P-Center em Julia

February 12, 2016

1 Trabalho de Implementação

1.1 INF2912 - Otimização Combinatória

1.1.1 Prof. Marcus Vinicius Soledade Poggi de Aragão

1.1.2 2015-2

1.1.3 Ciro Cavani

BigData / Globo.com Algoritmos de clusterização.

1.2 Conteúdo

Esse notebook tem o desenvolvimento e avaliação do algoritmo aproximado do P-Center (algoritmo Farthest-first traversal).

A avaliação do algoritmo é baseada em um mapeamento entre a maioria dos itens que foram atribuídos a um determinado cluster e o correspondente os valores verdadeiros gerados nesse cluster.

O P-Center teve resultados muito bons.

1.3 Dataset

```
In [1]: include("../src/clustering.jl")
        import Inf2912Clustering
        const Clustering = Inf2912Clustering
```

```
Out[1]: Inf2912Clustering
```

```
In [2]: dataset = Clustering.dataset_tiny()
        Clustering.summary(dataset)
        sleep(0.2)
```

```
Number of Groups: 3
Number of Features: 16
Number of Features (group): 3
Probability of Activation: 0.8
Number of Objects (total): 100
Number of Objects per Group (min): 20
Number of Objects per Group (max): 40
Number of Objects in 1: 36
Number of Objects in 2: 34
Number of Objects in 3: 30
```

1.3.1 P-Center - Problema de Localização de Centróides

Consiste em resolver o P-Center determinar os objetos representantes de cada grupo e classificar cada objeto como sendo do grupo com representante mais próximo

https://en.wikipedia.org/wiki/Metric_k-center
https://en.wikipedia.org/wiki/Farthest-first_traversal

```
In [3]: let
            k = 3
            data = map(first, dataset.data)

            centers = Array(Array{Int64,1}, 0)
            i = rand(1:length(data))
            push!(centers, data[i])

            min_dist(v) = minimum(map(c -> norm(c - v), centers))
            max_index() = indmax(map(min_dist, data))

            while length(centers) < k
                i = max_index()
                push!(centers, data[i])
            end

            cluster(v) = indmin(map(c -> norm(c - v), centers))

            assignments = zeros(Int, length(data))
            for (i, v) in enumerate(data)
                assignments[i] = cluster(v)
            end

            assignments
        end
```

```
Out[3]: 100-element Array{Int64,1}:
         2
         3
         1
         1
         2
         1
         2
         1
         1
         2
         1
         2
         1
         ⋮
         1
         3
         1
         1
         2
         2
         2
```

```
 1
 1
 1
 1
 2
```

In [4]:
```
"Algoritmo de clusterização P-Center (algoritmo Farthest-first traversal)."
function pcenter(dataset, k)
    data = map(first, dataset.data)

    centers = Array(Array{Int64,1}, 0)
    i = rand(1:length(data))
    push!(centers, data[i])

    min_dist(v) = minimum(map(c -> norm(c - v), centers))
    max_index() = indmax(map(min_dist, data))

    while length(centers) < k
        i = max_index()
        push!(centers, data[i])
    end

    cluster(v) = indmin(map(c -> norm(c - v), centers))

    assignments = zeros(Int, length(data))
    for (i, v) in enumerate(data)
        assignments[i] = cluster(v)
    end

    assignments
end

pcenter(dataset, 3)
```

Out[4]:
```
100-element Array{Int64,1}:
 1
 1
 3
 2
 2
 3
 2
 2
 2
 3
 3
 1
 2
 ⋮
 1
 3
 1
 1
 2
```

```
 3
 2
 1
 3
 2
 2
 3
```

In [8]:
```
import Clustering.mapping

"Algoritmo de clusterização P-Center (algoritmo Farthest-first traversal) \
aproximado para os grupos pré-definidos do dataset."
function pcenter_approx(dataset, k)
    assignments = pcenter(dataset, k)
    centermap = mapping(dataset, assignments, k)
    map(c -> centermap[c], assignments)
end

let
    k = dataset.groups
    prediction = pcenter_approx(dataset, k)
    Clustering.evaluation_summary(dataset, prediction; verbose=true)
    sleep(0.2)
end
```

```
Matriz de Confusão:

[25 4 7
 6 24 4
 2 1 27]

Tamanho: 100
Acertos: 76
Erros: 24
Accuracy: 76.0%

Cluster 1

Tamanho: 36
Accuracy: 81.0%
Precision: 75.76%
Recall: 69.44%
F-score: 0.72

Acerto positivo: 25 (69.44%)
Acerto negativo: 56 (87.5%)
Falso negativo: 11 (45.83%)
Falso positivo: 8 (33.33%)

Cluster 2

Tamanho: 34
Accuracy: 85.0%
Precision: 82.76%
Recall: 70.59%
```

```
F-score: 0.76

Acerto positivo: 24 (70.59%)
Acerto negativo: 61 (92.42%)
Falso negativo: 10 (41.67%)
Falso positivo: 5 (20.83%)

Cluster 3

Tamanho: 30
Accuracy: 86.0%
Precision: 71.05%
Recall: 90.0%
F-score: 0.79

Acerto positivo: 27 (90.0%)
Acerto negativo: 59 (84.29%)
Falso negativo: 3 (12.5%)
Falso positivo: 11 (45.83%)
```

```
In [6]: Clustering.test_dataset("small", pcenter_approx)
        sleep(0.2)
```

```
0.014421 seconds (37.55 k allocations: 11.260 MB)
Matriz de Confusão:

[407 0 1
 0 304 0
 0 1 287]

Tamanho: 1000
Acertos: 998
Erros: 2
Accuracy: 99.8%

Cluster 1

Tamanho: 408
Accuracy: 99.9%
Precision: 100.0%
Recall: 99.75%
F-score: 1.0

Acerto positivo: 407 (99.75%)
Acerto negativo: 592 (100.0%)
Falso negativo: 1 (50.0%)
Falso positivo: 0 (0.0%)

Cluster 2

Tamanho: 304
Accuracy: 99.9%
Precision: 99.67%
Recall: 100.0%
F-score: 1.0
```

```
Acerto positivo: 304 (100.0%)
Acerto negativo: 695 (99.86%)
Falso negativo: 0 (0.0%)
Falso positivo: 1 (50.0%)

Cluster 3

Tamanho: 288
Accuracy: 99.8%
Precision: 99.65%
Recall: 99.65%
F-score: 1.0

Acerto positivo: 287 (99.65%)
Acerto negativo: 711 (99.86%)
Falso negativo: 1 (50.0%)
Falso positivo: 1 (50.0%)
```

```
In [7]: Clustering.test_dataset("large", pcenter_approx)
        sleep(0.2)
```

```
0.200351 seconds (415.56 k allocations: 113.177 MB, 39.55% gc time)
Matriz de Confusão:

[2295 3 1
 5 3298 8
 5 1 4384]

Tamanho: 10000
Acertos: 9977
Erros: 23
Accuracy: 99.77%

Cluster 1

Tamanho: 2299
Accuracy: 99.86%
Precision: 99.57%
Recall: 99.83%
F-score: 1.0

Acerto positivo: 2295 (99.83%)
Acerto negativo: 7691 (99.87%)
Falso negativo: 4 (17.39%)
Falso positivo: 10 (43.48%)

Cluster 2

Tamanho: 3311
Accuracy: 99.83%
Precision: 99.88%
Recall: 99.61%
F-score: 1.0
```

```
Acerto positivo: 3298 (99.61%)
Acerto negativo: 6685 (99.94%)
Falso negativo: 13 (56.52%)
Falso positivo: 4 (17.39%)

Cluster 3

Tamanho: 4390
Accuracy: 99.85%
Precision: 99.8%
Recall: 99.86%
F-score: 1.0

Acerto positivo: 4384 (99.86%)
Acerto negativo: 5601 (99.84%)
Falso negativo: 6 (26.09%)
Falso positivo: 9 (39.13%)
```