Machine Learning Project 2010–2011

Meta-learning from an experiment database

Jonas De Greef Li Quan

For the machine learning project, we will study meta-learning from an experiment database[1, 8].

1 Introduction

Instead of treating this project as a general and vague overview of the capabilities and usage of all algorithms and datasets, we will focus this project on a more in-depth and less general case study of very large datasets.

Large datasets are very common in many software distributions that contain a fully developped and polished classifier to identify—for example medical treatments for a given set of symptoms.

2 Goals

We wish to process large datasets:

- without much side-effects (like overfitting and complexity);
- using a short amount of time and memory;
- maintaining a relative high accuracy.

3 Problems

Of course, this requires a lot of memory and computational complexity[4]. As already seen during this course, large datasets will likely lead to complex models[2, 6].

These problems evidently raise a few questions:

- Are the datasets artificially generated or "real-world" examples and does this influence the accuracy[7]?
- What is the effect of noise?

- Is it possible to preprocess the data (e.g. pruning sparse attributes) for memory and time issues; or is better to postprocess the data?
- How do we make the trade-off between time and accuracy?

It is our hope that by studying this very specific subdomain of the experiment database, we can offer a more satisfactory and in-depth answer to these questions.

For the recommender system, we will first try to choose between two or three algorithms; secondly, if this works well, we can try to search the optimal parameters for different datasets.

4 Data representation

For the data representation, we will extract relevant data from the experiment database into a single table[3] and mine that table using the WEKA data mining tool[5].

We already experimented with the WEKA tool and created some useful views for our goals. We can query the database for the accuracy of all experiments on datasets with a size larger than a certain treshold:

```
SELECT li.name, v.evalue, d.name
FROM experiment e, learner_application la, learner_implementation li, dataset d,
     data_property_value dp, evaluation v, evaluation_metric m,
     evaluation_metric_implementation mi
WHERE e.laid = la.laid and la.liid = li.liid and e.did = d.did and d.is_original='true'
     and v.eid = e.eid and v.emiid=mi.emiid and mi.emid=m.emid and m.name='
     predictive_accuracy' and d.did = dp.did and dp.dpid = 9 and dp.value > DATASETSIZE
ORDER BY v.evalue desc
```

Similary for the learning curve:

```
SELECT ev.evalue as pred_acc, (100 - pppv.value) as perc_size, li.name from experiment e
     , dataset di, dataset di2, data_property_value dp, preprocessing_step pps,
     preprocessor_application ppa, preprocessor_implementation ppi, preprocessor pp,
     preprocessor_parameter_value pppv, preprocessor_parameter ppp, evaluation ev,
     evaluation_metric em, evaluation_metric_implementation emi, learner_application la,
      learner_implementation li
WHERE e.laid = la.laid and la.liid=li.liid and li.name=ALGORITHM and e.eid = ev.eid and
     ev.emiid = emi.emiid and emi.emid=em.emid and em.name='predictive_accuracy' and e.
     did = di.did and di.preprocessed_by = pps.ppsid and pps.did_in=di2.did and dp.did =
      di2.did and dp.dpid = 9 and dp.value > DATASETSIZE and pps.ppaid = ppa.ppaid and
     ppa.ppiid=ppi.ppiid and ppi.name='weka.RemovePercentage' and ppa.ppaid=pppv.ppaid
     and pppv.pppid=ppp.pppid and ppp.alias='percentage
```

We can also show the impact of data property DP on the percentage of bias error in the total error for a number of different algorithms (the query is not shown here because it is too long).

5 Conclusion

In conclusion, we will try to find the best learning algorithm for large datasets. If possible, we wil determine the influence of the parameters on both accuracy and time.

References

[1] Blockeel, H. Experiment databases: A novel methodology for experimental research. In *Knowledge Discovery in Inductive Databases, 4th International Workshop, KDID'05, Revised, Selected and Invited Papers* (2006), vol. 3933 of *Lecture Notes in Computer Science*, Springer, pp. 72–85.

[2] Blockeel, H. *Machine learning and Inductive Inference*. Acco, 2010.

[3] Blockeel, H., and Vanschoren, J. Experiment databases: Towards an improved experimental methodology in machine learning. In *Knowledge Discovery in Databases: PKDD 2007, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Proceedings* (2007), vol. 4702 of *Lecture Notes in Computer Science*, Springer, pp. 6–17.

[4] Bottou, L., and Bousquet, O. Learning using large datasets. In *Mining Massive DataSets for Security*, NATO ASI Workshop Series. IOS Press, Amsterdam, 2008.

[5] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The WEKA data mining software: an update. *SIGKDD Explor. Newsl. 11* (November 2009), 10–18.

[6] Oates, T., and Jensen, D. Large datasets lead to overly complex models: An explanation and a solution. In *KDD* (1998), pp. 294–298.

[7] Pfahringer, B., Bensusan, H., and Giraud-Carrier, C. G. Meta-learning by landmarking various learning algorithms. In *Proceedings of the Seventeenth International Conference on Machine Learning* (San Francisco, CA, USA, 2000), ICML '00, Morgan Kaufmann Publishers Inc., pp. 743–750.

[8] Vanschoren, J., Van Assche, A., Vens, C., and Blockeel, H. Meta-learning from experiment databases: An illustration. In *Benelearn 2007, Annual Machine Learning Conference of Belgium and The Netherlands* (2007), pp. 120–127.