A General Approach to Incorporate Data Quality Matrices into Data Mining Algorithms

Ian Davidson
Computer Science
SUNY Albany
Albany, NY 12222
davidson@cs.albany.edu

Ashish Grover
GE R&D Center
GE Research
Niskayuna, NY 12309
agrover@rd.ge.com

Ashwin Satyanarayana
Computer Science
SUNY Albany
Albany, NY 12222
ashwin@cs.albany.edu

Giri K. Tayi
Business School
SUNY Albany
Albany, NY 12222
g.tayi@albany.edu

ABSTRACT

Data quality is a central issue for many information-oriented organizations. Recent advances in the data quality field reflect the view that a database is the product of a manufacturing process. While routine errors, such as non-existent zip codes, can be detected and corrected using traditional data cleansing tools, many errors systemic to the manufacturing process cannot be addressed. Therefore, the product of the data manufacturing process is an imprecise recording of information about the entities of interest (i.e. customers, transactions or assets). In this way, the database is only one (flawed) version of the entities it is supposed to represent. Quality assurance systems such as Motorola's Six-Sigma and other continuous improvement methods document the data manufacturing process's shortcomings. A widespread method of documentation is quality matrices. In this paper, we explore the use of the readily available data quality matrices for the data mining classification task. We first illustrate that if we do not factor in these quality matrices, then our results for prediction are sub-optimal. We then suggest a general-purpose ensemble approach that perturbs the data according to these quality matrices to improve the predictive accuracy and show the improvement is due to a reduction in variance.

Categories and Subject Descriptors

I.5.2 [Pattern Recognition]: Design Methodology - classifier design and evaluation

General Terms

Algorithms, Experimentation

Keywords

Data Quality, Ensemble Approaches, Six-Sigma, Classification, Decision Trees

KDD'04, August, 2004, Seattle, Washington, USA.

1. INTRODUCTION AND MOTIVATION

Data mining techniques have been widely employed to many industrial problems. The data mining process involves a) collecting data, b) transforming/cleansing the data, c) model building and d) model deployment/monitoring [1]. Most practitioners agree that data preparation, steps a) and b), consumes the majority of the data mining cycle time and budget [2].

Consider a set of entities such as customers whose information is recorded in a database. The process of recording this information can introduce errors. This view of data as being produced from a manufacturing process is common in the data quality literature [3]. We can consider that the aim of steps a) and b) is to restore the data to their "correct" values by removing these recording errors. While specific *detectable* errors can be identified and removed such as invalid ZIP codes prior to model building, for other errors we can only record their general properties. For example, it may be known that a particular ZIP code is often interchanged with another in the same state when entering data from forms into a database. We may be able to estimate the proportion of times this occurs (i.e. by polling a small sample), but it would be infeasible to verify and correct **all** cases. This is an example of an undetectable but documentable error.

Detectable errors can be corrected using ECTL (extract, clean, transform and load) tools that are widely available. However, this is time consuming particularly since it is an iterative process that involves all steps of the data mining cycle except d). Common repair and cleaning operations involve removing outliers, filling in missing values, removing nuisance columns and data aggregation to name a few. These operations must be proceduralized and applied to any new data that is collected and are an example of removing detectable errors.

A complementary approach to data cleaning and repairing that can address undetectable but documentable errors is to model the inherent quality levels of the data. This requires knowledge of the business processes that generated the data and involves modeling the known defects in these processes. For example, consider the customer gender field in an operational database that is populated by manual entry from hand written customer applications. An analysis of this process may reveal that 90% of the time the correct field value is entered. In 8% of the remaining situations FEMALE is entered as MALE in the other 2% MALE is entered as FEMALE. This data quality knowledge can be modeled as a data quality matrix [4]. Data quality matrices are commonly used for continuous data quality improvement and are readily available

in many organizations or can be generated using standard procedures. For example, the Six-Sigma methodology used by firms such as GE, Motorola [5] requires that such matrices be available and used.

In this paper, we introduce the idea of using data quality matrices with data mining algorithms. The idea behind this is work is as follows. If our data base is one flawed version of the entities, then recreating other versions, building models from them and combining their results can improve predictive accuracy. We propose a general purpose ensemble technique that exploits quality matrices and can be used with standard data mining algorithms such as a decision tree and naïve Bayes. The first section introduces different data quality dimensions followed by a discussion of how data quality matrices can model them. We provide examples of data quality matrices for the dimensions we will explore in this work with the remaining dimensions to be explored in the future. We then describe our ensemble technique and empirically verify its performance. We show for decision tree algorithms under a variety of conditions using real world data sets that our approach yields better predictive accuracies than using a single model or the ensemble technique bagging [6]. These conditions include the situation where the quality matrices are completely and partially accurate. We then show that the improvements are due to variance reduction. Finally, we discuss future work and summarize our paper and its contributions.

2. DATA QUALITY MATRICES

It is becoming more common for business organizations to view data or information they produce as being equivalent to a marketable product. This requires assessing the quality of data across multiple dimensions, evaluating the adequacy of the data quality for multiple uses and deploying techniques and methods for enhancing the data quality to higher acceptable levels. In order to assess the quality of data several practical approaches and tools are being used. One common approach is the use of the *data quality matrix* [4], which is a concise way of representing errors and deficiencies in the data along different dimensions. Specifically, the matrices enable us to quantify the errors in the data manufacturing process. Once the matrices are completed, quality enhancement can be undertaken by using a variety of analytical procedures. For an introduction see [7][8] which present various analytical models and procedures for data enhancement in database and data warehouse environments.

The basic forms of data quality matrices mentioned in this paper cannot model all of the quality dimensions. We provide examples for those that can be measured. We devote the rest of this section to describe how quality matrices can model the different types of errors. It is important to note the following. The matrices are created as part of quality assurance systems such as Six Sigma. Further, the matrices are typically calculated from samples of the data (i.e. customers are surveyed or additional third party data is purchased to verify personal information) and it is infeasible to correct all the errors in the data.

2.1 Accuracy

We can represent/model inaccuracies by taking a random sample and verifying the correctness of the recorded data values. For example, after such an analysis, it maybe determined that the gender and age fields could be modeled by the following data quality matrices.

Gender

	Male	**Female**
Male	0.95	0.05
Female	0.11	0.89

Age

Age	N(Age,stdev)

Table 1. Example quality matrices for accuracy.

If we consider Male as positive and Female as negative, this can be interpreted that the gender field has 5% false positives and 11% false negatives. For continuous attributes, such as age, approaches such as the Six-Sigma methodology can construct a probability distribution over the range of values. The mean of this distribution is the true value and the standard deviation a measure of variability due to errors. In this paper our focus is on using these quality matrices for data mining purposes. How these matrices are created is documented in the quality literature [4][5].

2.2 Contextual Quality

The contextual quality effectively measures the proportion of time a constraint or business rule is violated. Consider a customer database of sales transactions. Customers who purchase more than 1000 units of an item obtain a price discount which lowers the cost to below $10, whereas the typical price is more than $10. The following matrix illustrates the number of times this business rule is violated. We assume that when this situation occurs it is a data quality error not a transaction error.

	Quantity<=1000	**Quantity>1000**
Price<$10	0.01	0.99
Price>=$10	0.94	0.06

2.3 Semantic Interpretability

Consider a single field in the database labeled “Sales”. Unfortunately, the database is a combination of records where any sales tax is not added, only county sales tax is added or both county and state sales tax is added. The quality matrix approach can represent this situation as follows.

	Sales + 8%Tax	**Sales + 4% Tax**	**Sales No Tax**
Sale < $25	0.98	0.01	0.01
Sale>=$25	0.95	0.02	0.03

For all database records, eight percent sales tax should have been added. For each situation where this did not occur, we have the chance of the various errors (sales amounts) occurring.

3. AN ENSEMBLE APPROACH TO MINING WITH QUALITY MATRICES: EQPD

The view of data being manufactured by a process lends itself to the following pictorial representation of the process.

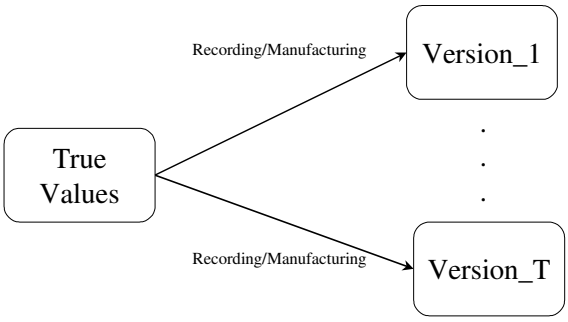

Figure 1. The view of data as being manufactured. Note each view is equally likely and the database is one such view.

Note that a version of the data is not dependent on any other version. Each version can be viewed as a stochastic perturbation of the true values and are captured/modeled by the quality matrices. We assume that true values are not known, therefore our mining approach must use a version of the data. If all *T* versions were available, then we could build a model for each and aggregate the predictions amongst the *T* models. However, we typically have only one version of the data. One way to *approximate* the various versions, is by perturbing the data available according to the quality matrices. The usefulness of the approximation, of course, depends on how close they are to the actual versions if they were available. How we approximate the versions is shown below. Note that our notation denotes all quality matrices as a single function *Q*. Future work will examine better approximations.

True Values	$\mathbf{R_1}$		$\mathbf{R_n}$
Database	$V_1=Q(R_1)$	…	$V_n=Q(R_n)$
Approx_Version$_1$	$AV_{1,1}=Q(V_1)$	…	$AV_{1,n}=Q(V_n)$
	.	…	.
Approx_Version$_T$	$AV_{T,1}=Q(V_1)$	…	$AV_{T,n}=Q(V_n)$

Table 2. The database of *n* records we have available is one (perturbed) version of the true values. From this, we can generate approximations of other versions.

We can now build models from all approximated versions and aggregate votes like the ensemble approach bagging. The pseudo code for the algorithm follows:

Algorithm: EQPD (Ensembles of Quality-Matrix Perturbed Data)

```
Input:D:TrainingSet,  T:#perturbations,  Q:
quality matrices, x: Test instance
Output:  M: The T models, Vi : Vote from
model i for test set instance
// Generate versions of data sets and build
models
For i = 1 to T
      Xi = PerturbData(D,Q)
      Mi = Classifier(Xi)
End For
// Predictions for test instance
For i = 1 to T
      Vi = Mi(x)
End For
```

4. EMPIRICAL RESULTS

In this section we wish to explore the properties of our approach to answer the following questions for decision tree classifiers:

1) Does the EQPD approach provide an improved accuracy over building a single model from the training data set?
2) Does the EQPD approach provide an improved accuracy over bagging the data set when using the same number of models (T)?
3) Does the performance of EQPD severely degrade if the quality matrices are incorrect?

We focus on three common data sets available from the UCI repository: Credit Screening, Breast Cancer and Contraceptive. Each recorded result is from 20 independent experiments of 10-fold cross validation. The statistically significant[1] best result for each data is shown in bold. If no result is statistically significant than another, then no entry is in bold.

Our first results shown in Table 3 illustrate the performance of a variety of techniques when less than 10% of discrete valued columns are perturbed by data quality matrices. We say that the quality matrices have a small variance (less than 0.1) when for binary matrices the variance is the weighted sum of the product of each row of the matrix. For example the GENDER quality matrix in Table 1 (assuming equal proportions of MALE and FEMALE) has a variance of $(0.5*0.95*0.05) + (0.5*0.89*0.11) \approx 0.07$

We find that the EQPD approach outperforms a single model and bagging when the variance is small.

	Error Single Model	**Error Bagging (T=750)**	**Error EQPD (T=750)**
Credit	22.8%	19.8%	**19.2%**
Breast	35.8%	34.4%	**33.1%**
Contraceptive	18.4%	16.5%	**15.9%**

Table 3. The performance of various approaches where the quality matrices are correct and their variability is < 0.1

Our next results shown in Table 4 illustrate the performance of a variety of techniques when less than 10% of discrete columns are

[1] Test of means at 95% confidence level

perturbed by data quality matrices but the quality matrices have a variance more than 0.1. We find in this situation our approach performs better than before when compared to bagging.

	Error Single Model	Error Bagging (T=1850)	Error EQPD (T=1850)
Credit	29.6%	29.2%	**27.1%**
Breast	38.4%	36.7%	**35.3%**
Contraceptive	27.1%	26.9%	**24.1%**

Table 4. The performance of various approaches where the quality matrices are correct and their variability is > 0.1

We now focus on the situation where the quality matrices are incorrect by a factor of 25%. For example, if the correct matrix value is 0.6 then a 25% error would produce the value of 0.4 or 0.8. We find (shown in Table 5) that the approach performs better than a single model all the time, but is not always better than bagging. As expected the performance improvement is not as great as before (Table 3 and Table 4).

	Error Single Model	Error Bagging (T=910)	Error EQPD (T=910)
Credit	25.6%	**24.3%**	25.1%
Breast	37.9%	37.1%	**36.9%**
Contraceptive	22.6%	20.1%	19.8%

Table 5. The performance of various approaches, the quality matrices are incorrect by 25% and their variability is < 0.1.

5. WHY THE APPROACH WORKS

The error of a mining tool can be decomposed into three components: a) noise, b) bias and c) variance [9]. As our approach outperforms the other approaches, then it must be better at reducing one of these three components.

The inherent noise in a problem reflects the "randomness" in the mapping between the independent and dependent variables. For example, consider the simple univariate case where the independent variable is SEX and the binary dependent variable is CHURN. Suppose each gender is apriori equally likely and 60% of males churn and 80% of females do **not** churn. Then the inherent noise in the problem 0.4*0.5+0.2*0.5 = 0.3. This noise is the Bayes error and no mining tool can reduce it.

The bias component of error reflects the systematic error due to an inappropriate model space for example. While the variance component of the error refers to the mining tools sensitivity to changes in the training data set.

For our previous empirical results we measure the variability over the twenty experiments. The variance over the training set error (for a 0-1 loss) is shown in Table 6, Table 7 and Table 8. We find that as expected bagging reduces the variability of the classifier. This is well known as bagging makes unstable learners stable [6]. However, EQPD further reduces the variance partially indicating why it outperforms the other approaches. Future work will look at determining if the bias is also reduced.

	Single Model	Bagging (T=750)	EQPD (T=750)
Credit	4.61	4.38	4.13
Breast	8.93	7.19	6.81
Contraceptive	3.79	3.42	3.11

Table 6. The variance of various approaches where the quality matrices are correct and their variability is < 0.1 (corresponds with Table 3).

	Single Model	Bagging (T=1850)	EQPD (T=1850)
Credit	6.71	5.48	4.53
Breast	9.43	8.34	6.99
Contraceptive	4.89	4.46	3.67

Table 7. The variance of various approaches where the quality matrices are correct and their variability is > 0.1 (corresponds with Table 4).

	Single Model	Bagging (T=910)	EQPD (T=910)
Credit	5.88	5.92	5.24
Breast	9.43	10.1	9.16
Contraceptive	4.91	4.59	3.67

Table 8. The variance of various approaches where the quality matrices are incorrect and their variability is < 0.1 (corresponds with Table 5).

6. FUTURE WORK

In this paper, we have not focused on the dimensions: timeliness, completeness and believability. As well as being more challenging than the dimensions investigated in this paper, the data mining and machine learning communities have tackled variations of these problems. Timeliness can be considered as an example of concept drift [10], completeness is an example of artificially producing surrogate variables [1] while believability has not been studied by either community. We intend to determine how making use of data quality matrices to model these dimensions lends itself to existing approaches or whether new approaches are required.

Our empirical results are encouraging and demonstrate the usefulness of the approach. However, like all ensemble techniques how many models to build remains an open and important question. In our approach we know that we are attempting to approximate the "true" ensemble. Furthermore, we have the quality matrices that are approximations to the generation mechanism that produced the true ensemble. We then can answer the question of how many models to build by replacing it with the question of how close to the true ensemble do we wish to be. Some of our previous work has developed bounds on the number of instances required to build a belief network [11]. We will

explore extending the approach described in that paper which uses Chernoff and Chebychev bounds to determine a bound for the number of ensembles to build.

7. CONCLUSION

Data quality is becoming an issue of growing importance to many information-intensive organizations. Accordingly current efforts to improve the quality of data have gone beyond procedures that involve simple detection and correction of data errors. Data quality enhancement efforts consider data as being a product of a data manufacturing process. Therefore a collection of data records generated from such a process may embody a variety of systemic errors and flaws. Although these errors are documentable, typically they are undetectable and hence require use of approaches such as quality matrices to model them. These matrices are useful and appropriate in capturing the stochastic nature of the data manufacturing process.

Our work tries to replicate the variability in the data using the data quality matrices. This naturally lends itself to building an ensemble of models. We propose the EQPD (Ensemble of Quality-Matrix Perturbed Data) approach that builds an ensemble of models from perturbations of the training data according to the quality matrix. Our empirical work shows that the approach leads to better accuracies than building a single model and better accuracies than bagging.

A partial explanation of why the approach works is that the variance of the EQPD ensemble error is less than a bagging an ensemble and a single model. Whether the approach also reduces bias we will address in future work.

8. REFERENCES

[1] M. Berry & G. Linoff. Mastering Data Mining, Wiley, 1999.

[2] D. Pyle. Data Preparation for Data Mining, Morgan Kauffman 1999.

[3] D. P. Ballou, R. Y. Wang, H. L. Pazer and G. K. Tayi, "Modeling Information Manufacturing Systems to Determine Information Product Quality". Management Science, Vol. 44, No. 4, 1998.

[4] E. M. Pierce, "Assessing Data Quality With Control Matrices". Communications of the ACM, Vol. 47, No. 2, 2004

[5] L. P. English, "Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits". John Wiley & Sons, Inc., 1999.

[6] L. Breiman. "Bagging predictors" Machine Learning, Vol. 26, No. 2, 1996.

[7] D. P. Ballou and G. K. Tayi, "Methodology for Allocating Resources for Data Quality Enhancement". Communications of the ACM, Vol. 32 No. 3, 1989.

[8] D. P. Ballou and G. K. Tayi, "Enhancing Data Quality in Data Warehouse Environments". Communications of the ACM, Vol. 42, No. 1, 1999.

[9] R. Kohavi and D. H. Wolpert. "Bias Plus Variance Decomposition for Zero-One Loss Functions". in Proceedings of the Thirteenth International Conference on Machine Learning, 1996, Morgan Kaufmann.

[10] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden context". Journal of Machine Learning, Vol. 23, No. 1, 1996.

[11] I. Davidson and M. Aminian, "Using the Central Limit Theorem for Belief Network Learning". The 8th Artificial Intelligence and Mathematics Symposium, 2004