THE PERILS OF IGNORING DATA SUITABILITY

The Suitability of Data Used to Train Neural Networks Deserves More Attention

Kevin Swingler
Computing and Maths, University of Stirling, Stirling, FK9 4LA, Scotland
kms@cs.stir.ac.uk

Keywords: Data Preparation; Machine Learning; Data Mining; Data Quality and Quantity

Abstract: The quality and quantity (we call it suitability from now on) of data that are used for a machine learning task are as important as the capability of the machine learning algorithm itself. Yet these two aspects of machine learning are not given equal weight by the data mining, machine learning and neural computing communities. Data suitability is largely ignored compared to the effort expended on learning algorithm development. This position paper argues that some of the new algorithms and many of the tweaks to existing algorithms would be unnecessary if the data going into them were properly pre-processed, and calls for a shift in effort towards data suitability assessment and correction.

1 INTRODUCTION

Neural networks are popular and well used machine learning techniques, and deserve their place in any data mining course, text book or software package. Algorithm research has expanded in recent years with authors producing thousands of papers either proposing new learning algorithms or improving existing ones.

However, there has not been a related explosion in research addressing the suitability of the data that these algorithms process and the issue is largely ignored by courses, books and software.

This paper argues that the preparation of data and the analysis of its suitability should receive the same attention that is afforded to algorithm development. The paper is not a criticism of algorithm development – there is still much work to do – rather it is a call to address the imbalance.

The paper starts by arguing that a robust set of methods for analysing and fixing the suitability of training data should be as much a part of the standard neural tool box as MLPs and RBFs. Section 2 demonstrates that this is not currently the case with a short analysis of data mining papers, popular text books and software packages, showing how each is biased towards learning algorithms at the expense of a treatment of data suitability. Section 3 mentions some general research in the area and the paper finishes with a short summary of some of the data suitability issues that deserve more attention.

2 DATA SUITABILITY IS LARGELY IGNORED

Machine learning algorithms, and neural networks in particular, owe their performance to three things: the data they are fed, the quality of the learning and inference algorithms and the expertise of the user. With existing algorithms, a little know-how and some trial and error it is reasonably easy to produce a correct solution from suitable data. However, many algorithms – and neural networks in particular – cannot compensate for unsuitable data, no matter how much expertise the user displays. It would therefore be sensible to use data suitability methods to fix or discard data prior to the application of a simple machine learning algorithm than to attempt to optimise the algorithm to work with data exhibiting a particular problem.

Research gains practical importance when it is applied, and it is most likely to become applied when it is taught in text books and courses and implemented in widely used software. In the next section we examine the treatment of data suitability by the data mining community, software packages and text books.

2.1 Data Mining

A recent survey paper (Wu et al., 2008) listed the top 10 data mining algorithms identified by the IEEE International Conference on Data Mining in

2006. Two things should be noted from this paper. There were no neural networks in the list – neither MLPs nor RBFs – and there were no data preparation techniques. The traditional text book criticism of neural networks is that they are a ‘black box’ technique. Generally this is cited as a reason why companies might shy away from using them for commercially important judgements, but it is also a weakness when relying on the machine learning technique itself to highlight problems in the original data.

Take a decision tree as an example. The explicit and accessible representation of knowledge allows users to trace the route to a classification and explore sensitivities (a small change in *x* would lead to a different classification, but the current class is insensitive to changes in *y*, for example). This is one text book explanation of the black box criticism, but it is possible to do the same thing with an analysis of the partial derivatives of an MLP. The real advantage of the decision tree’s structure is that it exposes problems that were hidden in the original data and allows the expert data miner to improve the model. This is not easy with neural networks and this is the main disadvantage of their black box nature.

Wu et al. found that the most popular classification methods were k Nearest Neighbours (kNN), Naïve Bayes (Hand, 2001), Support Vector Machines (Vapnik, 1995), and two tree building algorithms: C4.5 (Quinlan, 1993) and CART (Breiman, 1984). With the possible exception of SVM, these algorithms all share the common feature of allowing some problems with the training data to be fixed after model building or even at inference time. We argue that neural networks are not used for data mining as often as they once were partially for this reason. They hide the problems in the solution that were caused by problems with the data in such a way that no post-model building adjustments are possible.

2.2 Software Packages

Weka (Bouckaert, 2010) and Rapid Miner are two popular free data mining software packages. Both offer data visualisation, manipulation and attribute selection tools, but neither offers data quantity or quality analysis. SAS, which is a very popular commercial analytics package, offers neural networks amongst its data mining options but data quality processing is limited to outlier filtering. None of these software packages offers an analytic data quantity assessment tool.

Which neural networks do these software packages support? The three packages above offer RBFs and MLPs only. There have been many neural network architectures and algorithms designed since these two were invented, but these two persist as the only ones to make it into large scale data mining software packages.

To an extent, the field of data mining grew up with that of neural computing. Some early data mining software packages offered little more than neural networks – they could classify, predict and cluster and were viewed as something of a universal solution. As Wu et al. have shown, this is no longer the case, and we argue that the reason is that they hide the consequences of unsuitable training data.

If we are to see neural networks used for more commercial applications, we must address the issue of data suitability. This will take the field forwards faster than more incremental improvements in learning algorithm design.

There is a danger in the widespread practice of making an improvement to an existing algorithm and demonstrating that improvement on a benchmark data set. The danger is that we only see the successes, not the tweaks that produced no improvement. The risk is that the literature fills with algorithms that are suited to certain types of data or, worse, certain benchmark data sets. The practitioner is then faced with the impossible task of locating the right algorithm for their data. With better methods of understanding the data prior to learning, we could safely employ a smaller range of standard learning algorithms.

2.3 Text Books

A review of a number of data mining and neural computing text books further illustrates the point. Classic neural network texts such as (Hertz et al., 1991) and (Haykin, 1994) do not deal with data suitability issues at all. More recent neural network texts such as (Dreyfus, 2005), (Bishop, 2006) and (Tang et al., 2007) show a similar omission. (Swingler, 1996) dedicates a chapter to data quality and quantity but even the author admits that this is now out of date.

Data mining books should be better, but (Witten and Frank, 2005), which is a popular course text book offers two or three pages of vague advice on ensuring that data is suitable. Recently published (Du, 2010) has a chapter on data preparation but offers just a few pages on data quality and no analytic methods. There are a few specific books covering data quality and preparation: (Pyle, 1999) is good and (Dasu and Johnson, 2003) has some useful content but such books are rare compared to the number of data mining and neural network algorithm books on the market.

2.4 First Conclusion

The research that is being carried out on data suitability has less chance of being applied because books, courses and software packages are not treating it with the importance it deserves.

3 RESEARCH

We are not suggesting that data quality issues are ignored by researchers. The recent launch of the ACM Journal of Data and Information Quality (Madnick et al, 2009) is an encouraging development, though data preparation for machine learning is a small aspect of its overall remit. Much work on data quality has focused on management information systems and their need for data integrity. Data cleansing for machine learning presents an additional set of challenges.

Some authors (Zhu et al, 2007) have pointed out that data quality issues consume the majority of time and budget for commercial data mining projects. They also point out that data cleansing often focuses on incomplete, imprecise or uncertain data – errors in other words – rather than a more general question of data suitability for the machine learning task.

4 CALL TO ACTION

Poor data suitability can be difficult to detect. Problems range from simple data entry errors or missing values through outliers and minority values to multi-dimensional interactions such as correlated inputs and the many varieties of the curse of dimensionality.

The effects of poor data quality can be difficult to predict and detect and we have already mentioned that 'black box' neural networks are particularly susceptible to them. A set of methods for the analysis and correction of data suitability for neural network training and data mining in general is needed. Algorithms need to be developed, reported in text books and lecture courses, and embedded in data mining software packages. Assessment of data prior to the application of machine learning algorithms needs to gain an importance equal to that of those algorithms themselves.

There is, of course, active research into many aspects of data suitability. Some of the larger fields include data imputation, feature selection, and abnormality detection. Our argument is that there needs to be more of it and that it needs to be taken more seriously both by the research community and in textbooks and courses.

We need to identify and catalogue the problems that can be found in data sets destined for machine learning algorithms. We need automated methods for detecting, alerting and where possible correcting for these problems before the process of learning begins.

4.1 Making a Start

Much data quality research is concerned with data governance – that is, ensuring data is recorded, notated and audited correctly. Such assurances are comforting for the data miner, but it is not this type of data quality that interests us in this case. We are concerned with the qualities of a data set that make it suitable (or otherwise) as the raw ingredient for a machine learning project – hence our use of the term data suitability.

At a minimum, we suggest that no course, text book or software package about data mining should lack a detailed consideration of how the following impact on data quantity requirements and model quality:

4.1.1 Data Distribution

The distribution of the training data has a large impact on the quality of a learned model. The problem of imbalanced target classes is perhaps the best studied aspect of this – see (Japkowicz and Stephen, 2002) for an overview. The distribution of data also has an important impact on required training set size, feature selection, error detection and the risk of over-fitting. This is true for both numeric and nominal data types, for inputs and outputs.

Univariate histograms are a useful tool for early feature selection, but more work is need on automated distribution based data quality and selection methods. Features such as outliers, isolated data points and variables with too few or too many discrete values should be considered.

4.1.2 Missing Data and Errors

Imputation of missing data is well studied, with many algorithms available for this task (Little and Rubin, 2002) give a good overview. Imputing missing values has an impact on required data quantity, risk of over-fitting, data distribution and learning algorithm performance. Errors in the data are more difficult to spot but some of the methods used for data imputation can also be used for error detection.

4.1.2 Feature Selection

Feature selection is another well studied field with many proposed techniques, see (Gheyas and Smith, 2010) for a recent example. We suggest that these methods would benefit from being viewed in the light of the other data suitability issues listed here. In this we include other considerations such as feature independence.

4.1.3 Data Quantity

The issues listed above all have an impact on the quantity of data required for a successful machine learning project. Although it is true that solving the problems of data quality would mean that data quantity is not an issue in itself, it is certainly a useful measure of suitability when other aspects of data quality are only partially understood.

5 CONCLUSION

The majority of time and resources on most professional data mining projects is consumed by data preparation. This deals with outliers, missing values, abnormal distributions, data errors, insufficient data quantities, ill-posed data, co-dependent inputs and a list of other issues.

This paper does not argue that such data preparation, cleaning and verification does not take place, neither does it argue that the issue is ignored by the research community. It argues that algorithms for dealing with these issues are as important as algorithms for machine learning and inference, and so should constitute much more of the research in that field and a larger proportion of the content of teaching, text books and software.

We would like to see the data mining community make more use of neural computing based methods and we believe that an improved approach to data suitability will encourage that to happen.

ACKNOWLEDGEMENTS

Thanks to Prof. Leslie Smith for his help in preparing this paper

REFERENCES

Bishop, C.M., 2006. *Pattern recognition and machine learning*. Springer.

Bouckaert, R. R, Frank, E., Hall, M. A., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H., 2010. WEKA-experiences with a java open-source project. *Journal of Machine Learning Research, 11:2533-2541.* JMLR

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J., 1984. *Classification and regression trees.* Wadsworth.

Dasu, T., Johnson, T., 2003. *Exploratory data mining and data cleaning.* Wiley-Interscience.

Dreyfus, G., 2005. *Neural networks: methodology and applications*. Springer.

Du, H., 2010. *Data Mining Techniques and Applications: An Introduction*. Cengage Learning.

Gheyas, I. A. and Smith, L. S., 2010. Feature subset selection in large dimensionality domains. *Pattern Recognition. 43.* Elsevier.

Hand, D. J., Yu, K., 2001. Idiot's Bayes—not so stupid after all?. *Int. Stat. Rev. 69:385–398.* International Statistical Institute.

Haykin, S.S., 1994. *Neural networks: a comprehensive foundation.* Macmillan.

Hertz, J., Krogh, A. and Palmer, R.G., 1991. *Introduction to the theory of neural computation. Santa Fe institute studies in the sciences of complexity: Lecture notes.* Westview Press.

Japkowicz, N. and Stephen, S., 2002. The class imbalance problem: A systematic study. Intel. Data Anal. 6 pp. 429–449.

Little, R. J. A. and Rubin, D. B., 2002. *Statistical Analysis with Missing Data*. Wiley.

Madnick S. E., Wang, R. Y., Yang, W. L. and Hongwei, Z., 2009. Overview and Framework for Data and Information Quality Research. *ACM Journal of Data and Information Quality. 1,1.* ACM.

Pyle, D., 1999. *Data preparation for data mining.* Morgan Kaufmann.

Quinlan, J. R., 1993 *C4.5: Programs for machine learning*. Morgan Kaufmann.

Swingler, K., 1996. *Applying neural networks: a practical guide.* Academic Press.

Tang, H., Tan, K.C. and Zhang, Y., 2007. *Neural networks: computational models and applications.* Springer.

Vapnik, V., 1995. *The nature of statistical learning theory*. Springer.

Witten, I.H. and Frank, E., 2005. *Data mining: practical machine learning tools and techniques.* Morgan Kaufman.

Wu, X., Kumar, V., Quinlan, R. J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng A., Liu, B., Yu, P. S., Zhou, Z. H., Steinbach, M., Hand, D., J. and Steinberg, D., 2008. Top 10 algorithms in data mining. *Knowl. Inf.* Syst. *14, 1:1-37.* Springer-Verlag.

Zhu, X., Khoshgoftaar, T.M., Davidson, I. and Zhang, S., 2007. Editorial: Special issue on mining low-quality data. Knowl. Inf. Syst. 11,2: 131–136. Springer-Verlag.