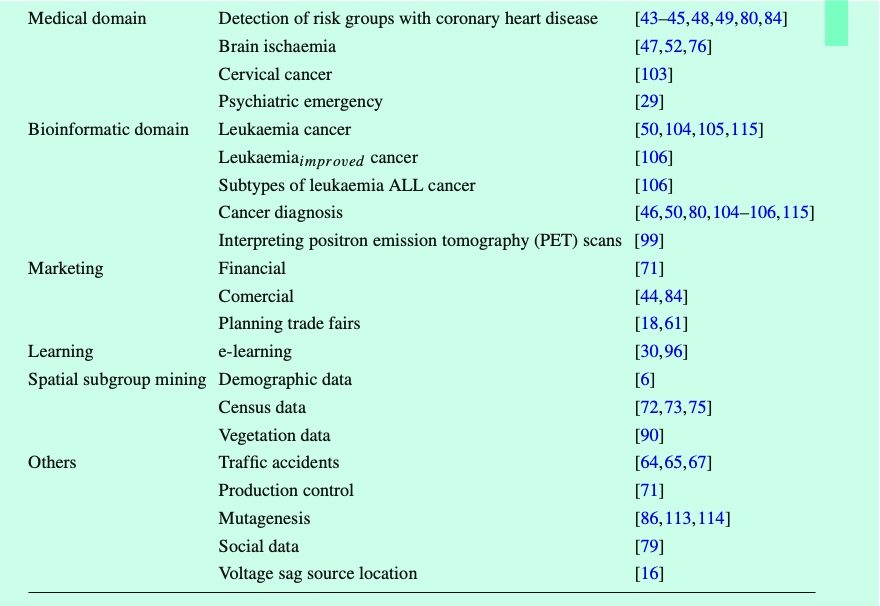
Highlight : Semantic Subgroup Discovery *\cite{herrera2011overview}, p.17)*

Highlight : Domain Knowledge *\cite{herrera2011overview}, p.17)*

Highlight : The visualisation of results is very important for the usability of the knowledge extracted. Subgroup discovery algorithms are often used to present the results to an expert, who will take decisions based on these data. – In [9], different visualisation methods supporting explorative and descriptive data mining were presented *\cite{herrera2011overview}, p.18)*

Highlight : – In [62,89], a visual interactive subgroup discovery procedure which allows the naviga- tion in the space of subgroups in a two-dimensional plot was shown. The authors used distribution rules for representing the knowledge.  *\cite{herrera2011overview}, p.18)*

Highlight : Table 7 summarises the real-world applications of algorithms of sub- group discovery.  *\cite{herrera2011overview}, p.18)*

Highlight  *\cite{herrera2011overview}, p.19)*

Highlight : This is an emergent field, and there are several open problems in subgroup discovery. An important problem to address is to determine which quality measures are more adapted both to evaluating the subgroups discovered and to guiding the search process. A wide number of measures have been used, but there is no current consensus in the field about which are the most suitable measures for both processes. *\cite{herrera2011overview}, p.23)*

Highlight : On the other hand, the discretisation of the continuous variables and its influence in the results of the subgroup discovery task is another open topic. It is unclear how the previous discretisation of continuous variables may affect the results of the subgroup discovery process, or the advantages of the subgroup discovery algorithms that use continuous variables without any prior discretisation *\cite{herrera2011overview}, p.23)*

Highlight : Another issue to be dealt with in more depth is the scalability of the subgroup discovery algorithms. Many of the subgroup discovery algorithms have a high computational cost when they are applied to large data sets. *\cite{herrera2011overview}, p.23)*

Highlight : Finally, the combination between the subgroup discovery task and other fields such as semantic data, contrast set, clustering, and so on is beginning to be used in this area. *\cite{herrera2011overview}, p.23)*

Highlight : 2. Agrawal R, Imieliski T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD international conference on management of data. ACM Press, pp 207–216 *\cite{herrera2011overview}, p.25)*

Highlight  *\cite{herrera2011overview}, p.25)*

Highlight : 9. Atzmueller M, Puppe F (2005) Semi-automatic visual subgroup mining using VIKAMINE. J Univers Comput Sci 11(11):1752–1765  *\cite{herrera2011overview}, p.25)*

Highlight : 10. Atzmueller M, Puppe F (2006) SD-Map—a fast algorithm for exhaustive subgroup discovery. In: Proceedings of the 17th European conference on machine learning and 10th European conference on principles and practice of knowledge discovery in databases, vol 4213. Springer, LNCS, pp 6–17 *\cite{herrera2011overview}, p.26)*

Highlight : 18. Berlanga FJ, del Jesus MJ, González P, Herrera F, Mesonero M (2006) Multiobjective evolutionary induction of subgroup discovery fuzzy rules: a case study in marketing. In: Proceedings of the 6th industrial conference on data mining, vol 4065. Springer, LNCS, pp 337–349 *\cite{herrera2011overview}, p.26)*

Highlight : 20. Boley M, Grosskreutz H (2009) Non-redundant subgroup discovery using a closure system. In: Proceedings of the European conference on machine learning and principles and practice of knowl- edge discovery in databases, vol 5781. Springer, LNAI, pp 179–194 *\cite{herrera2011overview}, p.26)*

Highlight : 27. Carmona CJ, González P, del Jesus MJ, Herrera F (2009b) Non-dominated multi-objective evolutionary algorithm based on fuzzy rules extraction for subgroup discovery. In: Proceedings of the 4th international conference on hybrid artificial intelligence systems, vol 5572. Springer, LNAI, pp 573–580 *\cite{herrera2011overview}, p.26)*

Highlight : 28. Carmona CJ, González P, del Jesus MJ, Herrera F (2010a) NMEEF-SD: Non-dominated multi-objec- tive evolutionary algorithm for extracting fuzzy rules in subgroup discovery. IEEE Trans Fuzzy Syst 18(5):958–970 *\cite{herrera2011overview}, p.26)*

Highlight : 40. Fayyad UM, Irani KB (1993) Multi-interval discretization of continuous-valued attributes for classifi- cation learning. In: 13th International joint conference on artificial intelligence, pp 1022–1029  *\cite{herrera2011overview}, p.27)*

Highlight : 41. Fayyad UM, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery: an overview. In: Advances in knowledge discovery and data mining. AAAI/MIT Press, pp 1–34  *\cite{herrera2011overview}, p.27)*

Highlight : 43. Gamberger D, Lavrac N (2002) Expert-guided subgroup discovery: methodology and application. J Artif Intell Res 17:501–527 *\cite{herrera2011overview}, p.27)*

Highlight : 54. Grosskreutz H, Rueping S (2009) On subgroup discovery in numerical domains. Data Mining Knowl Discov 19(2):210–216 *\cite{herrera2011overview}, p.27)*

Highlight : 55. Grosskreutz H, Rueping S, Wrobel S (2008) Tight optimistic estimates for fast subgroup discovery. In: European conference on machine learning and principles and practice of knowledge discovery in databases, pp 440–456 *\cite{herrera2011overview}, p.27)*

Highlight : 59. del Jesus MJ, González P, Herrera F (2007) Fuzzy sets and their extensions: representation, aggregation and models, vol 220, Springer, chap Subgroup Discovery with Linguistic Rules, pp 411–430 *\cite{herrera2011overview}, p.27)*

Highlight : 60. del Jesus MJ, González P, Herrera F (2007) Multiobjective genetic algorithm for extracting subgroup discovery fuzzy rules. In: Proceedings of the IEEE symposium on computational intelligence in multi- criteria decision making. IEEE Press, pp 50–57 *\cite{herrera2011overview}, p.27)*

Highlight : 61. del Jesus MJ, González P, Herrera F, Mesonero M (2007) Evolutionary fuzzy rule induction process for subgroup discovery: a case study in marketing. IEEE Trans Fuzzy Syst 15(4):578–592 *\cite{herrera2011overview}, p.28)*

Highlight : 66. Kavsek B, Lavrac N (2006) APRIORI-SD: adapting association rule learning to subgroup discovery. Appl Artif Intell 20:543–583 *\cite{herrera2011overview}, p.28)*

Highlight : 68. Kavsek B, Lavrac N, Jovanoski V (2003) APRIORI-SD: adapting association rule learning to subgroup discovery. In: Proceedings of the 5th international symposium on intelligent data analysis, vol 2810. Springer, LNCS, pp 230–241 *\cite{herrera2011overview}, p.28)*

Highlight : 70. Kloesgen W (1996) Explora: a multipattern and multistrategy discovery assistant. In: Advances in Knowledge discovery and data mining. American Association for Artificial Intelligence, pp 249–271 *\cite{herrera2011overview}, p.28)*

Highlight : 72. Kloesgen W, May M (2002) Census data mining—an application. In: Proceedings of the 6th European conference on principles of data mining and knowledge discovery, pp 65–79 *\cite{herrera2011overview}, p.28)*

Highlight : 83. Lavrac N, Zelezny F, Flach PA (2003) RSD: relational subgroup discovery through first-order feature construction. In: Proceedings of the 12th international conference inductive logic programming, vol 2583. Springer, LNCS, pp 149–165 *\cite{herrera2011overview}, p.28)*

Highlight : 85. Lavrac N, Kavsek B, Flach PA, Todorovski L (2004) Subgroup discovery with CN2-SD. J Mach Learn Res 5:153–188 *\cite{herrera2011overview}, p.28)*

Highlight : 88. Liu H, Hussain F, Tan C, Dash M (2002) Discretization: an enabling technique. Data mining Knowl Discov 6:393–423  *\cite{herrera2011overview}, p.29)*

Highlight : 91. Moreland K, Truemper K (2009) Discretization of target attributes for subgroup discovery. In: Pro- ceedings of the 6th international conference machine learning and data mining in pattern recognition, vol 5632. Springer, LNAI, pp 44–52 *\cite{herrera2011overview}, p.29)*

Highlight : 92. Mueller M, Rosales R, Steck H, Krishnan S, Rao B, Kramer S (2009) Subgroup discovery for test selection: a novel approach and its application to breast cancer diagnosis. In: Proceedings of the 8th international symposium on intelligent data analysis, vol 5772. Springer, LNCS, pp 119–130 *\cite{herrera2011overview}, p.29)*

Highlight : 101. Siebes A (1995) Data Surveying: foundations of an inductive query language. In: Proceedings of the 1st international conference on knowledge discovery and data mining. AAAI Press, pp 269–274 *\cite{herrera2011overview}, p.29)*

Highlight : 107. Umek L, Zupan B, Toplak M, Morin A, Chauchat JH, Makovec G, Smrke D (2009) Subgroup dis- covery in data sets with multi-dimensional responses: a method and a case study in traumatology. In: Proceedings of the 12th conference on artificial intelligence in medicine, vol 5651. Springer, LNAI, pp 265–274 *\cite{herrera2011overview}, p.29)*

Highlight : 108. Wrobel S (1997) An algorithm for multi-relational discovery of subgroups. In: Proceedings of the 1st European symposium on principles of data mining and knowledge discovery, vol 1263. Springer, LNAI, pp 78–87 *\cite{herrera2011overview}, p.29)*

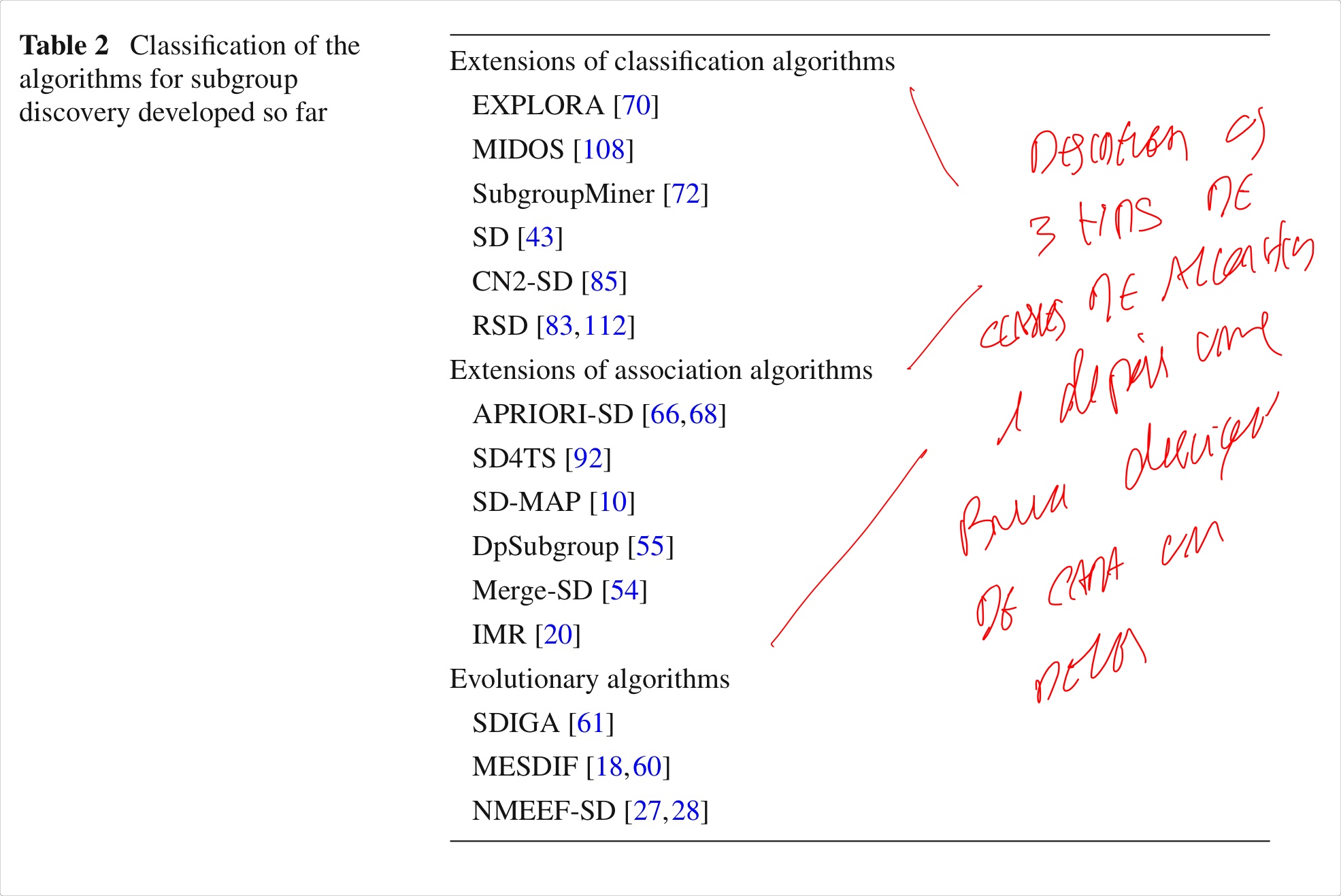
Highlight : 109. Wrobel S (2001) Inductive logic programming for knowledge discovery in databases. Springer, chap Relational Data Mining, pp 74–101 *\cite{herrera2011overview}, p.29)*

Highlight : 111. Zadeh LA (1975) The concept of a linguistic variable and its applications to approximate reasoning. Parts I, II, III. Inf Sci 8–9:199–249, 301–357, 43–80  *\cite{herrera2011overview}, p.30)*

Highlight : 112. Zelezny F, Lavrac N (2006) Propositionalization-based relational subgroup discovery with RSD. Machine Learning 62:33–63 *\cite{herrera2011overview}, p.30)*

Highlight : 116. Zembowicz R, Zytkow JM (1996) From contingency tables to various forms of knowledge in databases. In: Advances in knowledge discovery and data mining. AAAI/MIT Press, pp 329–349 *\cite{herrera2011overview}, p.30)*

**Notes in Workspace:**

Excerpt:  *\cite{herrera2011overview}, p.9)*

Excerpt: – MESDIF [18,60] is a multi-objective genetic algorithm for the extraction of fuzzy rules which describe subgroups. The algorithm extracts a variable number of different rules expressing information on a single value of the target variable *\cite{herrera2011overview}, p.15)*