Implementação e experimentação de algoritmo 2-aproximado para o problema dos k-centros

Alexis D G Mariz

Departamento de Ciência da Computação – Universidade Federal de Minas Gerais(UFMG) – Belo Horizonte – MG – Brazil

adgmariz@ufmg.br

Abstract. *This paper presents the implementation and evaluation of a 2-approximate algorithm for the k-centers problem. Experiments were conducted on 10 datasets from the UCI Machine Learning Repository, varying the 'p' parameter of the Minkowski distance function. The effectiveness of the algorithms was assessed based on the quality of the solutions and computational time in comparison with the K-Means algorithm. Although the 2-approximate algorithm showed a higher computational time, it provided solutions with an acceptable radius. However, in general, its silhouette and ARI values were significantly worse than those of K-Means.*

Resumo. *Este artigo apresenta a implementação e avaliação de um algoritmo 2-aproximado para o problema dos k-centros. Os experimentos foram realizados em 10 conjuntos de dados da UCI Machine Learning Repository, variando o parâmetro 'p' da função de distância de Minkowski. A eficácia do algoritmo foi avaliada com base na qualidade das soluções e no tempo computacional comparando com o algoritmo K-Means. Além do algoritmo 2-aproximado ter apresentado um tempo computacional maior, ele forneceu soluções com raio aceitável. No entanto, em geral, seus valores de silhueta e IRA foram significativamente piores que o K-Means.*

1. Introdução

O objetivo deste artigo é implementar e avaliar um algoritmo 2-aproximado para o problema dos k-centros, um problema NP-difícil notável na área de aprendizado de máquina e agrupamento de dados. Esse problema consiste na divisão de um conjunto de dados em ‘k’ grupos, de modo que a distância máxima de qualquer ponto de dados para o centro ao qual pertence seja minimizada.

Os experimentos foram realizados em 10 conjuntos de dados obtidos da UCI Machine Learning Repository, um repositório de conjuntos de dados para pesquisa em aprendizado de máquina. Para cada conjunto de dados, realizamos uma série de experimentos variando o parâmetro 'p' da função de distância de Minkowski, com 'p' assumindo os valores 1, 2 e 3. Para cada valor de 'p', realizamos 30 experimentos usando o algoritmo 2-aproximado e um experimento usando o K-Means, totalizando 93 execuções para cada conjunto de dados.

Neste trabalho, foi comparada a eficácia do algoritmo 2-aproximado com o algoritmo clássico de agrupamento, o K-Means. Além da qualidade das soluções de cada algoritmo, também foi comparado o tempo computacional gasto por cada algoritmo. Esta comparação não se limita apenas à qualidade das soluções de cada algoritmo, mas também ao tempo computacional necessário para produzir essas

soluções. Para avaliar a qualidade das soluções, utilizamos três métricas: o raio da solução, a silhueta e o índice de Rand ajustado.

2. Métodos e métricas:

Neste trabalho, foram utilizados dois métodos para resolver o problema dos k-centros: o algoritmo 2-aproximado e o algoritmo K-Means. O algoritmo 2-aproximado é uma solução heurística para o problema dos k-centros que garante uma solução no máximo duas vezes pior que a solução ótima. Este algoritmo seleciona o primeiro centro aleatoriamente entre os pontos e seleciona o próximo centro como o ponto que está mais distante do centro mais próximo já selecionado. Este processo é repetido até que 'k' centros tenham sido selecionados. Já o algoritmo K-Means é uma solução clássica para o problema dos k-centros. Primeiro, ele inicializa 'k' centróides selecionados aleatoriamente a partir dos pontos de dados. Em seguida, atribui cada ponto de dados ao centróide mais próximo, formando 'k' centros. Após a atribuição dos pontos, o algoritmo recalcula os centróides como a média de todos os pontos de dados em cada centro. Esse processo é repetido até que os centróides não mudem significativamente ou até que um número máximo de iterações ocorra.

Para calcular as distâncias entre os pontos, foi utilizada a função de distância de Minkowski. Essa métrica possui um parâmetro ‘p’, onde ‘p = 1’ representa a distância Manhattan e ‘p = 2’ a distância de euclidiana. Nesse trabalho foram utilizados 1,2 e 3 como valores de ‘p’.

As métricas de qualidade das soluções utilizadas foram: raio da solução, silhueta e índice de Rand ajustado(IRA). O raio da solução é a maior distância de um ponto de dados para o centro ao qual pertence. Esta métrica fornece uma medida da dispersão dos centros. A silhueta é uma medida de quão semelhante um posto está com seu próprio centro em comparação com outros centros. Os valores da silhueta variam de -1 a +1, onde um valor alto indica que o ponto está bem combinado com seu próprio centro e mal combinado com os centros vizinhos. O índice de Rand ajustado(IRA) é uma medida da similaridade entre duas atribuições de dados, neste caso, a atribuição de dados aos clusters produzidos pelo algoritmo e a atribuição verdadeira de dados aos clusters. O índice de Rand ajustado leva em conta a aleatoriedade e tem um valor máximo de 1 quando as duas atribuições de dados são idênticas.

3. Implementação

A implementação deste trabalho foi dividida em duas partes principais: a implementação do algoritmo 2-aproximado e a realização dos experimentos. O arquivo ‘k_centers.py contém a implementação do algoritmo 2-aproximado. Ele inclui funções para calcular a distância de Minkowski entre dois pontos, calcular a matriz de distância para um conjunto de dados, encontrar o ponto mais distante de um conjunto de centros, calcular os k-centros, calcular o raio dos k-centros e atribuir classes aos pontos de dados com base nos centros que eles pertencem. Para cada base de dados, foi criado um arquivo ‘[base-de-dados].py’ para realizar os experimentos. Isso foi feito para possibilitar a execução paralela dos experimentos das diferentes bases de dados. Foi

utilizado o algoritmo K-Means da biblioteca Scikit-learn. A função ‘calc_radius_kmeans’ é usada para calcular o raio da solução para o algoritmo K-Means. Por fim, os resultados são registrados em um arquivo CSV. Todos os arquivos estão disponibilizados em https://github.com/Adgmariz/k-centros.

4. Descrição dos experimentos e bases de dados

Os experimentos foram realizados com 10 bases de dados, onde para cada conjunto de dados, o algoritmo 2-aproximado foi executado 30 vezes para cada valor de 'p' (1, 2 e 3), onde 'p' é o parâmetro da função de distância de Minkowski. O algoritmo K-Means foi executado uma vez para cada conjunto de dados. As bases de dados utilizadas foram:
Experimento 1: Optical Recognition of Handwritten Digits: Foram usados os dados do arquivo 'optdigits.tra', que contém 3823 instâncias. O valor de ‘k’ utilizado foi 10.

Experimento 2: Blood Transfusion Service Center: Foram usados os dados do arquivo 'transfusion.data', que contém 748 instâncias. O valor de ‘k’ utilizado foi 2.

Experimento 3: Banknote Authentication: Foram usados os dados do arquivo 'data_banknote_authentication.txt', que contém 1372 instâncias. O valor de ‘k’ utilizado foi 2.

Experimento 4: Image Segmentation: Foram usados os dados do arquivo 'segmentation.test', que contém 2100 instâncias. O valor de ‘k’ utilizado foi 7.

Experimento 5: Red Wine Quality: Foram usados os dados do arquivo 'winequality-red.csv', que contém 1599 instâncias. O valor de ‘k’ utilizado foi 11.

Experimento 6: White Wine Quality: Foram usados os dados do arquivo 'winequality-white.csv', que contém 4898 instâncias. O valor de ‘k’ utilizado foi 11.

Experimento 7: Yeast: Foram usados os dados do arquivo 'yeast.data', que contém 1484 instâncias. O valor de ‘k’ utilizado foi 10.

Experimento 8: Diabetic Retinopathy Debrecen: Foram usados os dados do arquivo 'messidor_features.arff', que contém 1151 instâncias. O valor de ‘k’ utilizado foi 2.

Experimento 9: Mammographic Mass: Foram usados os dados do arquivo 'mammographic_masses.data', que contém 961 instâncias. Foram excluídas 162 instâncias com valores faltantes. O valor de ‘k’ utilizado foi 2.

Experimento 10: Website Phishing: Foram usados os dados do arquivo 'PhishingData.arff', que contém 1353 instâncias. O valor de ‘k’ utilizado foi 3.

5. Apresentação e análise dos resultados

A seguir os dados dos 10 experimentos. Para cada experimento, foram realizadas 1 execução do K-Means e 30 execuções do k-centros para cada p={1,2,3}. Note que o K-Means só foi realizado para ‘p=1’, pois sua implementação não aceita diferentes valores de ‘p’. Devido à extensão e à complexidade dos dados resultantes dos experimentos, as tabelas completas com os resultados não foram incluídas diretamente neste artigo para manter a concisão e a legibilidade. No entanto, todos os dados brutos e resultados completos dos experimentos estão disponíveis em:

https://docs.google.com/spreadsheets/d/1cXAN_IAmHarbQf5D1Zl-y-D1Ku1futGt_jfjww6T0uY/edit?usp=sharing

	P = 1					P = 2					P = 3			
	Raio	Silhueta	IRA	Tempo		Raio	Silhueta	IRA	Tempo		Raio	Silhueta	IRA	Tempo
K-Means	245.40	0.22	0.68	19.80										
k-centros	272.00	0.06	0.15	151.35	k-centros	52.97	0.08	0.26	84.28	k-centros	32.02	0.08	0.32	92.59
k-centros	280.00	0.09	0.26	142.22	k-centros	53.27	0.07	0.23	84.74	k-centros	32.64	0.05	0.20	97.96
k-centros	274.00	0.10	0.29	130.52	k-centros	53.81	0.09	0.28	84.83	k-centros	33.23	0.08	0.31	97.30
k-centros	276.00	0.07	0.22	125.10	k-centros	53.67	0.09	0.30	84.81	k-centros	33.79	0.03	0.16	97.82
k-centros	281.00	0.10	0.27	124.75	k-centros	54.10	0.05	0.14	84.57	k-centros	32.81	0.08	0.30	96.89
k-centros	278.00	0.06	0.22	124.12	k-centros	53.24	0.08	0.26	84.83	k-centros	32.97	0.07	0.24	118.87
k-centros	280.00	0.04	0.15	122.41	k-centros	53.49	0.08	0.24	84.72	k-centros	33.27	0.06	0.19	124.85
k-centros	283.00	0.06	0.21	122.27	k-centros	54.31	0.07	0.21	84.64	k-centros	32.99	0.06	0.16	125.14
k-centros	269.00	0.04	0.17	124.48	k-centros	53.07	0.07	0.16	84.65	k-centros	32.90	0.04	0.13	113.04
k-centros	270.00	0.11	0.33	122.60	k-centros	53.80	0.08	0.28	85.04	k-centros	34.16	0.04	0.23	113.87
k-centros	277.00	0.07	0.20	121.75	k-centros	53.81	0.07	0.24	84.69	k-centros	32.44	0.08	0.27	113.51
k-centros	272.00	0.03	0.11	112.64	k-centros	51.94	0.07	0.20	84.57	k-centros	32.77	0.07	0.24	114.09
k-centros	287.00	0.05	0.16	104.58	k-centros	52.27	0.10	0.24	84.35	k-centros	32.74	0.08	0.20	113.03
k-centros	279.00	0.07	0.17	105.18	k-centros	52.94	0.06	0.24	84.69	k-centros	33.59	0.06	0.23	112.36
k-centros	281.00	0.05	0.17	104.38	k-centros	51.82	0.09	0.32	84.79	k-centros	33.13	0.10	0.39	111.06
k-centros	276.00	0.10	0.29	104.40	k-centros	54.06	0.08	0.29	84.41	k-centros	32.89	0.09	0.25	111.69
k-centros	276.00	0.08	0.20	104.63	k-centros	53.03	0.07	0.20	84.30	k-centros	32.75	0.08	0.28	111.28
k-centros	278.00	0.08	0.24	104.64	k-centros	52.66	0.07	0.18	84.62	k-centros	33.10	0.05	0.19	111.15
k-centros	275.00	0.09	0.22	107.60	k-centros	53.80	0.06	0.16	84.36	k-centros	33.47	0.07	0.23	111.21
k-centros	281.00	0.06	0.20	104.36	k-centros	54.31	0.05	0.17	85.35	k-centros	32.72	0.08	0.27	115.16
k-centros	281.00	0.11	0.30	104.33	k-centros	54.10	0.07	0.22	86.34	k-centros	33.22	0.04	0.13	118.67
k-centros	273.00	0.07	0.18	104.70	k-centros	54.03	0.07	0.25	84.24	k-centros	31.95	0.05	0.20	97.43
k-centros	279.00	0.06	0.19	104.64	k-centros	52.28	0.10	0.31	84.35	k-centros	32.72	0.08	0.25	104.55
k-centros	277.00	0.10	0.28	104.71	k-centros	54.95	0.08	0.22	84.67	k-centros	33.29	0.06	0.24	125.95
k-centros	278.00	0.12	0.36	104.48	k-centros	52.79	0.07	0.19	86.17	k-centros	33.54	0.09	0.26	122.83
k-centros	271.00	0.09	0.21	105.50	k-centros	53.80	0.08	0.26	88.10	k-centros	33.74	0.07	0.29	129.20
k-centros	283.00	0.06	0.17	104.29	k-centros	53.24	0.06	0.23	87.95	k-centros	32.47	0.06	0.28	142.97
k-centros	276.00	0.05	0.14	98.31	k-centros	53.26	0.08	0.24	87.65	k-centros	32.40	0.08	0.26	147.83
k-centros	280.00	0.11	0.30	86.80	k-centros	53.94	0.06	0.16	87.65	k-centros	32.07	0.08	0.23	146.52
k-centros	289.00	0.07	0.20	86.62	k-centros	54.44	0.06	0.22	87.79	k-centros	32.83	0.08	0.21	145.39
Média	276.69	0.08	0.23	109.30	Média	53.44	0.07	0.23	85.27	Média	32.95	0.07	0.24	116.14
Desvio padrão	7.29	0.03	0.10	21.54	Desvio padrão	0.75	0.01	0.05	1.23	Desvio padrão	0.52	0.02	0.06	14.82

Figura 1. Tabela obtida no experimento 1

A tabela acima mostra os resultados obtidos no experimento 1, onde observa-se que o algoritmo K-Means obteve um valor de raio aproximadamente 12% melhor que a média dos resultados obtidos pelo algoritmo 2-aproximado k-centros. O K-Means também foi superior em silhueta e IRA, além de apresentar um tempo de execução 80% menor. Para o algoritmo k-centros, o valor ‘p=2’ obteve os melhores resultados dentre os valores de ‘p’ observados, com silhueta 0.07 e IRA 0.23.

	P = 1					P = 2					P = 3			
	Raio	Silhueta	IRA	Tempo										
K-Means	8627.53	0.70	0.08	0.67		Raio	Silhueta	IRA	Tempo		Raio	Silhueta	IRA	Tempo
Média	5784.44	0.85	0.04	5.37		5266.88	0.85	0.03	5.44		5433.33	0.85	0.03	5.20
Desvio padrão	732.35	0.03	0.01	1.22		694.99	0.01	0.01	0.44		670.61	0.00	0.00	0.02

Figura 2. Tabela reduzida do experimento 2

A tabela acima mostra os resultados obtidos no experimento 2, onde a média e o desvio padrão correspondem aos valores obtidos pelas 30 execuções do algoritmo k-centros. Observa-se que o algoritmo K-Means obteve um valor de raio maior que a

média do algoritmo k-centros, o que não era esperado. Isso pode ser explicado pelo fato do valor ‘k=2’. Com isso, o K-Means executado apenas uma única vez obteve um valor pior, pois o primeiro centróide escolhido é aleatório. Além disso, todas as execuções obtiveram valores de silhueta altos(próximos de 1).

	P = 1					P = 2					P = 3			
	Raio	Silhueta	IRA	Tempo										
K-Means	27.43	0.42	0.05	5.14		Raio	Silhueta	IRA	Tempo		Raio	Silhueta	IRA	Tempo
Média	28.38	0.43	0.07	19.18		15.96	0.45	0.07	16.55		14.32	0.44	0.07	16.52
Desvio padrão	3.21	0.04	0.02	1.93		1.59	0.03	0.02	0.19		1.21	0.04	0.02	0.45

Figura 3. Tabela reduzida do experimento 3

A tabela acima mostra os resultados obtidos no experimento 3. Ambos os algoritmos apresentaram desempenho parecido, exceto no tempo de execução, onde o k-centros gasta 4x o tempo do K-Means.

	P = 1					P = 2					P = 3			
	Raio	Silhueta	IRA	Tempo										
K-Means	806.96	0.31	0.34	9.66		Raio	Silhueta	IRA	Tempo		Raio	Silhueta	IRA	Tempo
Média	1116.19	0.19	0.13	40.17		443.28	0.28	0.12	33.42		414.25	0.30	0.09	32.55
Desvio padrão	107.47	0.07	0.03	4.02		1.65	0.08	0.04	2.40		0.00	0.06	0.03	0.24

Figura 4. Tabela reduzida do experimento 4

A tabela acima mostra os resultados obtidos no experimento 4. Observa-se que o K-Means obteve um valor de raio menor que a média do k-centros. Além disso, o k-centros obteve valores de silhueta melhores para ‘p=2’ e ‘p=3’.

	P = 1					P = 2					P = 3			
	Raio	Silhueta	IRA	Tempo										
K-Means	73.74	0.33	0.00	1.93		Raio	Silhueta	IRA	Tempo		Raio	Silhueta	IRA	Tempo
Média	78.19	0.09	0.02	20.03		59.16	0.07	0.02	18.98		56.89	0.11	0.02	19.35
Desvio padrão	2.85	0.04	0.01	1.06		5.04	0.03	0.01	0.25		8.11	0.07	0.01	0.59

Figura 5. Essa figura mostra a tabela reduzida do experimento 5

A tabela acima mostra os resultados obtidos no experimento 5. Houve pouca diferença entre os raios dos algoritmos. No entanto, o K-Means apresentou silhueta melhor. Não houve diferença significativa nas métricas do k-centro ao mudar o valor de ‘p’. Os valores de IRA próximos de 0 e silhueta pequena podem ser explicados com a natureza da base de dados de classificação de vinhos tintos, onde existem muito mais vinhos medianos do que excelentes ou ruins. A base de dados também não considera atributos como região e tipo de uva, fatores importantes para a qualidade de um vinho , considera apenas atributos físicos, como densidade e acidez.

	P = 1					P = 2					P = 3			
	Raio	Silhueta	IRA	Tempo										
K-Means	274.10	0.27	0.01	18.19		Raio	Silhueta	IRA	Tempo		Raio	Silhueta	IRA	Tempo
Média	144.16	0.08	0.01	176.86		109.37	0.24	0.01	155.79		102.51	0.27	0.01	151.93
Desvio padrão	11.90	0.09	0.00	16.13		7.23	0.06	0.01	5.04		8.49	0.06	0.01	7.44

Figura 6. Essa figura mostra a tabela reduzida do experimento 6

A tabela acima mostra os resultados obtidos no experimento 6. Observa-se que o algoritmo K-Means obteve um valor de raio maior que a média do algoritmo k-centros, o que não era esperado. Isso pode ser explicado pelo fato do valor 'k=2'. Com isso, o K-Means executado apenas uma única vez obteve um valor pior, pois o primeiro centróide escolhido é aleatório. Assim como no experimento 5, os valores de IRA próximos de 0 e silhueta pequena podem ser explicados com a natureza da base de dados de classificação de vinhos brancos, onde existem muito mais vinhos medianos do que excelentes ou ruins. A base de dados também não considera atributos como região e tipo de uva,fatores importantes para a qualidade de um vinho, considera apenas atributos físicos, como densidade e acidez.

	P = 1					P = 2					P = 3			
	Raio	Silhueta	IRA	Tempo										
K-Means	1.76	0.16	0.14	1.84		Raio	Silhueta	IRA	Tempo		Raio	Silhueta	IRA	Tempo
Média	1.34	0.19	0.04	20.93		0.73	0.25	0.04	22.25		0.60	0.23	0.03	28.67
Desvio padrão	0.14	0.07	0.03	1.19		0.02	0.08	0.03	0.97		0.03	0.07	0.04	7.25

Figura 7. Essa figura mostra a tabela reduzida do experimento 7

A tabela acima mostra os resultados obtidos no experimento 7. Observa-se que o algoritmo K-Means obteve um valor de raio maior que a média do algoritmo k-centros, o que não era esperado. Isso pode ser explicado pelo fato do valor 'k=2'. Com isso, o K-Means executado apenas uma única vez obteve um valor pior, pois o primeiro centróide escolhido é aleatório.

	P = 1					P = 2					P = 3			
	Raio	Silhueta	IRA	Tempo										
K-Means	480.39	0.35	0.00	5.72		Raio	Silhueta	IRA	Tempo		Raio	Silhueta	IRA	Tempo
Média	585.72	0.41	0.00	21.02		235.64	0.55	0.00	16.73		192.02	0.60	0.00	19.00
Desvio padrão	65.71	0.05	0.00	5.10		21.30	0.05	0.00	4.72		13.43	0.02	0.00	6.19

Figura 8. Essa figura mostra a tabela reduzida do experimento 8

A tabela acima mostra os resultados obtidos no experimento 8. Observa-se que o K-Means obteve um raio 18% menor que a média do k-centros. Além disso, o IRA obtido em todas as execuções foi 0, o que indica que os agrupamentos são aleatórios.

	P = 1					P = 2					P = 3			
	Raio	Silhueta	IRA	Tempo										
K-Means	57.82	0.50	0.14	4.20		Raio	Silhueta	IRA	Tempo		Raio	Silhueta	IRA	Tempo
Média	54.23	0.65	0.02	11.63		49.59	0.59	0.05	13.27		49.36	0.58	0.04	12.22
Desvio padrão	9.32	0.12	0.04	2.84		5.75	0.11	0.06	2.13		3.65	0.10	0.05	2.88

Figura 9. Essa figura mostra a tabela reduzida do experimento 9

A tabela acima mostra os resultados obtidos no experimento 9. Observa-se que o algoritmo K-Means obteve um valor de raio maior que a média do algoritmo k-centros, o que não era esperado. Isso pode ser explicado pelo fato do valor ‘k=2’. Com isso, o K-Means executado apenas uma única vez obteve um valor pior, pois o primeiro centróide escolhido é aleatório. Entretanto, a silhueta obtida em ambos os algoritmos foram próximas de 0.60, o que indica que os dados foram bem agrupados e cada ponto é bem semelhante ao seu grupo.

	P = 1					P = 2					P = 3			
	Raio	Silhueta	IRA	Tempo										
K-Means	7.34	0.30	0.24	1.57		Raio	Silhueta	IRA	Tempo		Raio	Silhueta	IRA	Tempo
Média	8.63	0.15	0.14	13.97		4.12	0.14	0.17	13.49		3.02	0.11	0.18	13.98
Desvio padrão	0.55	0.03	0.07	0.99		0.64	0.03	0.07	2.50		0.09	0.02	0.07	0.39

Figura 10. Essa figura mostra a tabela reduzida do experimento 10

A tabela acima mostra os resultados obtidos no experimento 10. Observa-se que ambos os algoritmos tiveram desempenho parecido, exceto pelo tempo de execução. Não foi observada diferença significativa na silhueta e no IRA nos diferentes valores de ‘p’.

6. Conclusão

Este trabalho apresentou a implementação e a avaliação de um algoritmo 2-aproximado para o problema dos k-centros, comparando-o com o algoritmo clássico de agrupamento K-Means. Os experimentos foram realizados em 10 conjuntos de dados distintos, variando o parâmetro 'p' da função de distância de Minkowski. Os resultados obtidos demonstram a eficácia do algoritmo 2-aproximado em comparação com o K-Means, principalmente em termos de qualidade da solução.

A análise dos resultados revelou que, embora o algoritmo 2-aproximado possa não fornecer a solução ótima, ele é capaz de fornecer uma solução com raio aceitável. No entanto, em geral, seus valores de silhueta e IRA foram significativamente piores que o K-Means. É importante notar que o algoritmo 2-aproximado apresentou um tempo computacional significativamente maior do que o K-Means. Além disso, a implementação é eficiente em termos de memória, pois utiliza uma matriz de distância para armazenar as distâncias entre os pontos calculada apenas uma vez, reduzindo assim a necessidade de cálculos de distância repetidos.

Em conclusão, este trabalho contribui para o entendimento e a aplicação de algoritmos aproximativos em problemas de agrupamento de dados. A implementação e os experimentos realizados fornecem uma base sólida para futuras atividades relacionadas à área de Machine Learning.

7. Referências

Alpaydin,E. and Kaynak,C.. (1998). Optical Recognition of Handwritten Digits. UCI Machine Learning Repository. https://doi.org/10.24432/C50P49.

Yeh,I-Cheng. (2008). Blood Transfusion Service Center. UCI Machine Learning Repository. https://doi.org/10.24432/C5GS39.

Lohweg,Volker. (2013). banknote authentication. UCI Machine Learning Repository. https://doi.org/10.24432/C55P57.

Image Segmentation. (1990). UCI Machine Learning Repository. https://doi.org/10.24432/C5GP4N.

Cortez,Paulo, Cerdeira,A., Almeida,F., Matos,T., and Reis,J.. (2009). Wine Quality. UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T.

Nakai,Kenta. (1996). Yeast. UCI Machine Learning Repository. https://doi.org/10.24432/C5KG68.

Antal,Balint and Hajdu,Andras. (2014). Diabetic Retinopathy Debrecen Data Set. UCI Machine Learning Repository. https://doi.org/10.24432/C5XP4P.

Elter,Matthias. (2007). Mammographic Mass. UCI Machine Learning Repository. https://doi.org/10.24432/C53K6Z.

Abdelhamid,Neda. (2016). Website Phishing. UCI Machine Learning Repository. https://doi.org/10.24432/C5B301.