KK-médias: teoria e comparação entre diferentes funções de kernels

Mateus Maia & Anderson Ara
Universidade Federal da Bahia
mateusmaia11@gmail.com

Resumo

O principal objetivo da clusterização é, em geral, um agrupamento ou a segmentação de determinadas observações em sub-grupos(clusters), considerando medidas de similaridade como critério de separação. Um dos exemplos clássicos é o K-Means que se utiliza a distância euclidiana entre as observações como medida de similaridade. Entretanto, este método têm algumas limitações, principalmente no que diz respeito a dados que não são linearmente separáveis. São utilizados então, Kernel K-Means para contornar esse problema. Ambos métodos foram implementados no R, e como resultado, notou-se que além de resolver bem a questão da falta da não linearidade o Kernel K-Means também apresenta sutis melhores resultados na clusterização quando comparado ao K-Means clássico.

Introdução

O objetivo do agrupamento é segmentar e dividir o conjunto de dados em subgrupos, usando uma medida de similaridade, a fim de agrupar observações semelhantes para tentar reconhecer padrões ou revelar *insights* de dados. Uma das técnicas de agrupamento mais clássicas é a K-Médias, onde muitos resultados, pesquisas e extensões envolvem esse método ao longo das últimas décadas (Steinley, 2006). Essa técnica operacionaliza a classificação minimizando as distâncias dentro dos grupos (Hartigan, 1975). Apesar da eficiência deste método, existem algumas limitações para identificar clusters separáveis não lineares. Analisando este problema, alguns autores propuseram algumas formas de lidar com o problema da não linearidade (Dhillon *et al.*, 2004). Uma abordagem para lidar com essa situação é o K-Médias Kernel, como uma extensão de K-Means, onde as observações são levadas para um espaço dimensional mais elevado, onde podem ser linearmente separáveis. Embora o truque do kernel tenha sido usado inicialmente no contexto do SVM (Vapnik, 1995), ele pode ser estendido a outros algoritmos baseados em distância (Schölkopf, 2001), que inclui a técnica de clustering K-Médias. Para determinar a eficiência do truque do kernel para resolver o problema de não linearidade na clusterização vários tipos de kernels foram avaliados, assim como diferentes hiperparâmetros para cada um a partir da construção de um grid que avaliava a acurácia para cada um desses hiperparâmetros.

K-Médias

Considerando um conjunto de dados $\mathbf{X} = (x_1, ..., x_n)$, onde $\mathbf{x}$ são as observações do dado, o objetivo principal do K-Means é minimizar

$$L = \sum_{j=1}^{K} \sum_{x_i \in C_j} ||x_i - m_j||^2, \tag{1}$$

em que K representa o número de clusters e

$$m_j = \sum_{xi \in C_j} \frac{x_i}{n_j} \tag{2}$$

refere-se ao centróide pertencente ao cluster j de n_j observações.

O Pseudo-Algoritmo foi construído da seguinte maneira

Algoritmo 1: K-Means

Input: A matriz de dados **X** e o número de clusters **K**,parâmetros da função kernel, critério de convergência **p**
Output: A matriz dos dados **X** e a classe referente a cada uma das observações

while $\varepsilon \geq p$ **do**
  Calcular randomicamente **K** centróides m_j;
  Calcular as distâncias D_{ij} de cada observação x_i em relação à cada centróide;
  Associar a menor distância D_{ij} ao cluster j ;
  Recalcular os centróides a partir da equação (2);
  Calcular ε, como sendo a distância entre entre o centróide anterior e o recém calculado;
end
Retorna a Matriz **X** e seus respectivos clusters.

Kernel K-Médias

O Kernel K-Means pode ser definido como uma extensão do K-Means clássico, em que utilizando-se do *kernel trick*, é possível levar as observações e os centróides para dimensões superiores onde as distâncias serão calculadas. Dessa forma, tomando um conjunto de observações $\mathbf{X} = (x_1, ..., x_n)$ e os clusters C_k em que $k = 1, ..., n$, existem centróides $\mathbf{m}_k$ para cada cluster C_k na dimensão $\mathbb{R}^n$, em que $\Phi : \mathbb{R} \to \mathbb{R}^n$.

Assim, para calcular as distâncias utiliza-se a Matriz-Gram (Schölkopf, 2001), também chama de Matriz Kernel, em que cada elemento é o produto interno entre as observações $\phi(x)$. E assim, a distância final pode ser calculada a partir de

$$D(\mathbf{x}_i, \mathbf{m}_k) = K(x_i, \mathbf{x}_i) - 2\frac{\sum_{x_j \in C_k} K(x_i, \mathbf{x}_j)}{|C_k|} + \frac{\sum_{x_l \in C_k} \sum_{x_j \in C_k} K(x_j, \mathbf{x}_l)}{|C_k|^2} \tag{3}$$

Assim, após associar a menor distância de cada observação para cada um dos k clusters, atualiza-se essa informação aos bancos, de modo a repetir esse processo até que ele atinja a convergência. O algoritimo criado para esta técnica seguiu o pseudo código:

Algoritmo 2: Kernel K-Means

Input: A matriz de dados **X** e o número de clusters **K**, **w** número máximo de iterações.
Output: A matriz dos dados **X** e seus respectivos clusters

Atribui-se clusters aleatórios para cada uma das observações $\mathbf{x}_i$;
Cálculo da matriz de Kernels **K** para o respectivo kernel;
while *A convergência não é atendida não atingiu-se o número máximo de iterações* **do**
  Calcular as distâncias $D(\mathbf{x}_i, \mathbf{m}_k)$ de cada observação x_i em relação à cada centróide;
  Associar a menor distância $D(\mathbf{x}_i, \mathbf{m}_k)$ ao cluster k ;
  Recalcular as distâncias $D(\mathbf{x}_i, \mathbf{m}_k)$;
end
Retorna a Matriz **X** e seus respectivos clusters.

Existem diversos tipos de Kernels que podem ser utilizados nessa técnica de clusterização, dentre os que mais utilizados três foram implementados sendo eles o *Gaussian Kernel*, *Polynomial Kernel*, *Linear Kernel*, apesar de não ser tão comum também foi utilizado o *Exponential Kernel* e *Cauchy Kernel*.

Tipos de Kernel	Gaussiano	Polinomial	Linear	Exponencial	Cauchy
Função Kernel	$e^{-\frac{\|\|x-y\|\|^2}{2\sigma^2}}$	$(x^\intercal y + c)^d$	$x^\intercal y + c$	$e^{-\frac{\|\|x-y\|\|}{2\sigma^2}}$	$\frac{1}{1+\frac{\|\|x-y\|\|^2}{\sigma}}$

Tabela 1: Tipos de Kernels e suas respectivas funções.

Resultados

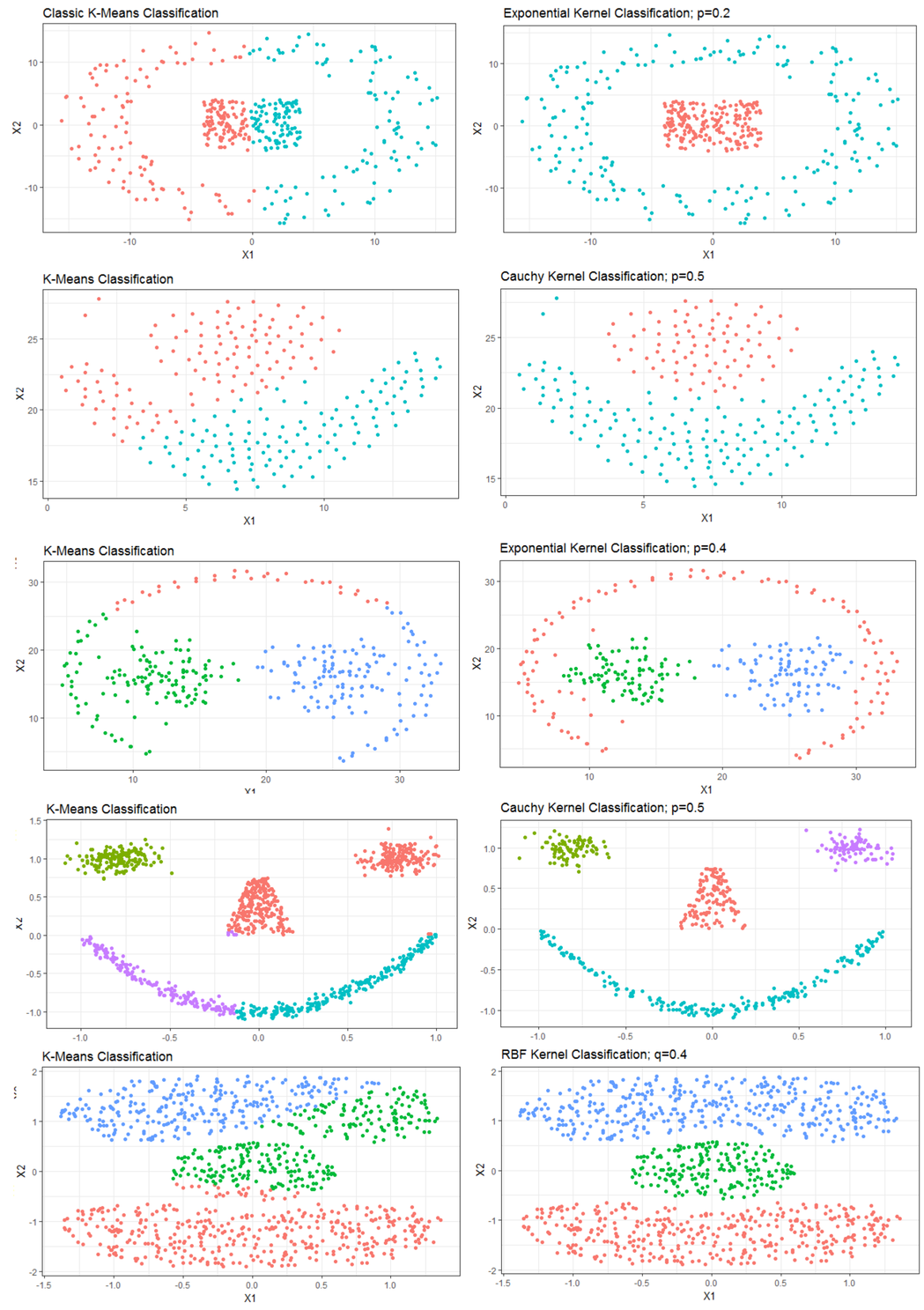

Figura 1: Comparação do K-Means clássico e o Kernel K-Médias que apresentou maior acurácia para cada base de dados.

Todos resultados foram resumidos na Tabela 2 que coloca a medida de acurácia em relação a cada um dos métodos para todas as bases de dados utilizadas. Pode-se perceber que para alguns datasets, principalmente os artificiais, os Kernels apresentaram melhora.A escolha de cada hiperâmetro foi feita após exaustiva avaliação da maioria acurácia variando os hiperparâmetros no intervalo $[0, 0.2]$ com variação de 0.1.

Tabela 2: Acurácia de cada método para cada base de dados

Banco de Dados	K-Means	RBF	Poly.	Lin.	Exp.	Cauchy
Flame	0,841	0,942	0,821	0,829	**0,975**	**0,975**
PathBased	0,743	0.853	0,797	0,773	**0.970**	0.953
Circles	0,531	0,917	**1,000**	0,579	**1,000**	0,912
Smiley	0,551	**1,000**	**1,000**	0,560	**1,000**	**1,000**
Cassini	0,854	**1,000**	0,928	0,928	**1,000**	**1,000**
Iris	0,920	0,893	0,853	0,887	**0,940**	0,900
Seeds	0,889	**0,909**	0,864	0,899	0,884	0,899
Glass	0,887	0,912	0,853	0,798	**0,921**	0,833

Considerações Finais

Através da comparação dos métodos de K-Médias e Kernel K-Médias pôde-se perceber que este segundo foi eficaz em lidar com o problema da não - linearidade comprovada pelo aumento na acurácia. Verificou-se também que a acurácia varia em função da função de kernel, assim como da escolha do hiperparâmetro. Dessa maneira é interessante desenvolver um *framework* na escolha de cada uma desses fatores. Para trabalhos futuros, busca-se desenvolver um *ensemble* com as funções bem kernel, além de outras metodologias para determinação dos hiperparâmetros selecionados.

Referências

Dhillon, I. S., Guan, Y., and Kulis, B. (2004). Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 551–556. ACM.

Hartigan, J. A. (1975). Clustering algorithms, new york: John willey and sons. *Inc. Pages113129*.

Schölkopf, B. (2001). The kernel trick for distances. In *Advances in neural information processing systems*, pages 301–307.

Steinley, D. (2006). K-means clustering: a half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, **59**(1), 1–34.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer science & business media.