Utilização do Algoritmo K-Means para encontrar centróides e utiliza-los como dados de treinamento para o KNN

Felipe Archanjo da Cunha Mendes[1]

[1]Bacharelado em Ciência da Computação – Universidade Tecnológica Federal do Paraná
Campo Mourão – PR – Brasil

felipemendes.1999@alunos.utfpr.edu.br

Abstract. *This work describes the utilization of centroids generated by the k-means algorithm in the problem of digit recognition. Five, ten, and twenty centroids were generated for each digit class, replacing the original training instances. Afterwards, a supervised classification model, KNN, was trained and tested with the generated centroids. The accuracy rates obtained were compared with the use of the original features, and the advantages and disadvantages of this approach were discussed. The results showed that the use of centroids significantly improved the accuracy rates compared to the original features, with the accuracy increasing as the number of centroids per class increased. Additionally, the use of centroids resulted in a reduction in processing costs. However, there was a partial loss of information and sensitivity to the initialization of the k-means algorithm.*

Resumo. *Este trabalho descreve a utilização de centróides gerados pelo algoritmo k-means no problema de reconhecimento de dígitos. Foram gerados 5, 10 e 20 centróides para cada classe de dígitos, substituindo as instâncias de treinamento originais. Em seguida, um modelo de classificação supervisionada KNN foi treinado e testado com os centróides gerados. As taxas de acerto obtidas foram comparadas com o uso das próprias features e foram discutidas as vantagens e desvantagens dessa abordagem. Os resultados mostraram que o uso de centróides melhorou significativamente as taxas de acerto em relação às features originais, sendo que a acurácia aumentou com o aumento do número de centróides por classe. Além disso, o uso de centróides proporcionou uma redução no custo de processamento. No entanto, houve perda parcial de informações e sensibilidade à inicialização do algoritmo k-means.*

1. Introdução

O reconhecimento de dígitos é um problema clássico na área de aprendizado de máquina, onde o objetivo é identificar corretamente os dígitos escritos à mão. Neste contexto, o algoritmo k-means pode ser utilizado para gerar centróides que representem as classes dos dígitos. Esses centróides podem ser utilizados para treinar um modelo de classificação supervisionada, como o k-vizinhos mais próximos (k-nearest neighbors - KNN). Este trabalho tem como objetivo descrever as taxas de acerto obtidas ao utilizar centróides gerados pelo algoritmo k-means no problema de reconhecimento de dígitos, comparando com o uso das próprias features, bem como discutir as vantagens e desvantagens dessa abordagem em termos de acurácia e custo de processamento.

2. Desenvolvimento

No código fornecido, o algoritmo k-means do pacote scikit-learn é utilizado para gerar centróides que substituirão as instâncias de treinamento. Para cada classe de dígitos, são

gerados 5, 10 e 20 centróides. Os dados de treinamento e teste são normalizados utilizando a técnica MaxAbsScaler. Em seguida, os modelos KNN são treinados e testados com os centróides gerados, calculando-se a acurácia obtida.

As taxas de acerto obtidas foram as seguintes:

- Utilizando 5 centróides por classe:

  Acurácia no conjunto de teste: 0.53333

- Utilizando 10 centróides por classe:

  Acurácia no conjunto de teste: 0.86667

- Utilizando 20 centróides por classe:

  Acurácia no conjunto de teste: 0.86667

Analisando os resultados, é possível observar que o uso de centróides melhora consideravelmente as taxas de acerto em relação ao uso das próprias features. A acurácia mais baixa obtida com 5 centróides por classe pode ser atribuída à menor quantidade de informações representadas por esses centróides em comparação com os outros casos. Já a acurácia obtida com 10 e 20 centróides por classe é igual e representa um resultado bastante satisfatório.

Em relação ao custo de processamento, utilizar centróides em vez de todas as amostras de treinamento reduz significativamente o tamanho dos dados utilizados para treinamento e teste. Isso resulta em um tempo de treinamento mais rápido, uma vez que menos exemplos são considerados. Além disso, o tempo necessário para calcular a distância entre as instâncias de teste e os centróides é menor do que o tempo necessário para calcular as distâncias para todas as instâncias de treinamento. Portanto, o uso de centróides apresenta uma vantagem em termos de custo de processamento.

Vantagens do uso de centróides:

- Redução da dimensionalidade dos dados: Utilizar centróides permite representar cada classe de dígitos com um número menor de instâncias, reduzindo a dimensionalidade do problema.
- Menor custo de processamento: O uso de centróides reduz a quantidade de dados utilizados no treinamento e teste, resultando em um tempo de processamento mais rápido.
- Generalização: Os centróides podem representar de forma eficiente as características principais de cada classe, permitindo uma melhor generalização do modelo.

Desvantagens do uso de centróides:

- Perda de informações: Utilizar centróides implica em substituir as instâncias de treinamento originais, resultando em uma perda parcial das informações contidas nos dados originais.
- Sensibilidade à inicialização: O algoritmo k-means depende da inicialização dos centróides. Diferentes inicializações podem levar a resultados diferentes, o que pode impactar na qualidade dos centróides gerados.

3. Conclusão

O uso de centróides gerados pelo algoritmo k-means no problema de reconhecimento de dígitos resultou em taxas de acerto superiores em comparação com o uso das próprias features. A acurácia obtida aumentou significativamente quando o número de centróides por classe foi aumentado de 5 para 10, e se manteve estável quando o número foi aumentado para 20. Além disso, o uso de centróides apresentou vantagens em termos de custo de processamento, reduzindo o tamanho dos dados utilizados e acelerando o tempo de treinamento. No entanto, é importante considerar as desvantagens, como a perda parcial de informações e a sensibilidade à inicialização do algoritmo k-means. Em resumo, o uso de centróides mostrou-se uma abordagem promissora para o problema de reconhecimento de dígitos, oferecendo melhorias nas taxas de acerto e no custo de processamento.