Algoritmos de Agrupamento Aplicados à Classificação de Precipitações

Emanuel Oliveira da Silva, Elloá B. Guedes, Maria Betânia Leal de Oliveira

[1]Escola Superior de Tecnologia
Universidade do Estado do Amazonas
Av. Darcy Vargas, 1200 – Manaus – Amazonas

{eos.eng, ebgcosta, mloliveira}@uea.edu.br

Abstract. *This work aims at proposing a rainfall classification criterion for the city of Manaus based on the adoption of clustering algorithms. In order to achieve such result, it was considered data from* 2010 *to* 2016*, adoption of* 5 *different clusters of events and the use of* k*-means and* k*-medoids algorithms. The results obtained with* k*-means, in particular, were more effective in capturing the higher frequency of low intensity rainfall events as well as the lower frequency of extreme rainfall events. The results obtained can be adoted as rainfall classification criterion for Manaus.*

Resumo. *Este trabalho tem por objetivo colaborar na proposição de um critério de classificação de precipitações para a cidade de Manaus baseado na utilização de algoritmos de agrupamento. Para tanto, considerou-se dados de eventos de precipitação dentre os anos de* 2010 *e* 2016*, a utilização de* 5 *categorias de eventos e a adoção dos algoritmos* k*-means e* k*-medoids. O algoritmo* k*-means, em particular, foi mais efetivo em capturar a maior frequência de eventos de baixa intensidade e a esporadicidade de eventos extremos. Os resultados obtidos podem ser utilizados para classificação de precipitações na cidade em questão.*

1. Introdução

Uma das atribuições da Meteorologia é propor métodos para *classificação de precipitações*, os quais visam criar categorias para a precipitação de acordo com o grau de intensidade. Esta é uma tarefa extremamente importante, pois estas categorias podem auxiliar em diversas outras análises, desde a prevenção de alagamentos até a administração de recursos hídricos. Apesar da sua importância, esta não é uma tarefa fácil. Conceber uma classificação de precipitações envolve aspectos meteorológicos, os quais possuem diversos fatores de natureza dinâmica, e também topologia, localização, etc. Levando em conta estes diversos fatores, que compõem características extremamente particulares, não se conhece uma metodologia geral para criação de sistemas de classificação de precipitações.

A cidade de Manaus, em particular, não possui um sistema de classificação de precipitações. Sioli afirma que as precipitações em Manaus são abundantes, não uniformes e caracterizam um fator importante que molda o clima na cidade [Sioli 1991]. A maior parte das chuvas em Manaus é resultado da influência de muitos sistemas de precipitações, sendo importante ressaltar que a cidade possui um clima de características tipicamente equatorais, com muita umidade e calor [da Silva 2012]. Considerando este contexto, a concepção e proposição de um sistema de classificação de precipitações para esta cidade é uma tarefa necessária, embora desafiadora.

Dada a lacuna observada, o objetivo deste trabalho é apresentar resultados de algoritmos de agrupamento, que seguem o paradigma não-supervisionado, na proposição de um método de classificação de precipitações em Manaus. Foram utilizados dados de 2371 eventos de precipitação registrados na cidade entre os anos de 2010 e 2016 e, como resultado, foram propostos dois critérios diferentes, considerando a utilização dos algoritmos k-means e k-medoids. Também foi feita uma análise comparativa entre os resultados obtidos a partir destas duas abordagens, onde foi possível identificar uma melhor adequação do critério proposto pelo k-means.

Para apresentar os resultados alcançados, este trabalho está organizado como segue: uma fundamentação teórica sobre agrupamento e sobre os algoritmos utilizados é apresentada na Seção 2; uma visão geral sobre o conjunto de dados e suas características é descrita na Seção 3; os resultados das abordagens propostas é apresentado na Seção 4, que também inclui uma análise comparativa; e, por fim, as considerações finais são apresentadas na Seção 5.

2. Agrupamento

Para problemas em que nao se conhece uma solução analítica, a *Aprendizagem de Máquina* (AM), subárea da Inteligência Computacional, permite a elaboração de soluções empíricas com a utilização de modelos e métodos computacionais que aprendem a partir de dados. Neste contexto, o termo "aprender" adquire uma conotação de aumentar a performance sobre uma tarefa em relação à um momento anterior [Witten et al. 2011].

Na AM consideram-se as *tarefas de previsão e de descrição*. Na previsão, a meta é encontrar um modelo, a partir dos dados de treinamento, que seja capaz de prever resultados para entradas anteriormente não conhecidas. Já nas tarefas de descrição, por sua vez, a meta é explorar ou descrever um conjunto de dados. Para tanto, segue-se o *paradigma de aprendizado não supervisionado*, cuja premissa é a detecção espontânea de padrões ou estruturas a partir dos próprios dados [Faceli et al. 2015].

O *agrupamento*, em particular, é um tipo de tarefa de descrição, segundo a qual os dados são agrupados de acordo com sua similaridade. Em um dado grupo, chamado de *cluster*, os dados ali contidos compartilham alguma característica ou propriedade relevante para o domínio do problema em questão [Faceli et al. 2015]. Pode-se também ver o agrupamento como uma partição do conjunto de dados em *clusters* de tal maneira que uma certa medida de similaridade entre qualquer par de observações associadas a um mesmo *cluster* minimiza uma certa função de custo [Haykin 2009].

Formalmente, entende-se por *cluster* um conjunto $\mathbf{C}_k = \{x_1, x_2, \ldots, x_{n_k}\}$ com n_k objetos, que pode ser visto como uma coleção de objetos próximos, ou que satisfazem alguma relação, ou que minimizam uma certa função de custo. O *centróide* do *cluster* $\mathbf{C}_k$, denotado por $\overline{x}^{(k)}$, é o objeto representativo que resume as informações contidas naquele *cluster*, que é dado por:

$$\overline{x}^{(k)} = \frac{1}{n_k} \sum_{x_i \in \mathbf{C}_k} x_i. \tag{1}$$

De maneira geral, a realização de um agrupamento contempla os passos a seguir:

1. **Extração e Seleção de Características**. Extrair e selecionar as características mais representativas do conjunto de dados;

2. **Algoritmo de Agrupamento**. Projetar ou escolher o algoritmo de clusterização de acordo com as caracteríistias do problema considerado. Medidas de distância e similaridade são as bases para a construção desses algoritmos. Para características quantitativas, a distância é preferível para reconhecer relações entre os dados. Já a similaridade é preferível quando se lida com característica qualitativas;
3. **Avaliação dos Resultados**. Avaliar o resultado da clusterização e julgar os resultados fornecidos pelo algoritmo;
4. **Explicação dos Resultados**. Fornecer uma explicação prática para os resultados obtidos na clusterização [Xu and Tian 2015].

Em relação ao algoritmo de agrupamento, necessário no Passo 2, a literatura dispõe de diversas soluções. De maneira geral, cada algoritmo é baseado em um critério de agrupamento, usa uma medida de proximidade e um método de busca para encontrar uma estrutura adequada que descreva os dados, de acordo com o critério de agrupamento adotado [Faceli et al. 2015]. Para os fins deste trabalho, os algoritmos k-means e k-medoids foram considerados, os quais são detalhados nas subseções a seguir.

2.1. Algoritmo *k*-means

Dado um número inteiro k, o *algoritmo k-means* particiona um conjunto de dados fornecidos em k *clusters*. Este algoritmo segue a abordagem particional e visa minimizar o erro quadrático, métrica considerada. Os *clusters* produzidos por este algoritmo são compactos, de formato esférico e podem ser desbalanceados.

Uma ideia geral do algoritmo k-means é dada como segue: seleciona-se aleatoriamente k objetos do conjunto de dados, que serão os centróides iniciais dos k *clusters*. Em seguida, por um processo de realocação iterativa, cada elemento do conjunto de dados é associado ao *cluster* mais próximo. Então, os centróides são recalculados. Este processo continua até que não haja mais alteração nos centróides [Faceli et al. 2015].

2.2. Algoritmo *k*-medoids

Assim como o k-means, o *algoritmo k-medoids* segue a abordagem particional e também visa minimizar uma métrica de distância. Porém, neste algoritmo considera-se encontrar uma quantidade k de *medoids*, isto é, identificar k elementos do conjunto de dados cuja dissimilaridade média dos demais pontos é mínima.

De maneira geral, o algoritmo k-medoids funciona como segue: seleciona-se aleatoriamente k pontos do conjunto de dados como sendo medoids. Em seguida, para cada medoid m e cada elemento do conjunto de dados d, computa-se o custo total desta configuração, isto é, a média de dissimilaridade de d a todos os dados associados a m. Seleciona-se então o medoid o como o menor custo para esta configuração. Os passos em questão são repetidos até que não haja mais alterações nos medoids.

3. Análise do Conjunto de Dados

O conjunto de dados considerado foi obtido do Laboratório de Instrumentação Meteorológica (LabInstru) da Escola Superior de Tecnologia da Universidade do Estado do Amazonas. Este conjunto contém todas as ocorrências de eventos de precipitação entre os anos de 2010 e 2016, nos quais foi registrada a data, horário de início, horário de término, volume de precipitação máxima em 10 minutos e volume de precipitação total. O conjunto de dados possui 2371 eventos de precipitação registrados.

Uma análise mais detalhada dos dados se fez necessária. Inicialmente, derivou-se o atributo duração, representado em minutos, que especifica o intervalo de tempo de

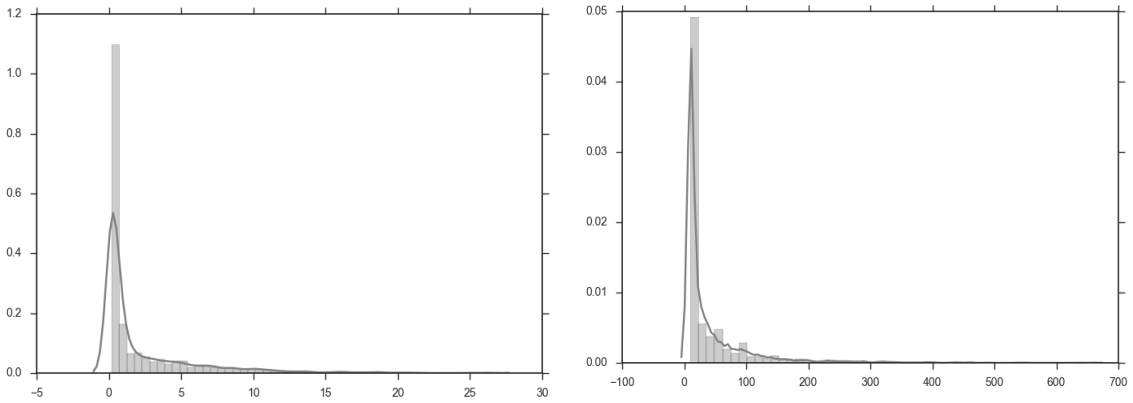
ocorrência do fenômeno de precipitação. Em seguida, considerando os demais dados disponíveis no conjunto de dados, foi realizada a estatística descritiva, sintetizada na Tabela 1. A partir da análise desta tabela, foi possível evidenciar que não há uma homogeneidade na distribuição das variáveis consideradas, possuindo grande dispersão e cauda longa. Estas considerações podem ser mais facilmente visualizadas na Figura 1, na qual é possível visualizar os histogramas das variáveis precipitação e duração.

Tabela 1: Estatística descritiva do conjunto de dados.

	Mediana	**Média**	**Desvio Padrão**	**Mínimo**	**Máximo**
Precipitação (mm)	0.8	5.60	11.79	0.2	125.6
Precipitação Máxima em 10 min (mm)	0.4	2.22	3.62	0.2	26.2
Duração (min)	10	40.77	63.46	10	660

Figura 1: Histogramas dos dados. O eixo x denota o volume em milímetros e o eixo y denota as ocorrências.

(a) Precipitação máxima em 10 min. **(b) Duração.**



Em relação à precipitação máxima em 10 minutos, os dados mostram uma grande quantidade de eventos de pouco volume. Vale ressaltar que nos dados também é possível identificar a existência de uma pequena quantidade de eventos de grande volume de precipitação em 10 minutos. A mesma analogia pode ser feita com os dados de duração. A precipitação máxima em 10 minutos foi escolhida em detrimento da precipitação por fornecer uma visão mais realística da intensidade. Por exemplo, se um volume de 26 mm ocorre em apenas 10 minutos, tem-se uma precipitação muito intensa, com potencial de causar estragos. Porém, se uma precipitação de 100 mm ocorrer ao longo de um dia, a intensidade da mesma estará dispersa no intervalo de tempo de sua duração, diminuindo o pontecial de danos.

Outras análises dos dados também foram realizadas, com o objetivo de identificar o turno de maior ocorrência de precipitações, avaliar o grau de correlação entre a precipitação e a duração, dentre outras. Ao apresentar as análises realizadas aos pesquisadores do LabInstru, utilizando conhecimentos da Meteorologia, a sugestão destes para a classificação dos eventos de precipitação em Manaus contemplou a existência de 5 classes de eventos: Fraco, Moderado, Forte, Muito Forte e Extremo. Porém, ainda restava

Tabela 2: Classificação dos eventos de precipitação segundo o algoritmo k-means.

Intervalos	Classificação
$[0.2, 1.6]$	Fraco
$[1.8, 4.6]$	Moderado
$[4.8, 8.6]$	Forte
$[8.8, 14.6]$	Muito Forte
$[15.0, 26.4]$	Extremo

saber como estas classes seriam organizadas. As propostas de classificação, utilizando algoritmos de agrupamento, são mostradas a seguir.

4. Agrupamentos de Eventos de Precipitação em Manaus

Os eventos de precipitação da cidade de Manaus no período de 2010 a 2016 foram tomados como base para a criação de uma classificação para eventos de precipitação na cidade. Considerou-se a existência de 5 classes de eventos de precipitação (Fraco, Moderado, Forte, Muito Forte e Extremo) correspondendo à intensidade do evento.

A identificação prévia da quantidade de classes dos eventos, realizada por meio de especialistas, é essencial para a execução dos algoritmos de agrupamento considerados, k-means e k-medoids, pois estes demandam a quantidade de classses como parâmetro de entrada.

4.1. Abordagem 1: Utilização do k-means

Ao submeter o conjunto de dados ao k-means, considerando $k = 5$ *clusters*, obteve-se os resultados mostrados na Tabela 2. De acordo com esta classificação, um evento será classificado como Moderado, por exemplo, se a precipitação máxima em 10 min for de 1.8 mm a 4.6 mm.

Em todos os *clusters* obtidos, foi possível perceber que há diferentes intervalos intra-*cluster* (diferença entre o limite superior e o limite inferior de cada *cluster*). Isto se deve ao número de classes e à distribuição do conjunto de dados. O fato do k-means gerar clusters compactos é um aspecto relevante para o domínio considerado, pois cada evento de precipitação irá possuir uma única classe.

Aplicando o critério de agrupamento proposto ao conjunto de dados original, obtém-se o resultado ilustrado na Figura 2. De acordo com o mesmo, aproximadamente 70% dos eventos foram agrupados como sendo da classe Fraco, o que significa que a grande maioria dos eventos de precipitação ocorridos em Manaus possui volume entre 0.2 mm e 1.6 mm máximos em 10 minutos. Os eventos do tipo Extremo, embora de grande intensidade, são de baixa ocorrência, representando cerca de 2% do total de dados.

4.2. Abordagem 2: Utilização do k-medoids

De maneira análoga à abordagem anterior, o conjunto de dados foi submetido ao algoritmo k-medoids considerando a existência de 5 classes diferentes para os eventos de precipitação. Os resultados obtidos encontra-se apresentados na Tabela 3.

De acordo com o agrupamento gerado com o k-medoids, é possível identificar um *gap* entre as classes Muito Forte e Extremo. Isto ocorreu em virtude da inexistência de eventos de precipitação maiores que 14.0 e menores que 14.4 no conjunto de dados.

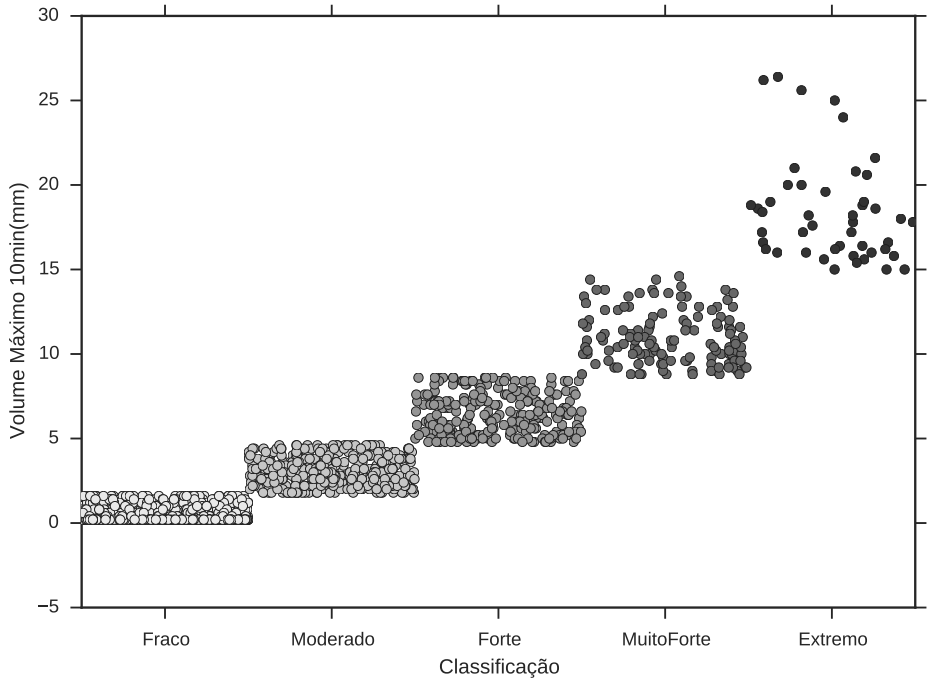


Figura 2: Classificação dos eventos considerando o agrupamento gerado pelo k-means.

Tabela 3: Classificação dos eventos de precipitação segundo o algoritmo k-medoids.

Intervalos	Classificação
$[0.2, 1.4]$	Fraco
$[1.6, 4.0]$	Moderado
$[4.2, 8.0]$	Forte
$[8.2, 14.0]$	Muito Forte
$[14.4, 26.4]$	Extremo

Assim como na abordagem anterior, o conjunto de dados original foi submetido ao agrupamento resultante do k-medoids, obtendo-se o resultado ilustrado na Figura 3.

4.3. Análise Comparativa

Comparando os agrupamentos gerados como os algoritmos k-means e k-medoids, é possível notar algumas similaridades e diferenças entre os agrupamentos gerados. Ambos os algoritmos resultam em agrupamentos com *gaps*. No caso do k-means, o *gap* reside entre as classes Muito Forte e Extremo. No caso do k-medoids, como mencionado, este *gap* reside entre as classes Muito Forte e Extremo. Idealmente, não se desejaria *gaps* entre os agrupamentos gerados, mas isto é uma consequência dos dados disponíveis.

O agrupamento do k-means gerou um *cluster* com maior abrangência para a classe Fraco do que o agrupamento gerado pelo k-medoids. A comparação entre a abrangência das demais classes varia entre os algoritmos. Em relação à eventos da classe Extremo, o k-medoids considerou um intervalo mais abrangente que o k-means, fazendo com que mais eventos fossem agrupados nesta classe.

As medidas de silhueta, que refletem a consistência entre os objetos e os *clusters* a

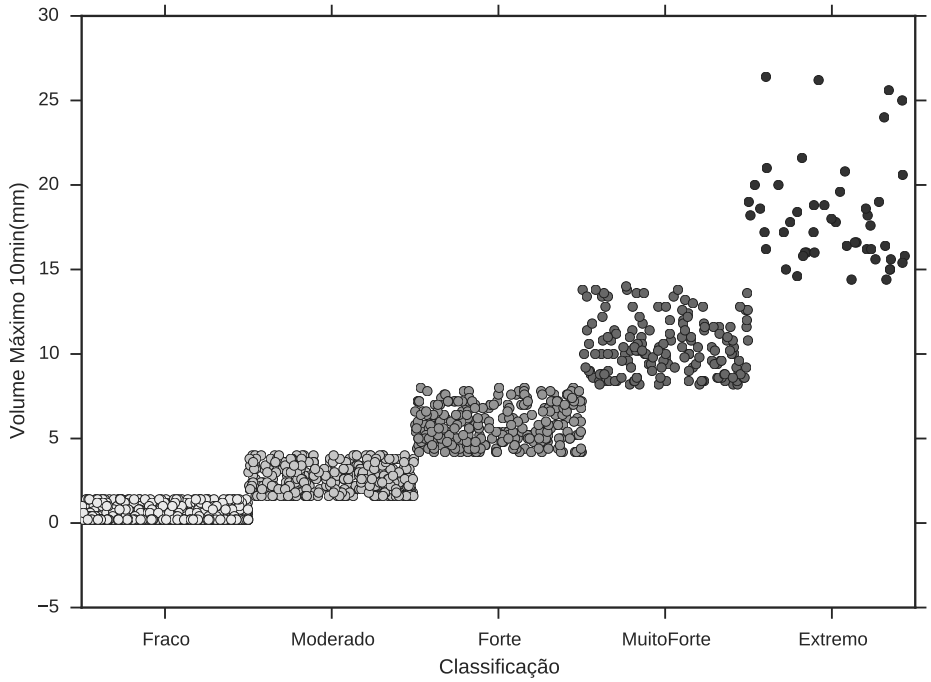


Figura 3: Classificação dos eventos considerando o agrupamento gerado pelo k-medoids.

que estão associados é similar para os agrupamentos gerados pelo k-means e k-medoids, sendo iguais a 0.75 e 0.74, respectivamente.

Considerando a esporadicidade de eventos da classe Extremo e a alta frequência de eventos da classe Fraco, o agrupamento gerado pelo k-means, quando comparado ao agrupamento gerado pelo k-medoids, reflete melhor o cenário realístico, gerando um agrupamento dos eventos de precipitação que pode vir a ser amplamente adotado.

5. Considerações Finais

Considerando dados de eventos de precipitação da cidade de Manaus coletados entre os anos de 2010 a 2016, o objetivo deste trabalho consistiu em propor critérios de classificação de precipitação baseados em algoritmos de agrupamento. Inicialmente, após uma análise estatística e discussão com especialistas da área de Meteorologia, considerou-se então a existência de 5 categorias de eventos de precipitação: Fraco, Moderado, Forte, Muito Forte e Extremo. Além disso, também foi identificado que a precipitação máxima em 10 min seria o atributo mais adequado para classificação desses eventos.

Os dados disponíveis foram agrupados com a utilização dos algoritmos de agrupamento k-means e k-medoids. Como resultado de cada um desses algoritmos, foi possível estabelecer um critério de classificação de precipitações. Este critério foi então aplicado ao conjunto de dados para que a distribuição dos eventos fosse examinada. Com base nestes resultados, foi possível perceber uma melhor adequação do k-means, por abranger uma grande quantidade de eventos do tipo Fraco e capturar adequadamente a baixa frequência de eventos do tipo Extremo.

Em trabalhos futuros, almeja-se utilizar outros algoritmos de agrupamento para criar novos critérios e efetuar outras comparações, eventualmente concebendo um critério de classificação de precipitações para a cidade de Manaus de natureza híbrida, resultado

da combinação de diferentes algoritmos. Além disso, deseja-se conceber meios de prever a ocorrência de eventos extremos, visando a diminuição dos prejuízos que estes podem gerar.

Agradecimentos

Os autores agradecem o apoio financeiro provido pela Universidade do Estado do Amazonas. O autor Emanuel Oliveira da Silva é bolsista do Programa de Apoio à Iniciação Científica da Universidade do Estado do Amazonas e FAPEAM edição $2016 - 2017$.

Referências

da Silva, D. A. (2012). Função da precipitação no conforto do clima urbano da cidade de Manaus. *Revista Geonorte*, 1(5):22–40.

Faceli, K., Lorena, A. C., Gama, J., and de Carvalho, A. C. P. L. F. (2015). *Inteligência Artificial – Uma abordagem de aprendizado de máquina*. Editora LTC, Rio de Janeiro.

Haykin, S. (2009). *Neural Networks and Learning Machines*. 3 edition.

Sioli, H. (1991). *Amazônia: Fundamentos da ecologia da maior região de florestas tropicais*. Vozes, Manaus.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data mining: practical machine learning tools and techniques*. Elsevier, 3 edition.

Xu, D. and Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Ann. Data. Sci.*, 2:165.