An Investigation of How Data Quality is Affected by Dataset Size in the Context of Big Data Analytics

Completed Research Paper

Philip Woodall
University of Cambridge
phil.woodall@eng.cam.ac.uk

Alexander Borek
IBM
alexander.borek@de.ibm.com

Jing Gao
University of South Australia
Jing.Gao@unisa.edu.au

Martin Oberhofer
IBM
martino@de.ibm.com

Andy Koronios
University of South Australia
Andy.Koronios@unisa.edu.au

Abstract

In the Big Data era the volume and availability of datasets are increasing massively throughout industrial organisations. These organisations, with this data, are using data analytics to provide business insights in a way that has never been exploited before. Despite its critical role in the past, however, the problems of data quality are sometimes being dismissed in this Big Data world as being irrelevant. For example, in a large sample of data, will the effects of any data errors be "scaled out" as we continue to add more data? The aim of this work was to determine, empirically, if and when this is the case. We investigated the problem of completeness on data mining classification as we increase the volume of records used to train the classifier. Our results indicate that data quality is even more important in the Big Data world of increased volume. We also found that there are opportunities for managers to improve their analytic results by combining, in the correction proportions, increasing dataset size with improvements to data quality.

Keywords: Data quality, Big Data, data analytics, data volume, data size, information quality.

Introduction

As we move into an era of Big Data, with the volume and availability of datasets expected to increase, the question of whether data quality matters in this context is being discussed (Mayer-Schonberger and Cukier 2013). In particular, as we scale data, will the effects of errors in the datasets be reduced to an extent where data quality assessment and improvement activities are no longer needed?

When datasets were small, we removed errors because it was feasible and important to do so; because when you have a small dataset any errors are much more likely to have a large impact on the result of an analysis/processing of the data (see for example the conclusions of (Gabrys and Petrakieva 2004)). As we move away from the small data world into widespread data availability, is the opposite true, and can we start to treat data problems as unavoidable and learn to live with them (Mayer-Schonberger and Cukier 2013)?

As an example, consider the question of determining the performance of a particular supplier to a manufacturing company. A series of transactions with the supplier could be recorded in a database indicating whether the supplier delivered parts to the company on-time or not (see Table 1). The company could calculate the supplier's performance using the number of on-time deliveries over the total number of deliveries for that supplier. If, however, the supplier is referred to by two different names (the quality error being that they should be consistently referred to), then some of the deliveries for that supplier may not be counted. This may give an inaccurate performance figure, as the example in Table 1 shows: the performance of the "Air parts Ltd" supplier would be counted as 75% (3/4). Correcting the error in the supplier names ("Air pts" to "Air parts Ltd.") would yield the actual performance to be, instead, 50% (3/6).

Overall, however, if the errors are not corrected and many more supplier transactions are recorded, then a few erroneous entries will not inflict such a large bias on the overall performance, and one would expect the result to tend to the correct performance figure as more transactions are used. Hence, we could learn to live with the errors and collect more data instead.

Company	On-time delivery?
Air parts Ltd.	y
Air parts Ltd.	y
Air parts Ltd.	y
Air pts	n
Air pts	n
Air parts Ltd.	n

Table 1: Supplier transactions

The question is whether this assertion is true, and in what cases does the effect of data quality diminish as we increase dataset size? And if it does, then the implication is that managers must start to consider what is the best way to obtain more accurate data analytics results for their businesses. Should they:

- improve data quality,
- increase dataset size,
- or apply both.

In an ideal world, applying both is likely to be the best option. However, in practice, companies have limited resources and time, and there is a need to use the approach that gives the maximum benefit for

the minimum effort. This, of course, may involve some aspect of both, and the question then arises of how much to focus on quality improvement and by how much should one increase their dataset. All this depends, of course, on how data quality is affected as we increase dataset size.

Aims of the Paper

This work aims to determine, empirically, whether, and in what particular cases, data quality matters when it comes to scaling data. The broad question, of which we only address part of in this paper, is: what effect does Big Data have on data quality?

In this case we refer to data quality not just in the literal sense, but also in the broadest sense. That is, including what effect Big Data has on existing data cleansing tools, techniques for assessment and improvement, underlying concepts such as definitions of dimensions etc. This is the broad question and the specific research question we address in this paper is:

How do missing values in a dataset affect the results from data analytics methods, such as data mining classification, as the number of records in the dataset increases?

We focus on the volume aspect of big data and investigate the effects of completeness on the result of a data mining classification task (see (Han and Kamber 2007)) as we increase the size of the training (input) dataset. Datasets can increase in size in two ways: appending of attributes (fields/columns) or appending of records (row/tuples) to the dataset. We investigate the latter in this paper, with future research plans to extend the investigation to consider attributes and tables. With the popularity of data science, many organisations are now applying data analytics methods to glean insights from their data. Data mining methods, such as classification, association rule learning, and clustering etc. have, hence, become the essential tools for the data scientist to uncover hidden correlations in their data. With its usage and importance set to increase, we chose the classification data mining method as a starting point for our investigation.

Our results confirm that, based on one of the examples studied, data quality (in particular, completeness) had an increasingly negative impact on the classification results as dataset size increased. However, critically, it is important to trade-off degrading data quality with any benefits obtained from using a larger dataset. The classification method, used in this case, produced better results with a larger input dataset to train the model, even with completeness errors. We therefore discuss a secondary research question, applicable to managers who need to act on these results:

Can increasing dataset size be used as an alternative to improving data quality to obtain better data for decision making?

Our initial results indicate that it is possible to use an increase in the dataset size as an alternative to improving data quality; although, these are preliminary results, and with having only experimented with a limited number of cases, we cannot generalise further at this stage.

This paper proceeds as follows: the next section discusses related work. Section 3 discusses the factors that can affect completeness as dataset size increases. Section 4 explains the experimental setup, how data quality errors were introduced into the dataset, the data mining classification methods used, and the experimental procedure. Section 4 presents the results of the experiments, and section 5 concludes with a discussion about the results with reference to the research questions.

Related Research

Existing research has considered various ways improve the performance of analytic, in this case data mining, methods. The relevant work in this area concerns efforts to determine the effect of dataset size (Brain and Webb 1999), and the general properties of datasets that influence the performance of classification methods (Kohavi and Wolpert 1996) (Langley et al. 1992). Although this work does not investigate the interplay of data quality and dataset size.

However, other research in this area, which is the most related, has looked specifically at the effect of null values on training analytical models (Ghahramani and Jordan 1995) (Gabrys and Petrakieva 2004). The

main approach used in this area of research is to address data problems by making improvements to the algorithms (processing method). Improving the tolerance of the processing method to data quality errors could also reduce the need for data quality improvement and/or increasing the dataset size to improve performance. For example, Google makes its search engine tolerant to errors because it can not feasibly correct all the data errors in web pages. Similarly, it may not be possible to have full control over the size of the dataset – especially if data is streamed over time, and, hence, one must wait before more data can be added.

This area of research differs from our line of investigation, which is to determine the trade-off between data quality and dataset size. Our aim is not to improve the processing method (whatever that may be: data mining classification, association rule learning etc.) – hence we refer to it in the general way – but rather we investigate the effect dataset size has on quality and whether one can be used as an alternative to the other in a more general sense.

Other research focuses on the data quality aspect alone and provides insights into how the results of association rule learning can be interpreted in light of quality errors in the input dataset (Berti-Équille 2007).

Factors Affecting Completeness as Datasets Increase in Size

Data quality can be measured using many different dimensions (Wand and Wang 1996) (Wang and Strong 1996), and since this a preliminary investigation, we deliberately chose completeness as a data quality problem to insert into the input datasets because it is simple to generate these types of errors. Completeness, in this context, is the number of missing data values in a dataset.

Datasets may have a certain frequency of errors spread around the attributes and records according to some distribution. The distribution may be completely random, i.e. with a 5% probability that a record (or attribute) will contain an error. Or there may be some systematic bias that causes errors to be placed not at random, but unevenly throughout the dataset, with a higher chance of occurring in particular records. Such a case could occur if two temperature sensors, at different locations, are used to insert data into a dataset and sensor 'A' produces a faulty reading more often than sensor 'B'. The temperatures from the location near sensor 'A' are therefore more likely to be incorrect in the dataset. The periods at which the sensor relay their readings could also bias the results because if sensor 'B' produces half the readings that sensor 'A' produces in the same time period, then there are likely to be even more errors in the final dataset for records about the location near sensor 'A'.

As the size of a dataset increases, therefore, the errors remain in proportion if there is randomness in the probability of an error occurring. Alternatively, however, due to biases, there may be a decrease or increase in the proportion of errors in a dataset as it expands in size, which is highly likely to decrease or increase respectively the errors in any analytical results obtained from processing the dataset.

For the completeness data quality problem, our assertion was that as the size of a dataset increases the frequency and magnitude of errors in the result is affected by the following factors:

- The proportion of errors in the dataset (i.e. the chance of an error occurring in each row and attribute),
- the distribution of how the errors are spread in the dataset,
- the robustness of the analysis/processing method, used to obtain the results, to errors.

Our experiments therefore took each of these factors into consideration. In this series of experiments, we have so far tested for the frequency of errors (not magnitude) that are randomly spread throughout the dataset for two different types of classification method (ID3 and Naïve Bayes).

Experimental Procedure

Our experimental procedure was governed by our research question: how do missing values in a dataset affect the results from data analytics methods, such as data mining classification, as the number of records in the dataset increases?

Therefore, the following subsections explain the choice of data analytics method with a brief explanation of how it works, the choice of dataset and how we increased the number of records, how we introduced missing values into the dataset, and the overall experimental procedure.

The Data Analytics Method Used: Data Mining Classification

A classification model (herein also referred to as a classifier) assigns a "label" to a record of data based on a pre-training of the classification model on a training dataset. The training dataset contains examples of records and their labels, and the classification model "learns" this association and can apply it to predict labels for any similar records. One example could be to predict whether a new customer of an organisation will be a "good", "medium" or "bad" customer for the organisation. This is based on training the classification model on the existing set of customers for which their "good", "medium" or "bad" label is known. There are many different methods that can perform classification and the ID3 and Naïve Bayes methods chosen and used in these experiments are two examples of different approaches.

ID3 was chosen because it is a widely known and established algorithm (Quinlan 1986) and is different from the Naïve Bayes method. The Naïve Bayes classifier is little affected by small changes in the training data, it being a high bias and low variance classifier (Brain and Webb 1999). These two methods are likely to exhibit different results and are therefore useful to help determine the effect of the robustness of the method to quality errors.

For our purposes, there are two steps when using a classifier:

- To train the classifier using a "training dataset" and
- To test the accuracy of the classifier using a "test dataset".

The inaccuracy of a classifier can be measured by counting the number of incorrect labels that the classifier assigns to the records in the testing dataset – providing that one knows the true labels of the testing dataset, so that the actual vs. assigned labels can be compared. We therefore deliberately performed our experiments with a dataset where the labels had been pre-assigned to each record, using the same dataset for training and testing.

The Dataset

To perform our experiments, we used the "Mechanical Analysis" dataset from (Bache and Lichman 2013). This contains data about mechanical pumps, which have been assigned classification labels according to various properties of the pump including, for example, the revolutions per minute of the pump and the component number etc. Altogether, the dataset contains eight attributes including one class label. The class label is simply an integer from 1 to 6. The dataset contains 6990 records, with no missing values, which makes it an ideal candidate for testing the completeness data quality dimension.

For our experiment, we divided this dataset into three: a small (2330 records), medium (4660 records) and a large size (6990 records) dataset. We could therefore train the classification model using either a small, medium or large size dataset. The "test dataset" always remained the same, and in this case was the large size dataset. Using two different methods for classification (ID3, Naïve Bayes), we measured the difference in accuracy between the small, medium and large size datasets for varying levels of completeness.

The Procedure

The overall experimental procedure is shown in Figure 1. First, a classifier is trained on a "training dataset", and then tested using the "test dataset". The input dataset of the testing phase has its labels removed, and then the model attempts to determine what the labels should be. The model therefore outputs the test dataset with the labels appended to each row. The accuracy of this model output is determined by comparing it to the correct answer (i.e. the original test dataset containing labels).

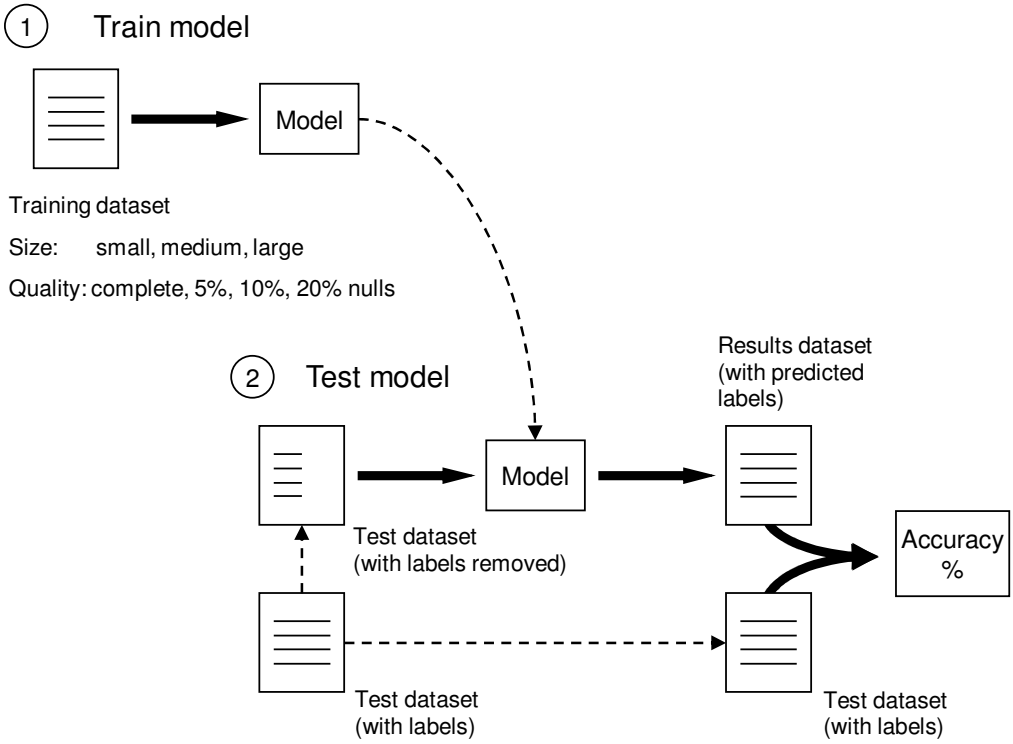

Figure 1: Experimental procedure

To make a comparison, this whole procedure was repeated for each size (small, medium and large) and quality (containing none, 5%, 10%, and 20% null values) of the training dataset. A comparison between these different results shows the differences in accuracy for the different dataset characteristics, and allowed us to determine how the accuracy of the result changes as the dataset size is increased for different levels of missing values.

One hundred iterations were performed to obtain the results (i.e. running the above tests 100 times with different random placement of the nulls into the training set). This ensured that the results were not biased based on a particular placement of nulls in the training dataset. Moreover, for each iteration, the records in the small and medium datasets were chosen randomly from the large dataset to ensure that the results were not skewed by a particular set of records. Hence, the mean percentage inaccuracies are presented in the results.

For each size of dataset we introduced no nulls (the complete dataset), 5% nulls, 10% nulls and 20% nulls. The nulls were inserted by selecting a record and an attribute at random, and then deleting that value. This was done repeatedly until the entire dataset contained, for example, 5% nulls (for the small dataset). So the small dataset with 5% nulls contained 932 nulls out of 18640 (2330*8) values.

Results

The following sections present the results for the investigation of how quality is affected by dataset size using the ID3 and Naïve Bayes classification methods.

ID3 Results

Figure 2 shows that as the dataset size increases, for all dataset sizes with and without null values, there is a reduction in misclassifications for the ID3 method; see for example the difference between the small, complete training dataset (approximately 57%) and the large, complete training dataset (approximately 2%).

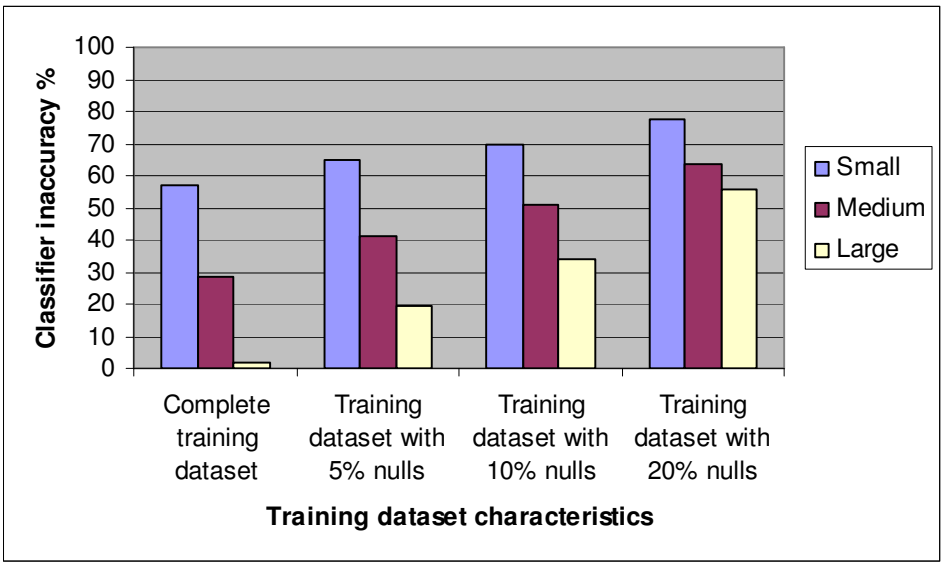

Figure 2: Classifier inaccuracy for ID3

However, as the dataset size increases, the difference in inaccuracy between classifiers trained on complete and incomplete datasets increases; this is shown in Figure 3 with the steady increase in percentage inaccuracy for each percentage of nulls. For the small size dataset, Figure 3 shows that the dataset containing 5% nulls is approximately 8% more inaccurate than the complete dataset (see the leftmost bar); this rises to 17% in the large dataset with 5% nulls. With 20% nulls, the small dataset has an inaccuracy of approximately 20% and this rises to over 50% as the dataset size increases. Hence, the accuracy gets worse as the dataset size increases.

However, this is not the whole story: although the accuracy is getting worse with larger datasets, it is doing so at a slower pace than the overall increase in performance of the ID3 method as the dataset size increases. This is to a point: the large dataset with 20% nulls returns approximately the same number of misclassifications as the small complete dataset (compare the leftmost bar and the rightmost bar in Figure 2). For example, starting with the small dataset with 20% errors, one can increase dataset size or improve quality to end up with the same level of classifier error (approximately 57%).

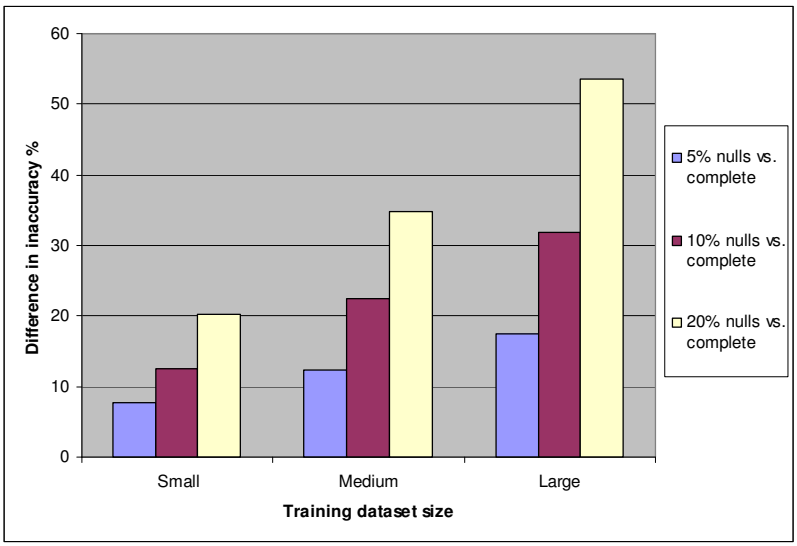

Figure 3: The difference in inaccuracy between a classifier trained on a complete dataset vs. a dataset with nulls, for the ID3 classifier

From any starting dataset with a particular size and % of nulls, to determine the benefits from improving quality, one should note the difference between the starting dataset and the bar of the same dataset size in the complete training dataset (see Figure 4). For example, starting with a small dataset with 10% nulls (70% inaccuracy), improving the quality of this dataset will yield a reduction of 13% inaccuracy (to 57%, shown by the bar for the complete, small dataset in Figure 2).

To determine the benefits from increasing dataset size, one should note the difference between the starting dataset and a bar of a larger size in the same % null group (see Figure 4). For example, starting with the same small dataset with 10% nulls (70% inaccuracy), increasing the size of this dataset will yield a reduction of 36% inaccuracy (to 34%, shown by the bar for the 10% nulls, large dataset). Interestingly, these results show that to improve the accuracy of the classifier, improving quality (giving a reduction of 13%) does not always outperform increasing the size of the dataset (giving a reduction of 36%).

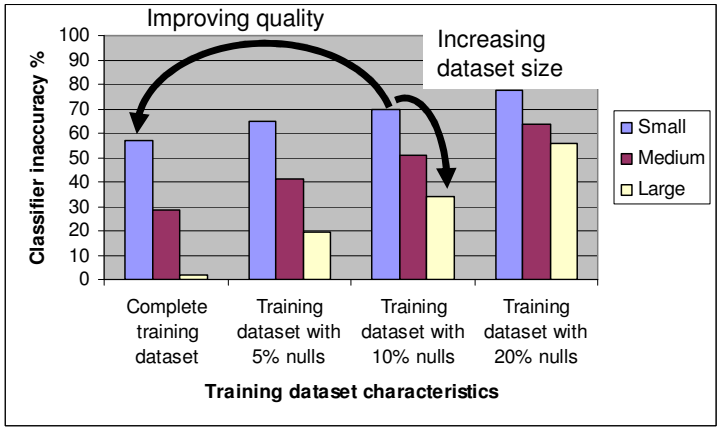

Figure 4: Improving quality or increasing dataset size

Clearly, however, it is better to both improve data quality and to increase dataset size to obtain the most accurate classifier, which has a 2% inaccuracy. Unfortunately, with constraints such as time, resources and costs, it is not always possible for organisations to perform both, and, often, a choice must be made as to what will provide the greatest benefit with the least amount of effort. Moreover, in some instances it may not be possible to obtain more data or correct the data, in which case only the available options can be performed.

In our case, we used the entire dataset to obtain these results, and hence could not add more records to determine whether increasing the dataset size will ever reach the 2% optimum performance level. In general, there may be a limit as to how many records can be added to a dataset, like in this case, which limits the quality of the results that can be obtained by increasing dataset size alone. This is analogous to the limit of correcting all errors in a dataset – once every error has been corrected, it is not possible to correct further data quality errors, and this also limits the quality of the results obtainable by improving data alone.

Note that there are other ways in which the results can be improved, such as by tuning parameters of the classification method. However, our analysis does not focus on this or other aspects of improving performance, but rather focuses solely on the analysis of data size vs. data quality trade-off.

Naïve Bayes Results

For the Naïve Bayes classification method, there is a small reduction in misclassifications as the dataset size increases for both datasets with and without null values (see Figure 5). Only the complete and 20% nulls datasets are shown because of the uniform nature of the results.

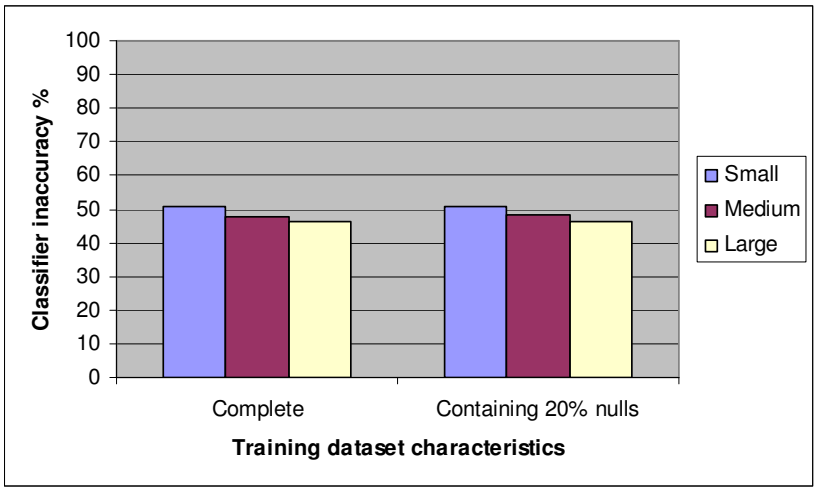

Figure 5: Classifier inaccuracy for Naïve Bayes

In the case of the Naïve Bayes method, there is only a very small increase in inaccuracy as dataset size increases, and there is no change when the null values are removed.

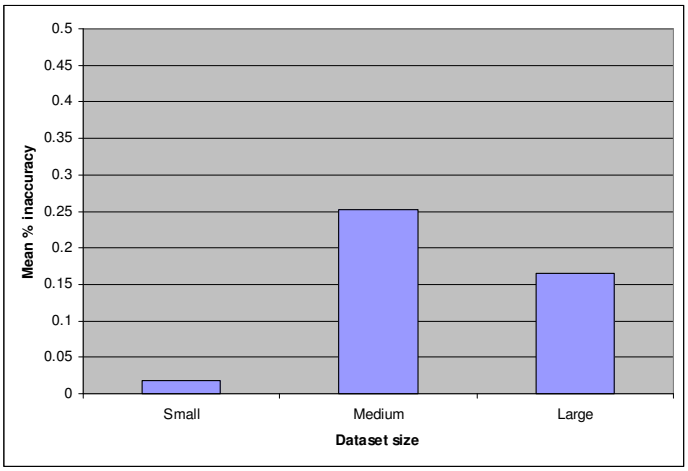

Figure 6: The difference in inaccuracy between a classifier trained on a complete dataset vs. a dataset with 20% nulls, for the Naïve Bayes classifier

Note that the maximum inaccuracy in Figure 6 is only 0.25%. The y-axis in Figure 6 is scaled differently to Figure 3, so the values in Figure 6 appear to be larger than they are relative to the values in Figure 3. Interestingly, the inaccuracy for the large dataset actually reduces from the medium dataset; although this is not significant being that these quantities are extremely small (approximately 0.09%).

Conclusion

An on-going trend is that data is not only used directly for operational decision making as in traditional business intelligence and enterprise information systems, but also large sets of data are analysed to isolate patterns, trends, correlations and irregularities to create new types of insight. For data quality practitioners, the new Big Data paradigm opens new questions about the role of data quality management that need to be addressed by experimental research. This research provides a first step by conducting experiments about the importance of data quality in the context of data mining of larger datasets as usually performed for Big Data analytics.

This paper investigated how missing values affect the results from data mining classification as the number of records in a dataset increases. We controlled the proportion and distribution of errors in the dataset, and used the ID3 and Naïve Bayes classification methods to establish how data quality is affected.

The improvement in classification results for the ID3 method, when increasing dataset size, reduces as data quality gets worse. This case shows that data quality still has an important role when dataset volume increases, and indicates that data quality could be even more significant in the Big Data world.

Although data quality has a larger negative impact as dataset size increases for the ID3 method, the reduction in misclassifications in the results outweighs the loss in general accuracy. These results show that it is possible to increase dataset size as an alternative to improving data quality to obtain better data for decision making. Our future work will aim to determine when, and in what cases, this alternative to data quality improvement can be applied successfully.

For the Naïve Bayes method, improving data quality made no difference to the results, whereas increasing the size of the dataset had a positive impact; however, this is only a very small difference. These results show that there may be cases where the only option is to increase dataset size (although in this case it is likely that the choice of method and its parameters is the correct way to improve classification performance).

Overall our results indicate that the role of data quality is becoming even more important in the Big Data era for two main reasons: firstly, in some cases, the impact of data quality is greater as dataset size

increases and, secondly, there is a critical choice to be made by managers about what amount of effort to invest in data quality improvement or whether to focus on other opportunities, such as increasing volume, to obtain better data for decision making.

References

Bache, K., and Lichman, M. 2013. "UCI Machine Learning Repository," Irvine, CA. University of California, School of Information and Computer Science.

Berti-Équille, L. 2007. "Data quality awareness: a case study for cost optimal association rule mining," *Knowledge and Information Systems* (11:2), pp. 191–215.

Brain, D., and Webb, G. I. 1999. "On the Effect of Data Set Size on Bias and Variance in Classification Learning," in *In Proceedings of the Fourth Australian Knowledge Acquisition Workshop*, University of New South Wales, pp. 117–128.

Gabrys, B., and Petrakieva, L. 2004. "Combining labelled and unlabelled data in the design of pattern classification systems," *International Journal of Approximate Reasoning* Integration of Methods and Hybrid Systems, (35:3), pp. 251–273.

Ghahramani, Z., and Jordan, M. I. 1995. "Learning from incomplete data," Technical Report No. CBCL 108, Massachusetts Institute of Technology.

Han, J., and Kamber, M. 2007. *By Jiawei Han - Data Mining: Concepts and Techniques: 2nd (second) Edition*, Elsevier Science.

Kohavi, R., and Wolpert, D. H. 1996. "Bias Plus Variance Decomposition for Zero-One Loss Functions," in *Machine Learning: Proceedings of the Thirteenth International Conference*, Morgan Kaufmann Publishers, pp. 275–283.

Langley, P., Iba, W., and Thompson, K. 1992. "An analysis of Bayesian classifiers," in *In Proceedings of the Tenth National Conference on Artificial Intelligence*, MIT Press, pp. 223–228.

Mayer-Schonberger, V., and Cukier, K. 2013. *Big Data: A Revolution That Will Transform How We Live, Work and Think*, John Murray.

Quinlan, J. R. 1986. "Induction of decision trees," *Machine Learning* (1:1), pp. 81–106.

Wand, Y., and Wang, R. Y. 1996. "Anchoring Data Quality Dimensions in Ontological Foundations," *Communications of the ACM* (39:11), pp. 86–95.

Wang, R. Y., and Strong, D. M. 1996. "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems* (12:4), pp. 5–34.