Using Bayesian Networks for Cleansing Trauma Data

Prashant J. Doshi
pdoshi@cs.uic.edu
Dept. of Computer Science
Univ of Illinois, Chicago, IL 60607

Lloyd G. Greenwald
lgreenwa@cs.drexel.edu
Dept. of Computer Science
Drexel Univ, Philadelphia, PA 19104

John R. Clarke
jclarke@gradient.cis.upenn.edu
Dept. of Surgery
Drexel Univ, Philadelphia, PA 19104

Abstract

Medical data is unique due to its large volume, heterogeneity and complexity. This necessitates costly active participation of medical domain experts in the task of cleansing medical data. In this paper we present a new data cleansing approach that utilizes Bayesian networks to correct errant attribute values. Bayesian networks capture expert domain knowledge as well as the uncertainty inherent in the cleansing process, both of which existing cleansing tools fail to model. Accuracy is improved by utilizing contextual information in correcting errant values. Our approach operates in conjunction with models of possible error types that we have identified through our cleansing activities. We evaluate our approach and apply our method to correcting instances of these error types.

Introduction

The activity of cleansing a raw data set involves detecting errant values, correcting the detected errors and optionally, filling in missing values. Data cleansing is viewed as a critical pre-processing step to data mining (Fayyad, Piatetsky-Shapiro, & Smyth 1996) and is performed either separately or in conjunction with other data mining activities. Insufficient data entry checks in legacy systems, disparate sources of data, and exigency in entering the data are some of the causes of errors in medical data sets. Current cleansing methods hinge on a manual inspection, "by hand", of the data to detect and correct errant values. The manual process of data cleansing is laborious, time consuming, and prone to errors thus necessitating automated methods of cleansing. A generic framework of automated data cleansing includes the definition and determination of error types, the identification of error instances, and the correction of the discovered error instances (Maletic & Marcus 1999).

Medical data is unique due to its large volume, heterogeneity and complexity (Cios & Moore 2002) making the task of cleansing a medical database unusually difficult. The complexity of medical syntax and semantics necessitates active and costly participation of a medical domain expert in the data cleansing effort, using existing methods.

We present a new approach to data cleansing using Bayesian networks (BN). This approach both improves the accuracy of data cleansing and reduces the costs. Accuracy is improved by designing Bayesian networks that capture the *context* of an errant value to disambiguate the error. Context may be used, for example, to detect and correct error instances in which a valid but incorrect value has been entered for an attribute. For example, a typographical error may cause entry of 870.0 instead of 87.0 as the diagnostic ICD-9 code for a patient with relapsing fever. Inspection of other related attributes of a patient record is the only way to identify and correct this class of error. Cost is reduced by building automated data cleansing filters from Bayesian networks that are partially designed by inducing models from data. Using supervised learning to partially induce models from data reduces the need for costly active medical domain expert participation.

Bayesian networks may be used to produce probability distributions over correct values. These distributions can capture and retain the uncertainty inherent in data cleansing. Furthermore, we may use these networks to generate the *most likely value* for missing attribute values.

Bayesian networks represent a mature technology that is flexible enough to be applied to most domains that require cleansing. Our approach operates in conjunction with models of possible error types that we have identified through initial cleansing activities. These error types include both medical-domain-specific error types, such as medical code substitutions, and non-domain-specific error types, such as typographical errors. Many of our models can be applied to other domains with minimal change, thus representing candidate generic data cleansing tools.

This paper is structured to present first the medical data in the next section, then to discuss general error detection techniques and error types. Our context-driven probabilistic approach to cleansing is given next followed by its evaluation. We then conclude this paper and outline future work.

Understanding the Medical Data

As part of a long-term project to improve emergency center trauma management for patients with multiple life-threatening injuries and specifically, better understand the effects of time delays on patient outcome (Clarke *et al.* 2002), we obtained a large file of patient data. The data file consists of medical records on 169,512 patients admitted directly to trauma centers in Pennsylvania between 1986 and 1999 and registered in the *Pennsylvania Trauma Sys-*

tems Foundation Registry. Each patient record is composed of 412 attributes of information. These attributes can be categorized in the following manner:

Patient General Information These attributes pertain to the institution number, trauma number allotted to the patient, general demographic data, background information on the care provided to the patient at the scene of injury, and patient insurance information.

Patient Clinical Data These attributes relate to patient vital signs such as pulse, respiratory rate, motor response, glasgow coma score, and revised trauma score. These vital signs are multiply recorded at the scene of injury, during transportation to the trauma center, and at the arrival of the patient in the trauma center. Drug screening information is also contained in these fields.

Process of Acute Care This is the largest category of attributes in a patient record. They relate to patient complications in the emergency department (ED), diagnosis of injury, treatment of the patient including non-operative and operative procedures, and final patient outcome

Timestamps This includes the date and time of various events that occur during the process of care of the patient. This includes the date and time of injury of the patient, scene care dispatch, scene care arrival, scene care leave, arrival of the patient at the ED, arrival of various medical personnel in the ED, and patient discharge from the ED.

We obtained this data as a single flat file of 69 million data points. From this file, we constructed a database of relational tables that house distinct categories of information. We were then able to utilize data query languages and other system tools to manipulate the data during the process of cleansing. After trimming the data base to patients and fields meeting the requirements for our planned studies we have a database with over 68 million data points spanning 166,768 patient records. Our first task was to clean this data, including estimating missing values.

Data Cleansing Preliminaries

Data entry and acquisition is inherently prone to errors. Though much effort is expended in implementing various data entry checks, most real world data sets suffer from errant values (Redman 1998) necessitating the use of cleansing techniques. The pre-processing step of data cleansing is critical to achieve standards of data quality and avoid compromising studies that may be performed on the data sets.

Many data cleansing efforts to date (e.g. (Lee *et al.* 1999; Hernández & Stolfo 1995)) have concentrated on the problem of detecting and removing duplicate records during integration of data from multiple sources. Our data cleansing effort concentrates on detecting and correcting data values from a single source. Kimball, R. (Kimball 1996) breaks down the process of cleansing into the six steps of elementizing, standardizing, verifying, matching, householding, and documenting. Elementizing breaks down the record into its atomic units for the subsequent steps. Standardizing converts the elements into a generic form. Verifying matches the data against known lists such as codes, product lists etc. Matching maps the correct values to the existing values in the data. Bookkeeping tasks are carried out during householding and documenting. Our cleansing effort incorporates each of these six steps. Specifically, standardizing and verifying are combined into a single step followed by matching and documenting.

Current cleansing tools (Rahm & Do 2000) are designed to aid the manual process of identifying and correcting errors. Such tools use statistical methods and subjective analysis to both detect and correct errors. This often entails substitution of the errant value with a single likely correct value computed through statistical techniques such as average and variance analysis. In addition to ignoring the uncertainty present in estimating the correct values, these tools are also incapable of directly capturing expert domain knowledge in arriving at their decisions. These limitations motivate our investigation of probabilistic models such as Bayesian networks that not only capture and model the uncertainty, but also permit assimilation of the knowledge of the domain expert. Rather than suggesting a single correct value, a BN generates a probability distribution over all possible correct values. The probability distribution represents the uncertainty implicit in the estimation of likely correct values. Either the entire distribution or the *most likely value* that has the highest probability may then be selected as the replacement for the errant value. In the subsections below we outline the process of detecting errors in our trauma data set, followed by an enumeration of the different error types that were discovered during this process.

Error Detection

Error detection at the record-level typically utilizes statistical outlier detection methods. These methods are collectively referred to as *data profiling*. Such methods compute for each attribute the length, value range, variance, and discrete values for that attribute, as well as the frequency and uniqueness of each value. Occurrences of null values and missing values for attributes are also noted. Domain-specific association rules (e.g. *hospital_stay = exit_date - entry_date*) are also mined and tested for potential errant values. Additionally, we developed partial ordering constraints for detecting anomalies in temporal attributes. These partial orderings amongst the temporal attributes are detailed in (Doshi 2002). Attribute values that violate the partial ordering constraints are isolated for potential correction.

Data cleansing efforts can benefit from any supplemental documentation that accompanies the raw data, including the *data dictionary*. Such literature, in addition to providing insight into the individual attribute semantics, serves as an invaluable reference source for verification.

Classification of Error in Trauma Data

During the process of data cleansing, medical domain experts are invaluable in deciphering medical jargon that is inherent in typical medical data sets, as well as elucidating relationships between attributes. Conversely, a large proportion of medical data contains non-domain specific entries that do not require the costly active participation of medical experts. To achieve these cost savings, we have classified error mechanisms into two classes: those that require medical knowledge and those that do not.

Non-domain specific errors The errors presented in this group are not unique to medical data sets and may occur in data in other domains as well.

- Common data entry errors: These include typographical errors such as missing decimal points in numeric and alphanumeric values, skipped characters and numbers, and presence of wayward special characters.
- Quantitative transformation errors: Attributes requiring values in a particular system of units may frequently be error traps. For example, the *body_temp* attribute requires body temperature in ^{o}C. If the temperature is measured in ^{o}F it must be manually converted to ^{o}C before entry. This type of transformation process is a source of numerous errors.

Domain-specific errors The errors in this group manifest due to the medical syntax and semantics.

- Code substitutions: These involve errant entry of invalid as well as other valid but incorrect codes in place of the correct medical codes.
- Temporal anomalies: Timestamp values present in temporal attributes may violate partial ordering constraints that are specific to the domain. Additionally, the time intervals derived from the timestamps may be above or below predefined limits.

Bayesian Networks for Data Cleansing

Our general approach to error correction for non-temporal data is to capture the set of attributes that can be used to disambiguate an error using a graphical probabilistic network, also known as a Bayesian network (Pearl 1988). A BN is a DAG in which nodes represent random variables and links between the nodes represent causal or statistical dependencies. Bayesian networks provide a graphical and intuitive method to capture the relationships between attributes in a task or domain. Attributes are represented with random variables, providing a representation for any uncertainty in attribute values. Domain knowledge is captured in the semantic structure of the network that illustrates the causal relationships amongst the attributes of the domain. Bayesian networks provide a compact representation by taking advantage of conditional independence and local structure, and permit efficient computation of joint probability distributions. In the subsections below, we introduce the generic BN at the core of our error correction method, followed by the procedure that was employed for the correction.

Error Correction Model

Figure 1 depicts the generic BN that was utilized during our approach. The random variable **Correct Value of Attribute X** captures the prior probability distribution over the range of correct values of the attribute X. **Error Mechanism** represents the different types of errors that may have occurred. The random variable **Entered Value of Attribute X** takes on the potentially errant values that have been entered into the database for the attribute X.

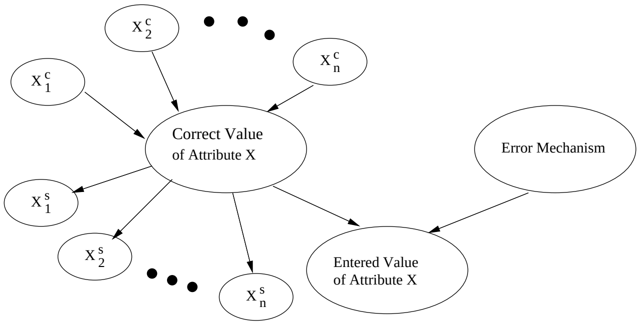

Figure 1: The Bayesian Error Correction Model.

Definition 1 *Let X be the attribute to be cleansed. Let* $\{X_i^c|i{=}1...n\}$ *and* $\{X_i^s|i{=}1...n\}$ *be the set of attributes that have a causal and a symptomatic relation respectively, with X. The* ***context*** *of X is the set* $\{\boldsymbol{X_i^j} \mid \textbf{i=1...n, j=c or s}\}$.

The context represents domain-specific information hardcoded in the network. For each attribute, whose errant values are to be corrected or missing values to be filled in, the context is identified using expert medical knowledge.

Definition 2 *Let X be the attribute to be cleansed. Let* C_X *be the context identified for X. A BN of the form given in Fig 1 containing* C_X *is called an* ***instantiated*** *BN for X.*

The process of entering values (evidence) for the contextual random variables in an instantiated BN is termed a *context check* for the attribute X. A context check, in addition to validating our assumption of the correct value, also acts as a strong inducer of the correct value.

Method of Correction

Errant and missing values may be substituted with the *most likely values*, or *all possible values*, or a *likely range of values* for that attribute. In our error correction we replace the errant value with either the *most likely value* or a *probability distribution* over all possible values of the attribute.

Our approach to correcting detected errant values for some attribute X involves a two-step procedure. The first step is to *instantiate* a BN for the attribute X, followed by isolating from the database, the chunk of errant data pertaining to the attributes that compose the BN.

The isolated chunk of data is randomly partitioned into a training set and a test (holdout) set of equal cardinalities. The training set is first cleansed manually by inspecting each patient record, specifically each errant value of the attribute is mapped to its likely correct value. The BN is trained using this set during which the conditional probability tables are populated with the appropriate probabilities. The training algorithm utilized a Bayesian MAP approach with Dirichlet functions as priors for learning the conditional probabilities. Assuming the prior distributions to be Dirichlet generally does not result in a significant loss of accuracy, since precise priors are not usually available, and Dirichlet functions can fit a wide variety of simple functions. The trained BN may now serve as an automated data cleansing filter. It is applied to the test set during which the errant and contextual information within each record is entered into the appropriate nodes of the BN and a probability distribution over the correct values is inferred. Either the *most likely value* or the entire probability distribution may be retained in the cleansed data. The tradeoffs involved in this decision are explored in (Doshi, Greenwald, & Clarke 2001). In instances where the distribution amongst the set of correct values is uniform, the errant value is left uncorrected in the data.

An important advantage of our BN is its dual use in correcting errant values as well as filling in missing values. Reliance of the cleansing process on the contextual information permits prediction of the correct value for missing values as well. A similar methodology as before is used to fill in missing values.

A special non-trivial class of error involves substitution of correct values with other valid but incorrect values. The abundance and complexity of medical codes facilitate frequent occurrences of such errors in medical data sets. One method of detecting and correcting such errors is a tedious manual inspection of each patient record. Such a method relies on expert domain knowledge coupled with contextual information in inferring the most likely correct value. Our BN by virtue of possessing both these entities represents an error correction model that is capable of correcting such errors. We give examples of such error corrections elsewhere.

Evaluation

In this section we demonstrate the application of our cleansing approach to the domain specific and non-domain specific classes of error outlined previously. Our approach to the first three classes of error is to produce a probabilistic error model that relates the data field in question with other fields in the data base. This provides a *context* that may be used, in conjunction with a model of possible error mechanisms, to compute a probability distribution over correct values for the data field in question. Our approach to the fourth class of errors is to develop a set of constraints that capture the partial ordering of all time and date fields in a record, and use these partial ordering constraints to detect and correct errant temporal data fields (Doshi 2002).

In the next subsection we give an example instantiated BN and test its predictive ability. We applied our trained BN models to the data sets that required cleansing. In the final subsection we show the predicted correct values for the corresponding errant values.

Performance

One of the attributes that required cleansing was *patient pre-existing conditions*. Figure 2 shows the BN instantiated for the attribute. The random variable **Error_Set** captures two causes of error: **MissDecPt** (missing decimal point) and **CodeSubst** (code substitution). The context consists of *on-scene pulse, on-scene systolic blood pressure, on-scene respiratory rate, and on-scene intubation*. The continuous values of each of these random variables have been suitably discretized using an entropy-based supervised discretization technique. The variables **Entered_Errant_Value** and **Correct_Value** represent the errant values and all possible correct values respectively.

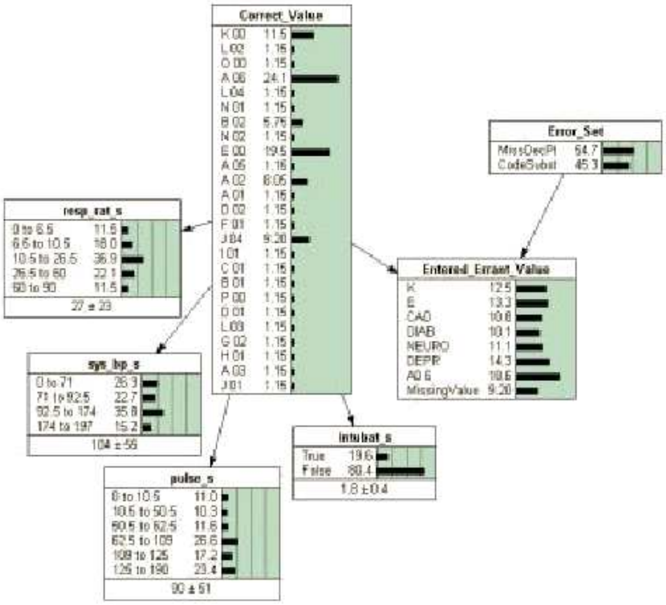

Figure 2: Example Instantiated BN.

Performance of the network was assessed using a 5-fold cross-validation technique. We isolated and manually cleansed 78 records containing errant values for the attribute. Our cross-validation technique encompassed 5 testing phases. In each phase, the data was randomly partitioned into 5 folds of which 4 folds were utilized in training and the remaining fold was used for testing the BN. During testing, contextual values as well as the errant value, if any, from the test record were entered and the most likely correct value and error mechanism were predicted. We observed an average error rate of 20% (ie. mispredictions on 3 cases) for predicting correct values and an error rate of 3.75% (ie. misprediction on 1 case) for predicting the error mechanism.

Predicted Correct Values

In Table 1 we specify the errant value followed by the most likely correct value as predicted by our instantiated BN models and relevant comments pertaining to the correction. For lack of space we show only a small sampling of corrections and hide the context of the corrected attribute value.

Rows 1-4 of Table 1 display frequently occurring errors in entry of medical codes such as misplacement of decimal points, missing decimal points and use of numeric characters in place of decimal points.

Mathematical transformations such as conversion of temperature from ^{o}F to ^{o}C become sources of error when such transformations are carried out manually. Rows 5-8 show corrections of errant *patient body temperature* values in the data. Inspection of contextual attributes such as *anatomic injury score*, *burn data*, and *patient pre-existing conditions* reinforced the belief in the likely correct value.

Errant entry of invalid or valid medical codes occurs frequently while recording medical data. In addition to absence of checks at the time of entry such errors are also caused by the preponderance of medical codes that reference the same

No.	Value	Replaced with	Reason for the action
1	A0.6	A.06	Misplaced decimal pt. Context check.
2	3310	33.10	Missing decimal pt. Freq(3.31-Spinal Tap,33.10-Incision of lung)=(356,1008).
3	34304	34.04	Suspect that decimal pt replaced with 3, Freq(34.04-Chest tube)=9780. Context check
4	57094	57.94	Suspect that decimal pt replaced with 0, Freq(57.94-foley)=15899. Context check.
5	104	40	Presumed to be in ’F, converted to ’C. Context check.
6	367	37	Missing decimal point and rounding. Context check.
7	-12	36	36 * 5/9 - 32 = -12 ’C temp. was further converted. Context check.
8	527	37	(980-32)* 5/9 = 527 Decimal pt. error. Context check.
9	870.3	87.03	Misplaced decimal pt. Frequency(87.03-CT head) = 48098. Context check.
10	CHF	A.03	**C**ongestive **H**eart **F** ailure, assigned code A.03. Context check.
11	CHOL	CHOL	Could imply Cholesterol or Cholecyctectory, both are valid preexisting conditions.
12	COPD	L.03	**C**hronic **O**bstructive **P**ulmonary **D**iseases, assigned code L.03. Context check

Table 1: Sample errors and their correction.

medical conditions. Rows 9-12 show the correction of medical code substitution errors in which the correct code is substituted with either an invalid or valid code in the data.

Using our error correction model we performed approximately 1,100 corrections. We have maintained extensive documentation on these changes to make them tractable. The documentation also serves as a source of reference for automating future cleansing activities.

Conclusion and Future Work

Data cleansing is now recognized as a consequential preprocessing step in data mining. Absence of formal methods of cleansing, insufficient documentation of retrospective cleansing activities, voluminous target data, and a lack of classification of candidate errors makes this task tedious and error-prone in itself. Furthermore, the uncertainty implicit in this process and its reliance on expert domain knowledge requires development of models that are able to capture both these qualities. BN represent an efficient candidate technology that meets these requirements and provides a pragmatic unambiguous approach to tackling the task.

We presented a generic BN model that utilizes the context of an attribute to infer a probability distribution over the correct values of the attribute as well as mechanisms of the error. The BN approach operates in conjunction with a set of possible domain-specific and non-domain specific error mechanisms that classify typical errors in medical data sets. The advantage of our approach lies in its intuitiveness and ability to capture uncertainty as well as domain knowledge to perform error correction. Furthermore the BN also addresses errors that involve one valid code substitution with another. Disadvantages involve instantiation of a BN for each attribute that must be cleansed, limitations of BN in modelling continuous-valued attributes (they must first be discretized), and poor results with small sizes of training sets. Many of these techniques were developed during a year-long effort to manually clean our target database. It is our hope that the methods we have enumerated here can help other efforts to better automate this process.

The cleansed data sets were employed in a successful statistical study (Clarke *et al.* 2002) that examined the effect of delay in the ED on patient outcome for a certain class of patients. Furthermore, we are testing our BN models on various other data sets as well.

Acknowledgements This research is sponsored in part by a Drexel/MCP-Hahnemann Research Synergies award.

References

Cios, K., and Moore, G. W. 2002. Uniqueness of medical data mining. *AI in Medicine* 26(1-2):1–24.

Clarke, J. R.; Trooskin, S. Z.; Doshi, P. J.; Greenwald, L. G.; and Mode, C. J. 2002. Time to laparotomy for intra-abdominal bleeding from trauma does affect survival for delays up to 90 minutes. *The Journal of Trauma: Injury, Infection, and Critical Care* 52(3):420–425.

Doshi, P. J.; Greenwald, L.; and Clarke, J. R. 2001. On retaining intermediate probabilistic models in data. In *AAAI Fall Symposium on Uncertainty in AI*, 47–48.

Doshi, P. J. 2002. Effective methods for building probabilistic models from large noisy data sets. Master’s thesis, Drexel University, Philadelphia, PA 19104.

Fayyad, U. M.; Piatetsky-Shapiro, G.; and Smyth, P. 1996. From data mining to knowledge discovery: An overview. *Advances in Knowledge Discovery and Data Mining, MIT Press/AAAI Press* 1–36.

Hernández, M. A., and Stolfo, S. J. 1995. The merge/purge problem for large databases. In Carey, M. J., and Schneider, D. A., eds., *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, 127–138.

Kimball, R. 1996. Dealing with dirty data. *Database and Management Systems* 9(10):55+.

Lee, M.; Lu, H.; Ling, T.; and Ko, Y. 1999. Cleansing data for mining and warehousing. In *Proceedings of the 10th Intl Conf on Database and Expert Systems*, 751–760.

Maletic, J., and Marcus, A. 1999. Progress rpt on automated data cleansing. Tech Rpt CS-99-02, U of Memphis.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Los Altos, California: Morgan-Kaufmann.

Rahm, E., and Do, H. H. 2000. Data cleaning: Problems and current approaches. *IEEE Data Engg Bulletin* 23(4).

Redman, T. 1998. The impact of poor data quality on the typical enterprise. *CACM* 41(2):79–82.