Análise de Métricas para Recuperação de Casos visando à aplicação no Problema de Condução de Trens de Carga

Matheus Streisky
Universidade Tecnológica Federal do Paraná
Av. Monteiro Lobato, s/n
Ponta Grossa – PR – Brasil
streisky@alunos.utfpr.edu.br

Richardson Ribeiro
Universidade Federal do Paraná
Departamento de Ciências Florestais, Campus Jd. Botânico – Curitiba – PR – Brasil
richardsonr@utfpr.edu.br

André Pinz Borges
Universidade Tecnológica Federal do Paraná
Av. Monteiro Lobato, s/n
Ponta Grossa – PR – Brasil
apborges@utfpr.edu.br

RESUMO

Este artigo propõe uma análise de métricas de similaridade para o algoritmo *k-NN*, a ser usado na etapa de recuperação de casos no problema de condução de trens de carga. A condução de trens é uma tarefa complexa que exige experiência e conhecimento das ações a serem executadas para uma viagem econômica. O conhecimento pode ser obtido de históricos de viagens armazenadas e reutilizadas no formato de experiências passadas (casos). Para tal, pode-se utilizar a técnica de Raciocínio Baseado em Casos (RBC), onde uma das etapas é a recuperação destas experiências. Na recuperação, métricas são usadas para avaliar qual dos casos passados mais se assemelham ao problema atual. Este artigo efetua uma análise de diferentes métricas de recuperação *k-NN* que podem ser usadas na etapa de recuperação. Dentre elas, a Distância de *Minkowski* mostrou-se eficiente para o problema estudado.

Palavras-chave

Raciocínio Baseado em Casos; *k-NN*; Condução de trens.

ABSTRACT

This article proposes an analysis of similarity metrics for *k-NN* algorithm, to be used in the case retrieval stage in the problem of conducting freight trains. Driving trains is a complex task that requires experience and knowledge of the actions to be performed for an economic trip. Knowledge can be obtained from stored travel histories and reused in the form of past experiences (cases). For this, the technique of Case-Based Reasoning (CBR) can be used, where one of the steps is the recovery of these experiences. In recovery, metrics are used to assess which of the past cases most resemble the current problem. This article examines the different *k-NN* recovery metrics that can be used in the recovery step. Among them, Minkowski's distance proved to be efficient for the problem studied.

Keywords

Case-based Reasoning; *k-NN*; Train driving.

1. INTRODUÇÃO

O Raciocínio Baseado em Casos (RBC) é um método de Inteligência Artificial que utiliza um conhecimento prévio adquirido em problemas passados para reaproveitar soluções bem sucedidas e evitar que falhas se repitam [1].

Em RBC descreve-se o conhecimento armazenado em memória como um caso. O caso consiste em representar um episódio em que um problema foi total ou parcialmente resolvido [6]. Em outras palavras, caso é um acontecimento problemático da vida de um ator ou objeto que pode ter sido solucionado ou não.

O ciclo do RBC é composto por 4 etapas principais (cf. Figura 1). Dado um novo problema, a partir da percepção de sensores de um ambiente, o objetivo é encontrar uma nova solução com uma consulta na base de casos (composta de vários casos) onde busca-se um ou mais casos semelhantes ao novo problema (geralmente os casos mais semelhantes são os melhores para o mesmo problema). É recuperada então, da base de casos, uma ou mais soluções para serem reutilizadas. Quando for necessário, é realizada uma adaptação (reuso) no caso escolhido para tentar solucionar o problema, visto que cada caso recuperado pode não ser completamente igual para todas as etapas do problema. Após a adaptação, o novo caso deve ser revisado para verificar se pode ser aplicado e, se resolver o problema de forma adequada, o caso é aplicado e retido na base de casos. A retenção dá-se para ajudar na resolução de problemas futuros. Se a solução adaptada não for aplicável, o caso deve ser novamente adaptado, modificando o caso [6].

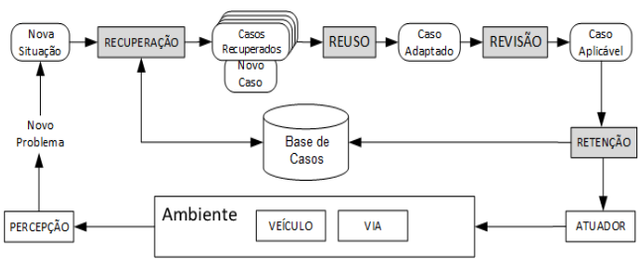

Figura 1 - Ciclo CBR [5].

Descobrir os casos com maior potencialidade é essencial para o sucesso de um sistema que usa RBC [6]. Logo, na etapa de recuperação de casos são usadas heurísticas e avaliações na seleção do caso mais semelhante, com o objetivo de diminuir o processo de busca. Esta etapa é importante no método RBC, pois filtra os casos mais semelhantes a um determinado problema em uma base de casos com *n* elementos [1].

É fundamental ter uma quantidade razoável de elementos registrados em uma base de casos, pois quanto mais casos estiverem disponíveis para recuperação, mais preciso será o resultado obtido nesta etapa. Bases de casos pequenas podem irão gerar resultados precisos, já que haverá uma gama de comparações menores a serem feitas ocorrendo, no pior caso, a recuperação de um resultado não satisfatório para o problema.

Entretanto, devido às heurísticas e avaliações necessárias na etapa

de recuperação de casos, uma base de casos grande traz consigo maior demanda de tempo e processamento, o que em algumas situações é inaplicável. Um dos desafios de usar algoritmos de recuperação eficientes é ter um bom desempenho quando aplicado a centenas ou milhares de casos sem diminuir sua precisão [8].
A recuperação também deve ser precisa e eficiente porque a mesma afeta as demais etapas presentes no método RBC. Para aplicar o reuso de casos, o caso recuperado passa por um processo de adaptação, modificando-o a fim de satisfazer o problema atual. Se o caso encontrado diferir do problema a ser solucionado, o processo de adaptação se torna demorado, pois é necessário "corrigir" as diferenças entre o caso e o problema.
Neste artigo serão analisadas métricas que podem ser usadas em conjunto com o algoritmo *k-NN* (*k-Nearest Neighbor*) utilizado na etapa de Recuperação. Tais métricas são usadas para calcular a distância, ou semelhança, entre o problema atual e os elementos de uma base de casos, onde *k* é o número de casos comparados ao problema atual. A distância de similaridade é uma medida numérica de quão parecidos dois casos/objetos são, sendo maior quando os casos/objetos são parecidos. O inverso é a medida de dissimilaridade, onde quanto maior seu valor, maior é a diferença entre dois casos/objetos [9].
Assim como mencionado por Zia *et al* (2014), *k-NN* é um dos algoritmos de recuperação mais difundidos pela literatura, pois possui uma fácil implementação e é menos susceptível a dados com ruídos. Dentre diversas técnicas de similaridade, a mais utilizada é a Distância Euclidiana para realizar a seleção dos vizinhos mais próximos. Porém, técnicas como a Distância do Cosseno e a Distância de *Mahalanobis* também podem ser usadas.

O presente artigo está organizado da seguinte forma. Na seção 2 há menções de trabalhos relacionados com métricas de avaliação e utilização do algoritmo *k-NN*. A seção 3 mostra a metodologia utilizada na realização deste artigo, analisando a importância da recuperação de casos e as vantagens e desvantagens de cada uma das métricas selecionadas. Logo em seguida, temos na seção 4 os resultados obtidos pela implementação destas métricas. Por fim na seção 5, as considerações finais e os trabalhos futuros.

2. TRABALHOS RELACIONADOS

RBC tem sido empregado em diversas pesquisas recentes. Vasconcelos (2016) comparou algoritmos de junção por similaridade em memória, utilizando o algoritmo *k-NN* para realizar consultas de similaridade entre dados multimídia (áudio, vídeo, imagem, textos longos, etc). Para calcular as similaridades foram utilizadas distâncias da família de *Minkowski* (Distância Euclidiana e de *Manhattan*). As consultas serviram para executar um filtro nas diversas informações contidas no banco de dados e recuperar os elementos mais similares entre si, para posteriormente executar um algoritmo de junção.
Na área de jogos, Bricce *et al* (2016) aplicou o algoritmo *k-NN* (aplicando Distância Euclidiana) para controlar os movimentos de *non-players characters* (NPCs) em um ambiente dinâmico, armazenando as jogadas do usuário para realizar o aprendizado de máquina. Isto permitiu que o NPC se torne mais inteligente de acordo com o tempo jogado pelo usuário, e essas informações armazenadas são recuperadas posteriormente para aplicar na jogabilidade do NPC.
Na condução de trens de carga, RBC é usado como uma técnica capaz de recuperar ações que maquinistas devem executar para percorrer determinados trechos de via, conforme descrito em Borges (2015). A técnica mostrou-se eficaz pois o compartilhamento das experiências dos maquinistas para conduzir um trem não é algo trivial de ser compartilhado e explicado. Logo, a partir de um conjunto de dados obtidos por sensores de locomotivas, é criada uma base de casos. A base é submetida ao ciclo de RBC, que retorna ações aplicáveis durante a condução.
Contudo, a eficácia de RBC passa pela recuperação de um conjunto de casos próximos aos dados lidos dos sensores. Se os casos recuperados forem distantes do problema atual, o esforço da etapa de adaptação aumenta. Sendo assim, a principal contribuição deste trabalho é analisar algumas métricas de similaridade a serem usadas na etapa de recuperação de casos, para verificar qual métrica melhor se encaixa no problema da condução de trens de carga.

3. METODOLOGIA

O que pode influenciar se a recuperação de casos será boa ou ruim é a escolha da métrica de similaridade aplicada nesta etapa. Esta escolha impacta em todo o sistema RBC. Na literatura não há um consenso sobre uma medida similaridade que seja aplicável a todos os tipos de variáveis que podem existir numa base de casos. Os pesquisados empregam medidas variadas, como: Distância Euclidiana, Distância Euclidiana Ponderada, Distância de *Manhattan* e Distância de *Minkowski*. Logo, tais medidas serão analisadas neste trabalho para uso futuro, sendo importante ressaltar que algumas medidas necessitam que os dados sejam normalizados antes da aplicação. A normalização consiste em colocar todas as variáveis que compõe um caso/objeto em uma mesma escala [9].

3.1 Distância Euclidiana

É uma medida de distância frequentemente empregada quando as variáveis são quantitativas [11]. Sua aplicação é recomendada para dados não padronizados, porém há uma desvantagem se houverem dados com diferença de escala (cm e km, por exemplo). No momento da transformação os resultados euclidianos sofrem uma influência das dimensões que possuem valores maiores. Apesar de ser uma distância de fácil entendimento e útil para aplicações de recuperação, ela não varia em um intervalo de 0 a 1 e, então, o valor obtido não é comparável aos valores alcançados por outras medidas de similaridade que variam entre 0 e 1. O cálculo desta distância é mostrado na Equação 1, onde *x* e *y* representam os atributos dos casos a serem comparados.

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1)$$

Um exemplo de utilização desta métrica é para calcular a rota de GPS de um helicóptero. Por se tratar de um veículo aéreo apenas um segmento de reta entre 2 pontos indica a distância a ser percorrida. Supondo que os pontos (4, 1) e (3, 7) representem *x* e *y* (origem e destino) respectivamente, realizando os cálculos tem-se a distância igual a 6,082.

3.2 Distância Euclidiana Ponderada

A Distância Euclidiana Ponderada é semelhante ao da Distância Euclidiana comum, porém com adição de pesos nas características dos dados relativos à sua importância. O cálculo é mostrado na Equação 2, onde *x* e *y* são os atributos dos casos a serem comparados e *w* é o peso atribuído à essas características.

$$d = \sqrt{\sum_{i=1}^{n} w_i\,(x_i - y_i)^2} \quad (2)$$

Exemplo: Um GPS de carro disponibiliza 2 opções de estrada, ambas de mesma distância, a percorrer do ponto de origem *x* (5, 5) para ponto destino *y* (2, 7). A primeira opção de estrada (*d1*) possui um pedágio de custo *w*=5, a outra estrada (*d2*) não possui pedágio, sendo assim assume valor de custo *w*=1. Realizando os cálculos tem-se os seguintes custos de cada estrada:

$$d1 = \sqrt{5.(5-2)^2 + 5.(5-7)^2} = \sqrt{65} = 8{,}062$$
$$d2 = \sqrt{1.(5-2)^2 + 1.(5-7)^2} = \sqrt{13} = 3{,}605$$

3.3 Distância de Manhattan

A Distância de *Manhattan*, possui funcionamento similar à Distância Euclidiana, mas não é influenciado pela diferença de escala sobre o resultado já que não há elevação ao quadrado das características a serem medidas.

Mesmo sendo mais fácil de calcular que as distâncias citadas anteriormente, esta métrica pode não ser adequada se os atributos estão correlacionados, porque não existe a garantia da qualidade dos resultados obtidos [4]. Para calcular a Distância de *Manhattan* utiliza-se da Equação 3, onde *x* e *y* representam os atributos dos casos a serem comparados.

$$d = \sum_{i=1}^{n} |x_i - y_i| \quad (3)$$

Para exemplificar, uma pessoa pega um taxi no centro de uma cidade no ponto *x* (1, 1) e deseja ir até o ponto *y* (3, 4). Calculando, obtemos a seguinte distância percorrida igual a 5.

3.4 Distância de *Minkowski*

Segundo Linden (2009), outra medida de distância utilizada é a Distância de *Minkowski*, que é uma generalização das distâncias Euclidiana e de *Manhattan*. A Distância de *Minkowski* pode ser calculada através da Equação 4, onde *x* e *y* representam os atributos dos casos a serem comparados e *q* é um inteiro positivo que, caso assuma valor igual a 1 resulta na Distância de *Manhattan*, caso assuma valor igual a 2 resulta na Distância Euclidiana e caso assuma valor igual a infinito resulta na Distância de *Chebyshev*, sendo esta última dada pela Equação 5.

$$d = \left(\sum_{i=1}^{n} |x_i - y_i|^q\right)^{\frac{1}{q}} \quad (4)$$

$$d = \max_k |x_k - y_k| \quad (5)$$

A escolha do valor de *q* depende exclusivamente da ênfase que se deseja dar a distâncias maiores. Quanto maior for o valor de *q*, maior será a sensitividade da métrica. Um exemplo simples é quando se calcula a distância entre os pontos *x* (0, 2, 1) e *y* (1, 18, 2). Usando diferentes valores de *q*, temos os seguintes valores para as distâncias, conforme Tabela 1.

Para aplicação do algoritmo *k-NN*, primeiro é necessário definir o valor de *k*. Recomenda-se que *k* possua um valor ímpar, assim a ocorrência de empates se tornam mais difícil de ocorrer. Será assumido os valores de *k*=3 e *k*=5 para um problema fictício a ser solucionado com valores iguais a [5.6, 3.7, 5.2, 0.4].

Tabela 1 – Cálculos das distâncias de Minkowski.

Valor de *q*	Cálculo
q=1	$d = [\|1-0\| + \|18-2\| + \|2-1\|] = 18$
q=2	$d = \sqrt{(1-0)^2 + (18-2)^2 + (2-1)^2} = 16{,}06$
q=3	$d = \sqrt[3]{(1-0)^3 + (18-2)^3 + (2-1)^3} = 16{,}002$
q=∞	$d = \sqrt[\infty]{(1-0)^\infty + (18-2)^\infty + (2-1)^\infty} = 16$

Estes valores contêm informações correspondentes a: velocidade, pressão dos freios, peso total dos vagões (em toneladas) e inclinação (subida, descida ou reta). Para aplicação das métricas no algoritmo *k-NN*, será utilizado uma base de casos contendo 150 elementos com valores simuladas de condução de trens de carga

Na utilização da Distância Euclidiana e da Distância de *Manhattan*, não há outros fatores a serem declarados, apenas a aplicação normal de suas fórmulas. Já a Distância Euclidiana Ponderada, será assumido os seguintes valores de *w* (peso) em cada uma das características dos casos: 4 para velocidade, 5 para a pressão dos freios, 3 para peso dos vagões e 1 para inclinação. Os valores foram atribuídos apenas para teste de forma empírica.

Na Distância de *Minkowski*, *q* (grau de sensitividade) assumirá valor igual ao infinito, pois se aplicado valores iguais a 1 ou 2, não serão obtidos resultados diferentes da Distância Euclidiana ou da Distância de *Manhattan*.

4. RESULTADOS

As tabelas 2 e 3 mostram os resultados do algoritmo *k-NN* para cada uma das métricas analisadas, onde *d* é o valor de dissimilaridade entre o problema a ser solucionado e os casos recuperados da base de casos.

Conforme possível analisar, os 3 primeiros casos recuperados não diferem quando o *k* assume valor igual a 3 ou igual a 5 em nenhuma das métricas aplicadas no algoritmo. Entretanto isso não significa que um valor maior que 3 não deve ser aplicado à *k*, pois se na pior das hipóteses nenhum dos 3 primeiros casos satisfazerem o problema, há mais opções de casos para reuso.

Também é possível analisar que mesmo métricas diferentes recuperaram casos iguais, porém em diferentes prioridades. A Distância Euclidiana Ponderada recupera o caso [6.0, 3.4, 4.5, 1.6] como o mais similar ao problema proposto, assim como a Distância de *Manhattan* recupera o mesmo caso como o terceiro mais similar. O mesmo ocorre no caso [6.1, 2.8, 4.7, 1.2], onde a Distância Euclidiana o indica como mais similar e a Distância de *Minkowski* o indica como o segundo mais similar. Para uma visão mais geral do problema, na Tabela 4 são representadas as médias de similaridade de cada métrica aplicada no problema.

A Distância de *Minkowski* obteve as menores médias de dissimilaridade entre todas as métricas analisadas, seguidas das distâncias Euclidiana, Euclidiana Ponderada e de *Manhattan*. Sendo assim, *Minkowski* se mostra como a métrica mais precisa e *Manhattan* como a menos precisa, para resolução deste problema dentre as quatro métricas investigadas.

5. CONSIDERAÇÕES FINAIS

O método RBC é uma ferramenta versátil e pode ser aplicada para diversas funcionalidades, sendo utilizada também em outras áreas fora a da computação, podendo ser moldada de acordo com as necessidades de quem a desenvolve.

Tabela 2 - Resultados obtidos com k=3.

Métrica	*k* = 3	
	Casos	*d*
Distância Euclidiana	[6,1; 2,8; 4,7; 1,2]	1,396
	[6,1; 3,0; 4,6; 1,4]	1,449
	[5,7; 3,0; 4,2; 1,2]	1,449
Distância Euclidiana Ponderada	[6,0; 3,4; 4,5; 1,6]	2,000
	[5,9; 3,0; 5,1; 1,8]	2,190
	[5,4; 3,0; 4,5; 1,5]	2,300
Distância de *Manhattan*	[5,6; 3,0; 4,5; 1,5]	2,500
	[5,9; 3,0; 5,1; 1,8]	2,500
	[6,0; 3,4; 4,5; 1,6]	2,600
Distância de *Minkowski*	[5,7; 2,8; 4,5; 1,3]	0,900
	[6,1; 2,8; 4,7; 1,2]	0,900
	[6,4; 2,9; 4,3; 1,3]	0,900

Tabela 3 - Resultados obtidos com k=5.

Métrica	*k* = 5	
	Casos	*d*
Distância Euclidiana	[6,1; 2,8; 4,7; 1,2]	1,396
	[6,1; 3,0; 4,6; 1,4]	1,449
	[5,7; 3,0; 4,2; 1,2]	1,449
	[5,4; 3,0; 4,5; 1,5]	1,493
	[5,7; 2,9; 4,2; 1,3]	1,568
Distância Euclidiana Ponderada	[6,0; 3,4; 4,5; 1,6]	2,000
	[5,9; 3,0; 5,1; 1,8]	2,190
	[5,4; 3,0; 4,5; 1,5]	2,300
	[6,1; 3,0; 4,6; 1,4]	2,351
	[6,0; 3,0; 4,8; 1,8]	2,351
Distância de *Manhattan*	[5,6; 3,0; 4,5; 1,5]	2,500
	[5,9; 3,0; 5,1; 1,8]	2,500
	[6,0; 3,4; 4,5; 1,6]	2,600
	[5,6; 3,0; 4,1; 1,3]	2,700
	[5,5; 2,6; 4,4; 1,2]	2,799
Distância de *Minkowski*	[5,7; 2,8; 4,5; 1,3]	0,900
	[6,1; 2,8; 4,7; 1,2]	0,900
	[6,4; 2,9; 4,3; 1,3]	0,900
	[6,2; 2,9; 4,3; 1,3]	0,900
	[5,6; 2,7; 4,2; 1,3]	1,000

Tabela 4 - Médias de dissimilaridade.

Métrica	*k* = 3	*k* = 5	Média (*k*=3 e *k*=5)
Distância Euclidiana	1,431	1,471	1,451
Distância Euclidiana Ponderada	2,163	2,238	2,201
Distância de Manhattan	2,533	2,620	2,577
Distância de Minkowski	0,900	0,920	0,910

A recuperação de casos é uma etapa importante em todo método RBC, pois esta etapa pode determinar se o problema enfrentando terá uma boa solução ou não. Implementando uma boa recuperação, a etapa de adaptação, se ocorrer, necessitará de menor esforço computacional e isso resultará em um método rápido e otimizado.

As métricas de dissimilaridade aplicadas no problema também impactam no desempenho do método, por mais que as diferenças entre as métricas analisadas neste trabalho sejam pequenas (uma diferença de 1,667 pontos de dissimilaridade entre a melhor e a pior métrica), em bancos de casos com milhares ou milhões de casos, tais métricas podem apresentar uma diferença significativa de dissimilaridade. É tarefa do desenvolver escolher a métrica a ser empregada em seu algoritmo, adotando a qual se encaixar melhor para solucionar o problema enfrentado.

Como este artigo propõe uma análise de diferentes métricas voltadas ao algoritmo *k-NN*, é interessante também examinar, em trabalhos futuros, outras métricas de dissimilaridade não citadas neste artigo no mesmo algoritmo. Outro trabalho futuro é analisar tais métricas utilizando outros algoritmos de recuperação existentes, os aplicando ao problema de condução de trens.

6. REFERÊNCIAS

[1] Aamodt, A.; Plaza, E. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. In: AICOM – Artificial Intelligence Communications. Vol. 7. Ed. IOS Press, EUA, 1994.

[2] Borges, A., P. Uma contribuição para geração de políticas de ações para condução de trens de carga usando Raciocínio Baseado em Casos. 2015. 220 f. Dissertação (Doutorado) – PUC-PR. Curitiba, 2015.

[3] Bricce, B. R., Silva, P. P., Martins, H. P., Silva, E. G. Aplicação do algoritmo knn para controle de movimentos de npcs em um ambiente dinâmico (jogo). REGRAD - Revista Eletrônica de Graduação do UNIVEM - ISSN 1984-7866, [S.l.], v. 9, n. 1, p. 18-32, Aug. 2016. ISSN 1984-7866.

[4] Cole, R., M. Clustering with Genetic Algorithms, M. Sc., Department of Computer Science, University of Western Australia, Australia, 1998.

[5] Corchado, J. M., Lees, B., Aiken, J. Hybrid instance-based system for predicting ocean temperature. International Journal of Computational Inteligence and Applications, v. 1, n. 1, p. 35-52, 2001.

[6] Kolodner, J. Reconstrutive memory: a computer model. Cognitive Science 7(4), 1983 apud KOLODNER 1993.

[7] Linden, R. Técnicas de Agrupamento. Revista de Sistemas de Informação de FSMA. n. 4, pp. 18-36, 2009.

[8] Abel, M. Um estudo sobre Raciocínio Baseado em Casos. 1996. 41 f. Dissertação (Pós-Graduação em Ciência da Computação) – UFRGS. Porto Alegre, 1996.

[9] Mitchell, T. Machine Learning. McGraw Hill, 1997.

[10] Vasconcelos, G. Q. Estudo comparativo de algoritmo de junção por similaridade em memória. 2016. 57 f. Dissertação (Graduação) – Curso de Bacharelado em Ciência da Computação, UEL, Londrina, 2016.

[11] Zia, S. S., Akhtar, P., Mughal, T. J., & Mala, I. Case Retrieval Phase of Case-Based Reasoning Technique for Medical Diagnosis. World Applied Sciences Journal, 32(3), 451-458. 2014.