Avaliação de Técnicas de Aprendizado e Classificação Automática em um problema de Classificação

Fernando Azevedo Peres
Vitória, Espirito Santo

Abstract

Este trabalho consiste em realizar uma comparação experimental entre um conjunto pré-definido de técnicas de aprendizado e classificação automática aplicadas a um problema de classificação. As técnicas escolhidas são: ZeroR (ZR), Naive Bayes Gaussiano (NBG), KMeans Centroides (KMC), K Vizinhos Mais Próximos (KNN) e Árvore de Decisão (AD).

Keywords: Classificação, ZR, NBG, KMC, KNN, AD

1. Introdução

Com o intuito de estudar algoritmos de classificação, foi proposta a realização de uma comparação experimental entre um conjunto de técnicas de aprendizado e classificação automática. As técnicas escolhidas para o estudo estão brevemente explicadas abaixo:

1. **ZeroR** [1] (ZR): ZeroR é o método de classificação mais simples que se baseia no alvo e ignora todos os preditores. O classificador ZeroR simplesmente prevê a categoria majoritária (classe). Embora não haja poder de previsibilidade no ZeroR, ele é útil para determinar um desempenho de linha de base como referência para outros métodos de classificação.
2. **Naive Bayes Gaussiano** [2] (NBG): O classificador Naive Bayesiano é baseado no teorema de Bayes com as suposições de independência entre os preditores. Um modelo Naive Bayesiano é fácil de construir, sem estimativas complicadas de parâmetros iterativos, o que o torna particularmente útil para conjuntos de dados muito grandes. Apesar de sua simplicidade, o classificador Naive Bayesian geralmente se sai surpreendentemente bem e é amplamente utilizado porque geralmente supera métodos de classificação mais sofisticados.
3. **K Means Centroid** (KMC): Este classificador é uma implementação própria seguindo alguns requisitos propostos pelo professor orientador. Será explicado na Seção 3.

4. **K Vizinhos Mais Próximos** [3] (KNN): K vizinhos mais próximos é um algoritmo simples que armazena todos os casos disponíveis e classifica novos casos com base em uma medida de similaridade (por exemplo, funções de distância).
5. **Árvore de Decisão** [4] (AD): A árvore de decisão constrói modelos de classificação ou regressão na forma de uma estrutura de árvore. Ele divide um conjunto de dados em subconjuntos cada vez menores enquanto, ao mesmo tempo, uma árvore de decisão associada é desenvolvida de forma incremental. O resultado final é uma árvore com nós de decisão e nós folha. Um nó de decisão tem duas ou mais ramificações. O nó folha representa uma classificação ou decisão. O nó de decisão mais alto em uma árvore que corresponde ao melhor preditor chamado nó raiz.

2. Base de dados

A base de dados escolhida para realizar essa análise foi Wine[1]. Os dados presentes nesse dataset são os resultados de uma análise química de vinhos cultivados na mesma região da Itália, mas derivados de três cultivares diferentes.

Descrição do Domínio. Todas as características presentes no dataset são contínuas.

Definição das Classes e das Características. Existem 3 classes presentes nesse dataset que representam os 3 cultivadores diferentes. Existem 13 características que representam a composição química dos vinhos.

Características:

1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline

[1]Informações do dataset estão disponíveis neste link

Número de Instâncias. O dataset Wine é composto por um total de 178 instâncias. A distribuição das instâncias por classes se dá da seguinte forma: Classe 0: 33.15%, Classe 1: 39.89%, Classe 2: 26.97%.

3. O método KMC

O classificador KMC utiliza um algoritmo de agrupamento para definir K grupos de exemplos de cada classe na base de treino. Assumindo que uma base de dados possui ncl classes, o algoritmo KMC forma inicialmente K*ncl grupos, sendo K grupos em cada uma das ncl classes. Em seguida, são calculados os centróides de cada um dos grupos (esses centroides são calculados a partir do algoritmo de clusterização KMeans [5]) e este centróide é associado a classe do grupo a partir do qual foi gerado. O método possui como hiperparâmetro o valor de K.

Para realizar uma classificação, o KMC verifica qual o centróide mais próximo do elemento a ser classificado e retorna a sua classe.

4. Descrição dos Experimentos Realizados e seus Resultados

Experimentos. Para o pré-processamento dos dados foi aplicado o método de normalização z-score, que consiste em calcular, para cada característica, quantos desvios padrões o valor difere da média.

Os algoritmos ZR e NBG são algoritmos que não possuem hiperparâmetros, logo foi realizado treino e teste com 3 rodadas de validação cruzada estratificada de 10 folds.

Os algoritmos KMC, KNN e AD precisam de ajuste de hiperparâmetros. Neste caso o procedimento de treinamento, validação e teste será realizado através de 3 rodadas de ciclos aninhados de validação e teste, com o ciclo interno de validação contendo 4 folds e o externo de teste com 10 folds. A busca em grade (Grid Search[2]) do ciclo interno considerou os seguintes valores de hiperparâmetros de cada técnica de aprendizado: KMC: [k = 1, 3, 5, 7], KNN: [n_neighbors = 1, 3, 5, 7], AD: [max_depth = None, 3, 5, 10].

Resultados. A Tabela 1 mostra os resultados obtidos para cada classificador, levando em conta média, desvio padrão, e intervalo de confiança 95% de significância da acurácia.

A Tabela 2 mostra os testes de hipótese entre os pares de métodos. Na matriz triangular superior são apresentados os resultados do teste t pareado (amostras dependentes) e na matriz triangular

[2]Método sistemático para otimização de hiperparâmetros.

Table 1: Resultados dos Classificadores

Método	Média	Desvio Padrão	Limite Inferior	Limite Superior
ZR	0.40	0.02	0.39	0.41
NBG	0.97	0.05	0.96	0.99
KMC	0.96	0.04	0.95	0.98
KNN	0.96	0.05	0.94	0.98
AD	0.89	0.08	0.86	0.92

inferior são apresentados os resultado do teste não paramétrico de wilcoxon. Os valores da célula da tabela que rejeitarem a hipótese nula para um nível de significância de 95% estão escritos em negrito.

Table 2: p-values pareados

ZR	**0.0**	**0.0**	**0.0**	**0.0**
0.0	NBG	0.296	0.053	**0.0**
0.0	0.291	KMC	0.844	**0.0**
0.0	0.115	0.988	KNN	**0.0**
0.0	**0.0**	**0.0**	**0.0**	AD

A figura 1 mostra um boxplot dos resultados de cada classificador em cada fold.

5. Conclusões

Análise Geral dos Resultados. A partir da Tabela 2, podemos observar que os resultados que rejeitam a hipótese nula são os que envolvem os classificadores ZeroR e AD. Podemos então concluir que o desempenho de ZeroR e AD é incomparável com os demais métodos. O baixo desempenho apresentado pelo ZeroR era esperado, já que é um classificador baseline. O classificador AD apresentou um resultado melhor que o ZeroR porém bem abaixo dos demais, não se mostrando uma boa escolha para este dataset.

Já em relação aos métodos NBG, KMC e KNN, podemos afirmar que seu desempenho é estatisticamente equivalente para o dataset usado neste experimento, onde qualquer um destes é uma boa escolha.

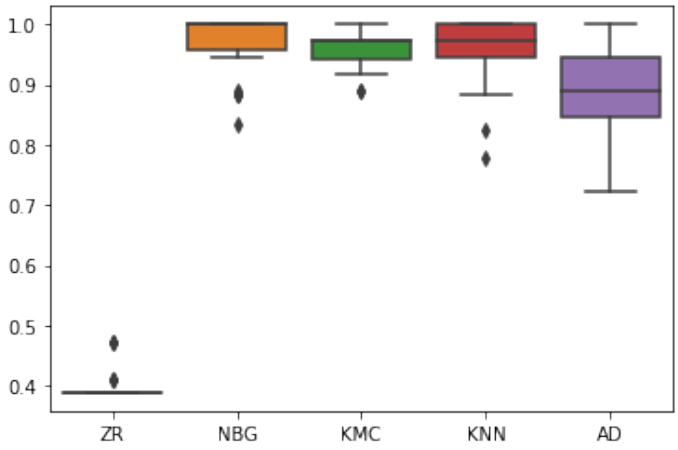

Figure 1: Boxplot comparativo dos Classificadores

Contribuições do Trabalho. Este trabalho contribuiu para a consolidação do conhecimento de todos os aspectos a respeito do uso de classificadores: pré-processamento de dados, escolha de hiperparâmetros, uso/implementação de classificadores, visualização e análise estatística dos resultados.

Melhorias e Trabalhos Futuros. Uma proposta de atividade para trabalhos futuros seria a variação dos hiperparâmetros do classificador AD, para verificar se seu desempenho continua abaixo do esperado. Outra proposta seria utilizar estes métodos em outros datasets com domínio igual, e numero de instancias similar ao Wine, para verificar se este resultado é generalizado para datasets com essas características.

References

[1] S. Sayad, Zero r, `https://www.saedsayad.com/zeror.htm` (Jun. 2022).

[2] S. Sayad, Gaussian naive bayes, `https://www.saedsayad.com/naive_bayesian.htm` (Jun. 2022).

[3] S. Sayad, K neighbors classifier, `https://www.saedsayad.com/k_nearest_neighbors.htm` (Jun. 2022).

[4] S. Sayad, Decision tree, `https://www.saedsayad.com/decision_tree.htm` (Jun. 2022).

[5] S. Sayad, K means, `https://www.saedsayad.com/clustering_kmeans.htm` (Jun. 2022).