A Relevância de Conjuntos Locais em Algoritmos de Aprendizado Baseado em Instâncias

Sandro Vieira Gonçalves & Maria do Carmo Nicoletti

Faculdade de Campo Limpo Paulista - FACCAMP
Campo Limpo Paulista - SP, Brasil

{sandro.tads15@gmail.com,carmo@cc.faccamp.br}

Abstract. *Instance-Based Learning (IBL) is a research area of Machine Learning (ML) that investigates the use of supervised ML algorithms that do not generalize the training set as usually ML algorithms do. Instead, the generalization process happens during the classification phase, when a new data instance should be classified, based on the stored training set. One of the problems associated with IBL algorithms is related to the volume of instances to be stored, which also causes, as a side effect, an increase in the processing time associated with the classification process. One of the strategies used to reduce storage volume and improve classification time is based on the concept of local set. This article addresses IBL algorithms in general and discusses the relevance of local sets in the context of automatic learning using IBL, focusing on the nearest neighbor (NN) algorithm, which is one of the most widespread algorithms of the ABI paradigm for the new instance classification task.*

Resumo. *Aprendizado Baseado em Instâncias (ABI) é uma área do Aprendizado de Máquina (AM) que investiga o uso de algoritmos de AM supervisionados que não realizam o processo de generalização, a partir do conjunto de treinamento. O processo de generalização acontece na fase de classificação, quando o algoritmo deve classificar uma nova instância, considerando o conjunto de treinamento armazenado. Um dos problemas associados a algoritmos ABI está relacionado ao volume de instâncias a ser armazenado que, também, provoca como efeito colateral o tempo de processamento associado à classificação de novas instâncias. Uma das estratégias utilizadas para reduzir o volume de armazenamento e melhorar o tempo de classificação é subsidiada pelo uso de conjunto local. Este artigo aborda algoritmos ABI em geral e discute a relevância de conjuntos locais no contexto de ABI, com foco no algoritmo vizinho mais próximo (NN), que é um dos algoritmos mais difundidos do paradigma ABI para a tarefa de classificação de novas instâncias.*

1. Considerações Iniciais

Dentre os vários grupos de algoritmos de aprendizado de máquina (AM) supervisionado, aquele constituído por algoritmos baseados em instâncias (ABI) é caracterizado por reunir algoritmos que armazenam as instâncias do conjunto de treinamento (CT), sem que antes tenha sido feito qualquer procedimento de generalização dessas instâncias. Em muitos domínios de conhecimento, entretanto, dependendo do volume do CT, o processo de armazenamento requer um espaço em memória considerável que, nem sempre, está disponível.

Algoritmos de ABI não generalizam o CT em expressões que representam as instâncias do conjunto, como usualmente a maioria dos algoritmos de AM faz. Via de regra o chamado 'conceito aprendido', para algoritmos ABI, consiste no próprio CT fornecido. O processo de generalização realizado por tais algoritmos é adiado até o momento em que uma nova instância de dado, de classe desconhecida, precisa ser classificada. No caso de algoritmos ABI, a tradicional 'fase de aprendizado' usualmente realizada por algoritmos de AM se resume apenas no armazenamento do CT.

Como destacado em [Mitchell 1997], o ABI inclui métodos de classificação e de regressão localmente ponderada de acordo com o(s) vizinho(s) mais próximo(s) e assumem que instâncias podem ser representadas como pontos em um espaço euclidiano, também incluem métodos de raciocínios baseado em casos (CBR) que usam representações simbólicas mais complexas para as instâncias.

O algoritmo NN (*Nearest Neighbor*) [Cover & Hart 1967] é um dos algoritmos mais utilizados de ABI para tarefa de classificação de novas instâncias, tal processo de classificação é feito por uma varredura do CT armazenado, com o objetivo de identificar a instância armazenada que seja mais 'similar' à nova instância. Uma vez identificada tal instância, a sua classe é atribuída à nova instância.

Se por um lado esse fato pode ser considerado vantajoso, uma vez que o tempo investido é apenas o tempo gasto no armazenamento, por outro, à medida que o volume de instâncias de treinamento cresce, as exigências para armazenamento de tal conjunto também crescem. O custo computacional envolvido na classificação de uma nova instância está fortemente relacionado ao volume do CT e, em uma situação em que tal volume é alto, o uso de ABI em tarefas de classificação pode ser desvantajoso, devido ao alto tempo envolvido.

Na literatura podem ser encontradas inúmeras referências que tratam de ABIs. Particularmente, fundamentais a qualquer pesquisa envolvendo ABIs são os trabalhos clássicos na área; entre eles estão [Aha *et al.* 1991] [Aha 1992] [Cover & Hart 1967] [Hart 1968].

2. Os Algoritmos Clássicos de ABI

Um dos algoritmos básicos de ABI é o NN que, além de ter sido pesadamente utilizado nos mais variados domínios de conhecimento, desde a sua proposta até os dias de hoje, foi também inspiração e base para muitos outros algoritmos ABI que se seguiram desde então.

Devido ao fato do algoritmo NN requerer uma capacidade de armazenamento alta, ele se torna não atrativo em muitas aplicações do mundo real. O algoritmo *Condensed Nearest Neighbour* (CNN) [Hart 1968] foi proposto com o objetivo de reduzir o conjunto de instância armazenado, preservando apenas instâncias do CT que discordam da classificação de seu vizinho mais próximo. Essa técnica de seleção de instâncias é conhecida como condensação. Gates em [Gates 1972] comenta que uma possível diminuição em eficiência (devido à redução do volume de instâncias armazenadas) do CNN pode ser compensada pelo menor espaço de armazenamento que o algoritmo requer, bem como pelo menor tempo de classificação utilizado pelo algoritmo, dado que busca a instância 'similar' em um conjunto de instâncias bem menor.

O algoritmo *Reduced Nearest Neighbour* (RNN) [Gates 1972] foi proposto como uma extensão do CNN, que ainda mantém o mesmo objetivo, *i.e.*, reduzir o conjunto de instâncias a ser armazenado. O RNN tem como primeiro passo a execução do algoritmo CNN em um conjunto original de instâncias. O conjunto reduzido, retornado pelo CNN é, então, em um segundo passo do RNN, entrada para um processamento adicional, que busca reduzir esse conjunto, removendo todas as instâncias que concordam com a classificação de seu vizinho mais próximo.

Outra técnica de seleção de instâncias, conhecida como edição, é proposta pelo algoritmo *Edited Nearest Neighbor* (ENN) [Wilson 1972]. A edição de instâncias do ENN busca eliminar as instâncias que discordam da classificação de seus vizinhos mais próximos e, neste contexto, tem o foco em preservar a precisão de classificação e, geralmente, reduzem muito pouco o volume de instâncias original. Algoritmos como o ENN e seus derivados, normalmente, são utilizados como filtro para instâncias com ruído em uma etapa de pré-processamento para outros algoritmos ABI.

Uma pesquisa relevante, com a finalidade de explorar os limites em ABI são os algoritmos pertencentes à família *Instance-Based Learning* (IBL), que constituem um conjunto de algoritmos IBI *i.e.*, os algoritmos nomeados de IB1, IB2, IB3, IB4 e IB5, propostos em [Aha *et al.* 1991] [Aha 1992]. Cada um deles pode ser considerado uma variante do algoritmo IB1. O IB1 pode ser caracterizado como uma versão do NN em que a cada instância foi associado um registro de desempenho em classificações anteriores, quando do armazenamento do conjunto de treinamento como expressão do conceito.

O algoritmo IB2 explora a abordagem baseada em instância do ponto de vista de redução do volume de instâncias, armazenando apenas as instâncias que discordam da classificação de seu vizinho mais próximo. Assim sendo, pode ser considerado uma variante do algoritmo CNN. O algoritmo IB3 é uma variante do IB1 proposto com o intuito de explorar o impacto de ruídos nas instâncias no resultado da classificação e, também, como uma tentativa de reduzir a sensibilidade do algoritmo IB2 a instâncias com ruído. O IB4 é uma variante do IB3 que busca contornar a sensibilidade a atributos irrelevantes na descrição das instâncias de dados e, finalmente, o IB5 é uma variante que busca explorar as possibilidades de um algoritmo IBI de considerar a introdução de novos atributos nas descrições das instâncias de dados.

3. Conjunto Local e Seu Uso no Processo de Redução do Volume de Instâncias em ABI

Uma maneira de tornar o ABI factível em domínios de conhecimento com um alto volume de instâncias de treinamento é por meio de um processo de redução do conjunto de instâncias disponibilizado, de maneira que sejam armazenadas apenas aquelas instâncias de dados consideradas efetivamente relevantes ao conceito representado pelo CT. No que segue é apresentado o uso do conceito de conjuntos locais por algoritmos ABI com vistas à redução do volume de instâncias de dados armazenadas.

3.1 Conjuntos Locais (CL) – Conceituação

Considere um conjunto de instâncias de dados X. O *conjunto local* (CL) de uma instância $x \in X$ é definido como segue. Seja y a instância mais próxima de x, com classe diferente da classe de x e que d = distância(x, y). O conjunto local de x (CL(x)) é o conjunto

de todas as instâncias de X que estão à uma distância de x menor ou igual a d. A Figura 1 mostra o conjunto de instâncias $\{x_1, x_2, x_3, x_4, x_5, x_6\}$ e exemplifica o conjunto local associado à instância x_2, sendo x_5 a instância mais próxima de x_2 cuja classe é diferente da classe de x_2.

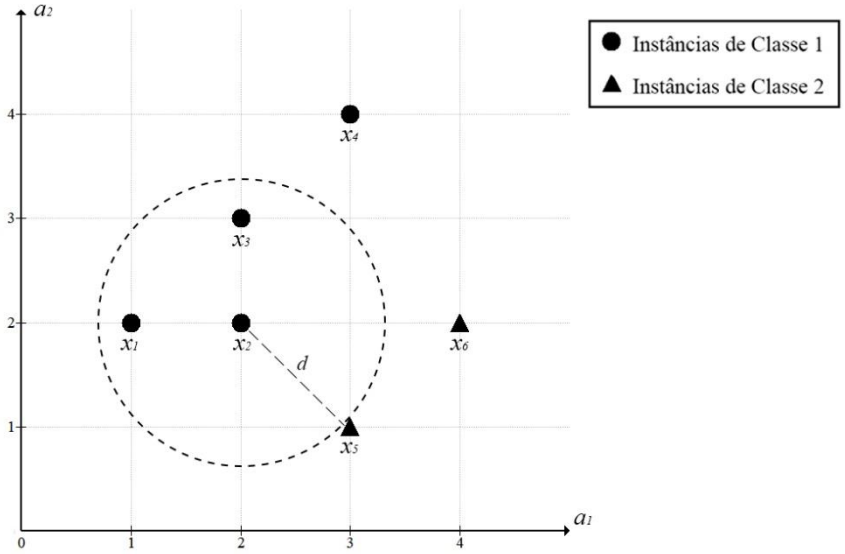

Figura 1 Conjunto local da instância x_2. Em que, d é a menor distância de x_2 à instância mais próxima de classe diferente, representada por x_5, e CL(x_2) = $\{x_1, x_2, x_3\}$.

O conceito de CL na classificação de novas instâncias de acordo com o vizinho mais próximo, apresentado em [Brighton 1996], é adaptado de um sub-paradigma de ABI conhecido como CBR (*Case Based Reasoning*) [Smyth & Keane 1995], em que um caso *i.e.,* uma instância com informações adicionais, pode ser excluído se houver outros casos que sobreponham sua informação no contexto geral.

3.2 O Uso de Conjuntos Locais em ABI

Uma característica comum em ABI é representar as n instâncias de um conjunto de treinamento como pontos em um espaço m-dimensional, em que m é o número de atributos que descrevem cada instância deste conjunto. O conceito de CL reforça essa noção intuitiva, considerando que CLs representam hiperesferas que delimitam instâncias que têm a mesma classe, fornecendo assim uma descrição de como as instâncias estão dispostas nos subespaços que delimitam no espaço m-dimensional associados às suas classes [Caises *et al.* 2011].

O processo de identificação de 'vizinhança' de uma instância pode ser calculado com um baixo custo computacional, tendo uma complexidade da ordem de $O(m \times n)^2$, considerando que informações podem ser facilmente obtidas no cálculo da matriz de distâncias, inerente do paradigma ABI.

A primeira proposta de algoritmo a utilizar a estratégia de CL para redução de instâncias foi apresentada em [Brighton & Mellish 1999]. O algoritmo, denominado como ICF (*Iterative Case Filtering*), após uma etapa de pré-processamento no ENN, descarta toda instância x, em que a cardinalidade do CL(x) for maior que a cardinalidade do CB(x), sendo CB o conjunto de todas as instâncias que tem x em seus respectivos CLs, *i.e*. a cobertura de x. Os resultados obtidos com o algoritmo ICF atingiram média de precisão de classificação superior ao NN e demonstraram bom critério para remoção de

instâncias supérfluas, que são instâncias que pouco contribuem na classificação de novas instâncias [Brighton & Mellish 2002].

Além da remoção de instâncias supérfluas, os CLs também podem ser utilizados para auxiliar na remoção de instâncias com ruído, *e.g.*, instâncias com ruído tendem a ter CL unitário. Neste contexto algoritmos ABI que utilizam o conceito de CL podem auxiliar na etapa de pré-processamento, como filtro para instâncias indesejáveis, das quais a maioria dos algoritmos de seleção de instâncias não são capazes de lidar, como é o caso do algoritmo LSSm (*Local Set-Based Smoother*) [Leyva *et al.* 2015].

Também podem ser empregados em estratégias mais extremas de redução do número de instâncias em relação ao conjunto original, mantendo apenas representações do centro de cada CL. O MRT (*Medoid Retention Technique*) [Caises *et al.* 2011] e o LSCo (*Local Set-Based Centroid Selector*) [Leyva *et al.* 2015] são exemplos de algoritmos que fazem uso dessa estratégia.

O conhecimento do CL de cada instância pode auxiliar também em algoritmos que buscam obter uma boa redução no volume de instâncias sem perder precisão de classificação em relação ao conjunto original. Como é o caso do algoritmo LSBo (*Local Set Border Selector*) [Leyva *et al.* 2015] que mantém apenas as instâncias de borda do CL, *i.e.*, instâncias com CL de menor cardinalidade. Por se tratar de uma estratégia de seleção de instâncias sensível a ruídos, o algoritmo LSBo necessita de uma etapa de pré-processamento com o algoritmo LSSm.

4. Considerações Finais

Este artigo tem por foco algoritmos de aprendizado baseado em instâncias (ABI) e, particularmente, algoritmos que fazem uso do conceito de conjuntos locais para promover tanto a redução do conjunto de treinamento armazenado quanto a preservação, ou mesmo melhoria, da precisão de classificação do algoritmo NN.

Neste contexto biobjetivo, os algoritmos que fazem uso do conceito de conjuntos locais apresentam propostas complementares, facilmente adaptáveis de acordo com o problema proposto. Característica não apresentada por algoritmos tradicionais de seleção de instâncias para o NN.

O custo computacional, para o calculo do conjunto local de cada instância, é praticamente o mesmo do requerido pelo algoritmo NN, que pode ser considerado um desempenho eficiente dentre os algoritmos de ABI. Além do que, uma vez calculado os CLs, as informações obtidas podem ser utilizadas tanto para o objetivo de redução do volume de instâncias quanto para a eliminação de instâncias indesejáveis, visando aumentar a precisão de classificação. Uma vantagem considerável para explorar diferentes soluções com um custo computacional factível.

Referências

[Aha *et al.* 1991] Aha, D. W.; Kibler, D.; Albert, M. K. (1991) Instance-based learning algorithms, *Machine Learning*, v. 6, pp. 37–66.

[Aha 1992] Aha, D. W. (1992) Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms, *International Journal of Man-Machine Studies*, 36, 267–287.

[Bottou & Vapnik 1992] Bottou, L.; Vapnik, V. (1992) Local learning algorithms, *Neural Computation*, v. 4, pp. 888 – 900.

[Brighton & Mellish 2002] Brighton, H.; Mellish, C. (2002) Advances in instance selection for instance-based learning algorithms, *Data Mining Knowlege Discovery*, v. 2, pp. 153–172.

[Brighton & Mellish 1999] Brighton, H. and Mellish, C. (1999) On the consistency of information filters for lazy learning algorithms. In Principles of Data Mining and Knowledge Discovery, 3rd European Conference, Prague, Czech Republic, J.M. Zytkow and J. Rauch (Eds.), pp. 283–288.

[Chang 1974] Chang, C.-L. (1974) Finding prototypes for nearest neighbor classifiers, IEEE Transactions on Computers, v. 23, no. 11, pp. 1179–1184.

[Cover & Hart 1967] Cover, T. M.; Hart, P. E. (1967) Nearest neighbor pattern classification. Institute of Electrical and Electronics Engineers Transactions on Information Theory, v. 13, no. 1, pp. 21–27.

[Domingos 1995] Domingos, P. (1995) Rule induction and instance-based learning: a unified approach. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada, pp. 1226–1232.

[Gates 1972] Gates, G.W. (1972) The reduced nearest neighbor rule, IEEE Transactions on Information Theory, v. 18, no. 3, pp. 431–433.

[Hart 1968] Hart, P. E. (1968) The condensed nearest neighbor rule, IEEE Transactions on Information Theory, v. 14, pp. 515–516.

[Leyva *et al.* 2013] Leyva, E.; González, A.; Pérez, R. (2013) Knowledge-based instance selection: a compromise between efficiency and versatility, Knowledge-Based System, v. 47, pp. 65–76.

[Leyva *et al.* 2015] Leyva, E.; González, A.; Pérez, R. (2015) Three new instance selection methods based on local sets, Pattern Recognition, v. 48, pp. 1523–1537.

[Ritter *et al.* 1975] Ritter, G. L.; Woodruff, H. B.; Lowry, S. R.; Isenhour, T. L. (1975) An algorithm for a selective nearest neighbor decision rule, IEEE Transactions on Information Theory, v. 21, no. 6, pp. 665–669.

[Wilson 1972] Wilson, D. L. (1972) Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics SMC-2, pp. 408-421.

[Wilson & Martinez 1997] Wilson, D. R.; Martinez, T. R. (1997) Instance pruning techniques. In: Fisher, D. (Ed.), Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97), San Francisco, CA: Morgan Kaufmann Publishers, pp. 403–411.

[Wilson & Martinez 2000] Wilson, D. R.; Martinez, T. R. (2000) Reduction techniques for instance-based learning algorithms, Machine Learning, v. 38, pp. 257–286.