Learning with Missing or Incomplete Data

Bogdan Gabrys

Smart Technology Research Centre
Computational Intelligence Research Group
Bournemouth University
bgabrys@bournemouth.ac.uk

Abstract. The problem of learning with missing or incomplete data has received a lot of attention in the literature [6,10,13,21,23]. The reasons for missing data can be multi-fold ranging from sensor failures in engineering applications to deliberate withholding of some information in medical questioners in the case of missing input feature values or lack of solved (labelled) cases required in supervised learning algorithms in the case of missing labels. And though such problems are very interesting from the practical and theoretical point of view, there are very few pattern recognition techniques which can deal with missing values in a straightforward and efficient manner. It is in a sharp contrast to the very efficient way in which humans deal with unknown data and are able to perform various pattern recognition tasks given only a subset of input features or few labelled reference cases.

In the context of pattern recognition or classification systems the problem of missing labels and the problem of missing features are very often treated separately.

The availability or otherwise of labels determines the type of the learning algorithm that can be used and has led to the well known split into supervised, unsupervised or more recently introduced hybrid/semi-supervised classes of learning algorithms.

Commonly, using supervised learning algorithms enables designing of robust and well performing classifiers. Unfortunately, in many real world applications labelling of the data is costly and thus possible only to some extent. Unlabelled data on the other hand is often available in large quantities but a classifier built using unsupervised learning is likely to demonstrate performance inferior to its supervised counterpart. The interest in a mixed supervised and unsupervised learning is thus a natural consequence of this state of things and various approaches have been discussed in the literature [2,5,10,12,14,15,18,19]. Our experimental results have shown [10] that when supported by unlabelled samples much less labelled data is generally required to build a classifier without compromising the classification performance. If only a very limited amount of labelled data is available the results based on random selection of labelled samples show high variability and the performance of the final classifier is more dependent on how reliable the labelled data samples are rather than use of additional unlabelled data. This points to a very interesting discussion point related to the issue of the trade-off between the information content in the observed data (in this case

P. Foggia, C. Sansone, and M. Vento (Eds.): ICIAP 2009, LNCS 5716, pp. 1–4, 2009.

available labels) versus the impact that can be achieved by employing sophisticated data processing algorithms which we will also revisit when discussing approaches dealing with missing feature values.

There are many ways of dealing with missing feature values though the most commonly used approaches can be found in the statistics literature. The ideas behind them and various types of missingness introduced in [20] are still in use today and the multiple imputation method is considered as state of the art alongside the Expectation Maximization (EM) algorithm [6,11,23,24]. In general the missing value imputation methods are the prevalent way of coping with missing data. However, as it has been pointed out in many papers [1,11,16,17,24] such a repaired data set may no longer be a good representation of the problem at hand and quite often leads to the solutions that are far from optimal.

In our previous work on the subject of learning on the basis of deficient data we have advocated a different, unified approach to both learning from a mixture of labelled and unlabelled data as well as robust approaches to using data with missing features without a need for imputation of missing values.

We argue that once the missing data has been replaced it can potentially result in overconfident decisions not supported by the discriminative characteristics of the observed variables and a different approach is needed.

One of the examples of such an approach not requiring imputation of missing values is based on hyperbox fuzzy sets and was presented in [8]. The General Fuzzy Min-Max (GFMM) algorithms for clustering and classification naturally support incomplete datasets, exploiting all available information in order to reduce a number of viable alternatives before making the classification decision. The GFMM algorithm is also able to learn from mixed supervised and unsupervised data, iteratively [9] or in an agglomerative manner [7] processing both types of patterns for adaptation and labelling of the fuzzy hyperboxes, thus falling into the semi-supervised category. The networks also posses the ability to quantify the uncertainty caused by missing data.

Such philosophy of dealing with both unlabelled and missing input data within a consistent, unified framework has also been pursued in our more recent work utilising the physical field based classifier, electrostatic charge analogy and a metaphor of data samples treated as charged particles which will be used here as the second example [3,4,22].

A number of simulation results for well-known data sets are provided in order to illustrate the properties and performance of the discussed approaches as well as facilitate the discussion.

References

1. Berthold, M.R., Huber, K.-P.: Missing values and learning of fuzzy rules. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 6(2), 171–178 (1998)
2. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on Computational learning theory, pp. 92–100. ACM, New York (1998)

3. Budka, M., Gabrys, B.: Electrostatic Field Classifier for Deficient Data. In: The sixth International Conference on Computer Recognition Systems, Jelenia Góra, Poland, May 25-28 (2009a)
4. Budka, M., Gabrys, B.: Mixed supervised and unsupervised learning from incomplete data using a physical field model. Natural Computing (submitted, 2009)
5. Dara, R., Kremer, S., Stacey, D.: Clustering unlabeled data with SOMs improves classification of labeled real-world data. In: Proceedings of the World Congress on Computational Intelligence, WCCI (2002)
6. Dempster, A., Laird, N., Rubin, D.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39(1), 1–38 (1977)
7. Gabrys, B.: Agglomerative Learning Algorithms for General Fuzzy Min-Max Neural Network. The Journal of VLSI Signal Processing 32(1), 67–82 (2002)
8. Gabrys, B.: Neuro-fuzzy approach to processing inputs with missing values in pattern recognition problems. International Journal of Approximate Reasoning 30(3), 149–179 (2002)
9. Gabrys, B., Bargiela, A.: General fuzzy min-max neural network for clustering and classification. IEEE Transactions on Neural Networks 11(3), 769–783 (2000)
10. Gabrys, B., Petrakieva, L.: Combining labelled and unlabelled data in the design of pattern classification systems. International Journal of Approximate Reasoning 35(3), 251–273 (2004)
11. Ghahramani, Z., Jordan, M.: Supervised learning from incomplete data via an EM approach. In: Cowan, J.D., Tesauro, G., Alspector, J. (eds.) Advances in Neural Information Processing Systems, vol. 6, pp. 120–127 (1994)
12. Goldman, S., Zhou, Y.: Enhancing supervised learning with unlabeled data. In: Proceedings of ICML (1998)
13. Graham, J., Cumsille, P., Elek-Fisk, E.: Methods for handling missing data. Handbook of psychology 2, 87–114 (2003)
14. Kothari, R., Jain, V.: Learning from labeled and unlabeled data. In: Proceedings of the 2002 International Joint Conference on Neural Networks, 2002. IJCNN 2002, vol. 3 (2002); Loss, D., Di Vincenzo, D.: Quantum computation with quantum dots. Physical Review A 57 (1), 120–126 (1998)
15. Mitchell, T.: The role of unlabeled data in supervised learning. In: Proceedings of the Sixth International Colloquium on Cognitive Science (1999)
16. Nauck, D., Kruse, R.: Learning in neuro-fuzzy systems with symbolic attributes and missing values. In: Proceedings of the International Conference on Neural Information Processing – ICONIP 1999, Perth, pp. 142–147 (1999)
17. Nijman, M.J., Kappen, H.J.: Symmetry breaking and training from incomplete data with radial basis Boltzmann machines. International Journal of Neural Systems 8(3), 301–315 (1997)
18. Nigam, K., Ghani, R.: Understanding the behavior of co-training. In: Proceedings of KDD 2000 Workshop on Text Mining (2000)
19. Pedrycz, W., Waletzky, J.: Fuzzy clustering with partial supervision. IEEE Transactions on Systems, Man, and Cybernetics, Part B 27(5), 787–795 (1997)
20. Rubin, D.: Inference and missing data. Biometrika 63(3), 581–592 (1976)
21. Rubin, D.: Multiple Imputation for Nonresponse in Surveys. Wiley-Interscience, Hoboken (1987)

22. Ruta, D., Gabrys, B.: A Framework for Machine Learning based on Dynamic Physical Fields. Natural Computing Journal on Nature-inspired Learning and Adaptive Systems 8(2), 219–237 (2009)
23. Schafer, J., Graham, J.: Missing data: Our view of the state of the art. Psychological Methods 7(2), 147–177 (2002)
24. Tresp, V., Ahmad, S., Neuneier, R.: Training neural networks with deficient data. Advances in Neural Information Processing Systems 6, 128–135 (1994)