Insper | Instituto de Ensino e Pesquisa

Ciência dos Dados

Relatório - Projeto 3

Turma B

Alexandre Young

Augusto Rissi

Gustavo Molina Freneda Benites

Repositório

https://github.com/gubenites/Ciencia-dos-Dados/tree/master/Projeto%203

Introdução

O projeto é um ensaio sobre a aplicação e funcionamento de algoritmos de clustering; este é um agrupamento de pontos de data considerados similares por uma determinada métrica em sets chamados de clusters.

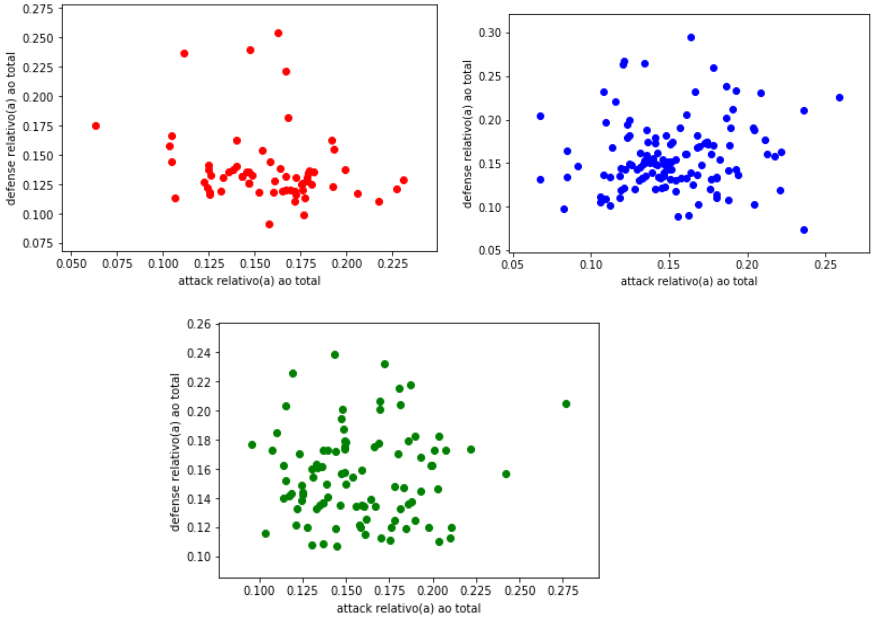
Para realizar o projeto decidimos escolher um dataset que possuísse ao mesmo tempo um leque de informações individuais tanto quanto uma grande quantidade de entradas. O dataset escolhido para o estudo foi do tema “Pokémon” pois este possuía grande variedade de informação numérica em grande quantidade (acima de 800 entradas de pokémons, cada um com vários atributos numéricos).

Desenvolvimento

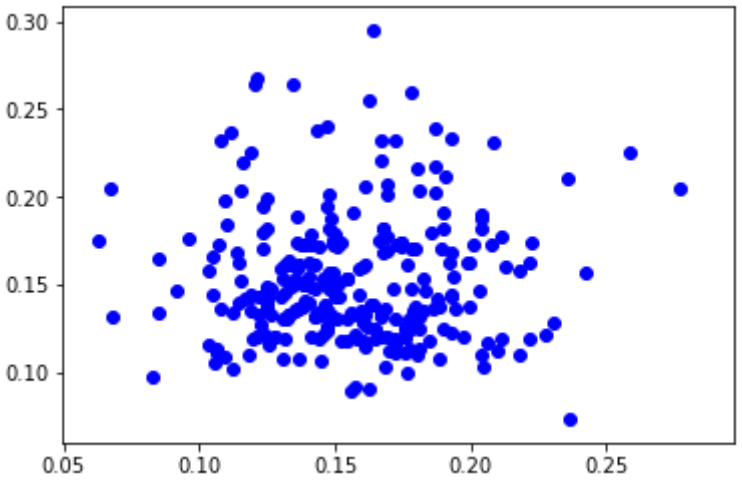
Demos início ao trabalho abrindo o dataset para avaliar as variáveis existentes. A partir dessa análise preliminar decidimos utilizar as grandezas ataque, defesa, ataque especial, defesa especial, velocidade e pontos de vida.

Para a utilização do material escolhido, tivemos que considerar como os atributos se aplicam na prática, sendo assim pesquisamos a fórmula de cálculo usada pelos jogos para definir os atributos reais de cada indivíduo , de tal forma que foi necessário assumir valores específicos relacionados a valores do jogo (I.V, Lv, etc).

Para a aplicação da clusterização procuramos separar pontos de data em três grupos diferentes separados pelo “tipo” de pokémon que representam: planta, fogo e água. Cada um dos três grupos foram “plotados” em um gráfico bidimensional, de ataque sobre defesa relativo à soma de todos os atributos:



O dataset definido pela união dos três datasets exibidos anteriormente será aquele utilizado para testar os algoritmos de clusterização. Quanto mais próximos forem os clusters separados pelos algoritmos em relação aos grupos iniciais que formaram o dataset, mais apropriado será o algoritmo para a finalidade que definimos. Segue o gráfico da união dos datasets

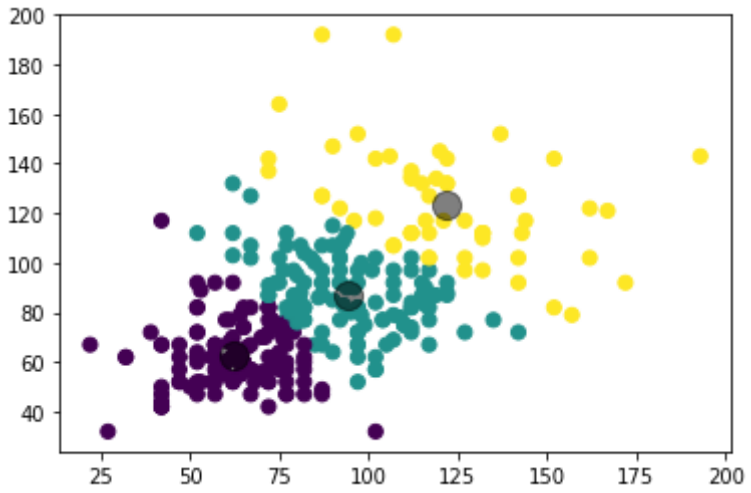


Com o dataset da união dos grupos anteriores aplicamos os algoritmos de clusterização KMeans, DBSCAN, Agglomerative Clustering e Spectral Clustering, com o objetivo de analisar qual desses métodos melhor separa a união dos conjuntos em seus tipos correspondentes.

KMeans

O algoritmo de clusterização KMeans procura separar os pontos de um dataset em um número previamente estabelecido de clusters definidos por um ponto centróide virtual e pontos do dataset de variância similar que pertencem a ele. Como critério de seleção, o algoritmo procura encontrar clusters cuja distância dos membros até os pontos centróides dos clusters seja mínima.

Aplicando o algoritmo no nosso dataset, obtivemos o seguinte resultado:



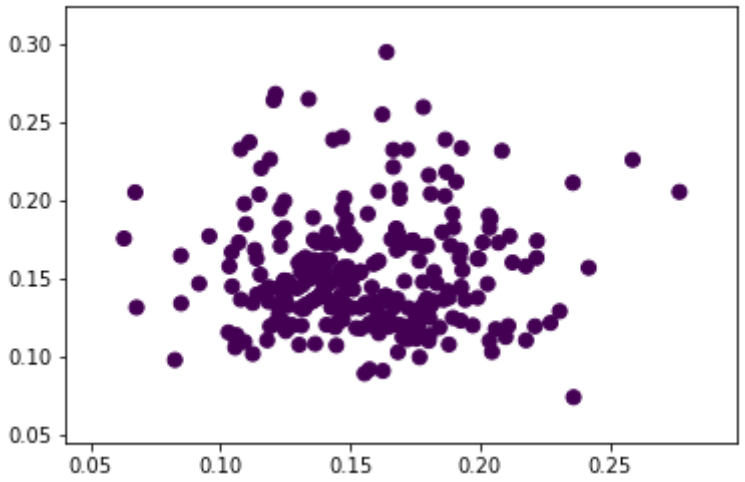
É razoável dizer que a separação realizada pelo algoritmo não foi adequada para separar nosso dataset nos grupos de interesse. Das recomendações de uso do algoritmo temos que seu resultado não descreve bem datasets cujos clusters tenham formas esparsas, irregulares ou alongadas, preferindo formas convexas e isotrópicas bem-definidas

É também prejudicial para a aplicação deste algoritmo que os membros de um mesmo cluster tenha uma grande variância da distância até os pontos centróides. Dos grupos separados vemos que os membros de tipo “água” tem uma variância grande na distribuição dos seus pontos, o que também contribui para a baixa performance do algoritmo.

DBSCAN

Density-Based Spatial Clustering of Applications with Noise, ou DBSCAN, é um algoritmo que separa os pontos de um dataset entre clusters definidos pela densidade dos pontos que o compõe

Aplicando o algoritmo obtivemos o seguinte resultado:



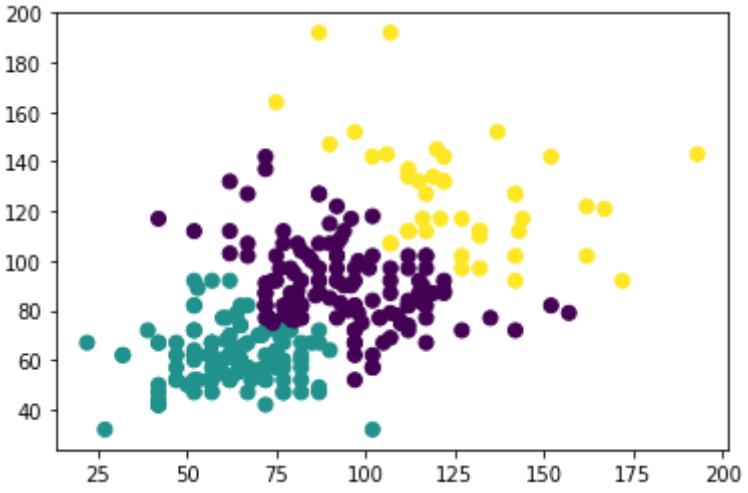
Como parâmetro de aplicação do algoritmo, uma distância máxima entre dois pontos de um mesmo cluster precisa ser provida explicitamente, no entanto não foi possível encontrar um valor adequado para clusterizar o dataset, pois a distribuição presente no nosso dataset é tal que para um valor proposto para essa distância o DBSCAN só será capaz de encontrar um grupo distinto ou nenhum grupo distinto.

Nosso dataset não é adequado para a aplicação deste algoritmo; idealmente para o DBSCAN teríamos áreas distintas e bem espaçadas de alta densidade que seriam visualmente identificáveis como múltiplos aglomerados. Nosso dataset apresenta apenas uma região de grande densidade que vai se espalhando sem definir novas aglomerações.

Agglomerative Clustering

O agglomerative Clustering é um algoritmo que executa um agrupamento hierárquico de baixo para cima de tal forma que cada observação começa em seu próprio cluster e os clusters sao sucessivamente incorporados a partir de 3 possibilidades:

O Agglomerative Clustering é um caso particular do Hierarchical Clustering, esse algoritmo cria clusters a partir de cada dado do dataset, para então formar clusters maiores até atingir um número "**n**" de clusters ("**n**" é um parâmetro do algoritmo). Os clusters sao unidos seguindo um processo de linkagem, a implementação utilizada pelo nosso grupo foi o "ward", e esse critério minimiza a variância da distância dos pontos em relação ao centro do cluster, critério similar ao KMeans.

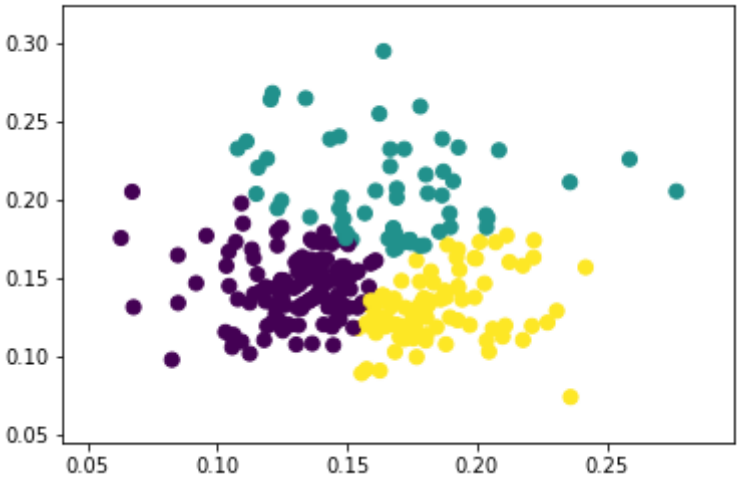


Ao observar o gráfico, pode-se perceber que os clusters resultantes sao próximos aos clusters obtidos pela aplicação do KMeans. Portanto, pelos mesmos motivos do KMeans, esse algoritmo não foi adequado para agrupar os pontos do dataset.

Spectral Clustering

Spectral Clustering pode ser explicado em dois processos sucessivos: primeiramente é construída uma medida de similaridade para o dataset, cujo intuito é definir em uma função como se dão as regras para que um ponto pertença a um determinado cluster (Para a nossa implementação foi utilizada a Radial Basis Function), essa função é então usada em uma redução dimensional com o intuito de tratar e reduzir grandezas no nosso dataset consideradas pouco significativas, tratando um set menor de variáveis, chamadas variáveis principais. Em seguida é aplicado o algoritmo de KMeans sobre o dataset obtido.

Uma vez aplicado o dataset, obtivemos o seguinte resultado:



Ainda que os clusters selecionados não sejam idealmente correspondentes aos grupos de interesse, é possível perceber uma maior similaridade entre a separação dos pontos com este algoritmo e os grupos previstos inicialmente.

Podemos atribuir essa melhor adequação dos resultados ao uso da Radial Basis Function como medida de similaridade.

Conclusão

Com a aplicação dos diferentes algoritmos de clusterização, aprendemos como a escolha de um algoritmo apropriado para as características do nosso dataset é vital para a separação dos sets de interesse. Com o uso do algoritmo de Spectral Clustering obtivemos o melhor resultado das tentativas que foram realizadas, que, apesar de não coincidir perfeitamente com nossos datasets iniciais, serve como um classificador automatizado probabilístico para os tipos de pokémons que quisermos analisar.

Para um melhor resultado da aplicação dos algoritmos de clusterização seria necessário realizar um estudo mais profundo de como os diferentes parâmetros dentro de um mesmo algoritmo usado afetam a clusterização da data. Seria teoricamente possível escrever nossa própria medida de similaridade para ser aplicada no Spectral Clustering, mas isso iria requerer tempo e conhecimentos além do escopo deste projeto.

Referências

https://jakevdp.github.io/PythonDataScienceHandbook/index.html
https://www.kaggle.com/rounakbanik/pokemon
http://scikit-learn.org/
https://bulbapedia.bulbagarden.net/wiki/Statistic