Rare Event Modeling: A data mining approach

Authors:
Bibhas Dey | Anindya Sengupta | Saurabh Bagchi

Table of contents

Abstract

Rare events are generally classified as events whose percentage and count of occurrence in the sample of study is relatively low. Some of the examples of rare events are disease incidents, natural disasters, banking frauds and machine component failures. Common statistical techniques do not work and/or give poor results on such type of analysis.

In this paper our modeling objective is to predict the likelihood of a rare event in the next t months (here is fixed). Traditionally, methods like logistic regression have been used in this analysis. This methodology suffers from the fact that it uses maximum likelihood method to estimate the risk factors due to which it is susceptible to small sample bias. Since such events occur rarely (1%-2% chance), we should have a sufficiently large sample to cover all variants of this event. This paper covers various approaches to tackle this issue.

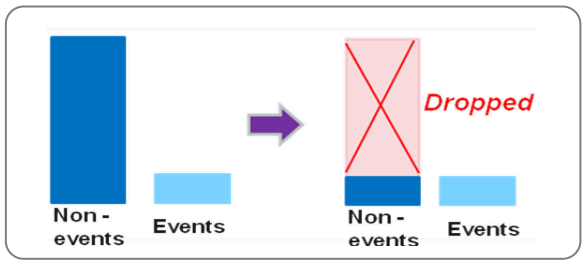
1. Data manipulation approach: Under-sampling the non-events

Under-sampling is an efficient method for classing-imbalance learning. This method uses a subset of the majority class to train the classifier. Since many majority class examples are ignored, the training set becomes more balanced and the training process becomes faster.

Steps:

a) Sample the data records from non-events class (Randomly, Near miss examples, Examples far from event class examples)

b) Introduce sampled data records into the original data set instead of original data records from the majority class



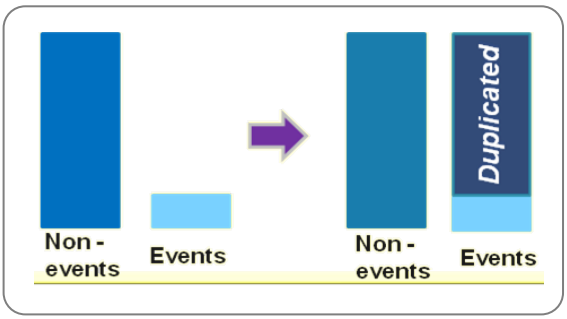
The main drawback of under-sampling is that potentially useful information contained in these ignored examples is neglected which results in a general loss of information and overly general rule.

2. Data manipulation approach: Over-sampling the events

The simplest method to increase the size of the minority class corresponds to random over-sampling, that is, a non-heuristic method that balances the class distribution through the random replication of positive examples.

Steps:

a) Make the duplicates of the rare events until the data set contains as many examples as the non-event class => balance the classes



Since this method replicates existing examples in the minority class, over-fitting is more likely to occur. It does not increase information but increases the misclassification cost.

3. Statistic based approach: Probability Distribution to detect Outliers

Basic Assumption: The count of non-events in the data is significantly larger than the number of events.

Distribution for the data D is given by

$$D = (1-\lambda).M + \lambda.A$$

Where, M – non-event distribution,
A - event distribution

M_t, A_t are sets of non-event, event elements respectively.

Steps:

a) Compute likelihood $L_t(D)$ of distribution D at time t

b) Measure how likely each element x_t is outlier:

i. $M_t = M_{t-1} \setminus \{x_t\}$, $A_t = A_{t-1} \cup \{x_t\}$

ii. Measure the difference $(L_t - L_{t-1})$

.

4. Distance based approach: Nearest Neighbour (NN) to detect Outliers

One of the most popular outlier detection techniques is distance-based outlier, introduced by Knorr and Ng (1998, 1999). The algorithm evaluates the local data distribution around a test data point and compares that distribution with the data distribution within the sample defined by its K nearest neighbours. The algorithm's success is based on the assumption that the distribution of distances between a true outlier and its nearest neighbours will be different from the distribution of distances among those neighbours by them. This assumption relies on the definition of an outlier as a point whose behaviour (i.e. the point's location in parameter space) deviates in an unexpected way from the rest of the data distribution.

Steps:

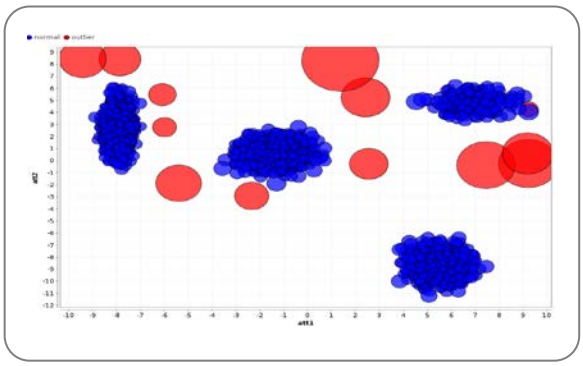
a) For each data point *d* compute the distance to the *k-th* nearest neighbour d_k

b) Sort all data points according to the distance d_k

a) Outliers are points that have the largest distance d_k and therefore are located in the more sparse neighbourhoods

b) Usually data points that have top *n*% distance d_k are identified as outliers

a) *n* – user parameter



However, as it computes all the dimensional distances of the points from one another, it is time-consuming if the available objects are of very great size. Besides, the direct application to high dimensional problems often results in unexpected performance and qualitative costs due to the curse of dimensionality.

5. Density based approach: Local Outlier Factor (LOF) to detect Outliers

We introduce a local outlier (LOF) for each object in the dataset, indicating its degree of outlier-ness. This is the first concept of an outlier which also quantifies how outlying an object is. The outlier factor is local in the sense that only a restricted neighbourhood of each object is taken into account. Our approach is loosely related to density-based clustering.

Steps:

a) For each data point q compute the distance to the k-th nearest neighbour (k-distance)

b) Compute reachable distance (reach-dist) for each data example q with respect to data example p as:

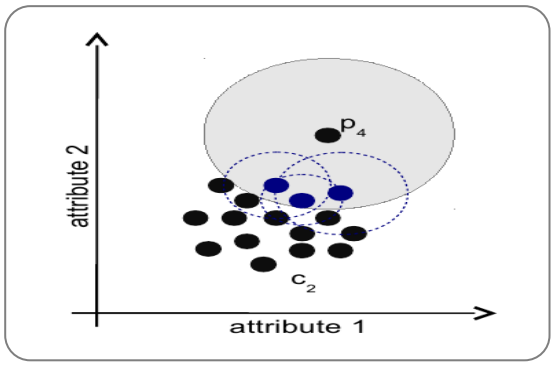
reach-dist (q, p) = max{k-distance(p), d(q,p)}

c) Compute *local reachable density* (*lrd*) of data example *q* as inverse of the average reachable distance based on the *MinPts* nearest neighbors of data example *q*

$$\mathbf{lrd(q)} = \frac{MinPts}{\sum_{p} reach\_dist_{MinPts}(q,p)}$$

d) Compute *LOF*(*q*) as ratio of average local reachable density of *q*'s *k*-nearest neighbors and local reachable density of the data record *q*

$$\mathbf{LOF(q)} = \frac{1}{MinPts} \cdot \sum_{p} \frac{lrd(p)}{lrd(q)}$$



6. Clustering based approach: K-nearest neighbour + canopy clustering to detect Outliers

The algorithm based on outlier removal and density calculation, also based on a simple and efficient data structure to find the k nearest neighbours, this data structure is called canopy which is simply a number of overlapped hyper spheres of the same radius cover the data space.

This data structure is used to partition the data space like the grid but here there is overlap between cells and the cells are not (hyper) rectangular regions but (hyper) spherical regions.

Steps:

a) Compute the sum of distances to the k nearest neighbours (k-NN) of each point

   i. Points in dense regions – small k-NN score

   ii. k has to exceed the frequency of any given attack type

   iii. Time complexity $O(n^2)$

b) Speed up with canopy clustering that is used to split the entire space into small subsets (canopies) and then to check only the nearest points within the canopies

c) Apply fixed width clustering and compute distances within clusters and to the centre of other clusters

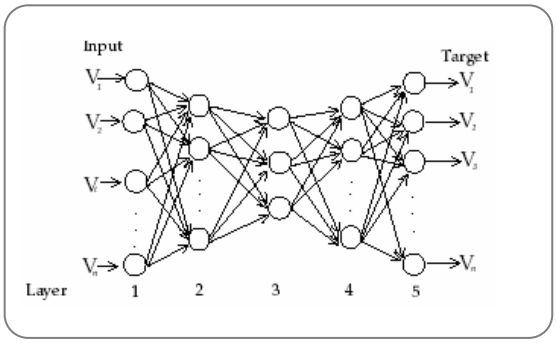
7. Model based approach: Neural Networks to detect Outliers

a) Use a replicator 4-layer feed-forward neural network (RNN) with the same number of input and output nodes

b) Input variables are the output variables so that RNN forms a compressed model of the data during training

c) A measure of level of outlying is the reconstruction error of individual data points.

8. Model based approach: Unsupervised Support Vector Machine (SVM) to detect



Outliers

Normal data records belong to high density data regions whereas anomalies belong to low density data regions

Steps:

a) Use unsupervised approach to learn high density and low density data regions

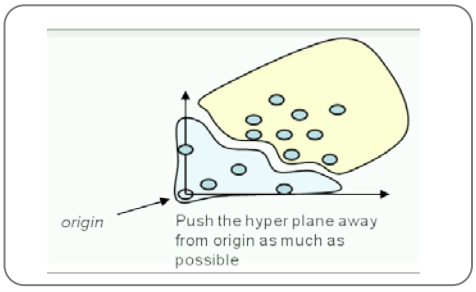
b) Use SVM to classify data density level

c) Separate the entire set of training data from the origin, i.e. to find a small region where most of the data lies and label data points in this region as one class. The parameters are:

i. Expected number of outliers

ii. Variance of rbf kernel (As the variance of the rbf kernel gets smaller, the number of support vectors is larger and the separating surface gets more complex)

d) Separate regions containing data from the regions containing no data



9. Logistic regression of rare events

Separation is a phenomenon associated with models for dichotomous or categorical outcomes, including logistic and prohibit regression.

Separation occurs if the predictor (or a linear combination of some subset of the predictors) is associated with only one outcome value when the predictor is greater than some constant.

If the outcome values are perfectly determined by the predictor then the condition "complete separation" is said to obtain. If instead there is some overlap then "quasi-complete separation" obtains. A 2 × 2 table with an empty cell is an example of quasi-complete separation.

A parameter in the model will tend to be infinite, if complete separation is observed. If quasi-complete separation is the case, the likelihood is maximized at a very large but not infinite value for that parameter.

Methods to fit these models include Firth logistic regression and Exact logistic regression, a bias-reduction method based on a penalized likelihood.

Firth:

i. The basic idea of the firth logistic regression is to introduce a more effective score function by adding a term that counteracts the first-order term from the asymptotic expansion of the bias of the maximum likelihood estimation—and the term will goes to zero as the sample size increases.

ii. In logistic regression, Firth's approach is equivalent to penalizing the likelihood by the Jeffrey's invariant prior. The attraction of this method is that it provides bias-reduction for small sample size as well as yields finite and consistent estimates even in case of separation.

iii. Firth's correction adds a fixed quantity to the likelihood, and in large samples the relative contribution of that quantity disappears at the rate of ***1/n*** dwarfed by the sample information.

Exact:

i. In case of small sample and/or very unbalanced binary data (20 cases out of 1000 samples) – 'exact logistic' regression is to be used

ii. Based on appropriate exact distribution of sufficient statistics for parameters of interest and estimates given by the technique do not depend on asymptotic results

iii. Unlike most estimators, rather than calculating coefficients for all independent variables at once, results for each independent variable are calculated separately with the other independent variables temporarily conditioned out

iv. The goal of the exact conditional analysis is to determine how likely the observed response y_0 is with respect to all 2^n possible responses $y = (y_1,\ldots,y_n)'$. One way to proceed is to generate every y vector for which $y'X_0 = t_0$, and count the number of vectors y for which y_0X_1 is equal to each unique t_1.

10. Conclusion

Data mining analysis of rare events requires special attention. Many real world applications exhibit "needle-in the-haystack" type of problem. There are many more methods that are used apart from the ones covered in this paper. But the current "state of the art" data mining techniques are still insufficient for efficient handling rare events. One cannot pin-point a particular technique to be applicable for all scenarios. To achieve a desired level of accuracy, the method used varies as per the problem statement at hand. It also differs based on the domain and nature of the business to which it needs to be applied. Given these facts, there is a growing need for designing better and more accurate data mining models to tackle rare event situation.

11. References

1. An introduction to the Analysis of rare events – Nate Derby

2. Data mining for Analysis of Rare Events: A case study in security, financial and medical applications – Aleksander Lazarevic, Jaideep Srivastava, Vipin Kumar

3. Anomaly detection using Data mining techniques – Margaret H. Dunham, Yu Meng, Donya Quick, Jie Huang and Charlie Isaksson

4. Method of multivariate analysis for imbalance data problem – Nikolai Gagunashvili

5. A Geometric Framework for unsupervised anamoly detection: Detecting Intrusions in Unlabeled data – Eleazer Eskin, Andrew Arnold, Michael Prerau, Leonid Portnov, Sal Stolfo

6. Nearest neighbor and clustering based anamoly detection algorithms for RapidMiner – Mennatallah Amer , Markus Goldstein

7. An application of cluster analysis in the financial services industry – Satish Nargundkar, Timothy J. Olzer

8. Exact Logistic Regression – Larry Cook

9. A solution to the problem of separation in logistic regression – Georg Heinze and Michael Schemper

10. An overview of classification algorithms for imbalanced datasets – Vaishali Ganganwar

11. LOF: Identifying Density-Based Local Outliers – Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, Jorg Sander

12. Appendix

Comparison table of the rare event modeling method

Methodology	Pros	Cons
Under sampling the non – events	• Entire response population constant • Higher accuracy	• Model results not stable • General loss of information and overly general rules
Over sampling the events	• Entire response population constant • Higher accuracy	• Over fitting likely to occur • Increases misclassification cost
Probability distribution to detect outliers	• Event distribution taken into account • MLE approach provides consistency in parameter estimation	• Mathematics involved is non-trivial • Heavily biased for small samples, sensitive to starting choice
Nearest neighbors to detect outliers	• Easy and simple to implement • Number of clusters can be variable	• Not suitable for datasets that have modes with varying density • Time consuming for high dimensions
LOF to detect outliers	• Localization of outliers • Density based clustering is used	• Difficult to visualize with greater attributes • Neighborhood is highly subjective
K nearest neighbor + Canopy Clustering	• Divide and conquer approach speeds up computation • Underlying data structure is simple and efficient	• Time complexity $O(n^2)$ • K has to exceed the frequency of any given attack type
Neural networks to detect outliers	• Mimic human understanding • Compressed data model formed	• Number of hidden nodes variable • Reconstruction error prone
Unsupervised SVM to detect outliers	• No labeling required in training data • Data density taken into consideration	• As variance of rbf kernel gets small separating surface becomes more complex
Logistic regression of rare events	• Easy to interpret and understand • Minimum of 50-100 observation suffices	• Memory intensive with more than one dependent variable • Separation may not be totally unambiguous

About Fractal Analytics

Fractal Analytics is a global analytics firm that serves Fortune 500 companies gains a competitive advantage by providing them a deep understanding of consumers and tools to improve business efficiency. Producing accelerated analytics that generate data driven decisions, Fractal Analytics delivers insight, innovation and impact through predictive analytics and visual story-telling.

Fractal Analytics was founded in 2000 and has 800 people in 13 offices around the world serving clients in over 100 countries.

The company has earned recognition by industry analysts and has been named one of the top five “Cool Vendors in Analytics” by research advisor Gartner. Fractal Analytics has also been recognized for its rapid growth, being ranked on the exclusive Inc. 5000 list for the past three years and also being named among the USPAACC’s Fast 50 Asian-American owned businesses for the past two years.

For more information, contact us at:

+1 650 378 1284

info@fractalanalytics.com

Follow us: