Universidade Federal de Pernambuco

Sistemas de Múltiplos Classificadores

Relatório - Lista 3

Hartur Barreto Brito

Sumário

1 Introdução **3**
1.1 Análise da base utilizada 3
1.1.1 Forest type mapping Data Set 3
1.2 Metodologia 3

2 Questão 1 **4**

3 Questão 2 **5**

4 Questão 3 **6**

5 Referências **9**

1 Introdução

Para implementação dos algoritmos de poda foi utilizada a linguagem *Python*, assim como no algoritmo de geração de *ensembles* de classificadores, aplicando o *Bagging*, que foi reaproveitado da solução desenvolvida durante a resolução da *Lista 1*.

O classificador utilizado no *ensemble* foi o SGD devido a sua simplicidade e velocidade de execução. Tendo em vista que o intuito dessa lista é o de observar o comportamento da técnica de poda, não foi feita uma análise aprofundada a respeito do classificador utilizado.

1.1 Análise da base utilizada

1.1.1 Forest type mapping Data Set

Essa base pode ser encontrada na base de dados UCI [1], mais especificamente em [3]. Originalmente, as classes dos dados são: 'd' (Mixed deciduous forest), 'h' (Hinoki forest), 'o' (Other non-forest land) e 's' (Sugi forest) mas, para possibilitar a utilização do *Sci-kit Learn*, elas foram convertidos em números, conforme mostrado na tabela 1. A tabela 1 também mostra como estão distribuídos as 523 amostras dessa base.

Código Original	Código Utilizado	Classe	Amostras
d	1	Mixed deciduous forest	159
h	2	Hinoki forest	86
o	3	Other non-forest land	83
s	4	Sugi forest	195

Tabela 1: Distribuição de amostras de cada classe da base *Forest type mapping*

1.2 Metodologia

Para obtenção de resultados foi seguida a mesma metodologia da primeira lista, que consiste em executar o algoritmo 10 vezes e utilizar a média dos resultados como resultado final. Os passos executados nessa lista são:

1. Mistura dos dados lidos da base;
2. Separação de dados de treino e teste;
3. Geração do *ensemble*;
4. Cálculo da precisão do *ensemble.*;
5. Separação dos dados de treino (anteriormente separado) entre dados para serem utilizados no treinamento dos classificadores do *Bagging* e dados para realização da poda;
6. Geração do *ensemble*;
7. Aplicação da poda;
8. Teste do *ensemble* após realização da poda para obter taxa de acerto.

2 Questão 1

O processo de poda consistem na exclusão de alguns classificadores do *pool* baseado na ideia de que nem sempre grandes *ensembles* são melhores [5], ou seja, "muitos" pode ser melhor que "todos" [6].

Algumas medidas que poderiam ser analisadas para comparar o desempenho de um *pool* de classificadores antes e após a aplicação da poda são:

- Velocidade de treinamento e execução dos classificadores:
  - Uma das principais motivações do processo de poda é a diminuição da quantidade de classificadores, diminuindo também o tempo necessário para o treinamento e execução.
- Taxa de acerto do *pool*:
  - O processo de poda é aplicado para que seja diminuída a quantidade de classificadores no *pool*, como consequência o tempo de treinamento e execução, sem afetar a taxa de acerto.

3 Questão 2

A técnica de poda é aplicada com o objetivo de aumentar o desempenho de um *pool* de classificadores. Para atingir esse objetivo pode ser aplicada utilizado uma medida de diversidade. Ao realizar o cálculo da diversidade de cada classificador, é gerando um *ranking* de classificadores de acordo com o valor obtido dessa medida. Uma das técnicas com essa abordagem é a *Keppa Pruning*, que utiliza a medida de diversidade *Keppa-Statistic* (k_p).

Entretanto, como observado na Lista 2, grande parte das medidas de diversidade não representam necessariamente a taxa de acerto dos classificadores, tornando-as insuficientes para determinar o "*ranking*" de um classificador.

Uma tentativa de corrigir esse problema, seria a adição da taxa de acerto P do classificador no conjunto de teste no cálculo do valor utilizado no *ranking*. Para adicionar a taxa de acerto, seria necessário inverter o valor obtido pela medida *Keppa-Statistic* $(1 - k_p)$, tendo em vista que quanto mais próximo de 1, mais similar são os classificadores. Também seria necessário além de dividir o resultado por 2, para que o valor final da contribuição C permaneça entre 0 e 1.

A representação dessa análise pode ser verificado na equação 1.

$$C = (1 - k_p) + P \tag{1}$$

A partir da soma da contribuição C, comparando todos os classificadores entre si, os classificadores seriam ordenados em ordem decrescente de contribuição ($ConInd$), como mostrado na equação 1 e seriam selecionados os M melhores classificadores.

$$ConInd_x = \sum_{i=1}^{M} C_i \tag{2}$$

4 Questão 3

O resultado do experimento realizado para verificação da eficiência do algoritmo *EPIC* pode ser analisado utilizando os gráficos das figuras 1, 2, 3, 4 e 5, que comparam a taxa de acerto do *pool* inicial, gerado utilizando o *Bagging*, e o *pool* gerado utilizando o *Bagging* e aplicando a poda *EPIC*.

Inicialmente o conjunto de amostras foi separado entre 2 grupos. 60% foi utilizado como treino do primeiro *pool* gerado com *Bagging* (utilizado para comparar resultados). Desse subconjunto, 60% foi utilizado para treino do *pool*, gerado utilizando o *Bagging*, e 40% utilizado para cálculo dos parâmetros do *EPIC* ($\alpha_{i,j}$, $\beta_{i,j}$, $\theta_{i,j}$).

Os outros 40% foi utilizado para realização de testes dos *pools* gerados (com e sem poda).

Analisando os gráficos, é observado que o *ensemble* gerado após a aplicação da poda obteve uma taxa de acerto média abaixo do *ensemble* gerado utilizando apenas o *Bagging*. Entretanto, a diferença entre eles foi abaixo de 5% em todos os casos.

O tempo de execução da classificação não foi obtido por não ter sido observado nenhuma diferença. Como a base utilizada não possui uma quantidade de parâmetros grande o suficiente, o tempo de execução foi insignificante, não possibilitando sua comparação.

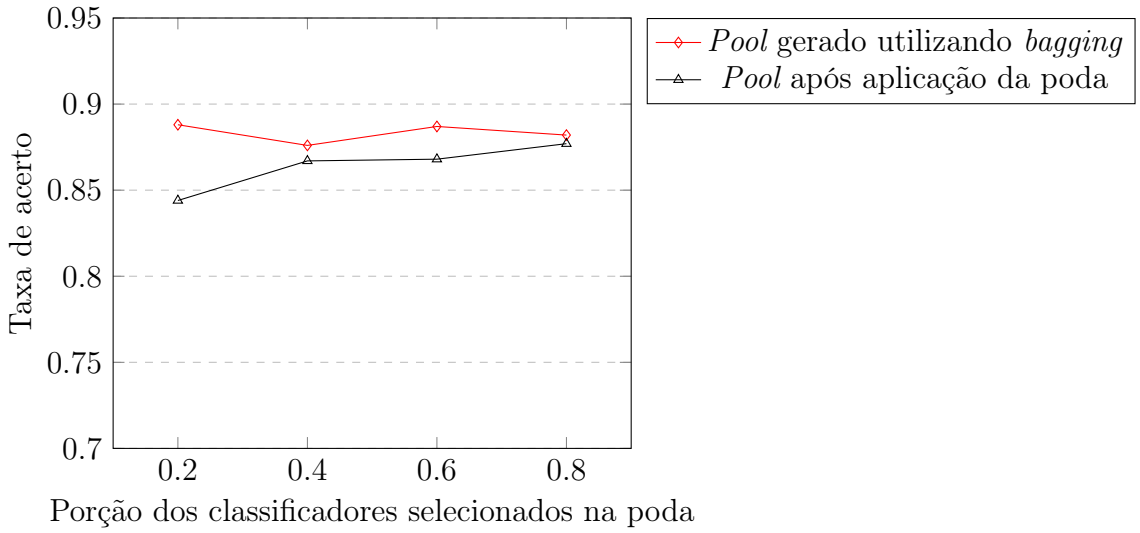

Figura 1: Comparação da taxa de acerto do *pool* com todos os classificadores e do *pool* após aplicação da poda, sendo o *pool* gerado utilizando *Bagging* com 20 classificadores.

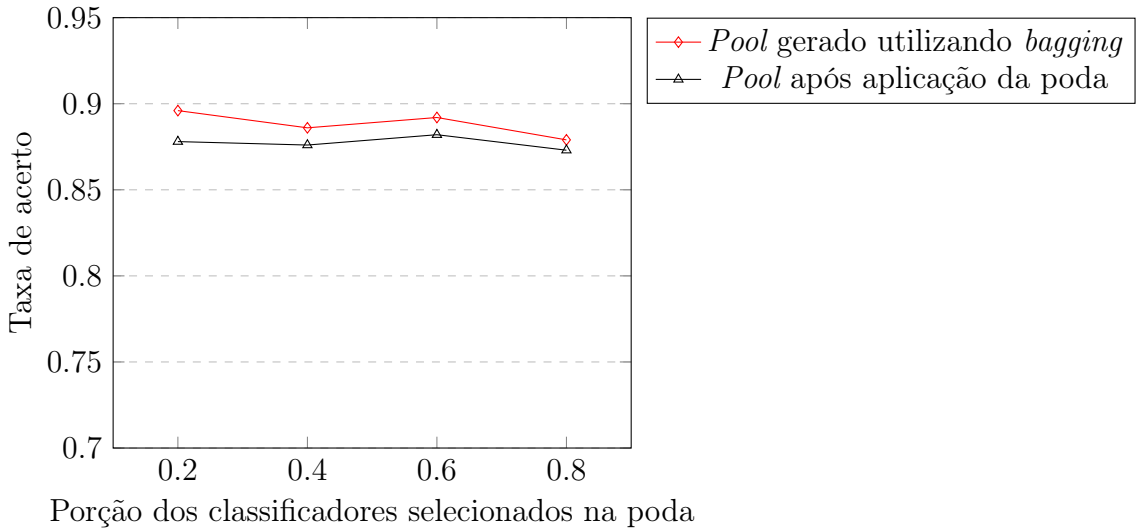

Figura 2: Comparação da taxa de acerto do *pool* com todos os classificadores e do *pool* após aplicação da poda, sendo o *pool* gerado utilizando *Bagging* com 40 classificadores.

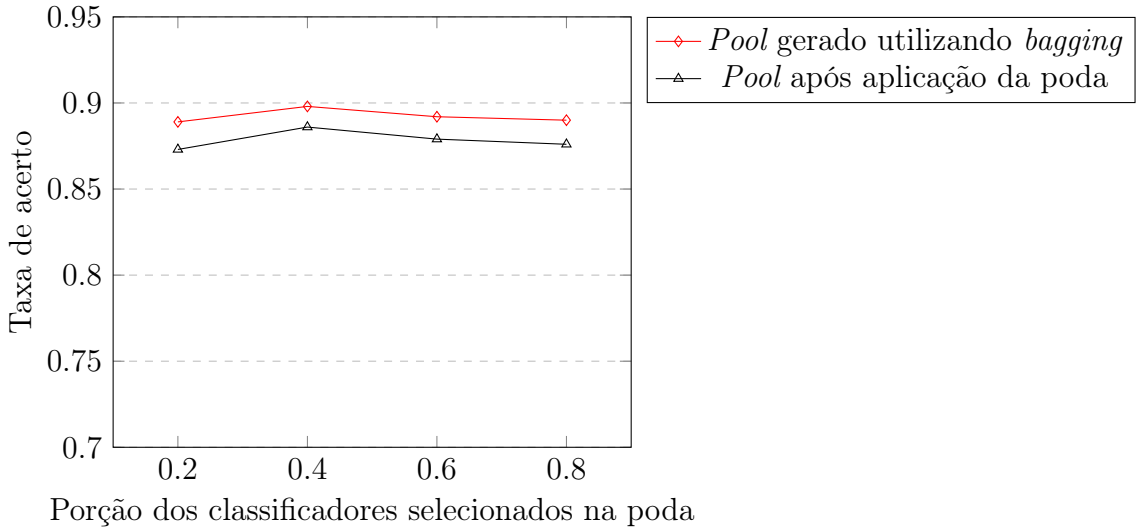

Figura 3: Comparação da taxa de acerto do *pool* com todos os classificadores e do *pool* após aplicação da poda, sendo o *pool* gerado utilizando *Bagging* com 60 classificadores.

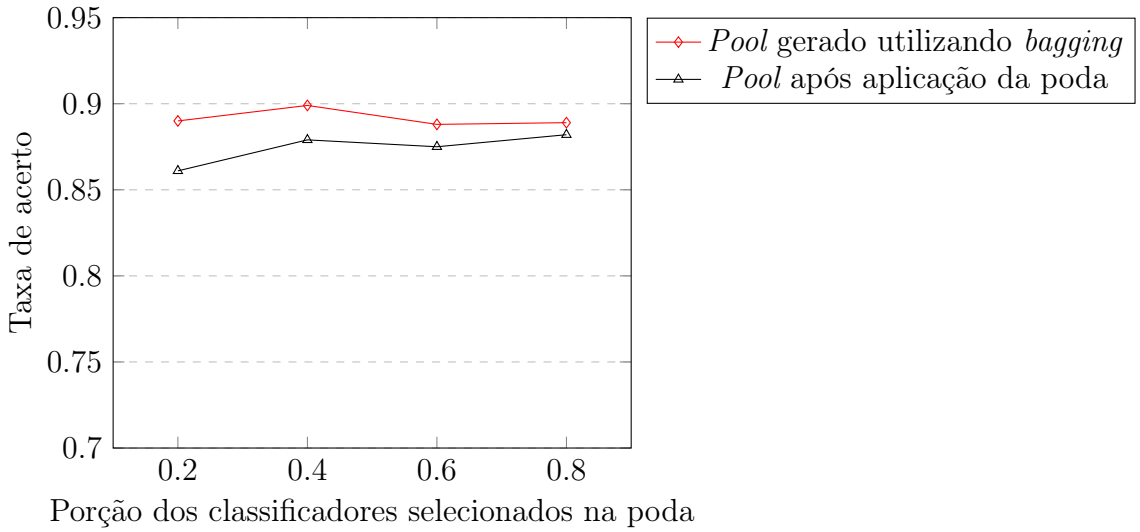

Figura 4: Comparação da taxa de acerto do *pool* com todos os classificadores e do *pool* após aplicação da poda, sendo o *pool* gerado utilizando *Bagging* com 80 classificadores.

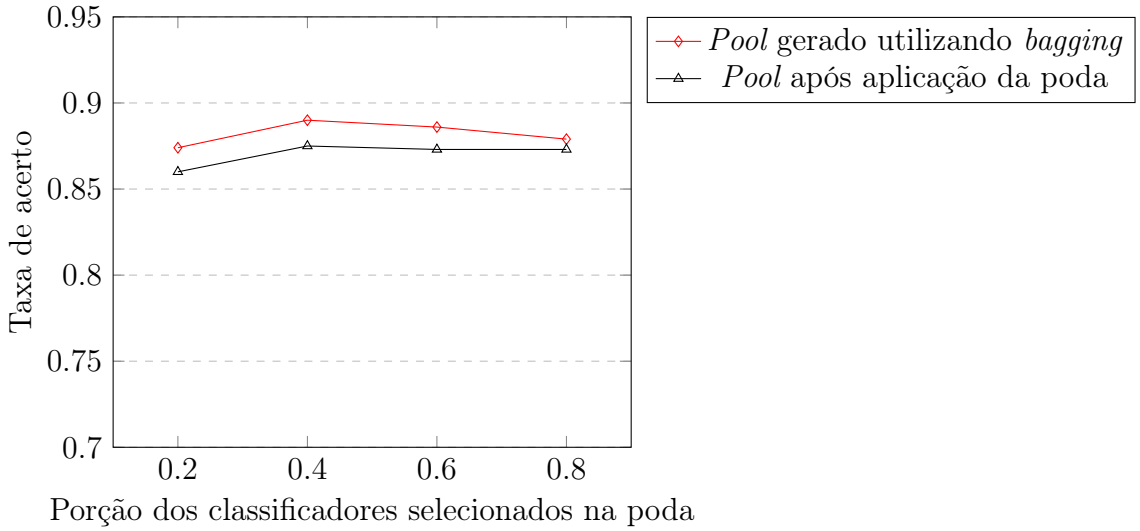

Figura 5: Comparação da taxa de acerto do *pool* com todos os classificadores e do *pool* após aplicação da poda, sendo o *pool* gerado utilizando *Bagging* com 100 classificadores.

5 Referências

1. https://archive.ics.uci.edu/ml/datasets.html

2. http://scikit-learn.org/stable/index.html

3. Forest type mapping Data Set - https://archive.ics.uci.edu/ml/datasets/Forest+type+mapping

4. Ludmila I. Kuncheva and Christopher J. Whitaker. 2003. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. Mach. Learn. 51, 2 (May 2003), 181-207.

5. L. I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms. John Wiley & Sons, 2014.

6. Z.-H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks:many could be better than all," Artificial intelligence, vol. 137, no. 1-2, pp. 239–263, 2002.

7. Z. Lu, X. Wu, X. Zhu, and J. Bongard, "Ensemble pruning via individual contribution ordering," in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 871–880, ACM, 2010.