Trabalho 1 - Comparação Experimental de Técnicas de aprendizado e classificação automática

Danilo Erler Lima

Vitória , Espirito Santo

[a] *Universidade Federal do Esprito Santo*

Abstract

O trabalho consiste na construção e analise de classificadores para o processo de detecção de classes de vinhos utilizando o wine dataset fornecido pela biblioteca sklearn. Dentro da proposta utilizamos os classificadores: Dummy Zero R (ZR), Naive Bayes Gaussian (NBG), K Nearest Neighbor (KNN), Decision Tree (DT), além da construção de um classificador baseado no uso do método KMeans, que o denominamos de: K Means Centroids (KMC). Para avaliação e comparação foram computadas métricas de acurácia dos classificadores. Cada classificador foi submetido a um processos de validação cruzada aninhada para a seleção dos melhores hiper-parametros, e depois a multiplas rodadas de execução. Os resultados obtidos demonstraram que os classificadores KNN, KMC e NBG obtiveram os melhores resultados (com acurácias médias de 0.96, 0.96 e 0.97 respectivamente).

1. Introdução

O trabalho desenvolvido procurou utilizar de técnicas de aprendizado de máquina [1] para o processo de classificação da base de dados wine fornecida pela biblioteca scikit learn. Para isso, lançamos de mão técnicas de classificação já difundidas, sendo elas: Naive Bayes Gaussian (NBG) [2], K Nearest Neighbor (KNN) [3] e Decision Tree (DT) [4].

Além disso, como proposta de desenvolvimento foi realizado a criação de um classificador baseado no método KMeans, que denominamos de K Means Centroids (KMC) explorando uma nova técnica de classificação.Ao final de tudo, utilizamos o classificador ingênuo Zero R (ZR) para realizar medidas comparativas.

Os experimentos realizados estão disponiveis no repositório publico no github:
https://github.com/daniloelima/Trab-01-IA

Dessa forma, o artigo está organizado segundo a seguinte estrutura: seção 2 Base de Dados, explicando sobre o wine data set e suas caracteristicas, seção 3 O Método KMC , com a descrição

do método desenvolvido, seção 4 Descrição dos Experimentos Realizados e seus Resultados, demonstra os resultados obtidos pelos classificadores fazendo comparações de desempenho, seção 5 Conclusão, apresenta as conclusões obtidas com o trabalho e indicando futuras propostas .

2. Base de Dados

A base de dados wine consiste na analise quimica de vinhos produzidos na mesma região da italia, contendo 178 instâncias de vinhos divididas em 3 classes que representam distintos cultivos, em que cada uma das instâncias contém 13 parametros.

2.1. Descrição do Domínio

Os dados extraídos são todos contínuos e com valores positivos.

2.2. Definição das Classes e das Características

O dataset é composto por 3 classes de diferentes cultivos, sendo extraidas por cada instâncias as seguintes características: 1) Álcool 2) Ácido málico 3) Cinzas 4) Alcalinidade das cinzas 5) Magnésio 6) Fenóis totais 7) Flavonóides 8) Fenóis não flavonóides 9) Proantocianinas 10) Intensidade da cor 11) Matiz 12)OD280/OD315 de vinhos diluídos 13) Prolina.

2.3. Número de Instâncias

O data set é composto por 178 instâncias divididas conforme a seguinte tabela:

Classes	class 0	class 1	class 2
Quantidade	59	71	48

Table 1: Divisão das instâncias por classe do wine dataset

3. O Método KMC

Como parte da proposta do trabalho foi desenvolvido um novo método para o processo de classificação, o K Means Centroids, nele o processo de fit consiste em aplicarmos o algoritmo KMeans para a construção de K centroids de cada uma das classes do dataset a ser classificada, já o processo de predição se da por meio de calcular a distância euclidiana do ponto analisado a todos os centroids antes estabelecidos, sendo a instância classificada conforme a classe do centroid mais próximo.

Para o uso do classificador utilizamos de recursos das bibliotecas scikit learn e scipy, como: base.BaseEstimator, cluster.KMeans e distance.euclidean

4. Descrição dos Experimentos Realizados e seus Resultados

O experimento nesse trabalho consiste no processo de classificação, que passou por 5 etapas: Coleta de Dados, Processamento desses Dados, Busca dos Hiperparametros, Técnicas de Classificação e Avaliação do Desempenho.

- Coleta de Dados:

  Utilizado o dataset fornecido pelo scikit learn descrito préviamente na seção 2 Base de Dados.

- Processamento dos Dados:

  Para o processamento dos dados foi utilizado o método StandartScaller, que processa as amostras conforme a fórmula:

  $z = (x - u)/s$

  No qual 'x' é amostra a ser processada, 'u' e 's' são respectivamente a média e o desvio padrão das amostras de treino.

- Busca dos Hiperparametros

  Esse passo foi utilizado apenas para os classificadores que possuiam hiperparametros a serem definidos, para ele foi utilizado a técnica de validação cruzada aninhada, com o uso do Grid Search. Os hiperparametros buscados por cada classificador foram: DT: (max-depth: 3, 5, 10), KNN (n-neighbors: 1, 3, 5, 7) e o KMC: (num-centroids: 1, 3, 5, 7).

- Técnica de classificação

  Para a classificação foram utilizados os classificadores Dummy Zero R (ZR), Naive Bayes Gaussian (NBG), K Nearest Neighbor (KNN), Decision Tree (DT), K Means Centroids (KMC), com o uso do RepeatedStratifiedKFold para realização de multiplos testes configurado com a divisão de 10 splits dos folds com 3 rodadas de repetição.

- Avaliação do Desempenho

  Como métrica de resultado computamos a acurácia dos classificadores computando ao final 30 resultados para cada um dos classificadores, a partir daí extraimos as estatísticas: média, desvio padrão, limite inferior e superior.

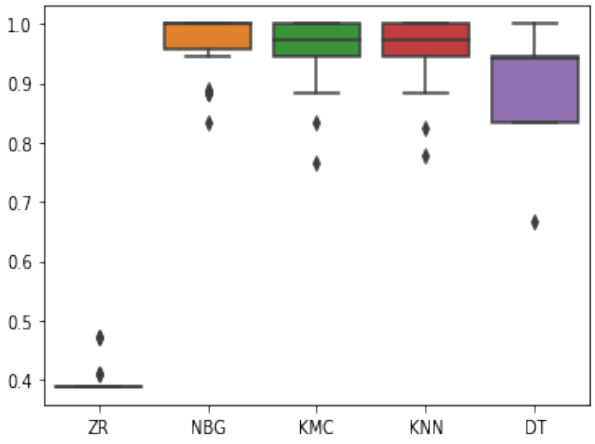

Figure 1: Boxplot com os resultados de cada classificador

Método	Média	Desvio Padrão	Limite Inferior	Limite Superior
ZR	0.39934	0.02471	0.39051	0.40819
NBG	0.97342	0.04820	0.95617	0.99067
KMC	0.96601	0.04808	0.94881	0.98322
KNN	0.96046	0.05319	0.94141	0.97949
DT	0.88965	0.07209	0.86385	0.91545

Table 2: Média, Desvio Padrão, Limite Superior e Inferior das acurácias de cada um dos classificadores

Foram obtidos os resultados conforme a tabela e o gráfico seguintes:

Computamos além disso os p-values das métricas 'wilcoxon' e 'ttest rel' de comparação entre os classificadores, em que aquelas com resultado inferior a 0.05, rejeitamos a ideia de "Null Hypothesis" assegurando que existe efeito significante.

ZR	**0**	**0**	**0**	**0**
0.0000008	NBG	0.2169484	0.0527121	**0**
0.0000011	0.3307432	KMC	0.3321132	**0.0000001**
0.0000012	0.1145886	0.4536952	KNN	**0.0000001**
0.0000015	**0.0000182**	**0.0000438**	**0.0000206**	DT

Table 3:

5. Conclusão

5.1. Análise geral dos resultados

De forma geral foi possivel observar que o classificador construido conseguiu obter ótimos resultados, bem próximos a um modelo ideal com média superior a 95% de precisão, tendo métricas similares a classificadores já consolidados como o Naive Bayes Gaussian e o K Nearest Neighbor e superando até mesmo o Decision Tree, muito embora seja evidente que a base de dados utilizada facilite esses resultados.

5.2. Contribuições do Trabalho

A principal contribuição do trabalho é a exploração de uma nova metodologia no processo de classificação, que embora em estados inicias já conseguiu obter resultados otimistas comparados com classificadores já consolidados.

5.3. Melhorias e trabalhos futuros

Dentre as propostas para trabalhos futuros, algumas seriam explorar outras métricas para o uso dos centroids, além da distância euclidiana e/ou utilizando não apenas o centroid mais proximo. Além de realizar experimentos utilizando outras bases de dados, com propriedades diferentes do wine dataset, por exemplo com dados desbalanceados.

Referencias Bibliográficas

References

[1] I. El Naqa, M. J. Murphy, What is machine learning?, in: machine learning in radiation oncology, Springer, 2015, pp. 3–11.

[2] K. P. Murphy, et al., Naive bayes classifiers, University of British Columbia 18 (60) (2006) 1–8.

[3] L. E. Peterson, K-nearest neighbor, Scholarpedia 4 (2) (2009) 1883.

[4] P. H. Swain, H. Hauska, The decision tree classifier: Design and potential, IEEE Transactions on Geoscience Electronics 15 (3) (1977) 142–147.