UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Marcelo Ladeira Marques

Uma abordagem baseada em classificadores de larga margem para geração de dados artificiais em bases desbalanceadas

Juiz de Fora

2017

UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Marcelo Ladeira Marques

Uma abordagem baseada em classificadores de larga margem para geração de dados artificiais em bases desbalanceadas

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Ciências Exatas da Universidade Federal de Juiz de Fora como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Orientador: Carlos Cristiano Hasenclever Borges
Coorientador: Saulo Moraes Villela

Juiz de Fora
2017

Marcelo Ladeira Marques

Uma abordagem baseada em classificadores de larga margem para geração de dados artificiais em bases desbalanceadas

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Ciências Exatas da Universidade Federal de Juiz de Fora como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Aprovada em 1 de Setembro de 2017.

BANCA EXAMINADORA

Prof. D.Sc. Carlos Cristiano Hasenclever Borges - Orientador
Universidade Federal de Juiz de Fora

Prof. D.Sc. Saulo Moraes Villela - Coorientador
Universidade Federal de Juiz de Fora

Prof. D.Sc. Raul Fonseca Neto
Universidade Federal de Juiz de Fora

Prof. Ph.D. Antônio de Pádua Braga
Universidade Federal de Minas Gerais

A minha família e amigos, pelo apoio e paciência incondicionais.

AGRADECIMENTOS

A minha família, em especial ao meus pais Lincoln e Rita e ao meu irmão Lincoln Netto, pelo apoio incondicional oferecido ao longo dos anos e pela compreensão e suporte em todos os momentos em que não pude estar presente, seja no trabalho ou em casa.

Aos professores Saulo e Cristiano, pela orientação, paciência e apoio oferecido desde o princípio.

A Karen Enes, minha orientadora não registrada no papel, pela amizade, pelo apoio oferecido ao longo do curso e principalmente por me convencer de que fazer o mestrado realmente era uma boa ideia, foi uma correria, mas no final com certeza valeu a pena.

A todos os meu amigos, que de maneira direta ou indireta possibilitaram a realização deste trabalho, seja rodando algum teste, seja conferindo o texto ou seja oferecendo uma palavra de apoio em um momento de necessidade.

A todos os meus parentes, pelo encorajamento e apoio.

Aos professores do Departamento de Ciência da Computação pelos seus ensinamentos e aos funcionários do curso, que durante esses anos, contribuíram de algum modo para o meu enriquecimento pessoal e profissional.

"Não fiz o melhor, mas fiz tudo
para que o melhor fosse feito.
Não sou o que deveria ser, mas
não sou o que era antes".
Martin Luther King

RESUMO

O presente trabalho tem como proposta o desenvolvimento de uma abordagem capaz de melhorar os resultados obtidos por algoritmos de classificação quando aplicados em bases desbalanceadas. O método, denominado Algoritmo de Balanceamento Sintético Incremental (*Incremental Synthetic Balancing Algorithm* – ISBA), realiza um procedimento iterativo baseado em classificadores de larga margem, visando gerar amostras sintéticas com o intuito de reduzir o nível de desbalanceamento. No processo são utilizados vetores suporte como referência para a geração das novas instâncias, permitindo posicioná-las em regiões com uma maior representatividade. Além disso, a estratégia permite que as novas amostras ultrapassem os limites das amostras utilizadas como referência para sua geração, o que possibilita uma extrapolação dos limites da classe minoritária, objetivando, assim, alcançar um maior reconhecimento dessa classe de interesse. São apresentados experimentos comparativos com demais técnicas, entre elas o *Synthetic Minority Over-sampling Technique* (SMOTE), os quais fornecem fortes evidências da aplicabilidade da abordagem proposta.

Palavras-chave: aprendizado desbalanceado, classificadores de larga margem, reamostragem, geração de dados artificiais.

ABSTRACT

In this work we propose the development of an approach capable of improving the results obtained by classification algorithms when applied to unbalanced datasets. The method, called Incremental Synthetic Balancing Algorithm (ISBA), performs an iterative procedure based on large margin classifiers, aiming to generate synthetic samples in order to reduce the level of unbalance. In the process, we use the support vectors as reference for the generation of new instances, allowing them to be positioned in regions with greater representativeness. Furthermore, the strategy allows the new samples to exceed the limits of the samples used as reference for their generation, which allows an extrapolation of the limits of the minority class, in order to achieve greater recognition of this class of interest. We present comparative experiments with other techniques, among them the Synthetic Minority Over-sampling Technique (SMOTE), which provide strong evidence of the applicability of the proposed approach.

Keywords: unbalanced learning, large margin classifiers, oversampling, synthetic sample generation.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SÍMBOLOS

LISTA DE ABREVIATURAS E SIGLAS

FMP *Fixed Margin Perceptron*

IMA *Incremental Margin Algorithm*

ISBA *Incremental Synthetic Balancing Algorithm*

RBF *Radial Basis Function*

SMOTE *Synthetic Minority Over-sampling Technique*

SVM *Support Vector Machine*

OPB *One Pass Balancing*

CONTEÚDO

1 INTRODUÇÃO

Todos os dias uma infinidade de dados são gerados e armazenados em bases ao redor do globo e, com o desenvolvimento e aplicação de novas tecnologias nos diferentes ramos da indústria, pesquisa e negócios, a tendência é que este volume de dados cresça em proporções cada vez maiores. Este crescimento ocorre, não apenas pelo surgimento de novas bases, mas em grande parte pelo aumento no tamanho das bases de forma individual, tanto pela quantidade de registros, quanto pelo número de atributos presente em cada registro.

Com o aumento na extensão e complexidade das bases, a realização de análises e interpretações manuais torna-se extremamente lenta, custosa e subjetiva, surgindo a necessidade de abordagens mais automatizadas para o processo de extração de conhecimento, como por exemplo a utilização de técnicas de aprendizado de máquina.

Segundo Marsland (2015), o aprendizado pode ser definido como o aprimoramento na realização de determinada tarefa a partir da prática, portanto, por extensão, o aprendizado de máquina pode ser entendido como a utilização das experiências contidas nos dados para o melhoramento da performance dos algoritmos, sendo possível categorizá-lo em quatro grupos principais (MARSLAND, 2015):

- Aprendizado supervisionado: um modelo é construído a partir de um conjunto de amostras e de seus respectivos rótulos, buscando assim generalizar os dados iniciais e possibilitar que novas entradas sejam automaticamente rotuladas de maneira correta.

- Aprendizado não supervisionado: os dados iniciais não possuem rótulos preestabelecidos, portanto o algoritmo busca características em comum entre as amostras, agrupando-as por similaridade.

- Aprendizado por reforço: o algoritmo recebe informações sobre a validade do resultado atual, dando sequência a busca por um processo de tentativa e erro, continuamente experimentando novas possibilidades, até atingir a solução correta.

- Aprendizado evolucionário: algoritmo baseado na evolução biológica, onde a qualidade da solução atual pode ser medida por uma função de aptidão.

No contexto deste trabalho, o aprendizado ocorre de forma supervisionada, sendo a classificação o problema de interesse, o qual tem como objetivo obter, a partir dos dados fornecidos como entrada, limites de decisão capazes de separar as amostras em diferentes classes.

Ao longo das últimas décadas diversos algoritmos de classificação foram propostos e estudados, a citar: árvores de decisão (MURTHY, 1998), abordagens baseadas em *Perceptron* (ROSENBLATT, 1962) (RUMELHART et al., 1985), *Radial Basis Function* (RBF) (HOWLETT; JAIN, 2013), *Support Vector Machines* (SVMs) (CORTES; VAPNIK, 1995) e inúmeros outros. Estes algoritmos, apesar de apresentarem boa capacidade de predição em uma extensa gama de aplicações, muitas das vezes, acabam atingindo resultados sub-ótimos em decorrência de determinadas particularidades das bases sujeitas ao aprendizado supervisionado. Dentre estas particularidades, o desbalanceamento entre as classes é uma área de especial interesse, devido a sua influência sobre os resultados obtidos pelos classificadores (CHAWLA et al., 2004).

O desbalanceamento da base caracteriza-se pela concentração de amostras em um grupo reduzido de classes, enquanto as demais possuem poucos exemplos. Nestes casos, existe pelo menos uma classe composta por um pequeno conjunto de treinamento, denominado minoritário, ficando as demais classes concentradas em outro conjunto, denominado majoritário (GANGANWAR, 2012).

O treinamento de classificadores em bases desbalanceadas pode apresentar comportamento bastante distinto em relação a bases que não apresentam este padrão devido à forma como os classificadores são construídos. Estes normalmente consideraram uma distribuição razoavelmente igualitária entre as classes, portanto, implicitamente assumem que todos os erros possuem o mesmo peso. O problema desta suposição é que, quando ela não se mostra verdadeira, o resultado do algoritmo tende a ser guiado pela classe majoritária, podendo levar a situação extrema onde os erros apresentados pela classe minoritária são considerados como ruído (LIU et al., 2007).

Tomando apenas a acurácia geral como medida, a princípio apresenta-se como uma solução razoável o enfoque na predição correta da classe majoritária, entretanto é importante ressaltar que é particularmente importante identificar de forma precisa as amostras pertencentes a classe minoritária. Esta preferência é desejável devido a natureza da classe minoritária, a qual, geralmente, contém as amostras de principal interesse, pois estas re-

presentam os eventos anormais, que são justamente o foco do aprendizado (POURHABIB et al., 2015).

A obtenção de bons resultados em problemas envolvendo bases desbalanceadas mostra-se especialmente relevante devido a sua aplicação em problemas reais, onde a dificuldade na obtenção de amostras de determinada classe muitas vezes ocasiona uma desigualdade na distribuição dos padrões entre os grupos (CASTRO; BRAGA, 2011), a citar: (i) detecção de fraudes em transações de cartão de crédito, resgate de seguros e utilização de serviços de telecomunicação (CHAN; STOLFO, 1998) (PHUA et al., 2004); (ii) estudos em bioinformática (RADIVOJAC et al., 2004); (iii) diagnósticos médicos (SUN et al., 2007) (MAZUROWSKI et al., 2008); (iv) categorização de texto (DUMAIS et al., 1998); (v) identificação de vazamento de óleo através de satélite (KUBAT et al., 1998); e previsão de colisões entre aeronaves (EVERSON; FIELDSEND, 2006).

1.1 TRABALHOS RELACIONADOS

As principais abordagens empregadas na busca por uma solução para o problema de classificação em bases desbalanceadas podem ser divididas essencialmente em duas categorias:

- A primeira categoria tem como foco a realização de modificações no próprio classificador, tornando-o compatível com a distribuição dos dados, a citar: (i) algoritmos que utilizam apenas amostras da classe de interesse durante o treinamento, como por exemplo *one-class* SVMs (RASKUTTI; KOWALCZYK, 2004) (MANEVITZ; YOUSEF, 2007); (ii) algoritmos baseados em *ensemble*, onde cada classificador é construído a partir de uma sub-amostragem balanceada da base original (TIAN et al., 2011); e (iii) estratégias de modificação na função de custo do classificador, as quais buscam atingir um *trade-off* ideal entre acurácia geral e melhor reconhecimento da classe minoritária (CASTRO et al., 2011) (TAO et al., 2007).

- A segunda categoria efetua um pré-processamento dos dados, balanceando a base através de uma reamostragem dos exemplos originais, posteriormente repassando a base modificada para um algoritmo de classificação. Os modelos de reamostragem comumente utilizados são *oversampling* (LIU et al., 2007), o qual gera novos exemplos da classe minoritária com base no comportamento das amostras originais e *undersampling* (YEN; LEE, 2009), onde amostras da classe majoritária são

descartadas com base em algum parâmetro de eliminação.

Dentre os métodos baseados em reamostragem, os mais simples elegem aleatoriamente amostras para serem duplicadas ou removidas da base. Estas abordagens, apesar de conseguirem reduzir a diferença de cardinalidade entre as classes, podem resultar em bases que afetam negativamente o funcionamento dos algoritmos de classificação. A duplicação de amostras de forma aleatória tende a tornar as regiões de decisão do classificador menores e mais específicas, podendo ocasionar *overfitting*, enquanto a remoção aleatória de exemplos, por não conferir a relevância das amostras antes de seu descarte, pode resultar na perda de informações vitais ao aprendizado (HAN et al., 2005). Devido às limitações apresentadas por estas abordagens mais simplistas, diversos estudos vêm sendo desenvolvidos na área, buscando assim melhorar os resultados obtidos pelos classificadores, sendo de especial interesse para este trabalho as propostas baseadas na geração de amostras artificiais a partir dos exemplos preexistentes na base.

A geração de dados sintéticos consiste na generalização dos padrões contidos nos dados reais, de tal forma que eles possam ser extraídos e replicados nos dados artificias, produzindo assim um maior número de instâncias para a classe em questão, as quais, por não serem réplicas idênticas das amostras reais, tendem a tornar as regiões de decisão menos específicas, evitando *overfitting* e, consequentemente, melhorando o resultado do classificador obtido (ZHANG et al., 2015). O *Synthetic Minority Over-sampling Technique* (SMOTE), proposto por Chawla et al. (2002), é um dos métodos mais amplamente utilizados, funcionando em linhas gerais da seguinte forma: para cada exemplo da classe minoritária, novos exemplos artificiais são gerados entre os limites estabelecidos por ele e um número pré-definido de seus vizinhos mais próximos. Devido à sua importância no contexto deste estudo, uma descrição mais detalhada do funcionamento do algoritmo será apresentada na Seção 3.1.

A utilização do SMOTE ou de técnicas similares, onde todas as amostras da classe minoritária (ou uma subamostragem aleatória desta) são utilizadas na geração de exemplos artificiais, apesar de apresentarem bons resultados em uma diversa gama de aplicações, podem resultar na geração de amostras em áreas pouco representativas, considerando que nenhuma avaliação adicional é utilizada para verificar a relevância dos exemplos gerados. Na tentativa de minimizar este problema, algumas abordagens têm sido propostas, sendo apresentados alguns exemplos relacionados ao presente estudo.

Han et al. (2005) propuseram o *Borderline*-SMOTE. O trabalho parte do pressuposto que as zonas limítrofes entre as classes são mais sujeitas a apresentarem erros durante o treinamento do classificador, sendo, portanto, as amostras pertencentes a estas áreas mais relevantes para o processo de reamostragem. Com o intuito de definir estas amostras de interesse, o algoritmo realiza um análise da classe dos vizinhos mais próximos de cada instância e, dependendo da relação entre vizinhos pertencentes às classes majoritárias e minoritárias, a amostra é categorizada segundo uma escala de interesse. Por fim, apenas as amostras classificadas como de maior interesse são utilizadas na geração dos novos exemplos, buscando produzir amostras mais representativas.

Outras formas de selecionar amostras de interesse também podem ser encontradas na literatura, como por exemplo o ADASYN (HE et al., 2008), que estabelece pesos para cada amostra segundo uma distribuição de densidade, e o MWMOTE (BARUA et al., 2014) que, na mesma linha, define parâmetros de ponderação para cada instância levando em conta o nível de dificuldade no aprendizado da amostra e sua proximidade com relação a classe majoritária. Através destes parâmetros os algoritmos buscam priorizar a escolha de exemplos mais relevantes e que apresentem maior dificuldade durante o processo de aprendizado.

Em determinadas situações, a escolha de amostras situadas nas zonas limítrofes entre as classes pode não ser suficiente para gerar novos exemplos representativos, como nos casos onde a classe minoritária apresenta uma distribuição muito densa. Nestas circunstâncias as novas amostras tendem a ser muito similares às utilizadas como referencia para sua criação, o que pode ocasionar *overfitting* no processo de aprendizado, assim como ocorreria com a utilização do SMOTE. Buscando contornar este problema, foi proposto por Koto (2014) o algoritmo SMOTE-*Out*, este baseia-se na utilização de amostras da classe majoritária como referência para a extrapolação dos limites da classe minoritária. Apesar de se apresentar como uma estratégia visando permitir a geração de exemplos que expandem a faixa de distribuição da classe minoritária, o controle desta expansão dependente de instâncias da classe majoritária pode distorcer a real distribuição dos dados.

De qualquer forma, a geração de instâncias que possam expandir a região de distribuição principalmente da classe minoritária mostra-se como um potencial caminho para melhor predizer estes dados. Porém, técnicas mais efetivas no processo de geração de tais instâncias são cruciais visando manter a distribuição padrão dos dados. Desta forma, o

uso de referências mais confiáveis para tal geração pode trazer um maior confiabilidade neste processo.

1.2 MOTIVAÇÃO

Nos últimos anos, a área de aprendizado em bases desbalanceadas vem recebendo crescente atenção, tanto em estudos teóricos, quanto em aplicações práticas. O esforço empregado buscando solucionar os problemas encontrados neste tipo de base tem resultado no desenvolvimento de diversas técnicas, as quais tem conseguido gradualmente melhorar os resultados obtidos pelos classificadores, entretanto as soluções apresentadas ainda possuem limitações, principalmente no entendimento e controle dos procedimentos aplicados, e casos de falha relevantes, sendo ainda necessários maiores estudos.

Abordagens de *oversampling* baseadas na seleção de amostras de especial interesse para servir de referência durante a geração dos novos exemplos, conseguem atingir melhores resultados quando comparadas a técnicas mais simples, como SMOTE e *oversampling* aleatório, mas ainda podem resultar em distorções na distribuição dos dados na base. Outras técnicas também foram propostas, buscando aumentar a diversidade da base minoritária, reduzindo *overfitting*, porém nestes casos não é possível garantir a validade dos novos exemplos gerados, o que pode, por sua vez, ocasionar distorções durante o processo de aprendizado.

Outro aspecto importante a ser ressaltado é que a maioria dos trabalhos desenvolvidos na área de dados desbalanceados têm como enfoque o desenvolvimento de novos algoritmos, mas pouco tem sido feito, em nível teórico, no sentido de entender os reais motivos que levam estes algoritmos a afetarem a qualidade de predição dos classificadores, sendo também necessária um maior enfoque nesta linha de pesquisa.

Levando em conta a existência de diversas questões ainda em aberto no que se refere ao treinamento de classificadores em bases desbalanceadas, o presente trabalho busca apresentar uma alternativa de solução, propondo uma abordagem que visa tratar o problema de forma mais abrangente, possibilitando tanto uma redução na probabilidade de ocorrência de *overfitting*, quanto evitando a geração de pontos irrelevantes ou que gerem distorções na distribuição dos dados. O estudo também tem como finalidade não apenas a proposta de um novo algoritmo, mas um entendimento aprofundado de seu funcionamento e das características construtivas que possibilitaram seu desenvolvimento.

No contexto desse trabalho, pretende-se utilizar como referência para geração de amostras sintéticas algum padrão construtivo de um possível classificador aplicado na discriminação de tais dados. Logicamente, a geração dos dados sintéticos deixa de ser um procedimento somente de pré-processamento, passando a ser incorporado em níveis distintos ao processo indutivo em questão. Deve-se ressaltar que o mais relevante nesse trabalho não é apresentar modelos viáveis deste acoplamento geração/indutor, sendo logicamente esperado uma melhoria na predição quando aplicado este padrão de geração direcionada.

1.3 OBJETIVOS

Os objetivos desse trabalho podem ser divididos em dois aspectos principais:

- Realizar um levantamento dos aspectos construtivos relevantes ao processo de geração de amostras artificiais;
- Propor uma abordagem para a geração de dados artificiais baseada em classificadores de larga margem que seja eficiente e robusta no processo de balanceamento das classes em relação à qualidade da predição a ser obtida.

1.4 CONTRIBUIÇÕES

As principais contribuições deste trabalho são:

- Levantamento de aspectos relacionados ao problema de bases desbalanceadas que não são normalmente abordados pela literatura, a citar: (i) entendimento com maior nível de detalhamento do processo de geração de pontos, em especial com relação ao SMOTE quando aplicado a classificadores de larga margem; (ii) estudo do comportamento do hiperplano de larga margem após a geração de novas amostras; (iii) análise do número de pontos a serem gerados dependendo das características da base; e (iv) um levantamento das principais precauções necessárias ao utilizar técnicas de balanceamento, em especial com relação ao processo de validação dos algoritmos;
- Desenvolvimento do Algoritmo de Balanceamento Sintético Incremental (*Incremental Synthetic Balancing Algorithm* – ISBA);

- Levantamento de aspectos relevantes e das dificuldade relacionadas a geração de amostras artificiais no espaço de características;
- Proposta de extensão do ISBA para situações onde é interessante a utilização de funções de mapeamento *kernel*.

1.5 ORGANIZAÇÃO

Essa dissertação está organizada em sete capítulos. Após o capítulo introdutório, uma fundamentação teórica sobre classificadores de larga margem é apresentada no Capítulo 2, enquanto no Capítulo 3 são fornecidos detalhes de técnicas de balanceamento, juntamente com uma explicação detalhada do algoritmo SMOTE, o qual é de fundamental importância para o entendimento do presente trabalho. Após esta revisão da literatura, nos Capítulos 4 e 5 são discutidos e apresentados os resultados do ISBA, contendo portanto as principais contribuições do estudos. Em seguida, no Capítulo 6 é introduzido os conceitos referentes a geração de amostras artificiais no espaço de característica, assim como uma discussão sobre a proposta de extensão do ISBA para os casos onde são empregadas funções *kernel*. Por fim, no Capítulo 7, são apresentadas as conclusões e algumas possibilidades de trabalhos futuros.

2 APRENDIZADO SUPERVISIONADO

2.1 CLASSIFICAÇÃO BINÁRIA

Uma tarefa de classificação binária pode ser definida da seguinte forma: considere o conjunto Z dos dados de entrada de cardinalidade m, denominado conjunto de treinamento, composto de um conjunto de vetores x_i e de um conjunto de escalares y_i. Definindo $Z = \{(x_i, y_i) \ : \ i \in \{1, 2, ..., m\}\}$ e f como a função a ser inferida a partir de Z. Cada componente dos vetores de entrada está inserido em um espaço de dimensão d, i.e., $x_i \in \mathbb{R}^d$. Esses componentes são chamados de atributos ou características e podem admitir valores reais ou discretos, onde cada vetor de entrada é rotulado por um escalar y_i. Especificamente no problema de classificação binária, os valores de y_i são mapeados em um conjunto discreto de classes, i.e, $y_i \in \{-1, +1\}$. Dado um conjunto de treinamento, um algoritmo de aprendizado é capaz de gerar um classificador, definindo uma hipótese (representada por um hiperplano) em relação a função $f : \mathbb{R}^d \rightarrow \{-1, +1\}$. A partir do classificador gerado, novas amostras x podem ter ser valores y preditos, sendo estas amostras adicionais denominadas conjunto de teste.

2.2 PERCEPTRON

Rosenblatt (1958) propôs um procedimento para a atualização de um vetor de pesos, baseando-se em um elemento processador com múltiplas saídas. A medida de avaliação utilizada é uma comparação do valor da saída com os valores desejados, sendo o modelo resultante denominado Perceptron. Este modelo, consiste na arquitetura mais simples de uma Rede Neural Artificial e é considerado como o primeiro modelo de aprendizado supervisionado.

Estruturalmente, o Perceptron realiza um mapeamento de um espaço de entrada de dimensão d para um espaço de saída de dimensão m, associando cada unidade de entrada, x_i, a um componente de um vetor de pesos, w_i de dimensão d, e uma camada de saída formada por m unidades. Quando empregado em problemas de classificação binária, é suficiente a existência de somente um elemento processador na camada de saída. A estrutura do Perceptron é apresentada na Figura 2.1.

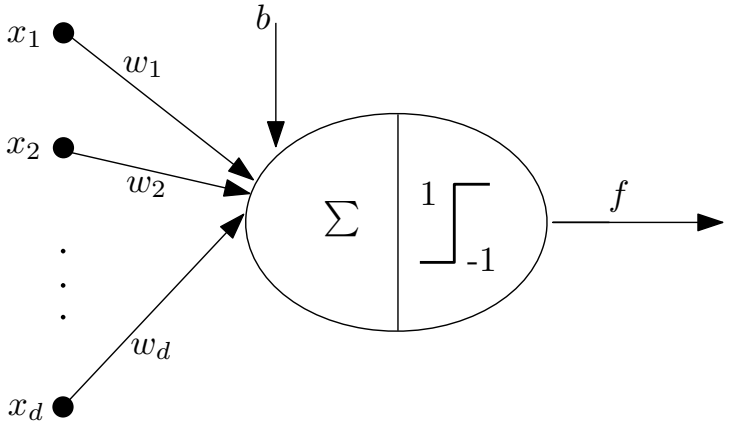


Figura 2.1: Topologia do Modelo Perceptron

O algoritmo desenvolvido por Rosenblatt pode ser utilizado para a determinação do vetor w em um número finito de iterações, desde que o conjunto de treinamento seja linearmente separável (NOVIKOFF, 1963). A quantidade de iterações está diretamente relacionada ao número de atualizações necessárias do vetor de pesos e, consequentemente, aos erros cometidos pelo algoritmo durante o treinamento. Dessa forma, quanto mais erros o algoritmo comete, mais atualizações são necessárias, bem como mais iterações do algoritmo. Neste caso, como o vetor de pesos w, que define o vetor normal ao hiperplano separador das classes, é determinado com base em sucessivas correções, tem-se o hiperplano separador construído de forma iterativa.

Com a limitação do modelo à classificação de dados linearmente separáveis, sua aplicabilidade mostra-se reduzida em casos onde os problemas são não linearmente separáveis. Por esta razão, foi proposto por Aizerman (1964) uma extensão do modelo, através da inserção de funções *kernel*, resultando na formulação dual do Perceptron. A formulação matemática do modelo Perceptron em variáveis primais e duais é detalhado a seguir.

2.2.1 FORMULAÇÃO PRIMAL

Matematicamente, o modelo Perceptron proposto por Rosenblatt consiste em encontrar um hiperplano separador dado pela solução do seguinte sistema de inequações lineares:

$$f(x_i) = \begin{cases} +1, & \langle w, x_i \rangle + b \geq 0 \\ -1, & \langle w, x_i \rangle + b < 0, \end{cases} \tag{2.1}$$

no qual w é o vetor de pesos e b o valor do viés (*bias*).

Uma amostra de treinamento (x_i, y_i) é uma instância incorretamente classificada se

$y_i(\langle w, x_i\rangle + b) < 0$. Enquanto uma amostra de treinamento for classificada incorretamente, a regra de correção deve ser aplicada até que não haja mais erros. A regra de correção é definida como:

$$\begin{aligned} w^{t+1} &\leftarrow w^t + \eta x_i y_i \\ b^{t+1} &\leftarrow b^t + \eta y_i, \end{aligned} \tag{2.2}$$

na qual η é a taxa de aprendizado e t a variável de iteração.

A resposta final do algoritmo de aprendizado é obtida quando o hiperplano solução é capaz de classificar todas as amostras de treinamento corretamente.

2.2.2 FORMULAÇÃO DUAL

Para modelar o Perceptron em termos de variáveis duais, torna-se necessário o mapeamento dos dados para um espaço de mais alta dimensão, denominado espaço de características, ou *ϕ-space*, comumente representado por F. Através do mapeamento $\phi : \mathbb{R}^d \times \mathbb{R}^d \rightarrow F$ é possível representar o conjunto de amostras não-linearmente separável em um espaço de mais alta dimensão, $x \rightarrow \phi(x)$, no qual o problema se torna linearmente separável. Essa nova modelagem permite a introdução de funções *kernel*, entretanto algumas modificações no modelo original do Perceptron são necessárias para viabilizar a construção de hipóteses lineares no espaço de variáveis duais, passando o vetor de pesos w a ser representado como um combinação linear dos vetores de entrada (x_i, y_i):

$$w = \sum_{j=1}^{m} \alpha_j y_j \phi(x_j), \tag{2.3}$$

na qual $\alpha \in \mathbb{R}^m$, $\alpha \geq \mathbf{0}$, é o vetor de multiplicadores ou variáveis duais associado ao conjunto de entrada.

Substituindo a expansão do vetor w, dada pela equação (2.3), na equação original de variáveis primais (2.1), a função f passa a ser definida da seguinte forma:

$$f(x_i) = \begin{cases} +1, & \sum_{j=1}^{m} \alpha_j y_j \langle \phi(x_i), \phi(x_j)\rangle + b \geq 0 \\ \\ -1, & \sum_{j=1}^{m} \alpha_j y_j \langle \phi(x_i), \phi(x_j)\rangle + b < 0. \end{cases} \tag{2.4}$$

A formulação dual do modelo Perceptron consiste em encontrar um hiperplano sepa-

rador, dado pela solução do seguinte sistema de inequações lineares:

$$f(x_i) = \begin{cases} +1, & \sum_{j=1}^{m} \alpha_j y_j K(x_j, x_i) + b \geq 0 \\ \\ -1, & \sum_{j=1}^{m} \alpha_j y_j K(x_j, x_i) + b < 0, \end{cases} \tag{2.5}$$

na qual b é o valor do viés (*bias*) e $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$.

Uma amostra de treinamento (x_i, y_i) é classificada incorretamente se:

$$y_i \left(\sum_{j=1}^{m} \alpha_j y_j K(x_j, x_i) + b \right) < 0. \tag{2.6}$$

Caso isso ocorra, a regra de correção do modelo deve ser aplicada até que não haja mais instâncias incorretamente classificadas. A regra de correção é então definida como:

$$\begin{aligned} \alpha_i^{t+1} &\leftarrow \alpha_i^t + \eta \cdot 1, \\ b^{t+1} &\leftarrow b^t + \Delta\alpha_i y_i, \end{aligned} \tag{2.7}$$

na qual $\Delta\alpha_i y_i$ refere-se ao somatório das correções nos valores dos multiplicadores considerando o respectivo sinal das classes, η é a taxa de aprendizagem e t a variável de iteração. O valor do viés (*bias*) pode ser computado separadamente em um esquema conforme o vetor de pesos. O hiperplano separador resultante é definido ao obter uma solução onde todas as amostras de treinamento são corretamente classificadas.

Uma grande diversidade de funções *kernel* já foram propostas pela literatura, entretanto, no contexto deste trabalho, três formulações são de especial interesse:

$$\text{Linear: } K(x_i, x_j) = \langle x_i, x_j \rangle + c \tag{2.8}$$

$$\text{Polinomial: } K(x_i, x_j) = (\langle x_i, x_j \rangle + 1)^d \tag{2.9}$$

$$\text{Gaussiano: } K(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2} \right), \tag{2.10}$$

onde c é uma constante, d o grau do polinômio e σ a largura da gaussiana.

Para o caso do *kernel* gaussiano, pode-se adotar um parâmetro $\gamma = 1/2\sigma^2$, obtendo:

$$K\left(x_i, x_j\right) = exp\left(-\gamma||x_i - x_j||_2^2\right). \tag{2.11}$$

2.3 MÁQUINA DE VETORES SUPORTES – SVM

A formulação original do Perceptron de Rosenblatt estabelece como critério de parada a correta classificação de todas as amostras do conjunto de treinamento, entretanto, para um mesmo problema, é possível obter diversas soluções que atendam a esta restrição. Como soluções diferentes podem apresentar diferentes qualidades de predição, outras abordagens mais restritivas foram propostas pela literatura, buscando assim obter o hiperplano que melhor generaliza o comportamento da base. Entre estas propostas, a Máquina de Vetores Suportes (*Support Vector Machine* – SMV) (BOSER et al., 1992) (CORTES; VAPNIK, 1995) destaca-se como um dos algoritmos mais utilizados no contexto de aprendizado de máquina, apresentando, de maneira geral, desempenho superior quando comparado aos demais métodos disponíveis na literatura (MARSLAND, 2014).

SVMs são classificadores de máxima margem que visam separar o conjunto de treinamento através do hiperplano que maximiza a distância entre as classes opostas, sendo este hiperplano obtido através da solução de um problema de otimização. Adotando $Z^+ = \{(x_i, y_i) \in Z : y_i = +1\}$ e $Z^- = \{(x_i, y_i) \in Z : y_i = -1\}$, o problema de classificação binária consiste em identificar um hiperplano, dado pelo seu vetor normal $w \in \mathbb{R}^d$ e pela constante $b \in \mathbb{R}$, de tal forma que o hiperplano separe os conjuntos Z^+ e Z^-. Com base nesta formulação também é possível introduzir o conceito de margem, que consiste em definir uma distância mínima $\gamma \geq 0$ entre o hiperplano separador e os conjuntos Z^+ e Z^-. Assim, o problema resume-se a definir (w, b) tal que:

$$y_i\left(\langle w, x_i\rangle + b\right) \geq \gamma, \quad \forall\left(x_i, y_i\right) \in Z \tag{2.12}$$

O SVM classifica as amostras através de um processo de otimização que identifica o hiperplano com a máxima margem entre as duas classes presentes no conjunto de treina-

mento, podendo este problema de otimização ser formulado como:

$$\begin{array}{ll} \min\limits_{w,b,\xi} & \frac{1}{2}w^T w + C\sum_i \xi_i \\ \text{sujeito a} & y_i\left(w^T\phi\left(x_i\right)+b\right) \geq 1-\xi_i,\ \xi_i \geq 0, \quad \forall i \in \{1,2,\ldots,m\} \end{array} \tag{2.13}$$

onde $\phi\left(.\right)$ é uma função de mapeamento para um espaço de características, ξ_i são as variáveis de folga e C é uma constante de penalização utilizadas no caso margem flexível, sendo os vetores suporte as amostras mais próximas do hiperplano separador.

Objetivando facilitar a solução desse problema, é conveniente relaxar as restrições das inequações por meio da introdução de um conjunto de multiplicadores de Lagrange não-negativos α_i, com $i \in \{1,2,\ldots,m\}$. Ao incorporar as restrições relaxadas, a função Langrangeana é dada por:

$$L\left(w,b,\alpha\right) = \frac{1}{2}w^T w - \sum_i \alpha_i y_i\left(\langle w, x_i\rangle + b\right) + \sum_i \alpha_i \tag{2.14}$$

Essa função deve ser minimizada em relação a w e b e maximizada em relação a α, sujeito a $\alpha_i \leq \mathbf{0}$ para todo $i \in \{1,2,...,m\}$. Essa solução pode ser obtida através da maximização da funções estritamente dual, na qual os parâmetros w e b são substituídos. Dessa forma, é possível obter a formulação dual do SVM estritamente em função dos valores dos multiplicadores α, como segue:

$$\begin{array}{l} \max\ L(\alpha) = \sum\limits_i \alpha_i - \frac{1}{2}\sum\limits_i\sum\limits_j \alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle \\ \text{sujeito a} \begin{cases} \sum_i \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq 0 \end{cases} \end{array} \tag{2.15}$$

Ao solucionar esse problema, os valores ótimos α^* são obtidos. Assim, é possível reconstruir o vetor normal w para essa solução da seguinte forma:

$$w^* = \sum_i \alpha_i^* y_i x_i \tag{2.16}$$

2.4 ALGORITMO DE MARGEM INCREMENTAL – IMA

Essa seção descreve o funcionamento do Algoritmo de Margem Incremental (*Incremental Margin Algorithm* – IMA), proposto por Leite and Fonseca Neto (2008). Trata-se de um algoritmo de aprendizado de larga margem incremental baseado em uma extensão do algoritmo Perceptron, o Perceptron de Margem Fixa (*Fixed Margin Perceptron* – FMP).

O FMP encontra a solução de um problema linearmente separável dada uma margem γ_f. O algoritmo é capaz de determinar um hiperplano separador, caso exista, com o valor da margem fixa pré-definido, tal que:

$$y_i \left(\langle w, x_i \rangle + b \right) \geq \gamma_f ||w||_2, \quad \forall i \in \{1, 2, \ldots, m\} . \tag{2.17}$$

Para a implementação do IMA, Leite and Fonseca Neto (2008) propuseram uma nova formulação para o problema de maximização de margem baseada em duas premissas básicas: (i) os pontos ou vetores suporte de classes contrárias se encontram à mesma distância do hiperplano separador, isto é $\gamma^+ = \gamma^-$ (equação 2.18); (ii) soluções de larga margem podem ser obtidas, em um número finito de correções, através do FMP.

$$\begin{aligned} \gamma^+ &= \min \; y_i \left(\langle w, x_i \rangle + b \right), \quad \forall x_i \in Z^+ \\ \gamma^- &= \min \; y_i \left(\langle w, x_i \rangle + b \right), \quad \forall x_i \in Z^- \end{aligned} \tag{2.18}$$

Assim, a nova formulação para o problema de maximização de margem consiste na maximização direta da margem geométrica definida pelo FMP, a partir da solução do seguinte problema de otimização:

$$\begin{gathered} \max_w \gamma_g \\ \text{Sujeito a} \\ y_i \left(\langle w, x_i \rangle + b \right) \geq \gamma_g ||w||_2 \end{gathered} \tag{2.19}$$

O algoritmo computa diretamente o valor da maior margem geométrica, a qual se aproxima suficientemente da margem ótima do problema, garantindo a construção de um classificador de larga margem.

A técnica de solução do IMA consiste então em uma estratégia de aprendizado incremental, através da qual são obtidas sucessivas soluções do FMP para valores crescentes de margem, que se aproxima suficientemente da margem ótima do problema, garantindo a

construção de um classificador de larga margem. Iterativamente, soluciona-se o problema de inequações não lineares da equação 2.20, sendo cada solução equivalente à solução do FMP:

$$y_i \left(\langle w, x_i \rangle + b \right) \geq \gamma_f ||w||_2, \quad i \in \{1, \ldots, m\} . \tag{2.20}$$

Villela et al. (2016) estenderam o algoritmo FMP para possibilitar, além do uso de um valor de norma diferente de 2 (podendo variar de 1 a ∞), permitir a flexibilização da margem, tornando possível a solução de problemas não linearmente separáveis. Os algoritmos resultantes dessa extensão foram denominados de Perceptron de Margem Fixa com norma p (*Fixed p-Margin Perceptron* – FMP_p) e Algoritmo de Margem Incremental com norma p (*Incremental p-Margin Algorithm* – IMA_p). Para a flexibilização da margem, os autores introduziram variáveis de folga, da seguinte forma:

$$y_i \left(\langle w, x_i \rangle + b \right) \geq \gamma_f ||w||_2 - \alpha_i \lambda, \quad i \in \{1, \ldots, m\} , \tag{2.21}$$

sendo α o vetor de multiplicadores e λ um parâmetro de controle.

Assim, sempre que ocorrer um erro na amostra x_i, o vetor de multiplicadores deve ser escalonado na forma:

$$\alpha^{t+1} \leftarrow \alpha^t \left(1 - \eta \gamma_f / ||w||_2 \right) , \tag{2.22}$$

sendo seguida de uma correção no valor do multiplicador associado à amostra:

$$\alpha_i = \alpha_i + \eta \cdot 1. \tag{2.23}$$

Para a atualização, a cada iteração do IMA, do valor da margem fixa, duas regras de correção podem ser adotadas: (i) caso a solução do problema forneça margens negativa e positiva diferentes, corrige-se o valor da margem fixa na forma da equação 2.24, onde γ^+ e γ^- são os valores relacionados, respectivamente, às menores distâncias projetadas dos pontos do conjunto Z^+ e Z^+ ao hiperplano separador da t-ésima iteração; (ii) caso a solução do problema forneça margens negativa e positiva iguais, torna-se necessário garantir um acréscimo no valor da nova margem fixa de acordo com a equação 2.25, onde

Δ é uma constante de incremento positiva.

$$\gamma_f^{t+1} = \frac{\gamma^+ + \gamma^-}{2}, \tag{2.24}$$

$$\gamma_f^{t+1} = \gamma_f^t + \max\left\{\Delta, \frac{\gamma^+ + \gamma^-}{2} - \gamma_f^t\right\}, \tag{2.25}$$

O critério de parada do algoritmo é baseado na definição de um número máximo de iterações do algoritmo de treinamento. Assim, caso não haja uma nova solução do problema FMP, adota-se como margem obtida o valor anterior de margem fixa, relacionado à última solução encontrada. A margem geométrica final estabelecida pelo processo incremental pode ser definida como a margem de parada do algoritmo FMP na última iteração.

3 TÉCNICAS DE BALANCEAMENTO

Esse capítulo é destinado a descrição do processo que resultou no desenvolvimento do Algoritmo de Balanceamento Sintético Incremental. São apresentadas as características das técnicas que precederam a abordagem proposta, servindo assim, tanto de referência, quanto de motivação para o presente trabalho. Inicialmente é apresentada uma estratégia conhecida como SMOTE, considerada bastante efetiva na geração de dados sintéticos, sendo, inclusive, objeto de diversas adaptações e usualmente tomada como referência em estudos na área. Além disso, tem-se a motivação adicional de tomar como base a técnica de posicionamento das instâncias sintéticas usada no SMOTE para a construção do modelo a ser desenvolvido.

3.1 SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE – SMOTE

Dentre os algoritmos de reamostragem baseados na geração de novas instâncias, o SMOTE (CHAWLA et al., 2002) destaca-se por sua ampla utilização e pelos bons resultados obtidos. Esta técnica consiste na geração de exemplos artificiais da classe minoritária através da interpolação das amostras pré-existentes, sendo aplicada, em linhas gerais, da seguinte forma: para cada amostra da classe minoritária é escolhido aleatoriamente um dos k vizinhos mais próximos da mesma classe, determinados por distância euclidiana. Em seguida, para cada dimensão da instância, é gerado um valor aleatório entre a amostra de origem e o vizinho escolhido, sendo estes valores utilizados na construção do novo exemplo artificial. O pseudocódigo apresentado nos Algoritmos 1 e 2 ilustram o processo de geração destas instâncias. Dependendo do nível de desbalanceamento e, consequentemente, do número de amostras a serem geradas, cada amostra da classe minoritária pode ser utilizada para gerar diversos exemplos artificiais, geralmente usando vizinhos distintos no processo construtivo.

Um aspecto importante a ser ressaltado é que as amostras artificiais são geradas através de um processo independente para cada atributo, o qual percorre as dimensões da base sorteando valores diferentes para a variável *gap* a cada novo passo. Este procedimento estabelece limites geométricos para a geração das novas amostras que podem, em alguns casos, extrapolar o raio da base original (raio da menor hiperesfera capaz de circunscrever

Algoritmo 1: Geração de amostra artificial

Entrada: amostra: A;
vizinho escolhido: B;
Saída: amostra artificial: G;
início
  $d \leftarrow$ dimensão do problema;
  para i **de** 1 **até** d **faça**
    $dif \leftarrow B_i - A_i$;
    $gap \leftarrow \text{rand}[0, 1]$;
    $G_i \leftarrow A_i + gap \times dif$;
  fim para
fim

a base). Isto ocorre pois, ao contrário do que pode parecer a princípio, a nova amostra não é gerada necessariamente no segmento de reta que liga as duas amostras de referência e sim dentro do hipercubo que tem como arestas a diferença entre os atributos da amostra de referência e seu vizinho escolhido. A Figura 3.1 ilustra a diferença entre as duas situações, questão muitas vezes não muito clara na literatura, apesar da definição correta do algoritmo, podendo levar a erros de implementação.

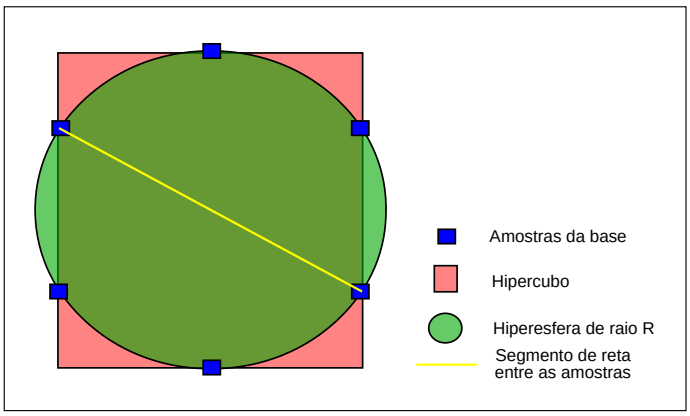


Figura 3.1: Representação dos limites SMOTE. Áreas em vermelho representam as regiões onde o SMOTE pode extrapolar os limites do raio original da base.

3.2 APRENDIZADO EM BASES ARTIFICIALMENTE BALANCEADAS

Uma das alternativas de interesse na área de aprendizado em bases desbalanceadas têm como proposta a geração de dados artificiais, sendo o SMOTE um dos principais representante desta classe de soluções. Apesar dos bons resultados obtidos pelo algoritmo em

Algoritmo 2: Synthetic Minority Over-sampling Technique

Entrada: número de amostra classe minoritária: T;
porcentagem das amostras utilizadas no balanceamento: N;
número de vizinho mais próximos: K;
Saída: conjunto de amostras artificiais geradas: $ConjG$;
início
  se $N < 100$ **então**
    seleciona aleatoriamente apenas uma porcentagem $N\%$ da base para ser utilizada durante o processo;
  fim se
  $N \leftarrow N/100$;
  para i **de** *1* **até** N **faça**
    para j **de** *1* **até** T **faça**
      $A \leftarrow$ amostra na posição j da classe minoritária;
      $B \leftarrow VizinhoMaisProximo(A, K)$;
      $ConjG \leftarrow$ *Geração de amostra artificial*(A, B);
    fim para
  fim para
fim
$//VizinhoMaisProximo()$ retorna aleatoriamente um dos K vizinhos mais próximos da amostra passada como parâmetro

diversos ensaios, existem situações onde as amostras geradas podem levar o classificador a produzir hipóteses pouco efetivas. Este comportamento indesejado ocorre, normalmente, quando os exemplos gerados apresentam um dos seguintes padrões: (i) possuem pouca variabilidade com relação a base original; (ii) estão localizados em regiões de pouca relevância; (iii) ou encontram-se em regiões pertencentes a classe majoritária.

3.2.1 GERAÇÃO DE AMOSTRAS COM POUCA VARIAÇÃO

A principal vantagem do SMOTE em relação a outras técnicas de balanceamento é a sua capacidade de gerar amostras que produzem um aumento na variabilidade da base, o que tende a evitar a ocorrência de *overfitting*. Isto ocorre pois, ao distribuir as novas amostras de forma mais uniforme pelo espaço, o algoritmo reduz a probabilidade do classificador inferir hipóteses muito especificas, as quais, apesar de classificarem corretamente os exemplos de treinamento, tendem a errar novas instâncias, mesmo que estas apresentem um comportamento similar ao desejado. Este conceito mostra-se válido tanto para a utilização de margem flexível, quando para margem rígida, porém sua ocorrência é mais acentuada e, mais facilmente observada em situações onde a margem rígida é adotada, levando em

conta que, nestes casos, o hiperplano tem que discriminar corretamente ambas as classes sem permitir nenhum erro durante o aprendizado, o que tende a resultar em hipóteses mais específicas. As Figuras 3.2 e 3.3 exemplificam a ocorrência deste aumento de variabilidade em uma base de duas dimensões, onde é possível observar um maior poder de generalização quando as amostras artificiais são consideradas.

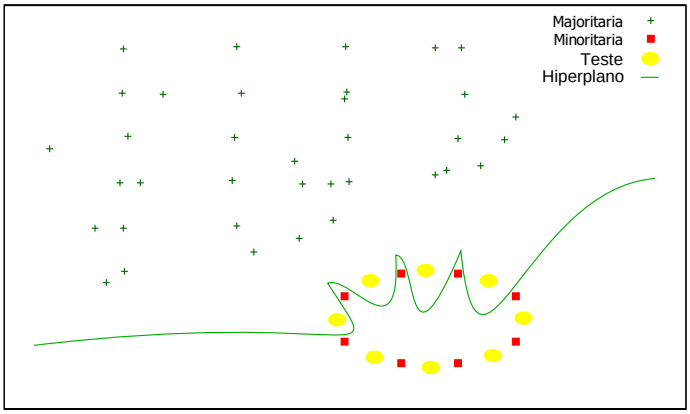


Figura 3.2: Hiperplano inferido utilizando apenas a base original. Neste caso, é possível observar que, mesmo em situações onde o hiperplano é capaz de classificar corretamente todas as amostras de treinamento, a ocorrência de *overfitting* pode levar a uma baixa taxa de acerto em novos exemplos, mesmo que estes apresentem o padrão desejado.

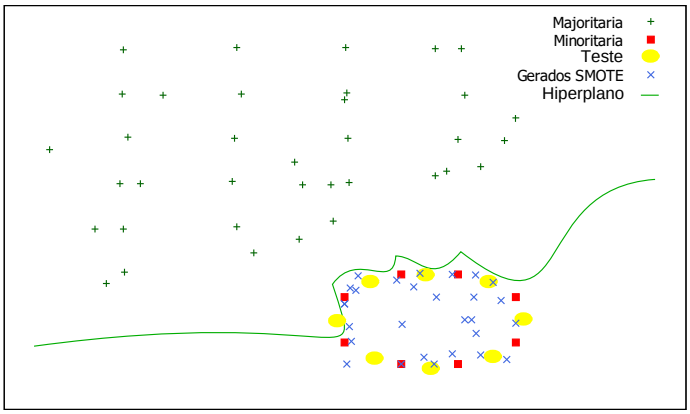


Figura 3.3: Hiperplano inferido utilizando a base original em conjunto com amostras geradas pelo SMOTE. Neste caso, as amostras artificiais evitam a construção de um hiperplano sobreajustado, o que possibilita uma classificação mais correta das amostras de teste.

Conforme apontado anteriormente, a utilização do SMOTE em situações onde as amostras são distribuídas de forma mais compacta pode resultar em exemplos artificiais muito similares as amostras de referência utilizadas em sua geração. Nestes casos, mesmo após o balanceamento, o processo de aprendizado pode continuar inferindo hipoteses com pouco

poder de generalização. As Figuras 3.4 e 3.5 apresentam um exemplo de como o SMOTE, em alguns casos, pode produzir amostras sem representatividade, resultando em uma base de treinamento incapaz de alterar o posicionamento do hiperplano de forma significativa. Apesar do exemplo ter sido construído baseando-se no funcionamento do SMOTE, o comportamento apresentado pode ser estendido a outros algoritmos que adotam a região entre os pontos de referência como limites para a geração de amostras artificiais, como o *Borderline*-SMOTE, MWMOTE e o ADASYN.

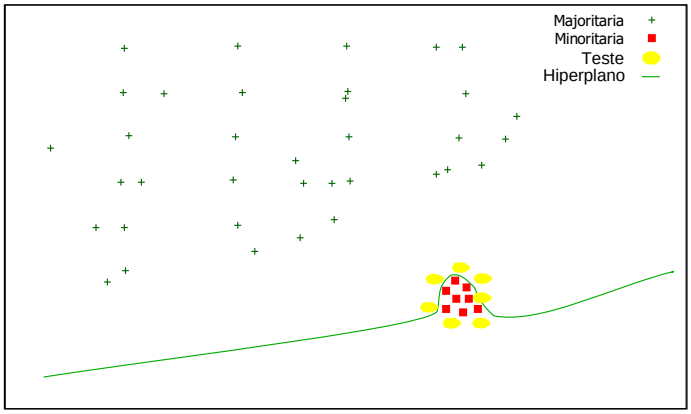


Figura 3.4: Hiperplano inferido utilizando apenas base original.

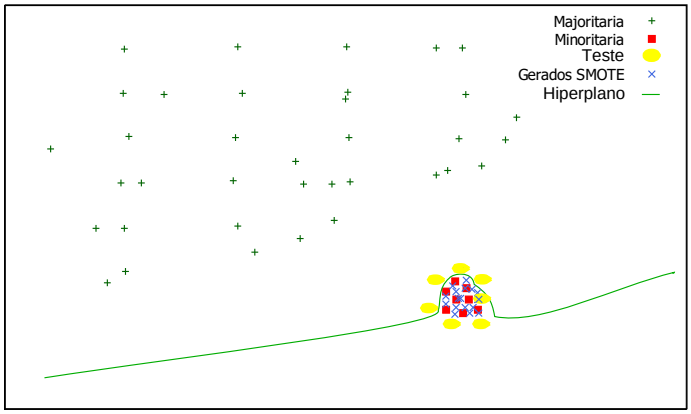


Figura 3.5: Hiperplano inferido utilizando base original em conjunto com amostras geradas pelo SMOTE. É possível observar que, mesmo com a utilização do SMOTE, o classificador continua gerando uma hipótese muito específica, resultando em uma baixa taxa de acerto nos pontos de teste.

3.2.2 GERAÇÃO DE AMOSTRAS COM POUCA RELEVÂNCIA

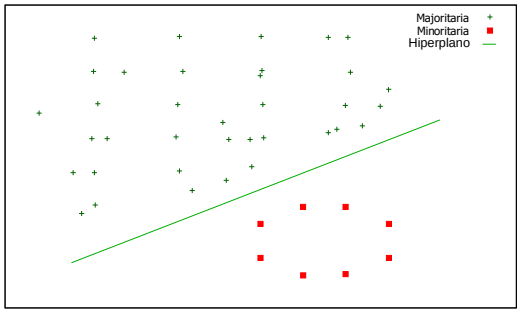
O SMOTE não define nenhum parâmetro para medir a relevância das instâncias pertencentes a base original, considerando todas como igualmente importantes e, portanto,

selecionando de forma aleatória combinações de vizinhos mais próximos para serem utilizados como referência para a geração dos novos exemplos. Esta abordagem não leva em conta que amostras localizadas nas zonas de decisão do problema apresentam maiores chances de serem classificadas incorretamente, sendo normalmente priorizadas durante o processo de aprendizado. Como resultado desta priorização, boa parte das amostras geradas pelo SMOTE acabam tendo pouca interferência na construção do classificador, podendo, muitas das vezes, serem até totalmente desconsideras, como nos casos baseados em vetores suporte.

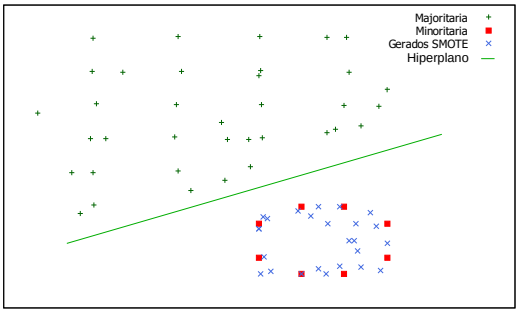
As Figuras 3.6 e 3.7 apresentam alguns exemplos de soluções SVM geradas a partir de bases com diferentes níveis de correção do desbalanceamento. Em 3.6(a) e 3.7(a) a base original não foi modificada, em 3.6(b) e 3.7(b) a base foi totalmente balanceada pelo SMOTE e em 3.6(c) e 3.7(c) apenas alguns exemplos artificiais foram gerados utilizando como referência amostras de interesse previamente selecionadas. A análise dos hiperplanos nas diferentes figuras permite observar que a utilização do SMOTE é capaz de promover consideráveis mudanças no hiperplano, entretanto soluções similares também podem ser atingidas através da utilização de um subgrupo de amostras artificiais, desde que estas sejam geradas em regiões que sejam relevantes para o posicionamento do hiperplano, fornecendo evidencias da falta de representatividade de um grande parcela dos exemplos artificiais produzidos pelo SMOTE.

Alguns trabalhos têm sido propostos pela literatura com o intuito de produzir exemplos artificiais com maior relevância através de uma seleção mais assertiva das amostras de referência. Em geral as abordagens baseiam-se na atribuição de pesos às instâncias da base original, sendo as amostras com um maior peso consideradas mais relevantes e, portanto, utilizadas com maior frequência como referência para a geração dos novos exemplos. Apesar destes algoritmos diferirem do SMOTE quanto a seleção das instâncias de referência, após esta etapa, a geração das amostras artificiais funciona de maneira similar à proposta pelos SMOTE, o que pode continuar resultando na criação de exemplos com pouca representatividade, pois estes ainda encontram-se limitados pelas amostras de referência.

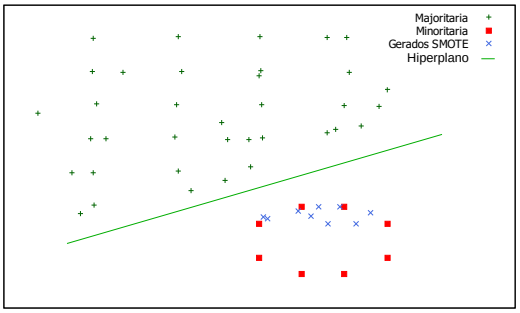
As amostras artificias geradas pelo SMOTE encontram-se sempre entre os limites estabelecidos por cada uma das características pertencentes as instâncias utilizadas como referência para sua criação. Esta abordagem faz com que sejam raros os casos onde as



(a) Base original.



(b) Base original mais amostras artificiais.

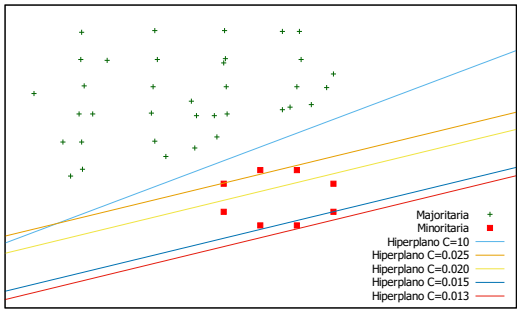


(c) Base original e subgrupo de amostras artificiais geradas a partir de instâncias consideradas relevantes

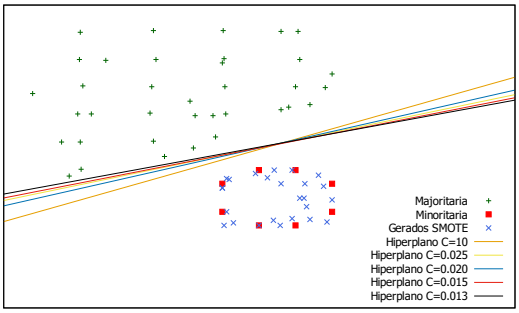
Figura 3.6: Soluções SVM Margem Rígida. Neste caso o SMOTE modifica levemente o hiperplano em relação a base original, entretanto o mesmo posicionamento também pode ser obtido a partir de um conjunto reduzido de amostras artificiais, desde que estas estejam localizadas em áreas de relevância.

novas amostras ultrapassem as extremidades da classe minoritária. Apesar de garantir certa robustez, este comportamento pode não ser tão adequado, visto que, desde que corretamente direcionada, uma extrapolação destas extremidades poderia levar a um reposicionamento do hiperplano separador, melhorando assim seu poder de generalização em relação a classe minoritária. A Figura 3.8 ilustra alguns destes conceitos em uma base de duas dimensões.

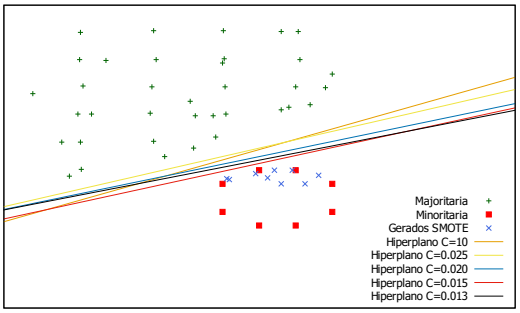
Conforme análise dos exemplos sugere, a utilização da abordagem adotada pelo SMOTE para a geração de novas amostras pode resultar em modificações irrelevantes na posição do hiperplano obtido em comparação ao gerado através da base original. Isso ocorre devido aos vetores suporte raramente serem modificados em relação aos vetores obtidos com a base original, considerando que as novas instâncias dificilmente são gerados fora das extremidades da classe minoritária (área vermelha 3.8). Como a construção do classificador é baseada nos vetores suporte, os hiperplanos gerados com e sem as amostras artificiais permanecem similares, com um exemplo deste comportamento sendo visto na Figura 3.6.



(a) Base original.



(b) Base original mais amostras artificiais.



(c) Base original e subgrupo de amostras artificiais geradas a partir de instâncias consideradas relevantes.

Figura 3.7: Soluções SVM Margem Flexível. Neste caso o SMOTE é capaz de evitar que a classe minoritária seja considerada como ruído durante o aprendizado, entretanto um resultado similar também pode ser atingido a partir de um conjunto reduzido de amostras artificiais, desde que estas estejam localizadas em áreas de relevância.

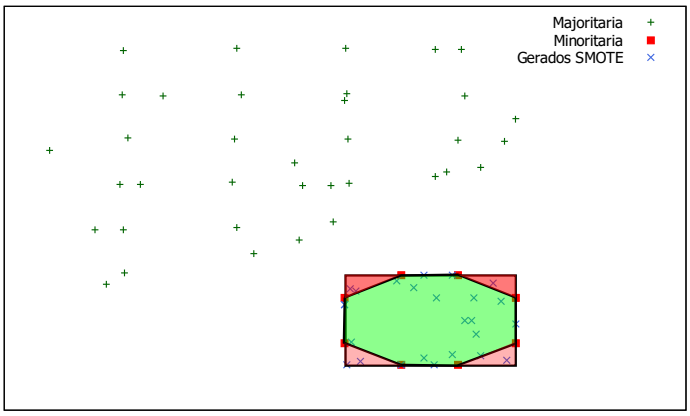


Figura 3.8: SMOTE base 2D. O retângulo externo representa os limites estabelecidos pelas características da base original, contendo o espaço onde o SMOTE pode gerar novas amostras artificiais; o polígono em verde representa as extremidades da classe minoritária; e os triângulos em vermelho representam as áreas onde o SMOTE é capaz de extrapolar as extremidades da classe minoritária

Outro aspecto importante é que, quando uma suavização da margem é permitida, o SMOTE é capaz de evitar que a classe minoritária seja tratada como ruído, garantindo

assim uma melhora significativa dos resultados. Entretanto, sua utilização dificilmente possibilita uma maior movimentação do hiperplano em direção à classe majoritária (vide 3.7), o que seria desejável na tentativa de priorizar a classe de maior interesse. É válido ressaltar que, no exemplo fornecido, uma solução simples para atingir um posicionamento desejável seria uma melhor parametrização da constante de flexibilização C, porém o mesmo não ocorre em bases onde a separação não é tão bem definida, sendo então de vital importância a utilização de métodos capazes de reposicionar o hiperplano baseando-se em outro procedimento mais eficiente.

Conforme apontado anteriormente, mesmo os algoritmos que apresentam propostas que diferem do SMOTE no que tange a seleção das amostras de referência, normalmente utilizam abordagens parecidas para a geração dos novos exemplos. Desta forma, apesar dos exemplos deste capítulo terem sido construídos através da utilização do SMOTE, seus conceitos também podem ser estendidos a estas outras técnicas, a citar: *Borderline*-SMOTE, MWMOTE e ADASYN.

3.2.3 GERAÇÃO DE AMOSTRAS INCORRETAS

A literatura contém poucos trabalhos voltados a geração de amostras artificiais fora dos limites estabelecidos pelas amostras da classe minoritária, entretanto alguns estudos podem ser mencionados, como o SMOTE-*Out* e uma variação do *Borderline*-SMOTE. Em ambos os casos a abordagem é baseada na seleção de um ponto da classe minoritária e de um de seus vizinhos mais próximos na classe majoritária, sendo o restante do processo similar ao SMOTE. O objetivo destes algoritmos é possibilitar um reposicionamento do hiperplano, buscando assim atingir uma melhor generalização da base, porém, a simples extrapolação dos limites da classe minoritária, sem a realização de nenhum tipo de verificação, pode levar a geração de amostras que não apresentem o padrão desejado. Nestas circunstâncias, as novas instâncias podem ocasionar uma descaracterização do comportamento da base, resultando na produção de um classificador sem nenhuma representatividade. As técnicas acima mencionadas buscam minimizar este risco limitando o valor do parâmetro randômico do SMOTE (variável *gap* do Algoritmo 1), mas só isso não é suficiente para garantir a efetividade das amostras, sendo assim considerado como importante neste trabalho, a proposição de estratégias mais eficazes, principalmente para avaliar um possível descarte de instâncias geradas e consideradas potencialmente distorcidas.

4 ALGORITMO DE BALANCEAMENTO SINTÉTICO INCREMENTAL – ISBA

O presente trabalho propõe uma nova abordagem para a geração de amostras artificiais, buscando produzir uma solução robusta o suficiente para tratar ao mesmo tempo todos os desafios mencionados nesse capítulo. Para tal, a proposta apresentada baseia-se essencialmente em dois princípios: possibilidade de extrapolação dos limites da classe minoritária e utilização de informações provenientes de uma aplicação prévia de classificadores de larga margem.

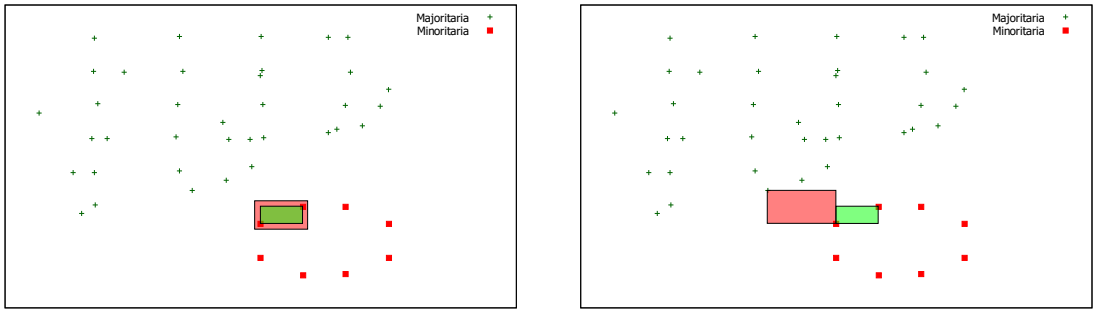
Em relação ao processo de extrapolação, propõe-se a introdução de um novo parâmetro de entrada, denotado por τ, que representa um porcentagem de possibilidade de extrapolação da diferença entre os atributos em questão. Esta modificação tem como objetivo redefinir os limites impostos pelo SMOTE na geração dos dados artificiais, consistindo basicamente na alteração da equação de cálculo da variável *gap* contida no Algoritmo 1, a qual passa a ser reescrita da seguinte forma:

$$\text{gap} \leftarrow \text{rand}[(0-\tau),(1+\tau)]$$

Com a inserção deste parâmetro, as novas amostras, não mais encontram-se delimitadas pela diferença entre os atributos das instâncias que lhe deram origem, tornando possível uma expansão da região definida pelo hipercubo previamente descrito. Devido a esta alteração na construção das amostras artificiais, o processo de aprendizado tende a ser positivamente afetado. Em primeiro lugar, a possibilidade de extrapolação permite um aumento na variação da classe minoritária, o que reduz a probabilidade de ocorrência de *overfitting*, diminuindo a questão apresentada na seção 3.2.1. Além desta vantagem, a expansão da classe minoritária também aumenta as chances de que as novas amostras sejam geradas em regiões mais próximas a classe majoritária, provocando um reposicionamento do hiperplano separador e, consequentemente, aumentando o poder de generalização do classificador em relação as amostras de interesse, assunto este abordado na seção 3.2.2.

É importante ressaltar que existem outras técnicas de reamostragem baseadas na extrapolação da classe minoritária, entretanto este trabalho difere das demais técnicas com

relação ao procedimento adotado. As abordagens encontradas na literatura utilizam uma amostra da classe majoritária como referência para a geração do novo exemplo, enquanto o algoritmo proposto limita sua seleção às amostras pertencentes a classe de interesse, porém permitindo que o novo exemplo ultrapasse os limites das instâncias de referência em uma determinada porcentagem τ. Esta diferença de abordagem tem como objetivo obter uma expansão da classe minoritária de forma mais natural, evitando uma descaracterização do seu padrão de distribuição. A Figura 4.1 ilustra a diferença entre as duas abordagens.

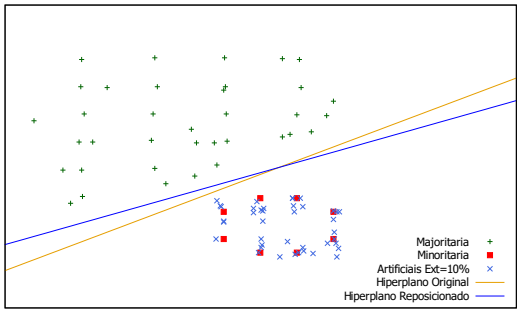


(a) Abordagem proposta. (b) Demais abordagens da literatura.

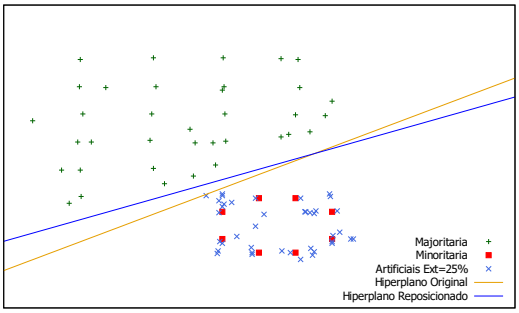
Figura 4.1: A área em verde representa a região onde as amostras artificiais poderiam ser geradas pelo SMOTE, enquanto as áreas em vermelho representam as regiões de extrapolação proposta pelos algoritmos.

Outro aspecto a ser observado é que a correta parametrização de τ é de fundamental importância para um bom funcionamento do algoritmo, visto que em casos extremos sua utilização também pode resultar na geração de amostras que não possuem o mesmo comportamento da classe minoritária real. Durante os experimentos o parâmetro τ foi calibrado através da análise do número de amostras descartadas durante o processo de geração, conceito que será introduzido mais a frente neste capítulo. A Figura 4.2 apresenta alguns exemplos da utilização da extrapolação com diferentes valores de τ.

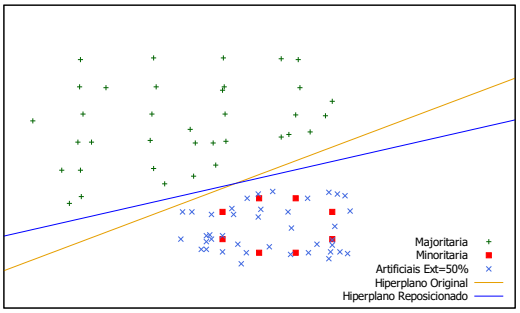
A possibilidade de extrapolação dos limites da classe minoritária, apesar de apresentar possíveis vantagens, ao ser utilizada isoladamente ainda tende a gerar um grande volume de amostras sem representatividade. Isto ocorre porque boa parte dos exemplos utilizados como referência durante o processo de geração continuam sendo escolhidos entre um conjunto de amostras com pouca relevância. Na tentativa de obter amostras de referência mais relevantes, um novo componente foi proposto no método, objetivando que o algoritmo passe a utilizar informações provenientes de classificadores de larga margem



(a) $\tau = 10$



(b) $\tau = 25$



(c) $\tau = 50$

Figura 4.2: Extrapolações com diferentes valores de τ. Conforme esperado, a geração de amostras fora dos limites originais da classe minoritária levam a um reposicionamento do hiperplano em direção a classe majoritária.

para selecionar amostras de interesse, diferindo, neste aspecto, das demais técnicas encontradas na literatura. Neste trabalho a noção de relevância não é definida com base na distância entre a amostra e os exemplos da classe majoritária que a rodeiam, mas sim no quão importante a amostra realmente é para a construção do hiperplano atual. Com esta informação espera-se obter referências mais assertivas e consequentemente, gerar novos exemplos artificiais com uma maior representatividade. A Figura 4.3 apresenta um exemplo da aplicação deste conceito.

O Algoritmo de Balanceamento Sintético Incremental (*Incremental Synthetic Balancing Algorithm* – ISBA) baseia-se na realização das seguintes etapas:

1. Execução de um classificador de larga margem;

2. Armazenamento dos vetores suporte utilizados como base para a construção do hiperplano;

3. Geração de novas amostras artificiais, selecionando como referência um dos vetores suporte e um dos seus k vizinhos mais próximos, sendo que o mesmo vetor suporte

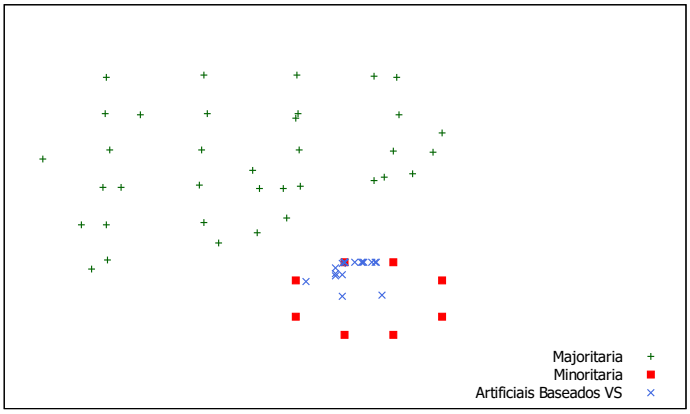


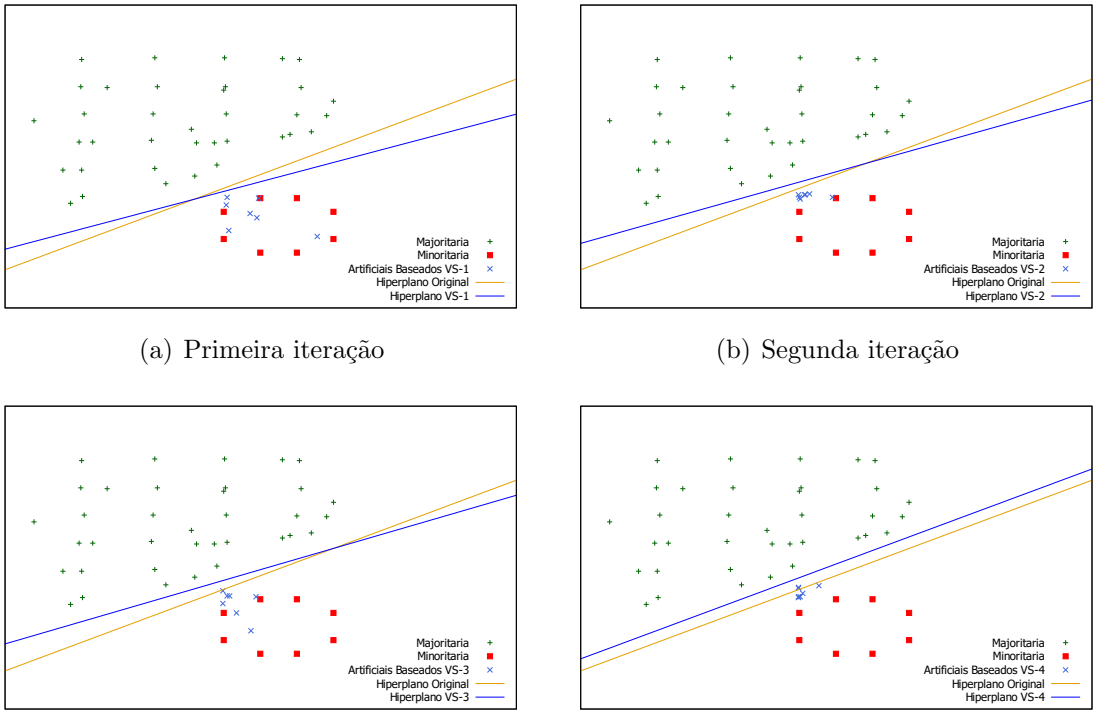
Figura 4.3: Reamostragem baseada em classificadores de larga margem, sem permitir extrapolação. É possível observar que as amostras geradas encontram-se nas áreas próximas a região de decisão, sendo portanto consideradas de especial relevância.

pode ser utilizado mais de uma vez, dependendo do número de amostras a serem geradas.

O balanceamento desejado pode ser atingido a partir de uma única execução destas etapas, ou seja, aplicando-se o classificador de larga margem somente no início do processo. Entretanto o método proposto baseia-se em uma abordagem incremental, onde um conjunto reduzido de instâncias é gerado a cada iteração, permitindo que o hiperplano e os vetores suporte sejam gradualmente atualizados, fazendo com que a solução seja direcionada de uma forma mais natural até o resultado desejado. A Figura 4.4 ilustra este comportamento ao longo de algumas iterações em um exemplo de duas dimensões.

A partir da utilização dos vetores suporte na geração das novas instâncias é possível obter amostras mais representativas, movimentando o plano em direção à classe majoritária. É importante destacar que a base original não foi descaracterizada, dado que as novas instâncias encontram-se sempre em suas redondezas. Outro aspecto a ser ressaltado é a concentração das amostras geradas, que ocorre quando se tem um reduzido número de vetores suporte, fato que seria minorado em casos com um número maior de dimensões.

Apesar do modelo de geração descrito poder ser aplicado uma única vez, utilizando a base original para a obtenção dos vetores suporte que servirão de referência, apresenta-se como uma estratégia mais interessante a utilização destes vetores suporte no decorrer do processo de introdução das amostras artificiais. Porém, um método de larga margem com solução em lote para tal fim demandaria um grande incremento no custo computacional. Desta forma, um método de larga margem construído de forma iterativa apresenta-se



(a) Primeira iteração

(b) Segunda iteração

(c) Terceira iteração

(d) Quarta iteração

Figura 4.4: Reposicionamento dos hiperplanos e geração de amostras artificiais. É possível observar a gradual movimentação do hiperplano e das novas amostras em direção a classe majoritária, levando a um aumento do poder de generalização do classificador com relação as amostras de interesse.

como uma opção mais adequada. Assim, foi feita a opção pela utilização do IMA$_p$, considerando que após o acréscimo dos novos exemplos ao problema, o algoritmo é capaz de recuperar a última solução, continuando o aprendizado deste ponto, não sendo necessária a reinicialização do processo de otimização, o que seria altamente custoso computacionalmente. Outro aspecto importante sobre a utilização do IMA$_p$ é que sua implementação possibilita estabelecer uma ordem para a verificação de classificação das amostras durante o treinamento, o que permite priorizar as novas amostras geradas. Como estas amostras tendem a ser justamente as que estão incorretas, provocando assim uma necessidade de reposicionamento do hiperplano, a sua seleção nas primeiras etapas do treinamento permite uma aceleração na convergência do algoritmo.

O novo método de reamostragem proposto, apesar de ser capaz de obter amostras mais representativas, também aumenta o risco de geração de amostras distorcidas. Visando minimizar este risco, um último aprimoramento foi inserido na abordagem, objetivando com isto realizar uma verificação de cada novo exemplo artificial gerado, descartando os que

forem considerados possivelmente nocivos ao processo de aprendizado. Esta verificação é baseada no conceito de margem e de valor funcional das amostras, sendo a margem a distância entre o hiperplano e a amostra mais próxima e o valor funcional a distância entre a amostra e o hiperplano (vide Figura 4.5). Ressalta-se que tanto o valor da margem, quanto do valor funcional, são atualizados a cada iteração do IMA_p, podendo então ser utilizados como critério de descarte. O Algoritmo 3 descreve o procedimento utilizado para a validação dos novos exemplos gerados.

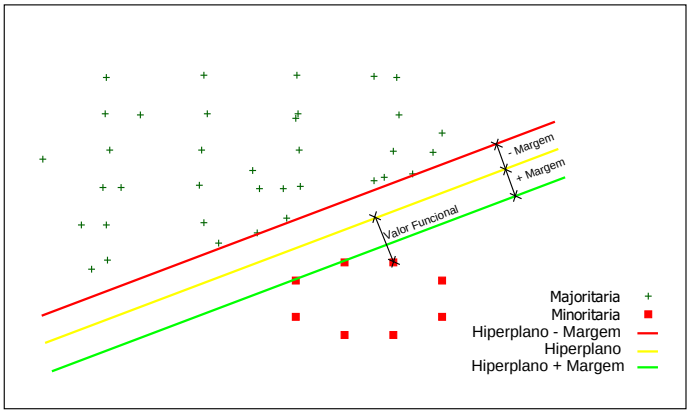


Figura 4.5: Margem e valor funcional. Ao utilizar apenas amostras com valores funcionais maiores que -Margem, o algoritmo não permite a mistura entre classes, ressaltando que nos casos onde é adotada margem flexível os novos pontos poderão ser gerados até o limite permitido pela flexibilização.

Algoritmo 3: Validação de amostra sintética baseada em margem

Entrada: amostras de referência: A e B;
amostra gerada: G;
margem: M;
hiperplano: H;
Saída: amostra artificial *aceita* ou *descartada*
início
  se *G não extrapolar os limites definidos por A e B* **então**
    retorna *amostra aceita*
  senão
    $ValorFuncional \leftarrow$ distância entre G e H;
    se $ValorFuncional > -M$ **então**
      retorna *aceita*
    senão
      retorna *descartada*
    fim se
  fim se
fim

O Algoritmo 3 é executado para cada amostra gerada, com isto espera-se reduzir o risco da utilização da extrapolação, pois a margem funciona como um limite que é automaticamente adaptado para cada problema, dado que, quanto maior a margem, mais seguro é extrapolar os limites da classe minoritária, sem que haja distorções em relação a outra classe. O número de amostras descartadas durante o processo também pode ser utilizado no ajuste de τ, levando-se em conta que no caso de altas taxas de descarte, uma redução no valor de τ é aconselhável.

Finalizando, todos os componentes visando construir um modelo de geração de dados artificiais baseados em procedimento incremental de larga margem são concatenados adequadamente formalizando o processo construtivo de geração de amostras, que é devidamente descrito no Algoritmo 4.

Algoritmo 4: Algoritmo de Balanceamento Sintético Incremental

Entrada: número de amostras a serem geradas para cada vetor suporte: X;
porcentagem de extrapolação: τ;
número de vizinhos mais próximos utilizados: K;
Saída: amostras artificiais geradas;
início
    executar IMA$_p$;
    $M \leftarrow$ margem do primeiro IMA$_p$ (antes início balanceamento);
    $H \leftarrow$ hiperplano gerado pelo IMA$_p$;
    enquanto *base desbalanceada* **faça**
        $VetorVS \leftarrow$ vetores suportes encontrados pelo IMA$_p$;
        para i **de** *1* **até** *total de vetores suportes* **faça**
            $A \leftarrow i$-ésimo elemento de $VetorVS$;
            para j **de** *1* **até** X **faça**
                $B \leftarrow VizinhoMaisProximo(A, K)$;
                $G \leftarrow$ *Geração de amostra artificial com extrapolação*(A, B, τ);
                $R \leftarrow$ *Validação de amostra sinstética baseada em margem*(A, B, G, M, H);
                se $R =$ *aceita* **então**
                    inserir amostra na base;
                senão
                    $j \leftarrow j - 1$;
                fim se
            fim para
        fim para
        executar IMA$_p$;
        $M \leftarrow$ margem atualizada do IMA$_p$;
        $H \leftarrow$ hiperplano gerado pelo IMA$_p$;
    fim enqto
    //$VizinhoMaisProximo()$ retorna aleatoriamente um dos K vizinhos mais próximas da amostra passada como parâmetro
fim

5 EXPERIMENTOS E RESULTADOS

Para avaliar a performance do método proposto foram utilizadas cinco bases binárias desbalanceadas e linearmente separáveis: Colon, Colon100, DLBCL, Leukemia e Sonar. As bases estão contidas no Repositório de Aprendizado de Máquinas da UCI (BACHE; LICHMAN, 2013) ou referenciadas por Golub et al. (1999) ou Alon et al. (1999). A base Colon100 foi gerada a partir da seleção de 100 atributos da base Colon. Os testes também foram realizados em cinco bases não linearmente separáveis retiradas do *KEEL-dataset repository* (SÁNCHEZ et al., 2011): Abalone9-18, Ecoli2, Vowel0, Yeast1 e Yeast5. Nas bases não linearmente separáveis foi adotado o menor valor de flexibilização onde é possível atingir uma convergência, partindo de 0,01 e aplicando acréscimos de potências de 10. Em todos os experimentos foi adotada a opção de divisão em treino e teste através da realização de *5-fold-cross-validation*, sendo que, para os casos onde aplicam-se variáveis randômicas (SMOTE e ISBA), cada *cross-validation* foi executado 5 vezes, buscando assim reduzir possíveis distorções.

Considerando que medidas simples de acurácia tendem a mascarar os resultados obtidos em bases desbalanceadas, as métricas adotadas foram as medidas de *precision* (P), *recall* (R), *F-measure* (F) e F_β*-measure* (F_β), sendo a F_β*-measure* a mais significante delas, pois permite um certo balanceamento entre *precision* e *recall*, porém dando preferência ao *recall*, que é a medida responsável por indicar a ocorrência de falso negativo, considerado o caso mais grave de erro, uma vez que representa classificar uma instância da classe minoritária como sendo da majoritária. Levando em conta para os cálculos o número de ocorrências de verdadeiros positivos (VP), falsos positivos (FP) e falsos negativos (FN), estas medidas podem ser definidas como:

$$P = \frac{VP}{VP + FP} \tag{5.1}$$

$$R = \frac{VP}{VP + FN} \tag{5.2}$$

$$F = 2 \cdot \frac{P \cdot R}{P + R} \tag{5.3}$$

$$F_\beta = (1 + \beta^2) \cdot \frac{P \cdot R}{(\beta^2 \cdot P) + R}, \tag{5.4}$$

com $\beta = 2$ para estabelecer maior prioridade para o valor de *Recall*.

Buscando definir valores para os parâmetros de entrada do ISBA capazes de produzir melhores resultados durante o processo de aprendizado, foi desenvolvido um estudo empírico sobre a relação entre estes parâmetros e a qualidade dos hiperplanos gerados. Através deste estudo foi possível obter as seguintes conclusões:

- Em alguns casos, a utilização da mesma flexibilização para a geração das amostras artificiais e para a construção do hiperplano solução não mostra-se uma boa abordagem. Isto ocorre pois o posicionamento das amostras geradas é limitado pelo valor de margem obtido durante os diversos aprendizados que ocorrem internamente ao ISBA, sendo assim, a escolha de um valor baixo de flexibilização, apesar de possibilitar a convergência, tende a obter uma margem pequena, o que pode ocasionar uma restrição muito forte durante a geração das amostras, tornando-as pouco representativas. Buscando solucionar este problema, foi definida uma constante calculada a partir da raiz quadrada da mediana entre as distâncias das amostras da classe minoritária e as amostras da classe majoritária. A flexibilização do ISBA passa a ser definida através da multiplicação desta constante pelo valor de flexibilização desejado para a construção do hiperplano solução, permitindo a obtenção de valores de margem adaptáveis a cada processo de balanceamento.

- A definição manual da porcentagem de extrapolação também mostrou-se pouco efetiva, devido a variação de comportamento entre as bases. Como solução foi proposta uma parametrização baseada no raio da classe minoritária, ficando a extrapolação definida como $\tau = 4 \cdot \sqrt{\text{raio da classe minoritária}}$, onde o raio é a distância entre o centro da classe e a amostra mais distante.

- A utilização de um valor fixo de vizinhos mais próximos (K), conforme proposto pelo SMOTE, também não obteve bons resultados no ISBA, sendo substituído por uma porcentagem do número de amostras da classe minoritária. A cada iteração o valor de K e atualizado para 10% do número de amostras minoritárias, porém garantindo um valor mínimo de $K = 3$, para evitar erros nos casos onde a classe minoritária possui um número de amostras muito reduzido.

- Durante os testes foi observado que o número de vetores suporte ($\#VS$) tende a aumentar significativamente conforme o ISBA gera novas amostras, com isto foi necessário inserir uma nova regra para definir quantas amostras são geradas a partir de cada vetor suporte (parâmetro X), evitando assim uma explosão repentina no número de amostras geradas. Com esta nova regra, o número de amostras geradas passa a ser limitada com base na cardinalidade da classe minoritária original ($\#Min_{in}$), ficando X redefinido como: (i) no caso de existir apenas um vetor suporte, $X = \#Min_{in}$; (ii) caso contrário, $X = \frac{\#Min_{in}}{\#VS}$.

- Devido a maior importância da classe minoritária, um dos comportamentos desejáveis é a movimentação do hiperplano gerado na direção da classe com mais amostras. O ISBA propões obter este resultado através da inserção do conceito de extrapolação dos limites da base original, entretanto, devido a complexidade inerente ao número de dimensões dos problemas, muitas da vezes, a geração de novas amostras resulta apenas em uma rotação do hiperplano, sem que isto ocasione uma aproximação deste da classe majoritária. Para minimizar este problema foi necessária um redefinição do conceito de balanceamento ideal, passando este a ser medido não apenas pela igualdade de cardinalidade entre as classes, mas também em relação ao número de dimensões do problema (d). Assim o critério de parada utilizado passa a ser uma comparação entre o número de amostras da classe minoritária ($\#Min$) e o número de amostras da classe majoritária ($\#Maj$) multiplicada por um fator de correção (FC): $\#Min > \#Maj \cdot FC$, com $FC = \sqrt{\sqrt{d}}$ e $FC = \sqrt{\sqrt{\sqrt{d}}}$ no caso específico das bases de *microarray*.

Como *baselines* de comparação foram utilizadas três possibilidades de solução: (i) treinamento a partir dos dados originais; (ii) treinamento a partir da base balanceada pelo SMOTE; (iii) e utilização dos dados originais no treinamento, porém somando o valor da margem (γ) ao hiperplano obtido, movendo-o assim o máximo possível em direção a classe majoritária sem modificar sua direção. Todos os processos de aprendizado foram realizados através da execução do IMA$_p$.

Por fim, buscando assegurar resultados confiáveis, mais alguma medidas mostraram-se necessárias, como a separação das bases em treino e teste de forma estratificada, ou seja, mantendo a proporção entre as classes presente na base original. Procedimento que, apesar de também ser recomendado em situações onde a base é balanceada, no caso específico

de bases desbalanceadas, mostra-se indispensável, visto que um conjunto de teste mal construído pode enviesar totalmente os resultados obtidos. Outro ponto importante é a atenção para a realização de *k-fold*, pois as amostras artificiais geradas não podem ser consideradas como possíveis pontos de teste, sendo portanto necessário sempre separar os conjuntos de teste antes do início do processo de balanceamento.

As tabelas 5.1 e 5.2 contém os resultados obtidos através dos experimentos realizados, sendo importante ressaltar que, nos casos onde a flexibilização foi adotada, todas as bases convergiram através da adoção de uma constante de flexibilização igual a 0,01.

Tabela 5.1: Bases linearmente separáveis

Base	Método	Precision	Recall	F-Measure	F_β-Measure
Colon	IMA_p	0,7667 (0,25)	0,6200 (0,38)	0,6376 (0,30)	0,6221 (0,35)
	$IMA_p+\gamma$	0,5717 (0,26)	0,9000 (0,22)	0,6815 (0,22)	0,7900 (0,21)
	SMOTE	0,7667 (0,23)	0,6200 (0,34)	0,6376 (0,27)	0,6221 (0,32)
	ISBA	0,7581 (0,22)	0,6360 (0,35)	0,6382 (0,27)	0,6309 (0,32)
Colon100	IMA_p	0,8267 (0,12)	0,7600 (0,26)	0,7736 (0,17)	0,7620 (0,22)
	$IMA_p+\gamma$	0,5445 (0,07)	0,8800 (0,11)	0,6686 (0,06)	0,7790 (0,07)
	SMOTE	0,8739 (0,11)	0,7360 (0,21)	0,7758 (0,12)	0,7471 (0,17)
	ISBA	0,7790 (0,14)	0,7520 (0,23)	0,7409 (0,14)	0,7424 (0,19)
DLBCL	IMA_p	0,9500 (0,11)	0,9000 (0,14)	0,9214 (0,11)	0,9079 (0,13)
	$IMA_p+\gamma$	0,5222 (0,10)	1,0000 (0,00)	0,6819 (0,08)	0,8403 (0,05)
	SMOTE	0,9400 (0,11)	0,9000 (0,13)	0,9171 (0,11)	0,9063 (0,12)
	ISBA	0,9900 (0,05)	0,9000 (0,13)	0,9386 (0,08)	0,9142 (0,11)
Leukemia	IMA_p	0,9667 (0,07)	0,8800 (0,18)	0,9096 (0,10)	0,8894 (0,15)
	$IMA_p+\gamma$	0,5381 (0,06)	1,0000 (0,00)	0,6979 (0,05)	0,8514 (0,03)
	SMOTE	0,9667 (0,07)	0,8800 (0,16)	0,9096 (0,09)	0,8894 (0,14)
	ISBA	0,9400 (0,08)	0,8800 (0,16)	0,8951 (0,08)	0,8833 (0,13)
Sonar	IMA_p	0,4556 (0,40)	0,4000 (0,38)	0,3574 (0,26)	0,3695 (0,31)
	$IMA_p+\gamma$	0,5030 (0,32)	0,7000 (0,33)	0,4810 (0,15)	0,5725 (0,22)
	SMOTE	0,6287 (0,35)	0,5100 (0,28)	0,4655 (0,18)	0,4732 (0,22)
	ISBA	0,6216 (0,32)	0,5888 (0,29)	0,5136 (0,19)	0,5367 (0,23)

Uma análise dos experimentos realizados nas bases linearmente separáveis indica uma grande proximidade entre os resultados obtidos através das quatro abordagens. Uma possível justificativa para este comportamento é o menor impacto negativo do desbalanceamento nos casos onde é possível obter soluções de margem rígida, pois ao não permitir uma flexibilização, o algoritmo evita também que a classe minoritária seja considerada como ruído, o que torna o desbalanceamento da base menos nocivo. Nestas situações a aplicação de técnicas mais sofisticadas de reamostragem mostra-se menos relevante, entre-

Tabela 5.2: Bases não linearmente separáveis utilizando flexibilização

Base	Método	Precision	Recall	F-Measure	F_β-Measure
Abalone	IMA_p	0,8350 (0,21)	0,6000 (0,12)	0,6906 (0,13)	0,6320 (0,12)
	$IMA_p+\gamma$	0,7178 (0,20)	0,7389 (0,14)	0,7265 (0,17)	0,7335 (0,15)
	SMOTE	0,2776 (0,04)	0,9144 (0,05)	0,4243 (0,05)	0,6237 (0,05)
	ISBA	0,3660 (0,14)	0,8956 (0,07)	0,5036 (0,13)	0,6704 (0,09)
Ecoli	IMA_p	0,6554 (0,14)	0,8091 (0,06)	0,7199 (0,11)	0,7693 (0,08)
	$IMA_p+\gamma$	0,6534 (0,12)	0,9036 (0,00)	0,7533 (0,08)	0,8348 (0,04)
	SMOTE	0,5656 (0,09)	0,9236 (0,04)	0,6978 (0,08)	0,8160 (0,06)
	ISBA	0,5944 (0,10)	0,9204 (0,07)	0,7188 (0,08)	0,8260 (0,07)
Vowel	IMA_p	0,8814 (0,09)	0,8111 (0,08)	0,8386 (0,03)	0,8206 (0,06)
	$IMA_p+\gamma$	0,8759 (0,10)	0,8333 (0,08)	0,8478 (0,03)	0,8377 (0,05)
	SMOTE	0,7420 (0,14)	0,9578 (0,04)	0,8286 (0,09)	0,8990 (0,05)
	ISBA	0,7102 (0,10)	0,9356 (0,06)	0,8019 (0,06)	0,8752 (0,04)
Yeast1	IMA_p	0,7913 (0,21)	0,1374 (0,07)	0,2193 (0,09)	0,1611 (0,08)
	$IMA_p+\gamma$	0,7406 (0,17)	0,1934 (0,06)	0,2955 (0,08)	0,2240 (0,07)
	SMOTE	0,4297 (0,02)	0,8591 (0,06)	0,5720 (0,02)	0,7150 (0,04)
	ISBA	0,5724 (0,05)	0,5468 (0,07)	0,5553 (0,04)	0,5492 (0,06)
Yeast5	IMA_p	0,6783 (0,19)	0,4028 (0,23)	0,4566 (0,18)	0,4196 (0,21)
	$IMA_p+\gamma$	0,7406 (0,17)	0,1934 (0,06)	0,2955 (0,08)	0,2240 (0,07)
	SMOTE	0,4131 (0,05)	0,9906 (0,03)	0,5809 (0,05)	0,7709 (0,04)
	ISBA	0,4351 (0,11)	0,9222 (0,10)	0,5825 (0,10)	0,7422 (0,09)

tanto é importante ressaltar que o ISBA foi capaz de manter a qualidade dos resultados obtidos pelas outras abordagens, fornecendo indícios de sua estabilidade.

Utilizando F_β como medida de comparação, os experimentos realizados nas bases não linearmente separáveis permite apontar os seguintes comportamentos:

- O ISBA é o único dos algoritmo que consegue, em todos os casos, garantir resultados melhores do que os obtidos pela utilização do IMA_p sem nenhuma modificação, evidenciando a confiabilidade do método, a qual se deve em grande parte a inserção da etapa de validação das amostras geradas no processo.

- A movimentação do hiperplano baseada apenas na margem, apesar de apresentar bons resultados em boa parte das bases, mostra-se como uma alternativa de alto risco, visto que nenhuma conferência é realizada, podendo levar a casos de falha, como ocorre na base Yeast5. Justificando assim a utilização de uma abordagem mais elaboradas, como a proposta pelo ISBA.

- Uma comparação do ISBA com o SMOTE, o qual é o *baseline* normalmente adotado na literatura, permite observar que o método proposto consegue melhores resultados em 2 das 5 bases, obtendo resultado piores, porém semelhantes em outras 2, indicando a validade da abordagem. Também é importante destacar que, de forma geral, os resultados do ISBA encontram-se próximos aos do SMOTE, sendo ainda necessária uma melhor parametrização do método.

De forma geral, é possível concluir que o ISBA, apesar de não apresentar os melhores resultados em todas as bases, possui uma característica de fundamental importância, que é a confiabilidade das bases obtidas através da sua utilização. Este aspecto deve-se a adoção dos conceitos referentes aos classificadores de larga margem para a geração das amostras artificiais e pode ser observado na estabilidade dos resultados obtidos. Também é importante destacar que melhores resultados podem ser atingidos, ainda mantendo esta confiabilidade, desde que parâmetros mais adequados sejam definidos, sendo portanto de extrema relevância a realização de maiores estudos neste sentido.

6 BALANCEAMENTO NO ESPAÇO DE CARACTERÍSTICAS

A proposta inicial deste trabalho, a qual resultou no desenvolvimento do ISBA, é baseada em uma abordagem voltada a solução do problema de classificação no espaço de entrada, sendo, portanto, importante ressaltar dois aspectos do método que mostram-se necessários para seu correto funcionamento: (i) todo o processo de geração de amostras, desde a escolha das instâncias de referência até a validação das amostras, deve ser realizado no espaço de entrada; (ii) todos os classificadores utilizados durante o processo devem adotar abordagens restritas ao espaço de entrada. Devido a esta característica, o algoritmo é limitado ao uso de constantes de flexibilização em aplicações não linearmente separáveis, não sendo possível a utilização de funções *kernel*.

A flexibilização, apesar de viabilizar o aprendizado em bases não linearmente separáveis sem a necessidade de um mapeamento das amostras para um espaço de maior dimensionalidade, não necessariamente garante os melhores resultados para todos os casos e, dependendo da base, nem mesmo consegue uma solução adequada. Desta forma considera-se relevante, ainda no escopo deste trabalho, o estudo da aplicação de funções *kernel* durante o processo de construção do hiperplano, assim como a avaliação da possibilidade de utilização destas funções em conjunto com o ISBA.

Uma solução simplista para estender o ISBA para os casos onde mostra-se interessante a aplicação de funções *kernel* seria a substituição do IMA_p por sua formulação dual, proposta por Leite and Fonseca Neto (2008). Com esta modificação, os vetores suporte utilizados como referência durante a geração das amostras artificiais passariam a ser obtidos através da solução do problema no espaço de característica. Após esta etapa, o restante do processo de reamostragem poderia ocorrer normalmente no espaço de entrada, resultando, conforme desejado, na viabilização da utilização de funções *kernel* em conjunto com o método proposto.

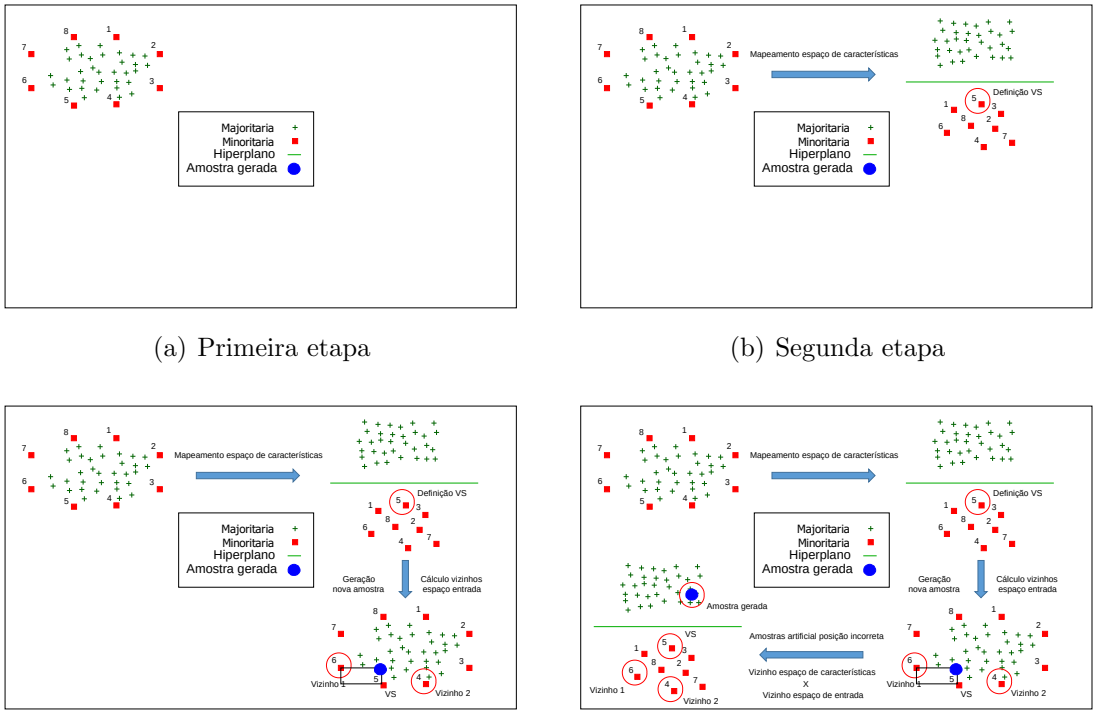
A substituição do IMA_p pela sua formulação dual possibilita a utilização do método também no espaço de características, entretanto esta solução apresenta alguns aspectos negativos, a citar:

- Uma das etapas do ISBA consiste na seleção de duas instâncias de referência, sendo uma delas um vetor suporte e a outra um de seus vizinhos mais próximos. Com a modificação para o IMA dual, a escolha do vetor suporte passa a ocorrer com base no seu posicionamento no espaço de características, entretanto a definição dos vizinhos mais próximos continua a ser calculada com base nas distâncias no espaço de entrada. O problema desta abordagem é a impossibilidade de garantir que a noção de distância utilizada no cálculo dos vizinhos mais próximos se estende ao espaço de características, o que pode levar a distorções em relação as amostras a serem utilizadas na geração dos dados artificiais. A Figura 6.1 procura ilustrar essa situação, sendo possível observar que as amostras definidas como vizinhos mais próximos no espaço de entrada encontram-se em extremidades opostas no espaço de características.

- Alguns trabalhos presentes na literatura têm apontado a ineficiência da geração de amostras artificiais no espaço de entrada, quando o objetivo é construir um classificador no espaço de características. Estes trabalhos argumentam que a geração de amostras com comportamentos similares no espaço de entrada não garante a manutenção desta similaridade em espaços de mais alta dimensão, consequentemente podendo resultar na geração de amostras sem nenhuma representatividade (PÉREZ-ORTIZ et al., 2013) (MATHEW et al., 2015). Novamente, a Figura 6.1 fornece um possível exemplo desta situação, ilustrando um caso onde a nova amostra artificial gerada não manteve o padrão de sua classe de origem, sendo mapeada para o lado errado do hiperplano após a utilização do *kernel*.

- Testes preliminares realizados em trabalhos anteriores com uma implementação análoga, porém baseada em SVM, também apontaram a imprevisibilidade da geração de amostras no espaço de entrada, para posterior mapeamento destas em um espaço de maior dimensão durante o treinamento do classificador (MARQUES et al., 2016).

Como a substituição do classificador primal pela sua formulação dual não resulta em uma extensão confiável do ISBA, em especial devido a imprevisibilidade resultante da geração de amostras no espaço de entrada em situações onde o hiperplano é construído no espaço de características, maiores estudos mostram-se necessários na busca por outras abordagens capazes de tratar este problema de forma mais eficiente. Neste contexto, uma



(a) Primeira etapa

(b) Segunda etapa

(c) Terceira etapa

(d) Quarta etapa

Figura 6.1: Geração de amostras no espaço de entrada e mapeamento no espaço de características.

possível técnica de interesse é apresentada por Mathew et al. (2015), onde é proposto o *Kernel-Based* SMOTE, algoritmo capaz de gerar amostras diretamente no espaço de características, evitando assim a distorção produzida durante o processo de mapeamento.

6.1 *KERNEL-BASED* SMOTE – K-SMOTE

Uma possível solução para o problema da distorção gerada pelo mapeamento das amostras artificiais do espaço de entrada para o espaço de características é a geração das amostras diretamente no espaço de características, conforme proposto por Mathew et al. (2015). Esta técnica utiliza apenas os produtos internos entre as amostras pertencentes à base original, armazenados em uma matriz *kernel*, que por sua vez é submetida a um processo de expansão, passando a representar também os pontos artificiais gerados. Para permitir esta alteração, com relação ao SMOTE original, algumas manipulações matemáticas são necessárias, sendo estas apresentadas nesta seção.

Primeiramente é necessário alterar o cálculo dos vizinhos mais próximos, para que está relação passe a representar a distância entre as amostras no espaço de características. A

distância euclidiana entre x_i e x_j mapeados a partir da função $\phi(.)$ pode ser obtida a partir da matriz *kernel* segundo a seguinte formulação:

$$d^{\phi}\left(x_i, x_j\right)^2 = \left\|\phi\left(x_i\right) - \phi\left(x_j\right)\right\|^2 \tag{6.1}$$

$$d^{\phi}\left(x_i, x_j\right)^2 = \mathbb{K}\left(x_i, x_i\right) - 2\mathbb{K}\left(x_i, x_j\right) + \mathbb{K}\left(x_j, x_j\right) \tag{6.2}$$

Considerando $S_{\min}$ como o conjunto das amostras pertencentes a classe minoritária, $S^*_{\min}$ como o conjunto das P amostras artificiais geradas e θ^{ij} como um valor aleatório sorteado entre $[0, 1]$, novos exemplos são gerados utilizando as instâncias pertencentes a $S_{\min}$ em conjunto com seus vizinhos mais próximos calculados a partir da equação 6.2. Supondo x_i como um ponto de referência e x_j como seu vizinho mais próximo selecionado, o exemplo artificial é gerado através de:

$$\phi\left(x^{ij}\right) = \phi\left(x_i\right) + \theta^{ij}\left(\phi\left(x_j\right) - \phi\left(x_i\right)\right) \tag{6.3}$$

Em 6.3 a geração da amostra depende do mapeamento $\phi(\cdot)$, entretanto a grande vantagem de um classificador *kernel* é que não é necessário conhecer o tipo de mapeamento ou a função ϕ explicitamente. Para tanto, utiliza-se uma função *kernel*, simétrica e semi-definida positiva, definida por $\mathbb{K} : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$. Os valores obtidos pela função $\mathbb{K}$ são correspondentes ao cálculo do produto interno dos vetores mapeados em um espaço de mais alta dimensão (KIVINEN et al., 2004). De forma conveniente, o produto interno das amostras originais pode ser representado a partir de uma matriz *kernel* K^1, onde $K^1_{kl} = \mathbb{K}\left(x_k, x_l\right) = \phi\left(x_k\right)^T \cdot \phi\left(x_l\right)$ com $x_k, x_l \in S$. Sendo então possível expandir K^1 para englobar também as novas amostras geradas, utilizando para isto apenas a relação entre as amostras de referência e as amostras geradas e os produtos internos armazenados em K^1. A nova matriz pode ser representada como:

$$K = \begin{bmatrix} K^1 & K^{2^T} \\ K^2 & K^3 \end{bmatrix} \tag{6.4}$$

onde,

$$K^1 \in \mathbb{R}^{N\text{x}N} \text{ com } \begin{matrix} K^1_{kl} = \mathbb{K}\left(x_k, x_l\right) \\ x_k, x_l \in S \end{matrix} \tag{6.5}$$

$$K^2 \in \mathbb{R}^{N\text{x}P} \text{ com } \begin{matrix} K^2_{kl} = \mathbb{K}\left(x_k, x^{ij}_l\right) \\ x_k \in S, x^{ij}_l \in S^*_{\min} \end{matrix} \tag{6.6}$$

$$K^3 \in \mathbb{R}^{P\text{x}P} \text{ com } \begin{matrix} K^3_{kl} = \mathbb{K}\left(x^{pq}_k, x^{ij}_l\right) \\ x^{pq}_k, x^{ij}_l \in S^*_{\min} \end{matrix} \tag{6.7}$$

sendo i e j os índices das amostras utilizadas como referência para a geração de l e p e q os índices das amostras utilizadas como referência para a geração de k.

Os elementos de K^2 podem ser calculados através de 6.10:

$$\mathbb{K}\left(x_k, x^{ij}_l\right) = \phi\left(x_k\right) \cdot \phi\left(x^{ij}_l\right) \tag{6.8}$$

$$\mathbb{K}\left(x_k, x^{ij}_l\right) = \phi\left(x_k\right) \cdot \Big[\phi\left(x_i\right) + \theta_{ij}\left(\phi\left(x_j\right) - \phi\left(x_i\right)\right)\Big] \tag{6.9}$$

$$\mathbb{K}\left(x_k, x^{ij}_l\right) = \left(1 - \theta_{ij}\right)\mathbb{K}\left(x_k, x_i\right) + \theta_{ij}\mathbb{K}\left(x_k, x_j\right) \tag{6.10}$$

Os elementos de K^3 podem ser calculados através de 6.13:

$$\mathbb{K}\left(x^{pq}_k, x^{ij}_l\right) = \phi\left(x^{pq}_k\right) \cdot \phi\left(x^{ij}_l\right) \tag{6.11}$$

$$\mathbb{K}\left(x^{pq}_k, x^{ij}_l\right) = \Big[\phi\left(x_p\right) + \theta_{pq}\left(\phi\left(x_q\right) - \phi\left(x_p\right)\right)\Big] \cdot \Big[\phi\left(x_i\right) + \theta_{ij}\left(\phi\left(x_j\right) - \phi\left(x_i\right)\right)\Big] \tag{6.12}$$

$$\begin{aligned} \mathbb{K}\left(x^{pq}_k, x^{ij}_l\right) = \; & \left(1 - \theta_{ij}\right)\left(1 - \theta_{pq}\right)\mathbb{K}\left(x_i, x_p\right) \\ & + \left(1 - \theta_{ij}\right)\left(\theta_{pq}\right)\mathbb{K}\left(x_i, x_q\right) \\ & + \left(\theta_{ij}\right)\left(1 - \theta_{pq}\right)\mathbb{K}\left(x_j, x_p\right) \\ & + \left(\theta_{ij}\right)\left(\theta_{pq}\right)\mathbb{K}\left(x_j, x_q\right) \end{aligned} \tag{6.13}$$

Por fim, a combinação das equações 6.2, 6.4, 6.10 e 6.13 permite a extensão do al-

goritmo SMOTE, de forma a gerar as amostras diretamente no espaço de características através de sua sua representação na matriz *kernel* resultante. Com isto é possível a geração de novos exemplos sem a necessidade dos riscos referentes ao mapeamentos.

6.2 PROPOSTA DE ABORDAGEM

A geração de amostras artificiais diretamente no espaço de características permite a redução da distorção, evitando assim mapeamentos inadequados durante o processo. Entretanto, a seleção de amostras de referência e a geração de novos exemplos pelo método K-SMOTE ainda funciona de forma análoga ao SMOTE, podendo então apresentar as mesmas deficiências listadas na seção 3.2. Desta forma, apresenta-se como um possibilidade promissora a proposta de uma solução híbrida, buscando integrar as vantagens do K-SMOTE a geração de amostras mais representativas obtida pelo ISBA.

Considerando a complexidade da abordagem híbrida mencionada, um estudo preliminar foi desenvolvido com o objetivo de realizar uma análise da validade da abordagem. Neste etapa foi implementada uma extensão do algoritmo K-SMOTE, denominada Balanceamento de Passo Único (One Pass Balancing – OPB), sendo inseridos os conceitos de extrapolação e de vetores suporte. Também foi feita a opção por utilizar, inicialmente, o SVM e não do IMA dual como classificador de referência. Esta opção inviabiliza computacionalmente o emprego de uma abordagem incremental, considerando que o SVM precisa realizar um novo processo de otimização para cada inserção de amostras artificiais. Entretanto o uso do SVM oferece algumas vantagens importantes, dado o caráter ainda incipiente do estudo: (i) Mathew et al. (2015) utiliza o SVM durante a realização de todos os testes, tornando uma implementação baseada na mesma formulação mais plausível de comparações; (ii) a utilização do SVM permite a observação do funcionamento do algoritmo em um ambiente mais controlado, considerando o aspecto determinístico do mesmo, enquanto o IMA, por tratar-se de um processo iterativo, apresenta variações no resultado para execuções distintas.

O primeiro passo para viabilizar o desenvolvimento do método proposto foi a implementação das etapas que constituem o próprio K-SMOTE, a saber: (i) o cálculo dos vizinhos mais próximos no espaço de características; (ii) a expansão da matriz *kernel*. Neste ponto é relevante ressaltar que a implementação do K-SMOTE apresenta uma variação importante com relação ao SMOTE. Conforme pode ser observado nas equações

6.10 e 6.13, no K-SMOTE um único valor de θ, sorteado entre $[0, 1]$, é utilizado para a geração da nova amostra, enquanto no SMOTE, um novo valor entre $[0, 1]$ é sorteado para cada dimensão. Esta modificação é necessária para que o método possa expandir a matriz *kernel*, porém também resulta em uma limitação das posições onde as novas amostras artificiais podem ser geradas. Tomando o exemplo da Figura 3.1 como referência, enquanto o SMOTE pode gerar amostras por toda a extensão do hipercubo em vermelho, o K-SMOTE se limita ao segmento de reta em amarelo.

No passo seguinte as modificações propostas para a construção do ISBA foram inseridas no algoritmo K-SMOTE. A obtenção dos vetores suporte no espaço de características é realizada em uma execução do SVM prévia ao balanceamento, através da qual é possível realizar uma análise dos valores de α, sendo todas as amostras com α maior que zero consideradas vetores suporte e armazenadas em um conjunto para servirem de referência para a geração das amostras artificiais. Durante esta etapa os limites de θ também foram alterados de $[0, 1]$ para $[(0 - \tau), (1 + \tau)]$, possibilitando assim que as amostras artificiais extrapolem seus exemplos de referência. Por último, também foi necessária a adoção de outra abordagem para o descarte de amostras distorcidas, pois diferente do que ocorre no IMA, o SVM não possibilita o cálculo dos valores funcionais das amostras, utilizados na validação inicialmente proposta no modelo primal. O novo método para validação das amostras artificiais é descrito no Algoritmo 5.

Conforme mencionado anteriormente, as amostras artificiais geradas pela abordagem proposta são representadas apenas pelos seus produtos internos na matriz *kernel*, não sendo, portanto, possível produzir uma representação gráfica do seu funcionamento. Visando um melhor entendimento, apenas a critério de exemplificação, um mapeamento direto foi realizado utilizando um *kernel* polinomial de grau dois, transformando uma base de duas para três dimensões. Neste exemplo foram destacados os vetores suporte, os exemplos gerados e o hiperplano construído, simulando assim os resultados que seriam obtidos pelo método proposto, porém remapeados para o espaço de entrada. A Figura 6.2 apresenta o resultado deste experimento, o qual tem como objetivo realizar, mesmo que em um problema simples, uma conferência dos conceitos utilizados na construção do método. Sua análise permite a observação de alguns aspectos importantes: (i) os vetores suporte estão corretamente localizados nas regiões mais próximas ao hiperplano separador, sendo portanto uma boa referência para a geração de novas amostras; (ii) as amostras geradas

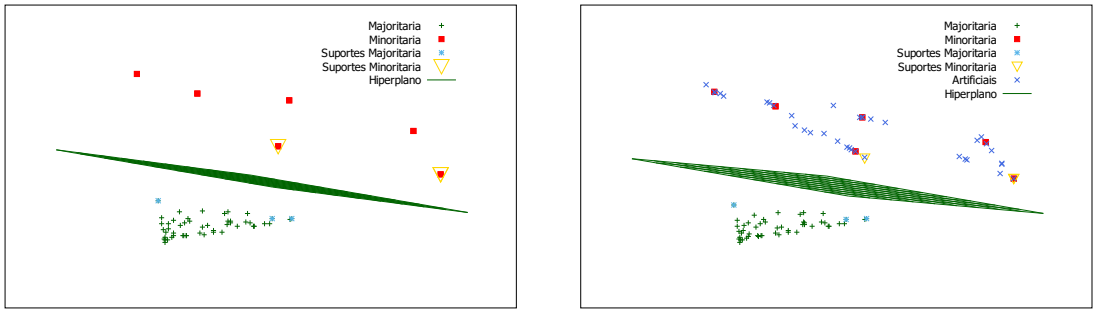
Algoritmo 5: Validação *kernel* de amostra sintética baseada em margem

Entrada: amostras de referência: A e V;
amostra gerada: G;
margem: M;
Saída: amostra artificial *aceita* ou *descartada*
início
  se *G não extrapolar os limites definidos por A e B* **então**
    retorna *amostra aceita*
  senão
    $dist_1 \leftarrow$ distância euclidiana entre A e G;
    $dist_2 \leftarrow$ distância euclidiana entre B e G;
    $dist \leftarrow \min\{dist_1, dist_2\}$;
    se $dist < M$ **então**
      retorna *aceita*
    senão
      retorna *descartada*
    fim se
  fim se
  `// distâncias euclidianas calculadas através da matriz kernel`
fim

são capazes de ultrapassar, mesmo que ligeiramente, os limites da base original, indicando um funcionamento correto da inserção da constante de extrapolação τ; (iii) uma das amostras artificiais externas a base original foi elegida como vetor suporte, ocasionando uma movimentação do hiperplano em direção a classe majoritária, comportamento desejável, considerando a maior importância da classe minoritária.

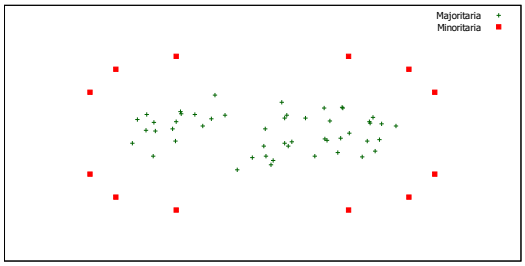
A implementação proposta resulta na obtenção da matriz *kernel* expandida, a qual pode ser utilizada no treinamento de classificadores baseados em uma formulação dual. Entretanto ainda é preciso viabilizar o processo de testes, para que a classe de novas instâncias possam ser inferidas. Com esse objetivo uma nova matriz *kernel* incluindo também as amostras de teste deve ser construída, porém atentando para o fato de que, apenas a matriz formada por K^1, K^2 e K^3 (equação 6.4) deve ser utilizada durante o treinamento. A Figura 6.3 ilustra a divisão das parcelas da matriz a serem utilizadas durante os processo de treino e teste.

Os estudos preliminares apresentados nesta seção fornecem indícios da aplicabilidade da abordagem proposta, sendo portanto justificável o desenvolvimento de um novo algoritmo iterativo similar ao ISBA, porém baseado em uma formulação dual: *Kernel-Based* ISBA – K-ISBA. A maior complexidade na adaptação reside no fato que a cada passo da iteração, a matriz expandida obtida na iteração anterior passa a ser considerada como



(a) Exemplo 3D com suportes

(b) Exemplo 3D + Amostras Artificiais



(c) Exemplo 2D

Figura 6.2: Exemplo do mapeamento polinomial entre os espaços.

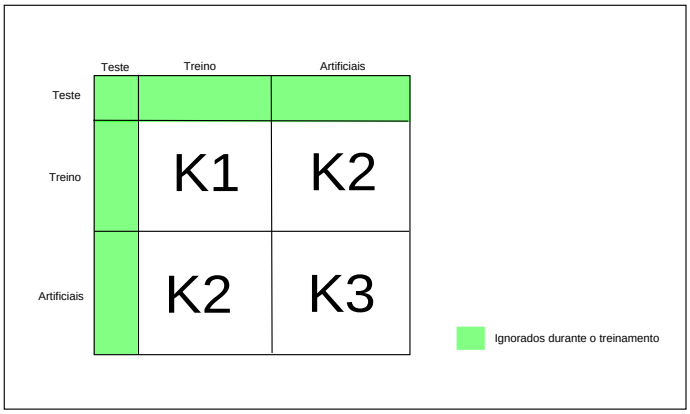


Figura 6.3: Matriz *kernel* expandida com incorporação dos dados de teste. K^1, K^2 e K^3 utilizados durante o treinamento e parcela em verde utilizada durante o teste.

a nova matriz K^1, possibilitando uma atualização das amostras a serem utilizadas como referência para a geração dos novos exemplos artificiais. Esta etapa ainda encontra-se em desenvolvimento, não sendo portanto inserida no escopo desse trabalho. Porém, experimentos utilizando a versão baseada no SVM foram realizados e servem como indicativo para mostrar o padrão do comportamento de um algoritmo de larga margem com geração de amostras artificiais de acordo com a estratégia descrita. Estes resultados serão apresentados a seguir.

6.3 EXPERIMENTOS

Para avaliar a performance do método proposto foram utilizadas três bases não linearmente separáveis: Abalone9-18, Ecoli2 e Yeast5, sendo tomadas as mesmas precauções adotadas durante os experimentos do ISBA para garantir a confiabilidade dos resultados, com a diferença de que a flexibilização foi substituída pela aplicação de funções *kernel*. Funções *kernel* polinomiais não foram capazes de atingir convergência, então em todos os testes foram utilizados o *kernel* gaussiano com 3 diferentes largura: 0,01, 1 e 100.

Quanto aos parâmetros adotados nos experimentos, o percentual de extrapolação foi fixado em 5%, o número de vizinhos mais próximos escolhido foi 5 e o critério de parada é o balanceamento igualitário da base. A tabela 6.2 contém os melhores resultados obtidos entre as 3 larguras, sendo estas larguras apresentadas na tabela 6.1. Os resultados completos encontram-se no apêndice A, assim como uma tabela contendo o percentual de convergência de cada método, o qual representa o número de validações cruzadas onde o algoritmo convergiu, em comparação ao total executado.

Tabela 6.1: Larguras adotadas

Base	γ
Abalone	100
Ecoli	0,01
Yeast5	0,01

Tabela 6.2: Bases não linearmente separáveis utilizando *kernel*

Base	Método	Precision	Recall	F-Measure	F_β-Measure
Abalone	SVM	0,3875 (0,34)	0,3889 (0,38)	0,3870 (0,36)	0,3878 (0,37)
	SMOTE	0,3568 (0,21)	0,6615 (0,25)	0,4528 (0,21)	0,5529 (0,22)
	K-SMOTE	0,3342 (0,26)	0,2975 (0,26)	0,3129 (0,26)	0,3031 (0,26)
	OPB	0,4556 (0,29)	0,4367 (0,32)	0,4439 (0,30)	0,4391 (0,31)
Ecoli	SVM	0,8022 (0,17)	0,7109 (0,16)	0,7526 (0,16)	0,7268 (0,16)
	SMOTE	0,6906 (0,07)	0,9198 (0,07)	0,7866 (0,06)	0,8607 (0,06)
	K-SMOTE	0,6772 (0,08)	0,9261 (0,07)	0,7796 (0,07)	0,8604 (0,06)
	OPB	0,6463 (0,11)	0,9126 (0,07)	0,7499 (0,08)	0,8372 (0,06)
Yeast5	SVM	0,6733 (0,25)	0,3583 (0,25)	0,4217 (0,23)	0,3783 (0,24)
	SMOTE	0,3947 (0,07)	0,9789 (0,05)	0,5602 (0,07)	0,7518 (0,06)
	K-SMOTE	0,4178 (0,05)	1,0000 (0,00)	0,5874 (0,05)	0,7792 (0,04)
	OPB	0,4057 (0,05)	1,0000 (0,00)	0,5753 (0,05)	0,7706 (0,04)

O foco principal dos experimentos encontra-se na comparação entre o OPB e o K-SMOTE. Esta comparação permite observar que, apesar de na maioria dos casos os algoritmos obterem resultados similares, em algumas situações o OPB é capaz de atingir melhores valores de F_β-*Measure*, como na base Abalone, o que indica um potencial a ser explorado através da utilização do método.

A obtenção de resultados similares na aplicação do K-SMOTE e do OPB, apesar de não ser um aspecto desejável, não invalida a abordagem, visto que a falta de melhora se deve em grande parte a utilização *kernel* gaussiano que, por chegar muito próximo da interpolação, não permite ao algoritmo gerar amostras artificiais mais relevantes. Com essa conclusão, mostra-se necessária a busca por outras possibilidades de *kernel*, que sejam capazes de atingir a convergência no processo de aprendizado, porém disponibilizando uma maior margem para que as amostras artificiais possam se aproximar mais da classe majoritária, movimentando assim o hiperplano na direção desejada. Outra abordagem de interesse é a utilização de outras medidas de distância para os cálculos envolvidos, considerando que a distância euclidiana tende a sofrer distorções em casos com muitas dimensões, o que ocorre com a utilização de *kernel*. Neste sentido, experimentos já foram realizados com a distância de tanimoto, mas ainda sem conseguir uma melhora significativa nos resultados.

7 CONCLUSÕES E TRABALHOS FUTUROS

O presente trabalho tem como principal contribuição o desenvolvimento de uma estratégia baseada em uma abordagem incremental usando classificadores de larga margem para a geração de dados artificiais, visando o balanceamento de bases onde a diferença de cardinalidade entre as classes é grande. Em uma primeira etapa, concentrou-se os esforços no desenvolvimento de uma estratégia específica para bases linearmente separáveis, sendo esta fase importante para o levantamento de conceitos que posteriormente seriam estendidos para casos não linearmente separáveis. O algoritmo propõe a utilização de vetores suporte como instâncias de referência no processo de geração dos novos exemplos, bem como um procedimento baseado na extrapolação dos limites da classe minoritária como meio de demarcar de maneira efetiva e confiável as zonas de fronteira entre as classes, buscando assim possibilitar uma melhoria no poder de generalização dos classificadores, principalmente em relação a classe minoritária. Durante o processo de reamostragem, a margem é utilizada como forma de validar as amostras geradas, fornecendo maior confiabilidade ao método.

Durante a fase de estudo sobre os processos de reamostragem e de classificação, foram levantados alguns aspectos funcionais importantes e que não são normalmente discutidos pela literatura, a saber: (i) a influência da região viável de geração de amostras artificiais do SMOTE, definida pelo hipercubo cujas arestas são formadas pela diferença entre os atributos das instâncias de referência; (ii) a compreensão de forma consolidada da razão de ineficiências que podem ocorrer durante o treinamento de classificadores usando bases balanceadas artificialmente, o que facilita a obtenção de alternativas capazes de propor procedimentos de geração mais eficazes. Neste caso identificou-se que a geração de amostras com pouca variabilidade, em locais pouco relevantes e em regiões incorretas foram os três principais motivos para a pouca efetividade dos modelos.

No decorrer dos experimentos, precauções foram necessárias para garantir a confiabilidade dos testes, como a necessidade de divisão das bases em treino e teste de forma estratificada, além da separação dos conjuntos de teste antes do início do processo de validação cruzada, cuidados fundamentais para que os resultados obtidos não fossem enviesados pelo desbalanceamento da base. Ainda durante os experimentos, um comportamento relevante

foi observado, levando a um nova proposta de balanceamento final. Identificou-se que a relação entre a diferença de cardinalidade das classes e o número de dimensões da base é crucial para o processo de balanceamento. Assim mostra-se interessante, para aumentar a eficiência, permitir que, no final da geração, a classe minoritária contenha mais amostras no total em relação a classe majoritária. Apesar de não ser o procedimento usual em modelos de balanceamento, neste caso, devido a alta dimensionalidade dos atributos, a mojoração da cardinalidade da classe minoritária em relação a majoritária torna-se quase mandatória. Porém, deve-se ressaltar que a determinação de qual deve ser o nível da majoração ainda carece de avaliações adicionais.

No que tange a opção adotada baseada na geração incremental das amostras artificiais, os experimentos mostraram evidências da validade desta abordagem. Os resultados indicaram uma melhora quando as instâncias são geradas de forma gradativa, o que permite uma atualização das amostras de referência, a saber, os vetores suporte, que vão se aperfeiçoando com o reposicionamento iterativo do hiperplano. Não foi possível superar os resultados obtidos para todas as bases de dados utilizadas, visto que o problema de classificação em bases desbalanceadas mostra-se extremamento complexo, e uma melhor parametrização do método proposto ainda é necessária para que um maior poder de generalização seja obtido. Finalizando, é importante pontuar que o procedimento de geração incremental só é viável computacionalmente quando baseado em classificador de larga margem iterativo. O uso do SVM, por exemplo, apesar de ser completamente viável em termos de qualidade de balanceamento, demandaria a execução de muitas atualizações das amostras de referência.

Em uma segunda etapa, objetivou-se adaptar a abordagem de balanceamento desenvolvida, visando sua aplicação em bases não linearmente separáveis, porém utilizando funções *kernel* e não flexibilização. Inicialmente, dado a maior complexidade no desenvolvimento de um modelo incremental no espaço de características aos moldes os ISBA, implementou-se um modelo de passo único baseado no SVM com geração diretamente no espaço de características. A geração das amostras artificiais diretamente no espaço de características é crucial para evitar distorções no mapeamento entre os espaços de entrada e características. O processo se dá diretamente nos dados da matriz *kernel*, sem necessidade da utilização dos valores dos atributos das amostras de referência. Porém, desta forma, tem-se uma limitação da região de geração das amostras, passando do hipercubo

definido pelas amostras de referência para a reta que une estas mesmas amostras. Esta abordagem, apesar de ainda não ter sido finalizada, mostrou-se um opção interessante, sendo portanto necessários maiores estudos.

As perspectivas de trabalhos futuros concentram-se principalmente na busca por uma melhor parametrização do ISBA, envolvendo: (i) a proposta de uma parametrização automática do valor de τ, que pode ser obtida através de métodos evolucionistas ou de um estudo mais aprofundado de sua relação com o valor da margem; (ii) a definição de critérios de parada mais eficazes, em especial envolvendo o valor da margem e o número de dimensões do problema; e (iii) uma parametrização também automática do número de amostras a serem geradas em cada iteração do algoritmo. Além destas possibilidades, a finalização da segunda etapa da pesquisa também é de especial interesse, possibilitando a utilização de funções *kernel* de forma iterativa, assim como é feito pelo ISBA. Ainda sobre esta extensão, é importante destacar que são necessárias maiores pesquisas sobre a utilização de outras funções *kernel*, outros critérios de descarte de amostras e a busca por medidas de distâncias mais eficazes no espaço de características, dado que a euclidiana tende a gerar resultados distorcidos. Por fim, também mostram-se relevantes estudos híbridos, envolvendo a utilização de funções *kernel* em conjunto com constantes de flexibilização, assim como a utilização de um classificador para a produção das amostras e de outro para a geração do hiperplano resposta.

REFERÊNCIAS

AIZERMAN, M. Theoretical foundations of the potential function method in pattern recognition learning. **Automation and remote control**, v. 25, p. 821–837, 1964.

ALON, U.; BARKAI, N.; NOTTERMAN, D. A.; GISH, K.; YBARRA, S.; MACK, D.; LEVINE, A. J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. **Proceedings of the National Academy of Sciences of the United States of America**, Department of Molecular Biology, Princeton University, Princeton, NJ 08540, USA, v. 96, n. 12, p. 6745–6750, 1999.

BACHE, K.; LICHMAN, M. **UCI Machine Learning Repository**. 2013. Disponível em: <http://archive.ics.uci.edu/ml>.

BARUA, S.; ISLAM, M. M.; YAO, X.; MURASE, K. Mwmote–majority weighted minority oversampling technique for imbalanced data set learning. **TKDE**, 2014.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: ACM. **Proceedings of the fifth annual workshop on Computational learning theory**, 1992. p. 144–152.

CASTRO, C. L.; BRAGA, A. P. Aprendizado supervisionado com conjuntos de dados desbalanceados. **RCA**, 2011.

CASTRO, H. N.; ABRIL, L. G.; BAHÓN, C. A. A post-processing strategy for svm learning from unbalanced data. In: **19th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning**, 2011. p. 195–200.

CHAN, P. K.; STOLFO, S. J. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In: **KDD**, 1998.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: synthetic minority over-sampling technique. **JAIR**, 2002.

CHAWLA, N. V.; JAPKOWICZ, N.; KOTCZ, A. Special issue on learning from imbalanced data sets. **ACM Sigkdd Explorations Newsletter**, ACM, v. 6, n. 1, p. 1–6, 2004.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, Springer, v. 20, n. 3, p. 273–297, 1995.

DUMAIS, S.; PLATT, J.; HECKERMAN, D.; SAHAMI, M. Inductive learning algorithms and representations for text categorization. In: **P 7th ICIKM**, 1998.

EVERSON, R. M.; FIELDSEND, J. E. Multi-objective optimisation for receiver operating characteristic analysis. **Multi-Objective Machine Learning**, Springer, v. 16, p. 533–556, 2006.

GANGANWAR, V. An overview of classification algorithms for imbalanced datasets. **International Journal of Emerging Technology and Advanced Engineering**, v. 2, n. 4, p. 42–47, 2012.

GOLUB, T. R.; SLONIM, D. K.; TAMAYO, P.; HUARD, C.; GAASENBEEK, M.; MESIROV, J. P.; COLLER, H.; LOH, M. L.; DOWNING, J. R.; CALIGIURI, M. A.; BLOOMFIELD, C. D.; LANDER, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. **Science**, v. 286(5439), p. 531–537, 1999.

HAN, H.; WANG, W.; MAO, B. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: **ICIC**, 2005.

HE, H.; BAI, Y.; GARCIA, E. A.; LI, S. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: IEEE. **Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on**, 2008. p. 1322–1328.

HOWLETT, R. J.; JAIN, L. C. **Radial basis function networks 2: new advances in design**, 2013.

KIVINEN, J.; SMOLA, A. J.; WILLIAMSON, R. C. Online learning with kernels. **IEEE transactions on signal processing**, IEEE, v. 52, n. 8, p. 2165–2176, 2004.

KOTO, F. Smote-out, smote-cosine, and selected-smote: An enhancement strategy to handle imbalance in data level. In: **ICACSIS**, 2014.

KUBAT, M.; HOLTE, R. C.; MATWIN, S. Machine learning for the detection of oil spills in satellite radar images. **ML**, 1998.

LEITE, S. C.; FONSECA NETO, R. Incremental margin algorithm for large margin classifiers. **Neurocomputing**, Elsevier, v. 71, n. 7, p. 1550–1560, 2008.

LIU, A.; GHOSH, J.; MARTIN, C. E. Generative oversampling for mining imbalanced datasets. In: **DMIN**, 2007.

MANEVITZ, L.; YOUSEF, M. One-class document classification via neural networks. **Neurocomputing**, Elsevier, v. 70, n. 7, p. 1466–1481, 2007.

MARQUES, M. L.; VILLELA, S. M.; BORGES, C. C. H. Uma estratégia de geração de dados artificiais para classificadores de largam margem aplicada em bases de dados desbalanceadas. **KDMILE - Symposium on Knowledge Discovery, Mining and Learning**, v. 4, p. 152–159, 2016.

MARSLAND, S. **Machine learning: an algorithmic perspective**, 2014.

MARSLAND, S. **Machine learning: an algorithmic perspective**, 2015.

MATHEW, J.; LUO, M.; PANG, C. K.; CHAN, H. L. Kernel-based smote for svm classification of imbalanced datasets. In: **IECON**, 2015.

MAZUROWSKI, M. A.; HABAS, P. A.; ZURADA, J. M.; LO, J. Y.; BAKER, J. A.; TOURASSI, G. D. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. **Neural networks**, Elsevier, v. 21, n. 2, p. 427–436, 2008.

MURTHY, S. K. Automatic construction of decision trees from data: A multi-disciplinary survey. **Data mining and knowledge discovery**, Kluwer academic publishers, v. 2, n. 4, p. 345–389, 1998.

NOVIKOFF, A. B. **On convergence proofs for perceptrons**, 1963.

PÉREZ-ORTIZ, M.; GUTIÉRREZ, P. A.; HERVÁS-MARTÍNEZ, C. Borderline kernel based over-sampling. In: SPRINGER. **International Conference on Hybrid Artificial Intelligence Systems**, 2013. p. 472–481.

PHUA, C.; ALAHAKOON, D.; LEE, V. Minority report in fraud detection: classification of skewed data. **Acm sigkdd explorations newsletter**, ACM, v. 6, n. 1, p. 50–59, 2004.

POURHABIB, A.; MALLICK, B. K.; DING, Y. Absent data generating classifier for imbalanced class. **JMLR**, 2015.

RADIVOJAC, P.; CHAWLA, N. V.; DUNKER, A. K.; OBRADOVIC, Z. Classification and knowledge discovery in protein databases. **Journal of Biomedical Informatics**, Elsevier, v. 37, n. 4, p. 224–239, 2004.

RASKUTTI, B.; KOWALCZYK, A. Extreme re-balancing for svms: a case study. **ACM Sigkdd Explorations Newsletter**, ACM, v. 6, n. 1, p. 60–69, 2004.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. **Psychological review**, American Psychological Association, v. 65, n. 6, p. 386, 1958.

ROSENBLATT, F. Principles of neurodynamics. Spartan Book, 1962.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. **Learning internal representations by error propagation**, 1985.

SÁNCHEZ, R. L.; ALCALÁ, F. J.; FERNÁNDEZ, H. A.; LUENGO, M. J.; DERRAC, R. J.; GARCÍA, L. S.; HERRERA, T. F. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. **JMLSC**, 2011.

SUN, Y.; KAMEL, M. S.; WONG, A. K. C.; WANG, Y. Cost-sensitive boosting for classification of imbalanced data. **PR**, 2007.

TAO, X.; JI, H.; XIE, Y. A modified psvm and its application to unbalanced data classification. In: **ICNC**, 2007.

TIAN, J.; GU, H.; LIU, W. Imbalanced classification using support vector machine ensemble. **NCA**, 2011.

VILLELA, S. M.; LEITE, S. C.; FONSECA NETO, R. Incremental p-margin algorithm for classification with arbitrary norm. **Pattern Recognition**, v. 55, p. 216–272, 2016.

YEN, S.; LEE, Y. Cluster-based under-sampling approaches for imbalanced data distributions. **ESA**, 2009.

ZHANG, X.; FU, Y.; ZANG, A.; SIGAL, L.; AGAM, G. Learning classifiers from synthetic data using a multichannel autoencoder. **arXiv**, 2015.

Apêndice A - TABELAS COMPLETAS DOS EXPERIMENTOS *KERNEL*

Tabela A.1: Base Abalone

γ	Método	Precision	Recall	F-Measure	F_β-Measure
0,01	SVM	0,6250 (0,53)	0,3403 (0,30)	0,4405 (0,39)	0,3743 (0,33)
	SMOTE	-	-	-	-
	K-SMOTE	-	-	-	-
	OPB	-	-	-	-
1	SVM	-	-	-	-
	SMOTE	-	-	-	-
	K-SMOTE	-	-	-	-
	OPB	-	-	-	-
100	SVM	0,3875 (0,34)	0,3889 (0,38)	0,3870 (0,36)	0,3878 (0,37)
	SMOTE	0,3568 (0,21)	0,6615 (0,25)	0,4528 (0,21)	0,5529 (0,22)
	K-SMOTE	0,3342 (0,26)	0,2975 (0,26)	0,3129 (0,26)	0,3031 (0,26)
	OPB	0,4556 (0,29)	0,4367 (0,32)	0,4439 (0,30)	0,4391 (0,31)

Tabela A.2: Base Ecoli

γ	Método	Precision	Recall	F-Measure	F_β-Measure
0,01	SVM	0,8022 (0,17)	0,7109 (0,16)	0,7526 (0,16)	0,7268 (0,16)
	SMOTE	0,6906 (0,07)	0,9198 (0,07)	0,7866 (0,06)	0,8607 (0,06)
	K-SMOTE	0,6772 (0,08)	0,9261 (0,07)	0,7796 (0,07)	0,8604 (0,06)
	OPB	0,6463 (0,11)	0,9126 (0,07)	0,7499 (0,08)	0,8372 (0,06)
1	SVM	-	-	-	-
	SMOTE	-	-	-	-
	K-SMOTE	0,5384(0,00)	0,6363(0,00)	0,58333(0,00)	0,6140(0,00)
	OPB	-	-	-	-
100	SVM	0,8412 (0,14)	0,5465 (0,19)	0,6527 (0,18)	0,5835 (0,19)
	SMOTE	0,7835 (0,13)	0,7033 (0,24)	0,7223 (0,19)	0,7082 (0,22)
	K-SMOTE	0,8429 (0,13)	0,5436 (0,17)	0,6499 (0,16)	0,5805 (0,17)
	OPB	0,8429 (0,13)	0,5436 (0,17)	0,6499 (0,16)	0,5805 (0,17)

Tabela A.3: Base Yeast5

γ	Método	Precision	Recall	F-Measure	F_β-Measure
0,01	SVM	0,6733 (0,25)	0,3583 (0,25)	0,4217 (0,23)	0,3783 (0,24)
	SMOTE	0,3947 (0,07)	0,9789 (0,05)	0,5602 (0,07)	0,7518 (0,06)
	K-SMOTE	0,4178 (0,05)	1,0000 (0,00)	0,5874 (0,05)	0,7792 (0,04)
	OPB	0,4057 (0,05)	1,0000 (0,00)	0,5753 (0,05)	0,7706 (0,04)
1	SVM	-	-	-	-
	SMOTE	-	-	-	-
	K-SMOTE	-	-	-	-
	OPB	-	-	-	-
100	SVM	-	-	-	-
	SMOTE	-	-	-	-
	K-SMOTE	-	-	-	-
	OPB	-	-	-	-

Tabela A.4: Porcentagens de convergência

Base	γ	Método	Convergência
Abalone	0,01	SVM	40%
		SMOTE	0%
		K-SMOTE	0%
		OPB	0%
	1	SVM	0%
		SMOTE	0%
		K-SMOTE	0%
		OPB	0%
	100	SVM	80%
		SMOTE	64%
		K-SMOTE	76%
		OPB	36%
Ecoli	0,01	SVM	100%
		SMOTE	88%
		K-SMOTE	96%
		OPB	72%
	1	SVM	0%
		SMOTE	0%
		K-SMOTE	0,04%
		OPB	0%
	100	SVM	100%
		SMOTE	100%
		K-SMOTE	100%
		OPB	100%
Yeast5	0,01	SVM	100%
		SMOTE	80%
		K-SMOTE	100%
		OPB	92%
	1	SVM	0%
		SMOTE	0%
		K-SMOTE	0%
		OPB	0%
	100	SVM	20%
		SMOTE	0%
		K-SMOTE	0%
		OPB	0%