Paula Amaral
Nova SST|FCT Nova e NovaMath CMA, Universidade Nova de Lisboa
email@email

Tiago Dias
Nova SST|FCT Nova e NovaMath CMA, Universidade Nova de Lisboa
email@email

UMA NUVEM DE ESFERAS – UM NOVO MODELO DE CLASSIFICAÇÃO AUTOMÁTICA

O modelo aqui apresentado foi proposto no artigo "A Classification Method Based on a Cloud of Spheres", publicado na revista *Euro Journal on Computational Optimization* e foi-lhe atribuído o Prémio Marguerite Frank para o melhor artigo de 2023 desta revista. Este modelo constitui uma nova proposta na classificação automática, com vantagens em termos de transparência face a outros métodos, que, apesar de produzirem bons resultados, funcionam como uma caixa preta. A ideia principal consiste na definição da superfície de separação entre duas classes, através de uma nuvem de esferas conectadas.

INTRODUÇÃO

A Classificação Automática Discreta Determinística (CADD) refere-se, de forma geral, à tarefa de atribuir uma classe k de um conjunto K a um objeto i representado por um vetor x de características ou atributos num espaço n-dimensional. Exemplos de CADD incluem a classificação de imagens (por exemplo, determinar se uma imagem contém um gato ou um cão) e a deteção de fraudes (por exemplo, classificar uma transação como fraudulenta ou não). Esta classe de modelos de classificação supervisionada pressupõe a existência de um histórico de dados, contendo registos/objetos/exemplos para os quais se conhece não só as características (*features*) mas também a etiqueta da correspondente categoria/classe (*label*). Estes dados são usados para treinar o modelo (*training phase*) e para testar o modelo (*testing phase*). A arquitetura de um algoritmo de classificação supervisionada pode variar, mas geralmente procura ajustar-se bem aos dados de treino, evitando o ajuste excessivo (*overfitting*) que compromete a capacidade de generalização para dados novos. Modelos de classificação, de maneira implícita ou explícita, minimizam uma função de erro que mede o quão bem o modelo classifica os dados de treino. Em tarefas de classificação supervisionada, assume-se a existência de separação entre as classes, sendo o objetivo identificar uma estrutura que represente da melhor forma essa separação. Essa estrutura pode estar embutida numa rede neuronal (*Neural Networks*) ou ser descrita por meio de regras (*Random Forest*) ou, implicitamente, pela definição de um modelo matemático cujos parâmetros são obtidos pela solução de um problema de otimização.

MODELOS CLÁSSICOS DE APRENDIZAGEM AUTOMÁTICA

A literatura sobre classificação é vasta. A título de exemplo, citam-se três referências sobre o tema [1, 2, 3]. De entre os métodos clássicos de classificação supervisionada, destacam-se: regressão logística, máquinas com vetores de suporte (SVM), k-vizinhos mais próximos (k-NN), árvores de decisão e redes neuronais. Esses métodos variam em complexidade computacional, capacidade de interpretar a regra subjacente à separação das classes (interpretabilidade) e capacidade de generalização. Recentemente, tem havido grande interesse no uso de técnicas baseadas na otimização de modelos de programação matemática, para

construção de classificadores com propriedades desejáveis, como margens máximas de separação ou minimização de funções de perda específicas. Veja-se, por exemplo, [4, 5, 6, 7]. Modelos baseados em programação inteira mista (MIP) e programação não linear inteira mista (MINLP) têm-se mostrado promissores, especialmente em contextos nos quais se pretende promover a interpretabilidade e o controlo sobre as características na fronteira da decisão, que permitem explicar por que um novo objeto é classificado numa classe e não noutra (*counterfactual analysis*). Esses modelos, embora mais pesados computacionalmente, permitem incorporar restrições explícitas ao processo de classificação, o que é útil em aplicações críticas, como diagnóstico médico e análise financeira.

ÉTICA E APRENDIZAGEM AUTOMÁTICA

É sabido que alguns algoritmos de aprendizagem automática (AA), em particular as redes neuronais artificiais (ANN), as redes neuronais profundas (DNN) e as florestas aleatórias (RF), têm a capacidade de representar associações não lineares complexas nos dados. No entanto, as ANN e as DNN são algoritmos de tipo "caixa preta", o que significa que não é possível compreender explicitamente o que é realmente aprendido nem como é aprendido, veja-se a este propósito a referência [7]. As falhas das ANN e das DNN têm sido reconhecidas e investigadas [8, 9]. A ausência de reflexão sobre estes algoritmos, bem como a sua aplicação indiscriminada e acrítica em situações do mundo real, pode conduzir a resultados muito indesejáveis, não éticos, injustos, sensíveis e até perigosos. Os modelos e algoritmos de AA evidenciam mesmo uma tendência para amplificar a discriminação com base no género [10] ou na raça [11], pelo que diversos autores têm alertado para esta problemática [12, 13, 14].

Além destas questões, as ANN e as DNN são difíceis de treinar [15]. A otimização dos parâmetros é demasiado complexa para se alcançar um ótimo global. Ainda assim, têm sido apresentados argumentos a favor da subotimização dos parâmetros, aludindo ao facto de que evitar o sobreajustamento pode até ser desejável. Além da otimização dos parâmetros, o próprio desenho da rede tem de ser avaliado. Com a complexidade das estruturas das redes neuronais, há demasiados componentes a considerar (regularização, taxa de aprendizagem, funções de ativação, número de camadas, unidades ocultas, nós intermédios). Dada a impossibilidade de analisar todas as combinações possíveis de estruturas e parâmetros, apenas uma parte dessas combinações é avaliada. O critério de escolha é muitas vezes sustentado por argumentos frágeis e demasiado dependentes dos dados de treino e teste, levantando questões quanto à sua adequação para novos dados.

Em SVM, são introduzidas funções (*kernels*) para lidar com a não linearidade. Embora os SVM com *kernel* sejam mais transparentes do que as DNN e tenham menos combinações de aspetos do modelo a estudar e a avaliar, ainda assim apresentam algum esforço de natureza combinatória no treino. A escolha correta da função *kernel*, juntamente com os parâmetros e metaparâmetros no caso de margens suaves (*soft margins*), pode ser complexa. Além disso, o tipo de linearidade que pode ser descrito é limitado, apresentando grandes dificuldades em aprender a não convexidade das curvas de separação das classes.

UM MODELO DE CLASSIFICAÇÃO BASEADO NUMA NUVEM DE ESFERAS CONECTADAS

As críticas às redes neuronais (ANN e DNN) e aos métodos de *kernel* de SVM motivaram a procura de métodos de classificação mais transparentes e simples. Com inspiração nos aspetos positivos do SVM, procurou-se um método baseado num modelo matemático, onde a relação entre o modelo e a solução pudesse ser claramente identificada, com apenas um número razoável de parâmetros a estimar, mas que fosse capaz de capturar estruturas de separação complexas nos dados.

Se considerarmos um problema de classificação com duas classes, encontrar uma superfície de separação pode ser difícil num espaço de alta dimensão, como $\mathbb{R}^n$ para valores elevados de n, dependendo da geometria dos objetos ou pontos. Se os pontos das duas classes forem linearmente separáveis, a maioria dos métodos clássicos propostos na literatura funcionará bem, como é o caso do SVM. No entanto, se os pontos não forem linearmente separáveis, encontrar uma superfície de separação pode tornar-se uma tarefa complexa. Normalmente, assume-se algum tipo de comportamento regular na distribuição espacial dos pontos (esférico, poliedral) para justificar a aplicação de um determinado método, mas especular sobre a geometria dos dados em espaços de alta dimensão pode ser enganador.

Nesse sentido, considerou-se que quanto mais flexível fosse a superfície de separação, melhor seria, permitindo assim a definição de formas diferentes e complexas. Se considerarmos, por exemplo, o caso bidimensional ilustrado na figura 1, a superfície de separação definida na figura 2 procura encapsular uma estrutura não linear e não convexa.

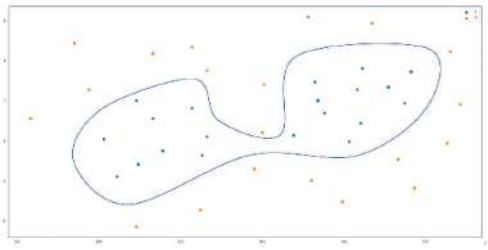

Figura 1. Separação de duas classes de pontos.

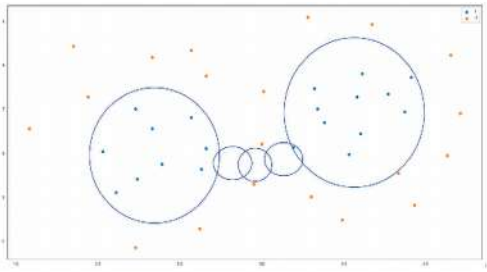

Figura 2. Nuvem de esferas.

O modelo proposto como uma nuvem de esferas conectadas (NEC) insere-se na linha de modelos que utiliza formulações de MINLP para gerar classificadores. Os parâmetros do modelo são obtidos obtendo a solução ótima deste problema. Nas secções seguintes, apresentamos alguns elementos sobre a formulação matemática, bem como um exemplo de aplicação.

FORMULAÇÃO MATEMÁTICA DA NEC

A formulação matemática da NEC tem como objetivo encontrar uma superfície baseada num conjunto de esferas para separar dois conjuntos de pontos de dados. Especificamente, pretende-se construir um conjunto conectado de esferas que cubra todos os pontos do primeiro conjunto, impedindo qualquer interseção com os pontos do segundo conjunto. O objetivo é alcançar esta separação com o menor número de esferas possível. O modelo tem diversas particularidades que não serão aqui descritas. O que se apresenta de seguida são apenas algumas das características do modelo. Como não se sabe à partida quantas esferas serão utilizadas para construir a nuvem, considera-se no modelo um limite superior para o número de esferas.

Como restrições, consideram-se, entre outros, os seguintes conjuntos:

1. Cobertura do primeiro conjunto: Cada ponto do primeiro grupo deve estar contido, pelo menos, numa das esferas.

2. Não cobertura do segundo conjunto: Nenhum ponto do segundo grupo pode estar contido em qualquer uma das esferas utilizadas na solução.

3. Conectividade das esferas: As esferas ativas devem formar uma estrutura conectada – ou seja, cada esfera deve intercetar, pelo menos, uma outra esfera e não podem existir subconjuntos de esferas desconectados dos restantes.

EXEMPLO

Vamos considerar três exemplos *Parallels* (figura 3), *Separation* (figura 4) e *Butterfly* (figura 5), onde os pontos da classe 1 (estrelas azuis) estão rodeados por pontos da classe 2 (cruzes vermelhas).

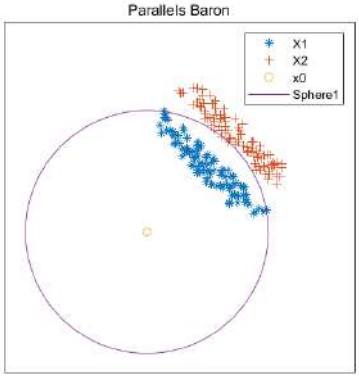

Figura 3. *Parallels.*

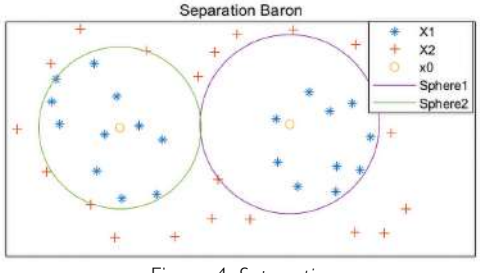

Figura 4. *Separation.*

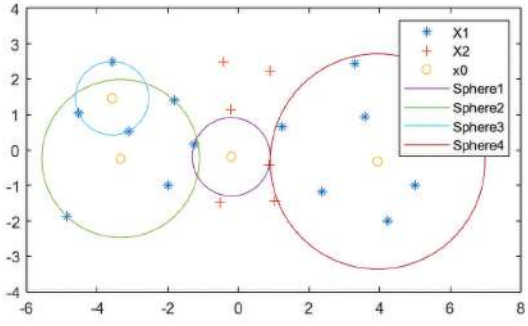

Figura 5. *Butterfly.*

O exemplo da figura 3 ilustra que este método pode ser aplicado mesmo em casos onde um separador linear funcionaria. Aqui uma única esfera é suficiente para definir a separação das duas classes. No exemplo da figura 5, os pontos da classe 1 estão mais afastados entre si do que no exemplo da figura 4. Ao resolver o modelo matemático, obtiveram-se soluções com nuvem definida por quatro e duas esferas, respetivamente. Serve o exemplo da figura 5 para ilustrar que para manter a conectividade das esferas pode ser necessário considerar esferas que não cobrem nenhum ponto da classe 1, as quais são designadas por esferas vazias.

Para determinar a solução ótima do modelo matemático utilizou-se um *solver* de otimização global [16]. No entanto, uma vez que estes problemas de otimização são, em geral, muito difíceis de resolver, para problemas de maior dimensão foram desenvolvidos métodos heurísticos. De seguida, é apresentada uma descrição geral de uma destas heurísticas.

HEURÍSTICA – MAIOR COBERTURA

Estrutura geral do algoritmo

Entrada: Conjunto de pontos de ambas as classes:

$$\text{Classe 1: } C_1 = \{x_{1i}, i = 1, ..., N_1\}$$

$$\text{Classe 2: } C_2 = \{x_{2j}, j = 1, ..., N_2\}$$

Saída: Conjunto de esferas S

Procedimento:

1. Inicializar o contador de iterações: $t \leftarrow 1$

2. Definir $I = C_1$ (ou seja, inicialmente todos os pontos da classe 1 estão por cobrir)

3. Enquanto I não estiver vazio, repetir os seguintes passos:

   a. Definir uma esfera candidata S_k para a iteração atual

   b. Garantir que S_k está conectada às esferas já definidas em S, por introdução, se necessário, de esferas vazias (sem pontos da classe C_1)

   c. Atualizar o conjunto S com S_k e atualizar I, removendo os pontos agora cobertos

Foi, de seguida, desenvolvido um estudo comparativo desta proposta com alguns dos métodos clássicos de AA.

COMPARAÇÃO COM MÉTODOS CLÁSSICOS PARA CONJUNTOS DE DADOS REAIS

Para este estudo comparativo foram utilizados oito conjuntos de dados reais provenientes do *UCI Machine Learning Repository* [17]. O objetivo foi comparar a abordagem heurística com algoritmos de classificação amplamente consolidados, tais como regressão logística (LR), SVM linear (L-SVM), SVM gaussiano fino (FG-SVM) e florestas aleatórias (RF). O protocolo experimental utilizou uma validação cruzada em dez subconjuntos (*10-fold cross validation*). Esta é uma técnica de avaliação de modelos de aprendizagem automática, onde os dados são divididos em dez partes (*folds*). Em cada iteração, nove partes são usadas para treino e uma para teste. O processo repete-se dez vezes, mudando o *fold* de teste de cada vez. No final, calcula-se a média dos resultados obtidos, proporcionando uma estimativa mais robusta do desempenho do modelo.

Tabela 1. Informação dos conjuntos de dados reais

Conjunto de dados	Número de *features*	Número de dados	% da classe 1
Banknotes	4	1372	44,46%
BCW Original	9	683	34,99%
BCW Diagnostic	30	569	37,26%
Ionosphere	34	351	64,10%
Mushrooms	22	8124	48,20%
Pima Diabetes	8	768	34,90%
Sonar	60	208	53,37%
Voting (US)	31	435	38,62%

O desempenho dos modelos é aferido com base na matriz de confusão (*confusion matrix*) que cruza as classes reais dos dados com as classes estimada pelo modelo.

Matriz de confusão	Classe estimada: Positiva	Classe estimada: Negativa
Classe real: Positiva	VP – Verdadeiro positivo	FN – Falso negativo
Classe real: Negativa	FP – Falso positivo	VN – Verdadeiro negativo

As métricas usadas para comparação do desempenho dos modelos foram:

Tabela 2. Resultados de classificação nos conjuntos de dados reais

Conjunto de dados	LR	L-SVM	FG-SVM	RF	Heurística	*Precision*	*Recall*	*F1 score*
Banknotes	98,91	98,40	99,93	94,97	94,66	91,87	96,65	94,16
BCW Original	96,63	96,78	97,07	97,51	93,64	91,64	90,14	90,82
BCW Diagnostic	95,25	99,65	97,89	95,96	90,18	88,09	85,85	86,69
Ionosphere	86,89	87,18	94,59	93,45	84,27	87,08	88,79	87,83
Mushrooms	100	99,90	100	100	99,12	98,38	99,81	99,09
Pima Diabetes	77,60	77,21	73,31	76,82	65,31	50,38	42,66	45,62
Sonar	75,96	77,40	83,65	84,62	62,88	68,37	57,96	62,07
Voting (US)	94,25	95,63	96,09	95,86	87,79	89,02	78,38	82,85

- Acurácia (*Accuracy*): mede a proporção de previsões corretas em relação ao total de previsões realizadas.

$$Accuracy = (VP + VN) / (VP + VN + FP + FN)$$

- Precisão (*Precision*): mede a proporção de estimativas positivas corretas em relação ao total de estimativas positivas feitas.

$$Precision = VP / (VP + FP)$$

- Revocação (*Recall*): mede a proporção de estimativas positivas corretas em relação ao total de instâncias positivas reais.

$$Recall = VP / (VP + FN)$$

- F1-Score: média harmónica entre a precisão e a revocação.

$$F1 = 2 * (Precision * recall) / (Precision + Recall)$$

A tabela 2 apresenta o resultado da comparação em termos de *Accuracy* da heurística proposta (coluna Heurística) com os métodos já referidos, nas colunas LR, L-SVM, FG-SVM e RF para os oito conjuntos de dados descritos na tabela 1. As colunas *Precision*, *Recall* e *F1 Score*, indicam os valores destas métricas apenas para a heurística.

CONCLUSÃO

O modelo da nuvem de esferas conectadas (NEC) apresenta potencial para se tornar uma metodologia viável e interessante para tarefas de classificação supervisionada no contexto da aprendizagem automática. Destaca-se pelo seu processo de classificação simples e transparente, bem como pelo reduzido número de parâmetros a estimar. Contudo, ainda existem desafios importantes a superar. A resolução exata do problema é difícil e, embora os métodos heurísticos consigam soluções interessantes, tendem a apresentar, em geral, um desempenho ligeiramente inferior quando comparadas com algoritmos de classificação consolidados na área de aprendizagem automática. Como perspetiva de trabalho futuro, pretende-se, entre outros, investigar formas de explorar melhor a formulação matemática deste modelo para desenvolver métodos exatos mais eficientes.

REFERÊNCIAS

[1] El Mrabet, M. A., El Makkaoui, K., & Faize, A. "*Supervised Machine Learning: A Survey*". In: *2021 4th International Conference on Advanced Communication Technologies and Networking* (CommNet), pp. 1-10, IEEE, 2021.

[2] Deng, L. & Yu, D. "*Deep Learning: Methods and Applications. Foundations and Trends® in Signal Processing*", 7(3-4), pp. 197-387, 2014. Now Publishers.

[3] Somvanshi, M., Chavan, P., Tambade, S., & Shinde, S. V. "A Review of Machine Learning techniques Using Decision tree andSupport Vector Machine". In: *Proceedings of the 2016 International Conference on Computing Communication Control and Automation* (ICCUBEA), pp. 1-7, IEEE, 2016.

[4] Carrizosa, E. & Romero Morales, D. "*Supervised Classification and Mathematical Optimization. Computers & Operations Research*", 40(1), pp. 150-165, 2013.

[5] Carrizosa, E., Molero-Río, C., & Romero Morales, D. "Mathematical Optimization in Classification and Regression Trees". *TOP,* 29(1), pp. 5-33, 2021.

[6] Astorino, A., Fuduli, A., & Gaudioso, M. (2016). "Nonlinear Programming for Classification Problems in Machine Learning". In *AIP Conference Proceedings*, 1776(1), 040004. AIP Publishing LLC.

[7] Malha, R., & Amaral, P. (2021). "A Maximal Margin Hypersphere SVM". In *International Conference on Computational Science and Its Applications* (pp. 304-319). Springer.

[8] Calado, M. C. A. *Explaining Incorrect Feature Learning in the Training of Deep Neural Networks.* Tese de mestrado, NOVA School of Science and Technology, NOVA University Lisbon, 2022.

[9] Goodfellow, I. J., Shlens, J., & Szegedy, C. *Explaining and Harnessing Adversarial Examples.* arXiv preprint *arXiv:1412.6572, 2014.*

[10] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. *Intriguing Properties of Neural Networks.* arXiv preprint *arXiv:1312.6199*, 2013.

[11] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. Advances in Neural Information Processing Systems*, 29, 2016.

[12] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. Machine Bias: *There's Software Used Across the Country to Predict Future Criminals, and it's Biased Against Blacks.* ProPublica, 2016. (Disponível em: *https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing*)

[13] Wang, S. & Gupta, M. "Deontological Ethics by Monotonicity Shape Constraints". In: *Proceedings of the International Conference on Artificial Intelligence and Statistics* (AISTATS), pp. 2043-2054, PMLR, 2020.

[14] Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Mehta, S., Mojsilovic, A., Nagar, S., et al. "Think Your Artificial Intelligence Software is Fair? Think Again." *IEEE Software*, 36(4), pp. 76-80, 2019.

[15] Livni, R., Shalev-Shwartz, S., & Shamir, O. *On the Computational Efficiency of Training Neural Networks. Advances in Neural Information Processing Systems*, 27, 2014.

[16] Sahinidis, N. V. BARON: "A General-Purpose Global Optimization Software Package". *Journal of Global Optimization*, 8, pp. 201-205, 1996.

[17] Dua, D. & Graff, C. *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences, 2017. *http://archive.ics.uci.edu/ml.*

SOBRE OS AUTORES

Paula Amaral é professora associada na FCT UNL, licenciada em Estatística e Investigação Operacional pela Universidade de Lisboa e com doutoramento em Matemática pela Universidade Nova de Lisboa. Os seus interesses de investigação incluem otimização, aprendizagem automática, rotas de veículos e planeamento de horários. Supervisionou diversas dissertações em ambiente empresarial (CTT, GALP, entre outras), incluindo o vencedor do Prémio de Melhor Tese de Mestrado da APDIO em 2009. As orientações de alunos de doutoramento focaram-se em Otimização para Aprendizagem Automática. Foi membro da direção da PT-MATHS-IN e da comissão executiva do centro de I&D NovaMath CMA. Atualmente, é membro da direção do EURO PT e cocoordenadora do Mestrado em Análise e Engenharia de Big Data.

Tiago Dias é aluno do Programa Doutoral em Matemática na FCT UNL, mestre em Gestão pela NOVA SBE. Os seus interesses de investigação incluem tópicos ligados com data science, em especial na interpretação matemática de modelos de machine learning. Atualmente, é economista na Comissão do Mercado de Valores Mobiliários (CMVM). Em 2023, juntamente com a sua orientadora, Paula Amaral, recebeu o Prémio Marguerite Frank pelo melhor artigo publicado na EJCO.

Secção coordenada pela PT-MATHS-IN, Rede Portuguesa de Matemática para a Indústria e Inovação *pt-maths-in@spm.pt*

NOTA DOS EDITORES

O presente trabalho é o primeiro que conta com a coordenação dos colegas Rui Borges Lopes (Universidade de Aveiro) e Lígia Henriques-Rodrigues (Universidade de Évora), a quem desejamos um bom trabalho. Agradecemos ainda o labor dos coordenadores cessantes, Filipe J. Marques e Ana Jacinta Soares.