Discovering Fuzzy and Statistical Patterns in Data: The `nuggets` R Package

BURDA Michal

University of Ostrava, Institute for Research and Applications of Fuzzy Modeling, NSC IT4Innovations, 30. dubna 22, 701 03 Ostrava 1
Czech Republic
E-mail: michal.burda@osu.cz

The discovery of hidden logical or statistical regularities in data has long been central to data mining and knowledge discovery. In the R ecosystem, however, most available tools for this purpose—such as `arules` and `arulesViz` [7, 8]—focus narrowly on classical *association rules* or frequent itemsets, operating on crisp, categorical data and using heuristic measures such as support and confidence. An R package for subgroup discovery (e.g., `rsubgroup` [2]) or non-R rule-based learning systems (e.g., `KEEL` [1], `FURIA` [9]) broaden the scope but are respectively confined to Boolean logic and predictive modeling. The `lfl` package [4] offers the `searchrules()` function for mining *fuzzy association rules*, which partially addresses the need for fuzzy reasoning in R; however, its capabilities are limited to this single pattern type and lack the extensibility required for more general fuzzy or statistical analyses.

The `nuggets` package for R statistical software is being developed to fill this methodological gap. It provides an extensible platform for pattern discovery that combines (1) *fuzzy logic–based conditions* applicable to numeric and linguistic variables, (2) *statistical evaluation of patterns* using hypothesis tests, (3) *interactive exploration and visualization* for human-guided discovery, and (4) *extensibility* for developing new pattern types and quality measures. Unlike previous tools, `nuggets` integrates the interpretability of linguistic rule systems with the flexibility of modern interactive Shiny-based data science interfaces. It thereby establishes a bridge between mathematical logic, fuzzy set theory, and practical data analysis within the R ecosystem.

Fuzzy Conditions and Linguistic Descriptions. The central concept in `nuggets` is the *condition*—a conjunction of predicates over the data. Conditions may be Boolean or fuzzy. Fuzzy predicates are constructed from numeric variables by membership functions (triangular, trapezoidal, raised-cosine) and evaluated via t-norms (Gödel, Goguen, Łukasiewicz). Each predicate corresponds to a linguistic label and defines a gradual transition between truth and falsity. This construction provides a mathematically rigorous link between data values and linguistic interpretation.

A particularly natural interpretation of fuzzy conditions appears in the tradition of *linguistic descriptions of data*. Here, discovered patterns are expressed as linguistic statements, e.g.,

"If income is high and age is middle, then occupation is likely IT."

Such expressions correspond formally to association rules or subgroup descriptions but use linguistic labels (*low, medium, high*) that mirror how humans reason about data. Because `nuggets` predicates are inherently fuzzy, they directly realize such linguistic terms, creating a

Acknowledgement The development of `nuggets` was supported by the project "Research of Excellence on Digital Technologies and Wellbeing CZ.02.01.01/00/22_008/0004583", which is co-financed by the European Union.

https://www.doi.org/10.15452/978-80-7599-515-5.2026.03

bridge between the association-rule community and researchers developing linguistic summaries of data. In practice, this makes it possible to produce outputs that are both computationally precise and linguistically interpretable.

Mathematical and Statistical Foundations. The `nuggets` package follows the general logical–combinatorial approach of GUHA-style methods, where patterns are defined as logical conditions satisfied by subsets of data. Each condition is evaluated primarily through its *support*—the proportion or total membership degree of observations fulfilling it. Support thus serves as the key criterion for candidate generation and pruning, ensuring that discovered conditions describe sufficiently large and relevant parts of the dataset.

Beyond support, `nuggets` enables the computation of optional *quality measures* or statistical tests that quantify the strength or distinctiveness of discovered relations. These may include classical metrics such as confidence, lift, or correlation, as well as conditional comparisons of subpopulations. Currently, the package implements statistical evaluation for *two-sample contrasts* (e.g., differences in means or proportions) and *conditional correlation-based patterns* between numerical attributes. Fuzzy-weighted versions of such tests are planned for future releases but are not yet implemented in the current version.

Pattern Types Beyond Rules. At present, `nuggets` supports several families of pattern discovery tasks unified by the concept of conditions and support-based evaluation. The fundamental pattern type corresponds to *association rules* [7], representing implications of the form $A \Rightarrow C$ derived from frequent conditions. In addition to these, the package includes functionality for *conditional contrasts*, where subgroups are compared with respect to a target variable using statistical descriptors such as mean differences or correlation coefficients. These contrasts can be interpreted as a lightweight form of subgroup discovery or contrast pattern mining [2, 3, 5].

The design of `nuggets` is modular and extensible, allowing researchers to define new pattern types by specifying how candidate conditions are processed within a user-defined callback R function. Planned extensions include emerging patterns, exceptional models, and fuzzy-weighted statistical evaluation, which will further expand the analytical capabilities of the framework while maintaining interpretability and logical consistency.

Interactive Exploration and Visualization. Pattern discovery is inherently exploratory. The `explore()` interface, implemented via `Shiny`, enables the analyst to browse the discovered rules and visualize their statistical characteristics. Filtering allows users to focus on specific conditions or variables. This interactive environment turns pattern discovery into a human–machine dialogue rather than a static list of results.

Extensibility and Research Applications. A central goal of `nuggets` is to provide an open experimental environment for methodological innovation. The framework separates the mining process from the evaluation phase, allowing users to attach their own analytical logic to the discovered conditions. Specifically, every generated frequent condition can be passed to a *user-defined R function*, which may compute arbitrary statistics, quality measures, or model diagnostics. This mechanism enables the creation of entirely new pattern types—ranging from domain-specific contrasts to model-based exceptions—without modifying the internal search engine. This makes `nuggets` a genuine research framework rather than a fixed toolbox. It allows computer scientists, statisticians, and fuzzy logicians to experiment with new ideas in data-driven reasoning—testing how alternative fuzzification strategies, evaluation functions, or statistical measures influence the discovery of interpretable patterns. The framework thus serves as a bridge between theoretical exploration and practical implementation.

Example Workflow. A minimal example illustrates the integration of fuzzy preprocessing,

statistical evaluation, and interactivity within the R session. As an example, a built-in **mtcars** dataset is used:

```
# Install the package (perform only once)
install.packages("nuggets")

# Load the library to the R environment
library(nuggets)

# Preprocess - dichotomize and fuzzify the numeric variables
cars <- mtcars |>
    partition(cyl, vs:gear, .method = "dummy") |>
    partition(carb, .method = "crisp", .breaks = c(0, 3, 10)) |>
    partition(mpg, disp:qsec, .method = "triangle", .breaks = 3)

# Search for associations
rules <- dig_associations(cars,
                          antecedent = everything(),
                          consequent = everything(),
                          max_length = 4,
                          min_support = 0.1,
                          measures = c("lift", "conviction"))

# Explore the found rules interactively
explore(rules, cars)
```

This workflow produces association rules based on both crisp and fuzzy predicates that are created from the original numeric variables of the **mtcars** dataset. The results can be visualized and inspected interactively, facilitating interpretation of association rules.

Conclusions and Outlook. The `nuggets` package provides a flexible and extensible framework for discovering interpretable data patterns based on frequent logical conditions. Its design unifies classical association-rule mining with linguistic and fuzzy representations, while enabling optional statistical evaluation for selected pattern types such as conditional contrasts and correlations. Pattern generation is driven by support, ensuring efficient mining of relevant conditions, whereas additional quantitative analyses or tests can be seamlessly attached when desired.

A major strength of `nuggets` lies in its extensibility. The framework allows users to define custom fuzzification schemes and to evaluate an arbitrary R function on every frequent condition, thereby enabling the creation of new, user-defined pattern types. This design encourages experimentation with alternative logical semantics, statistical measures, and application-specific evaluation criteria, making `nuggets` not only a tool for applied pattern discovery but also a research platform for developing new methods.

Future development will focus on incorporating fuzzy-weighted versions of statistical tests, expanding the library of built-in pattern types, and enhancing interactive visualization of complex fuzzy partitions. By combining formal clarity, methodological openness, and practical usability, `nuggets` contributes to the convergence of fuzzy logic, statistical reasoning, and interactive data exploration within the R environment. A detailed documentation of `nuggets` can be found at `https://beerda.github.io/nuggets/`.

References

[1] Alcalá-Fdez, J., et al.: *KEEL: A software tool to assess evolutionary algorithms for data mining problems.* Soft Computing **15** (2011) 1427–1454.

[2] Atzmueller, M.: *Subgroup discovery.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **5**(1) (2015) 35–49. doi: 10.1002/widm.1144.

[3] Bay, S. D., and Pazzani, M. J.: *Detecting group differences: Mining contrast sets.* Data Mining and Knowledge Discovery **5** (2001) 213–246.

[4] Burda, M.: *lfl: Linguistic Fuzzy Logic.* R package version 2.3.0 (2022). Available at `https://CRAN.R-project.org/package=lfl`.

[5] Chen, Y.: *Contrast Pattern Mining: A Survey.* arXiv preprint arXiv:2209.13556 (2022).

[6] Dong, G., and Li, J.: *Efficient mining of emerging patterns.* In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99) (1999) 43–52.

[7] Hahsler, M., Grün, B., Hornik, K.: *arules — A computational environment for mining association rules and frequent item sets.* Journal of Statistical Software **14**(15) (2005) 1–25.

[8] Hahsler, M.: *arulesViz: Interactive visualization of association rules with R.* R Journal **9** (2017) 163–175.

[9] Hühn, J., and Hüllermeier, E.: *FURIA: An algorithm for fuzzy unordered rule induction.* Data Mining and Knowledge Discovery **19** (2009) 293–319.