# Supervised Learning – A Study on Mining with Unbalance Dataset

**RamBabu Pemula[1], Bhukya Shankar Nayak[2], K. Surekha[3] and Myla Satish Babu[4]**
[1]Department of CSE, Nimra Institute of Engineering and Technology, Ongole, India
[2,3,4]Department of CSE & IT, Vishu Sree Institute of Technology, India

***Abstract*— Remote-sensing classification problems, targets the number of mines present are very less compared to with several number of clutter objects. One of the traditional classification techniques usually avoids the unbalance causing performance to suffer accordingly. In contrast, the recently developed infinitely imbalanced logistic regression (IILR) algorithm explicitly addresses class imbalance in its formulation. This algorithm gives the details necessary to employ it for remote-sensing data sets characterized by class imbalance. The method is applied to the problem of mine classification on three real, measured data sets. Specifically, classification performance using the IILR algorithm is shown to exceed that of a standard logistic regression approach on two land-mine data sets collected with ground-penetrating radar, and on one underwater-mine data set collected with side-scan sonar.**

***Keywords*—** Data Mining, Classification, Imbalance Data and Logistic Regression

## I. INTRODUCTION

Classification problem in remote-sensing, the number of targets present is very small compared with the number of clutter objects performed. For example, in land-mine detection applications, it is common to have nearly one hundred false alarms due to clutter for every real mine present. Similarly, in underwater-mine classification applications, the number of naturally occurring clutter objects that are detected typically far outweighs the relatively rare event of detecting a mine. Imbalance is often implicitly ignored when it comes time to use a classification algorithm to employ. As a result, the traditional classification approaches that do not account for severe class imbalance often lead to poor classification performance. Classification technique named infinitely imbalanced logistic regression explicitly acknowledges the problem of class imbalance in its formulation. Although the method was developed with involving rare events such as fraud detection or drug discovery, we apply the [1] technique to the problem of mine classification in remote-sensing data.

Ram Babu is with the Department of Computer Science Engineering, JNT University.
Shankar Nayak is with the Department of Computer Science Engineering, Vishnu Sree Engineering College JNT University.
Surekha is with the Department of Computer Science Engineering, Vishnu Sree Engineering College JNT University.
Satish Babu is with the Department of Computer Science Engineering, Vishnu Sree Engineering College, JNT University.

Specifically, we demonstrate the benefits of the proposed approach on three real data sets, measured remote-sensing data.

Various approaches have been developed to handle the issue of class imbalance in general. However, despite their relevance for many applications in the remote-sensing community, such methods have not been widely adopted. One notable exception to this is to classify agricultural crops in multispectral imagery by employing neural [2], networks that were specially designed for class imbalance. Other example is which used an ad hoc method for detecting oil spills in synthetic aperture sonar [3], imagery. This work is the first to address the issue of class imbalance for mine classification.

Mining highly unbalanced datasets, particularly in a cost sensitive environment, is among the leading challenges for knowledge discovery and data mining [1], [2].This problem arises when the class of interest is relatively rare as compared to the other. Without loss of generality assume that the positive class is the minority class, and the negative class is the majority class. Various applications demonstrate this characteristic of high class imbalance, such as bioinformatics, e-business, and information security, to national security.

## II. SECTION

***1) Related Work on Imbalance Data:*** Machine learning techniques involve imbalance data sets have developed different methods for solving the class unbalance problem such as adjusting cost of misclassification resizing training data sets

Resizing training data sets are over-sampling minority class examples and under-sampling the majority class. Over sampled the minority class by adding copies of the minority to training set. Over-sampling does not increase information but it does increase the cost of misclassification. Under-sampling removes randomly selected or near miss examples or examples that are far from minority class. Many examples for reducing majority class but down-sizing the majority class results in a loss of information.

Cost-sensitive classifier handles the problems with different misclassification error costs. Cost-sensitive classifiers may be imbalance data sets by setting a high cost to the misclassification of a minority class. Similar effect to over-sampling the minority class and many wide up with specific rules over fitting training.

***2) Logistic Regression:*** Logistic regression is used to model the classification of a categorical outcome with independent variables for data.

In national health survey by national center investigates the relationship between the health condition of demographic factors like race, gender and income levels of householders and general population

In logistic regression probabilities of outcome categories are assumed to be a function of linear combination of the explanatory variables is also called link function used to link functions are the cumulative logit function. It is classified logit function, probit function and complementary log-log function.

***3) Classifier Learning:*** Classifier learning is a fundamental technique in machine learning technique supposes the database is horizontally partitioned between the parties and participants learn the resulting decision tree, initially the classifiers were either true or false but recent research has expanded this representation to include real-valued, neural network and functional S-expression conditions. A classifier learning technique is an adaptive that learns to perform the best action given its input. By "best" is generally meant the action that will receive the most reinforcement from the system's environment. By "input" is meant the environment as sensed by the system, usually a vector of numerical values. The set of available actions depends on the decision context, for instance a financial one; the actions might be "buy", "sell", etc. In general, classifier learning is a simple model of an intelligent agent interacting with an environment. Ability to choose the best action improves with experience. The source of the improvement is reinforcement technically, payoff provided by the environment. In many cases, the payoff is arranged by the experimenter or trainer of the classifier. For instance, in a classification context, the payoff may be 1for "correct" and 0 for "incorrect". In a robotic context, the payoff could be a number representing the change in distance to a recharging source, with more desirable changes represented by larger positive numbers, etc. Often, systems can be set up so that effective reinforcement is provided automatically, for instance via a distance sensor.

## III. PROBLEM FORMULATION

***1) Unbalanced classifier Learning:*** Training data set consists of thousands of examples involves in real valued features are labeled as 1 are positive and -1 are negative. Missing of values in data sets

***2) Resampling Approach:*** synthetic minority over sampling is an over sampling designed to generate synthetic examples in a specific manner by operating in feature space rather than data space. The minority class is over sampled by taking each minority class is over-sampled by taking each minority class sample and k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. Synthetic samples are generated in the following way: take the difference between the feature vector sample under consideration and its nearest neighbor multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration.

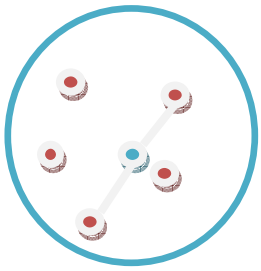

Fig. 1. Resampling Approach

The minority class is oversampled by taking each minority class sample synthetic examples blue circle along the line segments joining all of the k by default=5 minority class nearest neighbors in red circles.

***3) Land Mine Classification*:** In the test of United State Army desert contains two target land-mine datasets. Data was collected by a Mirage ground radar sensor located on an airborne platform. Sensor is a 300MHz to 3MHz system with polarimetric capability and resulting image resolution was 11.7cm in ground range and 4.77cm in azimuth. Two GPR

Each of the two images undergoes the same pre-processing. The detection stage proceeded as follows. The raw magnitude image was smoothed and dilated. Gradient blurred image was computed in the vertical and horizontal directions to form a gradient image. A filter template that mimics the response of a mine was then correlated with the gradient image to produce a filtered image. Energy locations in this filtered image were recorded as alarms. Observed that this detection procedure effectively rejects many false alarms due to vegetation that a simple energy detector applied to the raw magnitude image would flag as alarms

Feature extraction for each alarm from the detection stage a smaller image chip containing the object was then extracted from the original large raw image. Seven features were subsequently extracted for each alarm are energy score from the detection stage four moment based features of the pixel values of the magnitude chip such as mean variance and a measure of the spatial variance in the magnitude chip and the entropy of the pixel values.

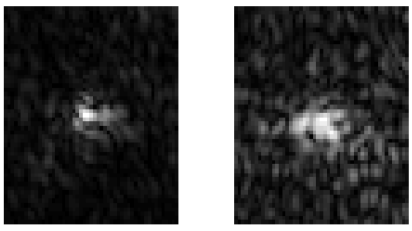

Fig. 2. Examples of Land Mine chips

Table 1: Mining data

| Data Sets | Number of Features | Number of Data Points | |
|---|---|---|---|
| LAND MINE A | 7 | 22 | 798 |
| LAND MINE B | 7 | 11 | 867 |
| UNDERWATER MINE | 22 | 16 | 2650 |

***4) Water Mine Classification:*** Under water mine data set from a site in the Gulf of Mexico data was collected as part of the technical cooperation program trail mongoose by Klein. The sensor has a center frequency of 455 KHz and bandwidth of 20KHz resulting image resolution was 3.29cm in range and approximately 10cm in azimuth.

The detection procedure followed the approach described four simple detectors are applied to the sonar imagery. Two detectors employ matched filters that search for an ideal target signature while the other two calculate certain statistical quantities indicative of targets and feature extraction for each alarm from the detection stage a smaller image chip containing the object was then extracted from the original swath image. Twenty two features related to geometrical or statistical quantities of the echo and shadow was extracted from each segment chip. Some example features include the angle and the ratio of the major and minor axes of an ellipse fit to the segmented pixels, the statistics of the height profile of the object, the length and width of the object, and a measure of the convexity of the segmented areas.

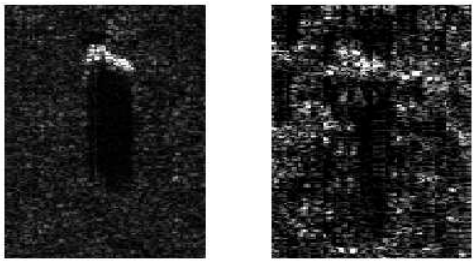

Fig. 3. Chips from water mine datasets

## IV. SIMULATION STUDIES

A Classification problem occurs standard classifiers tend to be overwhelmed by the large classes and ignore the small ones. In practical applications, the ratio of the small to the large classes can be drastic such as 1 to 100, 1to 1,000, or 1 to 10,000 (and sometimes even more). This problem prevalent in many applications, including: fraud/intrusion detection, risk management, text classification, and medical diagnosis/monitoring, but there are many others. It is worth noting that in certain domains the class imbalance is intrinsic to the problem. For example, within a given setting, there are typically very few cases of fraud as compared to the large number of honest use of preferred facilities. However, class imbalances sometimes occur in domains that do not have an intrinsic imbalance. This will happen when the data collection process is limited due to economic or privacy reasons. In addition, there can also be an imbalance in costs of making different errors, which could vary per case.

***1) Implementation Details:*** Classification of three datasets relevant are as under:

Bbaseline approach is logistic regression algorithm that does not account for class imbalance. Two different versions which differ only in the number of Gaussians that are used to model the distribution of the clutter class of data

The first version models the distribution of the clutter class of data using a single Gaussian (K = 1), while the second version uses a Gaussian mixture model with a maximum of K = 10 Gaussians. For three data sets, the same stratified five cross-validation training and testing procedure is employed. Specifically, the cross-validation was stratified in the sense that the proportion of samples from each class is approximately equal in each of the five folds. The imbalanced data sets considered, this means that each fold contains only 2-5 targets. The data points in one of the folds are treated as unlabeled testing data while the data points in the other folds are treated as labeled training data. The classifier is learned using the training data, and is then exploited to classify the data points in the testing fold. This process is then repeated so that each fold is treated as unlabeled testing data one time.

For each classification method, the output for a given unlabeled data point is related to the probability of belonging to the mine class. All data points for which this quantity is greater than or equal to some threshold  are classified as targets mines, while the other data points are classified as clutter. Varying the threshold Γ will effectively generate a receiver operating characteristic (ROC) curve. The area under the ROC curve (AUC), which provides a scalar summary measure of performance, can also be easily calculated. The AUC is given by the Wilcoxon statistic:

$$\mathrm{AUC} = \frac{1}{N_1 N_0} \sum_{i=1}^{N_1} \sum_{j=1}^{N_0} \mathbb{I}_{\psi_{1i} > \psi_{0j}}$$

where 1i are the classification scores of data belonging to the target class, 0j are the classification scores of data belonging to the clutter class, and I is an indicator function.

***2) Classification Results of Mine:*** classification methods considered in ROC curves and the performance in terms of AUC scores for the three data sets is shown in Fig. 4.

Performance in terms of AUC scores for the three data sets is compactly summarized in the Table 2. As can be seen Table 1, for all three data sets considered, the infinitely imbalanced logistic regression approach that accounts for class imbalance outperforms the standard logistic-regression approach. Additional results in terms of AUC using a leave-one-out procedure rather than the stratified five-fold cross-validation, not shown here due to space constraints, also support the claim that the IILR approach is superior to the logistic-regression approach. These results were made possible

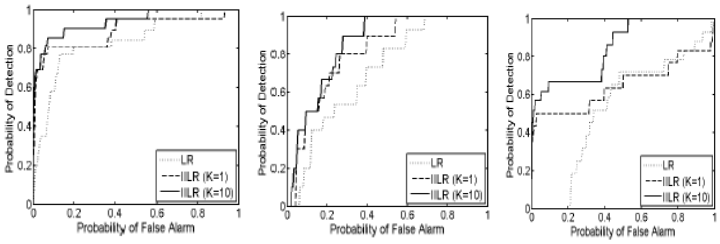


Fig. 4: Average ROC curves of the classification methods when using stratified five-fold cross-validation for the (a) Land Mine A, (b) Land Mine B, and (c) Underwater Mine data sets

Table 2: Classification methods using five-fold cross validation

| DATA SET | LR | IILR ($K = 1$) | IILR ($K = 10$) |
|---|---|---|---|
| LAND MINE A | $0.8367 \pm 0.0787$ | $0.8906 \pm 0.0791$ | $0.9417 \pm 0.0513$ |
| LAND MINE B | $0.7087 \pm 0.1287$ | $0.8137 \pm 0.1716$ | $0.8529 \pm 0.1245$ |
| UNDERWATER MINE | $0.5286 \pm 0.1634$ | $0.6497 \pm 0.2019$ | $0.8480 \pm 0.1393$ |

by the severe class imbalance of the data sets, in which the numbers of clutter data points far outweighed the numbers of target data points. However, a less obvious reason for the success of the IILR approach is the fact that the features of the target data points were well-represented by their mean values, ±x. In general, the IILR approach will not be suitable for data sets in which the data points of the rare class do not cluster tightly.

## V. CONCLUSION

In our recently-developed classification approach for data sets characterized by class imbalance has been described. The implementation details needed to employ the approach have also been provided. The utility of the infinitely imbalanced logistic regression algorithm for mine classification was demonstrated on three real, measured remote-sensing data sets. Several classification algorithms in use today are relatively similar to logistic regression. Therefore, after using this method to the community, we expect that other researchers will find the approach useful for other applications characterized by severe class imbalance. This methodology could be refined with previous information which cumulated by the retailers and would be implemented in future works.

## REFERENCES

[1] J. Bongard and H. Lipson. Automating genetic network inference with minimal physical experimentation using coevolution. In Proceedings of the 2004 Genetic and Evolutionary Computation Conference, pages 333{345. Springer, 2004.

[2] A. Owen, “Infinitely imbalanced logistic regression,” Journal of Machine Learning Research, vol. 8, pp. 761–773, 2007.

[3] L. Bruzzone and S. Serpico, “Classification of imbalanced remotesensing data by neural networks,” Pattern Recognition Letters, vol. 18, pp. 1323–1328, 1997.

[4] M. Kubat, R. Holte, and S. Matwin, “Machine learning for the detection of oil spills in satellite radar images,” Machine Learning, vol. 30, pp. 195–215, 1998.

[5] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” J. Royal Statistical Society B, vol. 39, pp. 1–38, 1977.

[6] N. Nasios and A. Bors, “Variational expectation-maximization training for Gaussian networks,” in Proceedings of the IEEE Workshop on Neural Networks for Signal Processing, 2003, pp. 339–348.

[7] M. Beal and Z. Ghahramani, “The variational Bayesian EM algorithm for incomplete data: Application to scoring graphical model structures,” Bayesian Statistics, vol. 7, pp. 453–464, 2003.

[8] V. Myers and Ø. Midtgaard, “Fusion of contacts in synthetic aperture sonar imagery using performance estimates,” in IOA International Conference on Detection and Classification of Underwater Targets, 2007, pp. 77–88.

[9] V. Myers, “Sonar image segmentation using iteration and fuzzy logic,” in CAD/CAC 2001 Conference Proceedings, 2001.

**RamBabu Pemula:** B.Tech (CSE), M.Tech Software Engineering from JNTU M.B.A from ANU. Currently working as a Asst Prof CSE dept at Nimra Institute of Engineering & Technology, Ongole. Having 5½ years of Experience in Academic interested areas include Data mining and Networks.

**Bhukya Shankar Nayak:** B.Tech CSE from JNTU Hyderabad, M.Tech CSE from IIT Madras, currently working as a Head of Department (CSE&IT) at Vishu Sree Institute of Technology. He is having 7½ years of experience in Industrial & Teaching guided many UG students and also attended many workshops & Conference at NIT's and IIT's

**K. Surekha:** Master of Computer Applications from Osmania University Hyderabad. She is having 3-years of experience in Academic guided many UG students, currently working as a Asst Prof at Vishu Sree Institute of Technology interest areas include Networks and Data mining

**Myla Satish Babu:** B.Tech CSE from JNTUH M.Tech CSE from JNTUH, currently working as a Asst Prof at Vishu Sree Institute of Technology in dept of CSE & IT. He is having 5 years of experience in Industry and Academic guided many UG students areas of interest include Computer Networks and Data Mining