IDENTIFICAÇÃO DE VARIEDADES DE TRIGO UTILIZANDO IMPLEMENTAÇÕES BASEADAS NO MÉTODO FEATORA

Chrystian Felipe Freitas (UEPG) E-mail: chrystian.felipe.freitas@gmail.com
Guilherme Lúdio Torquato de Souza (UEPG) E-mail: guilhermeludio@gmail.com
Leila Maria Vriesmann (UEPG) E-mail: lmvriesmann@uepg.br

Resumo: Variedades de trigo podem ser identificadas utilizando atributos de seus grãos. Para a identificação, pode-se utilizar o método FEATORA. O FEATORA é um método de Seleção Dinâmica de Subconjunto de Classificadores que faz uso de intervalos de valores de atributos contínuos e da probabilidade de reconhecimento de cada classificador. O presente trabalho propôs um novo esquema e um novo método baseados no método FEATORA. No novo esquema, foi utilizado o voto majoritário ponderado de esquemas do FEATORA. Já no novo método, os classificadores selecionados passaram por uma etapa adicional antes de ser fornecida a classe para a instância de teste. Os experimentos foram realizados. Atingiu-se uma taxa de reconhecimento mais próxima do valor do oráculo.

Palavras-chave: classificação, Seleção Dinâmica de Conjunto de Classificadores, grãos de trigo.

IDENTIFICATION OF WHEAT VARIETIES USING IMPLEMENTATIONS BASED ON THE FEATORA METHOD

Abstract: Wheat varieties can be identified using attributes of their grains. For identification, the FEATORA method can be used. FEATORA is a Dynamic Ensemble Selection method that makes use of ranges of continuous attribute values and the recognition probability of each classifier. The present work proposed a new scheme and a new method based on the FEATORA method. In the new scheme, the weighted majority vote of FEATORA schemes was used. In the new method, the selected classifiers went through an additional step before providing the class for the test instance. The experiments were performed. A recognition rate closer to the oracle value was achieved.

Keywords: classification, Dynamic Ensemble Selection of Classifiers, wheat grains.

1. Introdução

Variedades de trigo podem ser identificadas (ou classificadas) utilizando atributos dos grãos. É possível realizar esse processo de classificação computacionalmente. Há classificadores que podem ser adequados para um tipo de instância de grão de trigo, e inadequados para outros.

Quando são utilizados subconjuntos de classificadores, consegue-se uma diversidade de classificadores e há a combinação (ou a fusão) de votos. Assim, é possível extrair melhores taxas de reconhecimento do que com o uso de classificadores individuais em alguns experimentos.

Os classificadores considerados mais aptos para cada uma das instâncias são utilizados na Seleção Dinâmica de Subconjunto de Classificadores, ou Seleção Dinâmica de Conjunto de Classificadores (KO et al., 2008). O objetivo é tentar atingir o valor do oráculo, que ocorre quando as instâncias somente não são reconhecidas se não houver classificador no conjunto inicial que as reconheçam. O valor do oráculo pode ser utilizado como referência para comparar o quão bom o resultado desse tipo de classificação se apresenta, como pode ser observado em Vriesmann et al. (2012).

Um método de Seleção Dinâmica de Subconjunto de Classificadores é o FEATORA proposto por Vriesmann & Britto Jr. (2015). O método FEATORA utiliza, como critério para a seleção

de classificadores, as probabilidades de reconhecimento associadas aos intervalos de valores de atributos contínuos aos quais a instância pertence.

A identificação de variedades de trigo com o método FEATORA foi tratada em Vriesmann & Britto Jr. (2015), em Martins Neto et al. (2019) e em Padilha & Talignani (2019). Em Vriesmann & Britto Jr. (2015), foi proposto o método FEATORA. Já em Martins Neto et al. (2019), foram modificados cálculos estatísticos que têm relação com a seleção dos classificadores. Em Padilha & Talignani (2019), diferentes fórmulas para a determinação do número de intervalos de valores de atributos contínuos foram testadas. O valor do oráculo não foi obtido nesses trabalhos, que utilizaram a mesma base de dados.

O objetivo deste trabalho é realizar implementações baseadas no método FEATORA, de forma que a taxa de reconhecimento na classificação de variedades de trigo fique mais próxima do valor do oráculo. Para tanto, um novo esquema e um novo método serão propostos. Quanto ao novo esquema, votos de esquemas do FEATORA serão utilizados de forma ponderada. Quanto ao novo método, os classificadores pré-selecionados pelo método FEATORA, antes de atribuírem a classe final à instância, passarão por uma etapa adicional.

Detalhes sobre a base de dados utilizada, sobre os classificadores do conjunto inicial, sobre o método FEATORA, sobre o esquema e o método propostos podem ser encontrados na Seção 2. Já na Seção 3 são colocados os resultados obtidos pelo método FEATORA e pelos esquema e método propostos. Por fim, a Seção 4 conclui o trabalho e inclui sugestões de experimentos futuros.

2. Material e métodos

2.1. Base de dados e classificadores do conjunto inicial

A base de dados de grãos de trigo será a *seed* dataset, de Charytanowicz et al. (2010), que possui 3 classes distintas: *Canadian, Kama* e *Rosa*. Há 210 instâncias (70 de cada classe). Cada instância tem 7 atributos previsores: área, perímetro, compacidade, comprimento, largura, coeficiente de assimetria e comprimento do sulco do grão.

Os dados estarão normalizados. A ferramenta Weka (HALL et al., 2009), na versão 3.6.4, será usada. Com intuito de comparação de resultados entre o método FEATORA original e as implementações propostas no presente trabalho, serão utilizadas as mesmas definições de Vriesmann & Britto Jr. (2015), onde houve a divisão em 3 bases diferentes (*seed A*, *seed B* e *seed C*), cada uma com 70 instâncias, na base de grãos de trigo.

Dez classificadores *1*-NN (AHA et al., 1991) serão treinados e farão parte do conjunto inicial de classificadores. A diversidade dos classificadores será realizada por Subespaços Aleatórios (HO, 1998), com 5 de 7 atributos previsores.

Três execuções serão realizadas, altenando as bases de dados consideradas de treinamento, de validação e de teste entre *seed A*, *seed B* e *seed C*. Na primeira execução, o treinamento dos classificadores ocorrerá na base de dados *seed A*, o método de Seleção Dinâmica de Subconjunto utilizará cálculos de probabilidades obtidos dos dados da *seed B* (que também serão considerados dados de validação), e a base *seed C* será usada para teste. Na segunda execução, haverá o treinamento na base *seed C*, o método de Seleção Dinâmica de Subconjunto utilizará cálculos de probabilidades dos dados de *seed A*, e o teste será na base *seed B*. Já na terceira execução, *seed B* será utilizada para treinamento dos classificadores, enquanto que *seed C* para os cálculos de probabilidades para o método de Seleção Dinâmica de Subconjunto e *seed A* para teste. Será calculada a média e o desvio padrão das taxas de reconhecimento obtidas na base de teste nas 3 execuções.

2.2. Método FEATORA

O método FEATORA (VRIESMANN & BRITTO JR., 2015) cria uma tabela, para cada um dos atributos previsores da base de dados de validação, contendo seus intervalos de valores. Depois, cada um desses intervalos é associado com probabilidades de reconhecimento de cada um dos classificadores presentes no conjunto inicial.

Na fase de teste, com base no valor de cada atributo da instância, é feita a seleção dos classificadores que possuem uma determinada probabilidade mínima de reconhecimento. Esses classificadores ficarão em um subconjunto e, com base em um esquema, classificarão a instância.

Quatorze diferentes esquemas foram criados em Vriesmann & Britto Jr. (2015). Cada esquema possui, como parâmetro de entrada, um limiar superior. Alguns também podem utilizar um limiar inferior. Os nomes dos esquemas podem conter a letra W (quando o voto é ponderado pela probabilidade de reconhecimento dos classificadores) ou U (quando o voto é ponderado pela quantidade de seleções do classificador). Quando não possui nenhuma dessas letras em seu nome, o voto tem peso 1. De acordo com Martins Neto et al. (2019), complementando com Vriesmann & Britto Jr. (2015), os esquemas são:

- FEATORA-ELIMINATE e FEATORA-ELIMINATE-W: Os classificadores selecionados têm probabilidade de reconhecimento maior ou igual ao limiar superior em todos os atributos (ou no maior número de atributos, caso não houver classificador que tenha essa característica em todos os atributos);

- FEATORA-UNION, FEATORA-UNION-W e FEATORA-UNION-U: Os classificadores selecionados têm probabilidade de reconhecimento maior ou igual ao limiar superior, mesmo que em apenas um atributo;

- FEATORA-UNION-ELIMINATE, FEATORA-UNION-ELIMINATE-W e FEATORA-UNION-ELIMINATE-U: Os classificadores selecionados são obtidos pelo FEATORA-ELIMINATE e pelo FEATORA-UNION;

- FEATORA-UNIONL, FEATORA-UNIONL-W e FEATORAUNIONL-U: Os classificadores selecionados têm probabilidade de reconhecimento maior ou igual ao limiar superior (mesmo que em apenas um atributo), e não têm probabilidade menor que o limiar inferior;

- FEATORA-ELIMINATEL e FEATORA-ELIMINATEL-W: Os classificadores selecionados têm probabilidade de reconhecimento maior ou igual ao limiar superior (em todos os atributos ou no maior número de atributos), e não têm probabilidade menor que o limiar inferior;

- FEATORA-MAJ: A classe de cada instância é obtida pelo voto majoritário dos 13 esquemas anteriores.

2.3. Esquema e método propostos

2.3.1. Esquema FEATORA-MAJ-P

O FEATORA-MAJ-P (FEATORA com voto majoritário ponderado) foi criado como um novo esquema para o método FEATORA. Cada um dos 13 primeiros esquemas (Seção 2.2) propostos por Vriesmann & Britto Jr. (2015) fornece um voto ponderado. Esse peso é a soma

das probabilidades de reconhecimento dos classificadores do subconjunto que fornece a classe. A instância de teste recebe a classe que tem o maior somatório de probabilidades.

2.3.2. Método FEATORAGROUP_error

No método FEATORAGROUP_error existirá uma etapa adicional para avaliar a aptidão dos classificadores selecionados em cada um dos esquemas do FEATORA de Vriesmann & Britto Jr. (2015). Isso ocorrerá antes do fornecimento do voto final.

Os classificadores do subconjunto serão agrupados de acordo com o seu voto na instância de teste. A base de dados de grãos de trigo tem 3 classes distintas e, consequentemente, os classificadores serão divididos em 3 grupos. Os classificadores que votarem para a classe *Canadian* serão colocados no grupo 1, os classificadores que votarem para a classe *Kama* serão colocados no grupo 2 e os classificadores que votarem para a classe *Rosa* serão colocados no grupo 3.

O método utilizará o que é chamado na Estatística de teorema da multiplicação, de eventos sucessivos e de eventos independentes, e é definido quando o fato de acontecer um evento não altera a probabilidade de um outro evento ocorrer, de acordo com Iezzi et al. (2001). Nesse caso, o voto que um classificador do grupo fornece para uma instância é independente do voto de outro classificador. Por exemplo, o cálculo da probabilidade de erro, para um grupo que tem 2 classificadores, será dado por:

$$P(A \cap B) = p(A).\ p(B) \quad (1)$$

onde p(A) e p(B) serão as probabilidades de erro de reconhecimento dos classificadores A e B, respectivamente; e P(A∩B) será a probabilidade final obtida com a multiplicação.

No método FEATORAGROUP_error, em cada grupo, será calculada a probabilidade final com a multiplicação das probabilidades de erro de todos os seus classificadores. Escolhe-se o grupo com a menor probabilidade dos seus classificadores errarem a classe da instância de teste.

Um exemplo de funcionamento do método FEATORAGROUP_error pode ser explicado com base nas Figuras 1 e 2. Na Figura 1, foram pré-selecionados 7 classificadores (entre 10 classificadores disponíveis no conjunto inicial), sendo que 4 votaram na classe *Canadian*, 2 na classe *Kama* e 1 na classe *Rosa*. Em cada grupo, é mostrada a probabilidade de cada classificador errar as instâncias (obtém-se esse valor subtraindo-se a probabilidade de reconhecimento, de 100%), bem como os pesos (os quais estão apresentados na lateral superior direita de cada classificador).

Depois do cálculo da probabilidade de erro de cada classificador, faz-se a multiplicação em cada um dos grupos. Em esquemas fazem uso de pesos, o erro será elevado pelo peso. Por exemplo, se um classificador tiver peso 1, a probabilidade de erro será elevada a 1; e se tiver peso 2, a probabilidade de erro será elevada a 2 e assim sucessivamente.

O resultado (Figura 2) é a probabilidade de todo o grupo de classificadores errar. Quanto menor for esse valor, maior será a probabilidade dos classificadores reconhecerem a instância. Assim, escolhe-se a classe do grupo com o menor valor (classe *Canadian*, na Figura 2).

No método FEATORAGROUP_error, quanto mais classificadores existirem no grupo, menor será o resultado obtido com a multiplicação de erros. Considerando que o valor de erro de cada classificador estará no intervalo entre 0 e 1, a multiplicação desses erros tenderá a zero na medida em que aumentar a quantidade de classificadores no grupo. Da mesma maneira, os esquemas que trabalham com pesos terão o resultado da multiplicação de erros tendendo a

zero na medida em que os pesos forem maiores.

A probabilidade de erro do grupo será utilizada também como peso no esquema FEATORAGROUP_error-MAJ-P. Assim sendo, diferentemente do FEATORA-MAJ-P (Seção 2.3.1), será escolhida a classe com menor peso total.

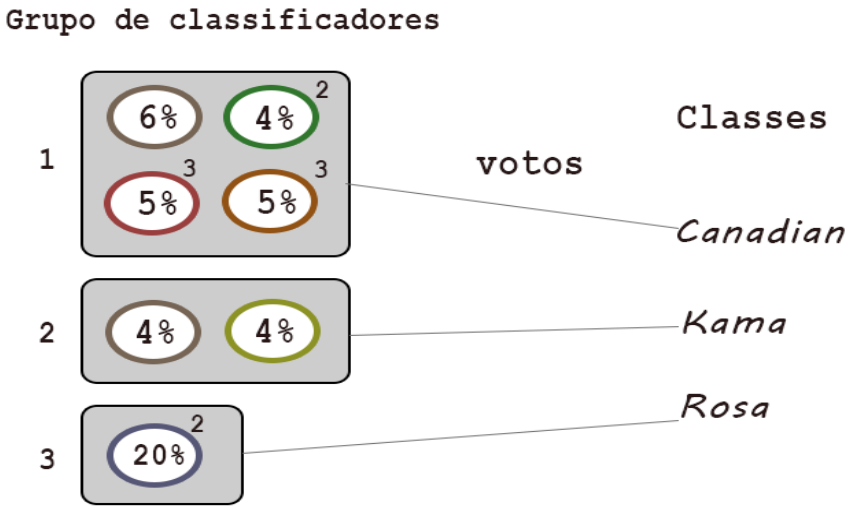

Figura 1. Probabilidade de erro dos classificadores

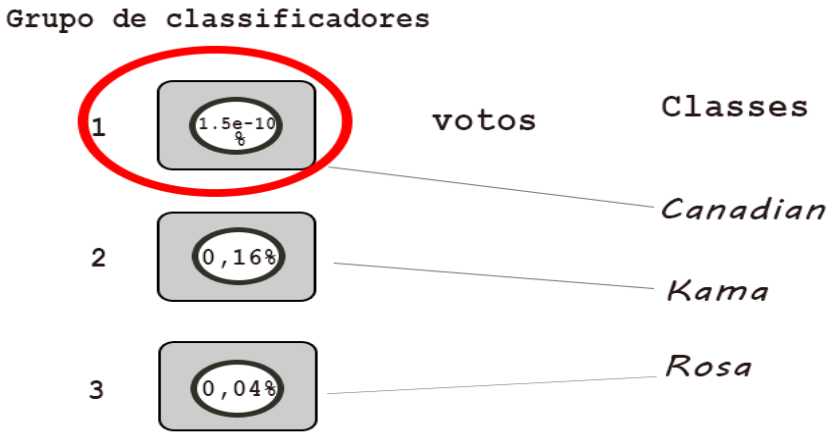

Figura 2. Probabilidade de erro de cada grupo

3. Resultados e discussão

O método e o esquema propostos foram implementados e testados. Os valores dos parâmetros de entrada foram idênticos aos utilizados em Vriesmann & Britto Jr. (2015). O valor inicial para o limiar inferior foi de 50% e o valor inicial para o limiar superior foi de 55%. Houve a adição de 5% nesses valores a cada nova execução, até chegarem a 95%. Também ocorreram execuções com o valor de 99% nesses limiares.

A maior taxa de reconhecimento que seria possível alcançar, chamada de oráculo, foi de 96,67%, com desvio padrão de 0,82, para as 3 execuções com conjunto inicial de 10 classificadores KNN (com K = 1) utilizando Subespaços Aleatórios, de acordo com Vriesmann & Britto Jr. (2015).

A Tabela 1 expõe as melhores taxas de reconhecimento (TR) e os valores de desvio padrão (DP) obtidos pelo método FEATORA em Vriesmann & Britto Jr. (2015), e pelo esquema FEATORA-MAJ-P. Também são colocados os valores de limiar inferior (LI) e de limiar superior (LS). Quando mais de um limiar atingisse a mesma taxa de reconhecimento, o desvio padrão era exposto como intervalo de valores. O FEATORA-UNIONL-U obteve a menor TR (91,90%), sendo o único esquema com resultado inferior ao FEATORA-MAJ-P (92,38%) proposto nesse trabalho. O esquema FEATORA-MAJ continuou com a maior taxa de reconhecimento (93,81%).

Tabela 1 – Melhores taxas de reconhecimento obtidas nos esquemas do método FEATORA e no esquema FEATORA-MAJ-P

Esquema	TR (%)	DP	LI	LS
FEATORA-ELIMINATE	93,33	0,82	-	99
FEATORA-ELIMINATE-W	93,33	0,82	-	99
FEATORA-UNION	92,86	1,43	-	99
FEATORA-UNION-W	92,38	0,82-2,18	-	85
FEATORA-UNION-U	92,38	0,82	-	85
FEATORA-UNION-ELIMINATE	92,86	1,43	-	99
FEATORA-UNION-ELIMINATE-W	92,38	0,82-2,18	-	99
FEATORA-UNION-ELIMINATE-U	92,38	0,82-2,18	-	85
FEATORA-UNIONL	92,38	2,18	50	90
FEATORA-UNIONL-W	92,38	2,18	70	70
FEATORA-UNIONL-U	91,9	2,18	75	80
FEATORA-ELIMINATEL	92,86	0	50	90
FEATORA-ELIMINATEL-W	92,86	0	50	90
FEATORA-MAJ	93,81	0,82	90	90
FEATORA-MAJ-P	92,38	0,82	50	70

Fonte: Os dados dos 14 primeiros esquemas foram obtidos de Vriesmann & Britto Jr. (2015)

A Tabela 2 apresenta os resultados do método FEATORAGROUP_error. Observa-se que as taxas de reconhecimento de cada esquema do FEATORAGROUP_error foram inferiores ou iguais às obtidas pelo esquema correspondente do FEATORA (Tabela 1), com exceção do esquema FEATORAGROUP_error-UNION-ELIMINATE-U (com 92,86%) e do esquema FEATORAGROUP_error-MAJ-P (com 95,24%). Esses dois últimos esquemas atingiram melhores taxas do que as obtidas na Tabela 1.

Tabela 2 – Melhores taxas de reconhecimento obtidas nos esquemas do método FEATORAGROUP_error

Esquema	TR (%)	DP	LI	LS
FEATORAGROUP_error-ELIMINATE	92,86	0,82	-	90
FEATORAGROUP_error-ELIMINATE-W	92,86	0,82	-	90
FEATORAGROUP_error-UNION	91,9	1,43	-	55
FEATORAGROUP_error-UNION-W	92,38	0,82-2,18	-	90
FEATORAGROUP_error-UNION-U	92,38	0,82	-	90
FEATORAGROUP_error-UNION-ELIMINATE	91,9	1,43	-	55
FEATORAGROUP_error-UNION-ELIMINATE-W	92,38	0,82-2,18	-	90
FEATORAGROUP_error-UNION-ELIMINATE-U	92,86	0,82-2,18	-	95
FEATORAGROUP_error-UNIONL	91,9	2,18	55	70
FEATORAGROUP_error-UNIONL-W	91,9	2,18	50	90
FEATORAGROUP_error-UNIONL-U	91,9	2,18	50	90
FEATORAGROUP_error-ELIMINATEL	92,38	0	50	90
FEATORAGROUP_error-ELIMINATEL-W	92,38	0	50	90
FEATORAGROUP_error-MAJ	93,81	0,82	50	90
FEATORAGROUP_error-MAJ-P	95,24	1,64	55	90

O esquema FEATORAGROUP_error-MAJ-P (Tabela 2) atingiu a maior taxa de reconhecimento dentre os esquemas e os métodos apresentados, ficando mais próximo do valor do oráculo. Esse resultado era esperado, pois cada um dos esquemas escolhe o grupo com a menor probabilidade de erro, e o peso do esquema fornece uma estimativa de erro do voto. Essa estimativa foi obtida com base na probabilidade de reconhecimento dos classificadores que efetivamente votaram na classe. Assim sendo, o esquema FEATORAGROUP_error-MAJ-P foi o mais adequado para a identificação de variedades de trigo da base de dados.

4. Conclusão

O presente trabalho tratou de implementações baseadas no método FEATORA para identificação de variedades de grãos de trigo. Primeiramente, propôs-se um novo esquema (FEATORA-MAJ-P). Depois, criou-se um novo método (denominado FEATORAGROUP_error) com uma etapa adicional, após a seleção do subconjunto de classificadores do FEATORA original. Nessa etapa adicional, os classificadores selecionados eram agrupados de acordo com a classe que votaram. Depois, foi obtida a probabilidade de erro de cada grupo. Essa probabilidade era calculada como eventos sucessivos e independentes de probabilidades de erro de seus classificadores. A classe final atribuída à instância de teste constituía-se da classe do grupo com a menor probabilidade de erro. Todos os esquemas do método FEATORA foram adaptados para o método FEATORAGROUP_error.

O valor do oráculo (96,67%) não foi atingido na identificação de variedades de trigo. Porém, o esquema FEATORAGROUP_error-MAJ-P obteve a maior taxa de reconhecimento (95,24%) entre todas as implementações baseadas no FEATORA.

Portanto, a etapa adicional implementada, juntamente com o voto ponderado de 13 esquemas, permitiu alcançar um valor mais próximo do oráculo para a base de dados de variedades de grãos de trigo do que os obtidos pelo método FEATORA original.

Como trabalhos futuros, sugerem-se experimentos com outras quantidades e outros tipos de classificadores no conjunto inicial, e também com outras bases de dados.

Referências

AHA, D. W.; KIBLER, D. & ALBERT, M. K. *Instance-based Learning Algorithms*. Machine Learning. Vol. 6, n.1, p.37-66, 1991.

CHARYTANOWICZ, M.; NIEWCZAS, J.; KULCZYCKI, P.; KOWALSKI, P. A.; ŁUKASIK, S. & ZAK, S. *Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images.* In Information technologies in biomedicine, Springer. p.15-24, 2010.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P. & WITTEN, I. H. *The Weka Data Mining Software: an update.* ACM SIGKDD explorations newsletter. Vol. 11, n.1, p.10-18, 2009.

HO, T. K. *The Random Space Method for Constructing Decision Forests.* IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 20, n.8, p.832-844, 1998.

IEZZI, G. et al. *Matemática - Ciências e Aplicações.* Atual, 2001.

KO, A. H.; SABOURIN, R. & BRITTO JR., A. S. *From Dynamic Classifier Selection to Dynamic Ensemble Selection.* Pattern Recognition. Vol. 41, n.5, p.1718-1731, 2008.

MARTINS NETO, A.; ROSAS, G. & VRIESMANN, L. M. *Melhoria na Probabilidade de Reconhecimento do Método FEATORA em Relação à Base de Dados de Grão de Trigo*. In II - Workshop Em Sistemas De Informação (II WSI 2019), Camboriú, SC, Brazil. 2019. Disponível em: <http://www.etic.ifc-camboriu.edu.br/2019/anais/24%20-%20WSI_2019_paper_18.pdf>. Acesso em 26 de maio de 2021.

PADILHA NETO, J. & TALIGNANI, L. F. P. *Método FEATORA na Classificação de Variedades de Trigo: uso de diferentes fórmulas na especificação da quantidade de intervalos de valores.* Trabalho de Conclusão de Curso, Bacharelado em Engenharia de Software, Universidade Estadual de Ponta Grossa, Ponta Grossa, PR, Brazil. 2019.

VRIESMANN, L. M.; BRITTO JR, A. S.; OLIVEIRA, L. E. S.; SABOURIN, R. & KO, A. H.-R. *Improving a Dynamic Ensemble Selection Method based on Oracle Information.* International Journal of Innovative Computing and Applications. Vol. 4, n.3/4, p.184-200, 2012.