# Subgroup Discovery for Defect Prediction

D. Rodríguez[1], R. Ruiz[2], J.C. Riquelme[3], and R. Harrison[4]

[1] Univ. of Alcalá, 28871 Alcalá, Spain
daniel.rodriguezg@uah.es*
[2] Pablo de Olavide Univ., 41013 Seville, Spain
robertoruiz@upo.es
[3] Univ. of Seville, 41012 Seville, Spain
riquelme@us.es
[4] Oxford Brookes Univ., Oxford OX33 1HX, UK
rachel.harrison@brookes.ac.uk

Although there is extensive literature in software defect prediction techniques, machine learning approaches have yet to be fully explored and in particular, Subgroup Discovery (SD) techniques. SD algorithms aim to find subgroups of data that are statistically different given a property of interest [1,2]. SD lies between predictive (finding rules given historical data and a property of interest) and descriptive tasks (discovering interesting patterns in data). An important difference with classification tasks is that the SD algorithms only focus on finding subgroups (e.g., inducing rules) for the property of interest and do not necessarily describe all instances in the dataset.

In this preliminary study, we have compared two well-known algorithms, the Subgroup Discovery algorithm [3] and CN2-SD algorithm [4], by applying them to several datasets from the publicly available PROMISE repository [5], as well as the Bug Prediction Dataset created by D'Ambros *et al.* [6]. The comparison is performed using quality measures adapted from classification measures. The results show that generated models can be used to guide testing effort. The parameters for the SD algorithms can be adjusted to balance the specificity and generality of a rule so that the selected rules can be considered *good enough* for software engineering standards. The induced rules are simple to use and easy to understand. Further work with more datasets and other SD algorithms that tackle the discovery of subgroups using different approaches (e.g., continuous attributes, discretization, quality measures, etc.) is needed.

## References

1. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In: Proceedings of the 1st European Symposium on Principles of Data Mining, pp. 78–87 (1997)
2. Herrera, F., Carmona del Jesus, C.J., González, P., del Jesus, M.J.: An overview on subgroup discovery: Foundations and applications. Knowl. Inf. Syst. (2010)

* Research supported by Spanish Ministry of Education MEC TIN 2007-68084-C02.

M.B. Cohen and M. Ó Cinnéide (Eds.): SSBSE 2011, LNCS 6956, pp. 269–270, 2011.

3. Gamberger, D., Lavrac, N.: Expert-guided subgroup discovery: methodology and application. Journal of Artificial Intelligence Research 17, 501–527 (2002)
4. Lavrač, N., Kavšek, B., Flach, P., Todorovski, L.: Subgroup discovery with CN2-SD. The Journal of Machine Learning Research 5, 153–188 (2004)
5. Boetticher, G., Menzies, T., Ostrand, T.: Promise repository of empirical software engineering data. West Virginia University, Department of Computer Science (2007)
6. D'Ambros, M., Lanza, M., Robbes, R.: An extensive comparison of bug prediction approaches. In: IEEE Mining Software Repositories (MSR), pp. 31–41 (2010)