

Margin-based ordered aggregation for ensemble pruning

Relatório do projeto da disciplina MCS

Leonardo Valeriano Neri¹

¹Centro de Informática – Universidade Federal de Pernambuco (UFPE)

{lvn}@cin.ufpe.br

1. Introdução

Este trabalho é o relatório do projeto da disciplina de *Multiple Classifier System* (MCS). A descrição do projeto é escolher um artigo publicado em periódicos na área de sistemas de múltiplos classificadores, que seja recente, e replicá-lo, implementando a proposta do artigo, os métodos utilizados na comparação e executar os experimentos descritos com as mesmas bases a fim de obter resultados próximos aos que foram publicados pelo autor. O artigo escolhido deve ser primeiramente aprovado pelo professor da disciplina, George Darmiton, para que se inicie o desenvolvimento do projeto. O projeto foi desenvolvido individualmente e a ferramenta de prototipação escolhida foi o Matlab.

Os sistemas de múltiplos classificadores têm como foco a combinação de classificadores construídos em ambientes heterogêneos ou homogêneos para solucionar o problema de decisão. O objetivo é gerar classificadores mais diversos possíveis para obter soluções mais precisas e robustas. Mais informações a respeito dos sistemas de múltiplos classificadores podem ser encontradas em [Michal Woniak 2014, Kuncheva 2004].

O artigo escolhido tem como título *Margin-based ordered aggregation for ensemble pruning*. Ele foi publicado no periódico *Pattern Recognition Letters*, no ano de 2013 pelos autores Li Guo e Samia Boukir [Li Guo 2013]. Ele trata do assunto de poda de classificadores, propondo um novo critério de ordenação dos classificadores gerados na construção do *pool* para depois selecionar uma determinada quantidade para o *ensemble* final.

A introdução do artigo escolhido [Li Guo 2013] traz uma breve revisão sobre sistemas de múltiplos classificadores e seus diferentes níveis de construção: nível dos dados, nível das características, nível de classificação e combinação. No nível dos dados, diferentes bases podem ser construídas para treinar os classificadores no *ensemble*. O nível de características pode construir diferentes bases selecionando quais características participarão do treinamento dos classificadores, ou ainda, criando novas características. No nível de classificação, determina-se quais tipos de classificadores e quantos deles farão parte do *ensemble*. No nível de combinação, determina-se qual maneira será utilizada para combinar as respostas dos classificadores. Ainda na introdução, o assunto de poda de *ensemble* de classificadores é revisado, mostrando as motivações principais para se utilizar métodos de poda. A poda de classificadores tem como objetivo reduzir o número de componentes do *ensemble* sem que seu desempenho seja deteriorado, e se possível aumentado [Li Guo 2013]. Mais informações sobre poda de *ensemble* podem ser encontradas em [Zhou 2012].

A proposta de [Li Guo 2013] é um método de poda utilizando o paradigma de

margem. É introduzido o conceito de margem do classificador, o qual é utilizado para ordenar os classificadores gerados e construir o *ensemble* no nível de classificação, através da seleção dos classificadores com menores margens.

O relatório é organizado como segue. Na Seção 2, é detalhado o conceito de margem para os métodos de *ensemble*. Na Seção 3, é explicado o método proposto por [Li Guo 2013] para poda. Na Seção 4, os experimentos são descritos, apresentando os resultados obtidos e comparações com os resultados do *pool* de classificadores e resultados de outro método de poda que também utiliza a ordenação de classificadores. Na Seção 5, são feitas as discussões e conclusões finais.

2. Margem para métodos de construção de *ensemble*

O conceito de margem foi originalmente usado para métodos de construção de *ensemble* para compreender o funcionamento dos mesmos [Schapire 1998]. Quanto maior a margem, maior a confiança na classificação. Quanto menor a margem, mais próxima é a relação *a priori* da instância em questão às fronteiras da classe a qual pertence.

Segundo o que é descrito na Seção 2 do artigo escolhido, no problema de classificação, as instâncias próximas às fronteiras das classes são interessantes pois contém informações significativas sobre as classes. Neste caso, os verdadeiros rótulos dessas instâncias não são informações importantes. O autor utiliza uma definição não supervisionada de margem para enfatizar esses casos, sendo calculada pela Equação 1. O termo c_1 é a classe mais votada para a instância x e v_{c_1} é o número de votos relacionados, c_2 é a segunda classe mais votada e v_{c_2} é o número de votos relacionados a ela. O intervalo da margem é entre 0 e 1. Este conceito de margem não depende do rótulo verdadeiro da instância x , com isso é esperado que seja robusto à ruídos, já que não é afetada pelo erro de classificação da instância.

$$margin(x) = \frac{v_{c_1} - v_{c_2}}{T} \quad (1)$$

onde T é o número de classificadores inicial do *pool*.

3. Método de poda de *ensemble* baseado no paradigma de margem

3.1. Critério de ordenação baseado em margem

A ideia proposta para alcançar eficiência nos problemas de classificação é focar em instâncias com margens pequenas, ou seja, próximas à regiões de fronteira entre classes, pois essas instâncias podem trazer informações significativas sobre as classes para melhorar a classificação [Guo 2010].

Através da Equação 1 é definido um critério para escolher os melhores classificadores do *pool*:

$$H(X) = -\frac{1}{|X|} \sum_{i=1}^{|X|} \log(margin(x_i)) \quad (2)$$

onde x_i é uma instância da base X .

Este critério é consistente com o que foi explicado anteriormente sobre a significância da informação que está relacionada com a margem das instâncias. quanto menor a margem de x_i , mais significativa é a informação que ela traz sobre as classes.

3.2. Ordenação dos classificadores do *pool* baseada em margem

O método proposto para poda baseado em ordenação utiliza a significância da informação calculada pela margem para cada classificador no *pool*, $C_t, t = 1, \dots, T$. Cada classificador do *pool* é utilizado para classificar o conjunto de instâncias separado para a poda V e é utilizado o critério de margem nesse conjunto de poda para cada classificador C_t :

$$H_t(V) = -\frac{1}{N} \sum_{i=1}^N \log(\text{margin}(x_i)) \forall (x_i, y_i) \in V / C_t(x_i) = y_i \quad (3)$$

onde x_i é uma instância de V que foi corretamente classificada por C_t .

Todos os classificadores são então ordenados de forma decrescente à significância da informação presente nas regiões de fronteira entre classes. Ou seja, quanto menor a margem do classificador C_t , melhor será seu posicionamento na lista ordenada. Por último, dependendo de quantos classificadores são desejados selecionar, são escolhidos os primeiros M classificadores da lista para compor o *ensemble* final. O valor de M pode ser calculado por meio de otimização de alguma função de avaliação aplicada sobre o conjunto de instâncias para poda. Como exemplo a precisão global, diversidade, entre outros.

3.3. Algoritmo

Para gerar o *pool* de classificadores é utilizado o método Bagging, utilizando árvores de decisão *Classification and Regression Trees* (CART) como os classificadores base. A função de avaliação para determinar a quantidade de classificadores selecionados é a precisão global. O algoritmo proposto pelo autor é definido nos passos abaixo:

1. Usar Bagging para gerar T árvores de decisão CART a partir da base de treino.
2. Classificar o conjunto de poda utilizando o *pool* e calcular a margem para cada instância de poda.
3. Computar o critério de margem definida na Equação 3 para cada classificador no *pool* e ordenar os classificadores de forma que $H_m(C'_t) > H_m(C'_{t+1}), t \leq (T-1)$.
4. Escolher os M primeiros classificadores que obtiverem a melhor precisão global no conjunto de poda para compor o *ensemble* final.
5. Avaliar o *ensemble* obtido no conjunto de testes.

4. Experimentos

4.1. Bases de dados

Foram utilizadas as 10 bases do repositório UCI do artigo escolhido para replicação dos experimentos. O artigo ainda utilizava outras duas bases que não são disponibilizadas em um domínio público, por causa disso, essas duas bases em questão foram descartadas dos experimentos realizados no projeto. Cada base foi dividida em três partes: treinamento, poda e teste, como a Tabela 1 descreve. As instâncias para cada conjunto foram obtidas através de seleções aleatórias sem reposição, utilizando o método *Leave M Out* de validação cruzada, o qual seleciona uma quantidade especificada de instâncias dentro do conjunto total de instâncias. O método *Leave M Out* é utilizado para retirar primeiramente as instâncias do conjunto de treinamento. Em seguida, sobre as instâncias restantes, são retiradas as instâncias para o conjunto de poda, e por último, seguindo a mesma estratégia,

são retiradas as instâncias de teste. As quantidades selecionadas para cada conjunto não foram justificadas pelo autor. Não é explicado o porquê de utilizar o conjunto de poda e de teste com o mesmo tamanho que o conjunto de treinamento. O *pool* de classificadores foi gerado com diversos tamanhos: 1, 51, 101, 151, 201, 251, 301, 351, 401, 451 e 501. O artigo escolhido só reporta os resultados obtidos com o tamanho 501, este relatório traz uma análise mais completa, observando todos os valores do intervalo.

Tabela 1. Bases de dados. Detalhes sobre a quantidade de instâncias utilizadas para cada conjunto, quantidade de atributos e classes.

Base de dados	Treino	Poda	Teste	Atributos	Classes
Connect-4	2000	2000	2000	42	3
Glass	72	71	71	9	6
Kr-vs-kp	1065	1065	1065	36	2
Letter	2000	2000	2000	16	26
Optdigits	1000	1000	1000	64	10
Pendigit	2000	2000	2000	16	10
Pima	256	256	256	8	2
Tic-tac	310	310	310	9	2
Waveform	1000	1000	1000	21	3
Wine quality-red	533	533	533	11	6

4.2. Comparação de desempenho entre métodos de poda

O artigo escolhido compara o desempenho do método proposto com os desempenhos obtidos pelo *pool* de classificadores e outro método de poda baseado em ordenação dos classificadores, nomeado como método de ordenação por precisão. Este método de poda é bem similar ao método proposto pelo autor, com diferença apenas no cálculo da precisão, onde utilizam-se todas as instâncias do conjunto de poda. Os resultados mostrados são um valor médio dos experimentos executados 10 vezes.

Tabela 2. Precisão obtida pelo *pool* gerado através do Bagging nos conjuntos de teste para cada tamanho de *pool*.

Base de dados / Pool	1	51	101	151	201	251	301	351	401	451	501
Connect-4	0.64	0.74	0.75	0.74	0.75	0.75	0.74	0.75	0.75	0.75	0.75
Glass	0.54	0.64	0.65	0.65	0.69	0.63	0.65	0.70	0.64	0.67	0.67
Kr-vs-kp	0.97	0.99	0.99	0.99	0.98	0.99	0.99	0.99	0.99	0.98	0.99
Letter	0.64	0.81	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.81	0.83
Optdigits	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Pendigit	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Pima	0.70	0.74	0.74	0.75	0.75	0.75	0.75	0.74	0.74	0.74	0.75
Tic-tac	0.81	0.93	0.92	0.92	0.91	0.92	0.92	0.93	0.94	0.93	0.93
Waveform	0.73	0.81	0.83	0.83	0.82	0.82	0.83	0.83	0.83	0.83	0.82
Wine quality-red	0.51	0.62	0.61	0.63	0.62	0.63	0.62	0.61	0.61	0.61	0.62

Tabela 3. Precisão obtida pelo método de ordenação por margem selecionando o melhor subconjunto de classificadores para formar o *ensemble*. São mostrados os conjuntos de teste e cada tamanho utilizado para gerar o *pool*.

Base de dados / Pool	1	51	101	151	201	251	301	351	401	451	501
Connect-4	0.64	0.74	0.75	0.74	0.74	0.75	0.74	0.74	0.75	0.75	0.75
Glass	0.54	0.65	0.64	0.63	0.71	0.65	0.69	0.68	0.65	0.68	0.68
Kr-vs-kp	0.97	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Letter	0.64	0.81	0.82	0.82	0.82	0.81	0.82	0.82	0.82	0.81	0.83
Optdigits	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Pendigit	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Pima	0.70	0.75	0.75	0.74	0.75	0.74	0.76	0.74	0.74	0.72	0.74
Tic-tac	0.81	0.95	0.94	0.95	0.95	0.96	0.95	0.96	0.97	0.96	0.95
Waveform	0.73	0.81	0.83	0.83	0.82	0.82	0.83	0.83	0.83	0.83	0.82
Wine quality-red	0.51	0.62	0.61	0.62	0.62	0.62	0.62	0.61	0.62	0.61	0.62

Tabela 4. Precisão obtida pelo método de ordenação por precisão selecionando o melhor subconjunto de classificadores para formar o *ensemble*. São mostrados os conjuntos de teste e cada tamanho utilizado para gerar o *pool*.

Base de dados / Pool	1	51	101	151	201	251	301	351	401	451	501
Connect-4	0.64	0.74	0.74	0.74	0.74	0.75	0.74	0.75	0.75	0.75	0.74
Glass	0.54	0.66	0.67	0.65	0.70	0.68	0.67	0.68	0.67	0.68	0.68
Kr-vs-kp	0.97	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Letter	0.64	0.81	0.82	0.82	0.82	0.82	0.81	0.82	0.82	0.81	0.83
Optdigits	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Pendigit	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Pima	0.70	0.74	0.74	0.75	0.74	0.74	0.76	0.74	0.75	0.73	0.75
Tic-tac	0.81	0.94	0.94	0.94	0.94	0.95	0.94	0.95	0.96	0.96	0.94
Waveform	0.73	0.81	0.83	0.82	0.82	0.82	0.83	0.83	0.82	0.82	0.82
Wine quality-red	0.51	0.61	0.60	0.63	0.61	0.61	0.62	0.61	0.62	0.61	0.62

4.2.1. Desempenho global de classificação

As Tabelas 2, 3 e 4 mostram a média da precisão global para cada base de teste e tamanho de *pool*, de forma individual para o método Bagging sem nenhuma poda, o método de poda baseado em margem e o método de poda baseado em precisão global, respectivamente. A quantidade de classificadores restantes após a poda é a que resulta na maior precisão global no conjunto de poda.

Ao analisar os resultados das tabelas citadas que mostram a média da precisão global, notamos que o método de ordenação por margem supera o método Bagging em 5 das 10 bases utilizadas, na maioria dos casos experimentados para o tamanho inicial do *pool*. Nos outros casos, os desempenhos são semelhantes, o que ainda favorece o método proposto de poda, pois ele alcança resultados semelhantes com um conjunto menor de classificadores.

Comparando com o método de ordenação por precisão, podemos observar que o método de ordenação por margem também é superior em 6 das 10 bases utilizadas para a maioria dos casos experimentados para o tamanho inicial do *pool*. Nos outros casos, os desempenhos são semelhantes ou o desempenho do método de ordenação por precisão é superior.

As comparações demonstram, segundo os experimentos realizados, que o método proposto de ordenação por margem consegue superar o método tradicional de construção de *ensemble* Bagging, como também consegue superar o método de poda que utiliza ordenação por precisão de classificadores para a maioria das bases utilizadas e na maioria dos casos de tamanho inicial do *pool* de classificadores.

Tabela 5. Máxima taxa de erro por classe obtida pelo *pool* gerado através do Bagging nos conjuntos de teste para cada tamanho de *pool*.

Base de dados / Pool	1	51	101	151	201	251	301	351	401	451	501
Connect-4	0.86	0.96	0.95	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.96
Glass	0.77	0.79	0.85	0.81	0.84	0.89	0.82	0.78	0.91	0.83	0.81
Kr-vs-kp	0.03	0.02	0.02	0.01	0.02	0.01	0.02	0.01	0.02	0.02	0.02
Letter	0.52	0.37	0.35	0.31	0.34	0.33	0.33	0.32	0.33	0.33	0.30
Optdigits	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0	0.00
Pendigit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pima	0.43	0.42	0.43	0.40	0.41	0.41	0.40	0.42	0.42	0.43	0.44
Tic-tac	0.26	0.16	0.14	0.19	0.20	0.18	0.14	0.15	0.12	0.12	0.13
Waveform	0.31	0.26	0.22	0.23	0.21	0.24	0.23	0.22	0.23	0.23	0.24
Wine quality-red	1.00	1.00	1.00	0.98	0.61	1.00	1.00	1.00	1.00	0.61	1.00

Tabela 6. Máxima taxa de erro por classe obtida pelo método de ordenação por margem selecionando o melhor subconjunto de classificadores para formar o *ensemble*. São mostrados os conjuntos de teste e cada tamanho utilizado para gerar o *pool*.

Base de dados / Pool	1	51	101	151	201	251	301	351	401	451	501
Connect-4	0.86	0.96	0.96	0.96	0.95	0.96	0.95	0.96	0.97	0.97	0.95
Glass	0.77	0.86	0.80	0.79	0.84	0.85	0.83	0.78	0.87	0.73	0.85
Kr-vs-kp	0.03	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Letter	0.52	0.36	0.35	0.31	0.34	0.33	0.32	0.33	0.33	0.34	0.29
Optdigits	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pendigit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pima	0.43	0.41	0.38	0.41	0.42	0.42	0.39	0.41	0.39	0.42	0.43
Tic-tac	0.26	0.10	0.11	0.11	0.12	0.09	0.10	0.10	0.07	0.08	0.09
Waveform	0.31	0.26	0.23	0.24	0.22	0.25	0.24	0.23	0.23	0.23	0.25
Wine quality-red	1.00	1.00	1.00	0.98	0.61	1.00	1.00	1.00	1.00	0.61	1.00

Tabela 7. Máxima taxa de erro por classe obtida pelo método de ordenação por precisão selecionando o melhor subconjunto de classificadores para formar o *ensemble*. São mostrados os conjuntos de teste e cada tamanho utilizado para gerar o *pool*.

Base de dados / Pool	1	51	101	151	201	251	301	351	401	451	501
Connect-4	0.86	0.95	0.94	0.96	0.95	0.96	0.95	0.96	0.96	0.96	0.96
Glass	0.77	0.84	0.85	0.82	0.87	0.82	0.82	0.78	0.74	0.76	0.80
Kr-vs-kp	0.03	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01
Letter	0.52	0.37	0.35	0.31	0.35	0.33	0.33	0.33	0.33	0.33	0.30
Optdigits	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pendigit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pima	0.43	0.42	0.42	0.41	0.39	0.42	0.41	0.42	0.38	0.41	0.41
Tic-tac	0.26	0.12	0.12	0.13	0.14	0.11	0.10	0.11	0.09	0.09	0.12
Waveform	0.31	0.26	0.24	0.24	0.21	0.25	0.24	0.23	0.25	0.25	0.25
Wine quality-red	1.00	0.98	1.00	0.98	0.61	1.00	1.00	1.00	1.00	0.61	1.00

4.2.2. Desempenho de classificação por classe

As Tabelas 5, 6 e 7 mostram a média da máxima taxa de erro por classe para cada base de teste e tamanho de *pool*, de forma individual para o método Bagging sem nenhuma poda, o método de poda baseado em margem e o método de poda baseado em precisão global, respectivamente. A quantidade de classificadores restantes após a poda é a que resulta na maior precisão global no conjunto de poda.

Ao analisar os resultados das tabelas citadas que mostram a média da máxima taxa de erro por classe, notamos que o método de ordenação por margem supera o método Bagging em 9 das 10 bases utilizadas, na maioria dos casos experimentados para o tamanho inicial do *pool*. Ou seja, a maior taxa de erro por classe obtida pelo método proposto pelo autor é menor que a maior taxa de erro obtida pelo método Bagging. Isso indica que a ideia de selecionar os classificadores que melhor se adequaram às regiões de fronteira fez com que o *ensemble* resultante do método proposto conseguiu classificar corretamente mais instâncias problemáticas do que o Bagging.

Comparando com o método de ordenação por precisão, podemos observar que o método de ordenação por margem também é superior em 5 das 10 bases utilizadas para a maioria dos casos experimentados para o tamanho inicial do *pool*. Isso indica que a ideia de selecionar os classificadores que melhor se adequaram às regiões de fronteira consegue, em alguns casos, mais precisão para as instâncias problemáticas do que a ideia de selecionar os classificadores de acordo com a precisão global.

As comparações demonstram que, segundo os experimentos realizados, o método proposto de ordenação por margem consegue superar o método tradicional de construção de *ensemble* Bagging, como também consegue superar em alguns casos o método de poda que utiliza ordenação por precisão de classificadores.

Tabela 8. Número de classificadores no *ensemble* resultante do método de ordenação por margem.

Base de dados / Pool	1	51	101	151	201	251	301	351	401	451	501
Connect-4	1	43	91	108	124	128	182	180	165	220	162
Glass	1	13	20	8	18	75	57	34	43	9	24
Kr-vs-kp	1	10	4	4	6	8	19	16	8	23	5
Letter	1	42	88	113	163	190	215	228	246	222	287
Optdigits	1	1	1	1	1	1	1	1	1	1	1
Pendigit	1	1	1	1	1	1	1	1	1	1	1
Pima	1	10	15	15	18	17	23	18	26	23	36
Tic-tac	1	18	21	21	25	33	32	35	55	28	52
Waveform	1	42	68	71	101	117	99	137	138	70	159
Wine quality-red	1	34	58	69	86	116	87	78	116	71	1

Tabela 9. Número de classificadores no *ensemble* resultante do método de ordenação por precisão.

Base de dados / Pool	1	51	101	151	201	251	301	351	401	451	501
Connect-4	1	29	28	31	37	48	71	51	60	66	74
Glass	1	8	15	9	15	18	15	13	16	12	8
Kr-vs-kp	1	11	2	8	12	6	6	12	40	7	4
Letter	1	44	58	95	106	132	100	147	116	101	154
Optdigits	1	1	1	1	1	1	1	1	1	1	1
Pendigit	1	1	1	1	1	1	1	1	1	1	1
Pima	1	16	15	14	14	14	16	13	10	22	20
Tic-tac	1	17	11	21	26	21	25	43	34	21	49
Waveform	1	25	35	34	31	52	41	48	37	46	33
Wine quality-red	1	19	20	18	37	29	36	42	28	28	33

4.3. Número de classificadores após a poda

As Tabelas 8 e 9 mostram a média do número de classificadores no *ensemble* após a poda, para cada base de teste e tamanho de *pool*, de forma individual para o método de poda baseado em margem e o método de poda baseado em precisão global, respectivamente. A quantidade de classificadores restantes após a poda é a que resulta na maior precisão global no conjunto de poda.

Comparando com o método de ordenação por precisão, podemos observar que o método de ordenação por margem não consegue construir um *ensemble* com menos componentes que o *ensemble* formado pelo método de ordenação por precisão. Este resultado diverge daquele reportado pelo artigo escolhido, onde é mostrado no caso do *pool* inicial com 501 classificadores, que o método baseado em margem seleciona menos classificadores para a maioria das bases testadas. Com uma análise mais completa para vários tamanhos iniciais de *pool*, observamos que a afirmação do autor não se sustenta.

Tabela 10. Valor de diversidade KW obtido pelo *pool* gerado através do Bagging nos conjuntos de teste para cada tamanho de *pool*.

Base de dados / Pool	1	51	101	151	201	251	301	351	401	451	501
Connect-4	0	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
Glass	0	0.17	0.16	0.16	0.17	0.17	0.16	0.16	0.16	0.15	0.16
Kr-vs-kp	0	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.02	0.01
Letter	0	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12
Optdigits	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pendigit	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pima	0	0.14	0.14	0.14	0.15	0.15	0.14	0.14	0.15	0.15	0.15
Tic-tac	0	0.13	0.13	0.13	0.12	0.13	0.13	0.13	0.13	0.13	0.13
Waveform	0	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
Wine quality-red	0	0.18	0.18	0.19	37.00	0.18	0.19	0.18	0.19	28.00	0.19

Tabela 11. Valor de diversidade KW obtido pelo método de ordenação por margem selecionando o melhor subconjunto de classificadores para formar o *ensemble*. São mostrados os conjuntos de teste e cada tamanho utilizado para gerar o *pool*.

Base de dados / Pool	1	51	101	151	201	251	301	351	401	451	501
Connect-4	0	0.17	0.17	0.17	0.17	0.15	0.17	0.18	0.14	0.15	0.10
Glass	0	0.11	0.06	0.03	0.05	0.08	0.07	0.05	0.05	0.01	0.03
Kr-vs-kp	0	0.09	0.03	0.02	0.03	0.03	0.04	0.04	0.02	0.04	0.01
Letter	0	0.16	0.16	0.17	0.17	0.18	0.18	0.18	0.15	0.16	0.17
Optdigits	0	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pendigit	0	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pima	0	0.11	0.09	0.06	0.06	0.04	0.05	0.03	0.04	0.03	0.05
Tic-tac	0	0.20	0.13	0.10	0.09	0.09	0.07	0.07	0.08	0.05	0.06
Waveform	0	0.19	0.18	0.19	0.16	0.15	0.15	0.12	0.14	0.08	0.13
Wine quality-red	0	0.17	0.16	0.14	37.00	0.14	0.09	0.09	0.09	28.00	0.04

4.4. Diversidade

Para medir a diversidade dos *ensembles* construídos, é utilizada a medida não pareada de Kohavi-Wolpert *variance* (KW) [Kuncheva 2003]. A medida é calculada como:

$$KW = \frac{1}{NL^2} \sum_{j=1}^N l(x_j)(L - l(x_j)) \quad (4)$$

onde N é o tamanho do conjunto de instâncias, L é a quantidade de classificadores do *ensemble* e $l(x_j)$ é a quantidade de classificadores que classificaram corretamente a instância x_j .

As Tabelas 10, 11 e 12 mostram a média do valor da medida de diversidade KW obtida para cada base de teste e tamanho de *pool*, de forma individual para o método Bagging, o método de poda baseado em margem e o método de poda baseado em precisão

Tabela 12. Valor de diversidade KW obtido pelo método de ordenação por precisão selecionando o melhor subconjunto de classificadores para formar o *ensemble*. São mostrados os conjuntos de teste e cada tamanho utilizado para gerar o *pool*.

Base de dados / Pool	1	51	101	151	201	251	301	351	401	451	501
Connect-4	0	0.18	0.12	0.11	0.10	0.10	0.12	0.08	0.08	0.08	0.08
Glass	0	0.08	0.07	0.03	0.04	0.04	0.03	0.02	0.02	0.02	0.01
Kr-vs-kp	0	0.09	0.02	0.05	0.05	0.02	0.02	0.03	0.06	0.01	0.01
Letter	0	0.15	0.18	0.17	0.17	0.17	0.13	0.14	0.11	0.11	0.14
Optdigits	0	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pendigit	0	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pima	0	0.14	0.08	0.06	0.05	0.04	0.04	0.02	0.02	0.03	0.03
Tic-tac	0	0.18	0.08	0.10	0.09	0.06	0.06	0.08	0.07	0.04	0.07
Waveform	0	0.19	0.15	0.13	0.09	0.12	0.08	0.09	0.06	0.07	0.04
Wine quality-red	0	0.13	0.09	0.06	0.08	0.05	0.06	0.05	0.03	28.00	0.03

global, respectivamente. A quantidade de classificadores restantes após a poda é a que resulta na maior precisão global no conjunto de poda.

Ao analisar os resultados das tabelas citadas que mostram a diversidade KW, notamos que o método de ordenação por margem constrói *ensembles* mais diversos do que aqueles que o método Bagging constrói em 5 das 10 bases utilizadas, na maioria dos casos experimentados para o tamanho inicial do *pool*.

Comparando com o método de ordenação por precisão, podemos observar que o método de ordenação por margem constrói *ensembles* mais diversos em 8 das 10 bases utilizadas para a maioria dos casos experimentados para o tamanho inicial do *pool*.

As comparações demonstram que, segundo os experimentos realizados, o método proposto de ordenação por margem consegue construir *ensembles* mais diversos do que aqueles que o método tradicional de construção de *ensemble* Bagging gera, como também consegue superar a diversidade obtida pelo método de poda que utiliza ordenação por precisão de classificadores para a maioria das bases utilizadas e na maioria dos casos de tamanho inicial do *pool* de classificadores. O fato do método de ordenação por margem não construir *ensembles* mais diversos que o Bagging na maioria dos casos apresenta divergência dos resultados obtidos pelo autor.

5. Conclusão

A proposta trabalhada nesse projeto foi a de um método de poda baseado em ordenação de classificadores, o qual utiliza o critério não supervisionado de margem de métodos de construção de *ensemble*. Ela considera que as instâncias com menor margem possuem as informações mais significativas para o problema de classificação e escolhe os classificadores que obtiveram melhores resultados de classificação para essas instâncias. O método demonstrou uma precisão superior que o método tradicional Bagging de geração de *pool*, como também apresentou precisão superior a outro método de poda baseado em ordenação por precisão global. O método consegue taxas máximas de erro por classe inferiores ao Bagging e também inferiores, em alguns casos, aos resultados alcançados pelo

método de ordenação por precisão. Esse último fato diverge dos resultados alcançados pelo autor, o qual mostra que o método de ordenação por margem é superior em todas as medidas de desempenho adotadas utilizando um *pool* com tamanho inicial de 501 classificadores. Na avaliação da quantidade de classificadores selecionados e a diversidade do *ensemble* construído, houveram divergências com os resultados apresentados pelo autor, onde o mesmo mostra que, para um *pool* com tamanho inicial de 501 classificadores, o método proposto seleciona uma quantidade menor de classificadores para a maioria das bases e, para a maioria dos casos de tamanho inicial de *pool*, utilizando a base *Pendigit's* de teste, o método proposto alcança um maior valor de diversidade que o Bagging e o outro método de ordenação. Já os resultados obtidos na avaliação deste projeto mostram que o método proposto pelo autor não consegue selecionar um número de classificadores menor que o método de ordenação por precisão e o *ensemble* criado a partir da poda não é mais diverso que o *pool* gerado pelo Bagging para a maioria das bases utilizadas.

Referências

- Guo, L., B. S. C. N. (2010). Support vectors selection for supervised learning using an ensemble approach. In *ICPR 2010, 20th IAPR International Conference on, Pattern Recognition*, pages 37–40.
- Kuncheva, Ludmila I.; Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. In Schapire, R. E., editor, *Machine Learning*, volume 51, pages 181–207. Kluwer Academic Publishers.
- Kuncheva, L. I. (2004). *Combining Pattern Classifiers Methods and Algorithms*. Wiley-Interscience.
- Li Guo, S. B. (2013). Margin-based ordered aggregation for ensemble pruning. *Pattern Recognition Letters*.
- Michal Woniak, Manuel Graña, E. C. (2014). A survey of multiple classifier systems as hybrid systems. *Information Fusion*.
- Schapire, R., F. Y. B. P. L. W. (1998). Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686.
- Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall CRC.