ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

Methodology and standards for data analysis with machine learning tools

Damien François *

Université catholique de Louvain - Machine Learning Group
av. G.Lemaître, 4, 1348 Louvain-la-Neuve - Belgium

Abstract. Many tools for data mining are complex and require skills and experience to be used successfully. Therefore, data mining is often considered an art as much as science. This paper presents some ideas on how to move forward from art to science, through the use of methodological standards and meta learning.

1 Introduction

Despite many years of scientific research, based on theoretical concepts borrowed from statistics, computer science, applied mathematics, information theory, etc., data analysis, or data mining, is still to be considered an art. Building a robust data-driven prediction model with high generalisation performances, identifying meaningful clusters in data, constructing valid decision rules, visualising high-dimensional data in a useful way, are all tasks that require skills and experience from the practitioner, which cannot be learnt directly from text books.

Many tools have been proposed in the literature, for addressing those tasks. Many of those tools come with a 'user manual' under the form of a scientific publication explaining the theoretical concepts and a functional description of the programs. Nevertheless, these are most of the time not sufficient to really unleash the full power of those tools. Very few of those user manuals actually provide hints, tips, and guidelines, about how to choose the values of the inherent parameters of the implemented method.

Furthermore, until recently, there was very few effort made to devise general methodological guidelines describing the overall process of data mining. While several proposals have been made in the literature, many practitioners are still unaware of these.

This paper summarizes these proposals and suggests a broader use of meta learning tools to aid the practice of data mining. It is organised as follows. Section 2 emphasizes the fact that data mining still needs standards and best practices. Section 3 introduces several methodological tools, such as methodological guidelines, and automatic assistants. Section 4 concludes and introduces further reading.

*This work is supported by a grant of the Walloon Region.

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

2 The art and science of data analysis

Systems identification is a mature research field with complete and self-sufficient theory, and well-established best practices. Nearly all one needs to know about system identification can be found in Ljung's book "System Identification - Theory For the User", which is considered a complete reference [1]. Along with his book, Ljung has developed and distributed a Matlab program for systems identification [2], which everybody can use with success after a lecture on the subject.

By contrast, computer programming is still considered an art. Indeed, the Bible of programming is the seminal series of monograph by Donald Knuth, entitled "The Art of computer programming". Covers of the third edition of Volume 1 quote Bill Gates as saying, "If you think you're a really good programmer [...] read (Knuth's) Art of Computer Programming [...] You should definitely send me a resume if you can read the whole thing."

Data mining, despite decades of scientific literature, is to be considered an art, too. Of course, data mining is based on solid grounds, with theoretical concepts that are strongly 'scientific'. Still, the all process of data analysis requires skills that cannot be formalised. The single fact that many data mining challenges are sill organised today shows that most of the time, people, not algorithms, are crucial to a achieve relevant and useful data mining.

The reason is that there is no single best way of analysing data; many *a priori* equivalent choices must be made and the gap between theory and practice is still huge; in theory, practice and theory are equal, but in practice...

Let us consider supervised learning for instance. Many regression tools, like artificial neural networks, support vector machines, trees, etc. depend on so many parameters no one can be considered the best of the best – The No Free Lunch Theorem. The choice of such model family is most of the time guided by elements that are exogenous to the problem at hand, be it the experience of the modeller, the historical inertia of his lab, the external constraints coming from clients, the lack of resources to actually search for the best model family, etc.

The choice of the model family is only one of the many choices to make;
- choice of the model family (ANN, SVM, Trees, etc)
- choice of the model structure optimisation methodology (CV, Bootstrap, etc)
- choice of the model parameter optimisation algorithm (joint gradients vs conjugate gradients,)
- preprocessing of the data (centring, reduction, functional reduction, log-transform, etc.) - how to deal with missing data (case deletion, imputation, etc.)
- how to detect and deal with suspect data (distance-based outlier detection, density-based, etc.)
- how to choose relevant features (filters, wrappers, embedded method ?)
- how to measure prediction performances (mean square error, mean absolute error, misclassification rate, lift, precision/recall, etc.)

No theoretical argument exists that can help making those choices in any situation. Very few guidelines are known. The experience of the practitioner is

often the only guide to make them, along with the application domain knowledge he can refer to.

3 Methodological tools

While no standard has been put forward up to know, efforts are made, both from academic and industrial practitioners, to set up best practices and well-thought methodological guidelines.

3.1 Methodology

Data mining methodology is often described as an ordered list of sub-tasks, or steps, to perform to fulfil the main task. This section presents and compares the steps described in four different data mining guidelines proposals from the literature. The proposals are named after the individuals or the research groups who have designed them.

Fayyad Usama Fayyad is Yahoo!'s executive vice president of Research & Strategic Data Solutions. Prior to joining Yahoo! Fayyad's experience also includes five years at Microsoft Research and at NASA's Jet Propulsion Laboratory (JPL). He has published over 100 technical articles in the fields of data mining and Artificial Intelligence, and has edited two books on data mining.

Fayyad's methodology begins with understanding the application domain and relevant prior knowledge, and defines "business" objectives. Then gather the data according to this objective, clean them, reduce them and explore them. Choose models and tools to test with respect to (1) the business objective and (2) the data at hand. Apply then the models and extract new knowledge, to finally interpret the results and act and make decision [3].

Cios Krzysztof J. Cios is professor at the University of Colorado in Denver. He leads the Data mining and Bioinformatics laboratory. He authored three books, and more than 150 papers on the subject of data mining.

Cios' methodology comprises fewer steps than Fayyad's but it also begins with understanding the application domain and needs, then data collection, understanding, cleaning and reduction. He suggests always performing simple statistics first, and then moving to more elaborate models. Cios mentions that backwards steps must be performed, for instance, data collection can be rethought after some interpretation of prediction models shows the need for it [4].

SEMMA SEMMA stands for Sample, Explore, Modify, Model, Assess. SEMMA was set up by statistical software vendor SAS to refer to the process of data mining. Although SAS explicitly denies SEMMA to be a data mining methodology, it sure resembles one. The SEMMA scheme first samples the data if the data are too voluminous for the analysis. Then, exploratory tools are used to gain basic knowledge of the data. Data are afterwards modified to focus the model selection process. The next step is to build prediction models and assess their performances. Once the optimal model is found, it is deployed to analyse new data [5].

		Fayyad	Cios	SEMMA	CRISP-DM
1	Objective determination	X	X		X
2	Data collection	X	X	X	X
3	Data cleansing	X	X	X	X
4	Data reduction	X		X	X
5	Problem reformulation	X			
6	Data exploration	X		X	
7	Tools selection	X		X	
8	Model construction	X	X	X	X
9	Model validation	X	X	X	X
10	Result interpretation	X	X	X	X
11	Deployment	X	X		X

Table 1: The steps involved in four proposed data mining methodologies.

CRISP-DM The CRISP-DM consortium (CRoss Industry Standard Process for Data Mining) is lead by SPSS, Daimler-Chrysler and NCR, and funded by the European Commission to develop a methodology for data mining based on best practices. The 1.0 version of the CRISP data mining guide was issued in 2000, based on the feedback of many practitioners around the world. The guidelines suggest first to understand the problem and define objectives, then understand the data and potential issues with cleaning, outliers, etc. Afterwards, models are built and tested before they are deployed [6].

Summary Several other data mining companies have set up standard methodologies, which they either have published through white papers or through their website. All the above-described data mining processes are rather similar, as illustrated in the following table, from which a complete data mining methodology can be devised (see Table 1.)

Fayyad's methodology is the most refined one. The SEMMA approach is focussed on the data analysis steps, while Cios'methodology and the CRIPS-DM framework are very similar, only the data reduction step is made explicit in the later while it is not the case in the former.

The methodology always begins with a business objective (step 1) and ends with a business action (step 11). In between, some steps are purely algorithmic (data cleansing, reduction, model construction), while others are more exploratory (step 6 and step 10). Finally, steps 2 and 10 serve as bridges forth and back between the business objective and the data mining problem definition.

Beyond these considerations, the RAMSYS system analyses the problem of distributed data mining, i.e. when the steps of Table 1 are performed at different places by different people [7].

3.2 Other standards

Along with methodological standards, which define the process of data mining, several other standards have been defined for more specific tasks.

Several proposals have been published to extend SQL, the de facto standard query language for relational databases, towards a data mining query language, e.g. DMX [8] and DMQL [9]. Rather than simply allowing for queries looking like 'Which customers churned last month', data mining query languages aim at formulating queries like 'Which customers are most likely to churn next month'.

Similarly, standard API's are arising, such as the JDM API [10]. API's serve the same purpose as query languages, except that they are targeted at developers rather than users. Using a standard API, a developer can access the data mining capabilities of one piece of software and process the results within another.

Another way of increasing data mining software inter-usability, is to develop common data and model storing and transfer formats. As such, PMML [11] allows building a prediction model using one given piece of software (for instance a development software) and using it in another one (for instance an online realtime system.)

3.3 Meta learning

While the overall process of data mining has been described in the literature, it does not allow yet the inexperienced user running successful data mining projects at once. Much more is needed, in terms of best practices and unspoken know-how. For instance, many newcomers in the data mining field forget to normalise features when building distance-based models (nearest neighbours, or models based on Gaussians for instance), or do not ensure that the way classes are coded matches the range of model output (for instance coding target as -1 and 1 while using a multilayer perceptron with an output node restricted to [0, 1]), or not making sure the order of the samples is appropriate, etc.

3.3.1 The process of meta learning

Meta learning can help formalizing this latent know-how [12]. Meta learning refers to the process of learning how learning algorithm perform on given datasets. While a typical learning problem involves a dataset and a set of candidate models from different families and/or structures, the meta learning problem considers pairs of dataset/champion model as a dataset from which new knowledge is to be inferred.

The idea behind meta learning is to learn which models perform well on which datasets, trying to infer rules or meta prediction models, which can guide the choice of a prediction model when facing a new dataset. The rationale is the very same as in traditional learning, i.e. similar models should perform similarly on similar dataset, just as similar data are mapped onto similar output values.

3.3.2 Dataset characterisation

One of the crucial points in meta learning is the dataset description, which requires defining and extracting relevant features from the dataset, just like features must be extracted from images before scene recognition can be performed,

for instance. The features extracted from datasets can be categorized as follows.

Statistical and entropic features. Much emphasis has been put on statistical and entropic features, such as the number of classes, of variables, skewness and kurtosis of the target variable, inter-correlation between variables, etc [13].

Features extracted from the structure of the champion prediction model. These features are extracted from the structure of prediction models built on the dataset, such as the optimal number of neurons in an artificial neural networks, the depth and balance of a regression or classification tree, etc. [14].

Landmarking The term "landmarking" refers to the fact of characterizing a dataset by the performances of several prediction models built on these datasets. Most of the time, simple models are considered, and chosen as different as possible in terms of learning schemes [15].

3.3.3 Tools and applications

This section presents some tools and applications of meta learning.

The **Metal Data Mining Assistant** [16] is the result of a European research project aiming at providing practitioners with tools guiding the process and the choices, in particular about what model selection concerns.

The **DCRanker** [17]and the **Intelligent Discovery Electronic Assistant (IDEA)** [18] both allow the user to specify the relative importance he associates with computation time and response accuracy. The system then suggests a ranking of the most relevant models according to these user preferences.

The **MiningMart** [19] is a project aiming specifically at providing recommendations about which preprocessing steps are the most appropriate

4 Conclusions

Data mining is slowly evolving from art to science. Practitioner skills are the most critical element in order to achieve successful and useful data mining; this will remain so as long as the know-how and the best practices are not properly formalized. Steps towards formalisation of explicit knowledge include the development of methodological standard guidelines, while implicit knowledge can be captured by meta learning tools.

Paths for further research should include random-based models. Random forests for instance are becoming very popular due to their robustness and ease of use. Extreme learning machine seem a very promising tool too [20], with very few parameters for the practitioner to choose. Quantifying the importance of the software versus the human also brings valuable insights [21]. For instance, the choice of the splitting of the data in training set, validation set and test set, is often done by the practitioner, while it could be optimised by the software to minimize bias [22]. The gap between the algorithmic tools and the human expert can be leaped with the aid of visualization tools [23], and by the domain knowledge that is available prior to modelling [24, 25].

References

[1] Lennart Ljung. *System identification: theory for the user*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1986.

[2] The Mathworks. *System identification toolbox data sheet*, 2007.

[3] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, 1996.

[4] K. Cios and L. Kurgan. *Advanced Information Systems*, chapter Trends in Data Mining and Knowledge Discovery Knowledge Discovery. 2002.

[5] SAS Institute. *SAS, The Enterprize Miner - SEMMA*.

[6] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. *CRISP-DM 1.0 - Step-by-step data mining guide*. The CRIPS-DM Consortium, 2000.

[7] S. Moyle and A. Jorge. Ramsys - a methodology for supporting rapid remote collaborative data mining projects. In C. Giraud-Carrier, N. Lavrač, S. Moyle, and B. Kavšek, editors, *ECML/PKDD'01 workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning: Internal SolEuNet Session*, pages 20–31, September 2001.

[8] Microsoft Corporation. *Data Mining Extensions (DMX) Reference*, September 2007.

[9] J. Han, Y. Fu, W. Wang, K. Koperski, and O. Zaiane. Dmql: A data mining query language for relational databases. In *SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'96)*, Montreal, Canada, June 1996.

[10] Mark F. Hornick, Erik Marcadé, and Sunil Venkayala. *Java Data Mining: Strategy, Standard, and Practice: A Practical Guide for architecture, design, and implementation*. Morgan Kaufmann Publishers Inc., 2006.

[11] The Data Mining Group. *The PMML Standard*, May 2007.

[12] R. Vilalta, Christophe Giraud-Carrier, Pavel Brazdil, and Carlos Soares. Using meta-learning to support data mining. *International Journal of Computer Science and Applications*, 1(1):31–45, 2004. DBLP.

[13] D. W Aha. Generalizing from case studies: A case study. In *n Proceedings of the Ninth International Workshop on Machine Learning*, pages 1–10. Morgan Kaufman, 1992.

[14] Y. Peng, P. Flach, P. Brazdil, and Soares C. Decision tree-based characterization for meta-learning. In *Proceedings of the ECML/PKDD'02 Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning*, pages 111–122, 2002.

[15] B. Pfahringer, H. Bensusan, and C Giraud-Carrier. Meta-learning by landmarking various learning algorithms. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 743–750, 2000.

[16] C. Soares P. Brazdil and J. Costa. Ranking learning algorithms. In *Machine Learning*, 2003.

[17] J. Keller, I. Holzer, and S Silvery. Using data envelopment analysis and cased-based reasoning techniques for knowledge-based engine-intake port design. In *In Proceedings of the Twelfth International Conference on Engineering Design*, 1999.

[18] Abraham Bernstein, Foster Provost, and Shawndra Hill. Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):503–518, 2005.

[19] K. Morik and M Scholz. The miningmart approach to knowledge discovery in databases. In N. Zhong and J Liu, editors, *ntelligent Technologies for Information Analysis*. Springer, 2004.

[20] Y. Miche, P. Bas, C. Jutten, O. Simula, and A. Lendasse. A methodology for building regression models using extreme learning machine: Op-elm. In M. Verleysen, editor, *ESANN 2008 : European Symposium on Artificial Neural Networks*, Bruges, April 23-25, 2008.

[21] C. Lemke and B. Gabrys. Do we need experts for time series forecasting? In M. Verleysen, editor, *ESANN 2008 : European Symposium on Artificial Neural Networks*, Bruges, April 23-25, 2008.

[22] M. Aupetit. Homogeneous bipartition based on multidimensional ranking. In M. Verleysen, editor, *ESANN 2008 : European Symposium on Artificial Neural Networks*, Bruges, April 23-25, 2008.

[23] T. Alhonnoro and M. Sirola. Feature selection on process fault detection and visualization. In M. Verleysen, editor, *ESANN 2008 : European Symposium on Artificial Neural Networks*, Bruges, April 23-25, 2008.

[24] E. Roglia, R. Cancelliere, and R. Meo. Classification of chestnuts with feature selection by noise resilient classifiers. In M. Verleysen, editor, *ESANN 2008 : European Symposium on Artificial Neural Networks*, Bruges, April 23-25, 2008.

[25] A. Pozdnoukhov and M. Kanevski. Geokernels: modeling of spatial data on geomanifolds. In M. Verleysen, editor, *ESANN 2008 : European Symposium on Artificial Neural Networks*, Bruges, April 23-25, 2008.