

# A Community-Based Platform for Machine Learning Experimentation

Joaquin Vanschoren and Hendrik Blockeel

Computer Science Dept., K.U. Leuven, Leuven, Belgium

**Abstract.** We demonstrate the practical uses of a community-based platform for the sharing and in-depth investigation of the thousands of machine learning experiments executed every day. It is aimed at researchers and practitioners of data mining techniques, and is publicly available at <http://expdb.cs.kuleuven.be>. The system offers standards and API's for sharing experimental results, extensive querying capabilities of the gathered results and allows easy integration in existing data mining toolboxes. We believe such a system may speed up scientific discovery and enhance the scientific rigor of machine learning research.

## 1 Introduction

Experimentation is the lifeblood of machine learning (ML) research. A considerable amount of effort and resources are invested in assessing the usefulness of new algorithms, finding the optimal approach for new applications or just to gain some insight into, for instance, the effect of a parameter. Yet in spite of all these efforts, experimental results are often discarded or forgotten shortly after they are obtained, or at best averaged out to be published, which again limits their future use. If we could collect all these ML experiments in a central resource and make them publicly available in an organized (searchable) fashion, the combined results would provide a highly detailed picture of the performance of algorithms on a wide range of data configurations, which can be tapped into to gain quick insights, reuse previous experiments, and increase the speed and depth of ML research.

In this paper, we demonstrate a prototype of a community-based platform designed to do just this. It allows researchers and practitioners to share their results and/or to keep an organized, detailed log of past experiments, even during early phases of algorithm development. This functionality can even be integrated seamlessly in DM toolboxes, without requiring any additional effort from the user. The biggest benefit, however, lies in the flexible querying capabilities it offers to browse and reorganize large amounts of publicly available (or locally stored) data. In the remainder of this paper we first describe the design of this platform in Section 2. Next, Section 3 illustrates the querying capabilities and their utility in ML research. Finally, Section 4 concludes with future work.

## 2 Anatomy of the Platform

The components of the platform are shown Fig. 1. To promote the free exchange of experiments, an extensible XML-based language for experiment exchange is introduced, dubbed ExpML, which is formally described in Vanschoren et al. [6], and adheres to an ontology of experimental concepts. It captures the basic structure of a machine learning experiment, yet allows each element (e.g. algorithms, kernels, datasets, evaluation metrics,...) to be described further to define the exact setup and cover task-specific properties. It currently supports a very wide range of classification and regression tasks, and is being extended further. It also supports the submission of new elements of any kind, verifies that experiments are reproducible, and allows each element to be ‘tagged’ with properties, such as dataset and algorithm characteristics. To facilitate the description of new experiments, a Java API is provided to compose experiments from scratch and submit them to the system. The same interface can also be used by existing DM tools to automatically stream new experiments to the database.

The experiments are then stored in an *experiment database* (ExpDB) [1,2]: a database designed to store all details of machine learning experiments in order to make them reproducible and clearly interpretable. It is designed to be very extensible, scale easily to large numbers of experiments and allow queries on practically any aspect of the experimental setup and outcome. As a result, its structure is very complex. Though we cannot include a complete discussion here, all the details of its design can be viewed at the project webpage. It currently contains over 600,000 experiments on 67 classification and regression algorithms, 149 different datasets and 2 data preprocessing techniques. Access is provided through web services for submitting experiments and launching queries.

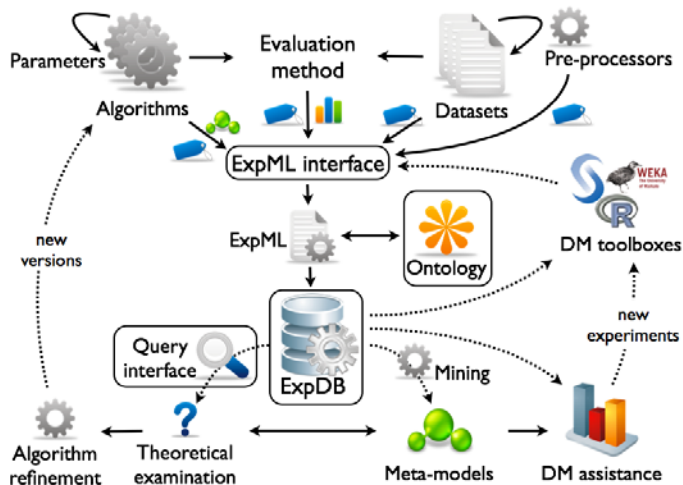


Fig. 1. The components of our platform and their use

Finally, the platform offers various ways to use the stored data. First, it offers two query interfaces to explore all the stored information. The first is an online interface, which offers an SQL query interface and basic visualizations of the returned results. The second is a stand-alone explorer tool offering more advanced querying and visualization features. Both are available at <http://expdb.cs.kuleuven.be>, including example queries and a video tutorial for using the explorer tool. Second, large sets of experiments can be downloaded to build meta-models of algorithm performance, e.g. in function of data characteristics. Finally, data can also be streamed to DM tools. For instance, experimental toolboxes can download experiments that have run before, visualization tools can download data points to provide new insights and DM assistance tools can check whether a certain approach is viable on a specific problem.

### 3 Benefits and Illustrations

The benefits of experiment databases have been listed systematically in previous work [5,2]. Here we simply illustrate these benefits with some example cases.

One of the biggest benefits is the reuse of previously run experiments. Collecting the experiments originally run to test various hypotheses, when combined, can be reused to test other hypotheses. As such, new experimental work can often be conducted by simply querying the database, instead of setting up new experiments; this is tremendously faster and easier. One clear example are benchmarking studies: competing algorithms probably have run before on the same datasets, and those experiments can therefore be simply downloaded instead of repeating them. The same strategy can also be used in algorithm development or refinement. For instance, we queried for the effect of the ‘gamma’ parameter in RBF kernels on several datasets, shown in Fig. 2. This analysis showed that the useful range of this parameter correlates with the number of attributes (shown in parentheses), which in turn led to an adjustment of the RBF function and improved the applicability of this particular algorithm [5].

Furthermore, it facilitates verification and iterative refinement of public knowledge. For instance, since a large number of bias-variance experiments

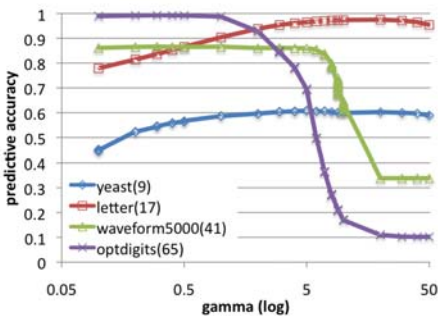


Fig. 2. The effect of kernel width in SVMs

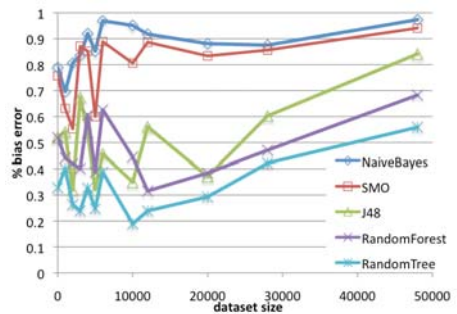


Fig. 3. Bias error ratio vs. dataset size

are stored, we could verify the claim that bias error becomes dominant on large datasets [3]. A single query reorganized the available experimental results and produced Fig. 3, showing that, at least for larger datasets, the percentage of bias error in the total error indeed increases with size. Again, not a single new experimental run was required; querying the stored experiments was sufficient.

Even when the stored experiments are not sufficient to answer a new experimental question, they may reduce the need for new experiments. New experimental runs may even be generated automatically as a query requires them (using, for instance, active learning principles), thus making the decision of which new experiments to perform in order to answer a new scientific question entirely transparent. Besides offering these practical advantages, the platform also simply serves as an ultimate reference, a ‘map’ of all known approaches and how they behave on certain types of datasets.

## 4 Conclusion and Future Work

Although standardization of experimental results has been a driving force in bio-informatics [4], our approach is, to the best of our knowledge, unique in ML and extends the latter by providing much more powerful querying abilities. Functioning as an open portal for ML experimentation, we are confident that this platform will be very useful for ML researchers, practitioners and students. Currently planned work includes a full integration in the WEKA toolbox [7] (beyond the interfaces available now), a graphical query interface, an account system and further work on the ontology used.

## Acknowledgements

This research is supported by GOA 2003/08 ‘Inductive Knowledge Bases’ and F.W.O.-Vlaanderen ‘Foundations of Inductive Databases for Data Mining’.

## References

1. Blockeel, H.: Experiment databases: A novel methodology for experimental research. In: Bonchi, F., Boulicaut, J.-F. (eds.) *KDID 2005*. LNCS, vol. 3933, pp. 72–85. Springer, Heidelberg (2006)
2. Blockeel, H., Vanschoren, J.: Experiment databases: Towards an improved experimental methodology in machine learning. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) *PKDD 2007*. LNCS (LNAI), vol. 4702, pp. 6–17. Springer, Heidelberg (2007)
3. Brain, D., Webb, G.: The Need for Low Bias Algorithms in Classification Learning from Large Data Sets. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) *PKDD 2002*. LNCS (LNAI), vol. 2431, pp. 62–73. Springer, Heidelberg (2002)

4. Stoeckert, C., Causton, H., Ball, C.: Microarray databases: standards and ontologies. *Nature Genetics* 32, 469–473 (2002)
5. Vanschoren, J., Pfahringer, B., Holmes, G.: Learning From The Past with Experiment Databases. In: Ho, T.-B., Zhou, Z.-H. (eds.) *PRICAI 2008*. LNCS (LNAI), vol. 5351, pp. 485–496. Springer, Heidelberg (2008)
6. Vanschoren, J., Blockeel, H., Pfahringer, B., Holmes, G.: Organizing the world's machine learning information. *Comm. in Comp. and Inf. Science* 17, 693–708 (2008)
7. Witten, I.H., Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn. Morgan Kaufmann, San Francisco (2005)