

Chapter 10

Discussion and conclusions

Anders Holst, Daniel Gillblad and Björn Levin

The main objectives of the DALLAS project has been to spread knowledge about the potential and state-of-the-art of learning systems to the industrial partners; to provide the academic partners with real world data to evaluate the methods on; and to build a contact network for learning systems between industry and academia. The means to achieve this was to try to apply the learning systems methods used at the participating institute and universities to seven tasks provided by the industrial partners. Four tasks were treated by all academic partners, two tasks were treated by a subset of the academic partners, and for one of the tasks it was impossible to provide any data.

Although the tasks were from quite different domains, they had many things in common. For example, most of the tasks involved time series, and the need to find relations between these. Some new methodology were also developed in this context, like the mutual information rate measure [Gillblad and Holst, 2001]. There were in total three conference contributions produced within the DALLAS project [Gillblad and Holst, 2001; Johansson and Niklasson, 2001, 2002].

10.1 Evaluation of the project

The industrial partners have had slightly different objectives when joining the project, but the main common reason for participation was knowledge transfer. Although some partners have looked upon the project also as a potential supplier of an algorithm that works on a single problem, getting an overview of current learning systems techniques and their potential was the most important goal. The partners feel that this has generally worked very well. Not only have the industries been introduced to new methods that were only vaguely known to them before, but they have also had the opportunity to see the results when these methods are applied to familiar problems within their own domains.

The knowledge transfer from the industry to the academics has also been substantial. The availability of a large number of real data sets has made it possible to find and deal with practical limitations and problems of the used methods. It has also given a thorough understanding for the practical problems involved in data analysis tasks. One example is the quality of real data, which is never as clean and straightforward as filtered or simulated data. The degree to which it contains noise, missing values, anomalies, and outright errors, is always much higher than expected.

The project has also compared the main methods represented by each academic partner to industrial standard methods such as PLS (partial least squares, widely used in the process industry). This has been very useful and interesting not only for the industrial partners, but also for the academics, since these methods are sometimes overlooked when working mainly with the development of ones own methods.

To summarize, the DALLAS project has given the participants, both industrial and academic, a good overview of methods and their strengths and limitations. The opportunity to see the methods being

tested on problems outside ones own specific domain has also increased everyones understanding of the methods and their possible applications.

Another very important effect of the DALLAS project is that it has created a lot of contacts between the industrial and academic partners. When one of the industrial partners face a problem that might benefit from the use of one of the methods presented in the DALLAS project or just a question about it, they can contact the academic partner that represented that method within the project.

The general agreement among the partners is that the DALLAS project has worked very well for this type of cooperation project, and it has even exceeded the expectations of several partners who have participated in similar projects before.

10.2 Comments on the project form

In a cooperation project like this, all partners must be prepared for the time needed to understand the different traditions in thinking and in how to express things. Early in the project the academic partners were not good at providing method descriptions on the appropriate level. This made it difficult for the industrial partners to get into the project in the beginning, and also to motivate participation in the project internally. This was considerably improved during the course of the project. It is also important that the companies reserve the time needed to absorb the results. Most companies in the DALLAS project did this very well.

Important is also how to transfer algorithms and software from the academic partners to the industry. It is clear that industry wants to and could benefit from being able to run the methods tested in the project themselves, on other kinds of data or new data sets. This is very difficult, though, since most software used by the academic partners are in the form of research prototypes, with very limited data import abilities and low flexibility. This kind of tools is necessary for method development, but not of very much use to the industry. The issue of who should adapt the research prototypes and create tools that are ready for industrial use must be considered in the future.

The ambition in the DALLAS project was to investigate the performance of a large number of algorithms on a large number of tasks. This of course resulted in very limited time for each academic partner to spend on each of the tasks. Because of this it was sometimes not possible to go as deep into each task as would have been preferred. Also, the focus had to be on applying existing algorithms rather than developing new ones.

The scheme with one academic partner responsible for each dataset worked out really well. It established close contacts between the academic partners and the industrial partners and made it easier to handle the sometimes large and complicated data sets.

Generally, the industrial partners were pleased with the ambition of the academic partners and the academic partners appreciated the commitment and interest from the industrial partners.

10.3 Conclusions

The main conclusions are perhaps two things that has been said many times before, but which should not be underestimated.

The first is that in all data analysis projects, the data collection, preprocessing, and the getting to know the domain and what the problem really is about, always takes a huge and by far the largest effort. It is also an iterative process that must be cycled a few turns before it gets right. Things that may happen is that something goes wrong with the data files, that there are too few effective data samples, or that data is selected in an unfortunate way. After a few iterations with new data sets it may also suddenly turn out that the initial formulation of the problem is not a good one, and that it has to be reformulated. In summary, a lot can go wrong with the data, and it is important not to give up at this

stage. Although all participants knew at the start that this stage would take a lot of time, everyone was still surprised to see that it took even longer than expected.

The second conclusion is that, once these initial obstacles are overcome, often the exact choice of which learning system to use is not so important since many methods perform roughly the same. The hardest part is the preprocessing, and once this has turned the data into something reasonable, it may suffice with rather simple methods to solve the real task. This also means that it often turned out that once the preprocessing was done, the results of a linear model was close to those of the non-linear models. This is true partly because, with the limited amounts of independent data, the number of free parameters had to be kept low. This in turn prevented too complicated models from being used.

Despite the above, some differences in the performance of the methods can still be seen. In for example the EKA case, two non-linear neural network models outperformed the other models, specifically the linear ones. Note also that there are many different kinds of linear models, of which for example partial least squares, Perceptron neural networks without hidden layers, and the naive Bayesian classifier, has been used in this project. Even when a linear model is sufficient to solve a task, all linear models do not perform the same. For example the naive Bayesian classifier performed somewhat better than PLS on the AstraZeneca task.

Thus, there is no single method that can be said to perform significantly better than all the rest for all problems. Some tendencies could be discerned however: the Bayesian methods worked better when there were extremely few data samples as in the AstraZeneca and NMA cases; the artificial neural networks worked better when there were large amounts of data and the nonlinearities could not be ignored as for EKA and SCA; and inductive logic methods worked best when there were a relatively small number of variables and a small number of classes as for Telia.

In conclusion, we feel that this has been a successful project, from which we have all gained valuable experience. It has given the academic partners a chance to test their methods on real data, and given important experience of working with industry. The industrial partners have learned about the potential of learning systems, and have got tasks relevant to them thoroughly analyzed. Finally, many useful contacts for possible future cooperation have been established.

10.4 Publications made during the project

- Gillblad D. and Holst A. (2001). Dependency derivation in industrial process data. In *Proceedings of 2001 IEEE International Conference on Data Mining (ICDM 2001)*, pp. 599–602. IEEE. San Jose, California.
- Johansson U. and Niklasson L. F. (2001). Predicting the impact of advertising: A neural network approach. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN2001)*, pp. 1799–1804. IEEE. Washington, DC, USA.
- Johansson U. and Niklasson L. F. (2002). Increased performance with neural nets – an example from the marketing domain. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN2002)*. IEEE. Honolulu, Hawaii, USA, (to appear).

