Extensão da medida de similaridade de Gower para dados categóricos ordinais

1 Definição do problema

Como medir a similaridade de dados categóricos ordinais?

Será utilizada como meio de partida e referência a medida de similaridade de Gower, que é dada pela equação

$$S_{ij} = \frac{\sum_{k=1}^{N} w_{ijk} s_{ijk}}{\sum_{k=1}^{N} w_{ijk}}$$

Ela pode ser aplicada a alguns tipos de variáveis, incluindo as qualitativas e quantitavas.

- Qualitativas (variáveis categóricas nominais e ordinais): para as nominais, a função de similaridade s_{ijk} em relação ao atributo k entre o objeto i e j é 1 se i e j concordam em relação ao valor do atributo k, e 0 caso contrário.
- Quantitativas (variáveis numéricas discretas e contínuas): para as variáveis com valores $x_1, x_2, ..., x_n$, a função de similaridade é $1 - |x_i - x_j|/R_k$. Sendo R_k o maior valor do intervalo do atributo k.

Em sua forma mais simples, a similaridade entre variáveis categóricas ordinais poderia ser uma alteração da fórmula das quantitativas, onde x_i seria o ranking atual do atributo k do objeto i, r_{ik}, e x_j, r_{jk}, resultando em $1 - |r_{ik} - r_{jk}|/R_k$. Essa solução, entretanto, não resolve o problema proposto completamente, porque ao comparar o objeto i ao j, necessita-se de um valor representando o quanto i é acima ou igual a j, e não o módulo da diferença.

Uma exemplificação poderá ajudar a entender. Uma empresa E abre inscrições para uma vaga de trabalho onde os requisitos consistem em ter uma proficiência avançada em Excel, intermediária em PowerPoint e nada para Word. Supondo que vários candidados aplicaram, como saber qual é o melhor?

No decorrer do texto será feito referência a essa exemplificação.

2 Propostas

Quatro propostas para a fórmula de similaridade foram criadas de forma incremental.

Todas as propostas trabalham com a transformação textual do nível de proficiência para um valor em uma escala, por exemplo, dado os níveis de proficiência "nada", "básico", "intermediário" e "avançado", atribuí-se um valor entre 1 e 4 para cada nível, resultando em 1 - "nada", 2 - "básico", 3 - "intermediário" e 4 - "avançado".

Tabela 1. Níveis de proficiência para a vaga da empresa E

	Excel	PowerPoint	Word
Empresa E	avançado (4)	intermediário (3)	nada (1)

Tabela 2. Níveis de proficiência dos candidatos A, B e C

	Excel	PowerPoint	Word
Candidato A	básico (2)	intermediário (3)	avançado (4)
Candidato B	avançado (4)	intermediário (3)	básico (2)
Candidato C	nada (1)	básico (2)	nada (1)

2.1 Medida de similaridade

A primeira abordagem consiste na atribuição de 1 ou 0 com base nos rankings de cada objeto de cada atributo e o nível de dificuldade do requisito. Dado o atributo k presente nos objetos i e j, e comparando i a j (candidato com empresa), s_{ijk} é definido como

$$s_{ijk} = \begin{cases} 0, & \text{if } \frac{r_{ik}}{R_k} < \frac{r_{jk}}{R_k} \\ 1, & \text{if } \frac{r_{ik}}{R_k} \geq \frac{r_{jk}}{R_k} \end{cases}$$

onde r_{ik} é ranking atual do atributo k do objeto i e R_k é o maior valor do intervalo (rank) do atributo k.

Com essa medida é possível ter um valor de match com a vaga de cada candidato, porém, caso seja necessário escolher o melhor candidado dentre os que deram match, uma outra medida para calcular o conhecimento, ou habilidade, a mais terá que ser definida.

2.2 Medida de habilidade

Para começar a medir o quanto a mais um candidato sabe, primeiro é calculado o quantidade de habilidade. A função de habilidade é dada por

$$h_{ik} = \frac{r_{ik}}{R_k}$$

Tabela 3. Habilidades dos candidatos A, B e C

	Excel	PowerPoint	Word
Candidato A	2/4	3/4	4/4
Candidato B	4/4	3/4	2/4
Candidato C	1/4	2/4	1/4

Tabela 4. Habilidades (nível de dificuldade) da empresa E

	Excel	PowerPoint	Word
Empresa E	4/4	3/4	1/4

Em relação à empresa, a média das habilidades (requisitos) é o nível de dificuldade da vaga da empresa.

2.3 Medida de mérito - versão 1

Com a função de habilidade definida, é então possível calcular o conhecimento a mais dos candidatos, e essa operação será feita pela função de mérito.

Um valor de mérito positivo (entre 0 e 1) significa que o candidato provavelmente deu match em todos os requisitos, e caso contrário, provavelmente não.

Nessa primeira versão, a medida é definida pela subtração da média das habilidades do candidato com o nível de dificuldade da empresa - a média das "habilidades" da empresa.

$$M_{ij} = M_i - M_j = \frac{\sum_{k=1}^{N} h_{ik}}{N} - \frac{\sum_{k=1}^{N} h_{jk}}{N}$$

onde i é o candidato e j é a empresa.

Média das habilidades dos candidatos e empresa

$$\text{Candidato A} \Rightarrow \frac{\frac{2}{4} + \frac{3}{4} + \frac{4}{4}}{3} = \frac{3}{4}$$

$$\text{Candidato B} \Rightarrow \frac{\frac{4}{4} + \frac{3}{4} + \frac{2}{4}}{3} = \frac{3}{4}$$

$$\text{Candidato C} \Rightarrow \frac{\frac{1}{4} + \frac{2}{4} + \frac{1}{4}}{3} = \frac{1}{3}$$

$$\text{Empresa} \Rightarrow \frac{\frac{4}{4} + \frac{3}{4} + \frac{1}{4}}{3} = \frac{2}{3}$$

Tabela 4. Méritos (v1) dos candidatos A, B e C

	Mérito
Candidato A	3/4 - 2/3 = 0.083
Candidato B	3/4 - 2/3 = 0.083
Candidato C	1/3 - 2/3 = -0.333

Essa medida tem uma falha, e é que ao nível do atributo não está sendo feito uma comparação de match. Como a média das habilidades é calculada e depois o nível de dificuldade da empresa subtraído, é possível que um candidato tenha rankings altos em alguns atributos e dar match nestes, mas em outros ter um valor baixo e não dar match, porém com a média, essa informação se perde e ocorre um contrabalanço dos valores. Isso é evidenciado no valor do mérito do Candidato A e B na tabela 4, onde o Candidato A no atributo Excel não deu match mas no atributo Word, sim; e o Candidato B que deu match em Excel e em Word, porém a medida do mérito das duas são iguais. O candidato B deveria ter mais mérito que o A.

Um outra medida de mérito é então apresentada para solucionar esse problema.

2.4 Medida de mérito - versão 2

Essa medida visa solucionar o problema evidenciado na medida anterior, e faz isso introduzindo um conceito de peso e removendo a comparação das habilidades ao nível de média.

Nela, para o mérito de cada atributo, a habilidade do candidato é multiplicada pelo nível de "habilidade" da empresa, denomidado de peso, e esse peso ao quadrado é subtraído depois. Como a média das habilidades do candidato é multiplicada pelo peso do atributo da empresa, o termo subtraído também deve ser multiplicado pelo peso, mas como o peso é igual à habilidade da empresa, esse termo se torna o peso ao quadrado.

$$M_{ij} = \frac{\sum_{k=1}^{N} h_{jk}h_{ik} - h_{jk}^2}{\sum_{k=1}^{N} h_{jk}} = \frac{\sum_{k=1}^{N} w_{ijk}h_{ik} - w_{ijk}^2}{\sum_{k=1}^{N} w_{ijk}}$$

Esse equação é a mesma de Gower porém com a subtração do peso ao quadrado no numerador.

Méritos dos candidatos11

$$\text{Candidato A} \Rightarrow \frac{(\frac{4}{4}\frac{2}{4} - \frac{4}{4}^2) + (\frac{3}{4}\frac{3}{4} - \frac{3}{4}^2) + (\frac{1}{4}\frac{4}{4} - \frac{1}{4}^2)}{\frac{4}{4} + \frac{3}{4} + \frac{1}{4}} = -0.156$$

$$\text{Candidato B} \Rightarrow \frac{(\frac{4}{4}\frac{4}{4} - \frac{4}{4}^2) + (\frac{3}{4}\frac{3}{4} - \frac{3}{4}^2) + (\frac{1}{4}\frac{2}{4} - \frac{1}{4}^2)}{\frac{4}{4} + \frac{3}{4} + \frac{1}{4}} = 0.031$$

$$\text{Candidato C} \Rightarrow \frac{(\frac{4}{4}\frac{1}{4} - \frac{4}{4}^2) + (\frac{3}{4}\frac{2}{4} - \frac{3}{4}^2) + (\frac{1}{4}\frac{1}{4} - \frac{1}{4}^2)}{\frac{4}{4} + \frac{3}{4} + \frac{1}{4}} = -0.469$$

Tabela 5. Méritos (v2) dos candidatos A, B e C

	Mérito
Candidato A	-0.156
Candidato B	0.031
Candidato C	-0.46875

Observa-se que o Candidato A e B não têm méritos iguais, sendo B maior que A. No caso do Candidato C e B, ambos não têm mérito positivo, e assim se entende que não deram match com a maioria dos requisitos ou que os que deram match o peso da empresa é muito baixo para poder influenciar positivamente o resultado final.

Na escala de [-1, 0], -1 significa que o candidato não teve match em nenhum requisito e os níveis de proficiência são os menores possíveis, e na medida que o mérito se aproxima de 0, quase ocorreu match em todos os requisitos, porém em alguns não. Na escala de [0, 1], 0 significa que teve match e há o mínimo de nível de proficiência necessário, e na medida que o mérito se aproxima de 1, o candidado tem match e um nível de proficiência muito alto em todos os requisitos.