synthetic minority oversampling using multivariate gaussian distribution

namagiri December 2023

1 Abstract

Highly imbalanced data is a common phenomenon from cases of fraudulent transactions to several medical datasets. Without rebalancing the dataset if it fits into any learning algorithm will result in a highly biased model towards the majority class to address this issue we have employed a technique of oversampling using a new smote(synthetic minority oversampling) strategy where k nearest neighbours are picked from the minority class and fitted to multivariate gaussian distribution to generate points which populate regions with high probability

2 Methodology

Partially guided synthetic minority oversampling using Gaussian kernel interpolation.

Let's consider the matrix X:

$$X = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{a1} & a_{a2} & \cdots & a_{na} \end{bmatrix}$$

The dimensions of matrix X are $n \times a$, where n refers to rows/instances and a refers to attributes.

Let t be the target attribute and $X[t] = \{0, 1\}$, representing a binary classification task.

$X = \mathrm{Maj} \cup \mathrm{Min}$, where Maj is the majority class and Min is the minority class.

$\mathrm{Maj}[t] = \{0\}$ and $\mathrm{Min}[t] = \{1\}$.

Let $\theta = \frac{n(\mathrm{Min})}{n(\mathrm{Maj})}$, where $\theta \in [0, 1]$.

$n()$ denotes the cardinality of the class, and $\theta \ll 1$.

Algorithm 1 Selecting ensemble classifier

Input: Minority data set Min ,Majority data set Maj, ratio balancer δ
Output: Ensemble classifier E
bounding the number of subsets $b = \lceil \frac{n(Maj)}{\delta * n(Min)} \rceil$
Selecting b subsets from $Maj \mid \bigcup_{i=1}^{b} Maj(i) = Maj \ and \ \bigcap_{i=1}^{b} Maj(i) = \emptyset, \forall i, \quad n(\text{Maj}(i)) > \delta * n(\text{Min})$
Number of synthetic samples to be generated
sample number $= \min_{i \in [1,b]} n(\text{Maj}(i)) - n(\text{Min})$
$Gen = Algorithm_2(Min, Samplenumber)$
$Min = Min \cup \text{GEN}$
$C_i(Maj(i), Min)$
Getting b classifiers and trained to evaluate the best classifier BEST $= decide(\{C_1, C_2, C_3, \ldots, C_b\})$
Return: BEST

Algorithm 2 Generating synthetic samples

Input: Minority class Min, Sample number
Output: Synthesized samples GEN
Start $s = sample\,size$ and $GEN \leftarrow \phi$
for $i \in$ Min **do**
  l = Sort(k nearest neighbours of i)
  $l\_first \leftarrow$ First $\frac{l}{4}$ elements of data
  $l\_last \leftarrow$ Last $\frac{l}{4}$ elements of data
  $l_1 = l\_last \cup l\_first$
  $f \leftarrow$ Multivariate gaussian distribution(l_1) ▷ Interpolated surface
  Evaluation:
  $l\_middle \leftarrow$ Middle $\frac{l}{2}$ elements of data
  for x in l_middle **do**
    $y \leftarrow$ Evaluate(f, x) ▷ Evaluate the function at the middle elements
  end for
  $s \leftarrow s - n(y)$
  $GEN \leftarrow GEN \cup y$
  $Min \leftarrow Min - l$
end for
Return: GEN

fitting the l1 vector into the distribution

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ be a vector-valued random variable with a multivariate normal distribution. It has mean $\boldsymbol{\mu} \in R^n$ and covariance matrix $\boldsymbol{\Sigma} \in S_n^{++}$.

The probability density function (pdf) of $\mathbf{X}$ is given by:

$$p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

To generate samples from this distribution, we can use the Cholesky decomposition. The Cholesky decomposition expresses the positive definite matrix $\boldsymbol{\Sigma}$ as the product of a lower triangular matrix $\mathbf{L}$ and its transpose:

$$\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$$

The algorithm for generating samples is as follows:

1. Ensure that $\boldsymbol{\Sigma}$ is positive definite.
2. Compute the Cholesky decomposition: $\mathbf{L} = \text{cholesky}(\boldsymbol{\Sigma})$.
3. Generate standard normal samples $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
4. Transform to multivariate normal samples: $\mathbf{X} = \boldsymbol{\mu} + \mathbf{z}\mathbf{L}^T$.