Implementação do algoritmo de clusterização K-means

Murillo Freitas Bouzon

***Resumo*—Para classificar dados onde as classes não são conhecidas, pode-se utilizar uma abordagem de clusterização para agrupá-los em diferentes grupos para facilitar a determinação de qual grupo cada observação pertence. Um método estatístico muito utilizado para esse tipo de problema é o *k-means*, onde calcula-se a distância dos dados para k centroides escolhidos aleatoriamente para definir a qual grupo cada dado pertence de acordo com o centroide mais próximo e atualiza-se os centroide iterativamente até convergirem. Neste trabalho foi implementado o método *k-means* para agrupar e classificar dados de 3 bases de dados diferentes. Os resultados obtidos mostraram que foi possível classificar a base *Iris* com uma acurácia média de 87% e a base *Wine* e *Divorce* com uma acurácia média de 96%.**

***Index Terms*—K-means, Aprendizado não-supervisionado, Machine Learning, Cluster**

I. INTRODUÇÃO

Em problemas de classificação, normalmente utiliza-se uma base de dados para realizar um aprendizado supervisionado, onde a classe que cada dado pertence é conhecida. Com isso, pode ser utilizado diversos métodos para aprender um modelo que melhor separe os dados de acordo com as classes.

No entanto, em problemas de classificação onde a classe dos dados não é conhecida, utiliza-se uma abordagem diferente para poder agrupar os dados semelhantes em diferentes grupos. Esta abordagem também é conhecida como clusterização.

Um algoritmo muito conhecido para resolver problemas de clusterização é o *K-means*, um método para encontrar *clusters* e os seus centroides em um conjunto de dados não-supervisionados de maneira iterativa.

Mesmo sendo um método antigo, o *K-means* ainda é utilizado em pesquisas recentes, como por exemplo no trabalho de [1] que para otimizar a seleção dos centroides iniciais, segmentando o conjunto de dados iniciais em k grupos e selecionando um ponto de cada grupo como centroide. Os experimentos foram realizados na base de dados IRIS e mostraram que o algoritmo proposto possui um tempo de convergência rápido e encontrou um valor de k próximo ao ideal.

Outro trabalho foi o de [2], onde foi proposto um novo algoritmo de clusterização para segmentação de imagem utilizando lógica *fuzzy* e *K-means*. Foram feitas análises qualitativas e quantitativas que mostram que o método conseguiu uma melhor performance na segmentação de diferentes tipos de imagem, resultando em uma qualidade visual melhor de segmentação comparado com outros algoritmos de clusterização.

Sabendo do potencial do algoritmo *K-means*, este trabalho tem como objetivo implementá-lo e verificar o seu funcionamento para clusterizar diferentes bases de dados e obter um melhor entendimento sobre o método.

II. CONCEITOS FUNDAMENTAIS

Nesta seção serão apresentadas as teorias relacionadas ao método desenvolvido neste trabalho

A. K-Means

O termo *K-means* foi apresentado em [3] e o método foi proposto por *Stuart LLoyd* em 1957, porém só acabou sendo publicado em [4].

O algoritmo *K-means* é um método iterativo que recebe como entrada um conjunto de dados e um valor k, representando o número de grupos que os dados serão separados. Para cada cluster, é definido um centroide aleatoriamente e então calcula-se a distância das observações para os centroides para definir a qual grupo cada observação pertence, de forma que a distância seja mínima para o centroide. Após isso, calcula-se um novo centroide a partir da média das amostras pertencentes para cada grupo, atualizando todos os centroides e recalculando as distâncias das observações em relação aos novos centroides. Essa etapa se repete até os centroides convergirem, ou seja, não houver mudança da posição dos centroides.

A Figura 1 mostra um exemplo da aplicação do algoritmo *K-means* em uma amostra de dados 2D.

III. METODOLOGIA

A metodologia deste trabalho foi dividida em 6 etapas.

1) Leitura da base
2) Seleção aleatória dos centroides
3) Cálculo das distâncias das observações para os centroides
4) Cálculo da média das observações para definir os novos centroides
5) Repetir as etapas 3 e 4 até convergir

A implementação foi feita em *Python*, utilizando a biblioteca *numpy* e a biblioteca *matplotlib* para plotagem de gráficos.

IV. BASE DE DADOS

Para validar se a implementação foi feita corretamente, foram realizados experimentos em três bases de dados, sendo descritas a seguir:

A. Iris Data Set

Uma das bases mais conhecidas na literatura utilizada para reconhecimento de padrões, sendo apresentada em [5]. Esta base possui 3 classes com 50 amostras cada classe e 5 características, sendo elas:

- Comprimento da sépala em cm

(a) Primeira iteração do método K-means.

(b) Segunda iteração do método K-means.

(c) Terceira iteração do método K-means.

Figura 1: Exemplo da aplicação do K-means em duas dimensões.

- Largura da sépala em cm
- Comprimento da pétala em cm
- Largura da pétala em cm
- Classes: *Iris Setosa, Iris Veriscolor, Iris Virginica*

B. Wine Data Set

Esta base de dados foi apresentada em [6], possuindo 178 observações e 13 características de análises químicas de 3 classes de vinhos cultivados em uma região da Itália, sendo 58 da classe 1, 71 da classe 2 e 48 da classe 3.

C. Divorce Predictors Dataset

A base *Divorce Predictor Dataset* foi criada em [7] para previsão de divórcios utilizando redes neurais. Esta base possui 54 características, sendo elas perguntas sobre o casal respondidas através de uma escala *likert*, e possui 170 amostras, sendo 86 casais casados e 84 divorciados.

V. EXPERIMENTOS E RESULTADOS

Para avaliar o método implementado, realizou-se dois experimentos. No primeiro experimento, foi utilizado o algoritmo implementado para clusterizar uma base dada em aula com o valor de $k = 3$. A Figura 2 apresenta os resultados do primeiro experimento. O resultado mostra que foi possível agrupar os dados em 3 grupos de uma forma em que eles ficaram bem discriminados visualmente.

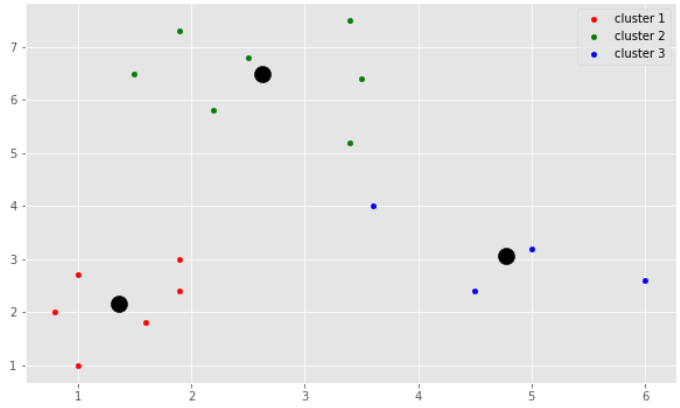

Figura 2: Resultado da aplicação do K-means na base dada em aula.

Para o segundo experimento, aplicou-se o método *K-means* para clusterizar e classificar os dados das 3 bases de dados apresentadas na Seção IV, utilizando um valor de k de acordo com o número de classes da base. A primeira base foi a *Iris Data Set* (Seção IV-A), onde foi utilizado um valore de $k = 3$. A Tabela I apresenta os resultados na base *Iris*.

Iris Dataset	Observada	Estimada	Acurácia
Iris Setosa	50	50	100%
Iris Versicolor	50	61	78%
Iris Virginica	50	39	78%

Tabela I: Resultados na base Iris.

Os resultados mostram que foi possível classificar a classe *Iris setosa* com 100% de acurácia e uma acurácia de 78% para as classes *Iris Versicolor* e *Iris Virginica*. Estes resultados ocorreram provavelmente devido a fácil discriminação entre a *Iris setosa* e as outras flores e a pequena semelhança entre a *Iris Versicolor* e a *Iris Virginica*. Em média, o algoritmo *K-means* conseguiu classificar a base Iris com uma acurácia de 87%.

Os resultados na base de dados *Wine* são apresentados na Tabela II, onde foi utilizado o valor de $k = 3$. Observando os resultados, nota-se que foi possível classificar as 3 classes

com uma taxa de erro pequena, obtendo uma acurácia média de 96%.

Wine Dataset	Observada	Estimada	Acurácia
Classe 1	58	62	93%
Classe 2	71	69	97%
Classe 3	48	47	98%

Tabela II: Resultados na base Wine.

Para a base de dados *Divorce* foi utilizado o valor de $k = 2$ para separar em grupos de casados e de divorciados. Os resultados são apresentados na Tabela III, onde é mostrado que foi possível obter uma acurácia média de 96%.

Divorce Dataset	Observada	Estimada	Acurácia
Casados	81	84	96%
Divorciados	86	90	96%

Tabela III: Resultados na base Divorce.

VI. CONCLUSÃO

Neste trabalho foi implementado o método estatístico *K-means* para agrupar um conjunto de dados em k grupos para poder classificá-los, sendo aplicado em 3 bases de dados diferentes.

Para validar o método foram feitos dois experimentos. No primeiro experimento foi aplicado o *K-means* em um conjunto de dados passado em aula para agrupá-los em 3 grupos. Os resultados mostraram que foi possível agrupar os dados de forma que facilitou a discriminação deles visualmente.

No segundo experimento, aplicou-se o *K-means* em 3 bases de dados diferentes. Para a base da *Iris* utilizou-se o valor de $k = 3$ e foi obtido uma acurácia média na classificação de 87%. Para as bases *Wine* e *Divorce* foi utilizado o valor de $k = 3$ e $k = 2$ respectivamente, obtendo uma acurácia média de 96% para as duas bases.

O algoritmo *K-means* mostrou ser uma boa abordagem para clusterizar dados e melhor visualizar dados não-supervisionados, podendo também ser aplicado em conjunto com o PCA. Também foi mostrado que o *K-means* pode ser utilizado para classificação quando os dados são supervisionados, obtendo uma boa performance na classificação.

REFERÊNCIAS

[1] Yin Cheng-xian. An improved k-means clustering algorithm. 2014.

[2] Siti Noraini Sulaiman and Nor Ashidi Mat Isa. Adaptive fuzzy-k-means clustering algorithm for image segmentation. *IEEE Transactions on Consumer Electronics*, 56, 2010.

[3] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press.

[4] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, March 1982.

[5] R. A. FISHER. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[6] B. Vandeginste. Parvus: An extendable package of programs for data exploration, classification and correlation, m. forina, r. leardi, c. armanino and s. lanteri, elsevier, amsterdam, 1988, price: Us $645 isbn 0-444-43012-1. 1990.

[7] Mustafa Kemal Yöntem, Kemal Adem, Tahsin İlhan, and Serhat Kılıçarslan. Divorce prediction using correlation based feature selection and artificial neural networks. *Nevşehir Hacı Bektaş Veli Üniversitesi SBE Dergisi*, 9:259 – 273, 2019.