Machine Learning
https://doi.org/10.1007/s10994-022-06296-4

A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning

Dina Elreedy[1] · Amir F. Atiya[1] · Firuz Kamalov[2]

Received: 11 July 2022 / Revised: 13 November 2022 / Accepted: 15 December 2022

Abstract

Class imbalance occurs when the class distribution is not equal. Namely, one class is under-represented (minority class), and the other class has significantly more samples in the data (majority class). The class imbalance problem is prevalent in many real world applications. Generally, the under-represented minority class is the class of interest. The synthetic minority over-sampling technique (SMOTE) method is considered the most prominent method for handling unbalanced data. The SMOTE method generates new synthetic data patterns by performing linear interpolation between minority class samples and their K nearest neighbors. However, the SMOTE generated patterns do not necessarily conform to the original minority class distribution. This paper develops a novel theoretical analysis of the SMOTE method by deriving the probability distribution of the SMOTE generated samples. To the best of our knowledge, this is the first work deriving a mathematical formulation for the SMOTE patterns' probability distribution. This allows us to compare the density of the generated samples with the true underlying class-conditional density, in order to assess how representative the generated samples are. The derived formula is verified by computing it on a number of densities versus densities computed and estimated empirically.

Keywords SMOTE · Class imbalance · Distribution density · Over-sampling · Minority class

Editors: Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michal Wozniak, Shuo Wang.

✉ Dina Elreedy
dinaelreedy@eng.cu.edu.eg

Amir F. Atiya
amir@alumni.caltech.edu

Firuz Kamalov
firuz@cud.ac.ae

[1] Computer Engineering Department, Cairo University, Giza 12613, Egypt

[2] Department of Electrical Engineering, Canadian University Dubai, Dubai 117781, UAE

Published online: 05 January 2023

1 Introduction

The class imbalance problem arises in various real world applications such as: medical diagnosis (Fotouhi et al., 2019), credit card fraud detection (Li et al., 2021), software testing (Balogun et al., 2020), e-commerce (Wu & Meng, 2016), and stock selection (Atiya et al., 1997). The unbalanced data problem occurs when one class is under-represented (the minority class), while the other class is over-represented in the data (the majority class). The class imbalance could be due the data collection process. For example, in medical diagnosis, normal cases could be larger than patients suffering from a certain uncommon disease (Liu et al., 2022; Fotouhi et al., 2019) although the target is to identify the minority class denoting the patients.

Standard machine learning classifiers such as Support vector machines (SVM) (Hearst et al., 1998), decision trees (Quinlan, 1996), and K-nearest neighbor (KNN) (Guo et al., 2003) generally assume at least implicitly an even class distribution. Thus, applying the standard approaches without handling the class imbalance could dramatically impact the classification performance since classifiers would be biased towards the over-represented (majority) class.

There are three major approaches for handling the class imbalance problem: the data level approach (Batista et al., 2004; Chawla et al., 2002; Guzmán-Ponce et al., 2021), the cost sensitive approach (Devi et al., 2022), and the algorithm level approach (Mullick et al., 2018; Buda et al., 2018; Ganaie et al., 2021).

The data level approach is the most prevalent paradigm in handling unbalanced data. Data level algorithms are sampling methods that apply data pre-processing before classification, typically by increasing the number of minority class samples which is known as over-sampling (Chawla et al., 2002; Koziarski et al., 2021). Conversely, some majority class samples could be excluded from the data, which is known as under-sampling (Chennuru & Timmappareddy, 2022; Vuttipittayamongkol & Elyan, 2020). A key advantage of the data level approach is its generality since it can be applied to any classifier.

Over-sampling can be performed using two main approaches. The first approach is replicating the original minority class samples such as: random over-sampling (Abd Elrahman & Abraham, 2013). However, this approach may result in over-fitting by over-emphasizing noisy minority samples. The second approach for increasing the number of minority class samples is to generate new synthetic minority class samples (Abd Elrahman & Abraham, 2013; Chawla et al., 2002; Wan et al., 2017; Goodman et al., 2022).

One of the most popular over-sampling methods is “Synthetic Minority Over-sampling Technique (SMOTE)" developed by Chawla et al. (2002). The SMOTE method generates synthetic data by applying linear interpolation between a minority class point and one of its K nearest neighbors. SMOTE is a powerful over-sampling method that has been widely adopted in many applications (Fernández et al., 2018; Ahsan et al., 2018; Kishor & Chakraborty, 2021). Furthermore, a plethora of SMOTE extensions have been developed such as: Borderline SMOTE (Han et al., 2005), Safe-level SMOTE (Bunkhumpornpat et al., 2009), ADASYN (He et al., 2008), SVM SMOTE (Nguyen et al., 2011), Localized Random Affine Shadowsampling (LoRAS) (Bej et al., 2021), CDSMOTE (Elyan et al., 2021), and Deep SMOTE (Dablain et al., 2022).

Another technique for synthetically generating minority class samples is to estimate the underlying minority class probability distribution, and generate samples from it such as: PDF oversampling (PDFOS) (Gao et al., 2014) and random walk oversampling (Zhang & Li, 2014). However, density estimation in case of scarce data samples would be inaccurate especially for high dimensional data. On the other hand, for high dimensional data such as images, Wan et al. (2017) develop a variational autoencoder for generating similar synthetic samples to the original ones.

In this work, we mainly investigate the SMOTE method due to its popularity and competitive performance. Despite the efficacy of the SMOTE over-sampling algorithm (Chawla et al., 2002), it has some limitations. For example, SMOTE oversamples noisy examples which could magnify the noise impact and degrade the classification performance. In addition, SMOTE could falsely generate synthetic samples in the majority class region misleading the classifier. Furthermore, SMOTE does not consider minority classes composed of several small disjuncts or sub-concepts (Prati et al., 2004). One of the main reasons for all of the aforementioned SMOTE pitfalls is the fact that the SMOTE patterns are not genuine as they are not generated from the original minority class distribution. It is important to establish that they are true representatives of the underlying class, by showing that they obey a similar distribution.

Another concern regarding SMOTE is that it is not sufficiently grounded on a solid mathematical theory. As a step towards this goal, this work aims to establish a mathematical foundation for analyzing the SMOTE algorithm. Specifically, this work derives a mathematical formulation for the probability distribution of the SMOTE synthetically generated samples. The benefit of this analysis is that it allows us to study how relevant are the generated samples, or how close are they in distribution to the true ones. Moreover, more better-suited SMOTE extensions could be constructed based on the insights gained from the theoretical analysis. Also, the analysis will shed some insight into the other SMOTE extensions.

The main contributions of this work are summarized as follows:

- In this work, we derive a mathematical formulation for the probability distribution of the SMOTE generated patterns. The presented theoretical formulation is general, and it can be applied to any class-conditional probability distribution. To the best of our knowledge, this is the first theoretical analysis deriving the probability density of the SMOTE generated patterns.
- As a follow-up test, we illustrate the general theoretical analysis by applying it to some distributions for verification of the main contribution.

The paper is organized as follows: Section 2 presents a literature review. Then, the mathematical derivation of SMOTE probability distribution is introduced in Sect. 3. After that, the experimental results are demonstrated in Sect. 4. Finally, Sect. 5 concludes the paper and presents some potential future research directions.

2 Related work

Handling unbalanced data has been extensively studied in the literature (Japkowicz & Stephen, 2002; Kamalov et al., 2022; Wang et al., 2018; Haixiang et al., 2017; Wang et al., 2021; Kaur et al., 2019). However, there are a few studies that provide theoretical or empirical analyses of the data sampling methods, in particular SMOTE. In this review, we focus primarily on these works. Elreedy and Atiya (2019) derive the expectation and covariance matrix of the SMOTE generated patterns. However, the analysis we present here is not restricted to the moments since we develop a mathematical formulation for the density function itself of the SMOTE generated samples. Another key distinction between this work and the work of Elreedy and Atiya (2019) is that the previous work assumes some approximations, while the work presented here is an exact formula. Studying the density characteristics of the SMOTE patterns could stimulate developing density oriented over-sampling methods (Yan et al., 2022; Mayabadi & Saadatfar, 2022). To the best of our knowledge, this is the first analytical formula for the density of SMOTE generated samples. We will limit this work on developing the analytical formula, rather than a complete analysis of SMOTE, as this is outside of the scope of the paper. For a comprehensive analysis of the features of SMOTE refer to the work of Elreedy and Atiya (2019).

Another theoretical analysis for resampling algorithms is performed by Moniz and Monteiro (2021). In particular, the authors apply no free lunch machine learning theorems to imbalanced learning. In addition, they provide a comparative empirical study for different resampling methods which are: random under-sampling, random over-sampling, importance sampling, SMOTE, and SMOTE combined with random under-sampling. The authors conclude that any two resampling strategies would have the same classification performance given no a priori knowledge or data assumptions.

Several empirical studies have been conducted to inspect sampling methods including SMOTE. For example, Luengo et al. (2011) analyze the behavior of different sampling methods, including SMOTE, one of its extensions called SMOTE-ENN, and an under-sampling method named EUSCHC (García & Herrera, 2009). The authors measure the impact of the different sampling methods on the shape of the processed data after sampling including: the overlapping between the different classes, and class separability and its geometrical properties. However, these measures do not consider the distribution of the generated examples.

Furthermore, the study introduced by Dudjak and Martinović (2020) develops a comparative analysis of the classification performance for diverse SMOTE extensions. The study classifies SMOTE extensions into three different categories according to the interpolation mechanism. The three categories are: SMOTE-like interpolation, range restricted interpolation, and multiple interpolations. SMOTE-like interpolation employs the same interpolation mechanism as SMOTE such as: Modified SMOTE (Hu et al., 2009). The range restricted interpolation elects only particular minority class samples for interpolation such as: Borderline SMOTE (Han et al., 2005). The multiple interpolations method adopts multiple neighbors for the interpolation process like Distance-SMOTE (De La Calleja & Fuentes, 2007).

Another piece of work analyzing resampling methods is introduced by García et al. (2010). This study investigates the impacts of the employed classifier and imbalance ratio on the classification performance. The authors recommend using over-sampling for low and moderate class imbalance ratios. Thabtah et al. (2020) provide another in-depth analysis of imbalance ratio and its effect on classifier accuracy using large scale experimental analysis. Moreover, Kamalov et al. (2022) attempt to determine the optimal sampling ratio using a large-scale study The study developed by Dubey et al. (2014) compares among different under-sampling methods, over-sampling methods, and combinations of both approaches. Their experimental analysis considers random over-sampling, SMOTE, random under-sampling, and K-Medoids under-sampling (Dubey et al., 2014). Their work assures that the sophisticated sampling methods such as: SMOTE and K-Medoids surpass random sampling methods.

Bolívar et al. (2022) conduct an empirical analysis evaluating the SMOTE performance on big data. Specifically, the authors consider high dimensional and sparse data. Their results indicate that the sparsity is more influential than dimensionality on the SMOTE performance on big data.

Several contributions have been devoted to determine the optimal over-sampling rate. For example, Weiss and Provost (2003) perform an experimental analysis to find the optimal class ratio from thirteen proposed class distributions by varying the minority class percentage in the training set. They conclude that the optimal balance is not necessarily achieved at full balance and it is a function of the underlying dataset. Albisua et al. (2013) extend the analysis developed by Weiss and Provost (2003) by conducting experiments on several sampling methods and different classifiers. Their experiments demonstrate that the optimal class balance depends not only on the data, but also on the employed classifier and the re-sampling method.

These works provide in-depth analysis of the functionality of SMOTE and other resampling approaches. These analyses are very useful for guiding the researchers for better usage of these algorithms. However, much of this analysis is empirical, and there is little theoretical analysis. This paper attempts to fill this void and provides an exact and full characterization of the probability density of SMOTE-generated patterns. This will help the researchers understand the functioning of SMOTE and find out how the different factors impact its performance.

3 SMOTE density analysis

3.1 SMOTE algorithm

In this section, we briefly describe the SMOTE over-sampling algorithm developed by Chawla et al. (2002). The SMOTE over-sampling algorithm proceeds as follows:

Algorithm 1 A description of SMOTE over-sampling algorithm developed by Chawla et al (2002).

```
repeat
    Select a minority sample $x_0$ at random.
    Identify the $K$-nearest neighbors of $x_0$ among the minority samples.
    Randomly choose one of the $K$ nearest neighbors of $x_0$, from the previous
    step and denote the chosen sample as $x_k$ where $k$ is the rank of the chosen
    neighbor.
    Perform linear interpolation between $x_0$ and the chosen neighbor $x_k$ to
    create a new synthetic sample $z$ according to:
    $z = x_0 + w(x_k - x_0)$ where $w$ is a uniform random variable in the range
    $[0, 1]$.
until $M$ synthetic samples are generated.
```

Figure 1 demonstrates the SMOTE generation mechanism. It can be noted from the figure that the SMOTE patterns lie on the connection lines between the minority class samples and their K nearest neighbors. The inward positioning of SMOTE patterns make them more contracted than the original distribution as inferred by Elreedy and Atiya (2019). This supports our argument that SMOTE patterns do not necessarily

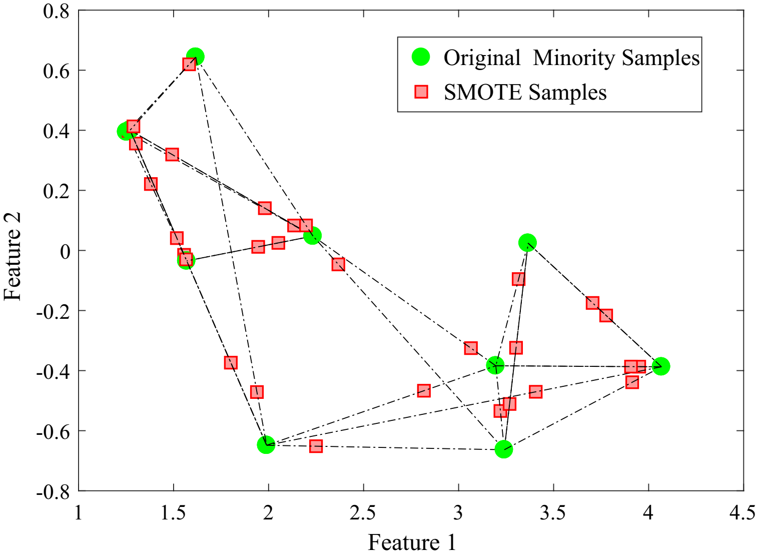

Fig. 1 The SMOTE interpolation mechanism displaying the original minority samples and the SMOTE generated patterns

follow the original minority class density. In this work, we analytically derive the probability density function of the SMOTE generated patterns.

3.2 Notation

In this section, we define the notation adopted in the theoretical analysis.

Let x_0 denote a candidate minority class sample for interpolation by SMOTE, and the original probability density of the minority class is denoted as $p_X(x)$. The synthetically generated sample created by SMOTE is defined as Z. Let d be the dimension of the class pattern.

The total number of minority class samples is defined as N. Let K represent the total number of neighbors used in SMOTE where different numbers of neighbors k are used each time, with k being randomly generated from 1 to K (Chawla et al., 2002). The Euclidean distance between the minority class sample x_0, and its chosen k^{th} neighbor is defined as r.

Let $B(x_0, r)$ define the spherical ball centered at x_0 with radius r enclosing all up to the the k^{th} nearest neighbor of x_0. The integral $I_{B(x_0,r)}$, called the coverage, defines the integral of the minority class density on the ball $B(x_0, r)$. The integral $I_{B(x_0,r)}$ is computed as follows:

$$I_{B(x_0,r)} = \int_{B(x_0,r)} pX(x)dx. \tag{1}$$

The incomplete beta function $B(q; a, b)$ (Dutka, 1981; Al-Sirehy & Fisher, 2013a, b), is defined as:

$$B(q;a,b) = \int_{t=0}^{q} t^{a-1}(1-t)^{b-1}\,dt \tag{2}$$

3.3 Theoretical analysis of SMOTE density

In this section, we introduce a mathematical analysis for the SMOTE oversampling method developed by Chawla et al. (2002). Specifically, we evaluate the probability density function of the SMOTE generated patterns $p_Z(z)$ for a general minority class density $p_X(x)$.

Theorem 1 *Let x be a random sample of a random variable X. Let Z be a random variable defined as a random linear interpolation between K-nearest neighbors of x_k as given in Algorithm* 1. *Then, the probability density of Z is given by*:

$$\begin{aligned} p_Z(z) &= (N-K)\binom{N-1}{K}\int_x p_X(x)\int_{r=\|z-x\|}^{\infty} p_X\left(x+\frac{(z-x)r}{\|z-x\|}\right)\left(\frac{r^{d-2}}{\|z-x\|^{d-1}}\right) \\ &\quad\times B\left(1-I_{B(x,r)};N-K-1,K\right)\mathrm{d}r\,\mathrm{d}x \end{aligned}$$

where $B\left(1-I_{B(x,r)};N-K,K\right)$ *is the incomplete beta function* (*Dutka*, 1981) *defined in Eq.* (2).

Lemma 1 *The conditional probability density function of the SMOTE generated patterns given a certain minority class sample x_0, $p_Z(z|x_0)$ is evaluated as*:

$$\begin{aligned} p_Z(z|x_0) &= (N-K)\binom{N-1}{K}\int_{r=\|z-x_0\|}^{\infty} p_X\left(x_0+\frac{(z-x_0)r}{\|z-x_0\|}\right)\left(\frac{r^{d-2}}{\|z-x_0\|^{d-1}}\right) \\ &\quad\times B\left(1-I_{B(x_0,r)};N-K-1,K\right)\mathrm{d}r \end{aligned}$$

This theorem essentially gives a full analytic solution for the probability density of the SMOTE generated samples. The formulas are *exact* and do not rely on any assumptions or approximations.

Proof of Theorem 1 and Lemma 1 The SMOTE algorithm first selects one of the minority samples x_0 randomly. Assume for the moment that this point x_0 is fixed or given. Then we select a neighbor (say x_k) randomly out of the K nearest neighbors of x_0. Then we pick a point z randomly from the line joining that neighbor x_k and x_0. This is given by the linear interpolation formula:

$$z = (1-w)x_0 + wx_k \tag{3}$$

where w is a uniform random number in [0, 1].

Consider a two-dimensional case for illustration and consider a probability mass located at area A in Fig. 2 which denotes an infinitesimal area element around the chosen neighbor x_k, at distance r from x_0. After applying SMOTE, namely Eq. (3), the probability mass is mapped into the yellow shaded area B, which is an infinitesimal circle sector reaching out till the k nearest neighbors of x_0. Then, the probability mass at area C, which is an

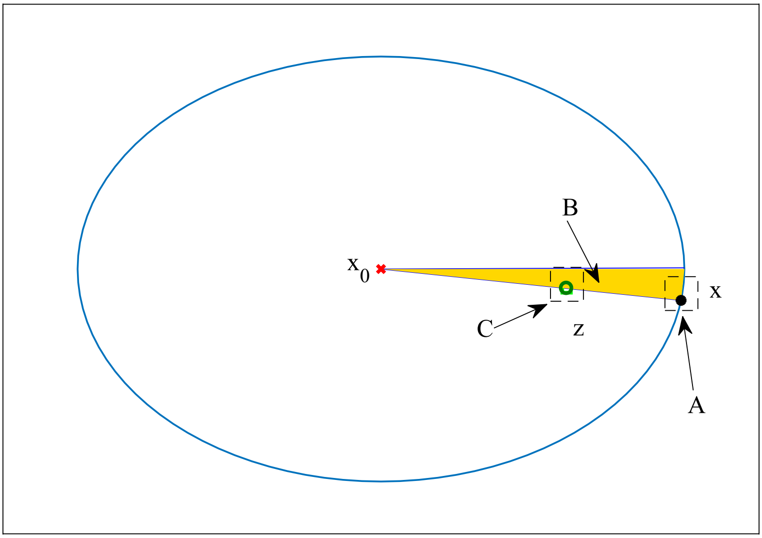

Fig. 2 SMOTE density mapping clarification, where x_0 is the minority class pattern on which SMOTE is applied, x_k denotes the randomly chosen neighbor, and z is the SMOTE generated sample

infinitesimal area element around the SMOTE generated pattern z (somewhere between x_k and x_0 according to the value of w) can be evaluated as:

$$p(z \in C) = \frac{p(z \in A)}{r \frac{\|z - x_0\|}{r}} \tag{4}$$

where r denotes the euclidean distance between x_0 and its chosen neighbor x_k, as defined in Sect. 3.2. The division by r is because the mass of A gets spread out to a region of length r, thus, diluting the density by that amount. Essentially, the probability mass that is concentrated in area A will become spread out to the whole circle sector B by the randomized linear interpolation procedure (Eq. 3). Thus, the density will be divided by r, due to the target area being bigger by that amount.

The division by $\frac{\|z - x_0\|}{r}$ is because the area at C is smaller than the area of A by that amount (due to the ratio of arc lengths: the arc length for C is $\|z - x_0\| d\theta$ (radius times $d\theta$), while for A it is $rd\theta$), so the probability mass gets more concentrated by that amount. This analysis can be generalized to the d-dimensional case as follows:

$$p(z \in C) = p(z \in A) \frac{r^{d-2}}{\|z - x_0\|^{d-1}} \tag{5}$$

According to the SMOTE method's geometry and as shown in Fig. 2, Z lies in the line connecting between x_0 and its the chosen neighbor x. Then:

$$r = \|x - x_0\| \geq \|z - x_0\| \tag{6}$$

The previous analysis assumes that the k^{th} neighbor is fixed and at a distance r from x_0, i.e. the probability density computed is conditioned on this assumption. Next step is to assume that r is random and take the expectation over r:

$$p_Z(z|x_0,k) = \int_{r=\|z-x_0\|}^{\infty} p_Z(z|x_0,r,k)p(r|x_0)\mathrm{d}r \tag{7}$$

Note that we used $r = \|z - x_0\|$ as the lower limit of the integral because this is implied in Eq. (6): the distance $\|z - x_0\|$ of the interpolated point is always smaller than or equal to the distance of the k^{th} neighbor r. The term $p(r|x_0)$ represents the probability density that a k^{th} nearest neighbor of x_0 is located at distance r away from x_0. This term will be evaluated later. Substituting from Eq. (5) into Eq. (7):

$$p_Z(z|x_0,k) = \int_{r=\|z-x_0\|}^{\infty} \frac{p_X\left(x_0 + \frac{(z-x_0)r}{\|z-x_0\|}\right)}{\int_{S(x_0,r)} p_X(x)\mathrm{d}x} \left(\frac{r^{d-2}}{\|z-x_0\|^{d-1}}\right) p(r|x_0)\mathrm{d}r \tag{8}$$

where $S(x_0, r)$ denotes a spherical shell around x_0 with radius r. The first quotient (p_X divided by the integral over the spherical shell) represents the probability density of a point at the location x_k, given that it occurs at a distance r from x_0. This means that the k^{th} neighbor (the point that is the target of the interpolation) is at distance r away from x_0. This quotient is obtained by straightforward application of Bayes probability rule, as follows:

$$p_X(x|x \in S(x_0,r)) = \frac{p_X(x, x \in S(x_0,r))}{p(x \in S(x_0,r))} = \frac{p_X(x)}{\int_{S(x_0,r)} p_X(x)\mathrm{d}x} \tag{9}$$

Note that the point x_k is written as $x_0 + \frac{(z-x_0)r}{\|z-x_0\|}$ according to Eq. (3) and enforcing the fact that it is a distance r away from x_0 (in other words that $x \in S(x_0, r)$ according to Eq. (9). The reason for writing it this way is that it has to be expressed in terms of z. So, x_k is written as x_0 plus the unit vector in the direction of z: $\frac{(z-x_0)}{\|z-x_0\|}$ multiplied by r so that it lands on the shell that is distance r away from x_0.

In summary, the derivation proceeds in several steps. In the first step we assume that the k^{th} neighbor is at a fixed distance r away from x_0, and evaluate the probability density (given r). Next step is to obtain the probability density of z given that the neighbor is a distance r away. If the neighbor is r distance away, then this means that the neighbor is located on a spherical shell of radius r. Using Bayes formula we obtain the quotient indicated in the formula. Next step we take the expectation over r, using the probability that the k^{th} neighbor is distance r away.

Evaluating $p(r|x_0)$ in Eq. (7):

According to the work of Fukunaga and Hostetler (1973), the coverage u of the k nearest neighbors is denoted as:

$$u = G(r) = \int_{B(x_0,r)} p_X(x)\,\mathrm{d}x \tag{10}$$

where $B(x_0, r)$ is the ball around x_0 enclosing up to the k^{th} nearest neighbor of x_0 and $p_X(x)$ represents the probability density function of the underlying distribution (from which the k neighbors are drawn).

Furthermore, according to the work of Fukunaga and Hostetler (1973), u follows a Beta distribution such that $u \sim Beta(u;k, N - k)$. Then, the density $p_U(u)$ is defined as:

$$p_U(u) = \frac{(N-1)!u^{k-1}(1-u)^{N-k-1}}{(k-1)!(N-k-1)!} \tag{11}$$

Using the theory of transformation of random variables (Magdon-Ismail & Atiya, 2002; Venkatesh, 2013), the probability density of r, the distance from x_0 to the k^{th} neighbor is given by:

$$p(r|x_0) = p_U(u)|\frac{du}{dr}| \tag{12}$$

Using Eq. (10), then Eq. (12) could be written as:

$$p(r|x_0) = p_U(u)G'(r) \tag{13}$$

where $G'(r)$ is the first derivative of $G(r)$.

Substitute from Eq. (11) into Eq. (13) yields:

$$p(r|x_0) \;=\; (N-1)\binom{N-2}{k-1}G^{k-1}(r)(1-G(r))^{N-k-1}G'(r) \tag{14}$$

Substituting from Eqs. (10) and (11) into Eq. (14), then the conditional probability density of r given a minority sample x_0, $p(r|x_0)$ is evaluated as:

$$\begin{aligned} p(r|x_0) \;&= (N-1)\binom{N-2}{k-1}\left(\int_{B(x_0,r)} p_X(x)\mathrm{d}x\right)^{k-1}\left(1-\left(\int_{B(x_0,r)} p_X(x)\mathrm{d}x\right)\right)^{N-k-1} \\ &\times \int_{S(x_0,r)} p_X(x)\mathrm{d}x \end{aligned} \tag{15}$$

where $G'(r)$ equals the integral over the shell of radius r: $\int_{S(x_0,r)} p_X(x)\mathrm{d}x$ because changing r to $r + dr$ in Eq. (10) will yield this integral over the shell. Substituting from Eq. (15) into Eq. (8) produces the following:

$$\begin{aligned} p_Z(z|x_0,k) =&(N-1)\binom{N-2}{k-1}\int_{r=\|z-x_0\|}^{\infty}\left(\int_{B(x_0,r)} p_X(x)\mathrm{d}X\right)^{k-1} \\ &\times\left(1-\left(\int_{B(x_0,r)} p_X(x)\mathrm{d}X\right)\right)^{N-k-1}\left[\int_{S(x_0,r)} p_X(x)\mathrm{d}x\right]\frac{p_X\left(x_0+\frac{(z-x_0)r}{\|z-x_0\|}\right)}{\int_{S(x_0,r)} p_X(x)\mathrm{d}x} \\ &\times\left(\frac{r^{d-2}}{\|z-x_0\|^{d-1}}\right)\mathrm{d}r \end{aligned} \tag{16}$$

Then, Eq. (16) can be simplified into:

$$\begin{aligned} p_Z(z|x_0,k) =&(N-1)\binom{N-2}{k-1}\int_{r=\|z-x_0\|}^{\infty}\left(\int_{B(x_0,r)} p_X(x)\mathrm{d}x\right)^{k-1} \\ &\times\left(1-\left(\int_{B(x_0,r)} p_X(x)\mathrm{d}x\right)\right)^{N-k-1} p_X\left(x_0+\frac{(z-x_0)r}{\|z-x_0\|}\right)\left(\frac{r^{d-2}}{\|z-x_0\|^{d-1}}\right)\mathrm{d}r \end{aligned} \tag{17}$$

So far we have assumed that we will take a fixed neighbor k. Since we select a neighbor at random among the K neighbors with probability $\frac{1}{K}$ next step we will take the expectation over this random selection over k. This results in the following:

$$
\begin{aligned}
p_Z(z|x_0) = \sum_{k=1}^{K} \frac{N-1}{K}\binom{N-2}{k-1}\int_{r=\|z-x_0\|}^{\infty}\left(\int_{B(x_0,r)} p_X(x)\mathrm{d}x\right)^{k-1} \\
\times\left(1 - \left(\int_{B(x_0,r)} p_X(x)\mathrm{d}x\right)\right)^{N-k-1} p_X\left(x_0+\frac{(z-x_0)r}{\|z-x_0\|}\right)\left(\frac{r^{d-2}}{\|z-x_0\|^{d-1}}\right)\mathrm{d}r
\end{aligned}
\tag{18}
$$

Let $I_{B(x_0,r)} = \int_{B(x_0,r)} p_X(x)dx$. Then:

$$
\begin{aligned}
p_Z(z|x_0) = \frac{N-1}{K}\int_{r=\|z-x_0\|}^{\infty}\sum_{k=1}^{K}\binom{N-2}{k-1} I_{B(x_0,r)}^{k-1}\left(1-I_{B(x_0,r)})\right)^{N-k-1} \\
\times p_X\left(x_0+\frac{(z-x_0)r}{\|z-x_0\|}\right)\left(\frac{r^{d-2}}{\|z-x_0\|^{d-1}}\right)\mathrm{d}r
\end{aligned}
\tag{19}
$$

Define $J(x_0, r)$ as follows:

$$
J(x_0, r) = \sum_{k=1}^{K}\binom{N-2}{k-1} I_{B(x_0,r)}^{k-1}\left(1-I_{B(x_0,r)})\right)^{N-k-1}
\tag{20}
$$

Equation (20) can be written as:

$$
J(x_0, r) = \sum_{m=0}^{K-1}\binom{N-2}{m} I_{B(x_0,r)}^{m}\left(1-I_{B(x_0,r)})\right)^{N-m-2}
\tag{21}
$$

From Eq. (21), $J(x_0, r)$ is a cumulative probability function $F(I_{B(x_0,r)}, N-2, K-1)$ for the Binomial distribution $\mathcal{B}(N-2, I_{B(x_0,r)})$.

The cumulative probability function $F(y, N, k)$ (Wadsworth, 1960) can be expressed as:

$$
F(y, N, k) = (N-k)\binom{N}{k}\int_{t=0}^{1-y} t^{N-k-1}(1-t)^k\, dt
\tag{22}
$$

Thus, $F(I_{B(x_0,r)}, N-2, K-1)$ is evaluated as follows:

$$
F(I_{B(x_0,r)}, N-1, K-1) = (N-K-1)\binom{N-2}{K-1}\int_{t=0}^{1-I_{B(x_0,r)}} t^{N-K-2}(1-t)^{K-1}\, dt
\tag{23}
$$

However, Eq. (23) can be expressed in terms of the incomplete beta function $B(1-I_{B(x_0,r)}; N-K-1, K)$ (Dutka, 1981), and defined in Eq. (2). From Eq. (23), $J(x_0, r)$ can be formulated as:

$$
\begin{aligned}
J(x_0, r) &= F(I_{B(x_0,r)}, N-1, K) \\
&= (N-K-1)\binom{N-2}{K-1} B\left(1-I_{B(x_0,r)}; N-K-1, K\right)
\end{aligned}
\tag{24}
$$

Substitute from Eqs. (20) and (24) into Eq. (19):

$$
\begin{aligned}
p_Z(z|x_0) &= \frac{N-1}{K}(N-K)\binom{N-2}{K-1}\int_{r=\|z-x_0\|}^{\infty} p_X\left(x_0+\frac{(z-x_0)r}{\|z-x_0\|}\right)\left(\frac{r^{d-2}}{\|z-x_0\|^{d-1}}\right) \\
&\times B\left(1-I_{B(x_0,r)}; N-K-1, K\right)\mathrm{d}r
\end{aligned}
\tag{25}
$$

Simplifying Eq. (25) results in:

$$\begin{aligned} p_Z(z|x_0) &= (N-K)\binom{N-1}{K}\int_{r=\|z-x_0\|}^{\infty} p_X\left(x_0+\frac{(z-x_0)r}{\|z-x_0\|}\right)\left(\frac{r^{d-2}}{\|z-x_0\|^{d-1}}\right) \\ &\quad \times B\left(1-I_{B(x_0,r)};N-K-1,K\right)\mathrm{d}r \end{aligned} \tag{26}$$

Accordingly, Eq. (26) proves Lemma 1.

Finally, taking expectation over x_0 yields the density of the SMOTE generated patterns $p(Z)$.

$$\begin{aligned} p_Z(z) &= (N-K)\binom{N-1}{K}\int_x p_X(x)\int_{r=\|z-x\|}^{\infty} p_X\left(x+\frac{(z-x)r}{\|z-x\|}\right)\left(\frac{r^{d-2}}{\|z-x\|^{d-1}}\right) \\ &\quad \times B\left(1-I_{B(x,r)};N-K-1,K\right)\mathrm{d}r\,\mathrm{d}x \end{aligned} \tag{27}$$

□

Consequently, Eq. (27) proves Theorem 1.

4 Experiments and results

In this section, we present the results of the numerical experiments conducted in support of our derived theoretical analysis presented in Sect. 3. To this end, we estimate the SMOTE patterns density $p(Z)$ using the developed theoretical analysis, and also evaluate the SMOTE density $p(Z)$ empirically. Then, we compare the two density estimates for verification.

In our experiments, we adopt two different distributions: multivariate uniform distribution over a disk, and multivariate Gaussian distribution. The uniform distribution is taken over a two-dimensional disk centered at the origin. The disk radius is set to 3, so the density equals $\frac{1}{9\pi}$ for $x_1^2+x_2^2 \le 3$ and zero otherwise.

We use a two-dimensional zero mean multivariate Gaussian distribution of the minority class samples, $\mu = [0,0]$, and we have examined different covariance matrices: the identity matrix $\Sigma = \mathbb{I}_2$, $\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 4 \end{bmatrix}$, and and $\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1.5 \end{bmatrix}$.

We perform the empirical density estimation of SMOTE patterns $p(Z)$ according to Algorithm 1 above. We have examined two different values for the number of original patterns for applying SMOTE at a single run, specifically, we tried $N = 30$ and $N = 50$ to mimic the scarcity of minority class patterns when applying SMOTE. For the K parameter in the K nearest neighbor applied in SMOTE, we have adopted two values: $K = 3$ and $K = 5$. The number of generated SMOTE patterns for the empirical density estimation is set to $M = 5 \times 10^7$ in order to obtain an accurate density estimate.

To implement the theoretically derived formula, we use a two-dimensional grid for integration. For display, we often hold one feature or dimension to be constant and evaluate the density of the other feature. This facilitates the presentation of the theoretical versus empirical densities on the same plots, as it may be hard to visualize two two-dimensional densities in a single plot.

We tested both results of the Theorem and the Lemma in the experiments. The first experiments, described this paragraph, verify the accuracy of $p_Z(x)$, while the next experiments verify the conditional distribution $p_Z(z|x_0)$. Figures 3 and 4 represent the marginal density of SMOTE generated patterns $p_Z(z)$ as given in Theorem 1 for the multivariate uniform distribution over a disk. The figures show the theoretical density estimated using the presented analysis, the empirical density, and the true original minority class density. As mentioned, these are one-dimensional sections in the 2-D for visualization purposes. The figures demonstrate the closeness between the theoretical and empirical density estimates which verifies the proposed analysis. Moreover, it can be noted from Figs. 3 and 4 that both of the theoretical and empirical SMOTE densities are close to the original minority class density in case of the multivariate uniform distribution over a disk. In other words, these results imply that the SMOTE patterns are adequate representatives for the original minority class in case of uniform distribution original density.

Figures 5 and 6 depict the SMOTE patterns density given a certain original data point x_0 (i.e. $p_Z(z|x_0)$) empirically and theoretically for the multivariate Gaussian distribution (as given in Lemma 1). The figures show the conformity between the the theoretical and empirical density estimates which confirms our introduced theoretical analysis. In these figures, the true density can not be plotted as we evaluate the conditional density of SMOTE samples given a particular original minority sample x_0.

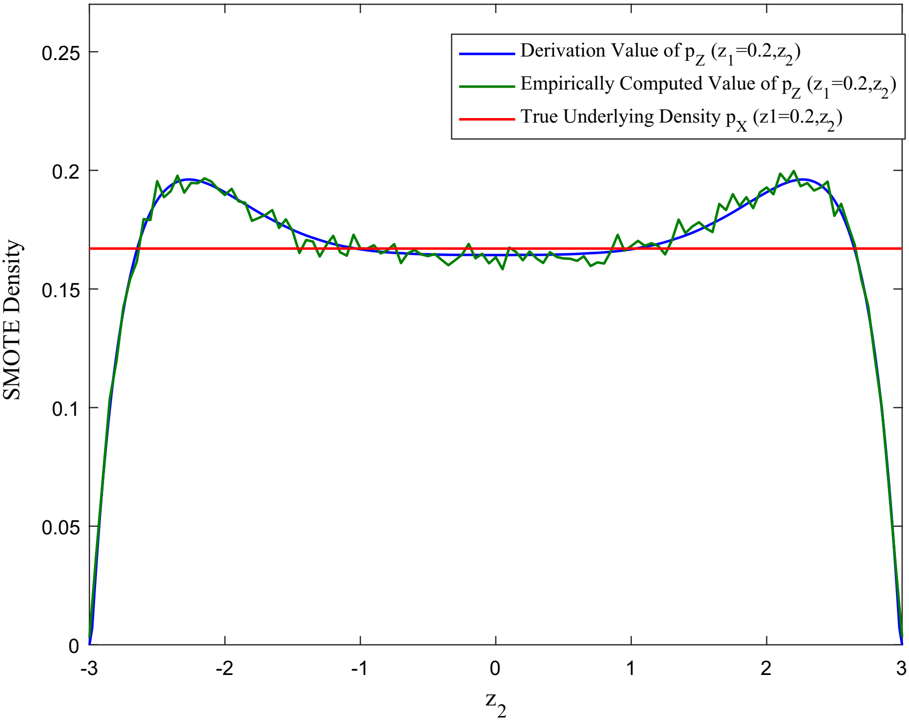

Fig. 3 Empirical versus theoretical densities of SMOTE patterns $p_Z(z)$ for 2-dimensional disk for $z_1 = 0.2$ using $N = 30$ and $K = 3$

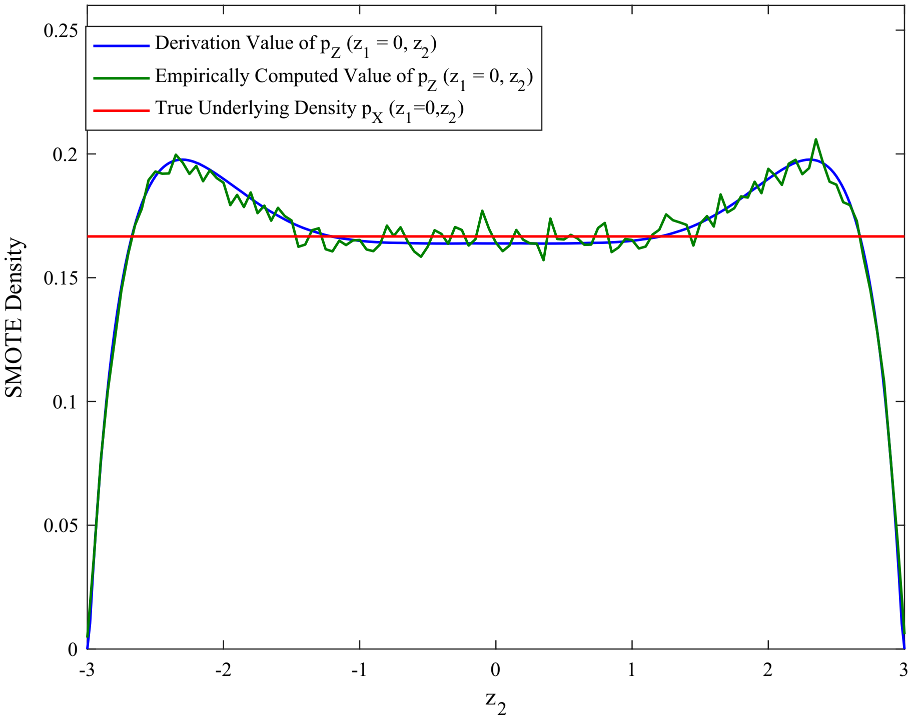

Fig. 4 Empirical versus theoretical densities of SMOTE patterns $p_Z(z)$ for 2-dimensional disk for $z_1 = 0$ using $N = 50$ and $K = 5$

Figure 7 demonstrates the two-dimensional density for SMOTE patterns estimated using the presented analysis. Similarly, Fig. 8 shows the SMOTE density estimated empirically. In order to obtain smooth results for the empirical estimation of the 2-dimensional density, we adopt the Parzen window density estimation (Parzen, 1962; Rosenblatt, 1956). We use Gaussian kernel, and the kernel width h is set to 0.05.

It could be observed from Figs. 7 and 8 that the SMOTE density is concentrated around x_0. This is reasonable as the SMOTE method places the synthetic patterns inwards around the minority class samples as presented in Fig. 1. These results are consistent with the argument of the contracteness behavior of the SMOTE oversampling method as raised by Elreedy and Atiya (2019). Specifically, the figures show that the SMOTE density estimates either theoretically or empirically are centered around the original minority sample x_0 on which SMOTE is applied.

In the next experiment we test whether the proposed formulas provide accurate estimates for the case of classification. Of course, obtaining the correct density is the building block of any further classification method, and therefore should guarantee accurate computation of any classification-based outcome such as classification accuracy. We considered a simple two-class classification problem, where the minority class is a uniform distribution over a two-dimensional disk centered at the origin, where the disk radius is set to 3, so the density equals $\frac{1}{9\pi}$ for $x_1^2 + x_2^2 \leq 3$ and zero otherwise. The majority class also has a uniform density over a disk of radius 3, but centered around a mean vector $\mu = (a,\ a)^T$, where a is a number that we vary. We considered a linear discriminant function classifier and applied

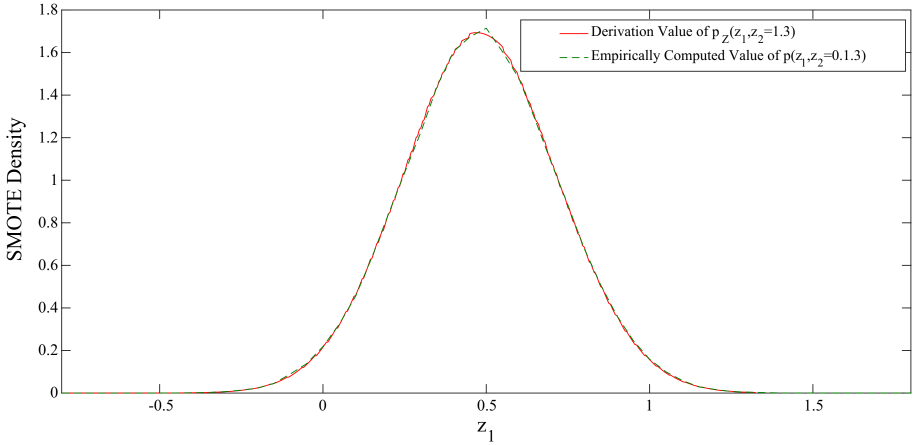

Fig. 5 Conditional empirical versus theoretical densities of SMOTE patterns $p_Z(z|x_0)$ for 2-dimensional Gaussian original distribution, $x_0 = [0.5, 1]$ and $z_2 = 1.3$ for $p_X(x) \sim \mathcal{N}([0, 0], [1\ 0.8; 0.8\ 4])$ using $N = 50$ and $K = 3$

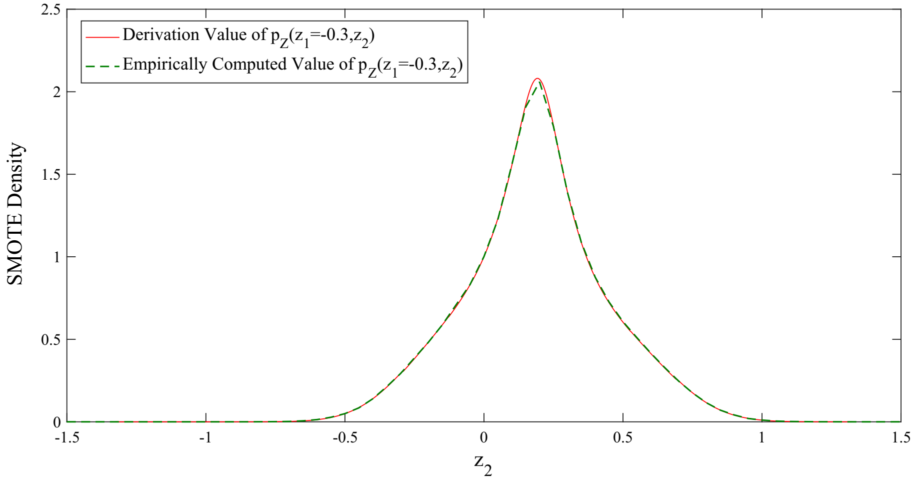

Fig. 6 Conditional empirical versus theoretical densities of SMOTE patterns $p_Z(z|x_0)$ for 2-dimensional Gaussian original distribution, $x_0 = [-0.2, 0.2]$ and $z_1 = -0.3, p_X(x) \sim \mathcal{N}([0, 0], [2\ 0; 0\ 1.5])$ using $N = 50$ and $K = 5$

our theoretical formulas versus the empirical way of generating SMOTE samples. We computed the geometric mean (G-mean) (Barandela et al., 2003), defined in Eq. (28) for both methods for several values of a. Table 1 shows the G-mean result. One can observe that both theoretical and empirical approaches produce very close numbers, indicating the accuracy of the developed formula.

$$Gmean = \sqrt{Sensitivity \times Specificity} \tag{28}$$

The sensitivity and specificity metrics are defined in Eqs. (29) and (30), respectively:

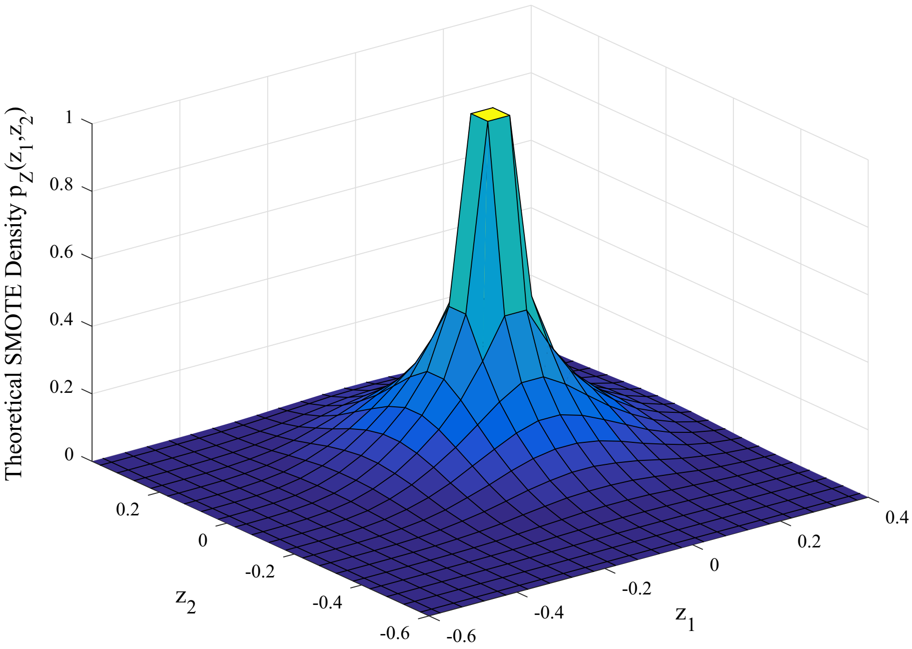

Fig. 7 Theoretical SMOTE density $p_Z(z_1, z_2)$ for 2-dimensional Gaussian original distribution, $x_0 = [0, 0]$ and for $p_X(x) \sim \mathcal{N}([0, 0], \mathbb{I})$ using $N = 50$ and $K = 3$

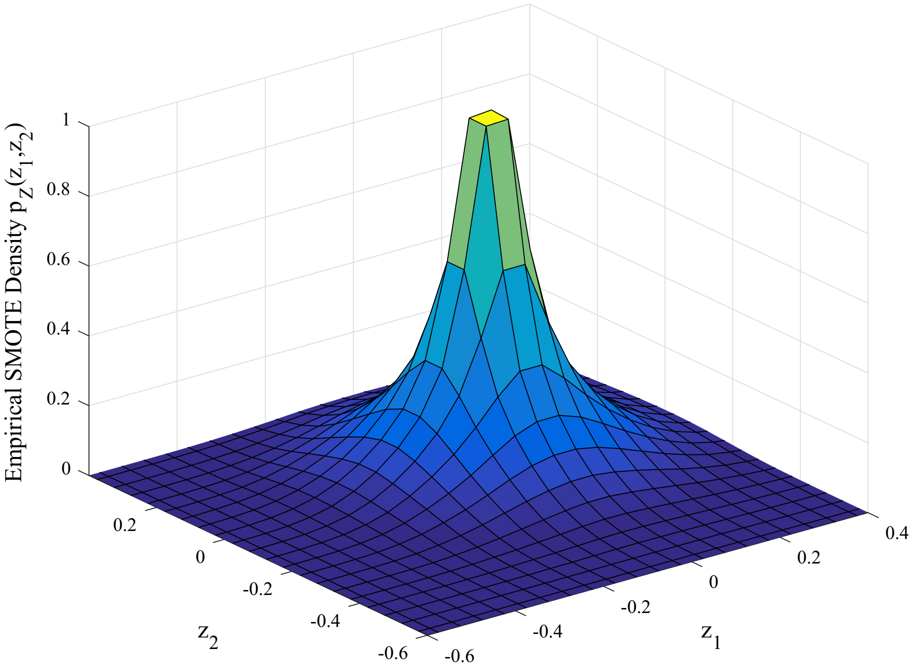

Fig. 8 Empirical SMOTE density $p(Z_1, Z_2)$ for 2-dimensional Gaussian original distribution using parzen window density estimation, $x_0 = [0, 0]$ and for $p_X(x) \sim \mathcal{N}([0, 0], \mathbb{I})$ using $N = 50$ and $K = 3$

Table 1 Theoretical and empirical Gmean values for different values of parameter a determining the mean of the majority class where $\mu = (a,\ a)^T$

a	Gmean-theoretical (%)	Gmean-empirical (%)
2	78.90070	78.90819
2.5	85.22475	85.22682
3	90.92420	90.92419
4	99.19205	99.18744

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{29}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \tag{30}$$

5 Conclusions and future work

In this paper, we develop a theoretical analysis of one of the most dominant over-sampling methods: the Synthetic Minority over-sampling TEchnique (SMOTE) method. The SMOTE algorithm is a very powerful over-sampling method for generating artificial minority class samples in order to balance the class distribution. However, the synthetic data generated by SMOTE may not exactly follow the original minority class distribution, which could impact the classification performance. Thus, this work theoretically analyzes the distribution of the synthetically generated patterns. Specifically, we introduce a full derivation of the probability density function of the SMOTE generated patterns. We applied the developed analysis to some distributions and verified correctness of the presented theoretical analysis by comparing with the empirical density estimates. The goal here has been to focus on deriving a complete and exact formula. Providing a theoretical formula would lay the groundwork for further analysis and guide further modifications of SMOTE in the future in the direction that improves its functionality. This is how this could potentially lead to improving the efficiency of classifiers and providing a better understanding of SMOTE. For example, this can lead to optimal formulas that set different weight to the original samples versus the weight given to the generated samples, in any classification scheme. Analyzing the behavior of SMOTE or quantifying the deviation of SMOTE from the true density could be performed in future work. Another potential future research direction could be investigating more complex minority class distributions such as multi-modal distributions, which present significant challenge for SMOTE in particular.

Author contributions Conceptualization: DE and AFA; Formal analysis: DE and AFA; Methodology: DE and AFA; Project administration: AFA and FK; Resources: AFA and FK; Software: DE and AFA; Supervision: AFA and FK; Validation: DE, AFA and FK; Writing: DE, AFA and FK.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB). This research received no external funding.

Data availibility The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Code availability The source code of the current work is available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors of this work declare no conflict of interest.

Consent for publication The presented work involves no experiments on individuals.

Ethical approval The experiments presented in this work do not involve individuals neither humans nor animals.

Consent to participate The presented work involves no experiments on individuals.

References

Abd Elrahman, S. M., & Abraham, A. (2013). A review of class imbalance problem. *Journal of Network and Innovative Computing, 1*(2013), 332–340.

Ahsan, M., Gomes, R., & Denton, A. (2018). Smote implementation on phishing data to enhance cybersecurity. In *2018 IEEE International Conference on Electro/Information Technology (EIT)* (pp. 0531–0536). IEEE.

Al-Sirehy, F., & Fisher, B. (2013). Further results on the beta function and the incomplete beta function. *Applied Mathematical Sciences, 7*(70), 3489–3495.

Al-Sirehy, F., & Fisher, B. (2013). Results on the beta function and the incomplete beta function. *International Journal of Applied Mathematics, 26*(2), 191.

Albisua, I., Arbelaitz, O., Gurrutxaga, I., Lasarguren, A., Muguerza, J., & Perez, J. M. (2013). The quest for the optimal class distribution: An approach for enhancing the effectiveness of learning via resampling methods for imbalanced data sets. *Progress in Artificial Intelligence, 2*(1), 45–63.

Atiya, A., Talaat, N., & Shaheen, S. (1997). An efficient stock market forecasting model using neural networks. In *Proceedings of International Conference on Neural Networks (ICNN'97)* (pp. 2112–2115). IEEE.

Balogun, A.O., Lafenwa-Balogun, F. B., Mojeed, H. A., Adeyemo, V. E., Akande, O. N., Akintola, A. G., Bajeh, A. O., & Usman-Hamza, F. E. (2020). Smote-based homogeneous ensemble methods for software defect prediction. In *International Conference on Computational Science and its Applications* (pp. 615–631). Springer

Barandela, R., Sánchez, J. S., Garcıa, V., & Rangel, E. (2003). Strategies for learning in class imbalance problems. *Pattern Recognition, 36*(3), 849–851.

Batista, G., Prati, R., & Monard, M. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter, 6*(1), 20–29.

Bej, S., Davtyan, N., Wolfien, M., Nassar, M., & Wolkenhauer, O. (2021). Loras: An oversampling approach for imbalanced datasets. *Machine Learning, 110*(2), 279–301.

Bolívar, A., García, V., Florencia, R., Alejo, R., Rivera, G., & Sanchez-Solis, J. P. (2022). A preliminary study of smote on imbalanced big datasets when dealing with sparse and dense high dimensionality. In *Mexican Conference on Pattern Recognition* (pp. 46–55). Springer.

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks, 106*, 249–259.

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 475–482). Springer.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Chennuru, V. K., & Timmappareddy, S. R. (2022). Simulated annealing based undersampling (SAUS): A hybrid multi-objective optimization method to tackle class imbalance. *Applied Intelligence, 52*(2), 2092–2110.

Dablain, D., Krawczyk, B., & Chawla, N. V. (2022). Deepsmote: Fusing deep learning and smote for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*. https://doi.org/10.1109/TNNLS.2021.3136503

De La Calleja, J., & Fuentes, O. (2007). A distance-based over-sampling method for learning from imbalanced data sets. In *FLAIRS Conference* (pp. 634–635).

Devi, D., Biswas, S. K., & Purkayastha, B. (2022). Correlation-based oversampling aided cost sensitive ensemble learning technique for treatment of class imbalance. *Journal of Experimental & Theoretical Artificial Intelligence, 34*(1), 143–174.

Dubey, R., Zhou, J., Wang, Y., Thompson, P. M., Ye, J., & Alzheimer's Disease Neuroimaging Initiative (2014). Analysis of sampling techniques for imbalanced data: An n= 648 ADNI study. *NeuroImage, 87*, 220–241.

Dudjak, M., & Martinović, G. (2020). In-depth performance analysis of smote-based oversampling algorithms in binary classification. *International Journal of Electrical and Computer Engineering Systems, 11*(1), 13–23.

Dutka, J. (1981). The incomplete beta function—A historical profile. *Archive for History of Exact Sciences, 24*, 11–29.

Elreedy, D., & Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Information Sciences, 505*, 32–64.

Elyan, E., Moreno-Garcia, C. F., & Jayne, C. (2021). Cdsmote: Class decomposition and synthetic minority class oversampling technique for imbalanced-data classification. *Neural Computing and Applications, 33*(7), 2839–2851.

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research, 61*, 863–905.

Fotouhi, S., Asadi, S., & Kattan, M. W. (2019). A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of Biomedical Informatics, 90*(103), 089.

Fukunaga, K., & Hostetler, L. (1973). Optimization of k nearest neighbor density estimates. *IEEE Transactions on Information Theory, 19*(3), 320–326.

Ganaie, M., Tanveer, M., & Alzheimer's Disease Neuroimaging Initiative (2021). Fuzzy least squares projection twin support vector machines for class imbalance learning. *Applied Soft Computing, 113*, 107933.

Gao, M., Hong, X., Chen, S., et al. (2014). Pdfos: Pdf estimation based over-sampling for imbalanced two-class problems. *Neurocomputing, 138*, 248–259.

García, S., & Herrera, F. (2009). Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. *Evolutionary Computation, 17*(3), 275–306.

García, V., Sánchez, J., & Mollineda, R. (2010). Exploring the performance of resampling strategies for the class imbalance problem. In *Trends in applied intelligent systems* (pp. 541–549).

Goodman, J., Sarkani, S., & Mazzuchi, T. (2022). Distance-based probabilistic data augmentation for synthetic minority oversampling. *ACM/IMS Transactions on Data Science (TDS), 2*(4), 1–18.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"* (pp. 986–996). Springer.

Guzmán-Ponce, A., Sánchez, J. S., Valdovinos, R. M., & Marcial-Romero, J. R. (2021). DBIG-US: A two-stage under-sampling algorithm to face the class imbalance problem. *Expert Systems with Applications, 168*(114), 301.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications, 73*, 220–239.

Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *International Conference on Intelligent Computing* (pp. 878–887). Springer.

He, H., Bai, Y., Garcia, E. A., & Li, S. A. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *IEEE International Joint Conference on Computational Intelligence, IJCNN* (pp. 1322–1328). IEEE.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications, 13*(4), 18–28.

Hu, S., Liang, Y., Ma, L., & He. Y. (2009). Msmote: Improving classification performance when training data is imbalanced. In *2009 Second International Workshop on Computer Science and Engineering* (pp. 13–17). IEEE.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis, 6*(5), 429–449.

Kamalov, F., Atiya, A. F., & Elreedy, D. (2022). Partial resampling of imbalanced data. arXiv preprint arXiv:2207.04631

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys (CSUR), 52*(4), 1–36.

Kishor, A., & Chakraborty, C. (2021). Early and accurate prediction of diabetics based on FCBF feature selection and smote. *International Journal of System Assurance Engineering and Management*. https://doi.org/10.1007/s13198-021-01174-z

Koziarski, M., Bellinger, C., & Woźniak, M. (2021). RB-CCR: Radial-based combined cleaning and resampling algorithm for imbalanced data classification. *Machine Learning, 110*(11), 3059–3093.

Li, Z., Huang, M., Liu, G., & Jiang, C. (2021). A hybrid method with dynamic weighted entropy for handling the problem of class imbalance with overlap in credit card fraud detection. *Expert Systems with Applications, 175*, 114750.

Liu, L., Wu, X., Li, S., Tan, S., & Bai, Y. (2022). Solving the class imbalance problem using ensemble algorithm: Application of screening for aortic dissection. *BMC Medical Informatics and Decision Making, 22*(1), 1–16.

Luengo, J., Fernández, A., García, S., & Herrera, F. (2011). Addressing data complexity for imbalanced data sets: Analysis of smote-based oversampling and evolutionary undersampling. *Soft Computing, 15*(10), 1909–1936.

Magdon-Ismail, M., & Atiya, A. (2002). Density estimation and random variate generation using multilayer networks. *IEEE Transactions on Neural Networks, 13*(3), 497–520.

Mayabadi, S., & Saadatfar, H. (2022). Two density-based sampling approaches for imbalanced and overlapping data. *Knowledge-Based Systems, 241*, 108217.

Moniz, N., & Monteiro, H. (2021). No free lunch in imbalanced learning. *Knowledge-Based Systems, 227*, 107222.

Mullick, S. S., Datta, S., & Das, S. (2018). Adaptive learning-based *k*-nearest neighbor classifiers with resilience to class imbalance. *IEEE Transactions on Neural Networks and Learning Systems, 29*(11), 5713–5725.

Nguyen, H. M., Cooper, E. W., & Kamei, K. (2011). Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms, 3*(1), 4–21.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics, 33*(3), 1065–1076.

Prati, R. C., Batista, G. E., & Monard, M. C. (2004). Learning with class skews and small disjuncts. In *Brazilian Symposium on Artificial Intelligence* (pp. 296–306). Springer.

Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys (CSUR), 28*(1), 71–72.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics, 27*, 832–837.

Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences, 513*, 429–441.

Venkatesh, S. S. (2013). *The theory of probability: Explorations and applications*. Cambridge University Press.

Vuttipittayamongkol, P., & Elyan, E. (2020). Neighbourhood-based undersampling approach for handling imbalanced and overlapped data. *Information Sciences, 509*, 47–70.

Wadsworth, G. P. (1960). Introduction to probability and random variables. Tech. rep.

Wan, Z., Zhang, Y., & He, H. (2017). Variational autoencoder based synthetic data generation for imbalanced learning. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1–7). IEEE.

Wang, L., Han, M., Li, X., Zhang, N., & Cheng, H. (2021). Review of classification methods on unbalanced data sets. *IEEE Access, 9*, 64606–64628.

Wang, S., Minku, L. L., & Yao, X. (2018). A systematic study of online class imbalance learning with concept drift. *IEEE Transactions on Neural Networks and Learning Systems, 29*(10), 4802–4821.

Weiss, G. M., & Provost, F. (2003). Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research, 19*, 315–354.

Wu, X., & Meng, S. (2016). E-commerce customer churn prediction based on improved SMOTE and AdaBoost. In *2016 13th International Conference on Service Systems and Service Management (ICSSSM)* (pp. 1–5). IEEE.

Yan, Y., Jiang, Y., Zheng, Z., Yu, C., Zhang, Y., & Zhang, Y. (2022). LDAS: Local density-based adaptive sampling for imbalanced data classification. *Expert Systems with Applications, 191*, 116213.

Zhang, H., & Li, M. (2014). RWO-sampling: A random walk over-sampling approach to imbalanced data classification. *Information Fusion, 20*, 99–116.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.