Journal of Biomedical Informatics 42 (2009) 356–364

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery

L.M. Taft [a,*], R.S. Evans [a,e], C.R. Shyu [a,c], M.J. Egger [a], N. Chawla [d], J.A. Mitchell [a], S.N. Thornton [a,e], B. Bray [a], M. Varner [b]

[a] Department of Biomedical Informatics, University of Utah Health Sciences Center, School of Medicine, 30 North 1900 East, Salt Lake City, Utah 84132, USA
[b] Department of Obstetrics and Gynecology, University of Utah, School of Medicine, USA
[c] Informatics Institute, University of Missouri Columbia, USA
[d] Department of Computer Science & Engg., University of Notre Dame, USA
[e] Department of Medical Informatics, Intermountain Healthcare, USA

ARTICLE INFO

Article history:
Received 2 June 2008
Available online 14 September 2008

Keywords:
Adverse drug events
Pregnancy
Labor and delivery
Oversampling
Data-mining

ABSTRACT

Background: The IOM report, *Preventing Medication Errors*, emphasizes the overall lack of knowledge of the incidence of adverse drug events (ADE). Operating rooms, emergency departments and intensive care units are known to have a higher incidence of ADE. Labor and delivery (L&D) is an emergency care unit that could have an increased risk of ADE, where reported rates remain low and under-reporting is suspected. Risk factor identification with electronic pattern recognition techniques could improve ADE detection rates.

Objective: The objective of the present study is to apply Synthetic Minority Over Sampling Technique (SMOTE) as an enhanced sampling method in a sparse dataset to generate prediction models to identify ADE in women admitted for labor and delivery based on patient risk factors and comorbidities.

Results: By creating synthetic cases with the SMOTE algorithm and using a 10-fold cross-validation technique, we demonstrated improved performance of the Naïve Bayes and the decision tree algorithms. The true positive rate (TPR) of 0.32 in the raw dataset increased to 0.67 in the 800% over-sampled dataset.

Conclusion: Enhanced performance from classification algorithms can be attained with the use of synthetic minority class oversampling techniques in sparse clinical datasets. Predictive models created in this manner can be used to develop evidence based ADE monitoring systems.

1. Background

The Institute of Medicine (IOM) in the report, *Preventing Medication Errors* [1] recommended the implementation of decision support tools derived from evidence based knowledge and patient information as part of the strategies to prevent medication errors (ME). The report also recommended the active monitoring of medication use to promote prevention strategies. Although medical research has actively pursued these problems, the reported incidence of ME is suspected to be under-estimated [1–3].

These IOM reports [1,2] define ME as avoidable errors occurring in the medication use process. Adverse drug event (ADE) is a more inclusive definition that covers both ME and adverse drug reactions.

Operating rooms, emergency departments and intensive care units are known to have a higher incidence of ADE [4]. Labor and delivery (L&D) areas are considered by quality assurance groups as special care units and pregnant women are considered by the FDA as a vulnerable group for ADE [1]. L&D provides emergency care and therefore should also be treated as a high risk area Studies published in the literature focus on specific drugs and anesthesiology events [5–9]. To the best of our knowledge there are no published studies of ADE as a general category in pregnant women. Our findings indicate an incidence of 0.34% of ADE in women admitted to L&D. This incidence is surprisingly low in a population that includes at least 10% of high risk pregnancies that require poly-pharmacy [10].

One of the most complex tasks in the design and development of automated decision support tools is evidence based rule generation and knowledge extraction from existing data [11]. The task is even more challenging in those cases where the class label of interest or ADE patients as in this case, has an incidence of 1% or less [12]. Datasets with these characteristics are also known as skewed or imbalanced datasets. The class of interest is relatively rare and there are important trade-offs in the decision between false negatives and/or false positives. Overall, it is more costly to have a false negative versus a false positive. More so in a medical application

* Corresponding author. Fax: +801 581 4297.
E-mail address: Laritza.Taft@hsc.utah.edu (L.M. Taft).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.09.001

where the interest is detecting patients with adverse outcomes that can be prevented. Without loss of generality, we will assume that the larger class or the majority class is the negative class and the class of interest is the minority (smaller) or positive class. We will use these terms inter-changeably in the paper. The use of machine learning algorithms in sparse datasets with class imbalance causes suboptimal classification performance as these techniques get overwhelmed by the majority class. Recent work has focused on sampling techniques that counter the problem of class imbalance by either oversampling the minority class or under-sampling the majority class [12–15].

In this paper, we focus on the application of the Synthetic Minority Over Sampling Technique (SMOTE). SMOTE works by generating new instances from the existing cases. SMOTE effectively counters the imbalance in data by not only solving the problem of high class skew but also the problem of high sparsity. It works in the "feature space" rather than "data space". The synthetic samples are created by taking each minority class sample and the *k* nearest neighbors. The synthetic sample shares features of both the chosen minority class sample and one or more of the nearest neighbors. This approach effectively forces the decision region of the minority class to become more general. The synthetic cases will not only increase the data space but will also amplify the features of the minority class without duplicating the original data. SMOTE's effectiveness has been shown in a variety of domains and with a variety of classifiers [15,16].

The objective of the present study was to apply SMOTE as an enhanced sampling method using a sparse dataset and to identify a prediction model for ADE in women admitted for L&D based on patient risk factors and comorbidities. We would like to note here that we tried other of oversampling methods like replication and random under-sampling but none of them resulted in improvement. Hence, for clarity of presentation in the paper, we only focus our discussion and results on using SMOTE.

1.1. Description of data-mining techniques

Machine learning techniques include both data sampling and learning algorithms. Over sampling techniques are applied to reuse the available data by dividing the dataset into three or more sets. Once the data sampling step is completed, the classification algorithms are applied to the resulting datasets. Subsequently, the performance of the classifiers is evaluated by comparison of the results in the training, testing and validation datasets.

SMOTE was used to generate new synthetic cases for this study. The computations for the new synthetic sample variables are based on Euclidian distance for continuous variables and the Value Distance Metric for the nominal features. The continuous variable values are created by taking the difference in distance between two existing minority class samples and multiplying that difference by a random number between 0 and 1. The resulting number is added to the feature value of the original sample and the result will be the value of that variable in the new synthetic sample. For nominal variables, the variable value is assigned by majority vote of the *k* nearest neighbors. As a result, the synthetic cases will have attributes with values similar to the existing cases and not just replications as provided with oversampling. The objective is to increase the representation of the minority class in the resulting dataset and reflect the structure of the original cases. By adding new samples of similar characteristics to the originals the decision region is amplified and there should be improvement of the evaluation measures: true positives and the area under the curve (AUC). The newly created cases are appended to the original dataset in 100% increments. Thus the "second" dataset will have 100% more minority class cases, the third 200% more minority class cases and so forth. This technique has proven to be useful in improving prediction of sparse datasets by other authors [14].

1.2. Classification algorithms

Naïve Bayes is a simple probabilistic classifier based on Bayes' theorem with strong (naive) independence assumptions. Bayes' theorem is based on the conditional probability theory; the posterior probability is proportional to the product of the prior probability and likelihood. With the independence assumption, the Naïve Bayes classifier over-simplifies the models. It avoids the complexity of producing the joint probabilities across features, which quickly becomes overwhelming by the large number of features. While the assumption of independence is "naïve", it has been shown to perform exceptionally well in classification in the medical field [17,18].

Decision trees are predictive models that allow the selection of an attribute that will serve as the root node for prediction. Based on the probability distribution chance of occurrence and gain or utility of the root nodes, the leaf nodes (or branching nodes) are created [17]. Decision trees are inductive learners that have proven to perform well in clinical research. The interpretation is facilitated for domain knowledge experts by the display in graphical form. C4.5 is a popular decision tree learning algorithm used in a multitude of domains. We used the WEKA [17] (Waikato environment for knowledge analysis) Open Source Software implementation of C4.5, namely JR48, in our experiments.

Naïve Bayes and decision trees were chosen as the classification algorithms for the experiments because the results are in a format that facilitates interpretation by domain experts. The graphical representation of the decision trees and the simplicity of the Naïve Bayes model are easily understood as opposed to the "black box" that other algorithms such as Neural Networks and Vector Machines generate [19].

2. Methods

2.1. Subjects

Records for the present study came from the Enterprise Data Warehouse (EDW) of Intermountain Healthcare in Salt Lake City, Utah. The EDW contains clinical care and coded data for billing and reporting. Data from 135,000 individual patients admitted for L&D during years 2002–2005 were extracted. The variables included demographic characteristics and discharge diagnosis as well as maternal and fetal outcomes and maternal comorbidities.

Inclusion criteria were post partum women with gestational ages between 20 and 44 weeks and birth weight between 500 and 4800 g. Two patient's records with maternal age above 55 were excluded as they were confirmed to be data entry errors. In patients with multifetal pregnancies, the outcome data of the first-born infant were selected for inclusion.

2.2. Data preprocessing

A classification methodology for outcomes and comorbidities was created based on the clinical classification of ICD9 codes for labor and delivery published by Yasmeen et al. [20] and on the reportable adverse events criteria published by the Joint Commission and the Utah Department of Health [21,22]. In interest of clarity we called these tables "published classifications".

The published classifications included ICD9 codes assigned to obstetrical diagnosis, pregnancy related comorbid diagnoses, procedures and for sentinel events. For example the diagnosis "diabetes mellitus" includes ICD9 codes: 250.xx, 357.2, 362.0, 648.0x. We