Universidade Federal de Pernambuco

Sistemas de Múltiplos Classificadores

Relatório - Lista 4

Hartur Barreto Brito

Sumário

1 Introdução

Para implementação dos algoritmos de seleção dinâmica de classificadores 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*

1.1.2 Breast Cancer Wisconsin (Original) Data Set

A base de dados relacionada à câncer de mama, que pode ser encontrada em [4], foi selecionada por conter uma quantidade reduzida de atributos (10) e classes (2), e uma quantidade grande de exemplos (699), o que aumenta a chance dos classificadores terem uma alta taxa de acerto.

Foi verificado que 16 amostras dessa base não apresentam o 7º atributo (*Bare Nuclei*). Como a quantidade de dados não informados é muito inferior à quantidade de dados existentes, foi calculada a moda desse atributo, resultando na substituição de todos os valores não informados por "1".

Além disso, também foi retirado o primeiro atributo, por ser apenas um número identificador e não ter influência sobre a classificação.

A distribuição das amostras com relação à classe pode ser verificada na tabela 2.

Código	Classe	Amostras
02	Benign	458
04	Malignant	241

Tabela 2: Distribuição de exemplos para cada classe da base de câncer de mama

1.1.3 Skin Segmentation Data Set

Essa base pode ser encontrada em [5]. Ela possui originalmente 245057 amostras, sendo 50859 classificadas como "skin" e 194198 classificadas como "non-skin", como pode ser visualizado na tabela 3. Para diminuir o custo computacional da execução dos classificadores, foram selecionadas 600 amostras aleatoriamente, sendo 300 da classe "skin" e 300 da classe "non-skin", como pode ser visualizado na tabela 4. Apesar de diminuir drasticamente a quantidade de exemplos, por possuir poucas classes, não houve prejuízo na classificação.

Essa base foi selecionada principalmente pela sua quantidade limitada de atributos (4), para que fosse aplicada a expansão por função, como será demonstrado na seção 3.

A distribuição das amostras utilizadas com relação à classe pode ser verificada na tabela 4.

Código	Classe	Amostras
01	Skin	194198
02	Non-skin	194198

Tabela 3: Distribuição de exemplos original para cada classe da base de identificação de pele

Código	Classe	Amostras
01	Skin	300
02	Non-skin	300

Tabela 4: Distribuição de exemplos utilizada para cada classe da base de identificação de pele

1.2 Metodologia

Para obtenção de resultados, os algoritmos foram executados 10 vezes, sendo o resultado final a média aritmética dos resultados parciais. Os passos executados nessa lista são:

1. Mistura dos dados lidos da base;
2. Separação de dados de treino, de seleção dinâmica e de teste;
3. Geração do *ensemble*;
4. Previsão do *ensemble*;
5. Previsão dos algoritmos KNORAU, OLA e LCA utilizando as 4, 6 e 8 amostras mais próximas;
6. Cálculo da precisão do *ensemble*;
7. Cálculo da precisão dos algoritmos KNORAU, OLA e LCA utilizando as 4, 6 e 8 amostras mais próximas;

2 Questão 1

A aplicação da seleção dinâmica de classificadores tenta diminuir ou retirar a contribuição de um classificador fraco na região do padrão a ser previsto com o objetivo de melhorar a taxa de acerto.

Um dos problemas da utilização de seleção dinâmica de classificadores é o aumento do custo computacional da execução de uma classificação.

3 Questão 2

Em todos os testes, os algoritmos KNORAU, OLA e LCA obtiveram resultados melhores que a abordagem utilizando todos os classificadores gerados pelo *Bagging*. Esse fato acontece pois os algoritmos selecionam o(s) classificador(es) mais competente(s) na região em que a amostra de teste se localiza, evitando a utilização de classificadores fracos nas regiões.

Pode ser observado também que os algoritmos OLA e LCA obtiveram resultados muito semelhantes e superiores aos do KNORAU. Uma explicação para esse acontecimento, seria que o KNORAU não elimina os classificadores que são fracos em uma determinada região, e sim atribui um peso menor em sua resposta.

Abaixo podem ser observados os gráficos dos resultados obtidos a partir da execução da geração de um *pool* utilizando *Bagging* com 20, 40, 60, 80 e 100 classificadores SGD, com a seguinte separação dos dados: 36% para treino, 24% para seleção dinâmica e 40% para teste.

3.1 Forest Type Mapping Data Set

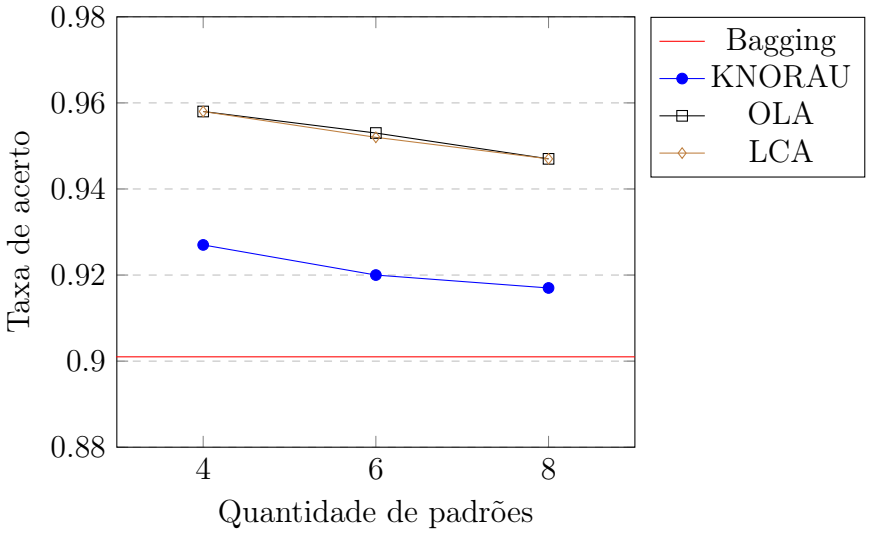

Figura 1: Taxa de acerto média de predições da base *Forest Types* por quantidade de padrões selecionados nos algoritmos KNORAU, OLA e LCA, utilizando *Bagging* com 20 classificadores SGD como gerador de *pool*.

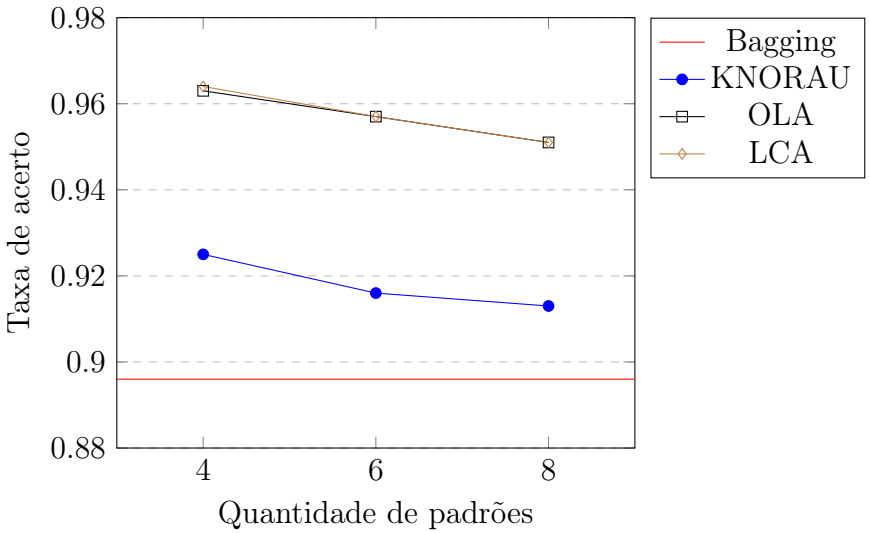

Figura 2: Taxa de acerto média de predições da base *Forest Types* por quantidade de padrões selecionados nos algoritmos KNORAU, OLA e LCA, utilizando *Bagging* com 40 classificadores SGD como gerador de *pool*.

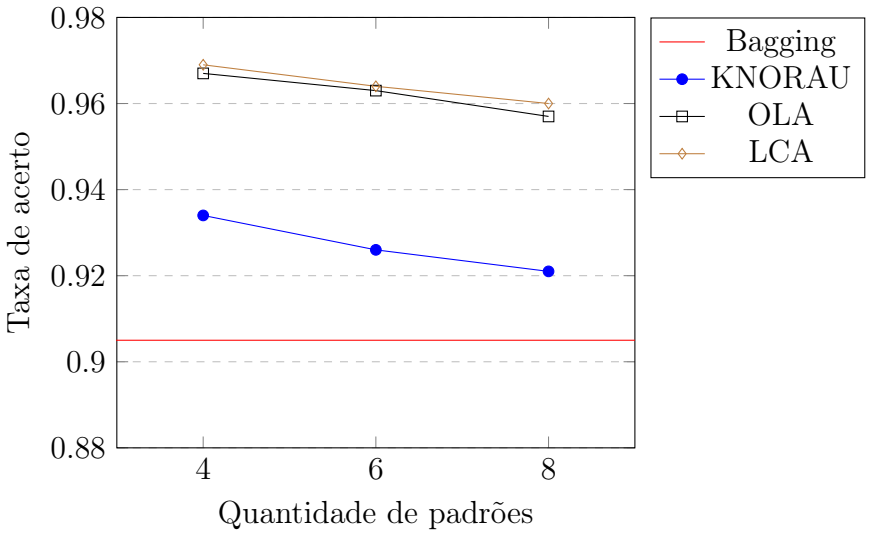

Figura 3: Taxa de acerto média de predições da base *Forest Types* por quantidade de padrões selecionados nos algoritmos KNORAU, OLA e LCA, utilizando *Bagging* com 60 classificadores SGD como gerador de *pool*.

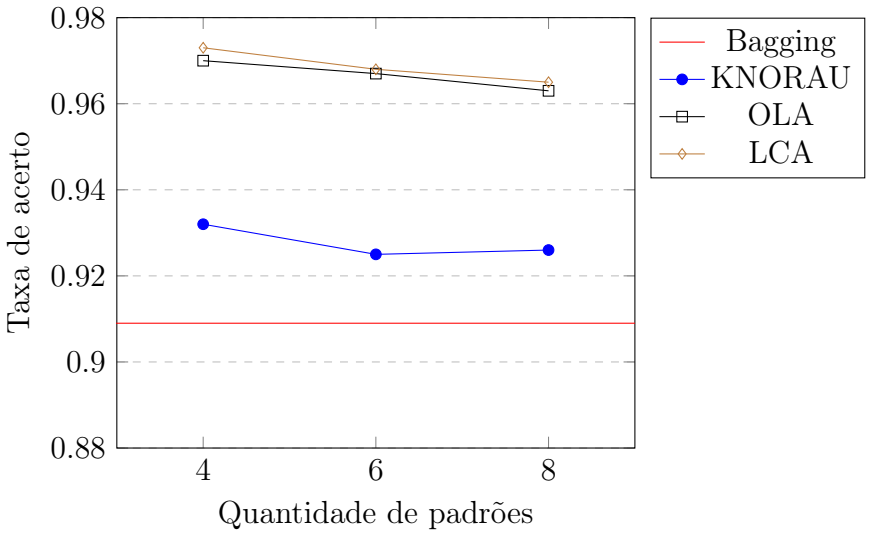

Figura 4: Taxa de acerto média de predições da base *Forest Types* por quantidade de padrões selecionados nos algoritmos KNORAU, OLA e LCA, utilizando *Bagging* com 80 classificadores SGD como gerador de *pool*.

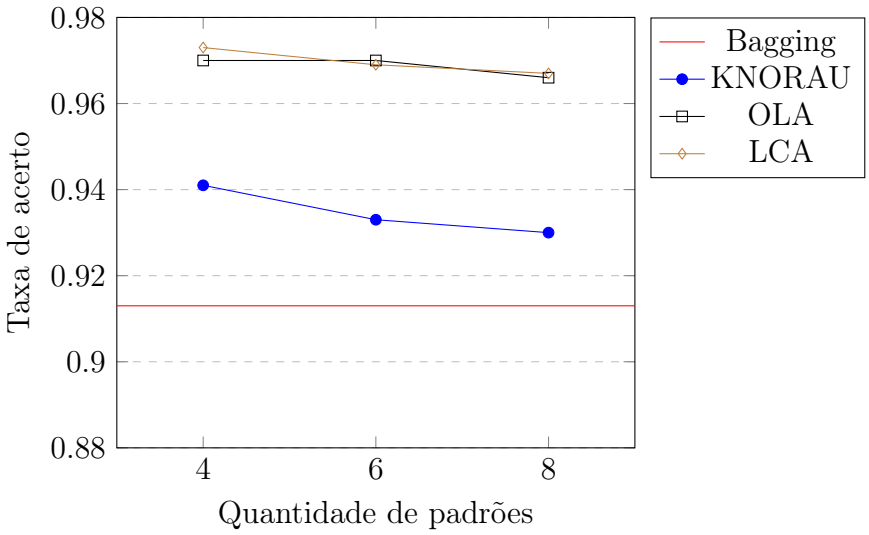

Figura 5: Taxa de acerto média de predições da base *Forest Types* por quantidade de padrões selecionados nos algoritmos KNORAU, OLA e LCA, utilizando *Bagging* com 100 classificadores SGD como gerador de *pool*.

3.2 Breast Cancer Wisconsin (Original) Data Set

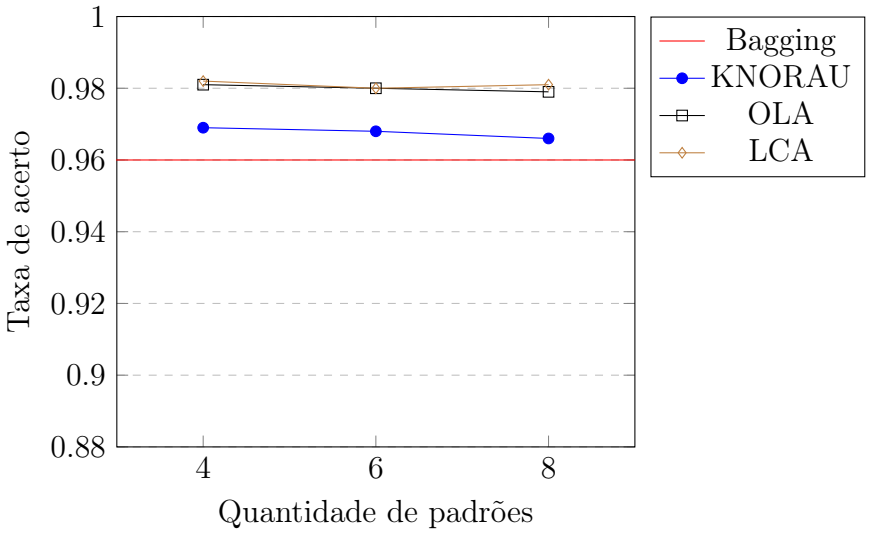

Figura 6: Taxa de acerto média de predições da base *Forest Types* por quantidade de padrões selecionados nos algoritmos KNORAU, OLA e LCA, utilizando *Bagging* com 20 classificadores SGD como gerador de *pool*.

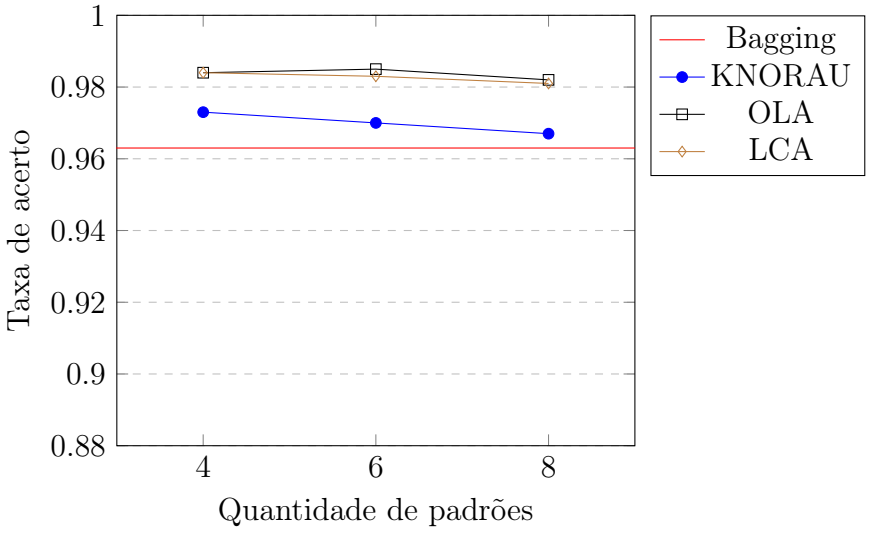

Figura 7: Taxa de acerto média de predições da base *Forest Types* por quantidade de padrões selecionados nos algoritmos KNORAU, OLA e LCA, utilizando *Bagging* com 40 classificadores SGD como gerador de *pool*.

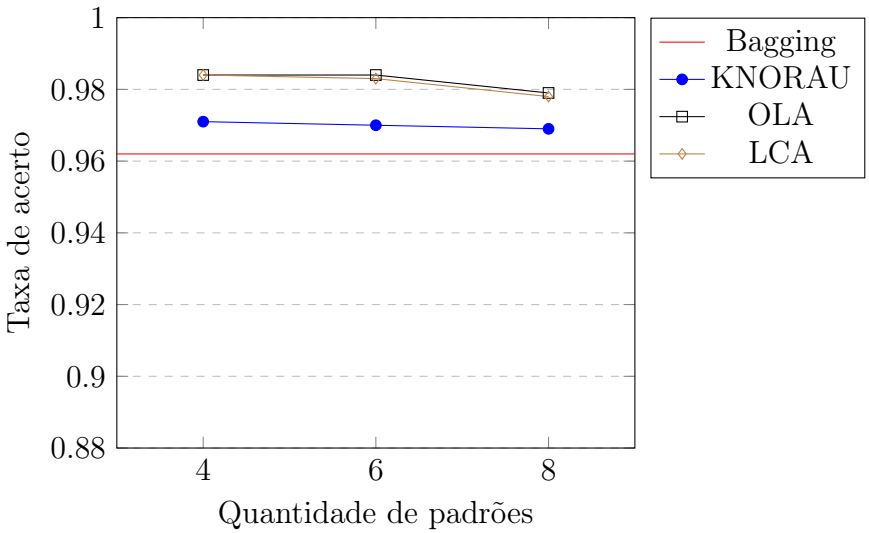

Figura 8: Taxa de acerto média de predições da base *Forest Types* por quantidade de padrões selecionados nos algoritmos KNORAU, OLA e LCA, utilizando *Bagging* com 60 classificadores SGD como gerador de *pool*.

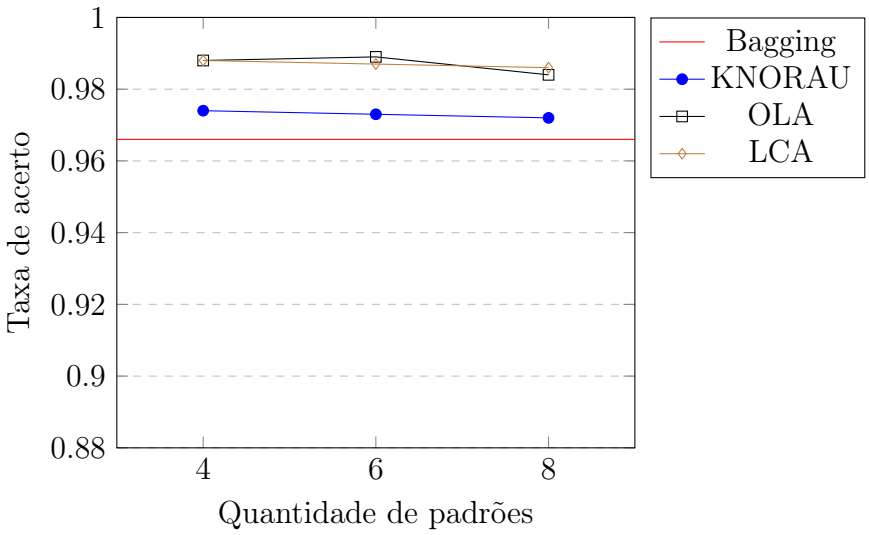

Figura 9: Taxa de acerto média de predições da base *Forest Types* por quantidade de padrões selecionados nos algoritmos KNORAU, OLA e LCA, utilizando *Bagging* com 80 classificadores SGD como gerador de *pool*.

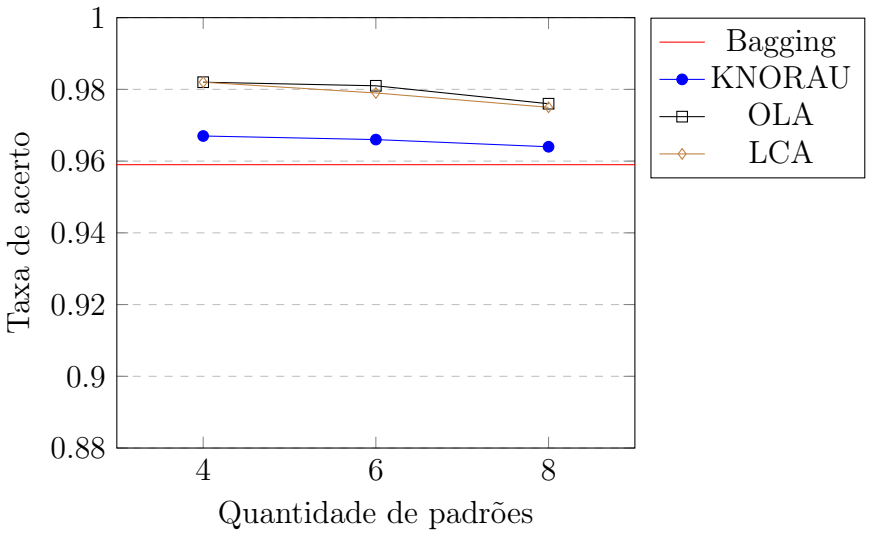

Figura 10: Taxa de acerto média de predições da base *Forest Types* por quantidade de padrões selecionados nos algoritmos KNORAU, OLA e LCA, utilizando *Bagging* com 100 classificadores SGD como gerador de *pool*.

3.3 Skin Segmentation Data Set

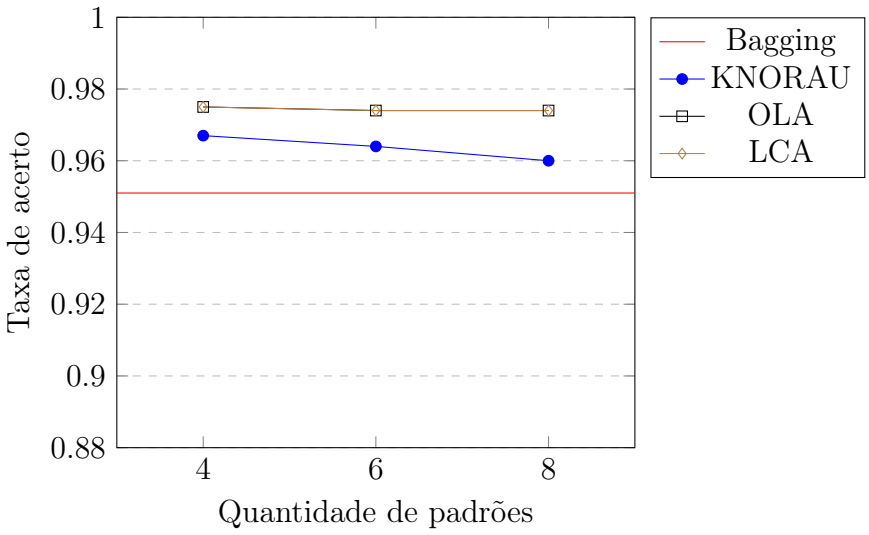

Figura 11: Taxa de acerto média de predições da base *Forest Types* por quantidade de padrões selecionados nos algoritmos KNORAU, OLA e LCA, utilizando *Bagging* com 20 classificadores SGD como gerador de *pool*.

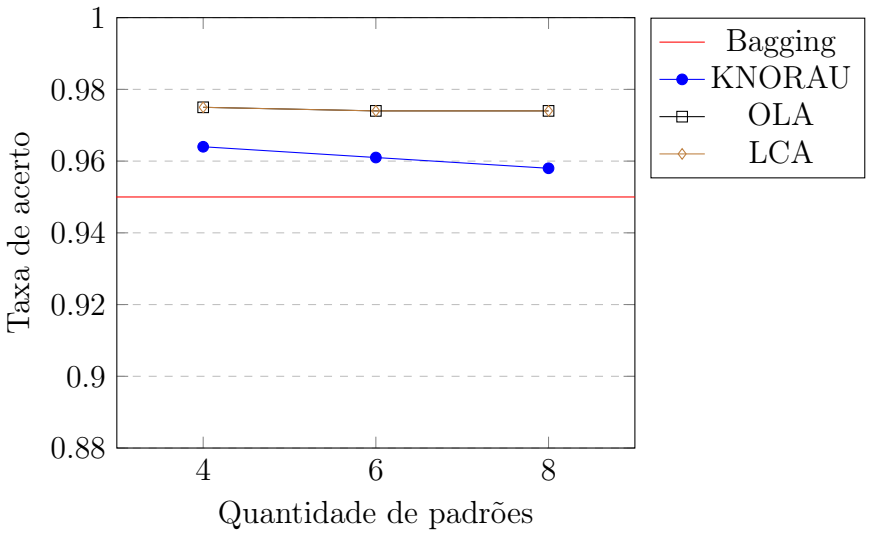

Figura 12: Taxa de acerto média de predições da base *Forest Types* por quantidade de padrões selecionados nos algoritmos KNORAU, OLA e LCA, utilizando *Bagging* com 40 classificadores SGD como gerador de *pool*.

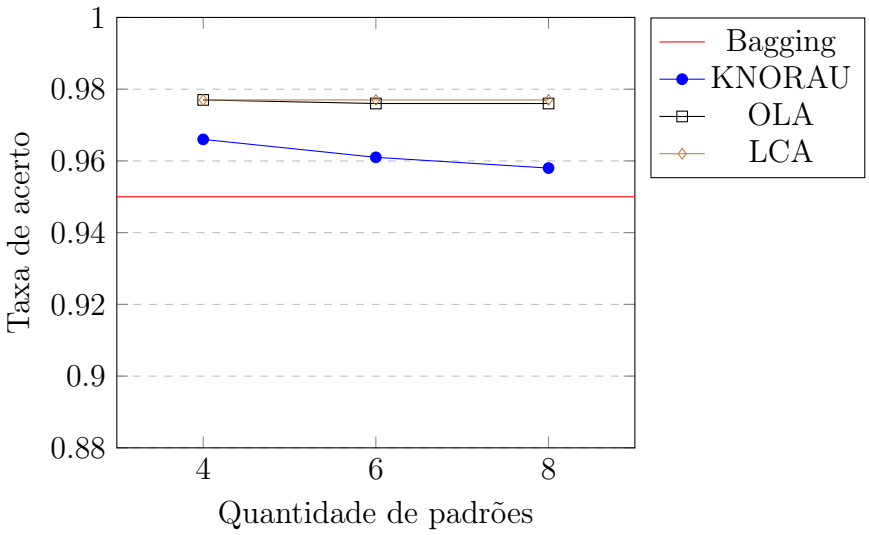

Figura 13: Taxa de acerto média de predições da base *Forest Types* por quantidade de padrões selecionados nos algoritmos KNORAU, OLA e LCA, utilizando *Bagging* com 60 classificadores SGD como gerador de *pool*.

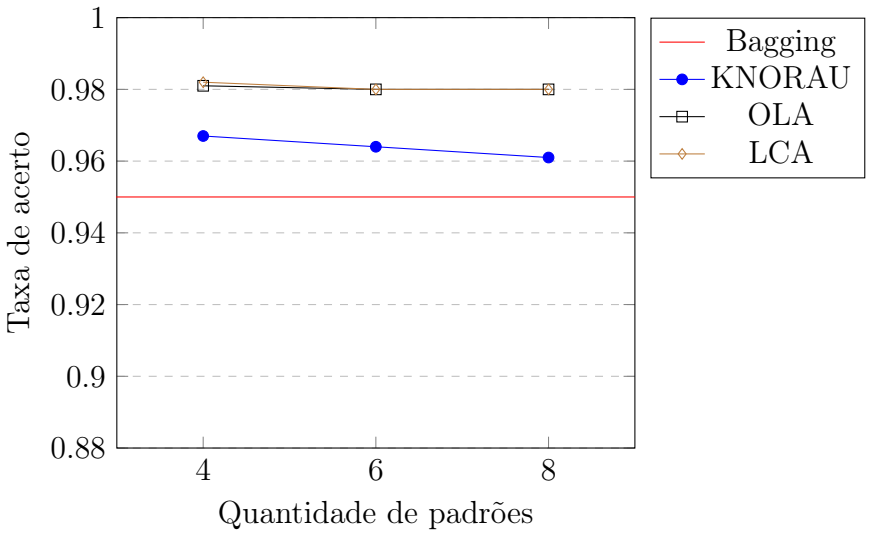

Figura 14: Taxa de acerto média de predições da base *Forest Types* por quantidade de padrões selecionados nos algoritmos KNORAU, OLA e LCA, utilizando *Bagging* com 80 classificadores SGD como gerador de *pool*.

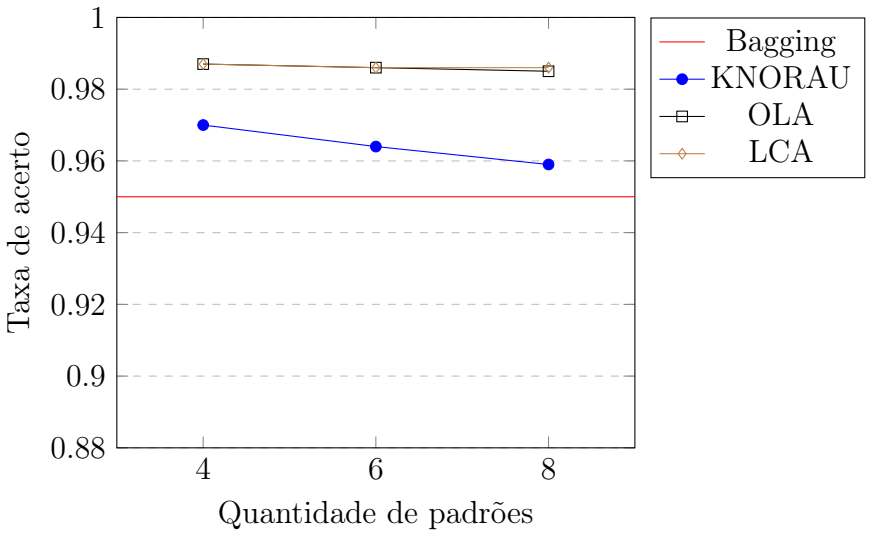

Figura 15: Taxa de acerto média de predições da base *Forest Types* por quantidade de padrões selecionados nos algoritmos KNORAU, OLA e LCA, utilizando *Bagging* com 100 classificadores SGD como gerador de *pool*.

4 Questão 3

Considero como "padrões fáceis" os padrões que possuem amostras utilizadas no treino mais próximas à ele de apenas uma classe ou que exista uma maioria de amostras de treino de uma das classes.

Já os "padrões difíceis", considero como sendo os que estão em uma região de indecisão, isto é, quando há um empate na quantidade de amostras de treino de duas ou mais classes. Além disso, considero como difíceis os padrões que estão muito distantes das amostras utilizadas durante o treino (distância maior do que a maior distância entre as amostras de treino).

A seleção dinâmica seria de grande ajuda quando se trata de amostras em regiões de indecisão (padrões "difíceis"), pois ele irá selecionar os classificadores mais adequados para classificar a amostra.

Quando se trata de um padrão "fácil", basta utilizar todo o *pool* de classificadores, pois o custo computacional é menor e, como o padrão é de fácil classificação, não haverá inconsistência entre os classificadores.

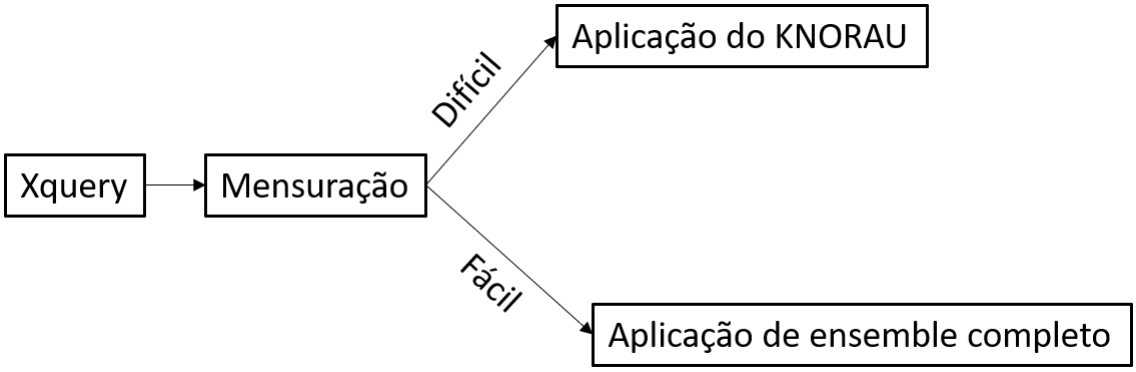

Figura 16: Arquitetura de classificação

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. Breast Cancer Wisconsin (Original) - https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29
5. Skin Segmentation Data Set - https://archive.ics.uci.edu/ml/datasets/Skin+Segmentation