# Improving Emergency Data Quality by Noisy and Error Identification

Gong Zhang, Trevor Strome, Simon Liao, Zhaopeng Fan

**Abstract**: For emergency department data, it is important to achieve the high quality data. Biomedical data errors or noise tend to fall in data entry inaccuracy, medical device limitations, data transmission errors, or man-made perturbations frequently result in imprecise or vague data. To find these errors or noise, a data mining algorithm was build to for identifying errors or noisy values and using the remaining correct data sets for subsequent modeling and analysis. This approach was to build a business rule based naïve bayes classifier. With this model, the error or noisy patterns can be discovered from datasets that were feeding to emergency room data repository.

## 1. INTRODUCTION

With the growing adoption of electronic health record (EHR) applications, the volume of medical data in repositories such as data warehouses is growing at a substantial rate. Medical Data repository collect and manage data from many diverse source systems, such as laboratory, pharmacy system, and electronic medical record (EMR) system, etc,. Through current information system can automatically retrieve information form data sources, error and noise information can still be feed into the medical data repository [1]. Because the medical data repository can be used in many purposes and needs, the highest accuracy diagnosis and data analysis are depend on the quality of data.

Zhu and Wu [2] state that the quality of a dataset can be characterized by namely attributes and class labels. The quality of the attributes is characterizing instances for classification purposes and the quality of class labels shows the correct assigning the class labels. Noise is often similarly divided into two major categories that are class noise. (Misclassifications or mislabeling) and attribute noise (errors introduced to attribute values). Many studies have shown that data quality is the main reason for inferior decision. Based on enhanced data consistency and less confusion among the underlying data, eliminating the data noise and errors leads to more accurate data mining [3], [4], [5], [6].

One standard algorithm was tested to evaluate how well it can satisfy our three desiderata for noise classification of accuracy, comprehensibility, and stability. This standard approach is Naïve Bayes [7], which is based on probability.

## 2. EMERGENCY DEPARTMENT DATA QUALITY

Many Emergency Departments (EDs) are adopting ED information systems (EDIS) to track patient flow and manage patient care. Due to the often frenetic pace of ED patient care, accurate data entry is (rightfully so) a lesser priority than patient care. Emergency Department patient care data available through EDIS systems is being used increasingly for quality and process improvement, sentinel disease surveillance, and clinical decision support applications.

There are several causes of data quality issues in data derived from EDIS systems:

1. Missing information because it is not available (i.e., health insurance information not available because patient did not bring health insurance card)
2. Missing information because it was not entered by user
3. Incorrect data not detected by user (or correct data not available)
4. Incorrect data that cannot be corrected due to limitations in the system

There are potentially serious consequences related to EDIS data quality. When used for critical clinical incident

reviews or infection control measure implementation or follow-up, “mundane” data like patient location and patient treatment/status data is vital to painting an accurate re-creation of the Emergency Department at some critical time. Missing health insurance information can prevent studies of between-hospital patient movements identify issues related to frequent ED users. Process improvement initiatives cannot be properly evaluated if data used to study patient flow is not reliable. See Table 1 for Four major errors of ED data quality issues and the impact on how the data can be utilized.

| | | |
|---|---|---|
| Data Omission | Missing key identifiers (i.e., health insurance information, health record number) | • Reduced ability to track multiple visits within and/or across sites.<br>• Potential impact on billing and revenue generation |
| Application Data Filtering / Validating | EDIS application does not filter items in drop-down menu to accommodate for different patient situations. | • List of non-admitted discharge types appear when selecting discharge disposition for admitted patients. |
| Selection Errors | Certain data in EDIS cannot be edited once selected:<br>• Status<br>• Location<br>• Providers | • Inaccurate determination of lengths of stay in locations, patient wait times, and provider performance. |
| User Variation | Users apply their own personal biases and/or bad habits when interacting with EDIS systems. | • Codified data fields in drop-down menus discarded in favor of free-text entry<br>• Differing interpretations of business-rules and standard work as they apply to the EDIS system |

Table 1 Four major errors of Emergency Department Information System data issues

Emergency Department data quality issues can be mitigated by several means. Primarily, EDIS application software can be designed to reduce and or eliminate opportunity for data entry errors through consistent use of data validation rules, user interfaces that permit (or make easier) the correction of incorrectly entered data, and intelligent list filtering.

## 3. PREPROCESSING THE DATA

There are two types of noise, attribute noise and class noise. Compared to class noise, attribute noise tend to happen more often in real world. If attribute values are predicable, attribute noise can be identified and error data sets can also be identified. However, for unpredictable attribute values, an algorithm has been developed to identify the suspicious noise instance.

Algorithm 1: Identifying noise for unpredictable attributes

Input: possibly noisy instance $D$; a set of business rules $L$;
Output: a set of suspicious instance $IS$;
Begin
  foreach instance $I_i \in D$
    flag = 0;
    foreach Rule $R_j \in L$
      if $I_i$ satisfies $R_j$ {flag = 1; break;}
    if flag == 0 {push $I_i$ into $IS$;}
End

When all suspicious instances have been identified, these instances will be input to a Naïve Bayes classifier for noise instance determination.

## 4. NAÏVE BAYES CLASSIFIER

With a set of noise instances as training data, the Naive Bayes classifier computes estimates of conditional probabilities and uses those probabilities in Bayes' rule to determine the most probable class noise instance.

Let $C_z$ be the target noise class among T (Total) classes. For an instance $x_k$, $P(C_z \mid x_k)$ is the predicated probability by the Bayes theorem. In practice, we propose to explore the values independently for each attribute. Let $P_j(C_z \mid x_k, )$ denotes the probability that

example $x_k$ belongs noise. The expected classification error can be minimized by choosing the maximal posterior probability,

The following formulation is to get maximal posterior probability of the instance $x_k$:

$$\arg\max_z P(C_z \mid x_k)$$

For $x_k$ belonging to the class $t$ ($t \neq z$), it could produce $P(C_z \mid x_k, b) > P(C_t \mid x_k)$.

A naive Bayes classifier assumes that all the attributes are independent given the noise. This assumption drastically reduces the necessary computations. Using the Bayes theorem, the computation for the conditional probability of a class $C_z$ is:

$$P(C_z \mid x_k) = \frac{P(C_z)P(X_j \mid C_z)}{P(X_j)} \quad (1)$$

Assuming the attributes $x_j = \{a_1, a_2, \ldots, a_m\}$ are independent given class label, the probabilities $P(X_j \mid C_z)$ can be decomposed into the product $P(a_1 \mid C_z)$ X $P(a_2, \mid C_z)$ X … X $P(a_M \mid C_z)$, M is the number of the attributes. The denominator of the equation above normalizes the result so that $\sum_z P(C_z \mid x_k)=1$.

With computation of $P(C_z \mid X = x_k)$, equation 1 can be written in the form of Equations 2:

$$P(C_z \mid x_k) = \frac{P(C_z)\prod_{j=1}^{M} P(X = a_j \mid C_z)}{P(X_j)} \quad (2)$$

3)

Equation 2 can be rewritten to (3) to avoid $\prod_{j=1}^{M} P(X = a_j \mid C_z) = 0$

$$L_z = \log(P(C_z)) + \sum_{j=1}^{M} \log(P(X = a_j \mid C_z)) \quad (3)$$

In practice, the attributes may not be independent, classification accuracy of Naïve Classifier still be good.

## 5. Results

To measure the reliability of our approach, we tested it on Winnipeg Regional Health Authority Emergency Program Data. This database consists of 10000 records.

A training set of 1000 error records with represented for the four major errors were chosen to train a naive Bayes classifier. The remaining examples constitute a test set.

Table 2: Identification accuracy for training and testing data

| Error/ Noises Types | Identification Accuracy for training Data | Identification Accuracy for testing Data |
|---|---|---|
| Data Omission | 95% | 93% |

| | | |
|---|---|---|
| Application Data Filtering | 80% | 74% |
| Selection Errors | 82% | 78% |
| User Variation | 75% | 69% |

If the error types can be defined by the clear business rules, the identification accuracy is high, such as data omission and selection errors. Otherwise, with many unpredicted error cases, the identification accuracies are relatively low, such as Application Data Filtering. For user variation, there is a wide variance on attribute values, such as inputting patient name into chief complain. Therefore, identification accuracy for user variation error is the lowest. Fortunately, user variation errors mainly happen in less important attributes. The overall accuracy is 79%.

Noise and error identification for Emergency data is a very important step of the data cleansing and consequently of great practical importance. Because the emergency data is collected from many data sources and involved a lot of human factors, there are the many technical challenges in this area and only a few researches has been done. For existing commercial software based on manually input rules it is very hard to set up and identification accuracy is low. In this paper, we presented an automated approach using business rule based naïve bayes classifier and the experiment result is encouraging. The future work is involved automatic identifying the uncertain instances for self leaning.

## REFERENCES


1. KOHN LT, CORRIGAN JM, DONALDSON MS, ED. To Err Is Human: Building a Safer Health System, National Academy Press Washington, DC 1999.
2. X. ZHU, and X. WU, “Class noise vs. attribute noise: a quantitative study of their impacts”, Artificial Intelligence Review 22 (3-4), 2004, pp. 177-210.
3. X. Zhu, X. Wu, and Q. Chen, “Eliminating class noise in large datasets,” Proc. ICML, 2003, pp. 920–927.
4. C. Brodley and M. Friedl, “Identifying mislabeled training data,” J. Artificial. Intelligence. Research, vol. 11, pp. 131–167, 1999.
5. D. Gamberger, N. Lavrac, and C. Groselj, “Experiments with noise filtering in a medical domain,” Proc. 16th ICML Conf., San Francisco, CA, 1999, pp. 143–151.
6. G. H. John, “Robust decision trees: Removing outliers from databases,” Proc. 1st Int. Conf. Knowl. Discovery Data Mining, 1995, pp. 174–179.
7. P. DOMINGOS AND M. PAZZANI, “On the optimality of the simple Bayesian classifier under zero-one loss,” Mach. Learn., vol. 29, no. 2/3, pp. 103–130, Nov./Dec. 1997.