ISSN:0975-9646

A System to Improve Data Quality in healthcare using Naïve Bayes Classifier

Shwetkranti N. Taware
PG Student, Computer Department
D.Y. Patil College of Engg.
Pune, India

Vaishali Kolhe
Assistant Professor, Computer Department
D.Y. Patil College of Engg.
Pune, India

Abstract—**Data accuracy is very important in modern database. Inaccurate data results in inaccurate decision. Data errors in some domains, such as medicine, banking may have particularly severe consequences. Data errors will arise at a range of points within the life cycle of information, from data entry, through storage, integration and cleaning. Every step presents a chance to deal with data accuracy. During data entry time, errors can catch and correct more fastly. The information community has paid comparatively very little attention to data quality at assortment time. Bayesian Network Model is an end-to-end system for form design, entry and data quality assurance. Using previous form submissions, system learns a probabilistic model over the fields of the form. Bayesian Network Model has been applied at every step of the data entry process to improve data quality. Before entry, it finds an ordering of form fields that promotes rapid information capture, driven by the greedy information gain and can statically reformulate form fields to give more accurate responses. During entry, it dynamically adapts the form based on entered values, facilitating re-asking, reformulation and real-time interface feedback in the spirit of providing appropriate entry friction. After entry, it automatically identifies possibly-erroneous inputs, guided by contextualized error likelihoods and re-asks those form fields, possibly reformulated, to verify their correctness. Bayesian Network Model is vulnerable to drawbacks which will overcome by proposed work. System has potential to improve data quality significantly at a reduced cost when compared to current approach.**

Keywords— Data accuracy; data entry; form Design; dataset; bayesian network; dynamic form; learning.

I. Introduction

Data quality is defined as in terms of accuracy, completeness, consistency, uniqueness and usability and so on. Numbers of Data quality improvement techniques were proposed by different authors. Important decisions are made by organizations and individuals based on inaccurate data stored in supposedly authoritative databases. Accurate Data is especially important for critical applications like healthcare, banking where even a single error can have drastic consequences. Improving data quality during data entry is best opportunities to assure data quality. Data entry is everywhere organizations all over the world rely on clerks to transcribe information from paper forms into supposedly authoritative databases. Maintaining high quality during data entry is challenge in many smaller institutes those operating in the developing world. Data entry operator fails to specify field constraints correctly and other validation logic due to lack of expert in form designing They also lack the resources needed for performing double entry. For low-resource organizations, data entry is the first and best opportunity to address data accuracy. In banking system, data entry operator may not know the form field's constraint. So that he will fill form number of times. As one record is stored in database two times it will create many drawbacks

1. Wastage of memory.
2. Cost to get account details is increased.

Data entry operator kept the diary for patient data. That means data in paper form. Data Entry operator may not expert in data entry. In survey methodology, soft constraints are given like as age =60, pregnant field still can take yes value. System gives only warning that means in accurate data is stored in database. There are many approaches presented by various researchers to provide the accuracy while data entry operations. Recently in [1], author presented the new approach called Bayesian Network Model. The foundation of this approach is a probabilistic representation of data collection forms induced from existing form values and their relationships between form fields. The Bayesian Network Model system trains and applies these probabilistic models in order to automatically derive process improvements in question flow, question wording, and re-confirmation. Since form layout and question selection is often adhoc, Bayesian Network Model optimizes form field ordering according to a probabilistic objective function that aims to maximize the information content of form answers as early as possible. Applied before entry, the model generates an entropy-optimal ordering, which focuses on important form field first. Applying its probabilistic model (Bayesian Network) during data entry, Bayesian Network Model can evaluate the conditional distribution of answers to a form field.

In current form filling, system has form fields with many extraneous choices; Bayesian Network Model can opportunistically reformulate them to be easier and more appropriate with the available information. The model is consulted to predict which responses may be erroneous, so as to re-ask those form fields in order to verify their correctness. Intelligent re-asking approximates the benefits of double entry at a fraction of the cost.

The proposed system reduces the amount of time of a data entry operator to fill out a form. System requires diabetes dataset to check quality of Bayesian and naïve byes algorithm. Also system uses naïve byes classifier for getting predicted values. System compares the accuracy (in terms of prediction) of Bayesian and naïve byes algorithm.

The rest of the paper is organized as follows: Section 2 discusses some related work and section 3 presents the design of system. The details of the results and some discussions on this approach are presented in section 4 as Results and Discussions. Section 5 elaborates hint of some extension of the approach as future work and conclusion.

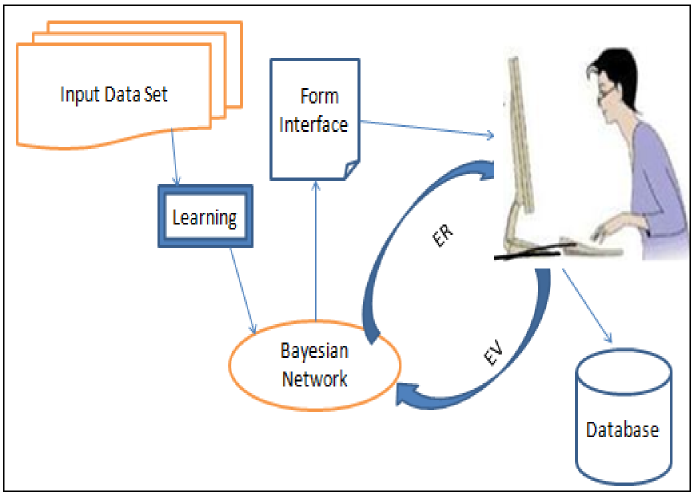
II. RELATED WORK

Most researchers adapted the data entry interface to increase user efficiency. Number of projects has used learning techniques to automatically fill or predict a top-k set of likely values. For example, Ali and Meek [6] predicted values for combo-boxes in web forms and measured improvements in the predictive model of form filling aims to reduce the amount of time a user spends filling out a form by predicting the values of fields on the form and using these predictions to make suggestions to the form filler. Ecopod[8] generated type-ahead suggestions that were improved by geographic information. They present their application EcoPod on mobile platform which replaces traditional paper field. It increases the efficiency of the identification process and reliability. The application takes as little information from the user as possible. It didn’t places no restrictions on the sequencing of the identification process. This approach is to make their solution attractive to both skilled person and professionals. Hermens [7] automatically filled leave-of-absence forms using decision trees and measured predictive accuracy and time savings. In these approaches, learning techniques are used to predict form values based on past data, and each measures the time savings of particular data entry mechanisms and/or the proportion of values their model was able to correctly predict. In Bayesian network that is focused here, gives probabilistic formalism based on learning relationships within the underlying data that guide the user towards more correct entries. This system develop and exploit probabilistic models of user errors and target a broader set of interface adaptations for improving data quality, including question reordering and re-asking and widget customizations that provide feedback to the user based on the likelihood of their entries.

III. PROPOSED METHOD

Proposed System takes input the form fields and prior data from Patient Dataset. The system gives the output improved dynamic form by applying question ordering before entry, question reasking & reformulation after and during entry.

Proposed system is probabilistic model of the data, represented as a Bayesian network over form form fields. This network captures relationships between a form’s fields in a stochastic manner. In particular, given input values for some subset of the form fields of a particular form instance, the model can infer probability distributions over values of that instance’s remaining unanswered form fields. The model is as shown in Figure 1.



ER: Error Likelihood EV: Entered values

Figure 1.Use of Bayesian Network Model

These phases show the use of Bayesian Network Model.

1. System takes the form fields of patient dataset and, previous records of patient dataset.
2. System learns the model and represent as bayesian network.
3. System order the question according to greedy Information gain principle.
4. System specify form interface to the data entry operator.
5. System enter data values in form fields and validate their constraint by previous information during entry.
6. System reask question if response will be erroneous
7. System predicts to data entry operator which entry cause to be erroneous by contextualized error likelihood principle.
8. Stop

Input Dataset (Diabetes Dataset)

Clinical data contains large amount of information about patients and their conditions. The diabetes patient has high blood sugar levels over long time. It has been found that 1.5 million deaths due to diabetes. Record set with medical attributes was collected from different hospital. It was survey data collected during pre–employment checkups conducted at the unit.

This dataset contains following form fields such as basic field’s patient id, Height, Weight, Gender, Age, FBSL, PPBSL, HDL, LDL, TCHOLE, BMI, SysBP, DiaBP.

Table I- Attributes of Diabetes Dataset

Sr No	Attribute	Description	Type
1	FBSL	Fast Blood Sugar Level (mg/dl)	Numeric
2	PPBSL	Post Prandial Sugar Level(mg/dl)	Numeric
3	HDL	High Density Lipoprotein(mg/dl)	Numeric
4	LDL	Low Density Lipoprotein(mg/dl)	Numeric
5	TCHOLE	Total cholesterol(mg/dl)	Numeric
6	BMI	Body Mass Index(kg)/(pounds)	Numeric
7	SysBP	Systolic Blood Pressure(mm/hg)	Numeric
8	DiaBP	Diastolic Blood Pressure(mm/hg)	Numeric

In proposed system there are three techniques: Form Field Reordering, Form field reasking and reformulation. These techniques are applied at the every step of data entry by serially.

Form field ordering before entry using entropy formula.

To get optimized form field reordering system uses the information entropy formula.The entropy of any field is given by following formula.

$$H(F_i) = -\sum_{fi} p(f_i) \log p(f_i)$$

The form fields are ordered by the highest conditional entropy(Information gain). As per as information gain calculation system generates ordering of form fields before entry.Due to this data entry worker can enter form field values fastly.

Form field values prediction using Bayesian Network Classifier.

The conditional entropy formula is as follows:

$$H(F_i|G) = -\sum_{g=(f_1,\dots f_n)} \sum_{fi} p(G=g, F_i=f_i) \log p(F_i=f_i|G=g)$$

By conditional entropy system predicts other field values approximately correct to avoid confusion of data entry operator.

Form field values prediction using Naïve Bayesian classifier.

The posterior probability is given as follows:

$$P(C|E) = \frac{P(E|C)P(C)}{P(E)}$$

C is predicted value for form field and E is known value of any form field where a dependance relations exists between C and E. By applying this fomula sysem predicts the values of other unfilled fom field values.Here system maximizes poserior hypohesis.

Form Field Reformulation during entry

During form filling according to previous value next question will be reformulated. As a simple example, conditioned on the answer for gender question being female , system can choose to reformulate a question about is preganant in binary form. Dynamic reformulation can be done by previous conditional expected answers. Another example for country field being mostly India, Australia but not mostly used Keniya. In combobox it dispalys only India, Australia. Keniya will be in More Option.This can be done according to probability distribution.For this system define threshold according to R& D department. If Probability distribution is above threshould that form field will be displayed by default and other option will get hide in more option to reduce cofusion for data entry user.

Form Field Re-asking after entry

After form filled reconfirm likely wrong responses. As a simple example if user enters age as 200 system will reask question. System will give some threshold value for each question. If value of that field is above the threshold, then system reask that question again. As system will not reask every question each time, the efficiency of system will be better.Here system uses the Betta distribution.

The complete working approach of model is represented in the algorithm 1 depicted below:

Algorithm 1

// input: Input Dataset

Form field Ordering

H (F$_i$) ⟵ Calculate entropy for each form field

If H (F$_i$) > T1(Threshold) then

Form field value= More Uncertain.

Else form field value= less uncertain can be predicted by previous value.

O/P: Reordered Form fields.

H (F$_i$ |G) ⟵ Calculate conditional entropy for each field

If H (F$_i$ |G) > T2 (Threshold) then

Get dependency between form fields. With this other form field values can be predicted.

System also predicts values by calculating maximum posterior hypothesis.

Form field Reformulation during entry.

System uses probability distribution to predict mostly used form field as discussed in previous example.

P(X) =number of occurrences of x/ number of tuples.

If P(X)> T3 (Threshold) then

X will be displayed by default in combo box and others will be hidden as more option.

Form field re-asking after entry.

System uses beta distribution to check correct values. It uses α and β as hyper parameters.

$\lambda \sim Beta(\alpha, \beta)$

If $\lambda == 1$ then inserted value is incorrect value

Else correct value.

// output: Improved dynamic form

IV. Results and Discussions

Some experiments are reported to show the effectiveness of the proposed system. Selecting a suitable dataset is a critical and important step in designing proposed system. Currently Diabetes dataset used for getting previous form fields & prior its values. System takes 1000 patient with 20 different attributes. These input dataset were then kept as source file to system. System uses 50% records for training and 50% for testing.

Practicability of System Demonstration

a) Reordering Experiments

Figure 2 shows the reordering experiment results. By reordering of form fields system requires only 0.5 seconds to fill data values. Without reordering of form fields system requires only 0.8 seconds to fill data values.

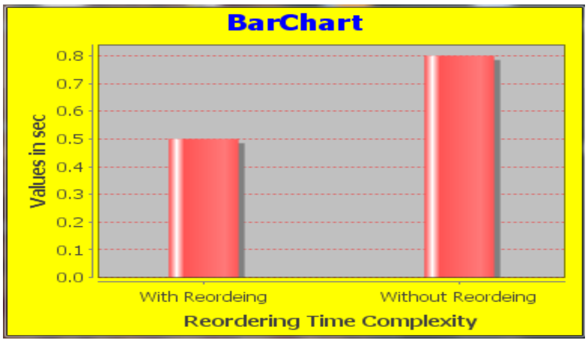


Figure 2. Result of reordering experiment

b) Reasking Experiments

Figure 3 shows the re-asking experiment results. By re-asking of form fields system gives 90% correct values and 10% gives wrong answer.

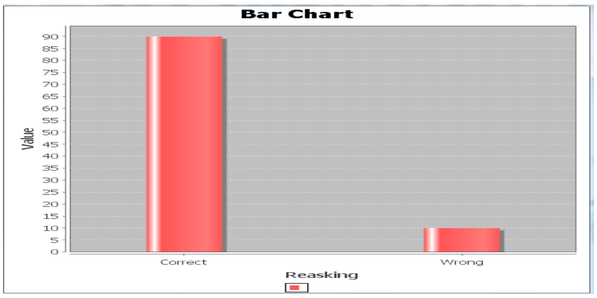


Figure 3. Result of reasking experiment

c) Comparison of Maximum Posterior Hypothesis and conditional entropy for prediction of form fields.

Figure 4 shows the comparison between Naïve byes and Bayesian network. Naïve byes use maximum posterior hypothesis formula to predict form field values. Bayesian network uses conditional entropy formula to predict form field values. This experiment shows that a naïve bye predicts 94% accurate values where as Bayesian network gives 90% accurate values.

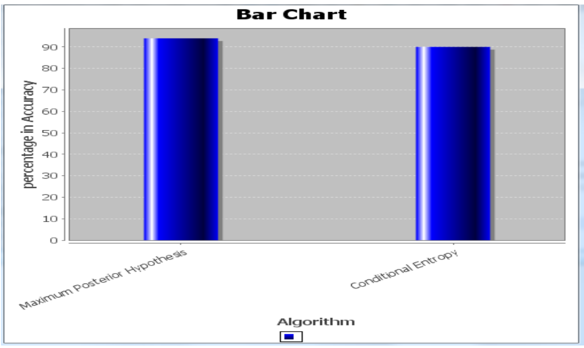


Figure 4. Result of comparison of Naïve byes (MPH) and Bayesian network(CE)

V. CONCLUSION

By using this system, data entry operator can fill form in less time as compare to current practice. Bayesian Network Model consists of data-driven insights to automate multiple steps in the data entry pipeline. Before entry, it finds an ordering of form fields that promotes rapid information capture, driven by the entropy principle and can statically reformulate form fields to promote more accurate responses. During entry, it dynamically adapts the form based on entered values, facilitating re-asking, reformulation and real-time interface feedback in the spirit of providing appropriate entry friction. After entry, it automatically identifies possibly-erroneous inputs, guided by contextualized error likelihoods and re-asks those form field values, possibly reformulated, to verify their correctness. The simulated empirical evaluations demonstrate the data quality benefits of each of these components: form field ordering reformulation and re-asking. Naïve byes (uses maximum posterior hypothesis) give more accurate predicted values as compare to Bayesian network (uses Conditional entropy).

REFERENCES

[1] J. M. Hellerstein. "Quantitative data cleaning for large databases" United Nations Economic Commission for Europe (UNECE),2009,pp. 21-32.

[2] K. Chen, H. Chen, N. Conway, T. S. Parikh, and J. M.Hellerstein, "Usher: Improving data quality with dynamic forms," In Proceedings of the International Conference on Data Engineering, 2011, pp. 596-614.

[3] S.Patnaik, E.Bruskill and W.Thies, "Evaluating the Accuracy of Data Collection on Mobile phones: A Study of Forms, SMS and Voice", Proc.IEEE (ICTD), 2009. pp. 200-214.

[4] K. Chen, H. Chen, T. S. Parikh, and J. M.Hellerstein, "Designing Adaptive Feedback for Improving Data Entry Accuracy," In Proceedings ACM (UIST), 2010, pp. 16-23.

[5] Erhard Rahm,Hong Hai Do. "Data Cleaning: Problems and Current Approaches", International Conference on Data Engineering, 2010, pp. 500-509.

[6] A.Ali and C.Meek, "Predictive models of Form Filling," Technical Report MSR-TR-2009-1v, Microsoft Research, Jan. 2009, pp. 23-29.

[7] L.A.Hermens and J.C.Schlimmer, "A Machine-Learning Apprentice for the Completion of Repetitive Forms," IEEE Expert: Intelligent Systems and Their Applications,vol. 9,no. 1,Feb. 1994, pp. 28-36.

[8] Y. Yu, J.A. Stamberger, A. Manoharan, and A. Paepcke, "Ecopod: A Mobile Tool for Community Based Biodiversity Collection Building," Proc. Sixth ACM/IEEE CS Joint Conf. Digital Libraries (JCDL), 2006, pp. 20-29.

[9] Chickering, D. M.,"Optimal structure identification with greedy search, "Journal of Machine Learning Research, Volume 3, (2002), pp. 547-554.

[10] C. Batini and M. Scannapieco, Data Quality: Concepts, Methodologies and Techniques.Springer, 2006.

[11] K. Kleinman, "Adaptive double data entry: a probabilistic tool for choosing which forms to reenter, " Controlled Clinical Trials, 2001.

[12] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning bayesian networks: The Combination of knowledge and statistical data," Machine Learning, vol.20, no.3, pp.197–243, 1995.