*MEDINFO 2010*
*C. Safran et al. (Eds.)*
*IOS Press, 2010*

*doi:10.3233/978-1-60750-588-4-856*

# A Model Driven Approach to Imbalanced Data Sampling in Medical Decision Making

**Hong-Li Yin, Tze-Yun Leong**

*Medical Computing Lab, School of Computing, National University of Singapore*

**Abstract**

*Classification is an important medical decision support function that can be seriously affected by disproportionate class distribution in the training data. In medical decision making, the rate of misclassification and the cost of misclassifying a minority (positive) class as a majority (negative) class are especially high. In this paper, we propose a new model-driven sampling approach to balancing data samples. Most existing data sampling methods produce new data points based on local, deterministic information. Our approach extends the idea of generative sampling to produce new data points based on an induced probabilistic graphical model. We present the motivation and the design of the proposed algorithm, and compare it with two representative imbalanced data sampling approaches on four medical data sets varying in size, imbalance ratio, and dimension. The empirical study helped identify the challenges in imbalanced data problems in medicine, and highlighted the strengths and limitations of the relevant sampling approaches. Performance of the model driven approach is shown to be comparable with existing approaches; potential improvements could be achieved by incorporating domain knowledge.*

***Keywords:***

Random sampling, Synthetic Minority Over Sampling (SMOTE), Model driven sampling, Imbalanced data learning

## Introduction

Classification is an important medical decision support function that is often seriously affected by disproportionate class distribution in the training data. A main motivation of this work is the data imbalance challenge we have encountered in building classifiers in many real-life medical domains, including head injury, asthma, breast cancer, etc. An imbalanced data set contains a disproportionately high number of data in one or more classes than those for a class that is of interest. Traditional machine learning methods cannot work well with such data to build an accurate classifier; they tend to bias toward the majority class data and result in a low positive rate. In medical decision making, the cost of misclassifying a minority (positive) class as a majority (negative) class is especially high. The imbalanced data problem is increasingly being actively addressed in the field; some recent work include: Mazurowski et al. [1] studied the effect of imbalanced data to neural networks in medical decision making; Cohen et al. [2] deployed existing imbalanced data learning techniques in the surveillance of nosocomial infection.

There are two main categories of approaches to address the imbalanced data problem: 1) Algorithmic level approaches, where new machine learning algorithms are proposed or standard machine learning algorithms are modified to accommodate imbalanced data, *e.g.* learning rare class only [3], cost sensitive learning [4], etc., and 2) Data level approaches, which re-sample the imbalanced dataset to produce a new training dataset with balanced class distribution. We focus on the data level approaches in this work.

In this paper, we propose a Model Driven Sampling (MDS) approach to solving the imbalanced data problem. Unlike existing sampling approaches that use local, deterministic information to generate new data points, MDS learns from the whole labeled data set and possibly domain knowledge to induce a probabilistic graphical network to generate new data points. We also examine the performance of MDS as compared with two representative sampling approaches - Random Sampling (RS), and Synthetic Minority Over Sampling (SMOTE), on imbalanced data sets with different characteristics in medicine. We compare the three approaches on four medical data sets varying in complexity or dimension, data size and imbalance ratio, using different machine learning algorithms to build the resulting classifiers.

## Related Work

Random sampling generates duplicated data without creating any new information. Random under-sampling randomly removes instances from the majority class to balance the class distribution. The disadvantage is that there is a risk in deleting useful information. Random over-sampling, on the other hand, randomly duplicates instances in the minority class. The disadvantage is that the decision region for the minority class may become more and more specific and possibly lead to data over fitting. In this paper, random sampling includes both random over-sampling and random under-sampling.

On the other hand, SMOTE creates synthetic data along the line between two nearest data points. In SMOTE, the minority class is over-sampled by taking each minority sample and introducing synthetic examples along the line segments joining with any of the k (k is 5 in the current implementation [5]) nearest minority neighbors. Chawla [5] showed that the synthetic examples generated by this technique cause the classifier to create larger and less specific decision regions as compared to random over-sampling.

Progressive sampling is systematically described by Foster et al. [6]. It was later used in [7, 8] for imbalanced data learning. However, the main advantage of progressive sampling is to improve the system efficiency by making use of minimal training data. The follow-on work [7, 8] either assumes there is sufficient training data, or use random sampling when there is insufficient data [7].

Random sampling is case duplicating based on one data point; SMOTE generates synthetic data based on two data points. They are representative of the sampling techniques that generate data based on local information. Recently, Liu et al.[9] proposed a generative oversampling approach which attempts to generate data based on the probability distribution of the minority data. This is similar to the idea presented in this work, with a major difference in the form of the probabilistic distribution generated.

## A Model Driven Sampling Approach

In many biomedical domains, minority data can be sparse, but domain knowledge is commonly available. Model Driven Sampling (MDS) is an approach to learning from the whole training data set (both minority and majority, learning from majority can prevent generating noisy minority), and supports incorporation of domain knowledge into the induced model.

In contrast to the generative sampling approach [9], which builds a probabilistic distribution as the generative model, We induce a probabilistic graphical model or Bayesian Network for data generation. Bayesian Network is a factored representation of a probability distribution, representing the probabilistic relationships among a set of random variables. For example, Figure 1 shows the commonly cited Asia network. In a Bayesian network, the nodes indicate the random variables; the arcs indicate conditional dependencies, and there is a conditional probability distribution embedded in each node, denoting the conditional relationships of the values of the random variables with respect to different configurations of its parent nodes. The advantage of Bayesian Network is that it can easily combine both observational (data) and domain knowledge. For example, the causal relationship between smoking and cancer can be added from domain knowledge if it was unknown from the training data.

The MDS algorithm is as follows: As shown in Figure 2, we first build a model A from the original data set; model A generates new data based on the characteristics of the whole data set; the generated data is combined with the original data to form a new training data set to train classifiers. Specifically, we learn a Bayesian model using the K2 structure learning algorithm [10] from the data set. Then we generate minority data from the model using the Markov Chain Monte Carlo (MCMC) method [11].

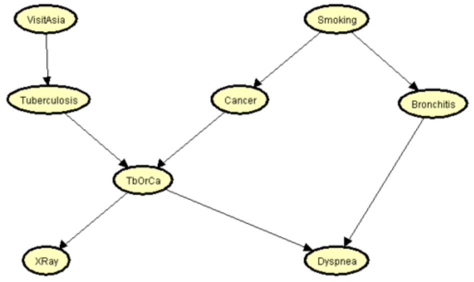

*Figure 1 - Bayesian Network model for the Asia data*

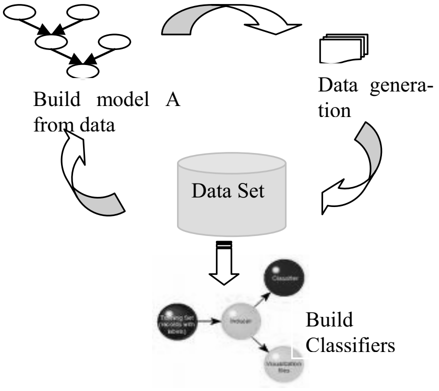

*Figure 2 - Work flow in Model Driven Sampling*

Comparing to other common sampling approaches, MDS generates new data based on the whole data set, thus the generated data would be more “meaningful” from a global perspective than the data generated randomly (random sampling) or data generated from local information (SMOTE). Due to the lack of domain knowledge and in fairness, we did not consider domain knowledge in the comparison analysis reported below, although we demonstrated the feasibility and promise of its inclusion.

## Datasets

The data sets selected for the analysis span a wide spectrum in terms of complexity or dimension, imbalance ratio, and size; they are meant to illuminate the strengths and limitations of the algorithms studied under different conditions in medical domains. The data sets are: Asia, Mammography, Indian Diabetes, Asthma First Visit data. The Asia data set is commonly used in machine learning communities as examples illustrating

Bayesian Network learning. The asthma data set is collected, under proper approval and usage guidelines, from the hospitals in Singapore. The other two data sets are from the UCI Machine Learning repository [12] which were used for imbalanced data learning in [5]. The characteristics of the data sets are: binary data, unevenly distributed with different imbalance ratios (IB) as shown in Figure 3 and Table 1. IB ratio is equivalent to the percentage of minority examples in the training data; the lower of the value the more imbalanced the data is.

*Table 1 - Class distributions (in numbers)*

| | Majority | Minority | IB ratio | Features |
|---|---|---|---|---|
| Asia | 530 | 42 | 0.073 | 7 |
| Indian Diabetes | 500 | 268 | 0.349 | 8 |
| Mammography | 10923 | 260 | 0.023 | 6 |
| Asthma First Visit | 678 | 213 | 0.239 | 40 |

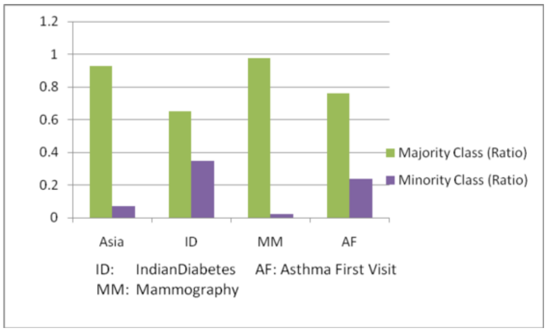

*Figure 3 - Data class distributions (in ratio).*

The Asia data set is about people who visited Asia and whether they had developed dyspnea or not. In our experiment, the Asia data set includes 42 positive cases, and 530 negative cases.

The Pima Indian Diabetes [12] data set includes 2 classes and 768 samples. The data is used to identify the positive diabetes cases in a population near Phoenix, Arizona. There are only 268 positive class samples.

The Mammography data set has a high skewed ratio: 10923 negative examples versa 260 positive examples. The trained classifier needs to be highly sensitive to detect the positive cases.

The Asthma First Visit data records the information when asthma patients visit the respiration centre for the first time. It has 40 attributes recording patients' general information, asthma history, treatment history, etc. There are 213 positive samples out of total 891 samples. The main problem is to determine whether a patient will encounter any control failure in the future based on the information provided on his first visit.

## Experiments

We conducted experiments on the four data sets using three classifiers (C4.5 decision tree, Bayesian Network, Support Vector Machine) and three data sampling techniques (RS - Random Sampling , SMOTE, MDS). The experiment design is as shown in Figure 4, and each experiment ran through 10 fold stratified cross validation. The original data was split into 10 folds, and in each fold, training data was sampled by various approaches to build a new model which would run on the testing data.

Prediction accuracy cannot be used as the evaluation criteria because it is shown to discriminate the minority classes [13]. As shown in Equation (1), we use the geometric mean(g-mean) [14] of the accuracies measured separately on each class (where $a^+$ is true positive rate and $a^-$ is true negative rate) as our evaluation criteria. The basic idea behind this measure is to maximize the accuracy on both classes.

$$\text{g-mean} = \sqrt{a^+ \times a^-} \qquad (1)$$

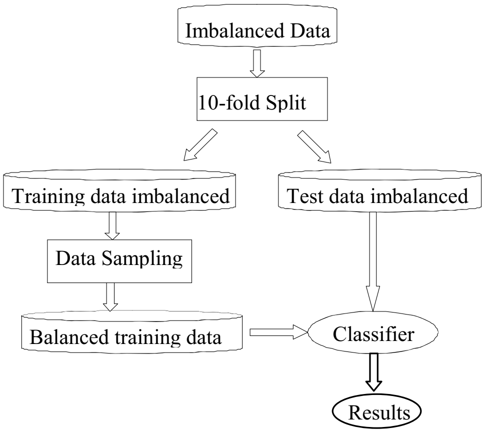

*Figure 4 - Experiment design*

Only the Bayesian Network (BN) results are reported in detail in this paper. This is because, as shown in Table 2, while the Bayesian Network performance may not always be the best in all experiments, it is generally more stable than others.

*Table 2 - Performance of various algorithms on MDS*

| | Asia | Asthma | Diabetes | Mammography |
|---|---|---|---|---|
| BN | 0.88 | 0.759 | 0.759 | 0.622 |
| C4.5 | 0.266 | 0.591 | 0.771 | 0.886 |
| SVM | 0.428 | 0.591 | 0.777 | 0.874 |

### Asia Data

The Asia data set has the lowest number of minority examples and the second lowest imbalance ratio 0.073. As shown in Table 3, the original data set without any sampling has a high prediction rate on its majority samples (98.7%), but a low prediction accuracy on its minority samples (7.1%), thus the overall performance is the lowest at 26.5%. Random sampling and SMOTE both significantly improve the predictions on minority samples and achieve a much better overall performance. MDS achieves the best performance 88% overall and 90.5% on minority data set.

*Table 3 - Asia data running results*

| | Original Data | RS | SMOTE | MDS |
|---|---|---|---|---|
| TP[1] | 0.071 | 0.881 | 0.69 | 0.905 |
| TN[2] | 0.987 | 0.863 | 0.925 | 0.856 |
| G-Mean | 0.265 | 0.872 | 0.799 | 0.88 |

### Indian Diabetes Data

Indian Diabetes data is a relatively balanced data set with the highest imbalance ratio at 34.9%. Therefore, without any sampling, the original data set can achieve a satisfying performance on minority data and a good overall performance. The three sampling approaches equally improve the performance especially on minority by11%. The overall performance is not much improved. (As shown in Table 4)

*Table 4 - Indian Diabetes data running results*

| | Original Data | RS | SMOTE | MDS |
|---|---|---|---|---|
| TP | 0.669 | 0.783 | 0.787 | 0.772 |
| TN | 0.836 | 0.741 | 0.745 | 0.745 |
| G-Mean | 0.748 | 0.762 | 0.766 | 0.759 |

### Mammography Data

Although Mammography data set has the lowest imbalance ratio 0.023, it is still relatively simple as it has only 6 features which result in a low data complexity. In

Table **5**, the original data set can achieve 85% overall performance. The other approaches can equally improve the minority prediction by 15%. SMOTE has the best overall performance 89%, and MDS has a comparable performance of 88.3%.

*Table 5 - Mammography data running results*

| | Original Data | RS | SMOTE | MDS |
|---|---|---|---|---|
| TP | 0.735 | 0.888 | 0.873 | 0.885 |
| TN | 0.981 | 0.857 | 0.908 | 0.881 |
| G-Mean | 0.849 | 0.872 | 0.89 | 0.883 |

### Asthma First Visit Data

The Asthma First Visit data has 40 features, the highest dimension among all. In Table 6, none of the approaches can achieve a good performance. Relatively, random sampling gives the best minority prediction (15% improvement) and the best overall performance (5% improvement). MDS approach ranks the second.

*Table 6 - Asthma First Visit data running results*

| | Original Data | RS | SMOTE | MDS |
|---|---|---|---|---|
| TP | 0.419 | 0.576 | 0.448 | 0.5 |
| TN | 0.852 | 0.732 | 0.805 | 0.775 |
| G-Mean | 0.598 | 0.649 | 0.6 | 0.622 |

## Discussion

There are three important challenges for learning with imbalanced data sets: 1) the imbalance ratio, 2) the absolute size of the minority data, 3) the dimension of the data set. The three factors are common in most medical data sets, and they vary among the four representative data sets chosen in this work. We have examined relatively easy problems which are less imbalanced, low dimensional, with sufficient minority samples (e.g., Indian Diabetes and Mammography datasets), to hard problems which are highly imbalanced, high dimensional (e.g., Asthma), or with scarce minority samples (e.g., Asia).

The three approaches considered represent a wide range of data sampling efforts in tackling the imbalanced problems. They can be categorized by their learning scopes. Random sampling duplicates the data without creating new information; SMOTE algorithm creates new synthetic data based on local information – the nearest neighbors; MDS approach generates data based on global information – the knowledge model built from the full training space. As illustrated in Figure 5, random sampling produces data from a single data point; SMOTE generates data over two data points; MDS generates data from a model built from all labeled data or domain knowledge.

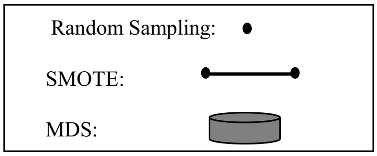

*Figure 5 - Learning scopes for 3 sampling approaches*

[1] TP is true positive rate for predicting minority samples.

[2] TN is true negative rate for predicting majority samples.

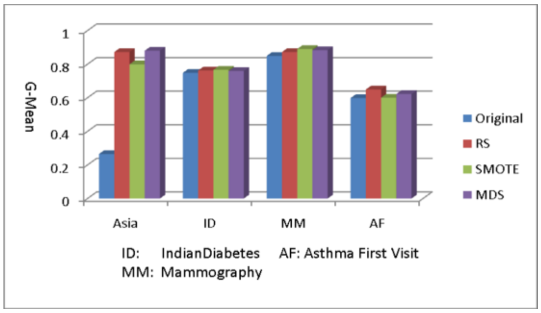

*Figure 6 - Overall performance (G-Mean) comparison*

Figure 6 shows that all three sampling approaches can improve classification performance on imbalanced data sets, especially on minority data. Comparing these three sampling approaches, random sampling is easy to implement and efficient; SMOTE will perform well especially when the minority data is tense; MDS will perform well when we have a reasonable accurate model to generate minority data, and this model could be from our medical domain knowledge or learning from existing data or both. Thus MDS can potentially address imbalanced problems with scarce or sparse minority data. In future work, we will incorporate domain knowledge into our model. This capability is a major difference from and a potential advantage over the other generative sampling approaches [9].

## Conclusions

This work examined and analyzed the challenges in the imbalanced data problems in medical decision support. We proposed a new approach – Model Driven Sampling that can potentially make use of all available data and domain knowledge to sample new data for balancing class distribution. We compared the performance of different major sampling approaches on four representative data sets. We showed that data sampling approaches can improve classification performance to a reasonable level most of the time. In particular, they can significantly improve predictions over the minority data, which is important in medical decision support. We showed that MDS is comparable in performance with respect to the other approaches considered, and can outperform them in certain cases. In future work, we will incorporate domain knowledge into MDS, and extend it to multi-class problems.

### Acknowledgment

We thank the reviewers for their valuable comments, and thank Prof Tow-Keang Lim for providing us the asthma data set. This work was partially supported by Academic Research Council grant R252-000-327-112 and grant R-252-000-309-112 from the Ministry of Education of Singapore.

## References

[1] Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural Networks. 2008 2008/4//;21(2-3):427-36.

[2] Cohen G, Hilario M, Sax H, Hugonnet S, Geissbuhler A. Learning from imbalanced data in surveillance of nosocomial infection. Artificial Intelligence in Medicine. 2006;37(1):7-18.

[3] Riddle P, Segal R, Etzioni O. Representation design and brute-force induction in a Boeing manufacturing design. . Applied Artificial Intelligence. 1994;8(125-147).

[4] Elkan C. The Foundations of Cost-Sensitive Learning. Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence; 2001; 2001. p. 973-8.

[5] Chawla NV, Bowyer KW, Hall LO, Kegel-meyer WP. SMOTE: Synthetic Minority Over-Sampling Technique Journal of Artificial Intelligence Research. 2002(16):321-57.

[6] Foster P, David J, Tim O. Efficient progressive sampling. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. San Diego, California, United States: ACM 1999.

[7] Willie N, Manoranjan D. An Evaluation of Progressive Sampling for Imbalanced Data Sets. Proceedings of the Sixth IEEE International Conference on Data Mining - Workshops: IEEE Computer Society 2006.

[8] Weiss GM. The effect of small disjuncts and class distribution on decision tree learning: Rutgers University; 2003.

[9] Liu A, Ghosh J, Martin C. Generative Oversampling for Mining Imbalanced Datasets. Proceedings of the International Conference on Data Mining; 2007 June 25-28; 2007. p. 66-72.

[10] Cooper GF, Herskovitz E. A Bayesian method for the induction of probabilistic networks from data. Machine Learning. 1992;9:309-47.

[11] Baysian Network in Java, http://bnj.sourceforge.net/.

[12] Blake C, Merz C. UCI Repository of Machine Learning Databases, 1998 "http://www.ics.uci.edu/~mlearn/~MLRepository.html".

[13] Kubat M, Holte RC, Matwin S. Learning when negative examples abound. Lecture Notes in Artificial Intelligence 1997: Springer; 1997. p. 146-53.

[14] Kubat M, Holte R, Matwin S. Addressing the Curse of Imbalanced Data Sets: One Sided Sampling. Fourteenth International Conference on Machine Learning; 1997; 1997. p. 179-86.

### Address for correspondence

Yin Hongli
Medical Computing Lab, School of Computing
Computing 1, 13 Computing Drive, National University of Singapore
Singapore 117417
Email: yinhl@comp.nus.edu.sg