Universidade Federal de Pernambuco

Sistemas de Múltiplos Classificadores

Relatório - Lista 1

Hartur Barreto Brito

Sumário

1 Introdução

Para a implementar os algoritmos necessários para a resolução das questões propostas pela primeira lista, foi utilizada a linguagem *Python*, usufruindo de sua biblioteca, a *Scikit-Learn* [2], que disponibiliza a implementação de diversos classificadores como, por exemplo, o Bayesiano, o *K-Nearest Neighbor* (KNN) e o *Stochastic Gradient Descent* (SGD), que foram os utilizados nas soluções.

Os classificadores utilizados foram escolhidos por sua simplicidade, uma vez que o foco da realização da lista é de analisar os conceitos discutidos em sala. São eles:

- Aplicação da técnica "*Bagging of Classifiers*"
- Aplicação da técnica "*Random Subspace*"
- Expansão de atributos das amostras

As bases de dados selecionadas foram: "*Breast Cancer Wisconsin (Original) Data Set*" [3] e "*Skin Segmentation Data Set* " [4]. Elas podem ser encontradas na coleção de bases UCI [1] aplicando os seguintes filtros de pesquisa:

- *Default task*: *Classification*;
- *Attribute type*: *Numerical*;
- *Data type*: *Multivariate*;
- *Format type*: *Matrix*.

Para realização de análise e ajustes nos dados disponibilizados pelas bases, como a procura e substituição de valores não informados, foi utilizado o *Excel*.

1.1 Análise das bases utilizadas

1.1.1 Breast Cancer Wisconsin (Original) Data Set

A base de dados relacionada à câncer de mama 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 1.

Código	Classe	Amostras
02	Benign	458
04	Malignant	241

Tabela 1: Distribuição de exemplos para cada classe da base de câncer de mama

1.1.2 Skin Segmentation Data Set

Essa base pode ser encontrada em [4]. Ela possui originalmente 245057 amostras, sendo 50859 classificadas como "skin" e 194198 classificadas como "non-skin", como pode ser visualizado na tabela 2. 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 3. 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 3.

Código	Classe	Amostras
01	Skin	194198
02	Non-skin	194198

Tabela 2: Distribuição de exemplos original para cada classe da base de identificação de pele

Código	Classe	Amostras
01	Skin	300
02	Non-skin	300

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

1.2 Metodologia

Nos experimentos, ao realizar a leitura dos dados das bases selecionadas, para cada vez que um classificador for ser avaliado (treinado e testado), as amostras lidas são trocadas de posição $N * 3$ vezes, sendo N a quantidade de amostras dessa base. Essa troca de posições é realizada a partir da obtenção de dois números aleatórios, para que os exemplos localizados nesses índices sejam trocados de posição.

A divisão de amostras que são utilizadas para treino e para teste é realizada da seguinte forma: São selecionados 60% dos exemplos das amostras de cada classe do problema apresentado para serem utilizadas no treino do classificador, e 40% das amostras dessa mesma classe para realização do teste desse classificador.

Para decidir a classificação do *ensemble*, foi utilizada a técnica de voto majoritário. Os valores de taxa de acerto apresentadas neste relatório foram resultados da média obtida a partir de 10 execuções da avaliação do classificador, conforme descrito acima.

2 Questão 1 - Bagging

Analisando os gráficos das figuras 1, 2, 3, 4 e 5 6 é possível observar quais classificadores se adaptaram melhor ao *Bagging*.

O classificador que obteve melhores resultados foi o SGD. A análise dos gráficos das figuras 3 e 6 mostra que, ao aplicar a técnica de *Bagging*, o classificador obteve uma taxa de acerto média maior do que a taxa utilizando apenas 1 SGD (apresentando todas as amostras de treinamento) na maioria dos casos, tendo um resultado inferior apenas na situação de 1 classificador no *Bagging*.

O resultado inferior ocorre devido ao fato de que, quando é alocado apenas 1 classificador no *Bagging*, seu treinamento será realizado com apenas um subconjunto das amostras separadas para treino, o que irá prejudicar sua avaliação. Quando são utilizados pelo menos 10 classificadores, as chances de existirem amostras de treino que não foram avaliadas pelos classificadores diminuem.

Apesar da característica de ter um resultado inferior quando utilizado apenas 1 classificador no *Bagging* ser a esperada, ela não foi observada nos classificadores KNN e SGD, como pode ser visto nos gráficos das figuras 1, 2, 4 e 5. Uma possível causa dessa característica seria o fato dos bancos terem poucas classes e serem simples de classificar, uma vez que a taxa de acerto, mesmo nas piores situações, estarem acima de 85% em todos os testes do *Bagging*.

Os resultados da execução do *Bagging* de classificadores Bayesianos, que podem ser observados nas figuras 1 e 4, obtiveram resultados melhores que o da execução de apenas 1 classificador Bayesiano utilizando todo o conjunto de treino, com exceção da execução com 20 classificadores, que teve uma taxa de acerto 0.001 menor.

Já o *Bagging* de KNN não obteve um bom resultado. Como pode ser visualizado no gráfico da imagem 2, nenhuma das combinação de KNN obteve um resultado melhor que 1 classificador KNN treinado com todas as amostras do conjunto de treino. O gráfico da figura 5 mostra que para a base de reconhecimento de pele, o *Bagging* de KNN também não conseguiu superar a implementação com apenas 1 KNN treinando com todas as amostras de treino, chegando, no máximo, a ficar com a mesma taxa de acerto (0.960).

As tabelas 4 e 5 demonstram que a aplicação dos classificadores Bayesiano, KNN e SGD separadamente já apresentam um resultado satisfatório na classificação das bases de dados. Para auxiliar na comparação de resultados entre a execução de apenas 1 classificador, sendo treinado por todo o conjunto de treino, e a execução de um *ensemble* de classificadores, aplicando a técnica *Bagging*, foram traçadas uma reta em vermelho nos gráficos das figuras 1 2 3 4 5 6.

2.1 Resultados da aplicação do Bagging na base Breast Cancer Wisconsin (Original)

Classificador	Média da precisão
Bayes	0.949
KNN	0.983
SGD	0.949

Tabela 4: Taxa de acerto média por classificador

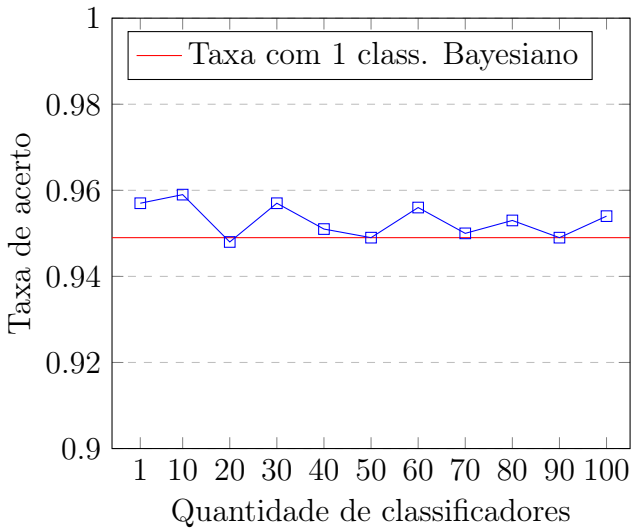

Figura 1: Taxa de acerto média de predições da base *Breast Cancer Wisconsin (Original)* por quantidade de classificadores utilizando *Bagging* com classificadores Bayesianos.

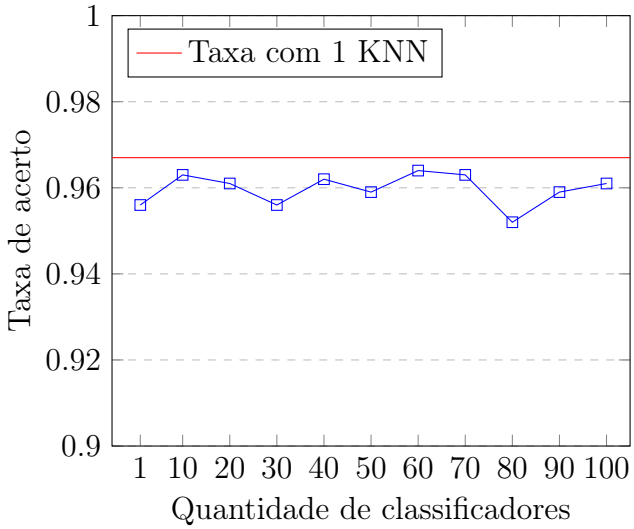

Figura 2: Taxa de acerto média de predições da base *Breast Cancer Wisconsin (Original)* por quantidade de classificadores utilizando *Bagging* com classificadores KNN.

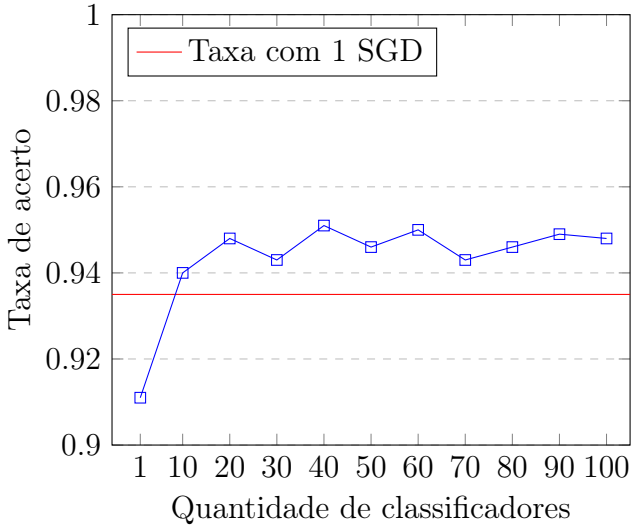

Figura 3: Taxa de acerto média de predições da base *Breast Cancer Wisconsin (Original)* por quantidade de classificadores utilizando *Bagging* com classificadores SGD.

2.2 Resultados da aplicação do Bagging na base Skin Segmentation

Classificador	Média da precisão
Bayes	0.888
KNN	0.960
SGD	0.900

Tabela 5: Taxa de acerto média por classificador

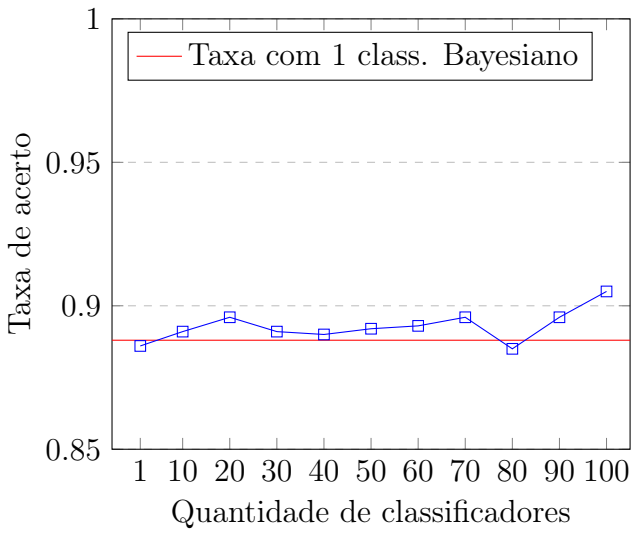

Figura 4: Taxa de acerto média de predições da base *Skin Segmentation* por quantidade de classificadores utilizando *Bagging* com classificadores Bayesianos.

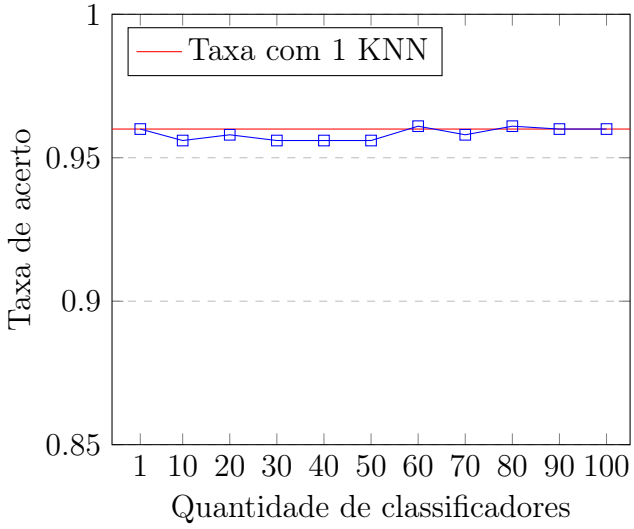

Figura 5: Taxa de acerto média de predições da base *Skin Segmentation* por quantidade de classificadores utilizando *Bagging* com classificadores KNN.

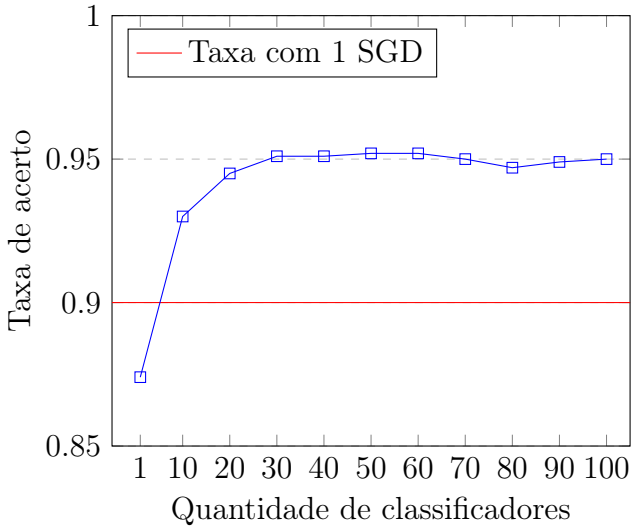

Figura 6: Taxa de acerto média de predições da base *Skin Segmentation* por quantidade de classificadores utilizando *Bagging* com classificadores SGD.

3 Questão 2 - Random Subspace

O resultado da expansão utilizando *Power Series* no banco *Breast Cancer Wisconsin (Original)* e da expansão *Trigonometric* no banco *Skin Segmentation* em conjunto com a aplicação da técnica de *Random Subspace* não obteve resultados satisfatórios pois, a maioria dos *ensembles* resultaram em uma taxa de acerto menor do que a taxa utilizando apenas 1 classificador, sendo ele treinado com todo o conjunto de treino.

As funções de expansão da família *Power Series* utilizadas foram:

- $F_3(x) = x^2$
- $F_4(x) = x^3$

As funções de expansão da família *Trigonometric* utilizadas foram:

- $F_2 = cos(\pi x)$
- $F_3 = sin(\pi x)$
- $F_4 = cos(2\pi x)$
- $F_5 = sin(2\pi x)$
- $F_6 = cos(3\pi x)$
- $F_7 = sin(3\pi x)$

O classificador que apresentou os melhores resultados com relação à taxa de acerto na classificação das amostras das bases *Skin Segmentation* e *Breast Cancer Wisconsin (Original)* nos experimentos da seção 2 foi o KNN.

Já com relação ao tempo para treino e geração de resultados, o classificador SGD foi o mais rápido, como pode ser visto nas tabelas 6 e 7.

Classificador	Tempo de execução
Bayes	325 segundos
KNN	422 segundos
SGD	53 segundos

Tabela 6: Tempo para gerar e testar todos os *pools* de classificadores (com 1, 10, 20, 30, ..., 100 classificadores) gerados com a técnica *Bagging* para o banco *Breast Cancer Wisconsin (Original)*

Classificador	Tempo de execução
Bayes	99 segundos
KNN	322 segundos
SGD	43 segundos

Tabela 7: Tempo para gerar e testar todos os *pools* de classificadores (com 1, 10, 20, 30, ..., 100 classificadores) gerados com a técnica *Bagging* para o banco *Skin Segmentation*

Ao contrário dos experimentos do *Bagging*, os gráficos apresentados nas figuras das sessões 3.1 e 3.2 apresentaram um crescimento gradativo da taxa de acerto conforme é aumentado a quantidade de classificadores no *ensemble*, com exceção dos gráficos das figuras 7 e 9.

Com relação à quantidade de atributos utilizados (30%, 40% e 50%), os classificadores não apresentaram grandes variações. Por exemplo, a análise dos gráficos das figuras 7, 8 e 9, mostra que a taxa de acerto de grande parte dos cenários estão entre 0.94 e 0.96.

Como nos gráficos do experimento da seção 2, foram traçadas retas em vermelho nos gráficos das seções 3.1 e 3.2 que representam a taxa de acerto quando é utilizado apenas 1 classificador treinado com todo o conjunto de treino para facilitar a comparação dos resultados.

3.1 Resultados da aplicação do Random Subspace na base Breast Cancer Wisconsin (Original)

3.1.1 Classificadores Bayesianos

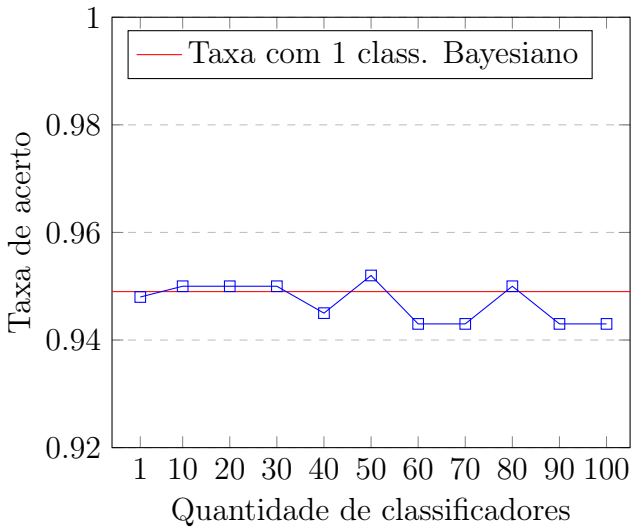

Figura 7: Taxa de acerto média utilizando *Random Subspace* aplicando a função de expansão de atributos *Power Series* com classificadores Bayesianos utilizando 30% dos atributos da base *Breast Cancer Wisconsin (Original)*

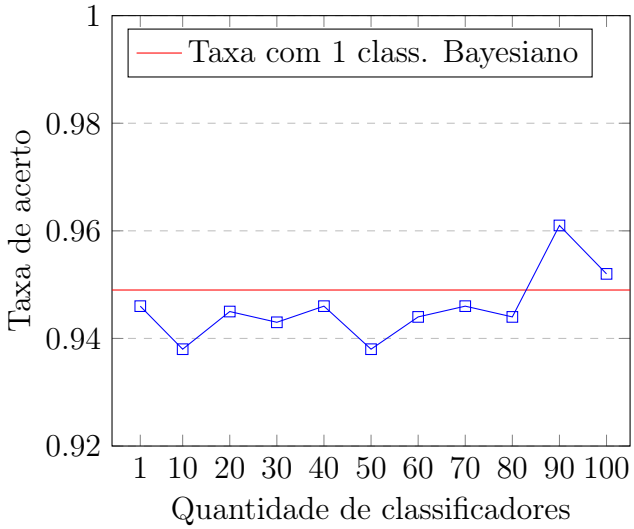

Figura 8: Taxa de acerto média utilizando *Random Subspace* aplicando a função de expansão de atributos *Power Series* com classificadores Bayesianos utilizando 40% dos atributos da base *Breast Cancer Wisconsin (Original)*

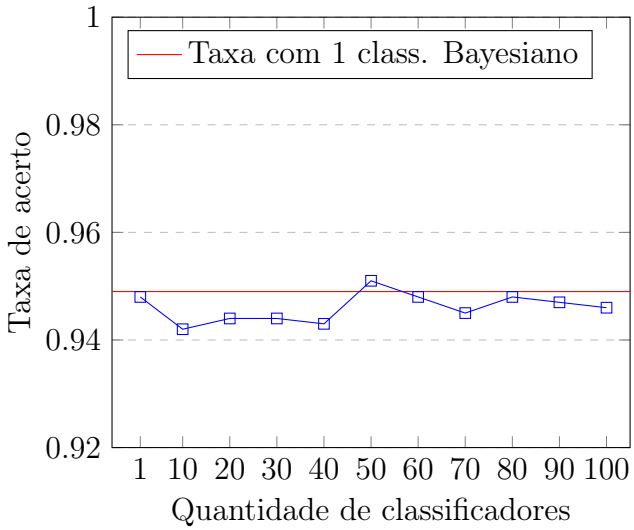

Figura 9: Taxa de acerto média utilizando *Random Subspace* aplicando a função de expansão de atributos *Power Series* com classificadores Bayesianos utilizando 50% dos atributos da base *Breast Cancer Wisconsin (Original)*

3.1.2 Classificadores KNN

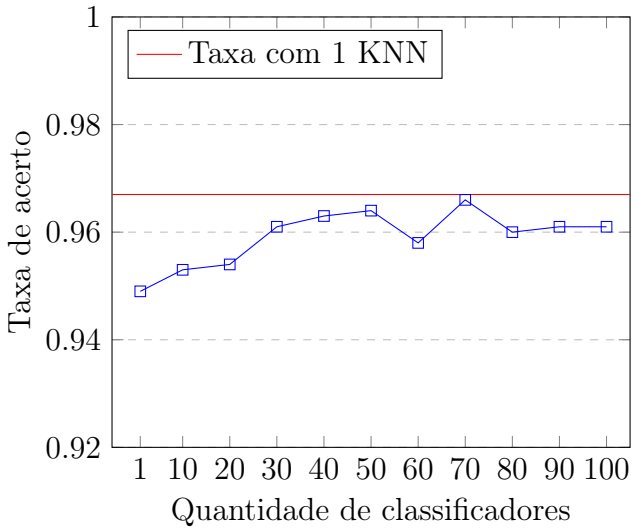

Figura 10: Taxa de acerto média utilizando *Random Subspace* aplicando a função de expansão de atributos *Power Series* com classificadores KNN utilizando 30% dos atributos da base *Breast Cancer Wisconsin (Original)*

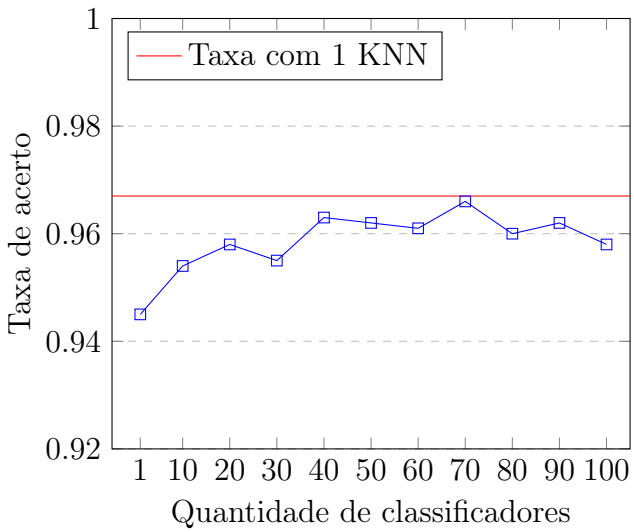

Figura 11: Taxa de acerto média utilizando *Random Subspace* aplicando a função de expansão de atributos *Power Series* com classificadores KNN utilizando 40% dos atributos da base *Breast Cancer Wisconsin (Original)*

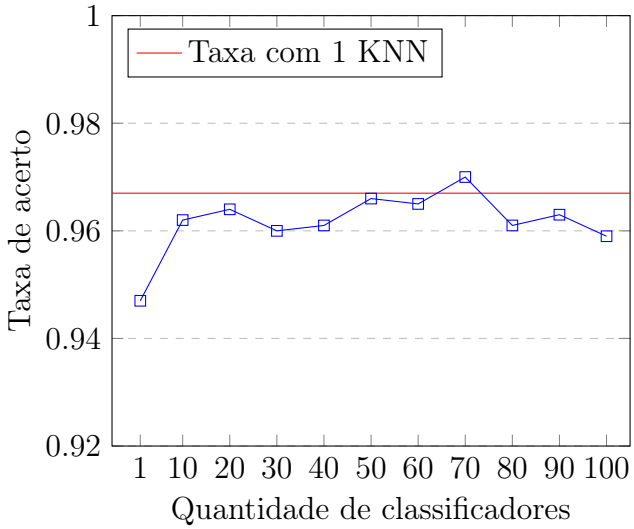

Figura 12: Taxa de acerto média utilizando *Random Subspace* aplicando a função de expansão de atributos *Power Series* com classificadores KNN utilizando 50% dos atributos da base *Breast Cancer Wisconsin (Original)*

3.1.3 Classificadores SGD

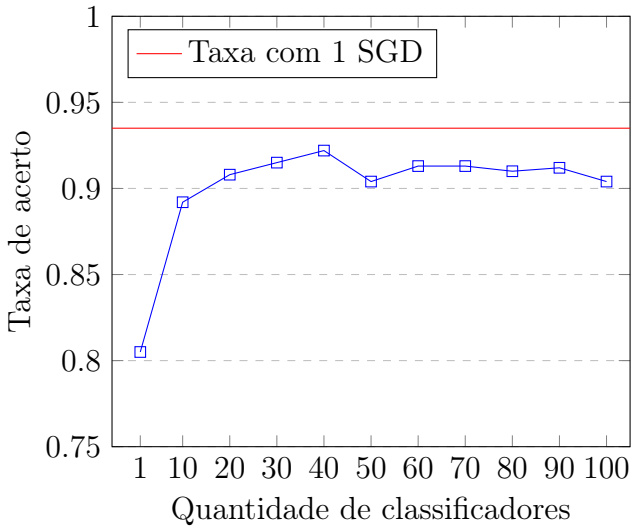

Figura 13: Taxa de acerto média utilizando *Random Subspace* aplicando a função de expansão de atributos *Power Series* com classificadores SGD utilizando 30% dos atributos da base *Breast Cancer Wisconsin (Original)*

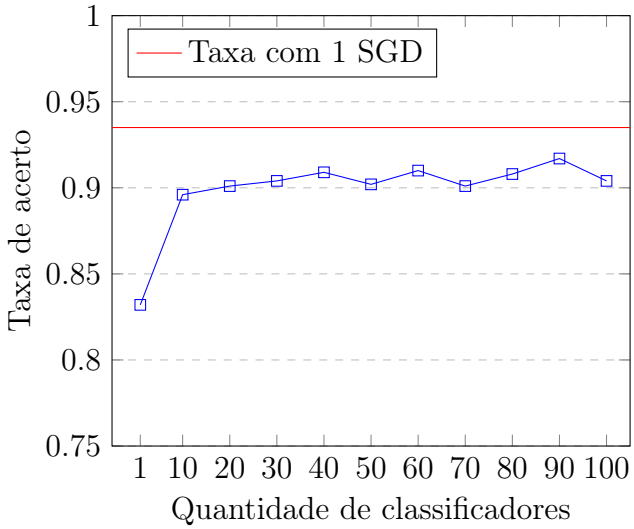

Figura 14: Taxa de acerto média utilizando *Random Subspace* aplicando a função de expansão de atributos *Power Series* com classificadores SGD utilizando 40% dos atributos da base *Breast Cancer Wisconsin (Original)*

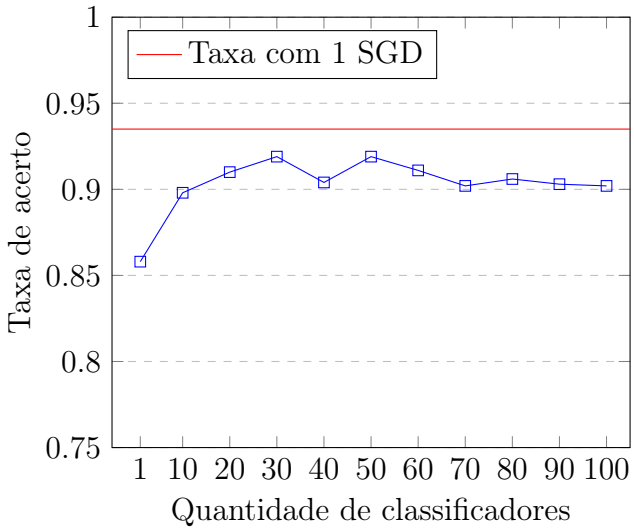

Figura 15: Taxa de acerto média utilizando *Random Subspace* aplicando a função de expansão de atributos *Power Series* com classificadores SGD utilizando 50% dos atributos da base *Breast Cancer Wisconsin (Original)*

3.2 Resultados da aplicação do Random Subspace na base Skin Segmentation

3.2.1 Classificadores Bayesianos

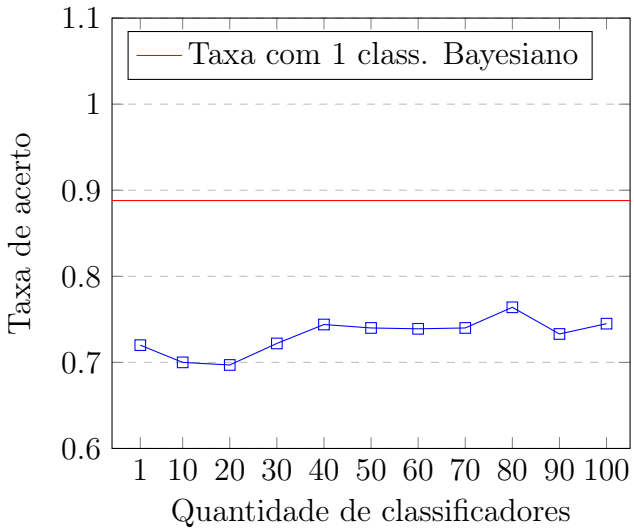

Figura 16: Taxa de acerto média utilizando *Random Subspace* aplicando a função de expansão de atributos *Trigonometric* com classificadores Bayesianos utilizando 30% dos atributos da base *Skin Segmentation*

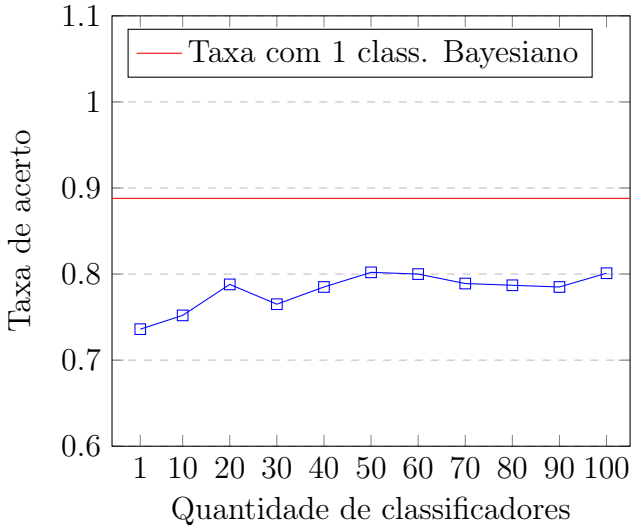

Figura 17: Taxa de acerto média utilizando *Random Subspace* aplicando a função de expansão de atributos *Trigonometric* com classificadores Bayesianos utilizando 40% dos atributos da base *Skin Segmentation*

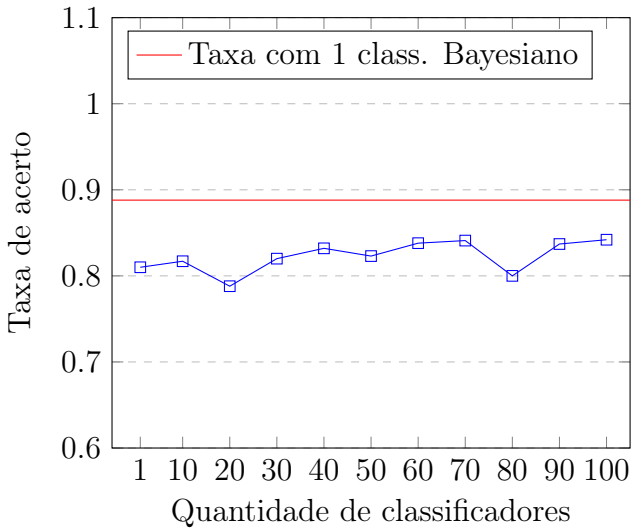

Figura 18: Taxa de acerto média utilizando *Random Subspace* aplicando a função de expansão de atributos *Trigonometric* com classificadores Bayesianos utilizando 50% dos atributos da base *Skin Segmentation*

3.2.2 Classificadores KNN

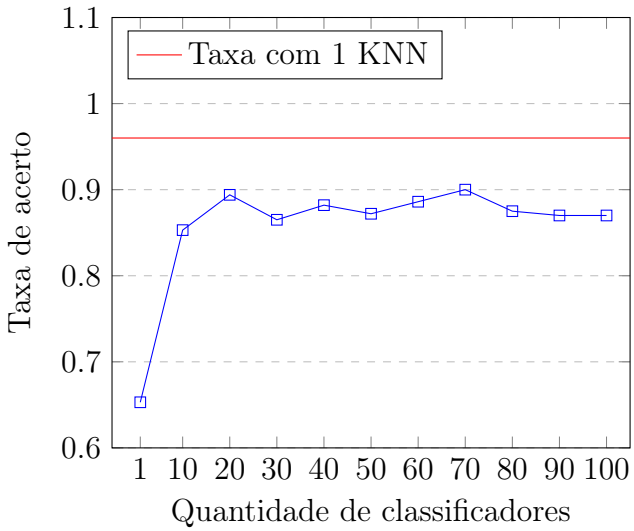

Figura 19: Taxa de acerto média utilizando *Random Subspace* aplicando a função de expansão de atributos *Trigonometric* com classificadores KNN utilizando 30% dos atributos da base *Skin Segmentation*

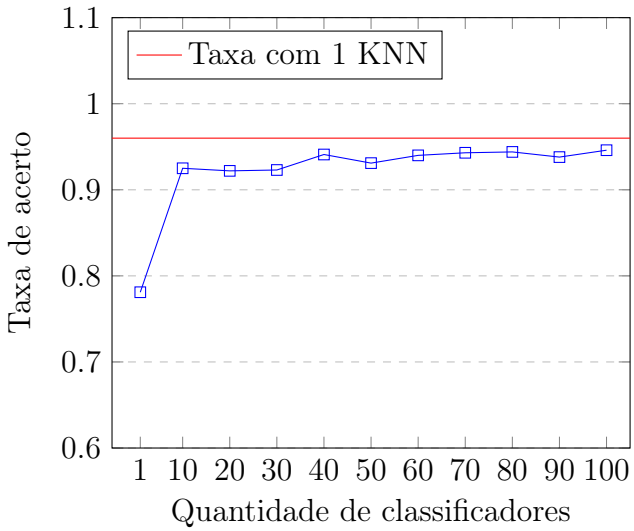

Figura 20: Taxa de acerto média utilizando *Random Subspace* aplicando a função de expansão de atributos *Trigonometric* com classificadores KNN utilizando 40% dos atributos da base *Skin Segmentation*

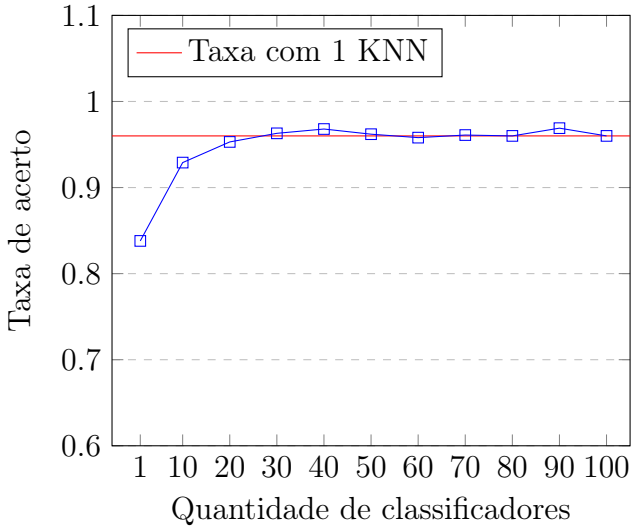

Figura 21: Taxa de acerto média utilizando *Random Subspace* aplicando a função de expansão de atributos *Trigonometric* com classificadores KNN utilizando 50% dos atributos da base *Skin Segmentation*

3.2.3 Classificadores SGD

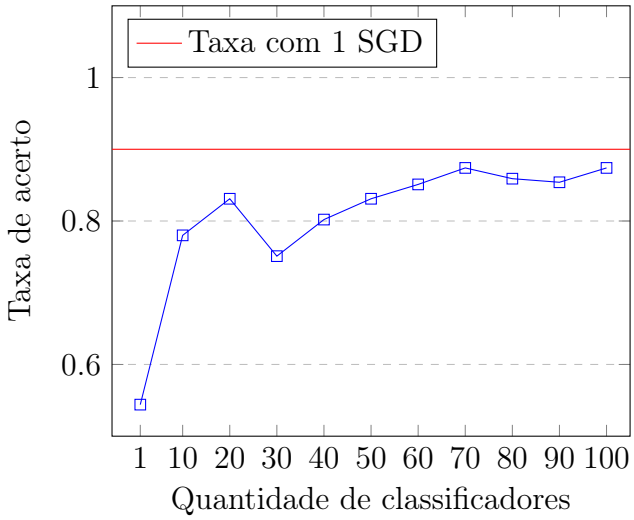

Figura 22: Taxa de acerto média utilizando *Random Subspace* aplicando a função de expansão de atributos *Trigonometric* com classificadores SGD utilizando 30% dos atributos da base *Skin Segmentation*

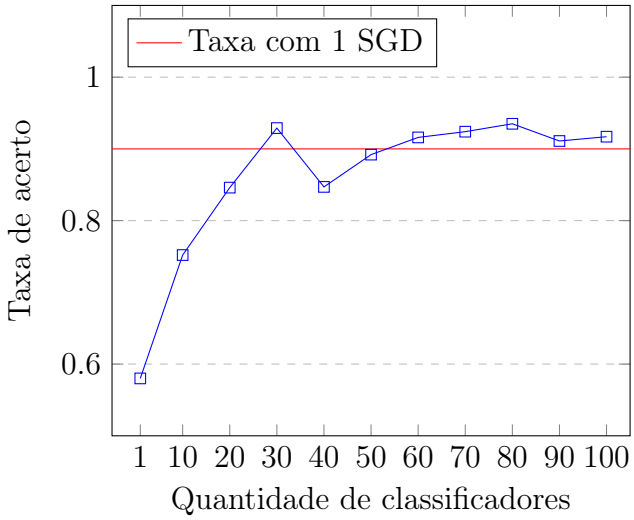

Figura 23: Taxa de acerto média utilizando *Random Subspace* aplicando a função de expansão de atributos *Trigonometric* com classificadores SGD utilizando 40% dos atributos da base *Skin Segmentation*

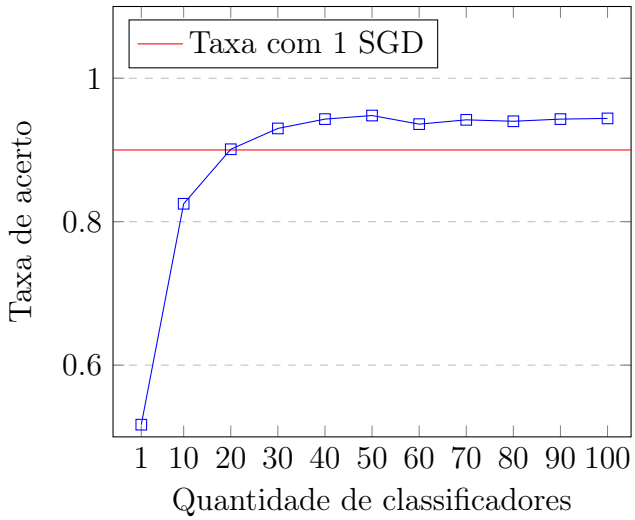

Figura 24: Taxa de acerto média utilizando *Random Subspace* aplicando a função de expansão de atributos *Trigonometric* com classificadores SGD utilizando 50% dos atributos da base *Skin Segmentation*

4 Questão 3

Um sistema de múltiplos classificadores (SMC) pode ser dividido em 3 fases, são elas: Geração de Classificadores, Seleção de Classificadores e Integração de Classificadores [7]. Durante a primeira fase, são gerados (treinados) os classificadores (*pool* de classificadores) que serão utilizados. Alguns exemplos de técnicas aplicadas durante essa fase, são: *Bagging of Classifiers* e *Random Subspace*.

A fase de seleção, que é a mais estudada dentre as 3 fases, é responsável por selecionar quais classificadores participarão do *ensemble* na classificação da amostra apresentada [7].

Durante a integração, os classificadores são combinados para que seja obtida uma resposta final. Exemplos de técnicas de integração são: voto majoritário, *tresholded voting*, *Borda count* e consenso unânime [8].

A fase de seleção pode ser dividida em 2 vertentes: *Static Ensemble Selection* e a *Dynamic Ensemble Selection* (DES) [6]. A abordagem estática consiste na aplicação da seleção dos classificadores durante a fase de treinamento do algoritmo, ou seja, a decisão da quando serão utilizados quais classificadores é realizada durante a fase de treinamento. Exemplos de *Static Ensemble Selection* são os algoritmos: *Decision Forests* e *Boosting* [6].

Já a abordagem dinâmica (DES), consiste em realizar a seleção de classificadores durante a apresentação das amostras. A maioria das técnicas de seleção são baseadas em estimativa de precisão local em pequenas regiões do *feature space* localizado ao redor do exemplo a ser avaliado. Essa região é chamada de região de competência [5].

A região de competência é amplamente utilizada nas soluções devido à sua capacidade de comparação dos resultados já apresentados durante o treinamento e as novas amostras a serem selecionadas, o que possibilita a inferência de quais classificadores seriam os mais indicados para prever o resultado dessa amostra.

Pesquisas mostram que a abordagens de seleção dinâmica de classificadores alcançam resultados melhores que as abordagens estáticas [5]. Entretanto, o maior desafio do DES é o de como definir um critério de medida do nível de competência da base de classificadores [5].

Um exemplo de seleção dinâmica é a utilização de *Meta-Learning* para selecionar quais classificadores melhores se encaixam em cada amostra apresentada. Como descrito em [5], essa técnica manuseia diferentes propriedades dos classificadores como, por exemplo, a precisão em uma determinada região, são extraídas do conjunto de treino e transformadas em *meta-features*. Com isso, ao ser apresentado um novo exemplo, as *meta-features* são extraídas baseando-se no conjunto de treino, e é utilizado um meta-classificador para selecionar os classificadores mais competentes nessa região de competência.

Uma abordagem simples de encontrar a quantidade de classificadores para prever o resultado de uma amostra, seria utilizando força bruta, i.e., realizar todas as combinações possíveis de classificadores, até encontrar a que melhor atende. Entretanto, essa abordagem é computacionalmente cara 7.

Para solucionar o problema do custo computacional, uma possibilidade é executar uma quantidade pré-definida de combinações aleatórias, limitando a quantidade de execuções realizadas, para que, a partir disso, seja selecionada a melhor combinação

dentre as testadas.

Com o intuito de tornar possível a utilização de força bruta na definição de quantos classificadores serão integrados, foi apresentada uma proposta em [7], dentro da abordagem de *Meta-Learning*, que pré-seleciona alguns classificadores para que seja aplicada a força bruta. Para selecionar esses classificadores, eles são avaliados utilizando *meta-regression*.

Entre as fases de seleção e integração, existe uma fase que serve para diminuir a quantidade de classificadores que farão parte do *ensemble* com o intuito de melhorar a predição. Essa fase se chama "Poda de classificadores".

Expandindo o teorema de "no free lunch" à definição de qual seria a melhor técnica de seleção de quantidade de classificadores, conclui-se que não é possível definir uma técnica que seja a mais eficiente para todos os tipos de bases. Entretanto, a técnica que chegaria mais próxima de tal generalização, seria a de força bruta, mas ela exige um custo computacional muito alto. Por outro lado, a abordagem proposta em [7] se torna bem promissora por tentar tornar a utilização da força bruta possível, aplicando-a a apenas um subconjunto de classificadores, que foram selecionados previamente utilizado *meta-regression*.

5 Referências

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

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

3. Breast Cancer Wisconsin (Original) - https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

4. Skin Segmentation Data Set - https://archive.ics.uci.edu/ml/datasets/Skin+Segmentation

5. CRUZ, Rafael M. O.; SABOURIN, Robert; CAVALCANTI, George D. C.. META-DES.H: A Dynamic Ensemble Selection technique using meta-learning and a dynamic weighting approach. 2015 International Joint Conference On Neural Networks (ijcnn), p.1-8, jul. 2015.

6. Rafael M.O. Cruz, Robert Sabourin, George D.C. Cavalcanti, META-DES.Oracle: Meta-learning and feature selection for dynamic ensemble selection, Information Fusion, Volume 38, 2017, Pages 84-103, ISSN 1566-2535

7. Anandarup Roy, Rafael M.O. Cruz, Robert Sabourin, George D.C. Cavalcanti. Meta-learning recommendation of default size of classifier pool for META-DES, Neurocomputing, Volume 216, pp. 351-362, 2016.

8. K. Woods , W.P. Kegelmeyer Jr. , K. Bowyer , Combination of multiple classifiers using local accuracy estimates, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 405–410 .