Relatório 04

Aprendizagem não supervisionada

O Método k-Means

Cristiano Lopes Moreira

Matrícula: 119103-0

Sumário

1. Introdução

O campo de estudo da engenharia e da ciência da computação denominado aprendizado de máquina é segmentado, de forma macro, em 3 categorias, aprendizado por reforço, supervisionado e não supervisionado. Esse último destina-se à tarefa de encontrar padrões em uma massa de dados, organizá-los, e transformá-los em informações pelo agrupamento de características intrínsecas medidas ou percebidas.

Na técnica de clustering, um conjunto de entrada é segmentado em grupos. Essa difere das técnicas de classificação, como o método da análise discriminante linear (LDA), por não ter a informação prévias de agrupamentos para auxiliar na classificação, o que faz essa técnica tipicamente um aprendizado não supervisionado.

O algoritmo k-Means, proposto por Stuart Lloyd em 1957, publicado como "k-means" por James MacQuem em 1967 (JAMES, 1967), é umas das formas de realizar a técnica de clustering pelo método da menor distância entre os elementos e o centro dos grupos (centroide).

O objetivo deste trabalho é implementar e verificar a eficiência do o algoritmo de clustering K-means utilizando bases de dados conhecidas para verificar a eficiência de agrupamento do algoritmo nas bases de dados propostas.

2. Desenvolvimento teórico

O método de agrupamento (clustering) de k-means é um método de quantização vetorial. Dado um conjunto de n pontos no espaço d-dimensional (padrões de entrada) e um inteiro k, para gerar um conjunto de pontos, chamados de centroides ou protótipos, também no espaço d-dimensional, busca-se minimizar a distância quadrática média de cada ponto ao centroide mais próximo.

Os agrupamentos podem ser inicializados, por exemplo, amostrando aleatoriamente os padrões de entrada. O critério de convergência pode ser certo

número pré-determinado de iterações ou então a ausência de modificações nos grupos.

O algoritmo k-médias funciona bem para problemas em que os grupos são compactos, bem separados e com formato hiper-esférico. Além de ser bastante simples, a complexidade computacional do k-médias é aproximadamente linear. No entanto, deve-se destacar também duas limitações do algoritmo: o número de grupos precisa ser definido previamente e o resultado do algoritmo depende da inicialização dos agrupamentos.

A qualidade da quantização, ao menos em um possível sentido, está diretamente associada ao objetivo do k-médias: minimizar a distância quadrática média de cada dado ao protótipo mais próximo.

2.1. Método k-means

1 Etapa escolha “k” pontos aleatórios como centros de custer/centroides.

2 Etapa: atribuir para cada ponto o cluster mais próximo calculando sua distância para cada centro (distância euclidiana).

3 Etapa: encontrar novo centro de cluster, tendo a média dos pontos atribuídos.

4 Etapa: repetir os passos 2 e 3 até que nenhum dos pontos do cluster mude.

2.1.1. Etapa 2

Distância Euclidiana, ou simplesmente distancia entre dois pontos obtida pelo teorema de Pitágoras

$$d = \sqrt[2]{\sum_{i=1}^{n}(p_i - q_i)^2} \qquad (1)$$

em que n é o número total de dimensões no conjunto de dados X_i, para i=1, ...n; e p e q são os pontos que se deseja calcular a distância .

3. Proposta de implementação Algoritmo k-means

N: número de padrões de entrada

k: número de agrupamentos

C: conjunto de padrões representado por cada protótipo

Pseudocódigo:

Inicialize os k agrupamentos;

Enquanto critério de convergência não for atingido do:

Para i = 1 até N faça:

Identifique o agrupamento mais próximo do padrão de entrada i;

Atualize C;

Finaliza Para

Reposicione cada agrupamento no centroide do subconjunto de padrões associados a ele;

Finaliza enquanto

Para verifica a eficiência do algoritmo pelas diversas dimensões será utilizada a ordenação e redução de dimensões dos dados de entrada pelo método da análise dos componentes principais (PCA).

4. Experimentação e Resultados

Para verificar o funcionamento do algoritmo de k-means, foi realizada a implementação em Python confrontando os resultados entre a classificação indicada na base de dados e o agrupamento proposto pelo algoritmo:

Ambiente:

PyCharm 2019.2.2 (Professional Edition) Build#PY-192.6603.34

Python 3.7.5 (tags/v3.7.5:5c02a39a0b, Oct 15 2019, 01:31:54) on win32

Bibliotecas:

matplotlib-3.1.1	(utilizado para plotagem de gráficos)
pandas-0.25.2	(suporte à plotagem de gráficos)
xlrd-1.2.0	(leitura de arquivos do Excel - base de dados)
numpy-1.17.4	(gestão de matrizes)

Base de Dados:

dbBase.xlsx	base	(Base exercício Aula)
	íris	http://archive.ics.uci.edu/ml/datasets/Iris
	wines	http://archive.ics.uci.edu/ml/datasets/Wine
	Walking and Talking with Someone	https://archive.ics.uci.edu/ml/datasets/Activity+Recognition+from+Single+Chest-Mounted+Accelerometer

4.1. Dados Exercício Aula

Base de dados:

		Classificação	
X	Y	Original	K-means
1.9	7.3	0	0
3.4	7.5	0	0
2.5	6.8	0	0
1.5	6.5	0	0
3.5	6.4	0	0
2.2	5.8	0	0
3.4	5.2	0	0
3.6	4	1	1
5	3.2	1	1
4.5	2.4	1	1
6	2.6	1	1
1.9	3	2	2
1	2.7	2	2
1.9	2.4	2	2
0.8	2	2	2
1.6	1.8	2	2
1	1	2	2

Resultados:

	Classificação		
Grupo	Qt Original	Qt-K-means	Precisão
0	7	7	100%
1	4	4	100%
2	6	6	100%
		Precisão Total	100%

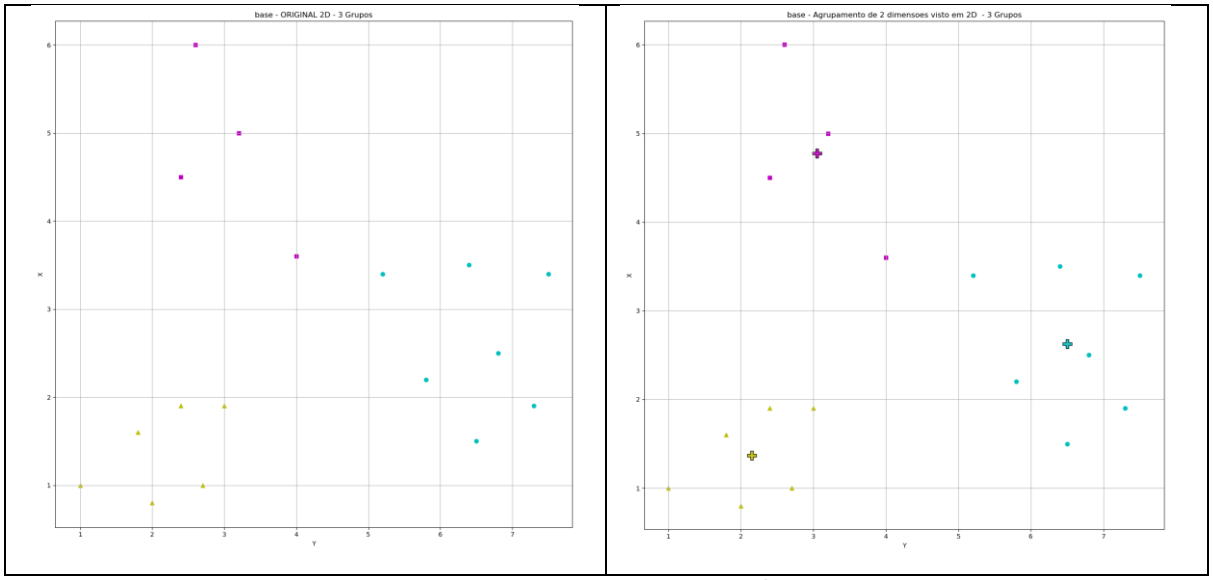

Fig1: Distribuição em 2D dos dados do Exercício em Aula

Pela Figura 1, gráfico de agrupamento de 2 dimensões 3 Grupos, pode-se observar os centroides, pela marcação '+', bem ao centro de cada grupo e equidistantes entre si.

O agrupamento para os dados do exercício 1 da aula pelo algoritmo de aprendizado não supervisionado k-means se mostrou extremamente assertivo.

Em comparação com a classificação esperada todos os grupos foram identificados corretamente pelo algoritmo. Observa-se que o algoritmo é eficiente para classificação de dados de grupos sem ruídos e esféricos.

4.2. Iris Fisher

Base de dados:

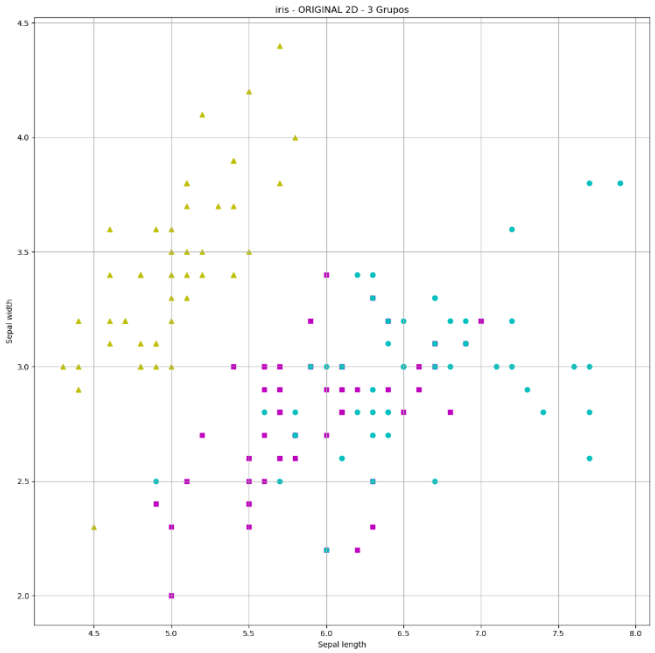

Fig2: Distribuição Original 3 dimensões visão em 2D da base de dados Iris-Fisher

O conjunto de dados contém 3 classes de 50 instâncias cada, em que cada classe se refere a um tipo de planta de íris, mostrados na Figura 2.

Uma classe é linearmente separável da outras duas 2 e as demais não são linearmente separáveis.

Atributo previsto: classe da planta íris.

Informações da base:

- comprimento da sépala em cm
- largura da sépala em cm
- comprimento da pétala em cm
- largura da pétala em cm
- 3 classes: (Setosa, Versicolour e Virginica)

Resultados:

	Classificação		
Grupo	Qt Original	Qt-K-means 4D	Precisão
0	50	50	100%
1	50	62	96%
2	50	38	72%
		Precisão Total	89.33%

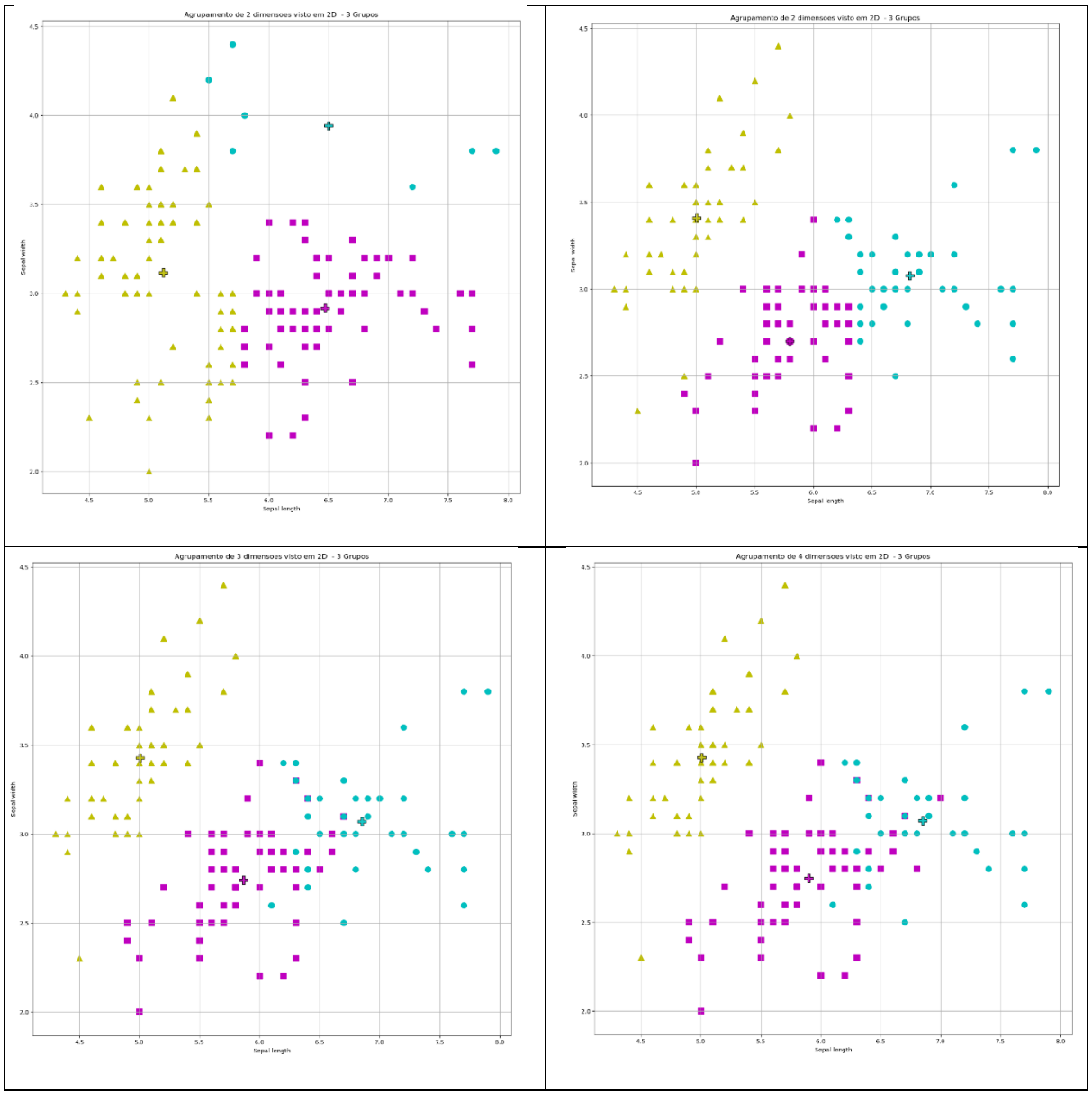

Fig3: Classificação k-means da base de dados Iris-Fisher – diversas dimensões – vista 2D

Pela figura 3, nos gráficos da primeira linha, agrupamento de 2 dimensões 3 Grupos, pode-se observar, pela diferença entre o gráfico da direita (com erro de classificação) e o da esquerda (mais próximo à classificação original) que, dependendo da posição inicial dos centroides propostos pelo algoritmo de k-means, ocorrem erros na classificação, exemplificado no centroide marcado em azul próximo à posição [6.5, 4.0]. Para mitigar esses erros é recomendado realizar diversas interações do algoritmo e obter a médias delas.

O algoritmo k-means apresentou 10.67% de erro em comparação com a classificação real informada pela base de dados, sendo o erro maior entre os grupos 1 e 2 que não são linearmente separáveis.

Observa-se que o algoritmo é eficiente para classificação dos dados que estão linearmente separados, observa-se também que as variáveis principais, demostradas nos gráficos de 2 dimensões, reduzidas pelo método PCA, possibilitam uma classificação semelhante quando comparada com as verificações em 3 e 4 dimensões, segunda linha da figura 3.

4.3. Wines

Base de dados:

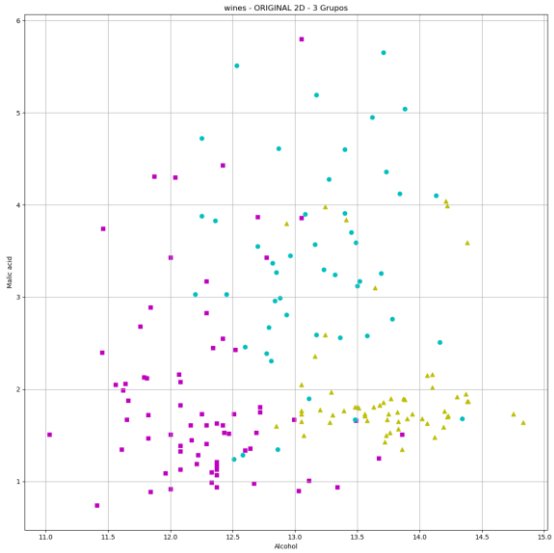

Fig4: Distribuição Original 3 dimensões visão em 2D da base de dados wines

Base de dados, figura 4, com os resultados de uma análise química de vinhos cultivados na mesma região da Itália, mas derivados de três variedades de uvas diferentes. A análise determinou as quantidades de 13 constituintes encontrados em cada um dos três tipos de vinhos.

Informações da base:

- Álcool
- ácido málico
- partículas
- Alcalinidade das cinzas
- Magnésio
- Fenóis totais
- Flavonóides
- Fenóis não flavonóides
- Proantocianinas
- Intensidade da cor
- Hue
- OD280 / OD315 de vinhos diluídos
- Prolina

Resultados:

	Classificação		
Grupo	Qt Original	Qt-K-means 3D	Precisão
0	59	64	86.44%
1	71	64	78.87%
2	48	50	68.75%
		Precisão Total	78.70%

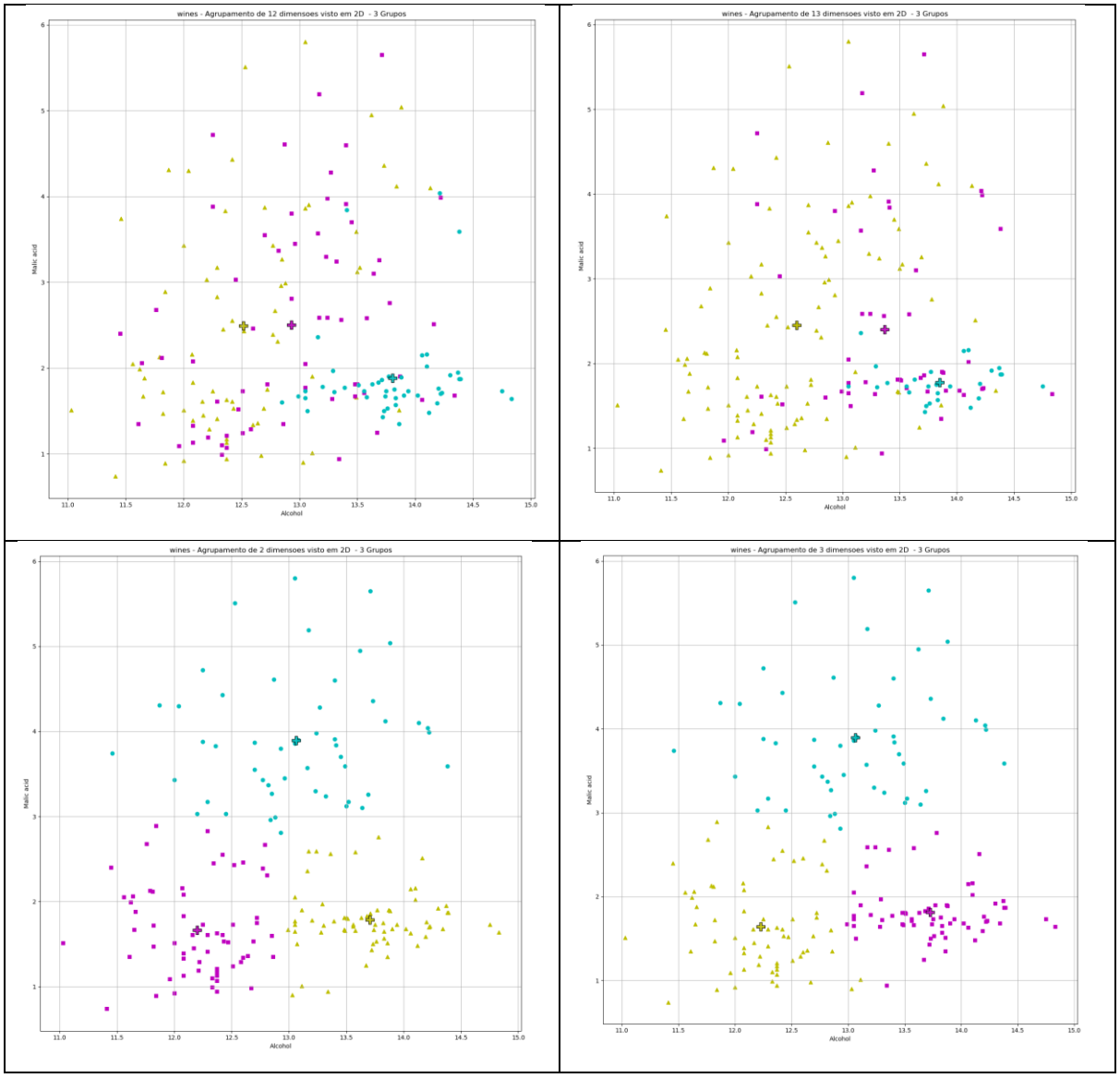

Fig5: Classificação k-means da base de dados wines – vista 2 e 3 dimensões

Pela figura 5, nos gráficos de agrupamento de 12 e 13 dimensões com 3 Grupos, pode-se observar que o algoritmo k-means não é eficiente quando existe muito ruído, tendo 46.6% de erro.

Pela redução de ruído utilizando a análise de componentes principais (PCA), para 2 ou 3 dimensões principais é possível aumentar a precisão do algoritmo de k-means obtendo um erro de 21.30% em comparação com a classificação real informada pela base de dados, sendo o erro maior entre os grupos 1 e 2 onde ocorre maior sobreposição dos dados.

4.4. Walking and Talking with Someone

Base de dados:

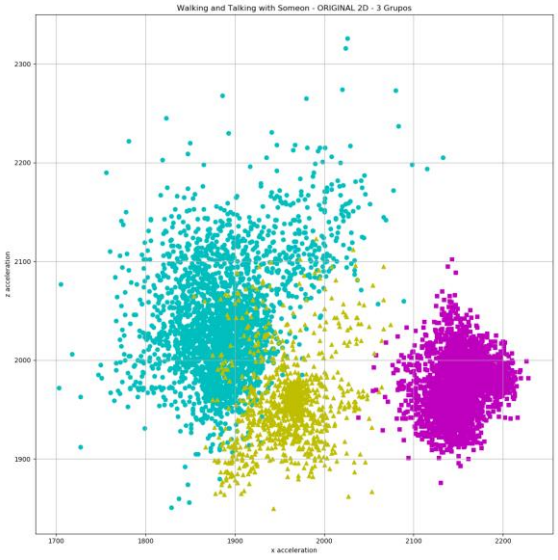

Fig6: Distribuição Original 3 dimensões visão em 2D da base de dados Activity Recognition from Single Chest-Mounted Accelerometer - 3 pessoas relação waking and talking with someone

Dados de acelerômetro não calibrado coletados de 3 participantes realizando atividade "Caminhando e conversando com alguém"

O conjunto de dados oferece desafios para identificação e autenticação de pessoas usando padrões de movimento.

Informações do conjunto de dados:

- Dados de um acelerômetro vestível montado no peito
- Frequência de amostragem do acelerômetro: 52 Hz
- Os dados do acelerômetro não são calibrados
- Número de participantes: 3 (1, 2 e 3 da base de dados original)
- Número de atividades: 1 (atividade 6: Caminhando e conversando com alguém)

Resultados:

	Classificação		
Grupo	Qt Original	Qt-K-means 2D	Precisão
0	2917	1971	65.23%
1	7100	7134	99.99%
2	1400	2312	93.28%
		Precisão Total	78.70%

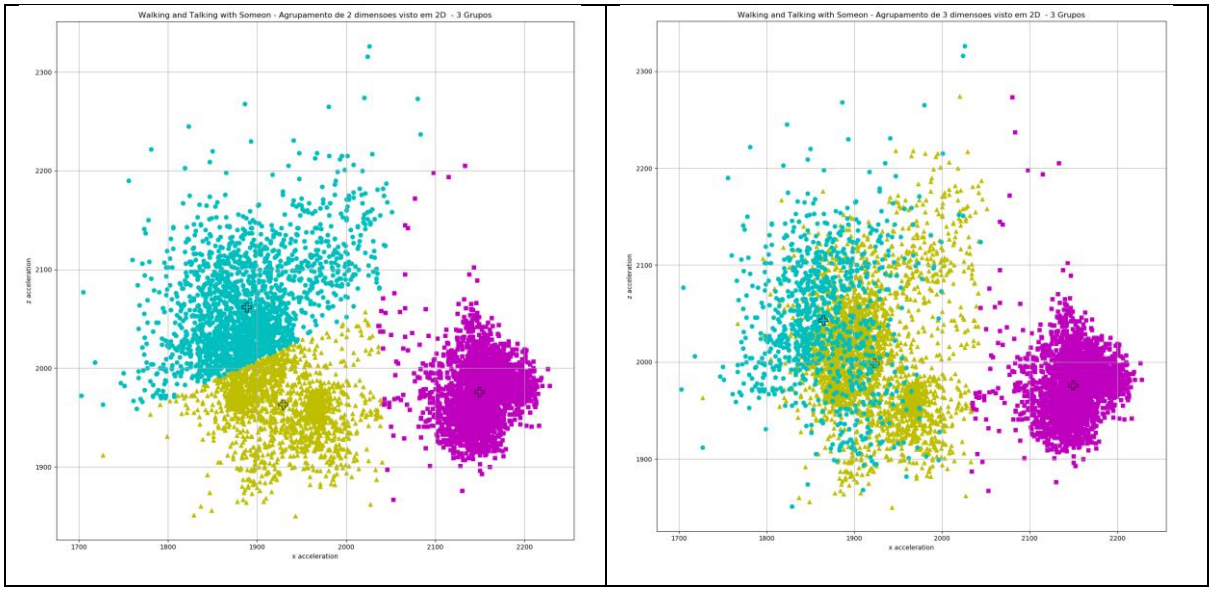

Fig7: Classificação k-means da base de dados Activity Recognition from Single Chest-Mounted Accelerometer

Pela figura 7, nos gráficos de agrupamento de 2 e 3 dimensões com 3 Grupos, pode-se observar que o algoritmo k-means não é eficiente quando existe muito ruído, com sobreposição de dados entre os grupos, tendo 34.77% de erro na classe 0 que muita sobreposição com a classe 1. Para os dados da Classe 2, linearmente separada das demais classes, observa-se uma grande eficiência na classificação.

Pela redução de ruído utilizando a análise de componentes principais (PCA), para 2 dimensões principais é possível aumentar a precisão do algoritmo de k-means reduzindo o erro geral de 19.27% para 9.71% em comparação com a classificação real informada pela base de dados, sendo o maior erro entre os grupos 0 e 1 onde ocorre maior sobreposição dos dados.

5. Conclusão

O método de agrupamento pela técnica de k-means mostra-se eficiente quando a massa de dados é compacta, com grupos esféricos, e baixo ruído, sendo possível utilizar da técnica PCA para reduzir os ruídos de uma massa de dados e aumentar a eficiência do algoritmo k-means.

Pelas interações com as diversas bases de dados foi possível observar que a em alguns casos a disposição inicial dos centroides geram distorções nos agrupamentos, sendo necessário e recomendável realizar várias interações e optar por aquelas de maior reincidência de agrupamentos/resultados (moda).

Conclui-se que agrupamento não supervisionado por clustering possibilita acertos na análise de dados, mas também pode gerar incoerências dependendo dos métodos e visões que se deseja buscar sobre uma massa de dados. As comunidades de aprendizado de máquina e reconhecimento de padrões precisam abordar uma série de questões para melhorar o entendimento do clustering de dados. “É importante lembrar que a análise de cluster é uma ferramenta exploratória; a saída dos algoritmos de agrupamento sugere apenas hipóteses.” (JAIN, 2010)

6. Referências

[1] JAMES. MacQueen. S**ome Methods for Classification and Analysis of Multivariate Observations**. In Proc. Fifth Berkeley Symp. Math. Statistics and Probability, pp. 281–296, 1967.

[2] JAIN, Anil K.. **Data clustering: 50 years beyond K-means**. Pattern Recognition Letters. Michigan, p. 651-666. 1 jun. 2010. Disponível em: <https://doi.org/10.1016/j.patrec.2009.09.011>. Acesso em: 24 nov. 2019.

[3] Dua, D. and Graff, C. (2019). **UCI Machine Learning Repository** [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.