Utilização de Centroides Gerados pelo *K-Means* como Conjunto de Treinamento para o Classificador *KNN*: Uma Avaliação de Precisão e *Trade-offs*

Gustavo Zanzin Guerreiro Martins[1]

[1]Departamento Acadêmico de Computação – Universidade Tecnológica Federal do Paraná (UTFPR)
Caixa Postal 86812-460 – Campo Mourão – PR – Brazil

gustavozanzin@alunos.utfpr.edu.br

Abstract. *This study aimed to use the centroids generated by the K-Means algorithm as the training set for the KNN classifier. A dataset of 1000 samples was extracted for training and testing, each containing 132 features of digits from 0 to 9. The samples were normalized and subjected to dimensionality reduction using PCA, resulting in two dimensions. Then, a variable number of centroids was generated using the K-Means algorithm to serve as the training set for the KNN classifier, with k = 5. However, the obtained results showed poor accuracy. Different values of centroids were tested, but all yielded very low accuracy. It is believed that the applied dimensionality reduction and possible errors in the conversion between the training and testing sets contributed to the unsatisfactory results. The experiment's code is publicly available for further analysis.*

Resumo. *Este trabalho teve como objetivo utilizar os centroides gerados pelo algoritmo K-Means como conjunto de treinamento para o classificador KNN. Foram extraídas 1000 amostras de treino e teste, cada uma contendo 132 características dos dígitos de 0 a 9. As amostras foram normalizadas e submetidas à redução de dimensionalidade usando o PCA, resultando em duas dimensões. Em seguida, uma quantidade variável de centroides foi gerada através do K-Means para ser o conjunto de treinamento do KNN, configurado com k = 5. No entanto, os resultados obtidos foram de baixa precisão. Diferentes valores de centroides foram testados, mas todos apresentaram acurácia muito baixa. Acredita-se que a redução de dimensionalidade aplicada e possíveis erros na conversão entre os conjuntos de treinamento e teste tenham contribuído para os resultados insatisfatórios.*

1. Introdução

O experimento realizado consistiu em extrair 1000 amostras de treino e de teste, cada uma com 132 características dos dígitos de 0 a 9, que em seguida foram nomalizados, e a partir disso aplicou-se a redução de dimensionalidade (*PCA*) para duas (2) dimensões e, finalmente, gerou-se uma quantidade *X* de centroides por meio do algoritmo *K-Means* a fim de que esses centróides fossem o próprio conjunto de treinamento para o classificador *KNN* – esse por sua vez, foi configurado com $k = 5$.

2. Resultados

Embora o tempo de execução de todo o processo explanado na Introdução deste relatório tenha sido satisfatório – não tendo ultrapassado a ordem de ~8 segundos -, os resultados de precisão se mostraram péssimos.

Nos testes realizados, foram abordados diferentes valores para as variáveis de centroides do algoritmo K-Means; 5, 10 e 20. O tal experimento gerou resultados de precisão ínfimos conforme mostra a tabela a seguir:

Tabela 1. Descrição dos resultados obtidos na classificação através do KNN para diferentes quantidades de centroides

K-Means Centroides	Valor de K (KNN)	Acurácia
5	**5**	**13%**
10	**5**	**11%**
20	**5**	**14%**

Acredita-se que tais resultados decorreram da significativa redução de dimensionalidade (PCA) aplicada tanto ao conjunto de treinamento, quanto ao conjunto de teste. Além disso, pode existir algum erro na conversão entre os formatos dos conjuntos, tanto de treinamento, quanto de teste em qualquer momento do algoritmo; o código do experimento está público e disponível em https://github.com/GustavoMartinx/Artificial-Intelligence/tree/main/K-means para eventuais análises.

3. Conclusão

Em suma, utilizar os centroides gerados pelo K-Means como conjunto de treinamento para algoritmos de aprendizado supervisionado, como o KNN utilizado nesse experimento, tem suas vantagens como redução da dimensionalidade, uma vez que os centroides representam um conjunto menor de pontos que resumem os agrupamentos encontrados, representando-os de forma compacta. Dessa forma, em comparação com o uso de todos os pontos do conjunto de treinamento original, utilizer os centroides pode resultar emu ma melhoria no desempenho computacional do algoritmo do aprendizado.

Por outro lado, a técnica mencionada apresenta perda de informações devido a essa redução da dimensionalidade.

Sendo assim, é importante avaliar cuidadosamente esses trade-offs e considerar a adequação dessa abordagem para o problema específico em questão.