

Nahim Alves de Souza

**Aumentando o poder preditivo de classificadores
lineares através de particionamento por classe**

Sorocaba, SP

25 de Janeiro de 2018

Nahim Alves de Souza

Aumentando o poder preditivo de classificadores lineares através de particionamento por classe

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação (PPGCC-So) da Universidade Federal de São Carlos como parte dos requisitos exigidos para a obtenção do título de Mestre em Ciência da Computação. Linha de pesquisa: Computação Científica e Inteligência Computacional.

Universidade Federal de São Carlos – UFSCar

Centro de Ciências em Gestão e Tecnologia – CCGT

Programa de Pós-Graduação em Ciência da Computação – PPGCC-So

Orientador: Profa. Dra. Tiemi Christine Sakata

Coorientador: Prof. Dr. Tiago Agostinho de Almeida

Sorocaba, SP

25 de Janeiro de 2018

Alves de Souza, Nahim

Aumentando o poder preditivo de classificadores lineares através de
particionamento por classe / Nahim Alves de Souza. -- 2018.
49 f. : 30 cm.

Dissertação (mestrado)-Universidade Federal de São Carlos, campus
Sorocaba, Sorocaba

Orientador: Profa. Dra. Tiemi Christine Sakata

Banca examinadora: Profa. Dra. Ana Lúcia Cetertich Bazzan, Profa. Dra.
Katti Faceli

Bibliografia

1. Classificação linear. 2. Agrupamento. 3. Aprendizado de máquina. I.
Orientador. II. Universidade Federal de São Carlos. III. Título.



UNIVERSIDADE FEDERAL DE SÃO CARLOS

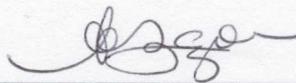
Centro de Ciências em Gestão e Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

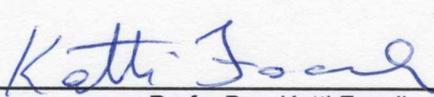
Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Nahim Alves de Souza, realizada em 25/01/2018:



Profa. Dra. Tiemi Christine Sakata
UFSCar



Profa. Dra. Ana Lúcia Cetertich Bazzan
UFRGS



Profa. Dra. Katti Faceli
UFSCar

Aos futuros leitores deste trabalho.

Agradecimentos

Agradeço,

a Deus, por toda orientação e pelas pessoas que Ele colocou no meu caminho para contribuírem de alguma forma com este trabalho.

aos meus pais João e Fátima, por todo o apoio e suporte nos meus estudos durante todos esses anos.

ao meu irmão Davy, por sempre ouvir minhas preocupações e reclamações.

ao Vinícius Angiolucci Reis, pela amizade e pelos conselhos que me encorajaram a concluir este trabalho.

ao Fellipe Augusto Ugliara, por toda colaboração e incentivo na escrita desta dissertação.

aos meus orientadores Tiemi e Tiago, pela ajuda e pela paciência comigo durante o desenvolvimento deste trabalho.

*“O segredo é quebrar os problemas em pequenos pedaços administráveis.
Se você lidar com eles, termina antes de saber disso.”
(Calvin e Haroldo, Bill Watterson)*

Resumo

Este trabalho descreve uma nova técnica de classificação, denominada *P2C - Partitioning to Classify*, cujo objetivo é obter bom desempenho de classificação usando métodos de predição lineares, mesmo em bases com dados não separáveis linearmente. A técnica proposta, inspirada na estratégia de divisão-e-conquista, aplica um método de agrupamento sobre cada partição, constituídas por amostras de mesma classe. Posteriormente, a união dos *clusters* em cada partição é realizada, criando uma única partição, onde cada grupo pode conter amostras separáveis linearmente. Em seguida, um ou mais classificadores lineares são treinados, de acordo com o número de grupos. Experimentos realizados com bases de dados que possuem diferentes estruturas e níveis de complexidade demonstram que o desempenho geral da predição é muitas vezes semelhante ou superior aos métodos de classificação não-lineares. As principais vantagens da técnica *P2C* são (i) a necessidade de menor esforço e recursos computacionais, e (ii) a possibilidade de tratar bases de dados grandes, devido à facilidade de paralelização das etapas.

Palavras-chaves: Classificação linear. Agrupamento. Aprendizado de máquina.

Abstract

This work describes a new classification technique called *P2C - Partitioning to Classify*. The main goal is to achieve reasonable classification performances using linear prediction methods, even on datasets with non-linear separable data. The proposed technique, inspired by the division-and-conquer strategy, applies a clustering method on each partition made of samples of the same class. Subsequently, the union among the clusters inside each partition is performed, creating a single partition, where each group can contain linearly separable samples. Then, one or more linear classifiers are trained, according to the number of groups. Experiments performed using datasets with different structural and complexity level indicate the overall performance of the prediction is similar or superior to well-known non-linear classification methods. The main advantages of *P2C* technique are (i) the need for less effort and computational resources, and (ii) the possibility of treating large datasets due to the ease of parallelization of the steps.

Key-words: Linear classification. Clustering. Machine learning.

Lista de ilustrações

Figura 1	– (a) Representação de uma base de dados com amostras de duas classes não separáveis linearmente. (b) A linha diagonal ilustra um classificador linear treinado sobre a base original. (c) A linha vertical indica a separação das amostras em dois grupos e as linhas diagonais representam os classificadores lineares gerados para cada um deles.	2
Figura 2	– Classificadores lineares em duas bases de dados: (a) bidimensional – uma reta separa as amostras e (b) tridimensional ou multidimensional – um hiperplano separa as amostras	6
Figura 3	– Base de dados bidimensional com classes não separáveis linearmente. Observa-se que em (a) o classificador linear não consegue separar as amostras corretamente, mas em (b) o classificador não-linear é capaz de separar as amostras de cada classe com maior precisão.	7
Figura 4	– Diagrama ilustrando as etapas principais da técnica P2C. A sigla <i>CLF</i> representa um classificador linear e <i>P2C(CLF)</i> representa a combinação final destes classificadores a partir da técnica P2C.	15
Figura 5	– Base de dados artificial utilizada para exemplificar as etapas da técnica P2C proposta neste trabalho. A base é bidimensional e contém amostras de duas classes não separáveis linearmente, representadas pelos círculos e quadrados.	16
Figura 6	– Particionamento por classe da base de dados ilustrada na Figura 5. Cada círculo representa um <i>cluster</i> de amostras encontrado pelo algoritmo de agrupamento. Neste exemplo, foram encontrados dois <i>clusters</i> em cada uma dos subconjuntos, formando duas partições.	16
Figura 7	– Ilustração da etapa de sobreposição de amostras. Os círculos ao redor das amostras representam os <i>clusters</i> . (a) Exemplo de dois <i>clusters</i> com alta taxa de sobreposição ($S_{c_i, c_l} \rightarrow 1$). (b) Exemplo de <i>clusters</i> cujas amostras que não possuem sobreposição ($S_{c_i, c_l} = 0$)	18
Figura 8	– Ilustração da proporção entre <i>clusters</i> . Os círculos ao redor das amostras representam os <i>clusters</i> . (a) Exemplo de dois <i>clusters</i> que possuem um número proporcional de amostras ($P_{c_i, c_l} \rightarrow 1$). (b) Exemplo de <i>clusters</i> desproporcionais em relação ao número de amostras ($P_{c_i, c_l} \rightarrow 0$)	19

Figura 9 – Ilustração referente à etapa de união de <i>clusters</i> formados após o particionamento (Figura 6). As figuras de (a)-(d) representam a comparação realizada entre os <i>clusters</i> , para determinar se eles devem, ou não, ser unidos. Cada uma delas apresenta duas representações da base: à direita, uma visão da base completa e, à esquerda, uma visão das partições que estão sendo comparadas. Note que os <i>clusters</i> destacados em (a) e (c) não serão unidos e os <i>clusters</i> destacados nas figuras (b) e (d) serão unidos.	20
Figura 10 – Resultado da união aplicada nos <i>clusters</i> das partições ilustradas na Figura 6. Na figura, os dois grupos estão separados mas, na prática, os dois grupos pertencem a uma única partição final.	21
Figura 11 – Ilustração dos classificadores lineares treinados para cada subconjunto da partição final produzida na etapa anterior (Figura 10).	21
Figura 12 – Modelo de classificação resultante da aplicação da abordagem proposta. A linha vertical representa a divisão entre os grupos da partição (indicados pelos centroides ‘×’) e as linhas diagonais representam os classificadores lineares treinados para cada grupo. Cada amostra nova de teste é classificada a partir do preditor correspondente ao grupo a que ela pertencer.	22
Figura 13 – Resumo das etapas da técnica P2C.	23
Figura 14 – Bases de dados utilizadas nos experimentos e reportadas na Tabela 5. As cores e formas representam amostras de classes distintas.	26
Figura 15 – Metodologia experimental. <i>CLF</i> corresponde a um método de classificação linear tradicional. <i>P2C(CLF)</i> representa a abordagem proposta neste trabalho, que é a aplicação da técnica sobre o classificador linear.	27
Figura 16 – Classificadores gerados utilizando o método BNB e LSVM. Nota-se que o método BNB obteve um baixíssimo desempenho, uma vez que os classificadores gerados não conseguiram separar as classes corretamente, o oposto do que ocorreu com o método LSVM. (a) <i>Plot</i> do classificador BNB aplicado sobre a base 4c1u . (b) <i>Plot</i> do classificador BNB aplicado sobre a base ds2c2sc13 . (c) <i>Plot</i> do classificador LSVM aplicado sobre a base 4c1u . (b) <i>Plot</i> do classificador LSVM aplicado sobre a base ds2c2sc13	31
Figura 17 – Classificadores gerados utilizando o classificador LSVM com a estratégia <i>one-vs-all</i> . (a) <i>Plot</i> do classificador LSVM aplicado sobre a base 3b1obs . (b) <i>Plot</i> do classificador LSVM aplicado sobre a base 4c1u . (c) <i>Plot</i> do classificador LSVM aplicado sobre a base 4b13c1	32

Figura 18 – Aplicação do método LSVM sobre a base de dados <code>engytime</code> , demonstrando que mesmo um método linear (sem aplicação do P2C) pode conseguir bons resultados nesta base de dados.	33
Figura 19 – Plot da base de dados <code>4b12c1</code> . As cores e formas representam amostras de classes distintas. (a) Classificador gerado utilizando o método RSVM sobre esta base. (b) Classificador final produzido com a aplicação da técnica P2C sobre o método LR.	34
Figura 20 – Etapas de particionamento por classe e união de <i>clusters</i> . (a) Particionamento realizado nas amostras da classe laranja, gerando 5 <i>clusters</i> . (b) Particionamento realizado nas amostras da classe verde, gerando 4 <i>clusters</i> . (c) Resultado da união entre os <i>clusters</i> de (a) e (b), produzindo 4 grupos, utilizados no treinamento do classificador LR.	34
Figura 21 – Classificadores gerados para cada grupo da Figura 20(c), após a união. Em (a) um classificador foi gerado com o método LR. Em (b),(c) e (d) a maior parte das amostras pertenciam a uma única classe, não sendo o treinamento do classificador, pois o mecanismo de classificação é a utilização da classe majoritária no grupo.	35
Figura 22 – Aplicação do método RSVM sobre a base de dados <code>2b14c1</code> . Nota-se que o classificador RSVM não conseguiu separar as classes adequadamente, devido à heterogeneidade entre as amostras das diferentes classes. . . .	36
Figura 23 – Aplicação dos classificadores linear e não-linear sobre a base <code>gauss4c1</code> . (a) Classificador P2C(LSVM) aplicado sobre a mesma base de dados. (b) Classificador RSVM aplicado sobre a mesma base de dados.	36
Figura 24 – Grupos formados após a etapa de união na base <code>gauss4c1</code> . As figuras (a), (b) e (c) mostram grupos formados com amostras de múltiplas classes, dificultando a classificação realizada por um método linear, que conseguiria um melhor resultado caso houvesse amostras de apenas duas classes. (d) Representa um grupo formado por amostras de uma única classe, cujo mecanismo de classificação será a utilização da classe majoritária como rótulo.	37

Lista de tabelas

Tabela 1	– Principais características dos classificadores lineares e não-lineares	8
Tabela 2	– Possíveis valores para o fator de união. A última coluna indica os casos em que a união é aplicada.	19
Tabela 3	– Bases de dados utilizadas nos experimentos. Cada linha representa uma base de dados. O “X” indica que a base possui a característica indicada na coluna.	25
Tabela 4	– Parâmetros de entrada necessários para aplicação da técnica P2C. A última coluna mostra o valor escolhido para as bases de dados utilizadas.	28
Tabela 5	– Desempenho dos classificadores nas bases de dados artificiais pela F-Medida (Macro). Em destaque, os melhores resultados para cada base de dados. Para comparação dos resultados com classificadores não-lineares, foi utilizado o SVM com <i>kernel</i> radial (RSVM).	30
Tabela 6	– Diferença dos valores da F-Medida com e sem a aplicação da técnica P2C, ou seja, o quanto o valor da F-Medida aumentou ou diminuiu. Em destaque, os classificadores que obtiveram maior aumento no desempenho depois da aplicação da técnica P2C. O crescimento médio da F-Medida foi de 0,17.	30

Lista de abreviaturas e siglas

BNB	<i>Bernoulli Naïve Bayes</i>
LR	<i>Logistic Regression</i>
LSVM	<i>Linear Support Vector Machine</i>
RSVM	<i>Radial Support Vector Machine</i>
CLF	Classificador linear
P2C	<i>Partitioning to Classify</i> - técnica de classificação proposta neste trabalho
P2C(CLF)	Aplicação da técnica P2C sobre um classificador linear
P2C(BNB)	Aplicação da técnica P2C sobre o método BNB
P2C(LR)	Aplicação da técnica P2C sobre o método LR
P2C(LSVM)	Aplicação da técnica P2C sobre o método LSVM
TN	<i>Tamanho normalizado - número de amostras de um cluster, normalizado entre 0 e 1.</i>

Lista de símbolos

γ	Limiar de união
K	Número de classes de uma base de dados
k	Número de <i>clusters</i>
k_i	Número de <i>clusters</i> da i -ésima partição
\mathcal{P}	Conjunto de partições
p_i	i -ésima partição
c_i	i -ésimo <i>cluster</i>
c_{i_j}	i -ésimo <i>cluster</i> da j -ésima partição
S_{c_i, c_j}	Sobreposição entre os <i>clusters</i> i e j
P_{c_i, c_j}	Proporção entre os <i>clusters</i> i e j
F_{c_i, c_j}	Fator de união entre os <i>clusters</i> c_i e c_j

Sumário

	Prefácio	1
1	CLASSIFICADORES LINEARES E NÃO-LINEARES	5
1.1	Classificação linear	5
1.2	Classificação não-linear	6
1.3	Comparação entre métodos lineares e não-lineares	7
2	COMBINANDO AGRUPAMENTO E CLASSIFICAÇÃO	9
2.1	Seleção de classificadores	9
2.2	Fusão de classificadores	11
2.3	Particionamento por classe	12
2.4	Particionamento, união e classificação	12
3	P2C - PARTITIONING TO CLASSIFY	15
3.1	Particionamento por classe	15
3.2	União de <i>clusters</i>	17
3.2.1	Sobreposição de amostras	17
3.2.2	Proporção entre <i>clusters</i>	18
3.2.3	Fator de união	18
3.3	Treinamento local	20
3.4	Classificação	22
4	EXPERIMENTOS	25
4.1	Bases de dados	25
4.2	Metodologia experimental	26
4.3	Análise dos resultados	28
4.3.1	Aumento do poder preditivo	29
4.3.2	Diminuição do poder preditivo	32
4.3.3	Equivalência no poder preditivo	32
4.3.4	Desempenho superior ao RSVM	33
4.3.5	Desempenho equivalente ao RSVM	34
4.3.6	Desempenho inferior ao RSVM	35
4.4	Avaliação estatística	37
4.5	Considerações Finais	38
	Conclusão	41

Referências 45

Prefácio

Nos últimos anos, diversos trabalhos têm surgido com o objetivo de oferecer métodos de classificação cada vez mais robustos para lidar com bases de dados complexas. A adoção de métodos de classificação não-lineares, tal como *deep learning* (KRIZHEVSKY; SUTSKEVER; HINTON, 2012; SCHMIDHUBER, 2014), vem tornando-se uma tendência, pois eles normalmente conseguem gerar modelos de predição extremamente sofisticados, robustos e com alto poder preditivo. Todavia, a principal desvantagem desses métodos é que a construção do modelo de predição pode demandar muitos recursos computacionais e tempo de processamento (BRITTO; SABOURIN; OLIVEIRA, 2014).

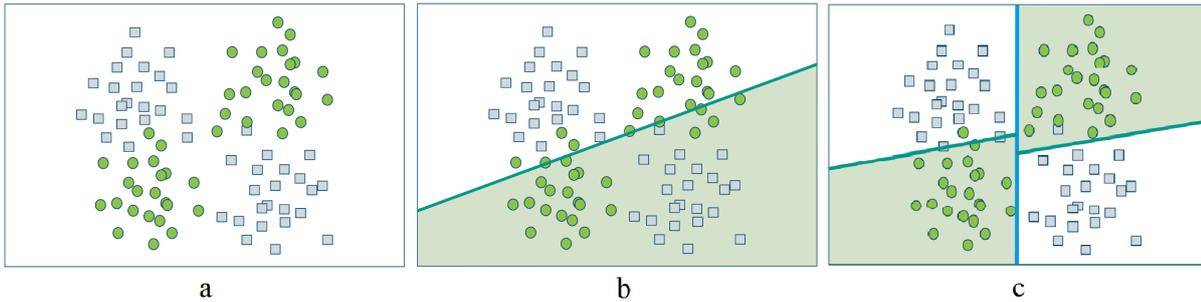
No mesmo cenário de classificação, outras abordagens também têm recebido destaque nesse período, como por exemplo, a fusão e a seleção de classificadores (BRITTO; SABOURIN; OLIVEIRA, 2014; KUNCHEVA, 2000; ALMEIDA; GALVÃO, 2016). Muitas delas são baseadas no princípio de divisão-e-conquista, através do particionamento da base de dados em vários grupos e do treinamento de classificadores especializados para cada um deles. Nesse escopo, Almeida e Galvão (2016) demonstraram que a combinação de vários métodos de classificação “fracos”, por meio de um *ensemble* (DIETTERICH, 2000), pode ser uma boa solução para aumentar consideravelmente o poder preditivo da classificação sem necessariamente aumentar muito o esforço computacional (LOCHTER; ZANETTI; ALMEIDA, 2015).

Em geral, para uma mesma base de dados, métodos de classificação lineares são treinados muito mais rapidamente do que métodos não-lineares (JOACHIMS, 2006), pois possuem uma complexidade menor em relação ao número de operações realizadas (FU; ROBLES-KELLY; ZHOU, 2010; FRADKIN, 2008). Consequentemente, a implementação de métodos lineares é mais simples e costuma ser menos custosa computacionalmente. Por outro lado, dependendo das características da base de dados, um modelo de predição linear pode apresentar desempenho insatisfatório na tarefa de classificação, pois as amostras podem ser inseparáveis linearmente. Neste cenário, este trabalho assume a hipótese de que é possível obter uma combinação de métodos lineares capaz de obter desempenho próximo ou superior a dos métodos de classificação não-lineares.

Para ilustrar uma possível combinação de classificadores lineares, a Figura 1(a) apresenta uma base de dados com duas classes, onde nenhum classificador linear seria capaz de separar corretamente as amostras. Provavelmente, um classificador linear qualquer geraria a hipótese ilustrada pela Figura 1(b). Neste caso, cerca de metade das amostras seria classificada corretamente. Entretanto, separando essas mesmas amostras em dois grupos, seria possível treinar dois classificadores lineares que, combinados, poderiam ser

capazes de classificar corretamente um número muito maior de amostras (Figura 1(c)).

Figura 1: (a) Representação de uma base de dados com amostras de duas classes não separáveis linearmente. (b) A linha diagonal ilustra um classificador linear treinado sobre a base original. (c) A linha vertical indica a separação das amostras em dois grupos e as linhas diagonais representam os classificadores lineares gerados para cada um deles.



Objetivos e contribuições

Este trabalho propõe uma abordagem chamada de *P2C - Partitioning to Classify*, baseada na estratégia de divisão-e-conquista (SMITH, 1983; CORMEN et al., 2009), cujo objetivo é encontrar grupos compostos por amostras que tenham maior possibilidade de serem linearmente separáveis. Inicialmente, as amostras de cada classe são separadas em subconjuntos, de modo que cada subconjunto contenha amostras de uma única classe. Em seguida, os subconjuntos são particionados, formando *clusters* que poderão ser unidos na etapa onde os *clusters* das diferentes classes tornam-se candidatos à união em grupos que serão utilizados na etapa de treinamento dos classificadores. Nesta etapa de união, os grupos com amostras muito próximas ou sobrepostas são mesclados para construir uma partição única, onde cada grupo pode conter amostras de uma ou mais classes. Ao final, para cada grupo, um classificador linear é treinado e, posteriormente, combinado com os demais com o intuito de obter melhor poder preditivo.

Uma das vantagens dessa abordagem está no fato de que a formação desses subconjuntos menores permite que a etapa de treinamento seja muito menos custosa dentro de cada grupo. Com isso, é possível construir um classificador robusto, a partir da combinação de classificadores lineares, com uma eficiência superior ou equivalente à obtida pelos métodos de classificação não-lineares. Além disso, é possível que a etapa de treinamento seja totalmente paralelizada, permitindo um enorme ganho de desempenho (FRADKIN, 2008; FU; ROBLES-KELLY; ZHOU, 2010).

Organização

O conteúdo deste trabalho está dividido da seguinte forma:

- O Capítulo 1 apresenta uma breve introdução sobre métodos de classificação lineares e não-lineares, citando suas principais características, destacando as suas vantagens e desvantagens;
- No Capítulo 2, os principais trabalhos da literatura são apresentados com o fim de exemplificar como a combinação de classificadores pode ser feita e demonstrar as principais motivações que conduziram à criação de uma nova técnica de classificação, proposta neste trabalho;
- O Capítulo 3 detalha a proposta desenvolvida, destacando cada uma das etapas da técnica P2C através de exemplos e ilustrações;
- O Capítulo 4 discorre sobre os experimentos realizados em diversas bases de dados e os resultados obtidos através da aplicação da técnica proposta neste trabalho.
- Por fim, são apresentadas as conclusões obtidas, as limitações da técnica proposta e os trabalhos futuros.

1 Classificadores lineares e não-lineares

No contexto de aprendizado de máquina, uma amostra corresponde a uma instância dos dados, geralmente representada por um vetor, cujos valores de cada posição correspondem às características da amostra (atributos) (MOHRI; ROSTAMIZADEH; TALWALKAR, 2012). O número de atributos determina a dimensionalidade da base de dados, portanto, uma base de dados pode ser denominada como unidimensional, bidimensional ou multidimensional, de acordo com o número de atributos utilizados para representar cada amostra.

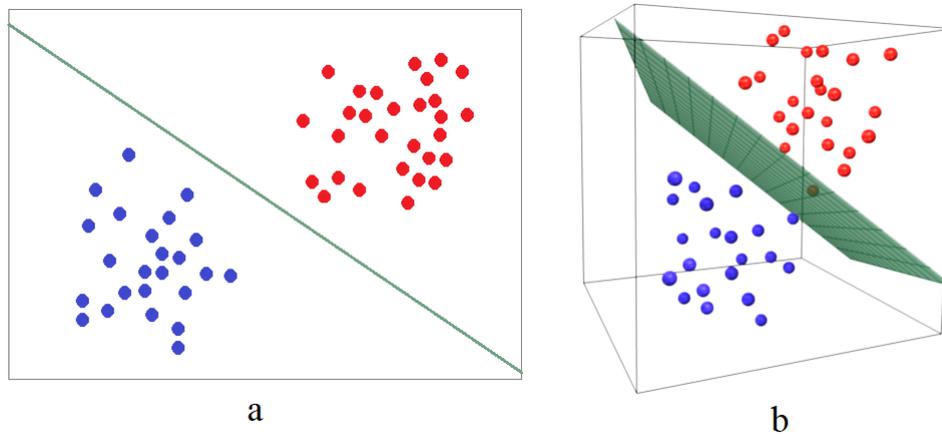
Em problemas de classificação de dados, além dos atributos, cada amostra também possui um rótulo, que representa uma classe ou categoria na qual a amostra está inserida. Um método de classificação, portanto, é um método capaz de identificar as características mais significativas de um conjunto de amostras e construir uma hipótese de classificação, de modo que, dada uma nova amostra, ela seja capaz de identificar a qual classe ela pertence.

1.1 Classificação linear

Um classificador linear é uma das ferramentas mais simples que podem ser utilizadas para classificar as amostras de um conjunto de dados. O objetivo de um método de classificação linear é encontrar uma função linear capaz de separar as amostras em suas respectivas classes, de acordo com as suas características, construindo assim, o classificador. Considerando, por exemplo, uma base de dados bidimensional e com duas classes linearmente separáveis, como na Figura 2(a), é possível encontrar uma reta que divida as amostras em duas classes. Portanto, um método de classificação linear, aplicado sobre essa base de dados, encontraria uma função linear capaz de produzir essa reta. No caso da Figura 2(b), uma base de dados tridimensional, com duas classes, foi utilizada demonstrando que a classificação linear também pode ser feita através de um hiperplano que separa as amostras no espaço.

Existem diversos métodos que podem ser utilizados para encontrar um classificador linear: Bernoulli Naïve Bayes (LANGLEY; IBA; THOMPSON, 1992; MCCALLUM; NIGAM, 1998), Regressão Logística (FACELI et al., 2011; CRAMER, 2002), SVM (com *kernel* linear) (VAPNIK; LERNER, 1963; CORTES; VAPNIK, 1995; BOSER; GUYON; VAPNIK, 1992). Embora esses métodos utilizem abordagens diferentes, todos possuem um mesmo objetivo: encontrar uma função linear que separe as amostras em suas classes, com a menor taxa de erro possível. Normalmente, essa taxa de erro é calculada por uma função denominada *função de custo*, cujo valor resultante corresponde ao erro quadrático

Figura 2: Classificadores lineares em duas bases de dados: (a) bidimensional – uma reta separa as amostras e (b) tridimensional ou multidimensional – um hiperplano separa as amostras



médio da distância entre os pontos calculados pela função linear e as amostras utilizadas para o treinamento. O algoritmo de classificação calcula a função linear iterativamente, durante a etapa de treinamento, variando os coeficientes dessa função, visando minimizar o valor resultante da função de custo a cada iteração.

1.2 Classificação não-linear

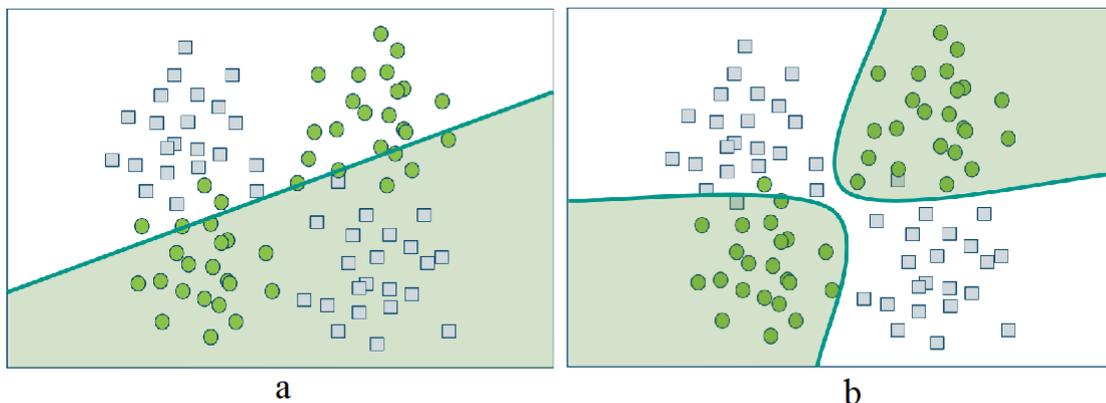
Embora os métodos de classificação lineares consigam obter bons resultados em diversos cenários, é importante notar que caso a base de dados possua amostras de classes que não sejam linearmente separáveis, um classificador linear não conseguirá realizar a classificação de forma eficaz. Além disso, observando os trabalhos publicados ao longo dos anos, é possível notar que a maior parte das bases de dados reais, utilizadas em problemas de classificação, não possuem classes linearmente separáveis (WOODS; KEGELMEYER; BOWYER, 1997; BRITTO; SABOURIN; OLIVEIRA, 2014; LOCHTER; ZANETTI; ALMEIDA, 2015). A necessidade da criação de métodos de classificação mais robustos tornou-se cada vez mais evidente e, a partir disso, começaram a surgir métodos de classificação não-lineares, capazes de gerar hipóteses mais complexas que oferecem melhor separação entre as classes. São exemplos de métodos não-lineares: SVM com *kernel* radial e polinomial (VAPNIK; LERNER, 1963; CORTES; VAPNIK, 1995; BOSER; GUYON; VAPNIK, 1992).

A Figura 3 ilustra uma base de dados com classes cujas amostras não são linearmente separáveis. Na Figura 3(a) é possível observar a aplicação de um método de classificação linear sobre esses dados, indicando que este é capaz de classificar corretamente cerca de metade das amostras. Por outro lado, o classificador não-linear, ilustrado na Figura 3(b),

é capaz de classificar corretamente quase todas as amostras através de uma fronteira de decisão não-linear.

A principal diferença entre um método linear e um não-linear está na função de classificação produzida durante a etapa de treinamento. Em uma base de dados bidimensional, como a que foi utilizada, um classificador linear tenta encontrar uma função linear que produza uma reta capaz de separar as amostras em suas classes (Figura 3(a)). Por outro lado, um classificador não-linear busca encontrar uma função não-linear para classificar as amostras e, desse modo, consegue produzir curvas que representam, com maior precisão, a estrutura das classes das amostras contidas na base (Figura 3(b)).

Figura 3: Base de dados bidimensional com classes não separáveis linearmente. Observa-se que em (a) o classificador linear não consegue separar as amostras corretamente, mas em (b) o classificador não-linear é capaz de separar as amostras de cada classe com maior precisão.



1.3 Comparação entre métodos lineares e não-lineares

A partir da descrição das seções anteriores, é possível destacar as características dos métodos e estabelecer uma comparação entre suas principais vantagens e desvantagens, sintetizadas na Tabela 1.

Os métodos de classificação não-lineares são capazes de produzir soluções mais robustas, uma vez que as funções não-lineares conseguem descrever as características dos dados com maior precisão. Entretanto, a complexidade da implementação desses métodos pode tornar o seu uso inviável em bases de dados muito grandes e com muitos atributos visto que, nestes casos, o tempo de processamento, durante a etapa de treinamento, pode ser muito grande. Assim, torna-se necessário avaliar a relação de custo-benefício da utilização desses métodos de acordo com o problema de classificação a ser resolvido.

Os métodos de classificação lineares possuem um baixo custo computacional e, geralmente, demandam pouco tempo de processamento devido a sua simplicidade de

Tabela 1: Principais características dos classificadores lineares e não-lineares

Lineares	Não-lineares
Implementação simples	Implementação complexa
Baixo custo computacional na etapa de treinamento	Alto custo computacional na etapa de treinamento
Tende a produzir um classificador com maior capacidade de generalização	Maior chance de produzir um classificador com <i>overfitting</i>
Pode ter bom desempenho em bases com muitos atributos	Pode ser computacionalmente custoso em bases com muitos atributos
Pode ser útil em problemas de larga escala, devido ao baixo custo computacional	Em bases de dados muito grandes, pode ser inviável devido ao tempo de processamento
Ineficaz em bases cujas classes não são linearmente separáveis	Produz soluções robustas para bases de dados complexas
Podem ser combinados para produzir soluções robustas e computacionalmente eficientes	A combinação de métodos não-lineares pode ser muito custosa computacionalmente

implementação (YUAN; HO; LIN, 2012). Alguns trabalhos demonstram que eles apresentam resultados promissores em bases de dados com muitos atributos como, por exemplo, bases de texto (YUAN; HO; LIN, 2012; LOCHTER; ZANETTI; ALMEIDA, 2015). Além disso, por causa da sua alta eficiência computacional, vários autores identificaram que a combinação desses métodos pode obter resultados promissores em bases de dados complexas, com desempenho semelhante à dos métodos não-lineares (DIETTERICH, 2000; BRITTO; SABOURIN; OLIVEIRA, 2014; LOCHTER; ZANETTI; ALMEIDA, 2015).

2 Combinando agrupamento e classificação

A combinação de métodos de aprendizado supervisionado e não-supervisionado para melhorar o desempenho na tarefa de classificação não é uma ideia nova. Nos últimos anos, diversos trabalhos têm aperfeiçoado essa ideia e mostrado que ela pode trazer resultados promissores nos mais diversos cenários (JACOBS *et al.*, 1991; KUNCHEVA, 2000; HSU; LIN, 2002; KYRIAKOPOULOU; KALAMBOUKIS, 2008; COLETTA *et al.*, 2014). Uma das principais motivações para a utilização dessa abordagem reside no fato de que a construção de um classificador único e com alto poder preditivo para uma base de dados complexa, geralmente, é computacionalmente custosa e, portanto, pode ser uma abordagem inviável para bases de dados grandes (BRITTO; SABOURIN; OLIVEIRA, 2014; ALMEIDA; GALVÃO, 2016).

Uma estratégia que pode ser utilizada para a solução de problemas complexos é a técnica conhecida como *divisão-e-conquista* (SMITH, 1983; SMITH, 1987). Essa técnica consiste, basicamente, em dividir um problema em instâncias menores (subproblemas), encontrar uma solução para cada subproblema e, por fim, combinar todas as soluções para resolver o problema original. O particionamento da base de dados e o tratamento dos seus subconjuntos como instâncias menores do problema de classificação são exemplos de uso da abordagem de divisão-e-conquista (FU; ROBLES-KELLY; ZHOU, 2010; FRADKIN, 2008; BRITTO; SABOURIN; OLIVEIRA, 2014; ALMEIDA; GALVÃO, 2016). O trabalho de Kuncheva (2000), por exemplo, demonstrou que um algoritmo de agrupamento pode ser utilizado para produzir uma partição da base dados, onde podem ser gerados classificadores especializados para cada grupo. Essa abordagem apresentou bons resultados e foi adotada em diversas outras pesquisas subsequentes que, a partir dos classificadores especializados, propuseram métodos de seleção dinâmica (WOODS; KEGELMEYER; BOWYER, 1997; BRITTO; SABOURIN; OLIVEIRA, 2014; KYRIAKOPOULOU; KALAMBOUKIS, 2008) e fusão de classificadores (RUTA; GABRYS, 2005) para construir soluções mais robustas.

2.1 Seleção de classificadores

O conceito de seleção de classificadores consiste, basicamente, no treinamento e seleção de modelos especializados de acordo com o conjunto de dados. Kuncheva (2000) demonstrou que a utilização de múltiplos classificadores, especializados em subconjuntos dos dados pode gerar um desempenho melhor ou igual ao uso de um único classificador para toda a base. Essa demonstração pode ser apresentada, resumidamente, da seguinte maneira: considerando uma base de dados qualquer, um classificador c é treinado para a mesma, obtendo um desempenho D . Em seguida, encontrando uma partição dessa mesma

base, formada por k subconjuntos, é possível treinar classificadores especializados para cada um deles. Caso o desempenho de um desses classificadores seja superior a D , para um determinado subconjunto, ele pode ser utilizado para classificar amostras que pertençam àquele subconjunto. Assim, é possível selecionar o classificador com maior acurácia para cada subconjunto, de modo que o resultado final será, sempre, igual ou melhor em comparação à utilização de um único classificador para toda a base.

Na literatura, também é possível encontrar vários autores que utilizaram a seleção de classificadores em diferentes contextos. Liu e Yuan (2001) realizaram a seleção com o auxílio de um algoritmo de agrupamento. Inicialmente, vários classificadores foram treinados em um mesmo conjunto de dados. Para cada um deles, as amostras classificadas incorretamente foram separadas e agrupadas. Numa etapa posterior, o método correlaciona esses grupos com as classes encontradas pelos classificadores, permitindo que as amostras sejam classificadas com maior acurácia, conforme pode ser observado nos resultados apresentados no trabalho. Um outro exemplo de aplicação foi apresentado no estudo de Saeedian e Beigy (2009), que recorreram à seleção de classificadores para solucionar um problema de classificação de e-mails como *spam* ou não-*spam*. Primeiramente, os e-mails foram agrupados em grupos disjuntos, de acordo com os tópicos abordados. Posteriormente, classificadores especializados foram treinados para cada um dos grupos formados. Os autores demonstraram que foi possível obter resultados superiores à técnica de voto majoritário, que é uma técnica onde a saída dos classificadores é combinada de modo que a saída que for mais comum é atribuída à amostra analisada (RUTA; GABRYS, 2005; LAM; SUEN, 1997). Todavia, apesar dos bons resultados, uma das desvantagens levantadas pelos próprios autores é que o número de agrupamentos utilizados em ambos os métodos foi calculado empiricamente e mantido fixo durante a realização dos experimentos, causando uma perda de generalidade do método.

Outra questão a ser destacada sobre a seleção de classificadores está na forma em que ela pode ser realizada – estática ou dinamicamente – conforme mostram os trabalhos de Britto, Sabourin e Oliveira (2014), Ruta e Gabrys (2005). A seleção estática é aquela onde o melhor classificador ou o melhor *ensemble* de classificadores (LOCHTER; ZANETTI; ALMEIDA, 2015) pode ser selecionado durante a etapa de treinamento. Por sua vez, a seleção dinâmica ocorre quando diferentes classificadores podem ser selecionados a partir das características de cada amostra nova, durante a classificação dos dados (CRUZ; SABOURIN; CAVALCANTI, 2015; KO; SABOURIN; BRITTO JR., 2008). Estas duas formas de seleção são bastante exploradas na literatura e ambas têm apresentado resultados promissores. Entretanto, em algumas ocasiões, a seleção dinâmica pode oferecer uma solução mais robusta e com melhor desempenho do que a seleção estática de classificadores (KO; SABOURIN; BRITTO JR., 2008; CAVALIN; SABOURIN; SUEN, 2013).

Embora existam diversas maneiras de realizar a seleção dinâmica, há várias ca-

racterísticas comuns. Por exemplo, é comum que essas técnicas sejam divididas em três fases (BRITTO; SABOURIN; OLIVEIRA, 2014): (1) Geração, (2) Seleção e (3) Integração. A partir da variação na implementação de cada uma dessas fases é que surgem as diversas técnicas conhecidas atualmente. Conforme apresentado por Britto, Sabourin e Oliveira (2014), os métodos de seleção dinâmica também podem ser identificados de acordo com as formas de medir a competência da classificação, que podem ser através de *ranking* (SABOURIN et al., 1993), acurácia (WOODS; KEGELMEYER; BOWYER, 1997), medidas de probabilidade (GIACINTO; ROLI, 1999), medidas de comportamento (GIACINTO; ROLI; FUMERA, 2000), medidas baseadas em *oracle*¹ (KO; SABOURIN; BRITTO JR., 2008), ou ainda, através de medidas de diversidade, ambiguidade e complexidade, como apresentado no trabalho de Britto, Sabourin e Oliveira (2014).

2.2 Fusão de classificadores

A fusão de classificadores é uma abordagem onde vários classificadores são gerados para uma mesma base de dados e, posteriormente, combinados através de um algoritmo de consenso, que define o modelo final da classificação. Na literatura, esta abordagem é frequentemente referida como um *Ensemble de Classificadores* (DIETTERICH, 2000), e ela tem sido utilizada em diversos trabalhos, devido ao seu bom desempenho na classificação de dados. Como exemplo, pode-se observar os trabalhos de Silva, Hruschka e Hruschka (2014) e de Lochter, Zanetti e Almeida (2015), que obtiveram resultados promissores ao aplicar um *ensemble* de classificadores para realizar análises de sentimentos no Twitter.

O trabalho de Acharya et al. (2011) também se destaca pela utilização de um *ensemble* de classificadores em conjunto com um algoritmo de agrupamento, através de um algoritmo de consenso, chamado de C³E (*Consensus between Classification and Clustering Ensembles*). Esse método assume que amostras com maior similaridade, possivelmente, pertencerão a uma mesma classe. Para cada amostra, um *ensemble* de classificadores calcula a probabilidade existente da amostra pertencer a cada classe. Paralelamente, um *ensemble* de agrupadores gera uma matriz de similaridades, formando grupos com as amostras semelhantes. Por fim, esses dados são combinados para formar uma nova distribuição de probabilidades para as amostras e, desse modo, cada amostra é rotulada de acordo com a probabilidade de pertencer a uma determinada classe. Devido aos bons resultados apresentados, Coletta et al. (2014) utilizou o método C³E, adaptando-o para o problema de análise de sentimentos. A partir dos experimentos, ambos os autores chegaram à conclusão de que a combinação de algoritmos de classificação e agrupamento apresentou uma acurácia maior do que a utilização de algoritmos de classificação isoladamente.

¹ O desempenho do *oracle* é estimado considerando que, se um dos classificadores treinados é capaz de classificar corretamente uma amostra, então é possível o grupo tomar uma decisão correta

Através da utilização de uma abordagem semelhante, [Kyriakopoulou e Kalamboukis \(2008\)](#) utilizaram um método de combinação de classificação e agrupamento de dados em uma aplicação de detecção de *spam*, e obteve resultados promissores. O método proposto foi dividido em três etapas: (1) agrupamento dos dados; (2) expansão: onde o resultado do agrupamento é utilizado para gerar *meta-atributos*²; (3) classificação: etapa final onde os dados expandidos com os meta-atributos são classificados.

2.3 Particionamento por classe

Como pôde ser observado nas seções anteriores, a estratégia de divisão e conquista resultou em diversos resultados promissores para o problema de classificação de dados, a partir da combinação entre técnicas de agrupamento e classificação. No entanto, como mostra o estudo de [Fradkin \(2008\)](#), a maior parte das propostas presentes na literatura não considera que o rótulo das amostras pode ser uma informação útil para gerar grupos melhores na etapa de agrupamento, ou seja, grupos com características que podem ser úteis na etapa de classificação. Assim, ele aplicou agrupamento dentro de cada conjunto de amostras separadas por classe. Essa proposta trouxe vantagens em relação a ideia de [Kuncheva \(2000\)](#), visto que a aplicação do agrupamento interno nas classes, usando o algoritmo *k*-médias, permitiu uma melhor identificação da estrutura das classes a partir da geração de grupos convexos ([FRADKIN, 2008](#)).

Para realizar as etapas de treinamento e classificação dos dados, [Fradkin \(2008\)](#) utilizou grupos gerados na etapa de agrupamento para rotular as amostras. A partir disso, para *K* classes, com *k* *clusters* gerados em cada uma delas, um único classificador é treinado para toda a base de dados, de modo que ele seja capaz de identificar todos os $k \times K$ novos rótulos. Seus experimentos demonstraram que esses rótulos foram identificados com maior acurácia pelos classificadores lineares do que com os rótulos originais, melhorando, portanto, o desempenho geral da classificação.

2.4 Particionamento, união e classificação

Observando os trabalhos apresentados, foi possível notar os principais pontos fortes e fracos de cada um deles e com isso propor uma nova abordagem para aumentar o poder preditivo dos métodos de classificação atuais. O resultado positivo apresentado por [Fradkin \(2008\)](#) foi uma das motivações decisivas para a adoção de uma abordagem semelhante proposta neste trabalho. Contudo, é necessário ressaltar que a técnica proposta por ele utiliza toda a base de dados como entrada única para o treinamento de um método capaz de lidar com múltiplas classes. Assim, com um método de classificação linear (no caso,

² Os *meta-atributos* são, basicamente, pesos dados a cada amostra de acordo com a quantidade de mensagens por grupo

o SVM Linear), múltiplos classificadores são gerados e combinados através da técnica *one-vs-all*. No final, é obtido um classificador para cada classe da base de dados (FU; ROBLES-KELLY; ZHOU, 2010).

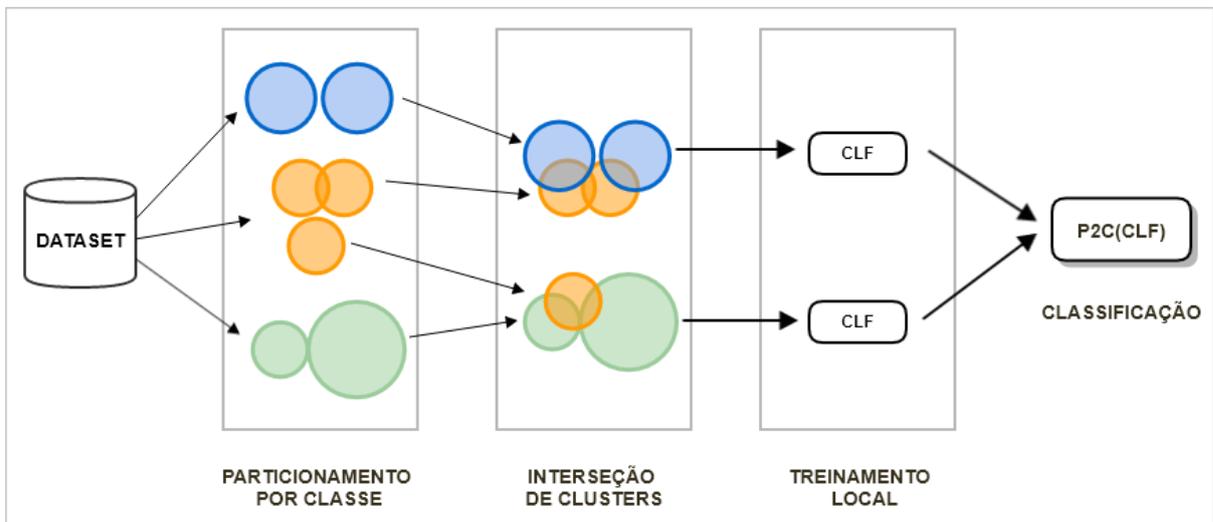
A técnica proposta neste trabalho também emprega a estratégia de divisão-e-conquista através do agrupamento dentro das classes, no entanto, diferente da abordagem de Fradkin (2008), os grupos formados não são utilizados para rotular as amostras. Ao invés disso, a etapa posterior ao agrupamento realiza a união entre os grupos das partições formadas, mantendo os rótulos originais e produzindo uma única partição final com diversos subconjuntos da base completa. O treinamento é realizado em cada subconjunto gerando classificadores lineares especializados. A abordagem proposta ainda produz vários classificadores no final, entretanto, como a quantidade de amostras em cada subconjunto é substancialmente reduzida, a etapa de treinamento tende a ser muito mais rápida (FRADKIN, 2008; BOLEY; CAO, 2004). Além disso, tanto a etapa de treinamento quanto de união podem ser paralelizadas para assegurar maior eficiência.

A proposta deste trabalho, portanto, incorpora as principais vantagens presentes nos trabalhos correlatos existentes na literatura, uma vez que a estratégia de agrupamento dentro das classes é utilizada em conjunto com as estratégias de geração de classificadores especializados e seleção de modelos. Com isso, espera-se que classificadores lineares treinados com diferentes visões localizadas das amostras possam, em conjunto, oferecer desempenho similar ao de um classificador não-linear, com a vantagem de ser mais eficiente em termos de complexidade computacional, além da possibilidade de paralelizar as etapas mais custosas.

3 P2C - *Partitioning to Classify*

A técnica proposta neste trabalho, nomeada de *Partitioning to Classify* (P2C) foi baseada na estratégia de divisão-e-conquista com o objetivo de encontrar grupos compostos por amostras que tenham maior possibilidade de serem linearmente separáveis. Ela pode ser dividida em quatro etapas, ilustradas no diagrama da Figura 4 e detalhadas nas seções seguintes: (i) particionamento por classe, (ii) união de *clusters*, (iii) treinamento local e (iv) classificação. Para ilustrar estas etapas, uma base de dados bidimensional foi gerada, com amostras de duas classes não separáveis linearmente, representada pela Figura 5. No final deste capítulo, a Figura 13 ilustra todas as etapas da técnica P2C, oferecendo uma visão geral da proposta apresentada.

Figura 4: Diagrama ilustrando as etapas principais da técnica P2C. A sigla *CLF* representa um classificador linear e $P2C(CL F)$ representa a combinação final destes classificadores a partir da técnica P2C.



3.1 Particionamento por classe

A primeira etapa da técnica P2C consiste na separação das amostras em subconjuntos de acordo com os seus rótulos (classes), seguida pela aplicação de um algoritmo de agrupamento sobre os dados de cada subconjunto. O objetivo desse procedimento é tentar identificar estruturas particulares de cada classe, que não seriam encontradas caso o agrupamento fosse aplicado sobre a base toda. A Figura 6 ilustra o resultado da aplicação de um algoritmo de agrupamento nos dois subconjuntos da base de dados ilustrada na Figura 5.

Figura 5: Base de dados artificial utilizada para exemplificar as etapas da técnica P2C proposta neste trabalho. A base é bidimensional e contém amostras de duas classes não separáveis linearmente, representadas pelos círculos e quadrados.

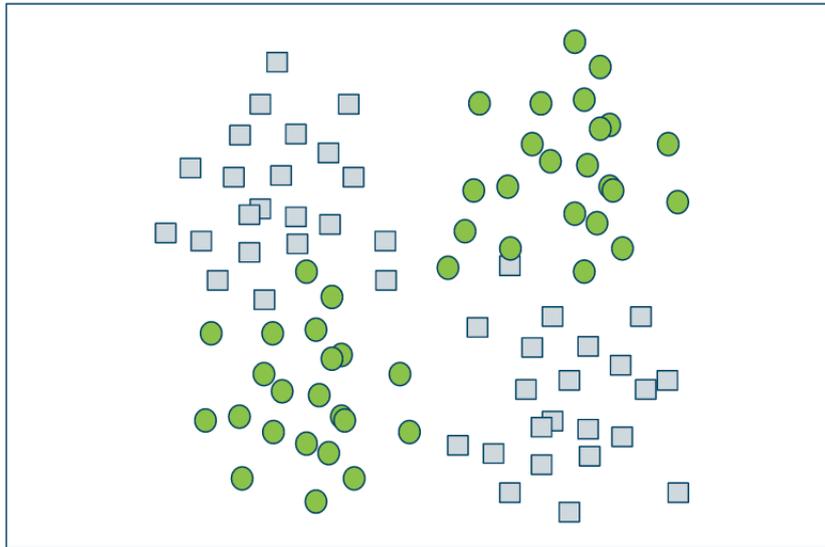
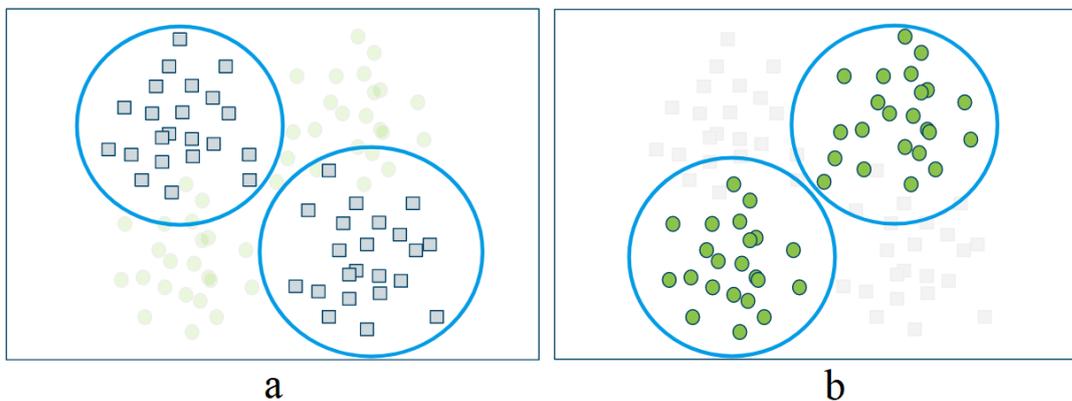


Figura 6: Particionamento por classe da base de dados ilustrada na Figura 5. Cada círculo representa um *cluster* de amostras encontrado pelo algoritmo de agrupamento. Neste exemplo, foram encontrados dois *clusters* em cada uma dos subconjuntos, formando duas partições.



O algoritmo de agrupamento escolhido para esta etapa foi o tradicional k -médias, por ser simples e normalmente apresentar bom desempenho, além de ser um dos mais indicados e utilizados na literatura (ARTHUR et al., 2007; FRADKIN, 2008). Por ser não determinístico, o k -médias deve ser executado várias vezes, de modo que o resultado selecionado será aquele que apresentar o menor erro quadrático médio da distância das amostras em relação ao centroide mais próximo (ARTHUR et al., 2007). O parâmetro k (quantidade de *clusters*) pode ser determinado pela técnica do cotovelo, que consiste em variar o valor de k em um intervalo e em seguida, escolher o valor que minimiza

significativamente a soma do erro quadrático médio, ou seja, o maior valor de k que adiciona no erro médio uma variância maior que um certo limiar.

Após a aplicação do algoritmo de agrupamento, é possível que existam *clusters* com um número muito pequeno de amostras, que precisam ser tratados para evitar o desbalanceamento na etapa de treinamento. Um limite mínimo de amostras por *cluster* foi definido como um parâmetro para evitar esse problema. Assim, os *clusters* considerados pequenos são unidos aos *clusters* mais próximos, de acordo com a distância euclidiana entre seus respectivos centroides.

3.2 União de clusters

Seja K o número de classes em uma determinada base de dados. Após a aplicação do algoritmo de agrupamento sobre as amostras de cada classe, são produzidas K partições $\mathcal{P} = \{p_1, \dots, p_K\}$. Uma partição p_i possui k_i *clusters* disjuntos $\{c_{i_1}, \dots, c_{i_{k_i}}\}$. A etapa de união tem o objetivo de identificar e unir *clusters* de diferentes classes cujas amostras estejam sobrepostas, para comporem o conjunto de treinamento de um mesmo classificador. Com isso, espera-se obter um classificador linear especializado na predição de amostras que estão localizadas em uma região de difícil separação (com sobreposição).

A união entre os *clusters* de partições de cada classe é feita em pares, comparando todos os *clusters* de uma partição com os *clusters* de outra. A união de dois *clusters* $c_{i_j} \in p_i$ com $c_{l_m} \in p_l$, sendo que $p_i \neq p_l$ e $p_i, p_l \in \mathcal{P}$, é realizada de acordo com um fator de união L que é calculado a partir dos valores de sobreposição S e proporção P entre cada par de *clusters* c_{i_j} e c_{l_m} .

3.2.1 Sobreposição de amostras

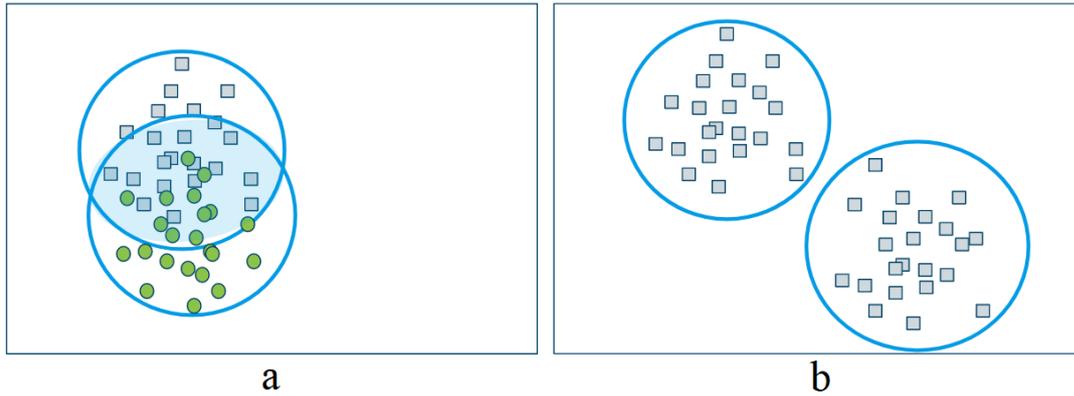
A *sobreposição* (S) entre dois *clusters* quaisquer c_i e c_l corresponde à razão entre a quantidade de amostras sobrepostas e o total de amostras $|c_i \cup c_l|$ (Equação 1). Para determinar a quantidade de amostras sobrepostas, utiliza-se o raio de maior comprimento de cada *cluster*, isto é, a distância euclidiana entre o centroide e a amostra do *cluster* que está mais distante dele. Isto quer dizer que, cada uma das amostras de c_i que esteja dentro do raio de maior comprimento de c_l é contada como sobreposta, e vice-versa.

$$S_{c_i, c_l} = \frac{\text{n}^\circ \text{ de amostras sobrepostas}}{|c_i \cup c_l|} \quad (1)$$

A Figura 7 ilustra dois exemplos de sobreposição de *clusters*. Caso o número de amostras sobrepostas entre os *clusters* c_i e c_l seja muito grande, o valor da sobreposição tenderá a 1 ($S_{c_i, c_l} \rightarrow 1$), conforme apresentado na Figura 7(a), onde o valor da sobreposição é $S_{c_1, c_2} = 25/44 = 0,57$. Por outro lado, caso o número de amostras sobrepostas seja

muito pequeno o valor de S_{c_i, c_l} tenderá a 0 ($S_{c_i, c_l} \rightarrow 0$), e caso não haja nenhuma amostra sobreposta, como na Figura 7(b), o valor de S_{c_i, c_l} será igual a 0.

Figura 7: Ilustração da etapa de sobreposição de amostras. Os círculos ao redor das amostras representam os *clusters*. (a) Exemplo de dois *clusters* com alta taxa de sobreposição ($S_{c_i, c_l} \rightarrow 1$). (b) Exemplo de *clusters* cujas amostras que não possuem sobreposição ($S_{c_i, c_l} = 0$)



3.2.2 Proporção entre *clusters*

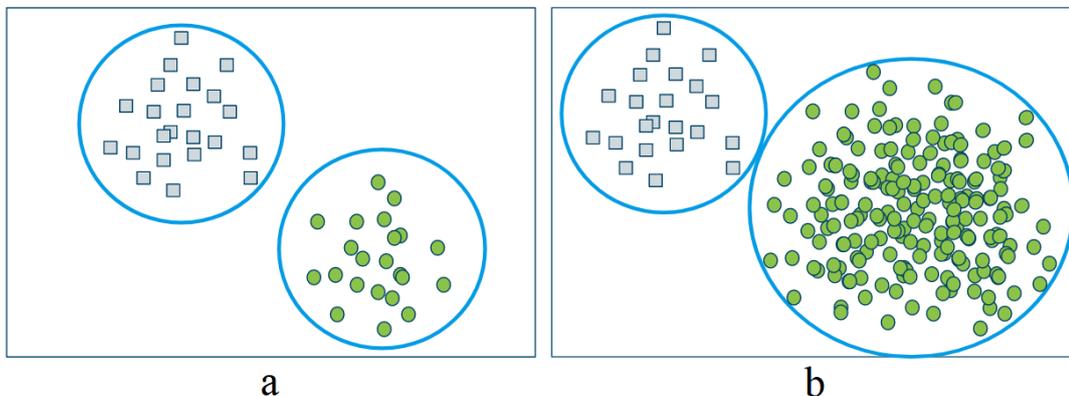
A *proporção* (P) entre dois *clusters* quaisquer corresponde à razão entre o tamanho normalizado (TN) do *cluster* menor (com menos amostras) pelo *cluster* maior (com mais amostras) (Equação 2). O tamanho normalizado (TN) de um *cluster* se refere ao número de amostras do *cluster*, normalizado entre 0 e 1, de modo que o *cluster* com mais amostras possua TN igual a 1. Com isso, se os *clusters* comparados forem proporcionais, o resultado de P será próximo de 1, caso contrário tenderá a 0. Na Figura 8(a) o TN dos *clusters* é igual a 1, pois os grupos são do mesmo tamanho, assim, a proporção será $P_{c_i, c_l} = 1$. No caso da Figura 8(b), o *cluster* menor possui 22 amostras e o *cluster* maior 110, neste caso, o TN deles será 1 e 0,2, respectivamente. Assim, a proporção entre os *clusters* da Figura 8(b) será $P_{c_i, c_l} = 0,2/1 = 0,2$, tendendo a 0 e indicando, portanto, que os *clusters* são desproporcionais.

$$P_{c_i, c_l} = \frac{TN \text{ de } |c_i|}{TN \text{ de } |c_l|}, \text{ sendo } |c_i| \leq |c_l| \quad (2)$$

3.2.3 Fator de união

O objetivo do fator de união é identificar os casos onde os *clusters* devem ser unidos. De acordo com este fator, os *clusters* devem ser unidos em um único subconjunto de amostras, que será utilizado na etapa seguinte durante o treinamento do classificador.

Figura 8: Ilustração da proporção entre *clusters*. Os círculos ao redor das amostras representam os *clusters*. (a) Exemplo de dois *clusters* que possuem um número proporcional de amostras ($P_{c_i, c_l} \rightarrow 1$). (b) Exemplo de *clusters* desproporcionais em relação ao número de amostras ($P_{c_i, c_l} \rightarrow 0$)



O fator de união F_{c_i, c_l} entre dois *clusters* c_i e c_l é calculado pela divisão do valor da sobreposição (S_{c_i, c_l}) pela proporção (P_{c_i, c_l}), conforme a Equação 3.

$$F_{c_i, c_l} = \frac{S_{c_i, c_l}}{P_{c_i, c_l}}, \text{ com } P_{c_i, c_l} \neq 0 \quad (3)$$

A Tabela 2 analisa todas as possibilidades de resultados para o fator de união e apresenta os casos considerados para união de c_i e c_l , sendo que $0 \leq S_{c_i, c_l} \leq 1$ e $0 < P_{c_i, c_l} \leq 1$. Conforme citado, valores de S_{c_i, c_l} tendendo a 1, indicam que há muita sobreposição entre as amostras, e valores tendendo a 0 indicam pouca sobreposição. Analogamente, valores P_{c_i, c_l} tendendo a 1 indicam que os *clusters* são proporcionais e valores tendendo a 0, que os *clusters* são desproporcionais.

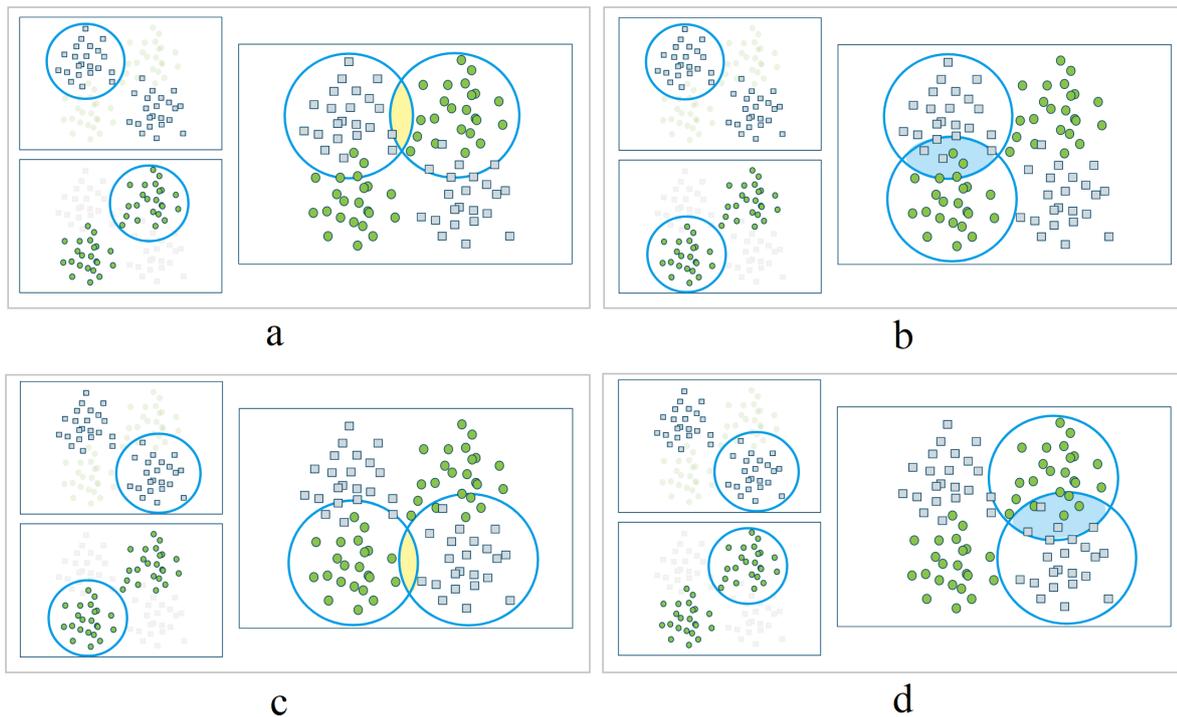
Tabela 2: Possíveis valores para o fator de união. A última coluna indica os casos em que a união é aplicada.

S_{c_i, c_l}	P_{c_i, c_l}	F_{c_i, c_l}	União
$S_{c_i, c_l} = 0$	$P_{c_i, c_l} > 0$	$F_{c_i, c_l} = 0$	Não
$S_{c_i, c_l} \rightarrow 0$	$P_{c_i, c_l} \rightarrow 1$	$F_{c_i, c_l} \rightarrow 0$	Não
$S_{c_i, c_l} \rightarrow 0$	$P_{c_i, c_l} \rightarrow 0$	$F_{c_i, c_l} \rightarrow 1$	Sim
$S_{c_i, c_l} \rightarrow 1$	$P_{c_i, c_l} \rightarrow 1$	$F_{c_i, c_l} \rightarrow 1$	Sim
$S_{c_i, c_l} \rightarrow 1$	$P_{c_i, c_l} \rightarrow 0$	$F_{c_i, c_l} \rightarrow +\infty$	Sim

A partir dos dados da Tabela 2, é possível notar que a união é aplicada sempre que há sobreposição de muitas amostras, ou ainda, quando há alguma sobreposição e o tamanho dos *clusters* é desproporcional (um deles muito maior que o outro). Em todos os

casos, o critério para união é definido por meio de um parâmetro de entrada (γ), que define um limiar de decisão para determinar se a união deverá ou não ser aplicada, $F_{c_i, c_l} \geq \gamma$.

Figura 9: Ilustração referente à etapa de união de *clusters* formados após o particionamento (Figura 6). As figuras de (a)-(d) representam a comparação realizada entre os *clusters*, para determinar se eles devem, ou não, ser unidos. Cada uma delas apresenta duas representações da base: à direita, uma visão da base completa e, à esquerda, uma visão das partições que estão sendo comparadas. Note que os *clusters* destacados em (a) e (c) não serão unidos e os *clusters* destacados nas figuras (b) e (d) serão unidos.

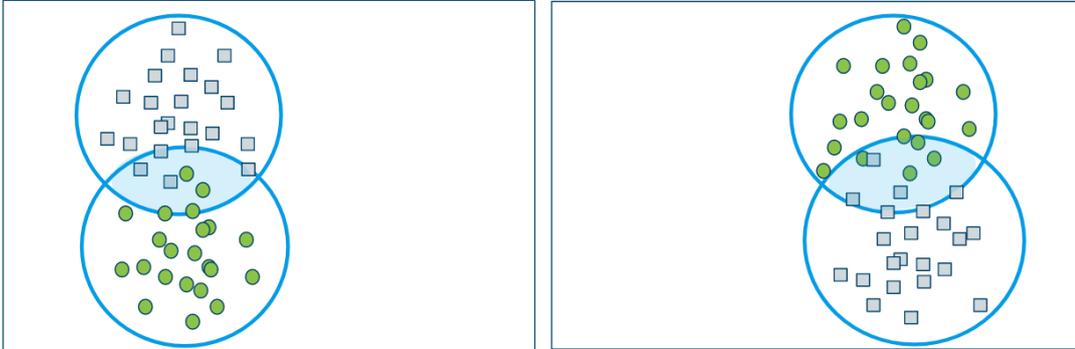


A Figura 9 ilustra como a comparação entre os *clusters* é realizada, considerando as duas partições geradas anteriormente (Figura 6). Cada uma das partes da Figura 9 (a – d), ilustra a comparação entre dois *clusters*, a partir de duas visões diferentes da base de dados, sendo uma visão geral (à direita) e uma visão das partições (à esquerda). Nota-se que, como neste caso os *clusters* são proporcionais ($P_{c_i, c_l} = 1$), a o fator de união foi definido pela sobreposição. É possível observar que os grupos nas Figuras 9(a) e 9(c) não foram unidos, ao contrário, os *clusters* ilustrados nas Figuras 9(b) e 9(d) foram unidos. No final, portanto, dois *clusters* distintos compuseram a partição resultante (Figura 10).

3.3 Treinamento local

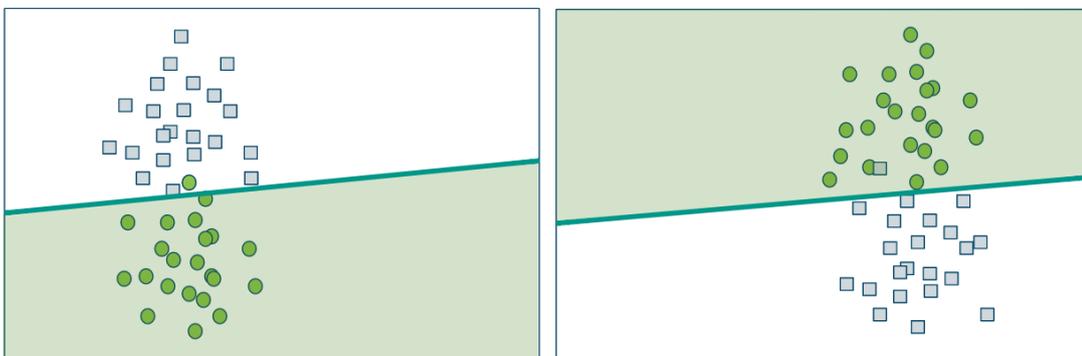
Na etapa de treinamento, um método de classificação é utilizado para gerar um classificador para cada um dos grupos da partição final resultante da etapa de união (Figura 10). O objetivo é produzir um classificador especializado nas amostras de cada

Figura 10: Resultado da união aplicada nos *clusters* das partições ilustradas na Figura 6. Na figura, os dois grupos estão separados mas, na prática, os dois grupos pertencem a uma única partição final.



grupo, e assim, poder oferecer melhor desempenho na predição das classes. Ao final desta etapa, portanto, será gerado um mecanismo de classificação para cada grupo, sejam classificadores lineares (como ilustra a Figura 11) ou a classificação pela classe majoritária.

Figura 11: Ilustração dos classificadores lineares treinados para cada subconjunto da partição final produzida na etapa anterior (Figura 10).



É possível que alguns dos grupos encontrados sejam constituídos por amostras de uma única classe. Neste caso, o próprio grupo torna-se em um classificador óbvio. Assim, se uma nova amostra de teste pertencer a algum desses grupos, ela é classificada como sendo da mesma classe de todas as amostras do grupo.

Caso existam grupos desbalanceados, isto é, onde o número de amostras de uma única classe é muito superior ao das outras classes, um algoritmo de balanceamento é aplicado para adicionar ou remover amostras e assim, balancear o grupo de modo que o treinamento não seja impactado negativamente. Para realizar esse balanceamento, foi utilizada a biblioteca *imbalanced-learn*³ (LEMAITRE; NOGUEIRA; ARIDAS, 2016).

³ Implementação da biblioteca *imbalanced-learn* disponível em: <<https://goo.gl/nfjZMX>>. Acesso em 29/05/2017.

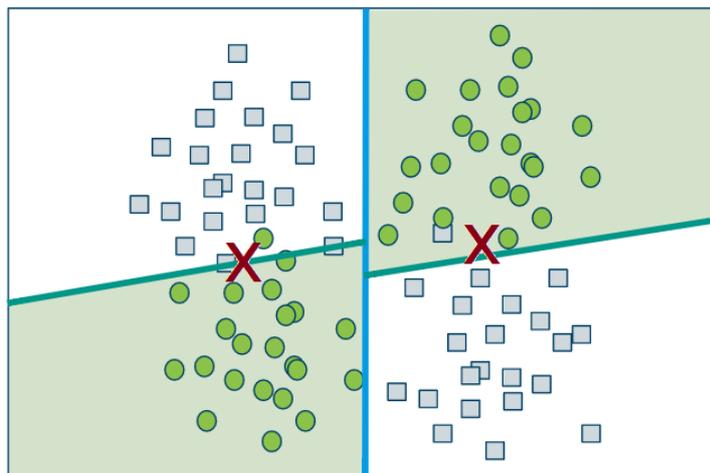
Contudo, caso existam grupos extremamente desbalanceados, de modo que o algoritmo de balanceamento não consiga tratá-los, nenhum classificador é treinado e a classe majoritária do grupo é utilizada para classificar, posteriormente as novas amostras.

3.4 Classificação

A fase de treinamento produz um mecanismo de classificação para cada grupo da partição final. Esse mecanismo pode ser a atribuição direta da classe do grupo ou o emprego de um classificador especializado para o grupo. Na fase de classificação, ao analisar uma nova amostra, primeiramente, é necessário identificar a qual grupo ela pertence. Isto é feito a partir do cálculo da distância euclidiana da amostra em relação aos centroides de cada grupo. Ao encontrar o centroide mais próximo, o mecanismo de classificação associado àquele grupo é utilizado para classificar a nova amostra.

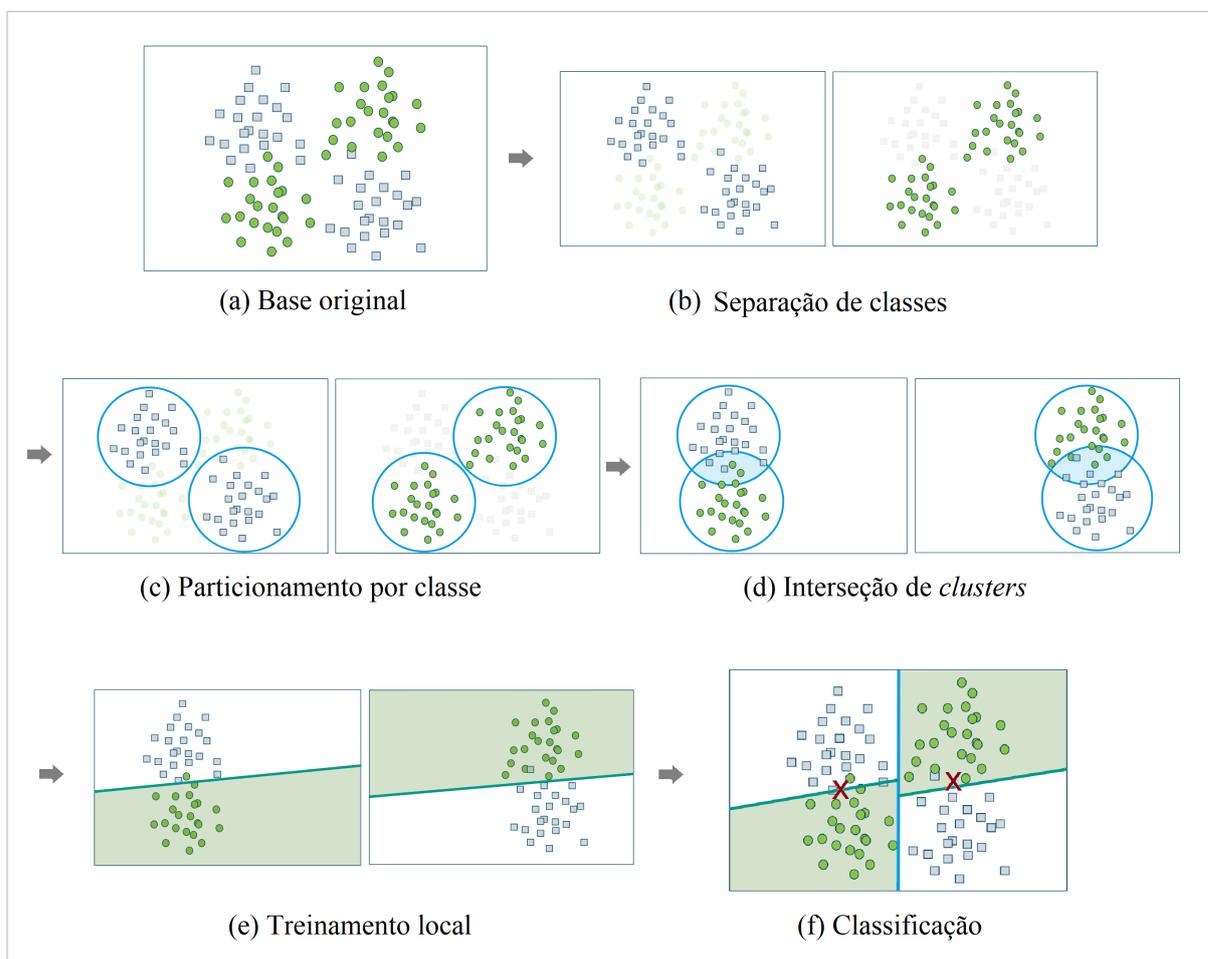
A Figura 12 ilustra os mecanismos de classificação definidos para a partição final, resultante do treinamento aplicado na partição apresentada na Figura 10. É possível notar que existem dois grupos, indicados pelos seus respectivos centroides (indicados por ‘×’). Para cada um deles, um classificador linear foi treinado, de modo que cada nova amostra de teste é associada ao centroide mais próximo e, conseqüentemente, a um dos classificadores especializados.

Figura 12: Modelo de classificação resultante da aplicação da abordagem proposta. A linha vertical representa a divisão entre os grupos da partição (indicados pelos centroides ‘×’) e as linhas diagonais representam os classificadores lineares treinados para cada grupo. Cada amostra nova de teste é classificada a partir do preditor correspondente ao grupo a que ela pertencer.



Por fim, a Figura 13 ilustra cada uma das etapas da técnica proposta, mostrando desde a base original até o classificador final gerado para esta base.

Figura 13: Resumo das etapas da técnica P2C.



4 Experimentos

Os experimentos descritos neste capítulo foram realizados com dois objetivos principais: (1) comparar o desempenho dos classificadores lineares com e sem o uso da técnica P2C, quando aplicada sobre bases de dados cujas amostras não são linearmente separáveis; e (2) investigar se a aplicação da técnica P2C conseguiria obter resultados similares ao desempenho de um classificador não-linear. As seções seguintes descrevem as bases de dados utilizadas, a metodologia empregada nos experimentos e, por fim, uma avaliação dos resultados obtidos.

4.1 Bases de dados

Para assegurar uma avaliação robusta, 11 bases de dados artificiais foram utilizadas⁴. A Tabela 3 descreve as principais características dessas bases de dados plotadas na Figura 14: todas elas são balanceadas e bidimensionais, a maioria possui poucas amostras, cinco delas são binárias e seis são multiclasse.

Tabela 3: Bases de dados utilizadas nos experimentos. Cada linha representa uma base de dados. O “X” indica que a base possui a característica indicada na coluna.

	binária	#amostras > 2K	disponibilizada por
2globs	X	-	Autor
monkey4	X	X	LASID
ds2c2sc13	-	-	LASID
2blobs	X	-	Autor
2bl4cl	-	-	Autor
gauss4cl	-	-	Autor
4bl2cl	X	-	Autor
engytime	X	X	FCPS
3blobs	-	-	Autor
4clu	-	-	Autor
4bl3cl	-	-	Autor

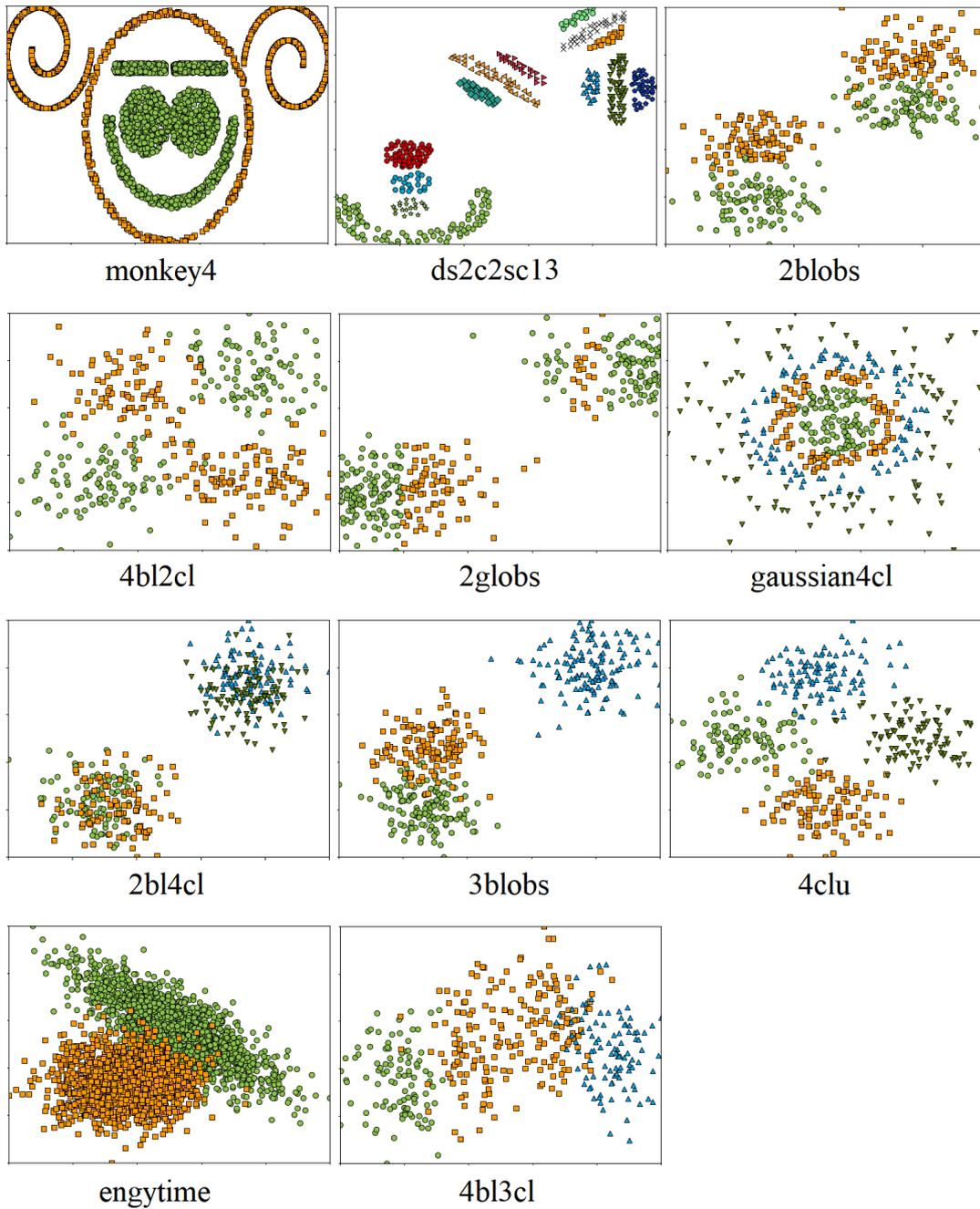
Nota: **LASID**: <<https://goo.gl/iWcQp2>>. Acesso em 28/12/2017.

FCPS: <<https://goo.gl/EMShXB>>. Acesso em 28/12/2017.

Autor: <<https://goo.gl/v214UB>>. Acesso em 28/12/2017.

⁴ O repositório com os códigos fontes e as bases utilizadas estão disponíveis em: <<https://bitbucket.org/nahimsouza/mestrado>>. Acesso em 29/05/2017.

Figura 14: Bases de dados utilizadas nos experimentos e reportadas na Tabela 5. As cores e formas representam amostras de classes distintas.



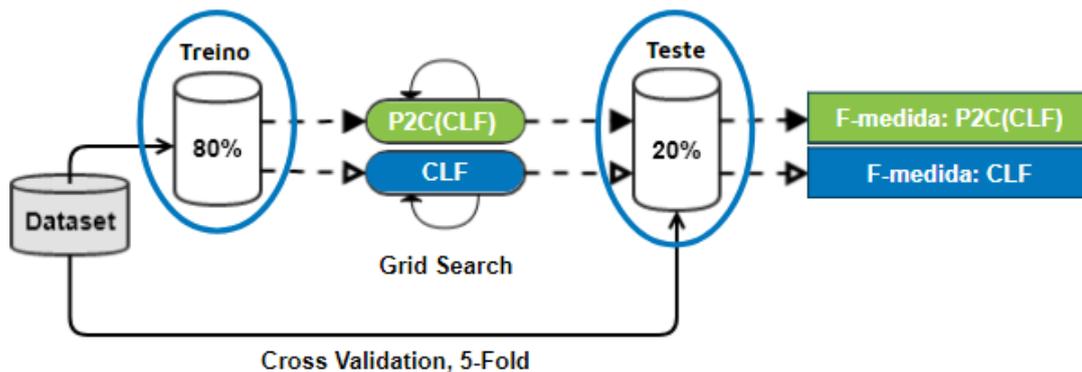
4.2 Metodologia experimental

Conforme mostrado na Figura 15, as etapas dos experimentos foram divididas em duas partes:

1. Realização de um *benchmark* com os métodos de classificação tradicionais (sem a aplicação do particionamento por classe) – representado na figura pela sigla *CLF*;

2. Execução de uma nova rodada de classificação, aplicando a P2C sobre os métodos de classificação lineares – representado na figura pela sigla $P2C(CLF)$.

Figura 15: Metodologia experimental. CLF corresponde a um método de classificação linear tradicional. $P2C(CLF)$ representa a abordagem proposta neste trabalho, que é a aplicação da técnica sobre o classificador linear.



Os seguintes métodos de classificação lineares foram utilizados: *Bernoulli Naïve Bayes* (BNB) (LANGLEY; IBA; THOMPSON, 1992; MCCALLUM; NIGAM, 1998), *Logistic Regression* (LR) (CRAMER, 2002; FACELI et al., 2011), *Linear SVM* (LSVM) (VAPNIK; LERNER, 1963; CORTES; VAPNIK, 1995; BOSER; GUYON; VAPNIK, 1992). Além disso, para verificar se a aplicação da técnica proposta conseguiria obter resultados semelhantes a um método de classificação não-linear, também foi utilizado o método *Radial SVM* (RSVM) (BOSER; GUYON; VAPNIK, 1992), que serviu como *baseline* de comparação.

A implementação dos métodos de classificação foi realizada com a utilização da biblioteca *scikit-learn*⁵ (PEDREGOSA et al., 2011), sendo que os meta-parâmetros foram ajustados através de *grid-search*: o método BNB foi treinado variando o limiar de suavização no intervalo entre 10^{-7} e 10^7 e o limiar de binarização entre 10^{-7} e 10^7 ; os métodos LR e LSVM foram avaliados variando o parâmetro de regularização (C) no intervalo entre 10^{-3} e 10^3 ; já o método RSVM foi avaliado variando os parâmetros C e $gamma$ no intervalo de 10^{-3} a 10^3 . A granularidade da variação dos parâmetros utilizada no *grid-search* foi baseada nos múltiplos de 10, ou seja, $\{10^n, 10^{(n+1)}, 10^{(n+2)}, \dots\}$.

A base de dados foi dividida em 5 fold, de modo que, o experimento fosse repetido 5 vezes, utilizando 80% para o treinamento e 20% para testes, em cada uma das repetições (Figura 15). O objetivo desta abordagem foi garantir que as amostras utilizadas durante o treinamento e o teste dos classificadores fossem as mesmas, permitindo uma melhor comparação entre os resultados do *benchmark*.

⁵ A biblioteca *scikit-learn* está disponível em: <<http://scikit-learn.org/stable/>>. Acesso em 29/05/2017.

Para a etapa de agrupamento, o k -médias foi executado 50 vezes, variando o número de grupos no intervalo entre 2 e 20. O limiar de variância empregado no método do cotovelo foi de 10%. Cada *cluster* resultante possui um número mínimo de 20 amostras, valor definido empiricamente, a fim de evitar a insuficiência de amostras durante o procedimento de validação cruzada no treinamento dos métodos de classificação. Esse número de amostras é verificado após a aplicação do agrupamento, caso um *cluster* possua menos de 20 amostras, ele é unido ao *cluster* mais próximo. O parâmetro de entrada do método referente ao limiar de união foi empiricamente definido como $\gamma = 0,5$.

Como pode ser observado, para a execução da técnica vários parâmetros de entrada precisam ser definidos. É necessário definir o número mínimo de amostras por *cluster* (ou a porcentagem mínima proporcional ao tamanho da base de dados), o limiar utilizado para avaliar o balanceamento dos grupos formados após a união, o número de iterações do k -médias para minimizar o problema do não-determinismo deste algoritmo, a porcentagem de minimização do erro médio quadrático (utilizada na técnica do cotovelo para determinar o número ideal de *clusters*) e, também, o limiar de união (γ), utilizado para avaliar a sobreposição e proporção entre os *clusters*. Estes parâmetros precisam ser ajustados de acordo com a base de dados onde a técnica será aplicada visto que o desempenho da técnica está diretamente relacionado à definição destes parâmetros. A Tabela 4 apresenta os parâmetros que são utilizados e o intervalo de variação utilizado nos experimentos apresentados neste trabalho.

Tabela 4: Parâmetros de entrada necessários para aplicação da técnica P2C. A última coluna mostra o valor escolhido para as bases de dados utilizadas.

Parâmetro	Intervalo de Variação	Valor Utilizado
Nº mínimo de amostras por <i>cluster</i>	{15, 20, 25, 30, 35}	20
% mínima de amostras por <i>cluster</i>	{5%, 10%, 20%}	5%
% Limiar de balanceamento	{15, 16, ..., 24, 25}	19
Nº iterações do k -médias	{20, 30, 40, 50, 100}	50
% ganho do erro quadrático	{10%, 20%, 30%}	10%
Limiar de união (γ)	{0,1, 0,2, 0,3, 0,4, 0,5, 0,6, 0,7}	0,5%

4.3 Análise dos resultados

As bases de dados foram utilizadas com o objetivo de validar o funcionamento da técnica P2C através da análise dos resultados e da avaliação estatística. Além disso, todas as bases utilizadas são bidimensionais, possibilitando a visualização dos dados para uma melhor compreensão de cada etapa da técnica proposta.

A Tabela 5 apresenta os resultados obtidos a partir da aplicação da técnica sobre cada uma das bases de dados. Esses valores foram obtidos a partir do cálculo da média da F-medida em cada um dos 5 fold. O desvio padrão foi de 0,04 e o coeficiente de variação⁶ foi 0,08, indicando que a variação dos resultados foi homogênea. A tabela está dividida em 3 partes: a primeira, mostra a F-Medida obtida pelos 3 métodos de classificação lineares utilizados no *benchmark* (BNB, LR e LSVM); a segunda, mostra o desempenho dos métodos combinados com a técnica P2C (P2C(BNB), P2C(LR) e P2C(LSVM)), que é a proposta deste trabalho; por fim, a última linha mostra o resultado do RSVM, que é um *baseline* de comparação, utilizado para verificar se o desempenho da aplicação da técnica P2C poderia oferecer desempenho próximo ao de um método de classificação não-linear.

A partir da Tabela 5, é possível observar que a combinação dos métodos lineares com o particionamento por classe conseguiu superar o desempenho obtido pelos métodos lineares de classificação em uma das bases, sendo que na maior parte das bases de dados, a aplicação do particionamento por classe obteve um desempenho muito similar ao método não-linear (RSVM). A Tabela 6 mostra a diferença dos valores da F-Medida dos métodos de classificação lineares com e sem a aplicação do particionamento por classe. Um valor positivo indica que houve um aumento no desempenho do classificador e um valor negativo indica que houve uma queda.

A análise dos resultados foi dividida em 6 subseções, que resumem os principais desempenhos obtidos: (i) Aumento do poder preditivo; (ii) Diminuição do poder preditivo; (iii) Equivalência no poder preditivo; (iv) Desempenho superior ao RSVM; (v) Desempenho equivalente ao RSVM; e (vi) Desempenho inferior ao RSVM.

4.3.1 Aumento do poder preditivo

De modo geral, os 3 métodos de classificação linear examinados (BNB, LR e LSVM) obtiveram ganhos de desempenho ao serem combinados com o P2C. Todavia, os resultados obtidos a partir da combinação da técnica P2C com o método BNB foram muito significativos, pois embora este método de classificação não tenha alcançado os melhores resultados, a aplicação do particionamento por classe fez com que o seu desempenho aumentasse visivelmente, em várias bases de dados.

O dois maiores ganhos na F-Medida foram obtidos a partir da aplicação da técnica P2C sobre o método BNB, como pode-se observar na Tabela 6 nas bases `4c1u` (ganho de 0,70) e `ds2c2sc13` (ganho de 0,52). No entanto essa variação ocorreu pelo fato de que o classificador gerado pelo método BNB produziu resultados muito ruins, não separando as classes corretamente, como demonstram as Figuras 16(a) e (b). Para efeito de comparação,

⁶ O coeficiente de variação é a razão entre o desvio padrão e a média, representando a variabilidade dos dados em relação a média. Fonte: <<http://leg.ufpr.br/~shimakur/CE055/node26.html>>. Acesso em 10/02/2018.

Tabela 5: Desempenho dos classificadores nas bases de dados artificiais pela F-Medida (Macro). Em destaque, os melhores resultados para cada base de dados. Para comparação dos resultados com classificadores não-lineares, foi utilizado o SVM com *kernel* radial (RSVM).

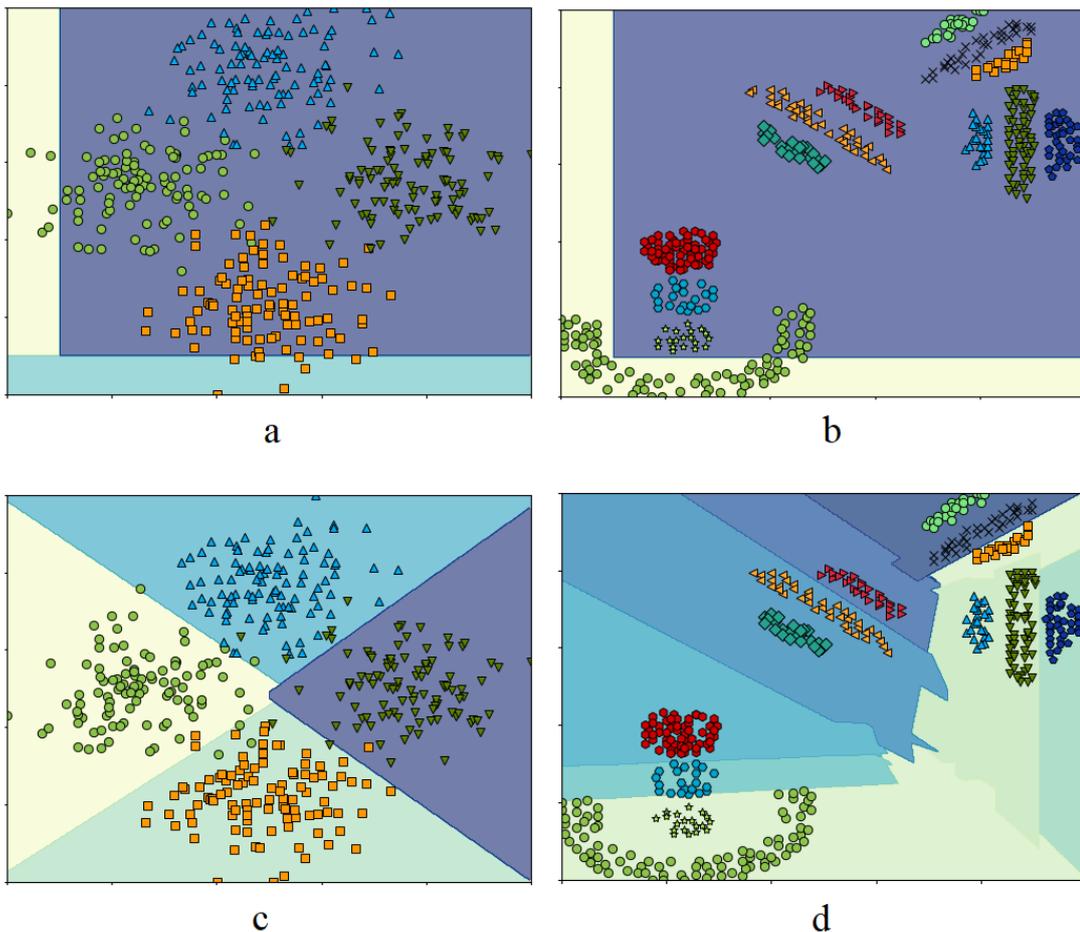
	2globs	monkey4	ds2c2sc13	2blobs	2bl4cl	gauss4cl	4bl2cl	engytme	3blobs	4cln	4bl3cl
BNB	0,41	0,58	0,09	0,45	0,24	0,18	0,39	0,37	0,31	0,24	0,35
LR	0,58	0,60	0,45	0,89	0,58	0,19	0,53	0,94	0,93	0,95	0,89
LSVM	0,41	0,52	0,89	0,85	0,59	0,32	0,63	0,94	0,94	0,94	0,91
P2C(BNB)	0,75	0,94	0,61	0,50	0,39	0,18	0,65	0,37	0,70	0,94	0,38
P2C(LR)	0,84	0,99	0,94	0,89	0,58	0,19	0,91	0,94	0,93	0,94	0,90
P2C(LSVM)	0,86	0,99	0,97	0,93	0,59	0,47	0,85	0,94	0,93	0,95	0,90
RSVM	0,89	1,00	0,98	0,93	0,59	0,63	0,89	0,95	0,93	0,95	0,92

Tabela 6: Diferença dos valores da F-Medida com e sem a aplicação da técnica P2C, ou seja, o quanto o valor da F-Medida aumentou ou diminuiu. Em destaque, os classificadores que obtiveram maior aumento no desempenho depois da aplicação da técnica P2C. O crescimento médio da F-Medida foi de 0,17.

	2globs	monkey4	ds2c2sc13	2blobs	2bl4cl	gauss4cl	4bl2cl	engytme	3blobs	4cln	4bl3cl
P2C(BNB) - BNB	0,34	0,36	0,52	0,05	0,15	0,00	0,26	0,00	0,39	0,70	0,03
P2C(LR) - LR	0,26	0,39	0,49	0,00	0,00	0,00	0,38	0,00	0,00	-0,01	0,01
P2C(LSVM) - LSVM	0,45	0,47	0,08	0,08	0,00	0,15	0,22	0,00	-0,01	0,01	-0,01

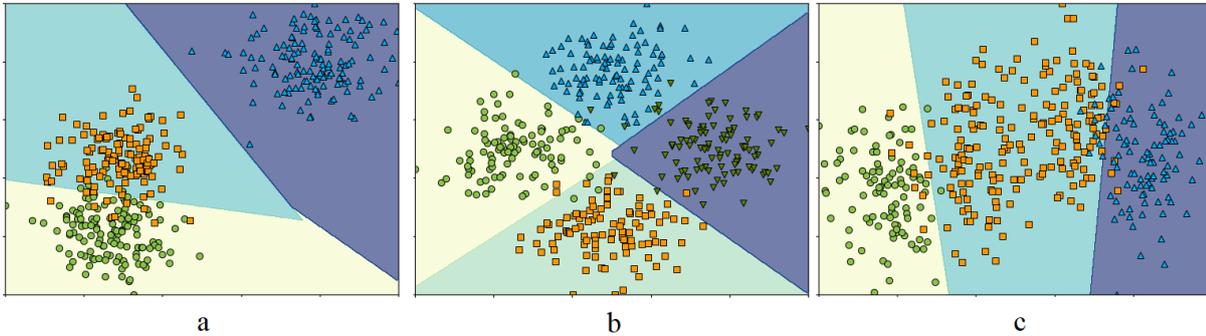
é plotado nas Figuras 16(c) e (d) a margem de separação produzida pelo LSVM.

Figura 16: Classificadores gerados utilizando o método BNB e LSVM. Nota-se que o método BNB obteve um baixíssimo desempenho, uma vez que os classificadores gerados não conseguiram separar as classes corretamente, o oposto do que ocorreu com o método LSVM. (a) *Plot* do classificador BNB aplicado sobre a base 4c1u. (b) *Plot* do classificador BNB aplicado sobre a base ds2c2sc13. (c) *Plot* do classificador LSVM aplicado sobre a base 4c1u. (d) *Plot* do classificador LSVM aplicado sobre a base ds2c2sc13.



Um fato que evidencia ainda mais que as grandes variações ocorreram na F-Medida devido ao mau desempenho do método BNB são os resultados apresentados pelas bases 3blobs, 4c1u e 4b13c1. Nestas bases, apenas o classificador P2C(BNB) conseguiu obter aumento no poder preditivo. Analisando a distribuição das amostras nessas bases, nota-se que as classes estão bem separadas e que a aplicação da estratégia *one-vs-all* foi suficiente para produzir um bom desempenho dos métodos LR e LSVM, visto que os resultados anteriores já haviam atingido resultados maiores ou iguais a 0,9 na F-Medida. A Figura 17 exemplifica este fato, mostrando os classificadores gerados a partir do método LSVM sobre essas bases de dados.

Figura 17: Classificadores gerados utilizando o classificador LSVM com a estratégia *one-vs-all*. (a) *Plot* do classificador LSVM aplicado sobre a base `3b1obs`. (b) *Plot* do classificador LSVM aplicado sobre a base `4c1u`. (c) *Plot* do classificador LSVM aplicado sobre a base `4b13c1`.



4.3.2 Diminuição do poder preditivo

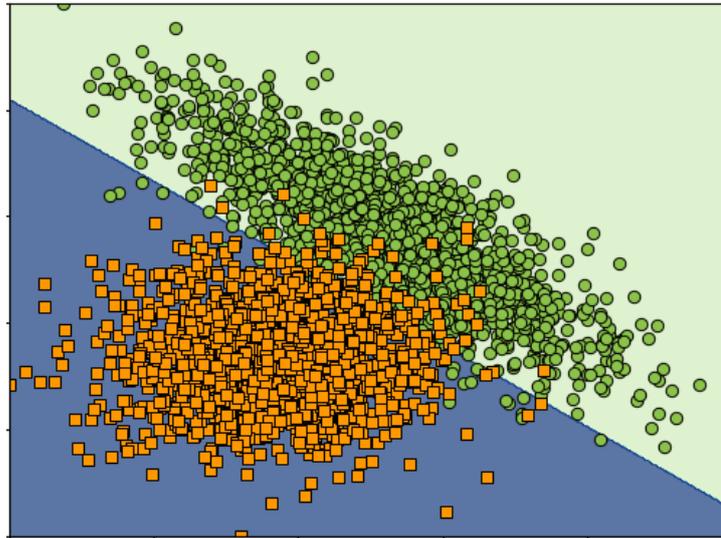
Observando os resultados da Tabela 6, é possível notar que a aplicação do método não causou reduções no desempenho dos métodos de classificação lineares, com exceção da diminuição de 0,01 na F-Medida ocorrida nas bases `3b1obs`, `4c1u` e `4b13c1`. Portanto, foi possível concluir que a aplicação da técnica P2C não reduziu significativamente o poder preditivo dos classificadores em nenhuma das bases de dados analisadas.

4.3.3 Equivalência no poder preditivo

A partir dos valores apresentados na Tabela 6, é possível observar vários casos onde a diferença da F-Medida foi igual a 0,00. Nestes casos, a aplicação da técnica P2C não conseguiu trazer um aumento no desempenho, mas também não causou uma diminuição no mesmo, ou seja, o desempenho do classificador após a aplicação da técnica foi equivalente ao desempenho do classificador linear.

O resultado obtido na base `engytime` destaca-se, pois para todos os métodos de classificação a diferença foi igual a 0,00. Neste caso, dois fatos devem ser mencionados: o primeiro, é que os classificadores lineares obtiveram bons resultados, visto que as amostras, originalmente já eram separáveis linearmente, como mostra a Figura 18; o segundo é que, como as amostras estão muito próximas umas das outras, mesmo após o particionamento por classe, a etapa de união de *clusters* fez com que todos os *clusters* fossem unidos em um único grupo (o mesmo que existia antes da aplicação do P2C). Isto fez com que um único classificador fosse gerado para toda a base, após a aplicação da P2C, resultando em um desempenho idêntico ao dos classificadores lineares.

Figura 18: Aplicação do método LSVM sobre a base de dados `engytime`, demonstrando que mesmo um método linear (sem aplicação do P2C) pode conseguir bons resultados nesta base de dados.



4.3.4 Desempenho superior ao RSVM

Um dos casos que se destacou na Tabela 5 foi o da base `4b12c1`, onde o desempenho da técnica P2C aplicada ao método de regressão logística (LR) foi levemente superior ao do método não-linear RSVM. A Figura 19(a) apresenta o resultado da aplicação do RSVM sobre a base, enquanto a Figura 19(b) apresenta o classificador final produzido com a aplicação da técnica P2C sobre o método LR. Através dessa figura é possível notar que o classificador gerado pela aplicação da técnica (P2C(LR)) é muito similar ao classificador gerado pelo método RSVM, e por isso, obteve um resultado muito semelhante a ele. As etapas da geração deste classificador (Figura 19(b)) podem ser vistas nas Figuras 20 e 21.

Na Figura 20(a) e (b) as duas classes de amostras foram separadas e em seguida o particionamento por classe foi aplicado, produzindo 5 *clusters* nas amostras da Figura 20(a) e 4 *clusters* nas amostras da Figura 20(b). Em seguida, a união entre esses *clusters* foi realizada, gerando os 4 grupos da Figura 20(c). Estes grupos, são aqueles que foram utilizados na etapa de treinamento do classificador.

Nota-se, no entanto, que entre os quatro grupos formados após a união (Figura 20(c)), um deles possui amostras majoritariamente da classe verde, e outros dois, majoritariamente amostras da classe laranja. Desse modo, não foi necessário realizar o treinamento do classificador LR para esses grupos (como se vê nas Figuras 21(b), (c) e (d)), economizando tempo de processamento computacional. Para o outro grupo, no topo da Figura 20(c), que continha amostras das duas classes, um classificador linear foi gerado, separando corretamente a maioria das amostras, como pode ser visto na Figura 21(a). Assim, a partir da combinação desses mecanismos de classificação, foi gerado o classificador

Figura 19: Plot da base de dados 4b12c1. As cores e formas representam amostras de classes distintas. (a) Classificador gerado utilizando o método RSVM sobre esta base. (b) Classificador final produzido com a aplicação da técnica P2C sobre o método LR.

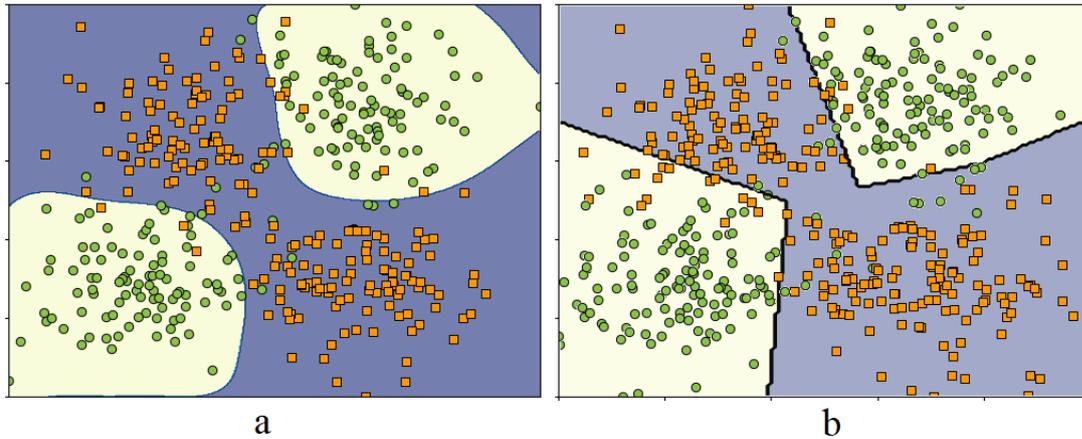
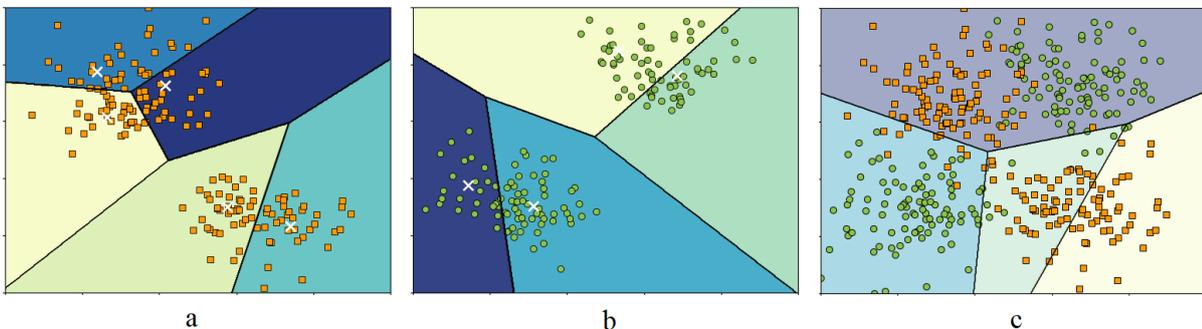


Figura 20: Etapas de particionamento por classe e união de *clusters*. (a) Particionamento realizado nas amostras da classe laranja, gerando 5 *clusters*. (b) Particionamento realizado nas amostras da classe verde, gerando 4 *clusters*. (c) Resultado da união entre os *clusters* de (a) e (b), produzindo 4 grupos, utilizados no treinamento do classificador LR.



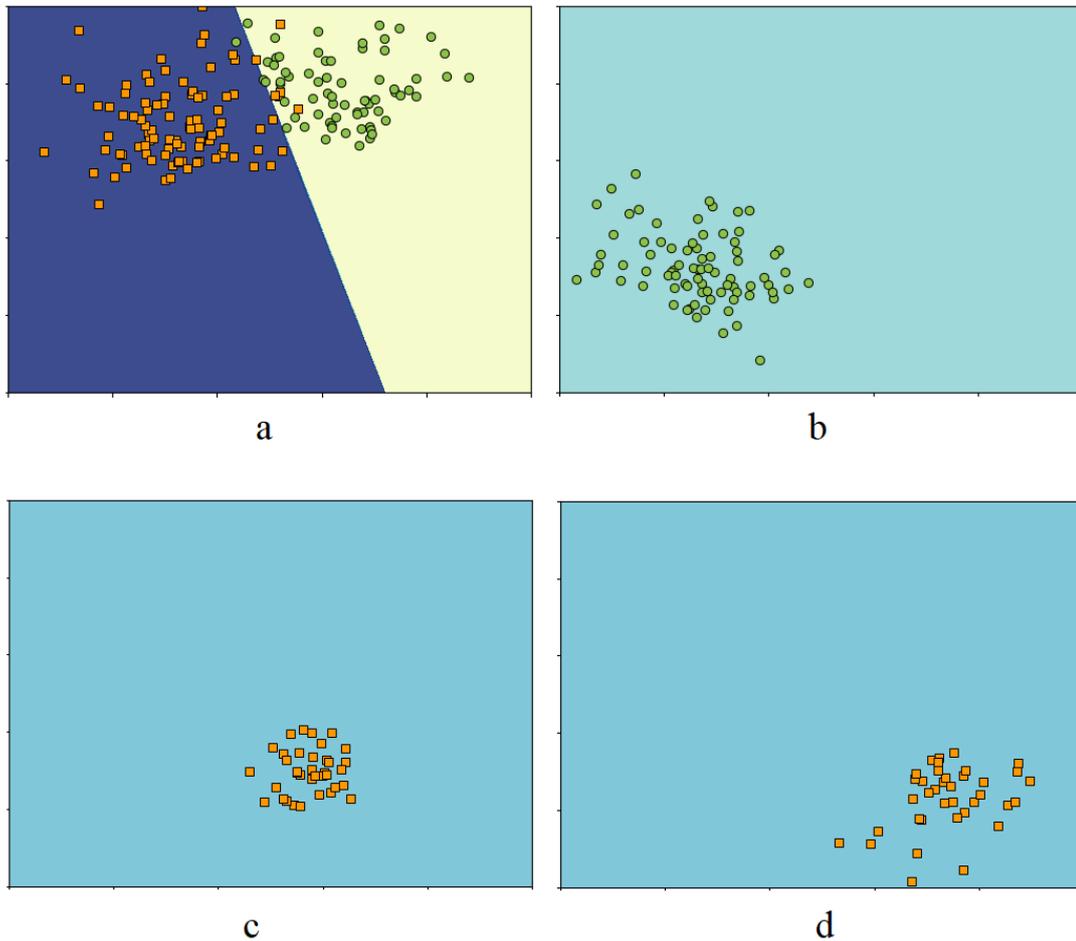
da Figura 19(b), que obteve desempenho ligeiramente superior ao método RSVM.

4.3.5 Desempenho equivalente ao RSVM

Nota-se nos resultados da Tabela 5 que, com exceção da base 4b12c1, os melhores resultados da F-Medida foram obtidos pelo método RSVM. Entretanto, a diferença média da F-Medida entre o método RSVM e a técnica P2C foi aproximadamente 0,02, demonstrando que a variação foi pequena e que, portanto, a técnica P2C conseguiu um desempenho equivalente a um método de classificação não-linear.

Um outro caso que merece destaque é o que ocorreu na base 2b14c1. Neste caso, todos os classificadores obtiveram resultados ruins. Isto ocorreu porque, como demonstra

Figura 21: Classificadores gerados para cada grupo da Figura 20(c), após a união. Em (a) um classificador foi gerado com o método LR. Em (b), (c) e (d) a maior parte das amostras pertenciam a uma única classe, não sendo o treinamento do classificador, pois o mecanismo de classificação é a utilização da classe majoritária no grupo.



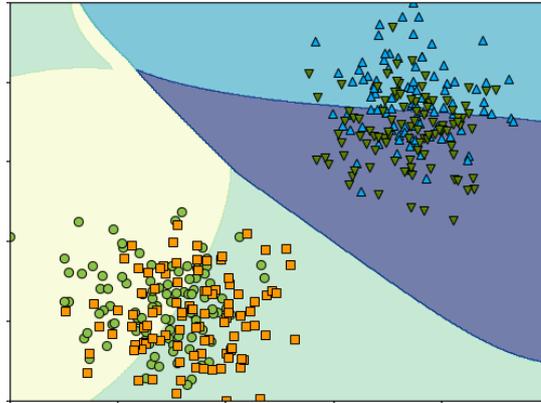
a Figura 22, as amostras de diferentes classes estavam muito mescladas, formando grupos heterogêneos. Desse modo, conclui-se que o resultado da técnica P2C também foi equivalente ao método RSVM, visto que nem mesmo o classificador não-linear conseguiu alcançar um bom desempenho nesta base de dados.

4.3.6 Desempenho inferior ao RSVM

A técnica P2C não ofereceu melhorias significativas quando aplicado na base gauss4cl. Neste caso, foi obtido apenas um pequeno aumento no desempenho do método P2C(LSVM). Além disso, nesta base de dados, o método RSVM conseguiu um desempenho muito superior aos métodos lineares, mesmo após a aplicação da técnica P2C.

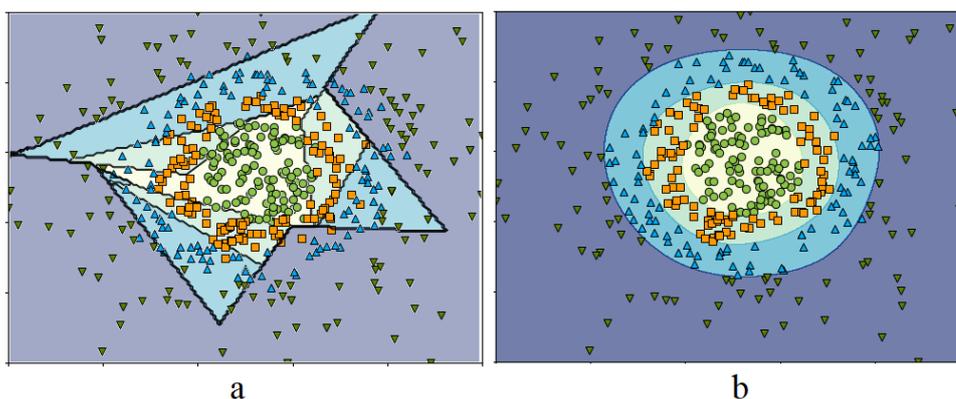
Na Figura 23(a), é possível notar que o classificador linear teve maior dificuldade de

Figura 22: Aplicação do método RSVM sobre a base de dados 2b14c1. Nota-se que o classificador RSVM não conseguiu separar as classes adequadamente, devido à heterogeneidade entre as amostras das diferentes classes.



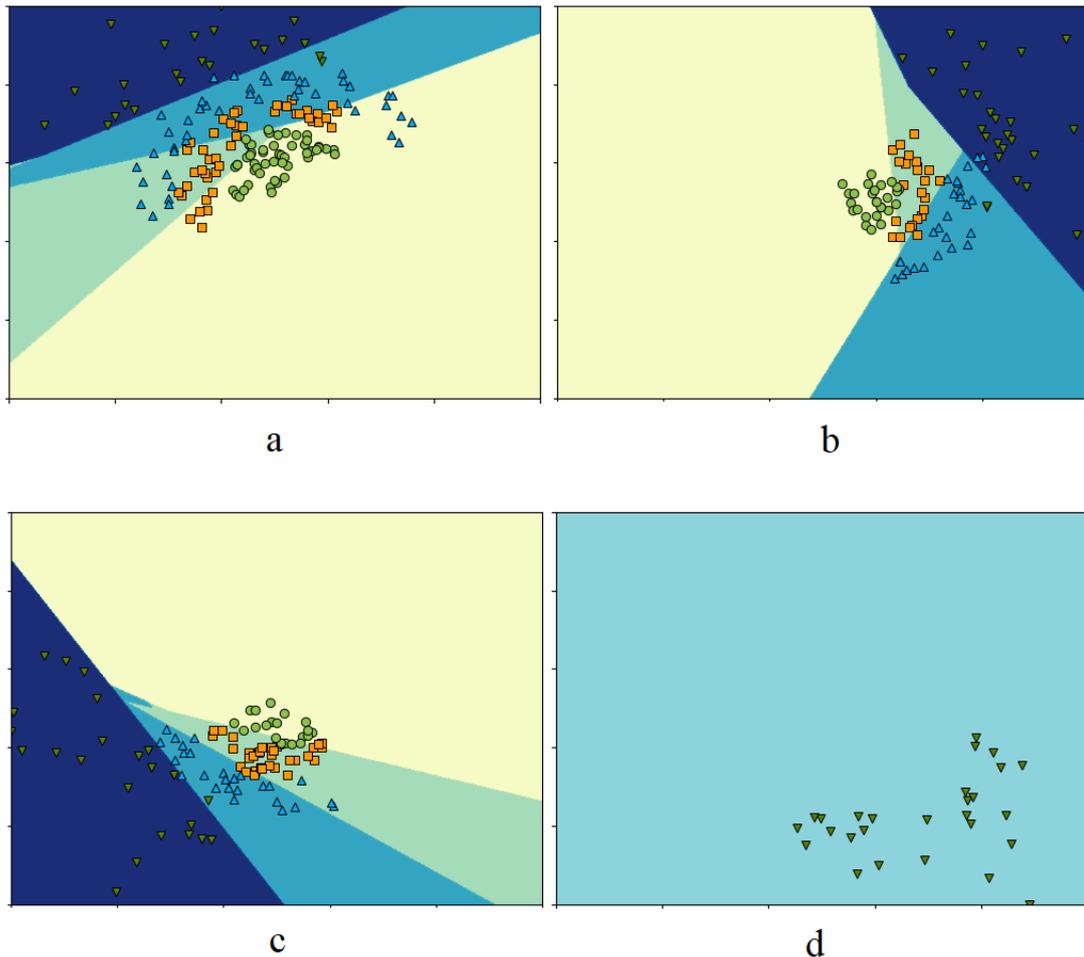
separar as amostras corretamente, enquanto o classificador RSVM, conseguiu classificar corretamente a maior parte das amostras (Figura 23(b)). Além disso, analisando a distribuição das amostras nos grupos, após a etapa de união (Figura 24), observa-se que os grupos formados possuíam amostras de diversas classes, dificultando o bom desempenho dos classificadores lineares.

Figura 23: Aplicação dos classificadores linear e não-linear sobre a base `gauss4c1`. (a) Classificador P2C(LSVM) aplicado sobre a mesma base de dados. (b) Classificador RSVM aplicado sobre a mesma base de dados.



Conforme apresentado no Capítulo 1, a função de decisão gerada por um classificador não-linear consegue representar curvas que um classificador linear não consegue. Portanto, essa característica fez com que o resultado do classificador RSVM fosse superior nesta base de dados.

Figura 24: Grupos formados após a etapa de união na base `gauss4c1`. As figuras (a), (b) e (c) mostram grupos formados com amostras de múltiplas classes, dificultando a classificação realizada por um método linear, que conseguiria um melhor resultado caso houvesse amostras de apenas duas classes. (d) Representa um grupo formado por amostras de uma única classe, cujo mecanismo de classificação será a utilização da classe majoritária como rótulo.



4.4 Avaliação estatística

Ao avaliar o desempenho de uma nova técnica de classificação, como a que foi proposta neste trabalho, é necessário considerar se os resultados obtidos são estatisticamente significantes, ou seja, deve-se garantir que os resultados não foram obtidos ao acaso. Para isso, os resultados obtidos devem ser avaliados através da utilização de *testes de hipótese* (DEMŠAR, 2006). Um teste de hipótese avalia a probabilidade de se cometer um *erro tipo I* (rejeição da hipótese nula quando ela é verdadeira). A hipótese nula, no caso deste trabalho, seria assumir que aumento no desempenho foi obtido ao acaso e não devido à aplicação da técnica proposta (P2C). Assim, o objetivo da avaliação estatística é verificar a probabilidade da ocorrência da hipótese nula e descartá-la, caso seja muito pequena.

Neste trabalho, foram utilizados o Teste de Wilcoxon e o Teste de Friedman para avaliar a significância dos resultados nas bases de dados *artificiais*. O Teste de Wilcoxon é um teste alternativo ao Teste-T, quando as amostras não seguem uma distribuição normal (DEMŠAR, 2006). Este teste foi utilizado para comparar a variação do desempenho aos pares, comparando os resultados com e sem a aplicação do particionamento por classe em cada um dos métodos de classificação avaliados (BNB, LR e LSVM), utilizando dados da Tabela 5. O Teste de Friedman, por sua vez, foi utilizado para avaliar a variabilidade entre os resultados obtidos pelos classificadores lineares reportados na Tabela 5, com e sem a aplicação da técnica P2C.

Para realizar o teste de hipótese, primeiramente, é necessário estabelecer um nível de significância (α), que corresponde à probabilidade de se cometer um erro tipo I. Em seguida, a estatística do teste e o *p-valor* são calculados. O *p-valor* representa a probabilidade de significância, caso o *p-valor* seja menor que o valor de α , o resultado é considerado como significativo e a hipótese nula é rejeitada.

Nos testes realizados, o nível de significância estabelecido foi de 5%, isto é, $\alpha = 0,05$. O aumento médio no desempenho relatado na Tabela 6 foi de 0,17. Tanto o Teste de Wilcoxon, quanto o Teste de Friedman apresentaram resultados indicando que esse aumento no desempenho foi estatisticamente significativo, com *p-valor* $< 0,05$, rejeitando a hipótese nula. Portanto, de acordo com os experimentos, há 95% de chances de se obter um aumento médio de 0,17 no desempenho (F-medida) a partir da aplicação da técnica proposta sobre os classificadores lineares avaliados neste trabalho.

4.5 Considerações Finais

Através dos resultados obtidos e da avaliação estatística realizada foi possível validar a aplicação da técnica de particionamento por classe. Como foi observado, a técnica P2C conseguiu oferecer bons resultados nas bases de dados artificiais, sendo que seu pior resultado foi obtido nas bases onde as amostras estavam muito mescladas, formando grupos heterogêneos onde qualquer classificador, linear ou não-linear, é incapaz de separar as amostras corretamente.

Observando o caso da base `gauss4c1`, onde o classificador RSVM conseguiu melhor resultado do que os classificadores lineares, nota-se a necessidade de uma maior investigação nas etapas de particionamento por classe e união de *clusters*, pois caso os grupos formados possuíssem amostras separáveis linearmente, o resultado poderia ter sido melhor. Dessa forma, é necessário avaliar a forma de se escolher o melhor número de grupos, bem como o método de agrupamento a ser utilizado.

A etapa de união de *clusters* deve ser melhor avaliada pois, em alguns casos – como ocorreu na base `engytime` – a união fez com que os *clusters* fossem unidos em um único

grupo, igual à base original, tornando a aplicação da P2C completamente ineficaz neste caso.

De modo geral, os experimentos demonstraram que existem diversos casos onde a aplicação da técnica proposta tem potencial para oferecer bons resultados. Em trabalhos futuros, as etapas de particionamento e união podem ser melhor avaliadas e ajustadas. Além disso, futuramente, uma melhor avaliação dos cenários pode ser realizada através da investigação de bases de dados reais.

Conclusão

Ao longo dos anos, diversos métodos de classificação têm sido criados e aperfeiçoados para lidar com bases de dados complexas. No entanto, grande parte dos métodos apresentados recentemente demandam muitos recursos computacionais e tempo de processamento, como por exemplo, os métodos de *deep learning* (KRIZHEVSKY; SUTSKEVER; HINTON, 2012; SCHMIDHUBER, 2014). Essas características podem tornar sua aplicação inviável em cenários reais e dinâmicos, principalmente se a capacidade computacional for limitada.

Neste contexto, diversos trabalhos surgiram com o propósito de combinar classificadores “fracos”, porém com baixa demanda de recursos computacionais, para gerar classificadores com maior poder preditivo. Na literatura, diversos autores mostraram que combinação desses classificadores pode oferecer soluções eficientes, com alto poder preditivo, sem aumentar significativamente o esforço computacional (BRITTO; SABOURIN; OLIVEIRA, 2014; KUNCHEVA, 2000; LOCHTER; ZANETTI; ALMEIDA, 2015). Baseado nisso, este trabalho foi proposto com a ideia de combinar classificadores lineares para oferecer um maior poder de predição em bases de dados cujas classes não são linearmente separáveis. O objetivo desta proposta é oferecer um mecanismo de classificação com alto poder preditivo e baixo custo computacional, mesmo em bases de dados grandes, onde a aplicação de métodos não-lineares de classificação pode ser inviável.

A nova técnica apresentada neste trabalho foi denominada *Partitioning to Classify* (P2C), ela é fundamentada na estratégia de divisão-e-conquista e pode ser utilizada em problemas de classificação de dados. Dois fatores principais motivaram a elaboração da proposta: (1) a especialização de classificadores em regiões da base original, permitindo que vários classificadores especializados sejam treinados e combinados para gerar uma solução mais robusta e com maior poder preditivo; e (2) a diminuição do tempo de treinamento do classificador, uma vez que o custo computacional para treinar um classificador em um grupo de amostras é menor do que o custo para treinar um classificador para toda a base.

Em linhas gerais, a técnica proposta consiste em separar as amostras por classe e gerar uma partição para cada um destes subconjuntos. Em seguida, é realizado o procedimento de união de *clusters*, que se resume em comparar os *clusters* de acordo com o seu tamanho e distribuição das amostras e verificar se existe uma união entre os *clusters* de diferentes classes. Ao final desta etapa, são formados diversos grupos, compondo uma única partição. Para cada um destes grupos, um classificador linear é treinado e a combinação desses classificadores especializados resultará em um classificador global.

A etapa de particionamento por classe permitiu que o problema fosse dividido

em instâncias menores e por meio da união de *clusters* foi possível realizar a etapa de “conquista” da estratégia de divisão-e-conquista. Esta segunda etapa permitiu a união de *clusters* cujas amostras estavam próximas, de acordo com o tamanho dos *clusters*. Assim, foi possível formar grupos com um menor número de amostras, permitindo que a etapa de treinamento fosse realizada com um menor esforço computacional e que classificadores especializados fossem gerados para a base de dados.

A partir dos experimentos realizados e análise estatística dos resultados obtidos, concluiu-se que a técnica proposta neste trabalho foi capaz de aumentar o poder preditivo dos métodos lineares de classificação avaliados. Como pôde ser observado, a aplicação da técnica P2C possibilitou que os classificadores lineares apresentassem desempenho próximo ao do SVM Radial para a maioria das bases de dados utilizadas. Além disso, a técnica proposta ofereceu uma nova forma de tratar problemas de classificação de dados por meio da estratégia de divisão-e-conquista. Fato este que pode ser de alta relevância em cenários reais, nos quais as bases de dados podem ser enormes e intratáveis por métodos não-lineares.

Limitações

As principais limitações da técnica P2C estão relacionadas aos parâmetros de entrada utilizados. Para a execução da técnica, é necessário definir o número mínimo de amostras por *cluster* (ou a porcentagem mínima proporcional ao tamanho da base de dados), o limiar utilizado para avaliar o balanceamento dos grupos formados após a união, o número de iterações do *k*-médias para minimizar o problema do não-determinismo deste algoritmo, a porcentagem de minimização do erro médio quadrático (utilizada na técnica do cotovelo para determinar o número ideal de *clusters*) e, também, o limiar de união (γ), utilizado para avaliar a sobreposição e proporção entre os *clusters*. Todos estes parâmetros podem influenciar no desempenho da classificação, pois eles determinam o tamanho e a distribuição das amostras nos grupos formados ao final da aplicação da técnica P2C. Assim, a definição destes parâmetros precisa ser realizada com cautela.

Trabalhos Futuros

Considerando as limitações observadas, torna-se necessário estudar a possibilidade de ajustar o método para que ele seja menos sensível aos parâmetros de entrada, visto que eles podem influenciar muito positiva ou negativamente os resultados. O ideal seria que a própria implementação da técnica conseguisse ajustar os parâmetros ideais para cada conjunto de dados, oferecendo soluções mais adaptáveis e robustas.

A partir dos experimentos realizados, nota-se, também, a necessidade de explorar a questão do particionamento por classe, uma vez que outros métodos de agrupamento

podem ser avaliados para possibilitar a formação de grupos com amostras que possam ser linearmente separáveis.

Outra questão que pode ser investigada futuramente está relacionada à paralelização das etapas de particionamento, união e treinamento, pois elas podem ser executadas em várias *threads*, reduzindo significativamente o tempo de processamento global.

A aplicação da técnica P2C sobre bases de dados reais também precisa ser avaliada, visto que os resultados apresentados foram obtidos com bases de dados artificiais. Dentro das bases de dados reais, podem ser avaliadas bases de dados simples e, posteriormente, bases de dados maiores e mais complexas, onde nem mesmo os classificadores não-lineares conseguem ser treinados.

De modo semelhante, outros métodos de classificação, além dos três avaliados (BNB, LR e LSVM) podem ser considerados futuramente, para verificar se a aplicação da técnica P2C também pode aumentar o seu poder de predição.

Referências

ACHARYA, A. et al. C3E: A Framework for Combining Ensembles of Classifiers and Clusterers. In: *Multiple classifier systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, (Lecture Notes in Computer Science, v. 6713). p. 269–278. ISBN 978-3-642-21556-8. Citado na página 11.

ALMEIDA, L. M.; GALVÃO, P. S. Ensembles with clustering-and-selection model using evolutionary algorithms. In: *Proceedings of the 5th Brazilian Conference on Intelligent Systems (BRACIS'16)*. Recife, Brazil: [s.n.], 2016. ISBN 9781509035663. Citado 2 vezes nas páginas 1 e 9.

ARTHUR, D. et al. k-means++: The advantages of careful seeding. In: *Proceedings of the 18th Annual ACM-SIAM symposium on Discrete algorithms (SODA'07)*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2007. v. 8, p. 1027–1035. ISBN 978-0-898716-24-5. Disponível em: <<http://portal.acm.org/citation.cfm?id=1283494>>. Citado na página 16.

BOLEY, D.; CAO, D. Training Support Vector Machine using Adaptive Clustering. In: *Proceedings of the 4th SIAM International Conference on Data Mining (SDM'04)*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2004. p. 126–137. Disponível em: <<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.120.8797&rep=rep1&type=pdf>>. Citado na página 13.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A Training Algorithm for Optimal Margin Classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*, v. 1, n. 1, p. 144–152, 1992. ISSN 0-89791-497-X. Citado 3 vezes nas páginas 5, 6 e 27.

BRITTO, A. S.; SABOURIN, R.; OLIVEIRA, L. E. Dynamic selection of classifiers—A comprehensive review. *Pattern Recognition*, v. 47, n. 11, p. 3665–3680, nov 2014. ISSN 00313203. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S0031320314001885>>. Citado 7 vezes nas páginas 1, 6, 8, 9, 10, 11 e 41.

CAVALIN, P. R.; SABOURIN, R.; SUEN, C. Y. Dynamic selection approaches for multiple classifier systems. *Neural Computing and Applications*, v. 22, n. 3-4, p. 673–688, mar 2013. ISSN 0941-0643. Disponível em: <<http://link.springer.com/10.1007/s00521-011-0737-9>>. Citado na página 10.

COLETTA, L. F. S. et al. Combining Classification and Clustering for Tweet Sentiment Analysis. In: *Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS'14)*. IEEE, 2014. p. 210–215. ISBN 9781479956180. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6984832>>. Citado 2 vezes nas páginas 9 e 11.

CORMEN, T. H. et al. *Introduction to Algorithms, Third Edition*. 3rd. ed. [S.l.]: The MIT Press, 2009. ISBN 0262033844, 9780262033848. Citado na página 2.

CORTES, C.; VAPNIK, V. Support-Vector Networks. *Journal of Machine Learning*, v. 20, n. 3, p. 273–297, aug 1995. ISSN 15730565. Citado 3 vezes nas páginas 5, 6 e 27.

- CRAMER, J. *The Origins of Logistic Regression*. [S.l.], 2002. Disponível em: <<https://EconPapers.repec.org/RePEc:tin:wpaper:20020119>>. Citado 2 vezes nas páginas 5 e 27.
- CRUZ, R. M. O.; SABOURIN, R.; CAVALCANTI, G. D. C. META-DES.H: A Dynamic Ensemble Selection technique using meta-learning and a dynamic weighting approach. In: *Proceedings of the International Joint Conference on Neural Networks (IJCNN'15)*. Killarney, Ireland: IEEE, 2015. v. 2015-Septe, p. 1–8. ISBN 978-1-4799-1960-4. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7280594>>. Citado na página 10.
- DEMŠAR, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, v. 7, p. 1–30, 2006. ISSN 1532-4435. Citado 2 vezes nas páginas 37 e 38.
- DIETTERICH, T. G. Ensemble Methods in Machine Learning. In: KITTLER, J.; ROLI, F. (Ed.). *Multiple classifier systems*. 1. ed. Cagliari, Italy: Springer Berlin Heidelberg, 2000. cap. 1, p. 1–15. ISBN 978-3-540-67704-8. Disponível em: <http://link.springer.com/chapter/10.1007/3-540-45014-9_1>. Citado 3 vezes nas páginas 1, 8 e 11.
- FACELI, K. et al. *Inteligência artificial: Uma abordagem de aprendizado de máquina*. 1st. ed. [S.l.]: LTC, 2011. Citado 2 vezes nas páginas 5 e 27.
- FRADKIN, D. Clustering Inside Classes Improves Performance of Linear Classifiers. In: *Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI08)*. IEEE, 2008. v. 2, n. Cic, p. 439–442. ISBN 978-0-7695-3440-4. ISSN 10823409. Disponível em: <<http://ieeexplore.ieee.org/document/4669806/>>. Citado 6 vezes nas páginas 1, 2, 9, 12, 13 e 16.
- FU, Z.; ROBLES-KELLY, A.; ZHOU, J. Mixing Linear SVMs for Nonlinear Classification. *Journal of IEEE Transactions on Neural Networks*, v. 21, n. 12, p. 1963–1975, dec 2010. ISSN 1045-9227. Disponível em: <<http://ieeexplore.ieee.org/document/5634127/>>. Citado 4 vezes nas páginas 1, 2, 9 e 13.
- GIACINTO, G.; ROLI, F. Methods for dynamic classifier selection. In: *Proceedings of the 10th International Conference on Image Analysis and Processing (ICIAP'99)*. Venice, Italy: IEEE Comput. Soc, 1999. p. 659–664. ISBN 0-7695-0040-4. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=797670>>. Citado na página 11.
- GIACINTO, G.; ROLI, F.; FUMERA, G. Selection of Classifiers Based on Multiple Classifier Behaviour. In: FERRI, F. J. et al. (Ed.). *Proceedings of the Joint international workshops on Syntactical and Structural Pattern Recognition & Statistical Pattern Recognition (SSPR'00)*. Alicante, Spain: Springer Berlin Heidelberg, 2000. p. 87–93. ISBN 3-540-67946-4. Citado na página 11.
- HSU, C.-W.; LIN, C.-J. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, v. 13, n. 2, p. 415–425, mar 2002. ISSN 10459227. Disponível em: <<http://ieeexplore.ieee.org/document/991427/>>. Citado na página 9.

- JACOBS, R. A. et al. Adaptive Mixtures of Local Experts. *Neural Computation*, v. 3, n. 1, p. 79–87, feb 1991. ISSN 0899-7667. Disponível em: <<http://www.mitpressjournals.org/doi/abs/10.1162/neco.1991.3.1.79>>. Citado na página 9.
- JOACHIMS, T. Training linear SVMs in linear time. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '06*. Philadelphia, PA, USA: ACM Press, 2006. p. 217–226. ISBN 1595933395. Disponível em: <<http://portal.acm.org/citation.cfm?doid=1150402.1150429>>. Citado na página 1.
- KO, A. H.; SABOURIN, R.; BRITTO JR., A. S. From dynamic classifier selection to dynamic ensemble selection. *Pattern Recognition*, v. 41, n. 5, p. 1718–1731, may 2008. ISSN 00313203. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S0031320307004499>>. Citado 2 vezes nas páginas 10 e 11.
- KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - NIPS'12*. [S.l.: s.n.], 2012. p. 1–9. ISBN 9781627480031. ISSN 10495258. Citado 2 vezes nas páginas 1 e 41.
- KUNCHEVA, L. Clustering-and-selection model for classifier combination. In: *Proceedings of the 4th International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies (KES'00)*. Brighton, UK: IEEE, 2000. v. 1, p. 185–188. ISBN 0-7803-6400-7. Disponível em: <<http://dx.doi.org/10.1109/KES.2000.885788>>. Citado 4 vezes nas páginas 1, 9, 12 e 41.
- KYRIAKOPOULOU, A.; KALAMBOUKIS, T. Combining Clustering with Classification for Spam Detection in Social Bookmarking Systems. In: *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD'08)*. Antwerp, Belgium: [s.n.], 2008. p. 47–54. Disponível em: <<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.183.3273>>. Citado 2 vezes nas páginas 9 e 12.
- LAM, L.; SUEN, S. Y. Application of majority voting to pattern recognition: An analysis of its behavior and performance. *Trans. Sys. Man Cyber. Part A*, IEEE Press, Piscataway, NJ, USA, v. 27, n. 5, p. 553–568, set. 1997. ISSN 1083-4427. Disponível em: <<http://dx.doi.org/10.1109/3468.618255>>. Citado na página 10.
- LANGLEY, P.; IBA, W.; THOMPSON, K. An Analysis of Bayesian Classifiers. In: *Proceedings of the 10th National Conference on Artificial, AAAI'92*. [S.l.: s.n.], 1992. v. 1, n. 1, p. 223–228. ISBN 0262510634. Citado 2 vezes nas páginas 5 e 27.
- LEMAITRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, v. 18, p. 1–5, sep 2016. ISSN 15337928. Disponível em: <<http://arxiv.org/abs/1609.06570>>. Citado na página 21.
- LIU, R.; YUAN, B. Multiple classifiers combination by clustering and selection. *Information Fusion*, v. 2, n. 3, p. 163–168, sep 2001. ISSN 15662535. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S1566253501000331>>. Citado na página 10.
- LOCHTER, J.; ZANETTI, R.; ALMEIDA, T. Detecção de Opinião em Mensagens Curtas usando Comitê de Classificadores e Indexação Semântica. In: *Anais do XII Encontro*

Nacional de Inteligência Artificial e Computacional (ENIAC'15). Natal, Brazil: [s.n.], 2015. p. 1–7. Disponível em: <<http://www.lbd.dcc.ufmg.br/colecoes/eniac/2015/007.pdf>>. Citado 6 vezes nas páginas 1, 6, 8, 10, 11 e 41.

MCCALLUM, A.; NIGAM, K. A Comparison of Event Models for Naive Bayes Text Classification. *Proceedings of the Workshop on Learning for Text Categorization, AAAI/ICML-98*, v. 1, n. 1, p. 41–48, 1998. ISSN 0343-6993. Disponível em: <<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.9324&rep=rep1&type=pdf>>. Citado 2 vezes nas páginas 5 e 27.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. *Foundations of Machine Learning*. 1st. ed. [S.l.]: MIT Press, 2012. 1–8 p. ISSN 03029743. ISBN 9783642341052. Citado na página 5.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 27.

RUTA, D.; GABRYS, B. Classifier selection for majority voting. *Information Fusion*, v. 6, n. 1, p. 63–81, mar 2005. ISSN 15662535. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S1566253504000417>>. Citado 2 vezes nas páginas 9 e 10.

SABOURIN, M. et al. Classifier combination for hand-printed digit recognition. In: *Proceedings of the 2nd International Conference on Document Analysis and Recognition (ICDAR'93)*. Tsukuba Science City, Japan: IEEE Comput. Soc. Press, 1993. p. 163–166. ISBN 0-8186-4960-7. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=395758>>. Citado na página 11.

SAEEDIAN, M. F.; BEIGY, H. Dynamic classifier selection using clustering for spam detection. In: *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining (CIDM'09)*. Nashville, TN: IEEE, 2009. p. 84–88. ISBN 978-1-4244-2765-9. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4938633>>. Citado na página 10.

SCHMIDHUBER, J. *Deep learning in Neural Networks: An Overview*. [S.l.], 2014. v. 4, 1–88 p. Citado 2 vezes nas páginas 1 e 41.

SILVA, N. F. da; HRUSCHKA, E. R.; HRUSCHKA, E. R. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, Elsevier B.V., v. 66, p. 170–179, oct 2014. ISSN 01679236. Disponível em: <<http://dx.doi.org/10.1016/j.dss.2014.07.003>>. Citado na página 11.

SMITH, D. R. *The structure of divide and conquer algorithms*. [S.l.], 1983. v. 27, n. 1, 1–38 p. Citado 2 vezes nas páginas 2 e 9.

SMITH, D. R. Applications of a strategy for designing divide-and-conquer algorithms. *Journal of Science of Computer Programming*, v. 8, n. 3, p. 213–229, 1987. ISSN 01676423. Citado na página 9.

VAPNIK, V.; LERNER, A. Pattern recognition using generalized portrait method. *Journal of Automation and Remote Control*, v. 24, n. 6, p. 774–780, 1963. ISSN 0005-1179. Citado 3 vezes nas páginas 5, 6 e 27.

WOODS, K.; KEGELMEYER, W.; BOWYER, K. Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 19, n. 4, p. 405–410, apr 1997. ISSN 01628828. Disponível em: <<http://ieeexplore.ieee.org/ielx1/34/12846/00588027.pdf?tp=&arnumber=588027&isnumber=12846>>. Citado 3 vezes nas páginas 6, 9 e 11.

YUAN, G.-x.; HO, C.-h.; LIN, C.-j. Recent Advances of Large-Scale Linear Classification. In: *Proceedings of the IEEE 100*. [s.n.], 2012. v. 100, n. 9, p. 2584–2603. ISSN 0018-9219. Disponível em: <<http://ieeexplore.ieee.org/document/6177645/>>. Citado na página 8.