Agrupamento de dados sobre cartas de Pokémon TCG com base em seus atributos e relações

Eric Pavarim Lima
IMECC
RA : 247471

18 de Junho de 2024

1 Introdução

O mundo dos card games é uma febre mundial há muitos anos desde jogos antigos e um dos mais populares é o conjunto de cartas inspirado na japonesa franquia Pokémon que reunie inúmeros fãs, tanto colecionadores quanto competidores em torneios esportivos do jogo.
O presente projeto tem como foco o estudo e revisão de métodos de aprendizado de máquina não-supervisionado, com foco em dados de rede, ou seja, dados que possuem alguma relação intrínseca. Para tal objetivo faz-se necessário estudo da teoria de novos métodos de clusterização mais sofisticados. [1].
Para essa análise será utilizado o método de Schotaschic Block Model (abreviado para SBM) que consiste na atribuição de elementos a certos blocos (clusters) seguindo um modelo estocástico com probabilidades de pertencimento e uma comparação póstuma com atribuição de etiquetas (como se houvesse a supervisão dos dados).
O conjunto de dados conta com 22 observações e foi extraído por meio de uma API do Pokémon TCG [1] (com uso de pacotes como jsonlite e httr para extração das cartas). Além disso, apenas uma expansão (conjunto de cartas temático lançado periodicamente) foi utilizada, com nome de *Double Crisis*, pois é um conjunto relativamente menor causando, consequentemente, uma melhor visualização dos dados embora qualquer outra expansão tenha tido resultados similares.
As variáveis existentes são inúmeras pois cada carta contém diversas informações como ataques, habilidades, cotação de mercado etc., entretanto, as mantidas foram : name (indicando o nome da carta), hp (indicando a quantidade de vida) e rarity (indicando a raridade da carta).

2 Materiais e métodos

A principal diferença na abordagem de dados relacionados é a necessidade de quantificar as interações entre elementos.
Para isso, faz-se necessário uma matriz A chamada de matriz de adjacências que é uma métrica de relações em que o elemento a_{ij} indica a relação do elemento i com o elemento j. No geral, trata-se de elementos binários, i.e, a_{ij} é igual a 1 se o elemento i interage com o elemento j e 0 caso não ocorra tal interação, embora não seja regra e possam existir matriz de adjacência ponderadas (os valores usualmente variando entre 0 e 1 quantificam quão forte é a relação).
No banco de dados em questão, fora criada uma função (para esse projeto) que relaciona as cartas pokémon i e j e seus respectivos pontos de vida HP_i e HP_j da seguinte forma:

$$\text{- } a_{ij} = 0 \text{ se } |\mathrm{HP}_i - \mathrm{HP}_j| \geq \tfrac{1}{22} \sum_{k=1}^{22} \mathrm{HP}_k$$
$$\text{- } a_{ij} = 1 \text{ caso contrário.}$$

isto é, o valor absoluto da diferença entre pontos de vida ser maior que um limiar (a média geral nesse caso), portanto, pontos de vida "distantes" e não ocorre a interação entre atores e vice-versa. A binaridade foi escolhida pois o pacote 'mixer' (de interesse) trabalha apenas com matrizes de adjacência binárias.

Voltando para a parte teórica, após a definição da matriz de adjacências A, ocorre a parte da modelagem estocástica, de fato. Assumindo que haja K blocos (clusters) em uma população, a probabilidade de um elemento pertencer a um bloco k é dada por τ_k. Além disso temos um vetor:

$$z = (z_{i1}\, z_{i2} \ldots z_{iK})^t$$

que é uma indicadora do pertencimento do elemento i aos blocos, sendo $z_{ik} = 0$ se i não pertencer ao bloco k e vice-versa.
O modelo, além disso, assume Θ uma matriz $K \times K$ chamada matriz de interação entre blocos em que os elementos θ_{kh} são as probabilidades de um ator no bloco k estar relacionado a outro no bloco h. De forma sucinta, se a diagonal de Θ for maior que qualquer elemento da respectiva linha, isso é um indício que o modelo tem clusteres em comunidades, isto é, a relação intrínseca (dentro do cluster) é maior do que relações de elementos entre clusteres. Daí:

$$P\{y_{ij} = 1|\Theta, z_{ik} = 1, z_{jh} = 1\theta_{kh}\} = z_i^T \Theta z_j$$

em termos intuitivos, a probabilidade de um elemento presente em dois blocos distintos terem relação é o produto das indicadores sobre a matriz Θ. Analogamente a probabilidade de não ocorrer interações dado a habitação de elementos em blocos distintos será $1 - \theta_{kh} = 1 - z_i^T \Theta z_j$.
Denota-se por equivalência estocástica a propriedade de que θ_{kh} só depende das indicações de pertencimento z_i e z_j e é definido como a probabilidade do ator i se relacionar e ser relacionado a qualquer outro no grupo ser a mesma probabilidade do ator j [9].

Além disso, tem-se que a forma completa da função de verossimilhança dos dados é dada por [3]:

$$L_K(\tau, \Theta) = P\{Y, Z|\tau, \Theta\} = P(Y|Z, \Theta)P(Z|\tau) = \\ \prod_{i,j}(z_i^T \Theta z_j)^{y_{ij}}(1 - z_i^T \Theta z_j)^{1-y_{ij}} \times \prod_{i=1}^{n} \prod_{k=1}^{K} \tau_k^{z_{ik}}$$

essa verossimilhança é responsável pela modelagem dos clusteres, visto que traz a informação contida nos dados sobre os parâmetros de interesse. A função de verossimilhança observada possui dificuldades na computação e manipulação, impossibilitando , assim, inferências de maximizá-la (como EMV).
Diversos autores propuseram métodos e estimadores para contornar esse problema de clusterização [4][6][7][8]. Para o uso nesse projeto, utilizou-se o pacote 'mixer' do R em conjunto com outros pacotes como 'network' e 'igraph', o método utilizado para clusterização chama-se "variational" e consiste, basicamente, em otimizar o limite inferior de $\log L_K$ [5] para obter um "intervalo" crescente e preciso.

3 Resultados

Como dito, anteriormente, fora aplicado a função mixer() do pacote de mesmo nome a fim de gerar um modelo de blocos estocásticos. Obteve-se assim, como visto na figura 1 e 2, as relações bem estabelecidas e o número ideal de clusters do método em $G = 3$.

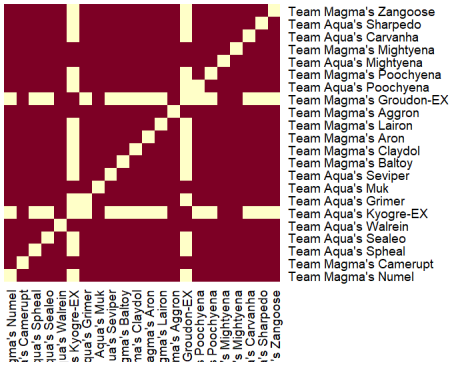

Figure 2: Gráfico de calor dos dados com suas relações estabelecidas pela matriz de adjacências

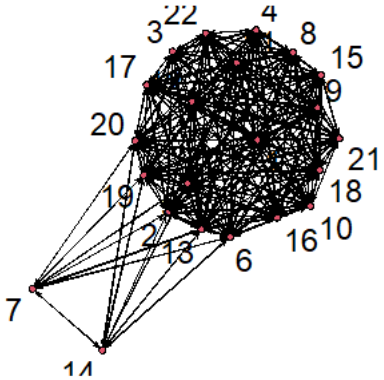

Figure 1: Grafo de redes sobre os dados com base na matriz de adjacências e seus vetores de relação

Após isso, de forma visual, o gráfico de redes com cores indicando os blocos é visto na figura 3. Além disso, obtem-se também os valores estimados de proporção por bloco (isto é, quantidade agregada em cada cluster sob o total) e a matriz Θ citada anteriormente.

$$\tau = \begin{pmatrix} 0,63 \\ 0{,}27 \\ 0{,}10 \end{pmatrix} \qquad \Theta = \begin{pmatrix} 0,999999 & 0,999999 & 0.000001 \\ 0,999999 & 0,999999 & 0,999999 \\ 0.000001 & 0,999999 & 0,999999 \end{pmatrix}$$

Eles indicam de forma clara que o primeiro cluster com o terceiro têm pouquissimas relações, isto é, são altamente distinguíveis, entretanto, as relações do segundo com o terceiro e do segundo com o primeiro são altíssimas. Para além disso, todos têm relações intrísecas altas entre si, evidenciando alta compatibilidade intra-cluster.

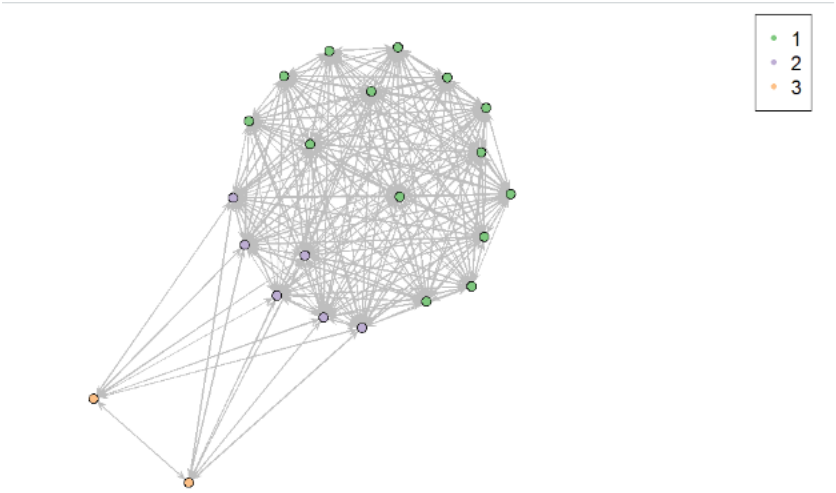

Figure 3: Grafo dos nós com cores indicando os devidos clusters

Para uma análise mais completa, vale da atribuição de etiquetas prévias, sendo elas os tipos de raridade das cartas, definidas em "Common" (cartas comuns, em geral de menor atributos), "Rare Holo" (cartas raras holográficos, tendo um amplo leque de características) e "Rare Ultra" (as cartas ultra raras, com maiores atributos e em pouquissimas quantidades). Para isso, a matriz de adjacência fora transformada em um objeto de classe 'network' (pelo pacote de mesmo nome) para que haja a devida atribuição.

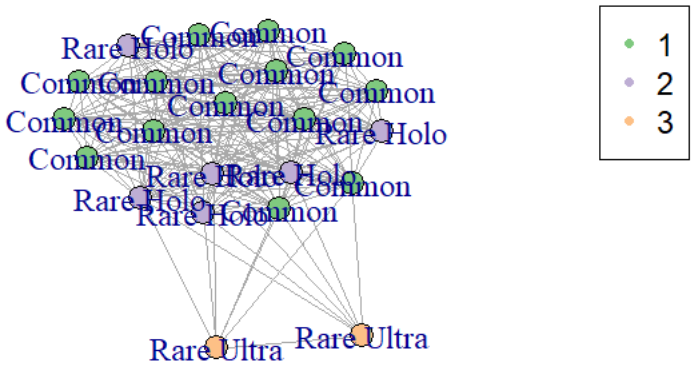

Figure 4: Gráfico de redes com cores indicando etiquetas pré-estabelecidas

	Common	Rare Holo	Rare Ultra
1	12	2	0
2	2	4	0
3	0	0	2

Table 1: Clusteres e suas devidas alocações por etiqueta

A atribuição das etiquetas leva à Tabela 1 em que apresenta a quantidade de atores em cada grupo dado sua etiqueta, evidenciando que, o primeiro grupo concentra as cartas comuns, o terceiro as ultra raras e as raras holográficas, mesmo concentradas no segundo grupo, ainda ocupam o primeiro de forma significativa.
Logo após isso, fora realizado o gráfico de redes por atributos (como segue na Figura 4), demonstra-se, então, a altíssima convergência entre os clusteres da Figura 3 e de seus locais na Figura 4, indicando que a clusterização serviu de forma efetiva como classificador (principalmente de cartas Ultra Raras e Comuns).

4 Conclusão

Em suma, a clusterização por modelo de blocos estocásticos mostrou-se eficaz no contexto de dissimilaridade entre dados. Toda a análise prévia indicou que, mesmo sem uma supervisão, o modelo conseguiu de forma efetiva classificar os dados tal que quando feito a adição de etiqueta sobre os dados, tal classificação se mostrou clara.
Além disso, o número de clusters e os parâmetros de relação obtidos pelos pacotes e método utilizados aparenta ser o ideal,

visto que, como dito, houvera a convergência de um panorama supervisionado.
O modelo utilizado, entretanto, não se limita a uma aplicação simples, visto que diferentes autores, como citado, trabalham em algoritmos de otimização dos estimadores ($\hat{\tau}$ e $\hat{\Theta}$), sendo assim, é um problema ainda atual e que abre a possibilidades para algoritmos de classificação e abordagens numéricas vastas.

Referências

[1] Pokémon TCG API :https://pokemontcg.io/
[2] Análise de dados com R : https://bookdown.org/ricardo$_{l}ehtonen/anlise_{d}e_{d}ados_{c}om_{r}/network.htmlpacote - igraph$
[3]$CharlesBouveyron_{G}illesCeleux_{T}.BrendanMurphy_{A}drianE.Raftery-Model-BasedClusteringandClassificationforDataSciencew$
[4]$K.Nowicki, T.A.BSnijders-EstimationandPredictionforStochasticBlockmodelsforGraphswithLatentBlockStructure$[1997]
[5]$J.JDaudin, F.Picard, S.Robin - Amixturemodelforrandomgraphs$[2008]
[6]$S.Gaucher, O.Klopp - Optimalityofvariationalinferenceforstochasticblockmodelwithmissinglinks$[2021]
[7]$H.Zanghi, C.Ambroise, V.Miele - Fastonlinegraphclusteringvia Erdős–Rényimixture$[2008]
[8]$P.Latouche, E.Birmelé, C.Ambroise-VariationalBayesianinferenceandcomplexitycontrolforstochasticblockmodels$[2012]
[9]$C.JAnderson, S.Wasserman, K.Faust - BuildingStochasticBlockmodels$[1992]