Desenvolvimento de um Algoritmo de Classificação de Dados Baseado na Relevância da Informação

Júlio César Guimarães Costa[a,b], Gleifer Vaz Alves[a], André Pinz Borges[a]

[a] *Grupo de Pesquisa, LaCA-IS, Departamento Academico de Informática, Universidade Técnologica Federal do Paraná (UTFPR), R. Doutor Washington Subtil Chueire, 330, Paraná, Brasil*
[b] *Autor para correspondência: juliocesargcosta123@gmail.com*

Resumo

Algoritmos de classificação são uma ferramenta poderosa, principalmente nos dias de hoje, quando os dados estão se tornando cada vez mais valiosos. Este artigo propõe um algoritmo chamado Prevh que baseia-se em uma nova forma de se visualizar as informações. Para tal foi escolhida uma metodologia qualitativa, básica e explicativa. Depois de ser desenvolvido o algoritmo Prevh, foi submetido a uma série de comparações com o algoritmo de classificação KNN o qual mais se assemelha. Uma delas pode ser inclusive observada na seção 4 deste artigo. Por fim foi observado que além de considerar a relevância da informação, o algoritmo Prevh foi capaz de ser conclusivo em situações nas quais o algoritmo KNN não foi.

Abstract

Classification algorithms are a powerful tool, especially today, when data is becoming increasingly valuable. This paper proposes an algorithm called Prevh, which is based in a new way of visualize information. For this purpose, a qualitative, basic and explanatory methodology was chosen. Initially, we have developed the algorithm, next we have run a series of comparisons between Prevh and KNN classification algorithm, since both share some similarities. In Section 4, we describe such similar features. Finally, we have observed that in addition to considering the information relevance, Prevh was able to be conclusive in some scenarios where KNN was not.

Palavras-chaves: Aprendizagem de Maquina, Mineração de Dados, Classificação de Dados

1. Introdução

Aplicações e tecnologias que envolvem problemas de classificação são mais comuns do que podemos imaginar nos dias de hoje, e estão presentes em uma infinidade de algoritmos que envolvam Aprendizagem de Máquina em geral [1]. Alguns exemplos podem ser notados em algoritmos de plataformas como Facebook, Amazon, Google, entre outros. Estas empresas trabalham com uma grande quantidade de dados e se utilizam dessas importantes ferramentas em traçamentos de perfil de clientes, classificação de categorias, entre outras aplicações [2].

Neste artigo é proposto o algoritmo Prevh que se mostrou uma ferramenta poderosa quando utilizada em conjunto com um outro algoritmo de classificação já consolidado, chamado KNN. Ambos algoritmos têm o objetivo de classificar uma entidade X que hipoteticamente surja em um espaço n-dimensional definido por K entidades. Será possivel observar que as situações de aplicação do algoritmo Prevh ocorrem principalmente em situações em que o algoritmo KNN é incerto quanto a classificação, o que será demonstrado e comprovado no decorrer do artigo.

Além da aplicação do algoritmo a relevância da informação pode ser definida como um valor ou peso que é calculado e atribuído tanto ao dado de treinamento quanto no dado que se pretende classificar, e que representa justamente a veracidade da informação, este recurso pode ser e será utilizado na previsão de diferentes cenários sequenciais definidos por processos estocásticos [3] que serão explorados em trabalhos futuros.

2. Definição do procedimento matemático

2.1. Terminologia

Inicialmente, apresenta-se a terminologia de alguns elementos usados neste trabalho.

Cluster C = Agrupamento de pontos representantes de um mesmo resultado.
Entidade = Ponto de coordenadas (x,y,...,n) posicionado no espaço n-dimensional.
Relevância = O quanto uma entidade representa um determinado *cluster*, variando de 0 a $\sqrt{nD}$, onde nD é o número de dimensões do sistema.
Entidade X = Entidade que se pretende classificar.
Entidade a = Entidade representante de um *cluster* C, proveniente dos dados de entrada.

2.2. Introdução ao procedimento matemático

Buscando facilitar o entendimento do conceito matemático abordado neste artigo é interessante que seja feita uma breve referência ao seguinte conceito - "Dois corpos atraem-se por uma força que é diretamente proporcional ao produto de suas massas e inversamente proporcional ao quadrado da distância que os separa" [4], partindo de tal conceito foi feita uma associação entre a massa e a informação, e consequentemente a força de atração e a relevância, foi então definido que a relevância da informação é diretamente proporcional ao índice de pertinência de uma entidade a um determinado *cluster*, e portanto ambas são inversamente proporcionais a distância que as separa de outras entidades.

Para a aplicação do algoritmo, é estritamente necessário que o universo estudado (incluso a entidade X) seja normalizado entre os valores 0 e 1, para que seja possível calcular a maior distância euclidiana no espaço. Tal regra pode ser observada com a aplicação do teorema de Pitágoras no espaço n-dimensional [5], e este valor será utilizado como referência para a maior relevância que uma instância pode ter no universo estudado.

3. Fórmulas matemáticas

A partir dos conceitos apresentados foram definidas as fórmulas a seguir:

Definição 1 (Média Ponderada da soma das Distâncias Euclidianas)**.**

$$Mdp^{K}_{X \to eC} = \frac{\sum_{a=1}^{neC} \sqrt{\sum_{i=1}^{nD} (Cx - Ca)^2} * R_a}{neC} \tag{1}$$

onde:

$Mdp^{K}_{X \to eC}$ = *Média Ponderada da soma das Distâncias Euclidianas.*
neC = *Número de entidades representantes do cluster* C*.*
a = *entidade do cluster* C*.*
nD = *Número de dimensões do sistema.*
Cx = *Coordenada da entidade* X*.*
Ca = *Coordenada da entidade* a*.*
Ra = *Relevância da entidade a que é definida nos dados de treinamento baseando-se na veracidade da informação.*

A média ponderada da soma das distâncias euclidianas entre a entidade X *e as entidades* A *representantes de um cluster* C *em um espaço delimitado por* K*, é calculada pelo somatório da distância euclidiana entre* X *e* A *representante de* C*, multiplicado por Ra. Tudo dividido pelo número de entidades de* C*.*

Definição 2 (Índice de pertinência de X em C no espaço K)**.**

$$P^{K}_{X \to C} = 1 - \frac{Mdp^{K}_{X \to eC}}{\sum_{i=1}^{neK} Mdp^{K}_{X \to eC}} \tag{2}$$

onde,

$P^{K}_{X \to C}$ = *Índice de Pertinência de* X *em* C *no espaço* K*.*
$Mdp^{K}_{X \to eC}$ = *Média Ponderada das Distâncias Euclidianas.*
neK = *Número de entidades no espaço delimitado por* K*.*

O índice de pertinência de uma entidade X *em* C *em um espaço delimitado por* K*, é definida por 1 menos a Média Ponderada da soma das Distâncias Euclidianas sobre a soma de todas as MDPs entre a entidade* X *e todas entidades no espaço delimitado por* K*.*

Definição 3 (Calculo da relevância de X em C no espaço K)**.**

$$R^{K}_{X \to C} = \sqrt{nD} * P^{K}_{X \to C} \tag{3}$$

onde,

nD = *Número de dimensões do sistema.*
$P^K_{x \to C}$ = *Índice de Pertinência de* X *em* C *no espaço* K.
$R^K_{x \to C}$ = *Relevância de* X *em* C *no espaço* K.

A relevância da entidade X *para um determinado cluster* C *em um espaço delimitado por* K, *é definida pela multiplicação entre a raiz do número de dimensões* nD *pelo índice de pertinência de* X *em* C *no espaço* K.

4. Comparativo entre o algoritmo de classificação KNN e o algoritmo Prevh

Foi proposto o seguinte teste comparativo: em um sistema de duas dimensões normalizado entre 0 e 1, pretende-se classificar uma entidade X de coordenadas (0.5,0.5), para um espaço delimitado por $K = 3$, e para tal foram aplicados, primeiramente o algoritmo KNN e posteriormente o algoritmo Prevh, após os resultados obtidos foram comparados. Os dados de entrada são apresentados na Tabela 1 e Figura 1.

Ca_x	Ca_y	Ca_C	Relevância (Ra)	Distância Euclidiana
0,1	0,1		1,41	0,56
0,1	0,2		1,41	0,5
0,4	0,4		1,41	0,14
0,4	0,9		1,41	0,41
0,5	0,8		1,41	0,3
0,5	0,6		1,41	0,1
0,6	0,4		1,41	0,14
0,8	0,3		1,41	0,36
0,9	0,7		1,41	0,44

Tabela 1: Representação descritiva dos dados de teste incluindo a distância euclidiana já calculada entre as instâncias e a instância que se pretende classificar X(0.5,0.5), com os K=3 vizinhos mais próximos marcados em amarelo.

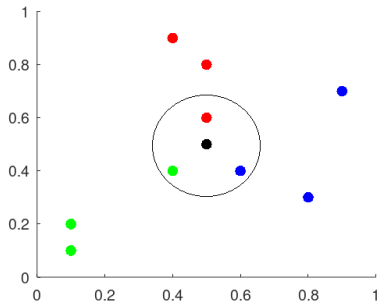

Figura 1: Representação gráfica dos dados de teste com a área de atuação limitada a K=3.

4.1. Aplicando o algoritmo KNN

Ao ser aplicado, o algoritmo KNN [6] para K = 3 foram obtidos os seguintes resultados:
0,33 ou 33% de chance de X ser uma entidade da cor azul.
0,33 ou 33% de chance de X ser uma entidade da cor vermelha.
0,33 ou 33% de chance de X ser uma entidade da cor verde.

4.2. Aplicando o algoritmo Prevh

Ao serem aplicadas as fórmulas matemáticas descritas anteriormente neste artigo, os resultados ilustrados na Tabela 2 foram obtidos.

C_c	$Mdp^K_{X \to eC}$	$P^K_{X \to C}$	$R^K_{X \to C}$
	0,1974	0,6316	0,8905
	0,141	0,7369	1,0390
	0,1974	0,6316	0,8905

Tabela 2: Resultados obtidos pelo algoritmo Prevh

0,6316 ou 63% de chance de X ser uma entidade da cor verde.
0,7369 ou 73% de chance de X ser uma entidade da cor vermelha.
0,6316 ou 63% de chance de X ser uma entidade da cor azul.

4.3. Conclusão dos testes

É possível perceber que enquanto o algoritmo KNN, foi inconclusivo para $K = 3$, o algoritmo Prevh conseguiu propor uma classificação vermelha a entidade X. Este fator ocorreu devido a diferente conotação que é dada a distância euclidiana em ambos os algoritmos, pois por mais que a relevância das informações presentes em $K = 3$ seja a mesma, o algoritmo Prevh utilizou-se diretamente da distância euclidiana no cálculo do índice de pertinência, enquanto o algoritmo KNN utilizou-se da distância euclidiana somente para definir o espaço K limitante.

5. Considerações Finais

As aplicações do algoritmo Prevh mais visíveis são em situações semelhantes as aplicações do algoritmo KNN, tanto quanto em situações que envolvam dados discretos definidos por processos estocásticos [3], pois, a relevância atribuída ao dado pode contribuir para a sequencialidade dos dados no tempo possibilitando a previsão de diversas direções para os quais os dados podem evoluir.

Referências

[1] A. Soofi, A. Awan, Classification Techniques in Machine Learning: Applications and Issues, Journal of Basic & Applied Sciences 13 (2017) 459 – 465. URL: https://pdfs.semanticscholar.org/2678/e213cec548d278879ceaf01582ee8913cc3f.pdf.

[2] B. Marr, M. Ward, Artificial Intelligence in Practice: How 50 Successful Companies Used AI and Machine Learning to Solve Problems, Wiley, 2019.

[3] M. KAC, J. LOGAN, Chapter 1 - fluctuations based in part on a series of lectures given at l'ecole polytechnique federale at lausanne, switzerland as a part of their troisieme cycle, summer 1974., em: E. MONTROLL, J. LEBOWITZ (Eds.), Fluctuation Phenomena, Elsevier, 1979, pg. 1–60.

[4] I. Newton, J. Rodrigues, F. C. Gulbenkian, S. de Educação e Bolsas, Princípios matemáticos da filosofia natural, Fundação Calouste Gulbenkian Serviço de Educação e Bolsas, Lisboa, 2010. OCLC: 959175701.

[5] J. Bezerra, Teorema de pitagoras no espaço, Revista do Professor de Matemática 79 (2012) 22–24.

[6] T. M. Mitchell, Machine learning, McGraw-Hill series in Computer Science, mcgraw-hill science/engineering/math; 1ª edição (1 março 1997) ed., McGraw-Hill, Porland, 1997.