Avaliação do Algoritmo de *Stacking* em Dados Biomédicos*

Maria Izabela R. Caffé[1,2] , Pedro Santoro Perez[1] & José Augusto Baranauskas[1]

[1]Departamento de Computação e Matemática — DCM
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto — FFCLRP
Universidade de São Paulo — USP
Av. Bandeirantes, 3900 - Ribeirão Preto, SP - 14040-901 - Brasil

[2]Faculdade de Medicina de Ribeirão Preto — FMRP
Universidade de São Paulo — USP
Av. Bandeirantes, 3900 - Ribeirão Preto, SP - 14040-901 - Brasil

mariairc@aluno.ffclrp.usp.br, pedrosperez@usp.br, augusto@usp.br

Abstract. *Stacking is a well studied ensemble technique, but some of its aspects still need to be explored, e.g., there are no recommendations on which and how many algorithms should be used at level-0 or even which algorithm should be used to compose the level-1 meta-classifier. The literature indicates the meta-algorithm at level-1 should be simple, and Naive Bayes has been typically used in studies. This study analyzed stacking in medical datasets, using three different paradigms of machine learning algorithms to compose the meta-classifier. The experiments indicate simple meta-algorithms do not provide good results, and therefore, the meta-classifier must have a degree of complexity for it to achieve a good performance.*

Resumo. *O stacking é uma técnica de combinação de classificadores bem estudada, mas ainda com muitos aspectos a explorar, e.g., não existem recomendações sobre quais e quantos algoritmos devem ser utilizados no nível-0 nem qual algoritmo deve compor o meta-classificador do nível-1. A literatura indica que o meta-algoritmo do nível-1 deve ser simples, sendo Naive Bayes geralmente utilizado nos estudos. Neste estudo analisou-se stacking em conjuntos de dados biomédicos, utilizando três paradigmas de aprendizado de máquina para o meta-classificador. Os experimentos mostram que meta-algoritmos simples não apresentam bons resultados, indicando que devem ter certo grau de complexidade para obterem bom desempenho.*

1. Introdução

Combinação de classificadores (*ensembles*) é um método de aprendizado de máquina no qual vários modelos (hipóteses) são combinados para gerar um único classificador final. Em geral, a utilização de *ensembles* tem a tendência de diminuir a taxa de erro, tornando o classificador final mais preciso, já que utiliza as predições de todos os algoritmos de

*Projeto de pesquisa realizado com apoio financeiro do CNPq/FAPEAM — INCT ADAPTA.

aprendizado envolvidos para gerar uma hipótese final para o mesmo conjunto de dados [Dzeroski & Zenko 2004].

Stacking [Wolpert 1992] é um dos métodos de combinação de classificadores heterogêneos. O método proposto é uma estrutura em duas camadas: no nível-0 vários algoritmos de aprendizado recebem o conjunto de treinamento, gerando os classificadores de nível-0; a camada seguinte (nível-1) tem como entrada as predições da camada anterior (nível-0), na qual um meta-algoritmo de nível-1 as combina para fornecer o meta-classificador final h^*, conforme é mostrado na Figura 1. Mais precisamente, suponha L diferentes algoritmos de aprendizado A_1, A_2, …, A_L e um conjunto de n exemplos $\{(x_1,y_2), (x_2,y_2), \ldots, (x_n,y_n)\}$, onde fica implícito o fato que cada elemento x_i é um vetor. Cada algoritmo de nível-0 é aplicado ao conjunto de treinamento, induzindo os classificadores $\{h_1, h_2, \ldots, h_L\}$ deste nível. Cada classificador do nível-0 é então utilizado para rotular os exemplos. Isso implica que, para cada exemplo x_i, uma tupla é formada, composta pela classe predita por cada um dos classificadores de nível-0 juntamente com a classe verdadeira daquele exemplo, isto é, $(h_1(x_i), h_2(x_i), \ldots, h_L(x_i), y_i)$. Essas tuplas constituem o conjunto de treinamento de nível-1, cujos atributos são as classes preditas por cada um dos L classificadores. Com isso, é possível aplicar o meta-algoritmo de aprendizado aos exemplos de nível-1 para aprender o meta-classificador h^*.

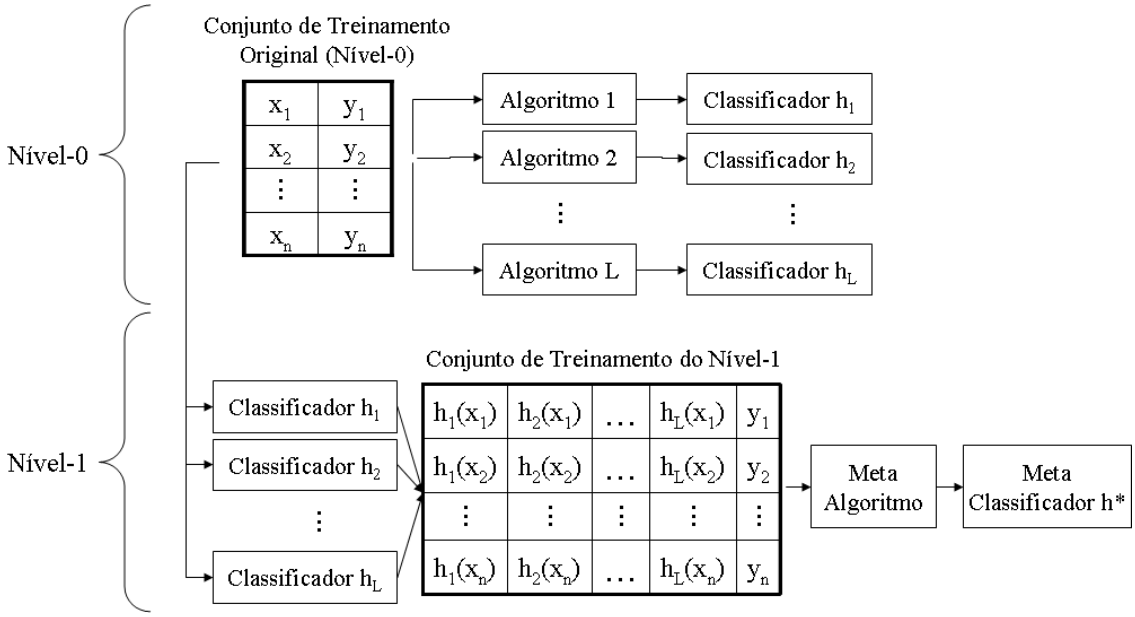

Figura 1: *Stacking*

Uma das explicações para o uso de *stacking* é que ao passar pelas camadas, o meta-classificador aprende os erros dos classificadores anteriores. [Wolpert 1992] relata que muito aspectos sobre *stacking* ainda são desconhecidos (*black art*) no sentido que não há recomendações sobre quais e quantos algoritmos devem ser utilizados no nível-0 nem qual o algoritmo que deve compor o meta-classificador do nível-1. Em geral, os trabalhos na literatura utilizam um número variável de algoritmos no nível-0 e para gerar o meta-classificador o algoritmo Naive Bayes.

Normalmente, dados biomédicos estão associados a uma grande quantidade de classes, quantidade de exemplos muito variada, excesso de atributos e exemplos com dados faltantes e redundantes, o que tem motivado a aplicação de algoritmos de aprendizado de máquina nesse domínio [Pereira & Schmitz 2010, Pollettini *et al.* 2009]. Neste sentido, o estudo de [Tanwani *et al.* 2009] buscou traçar um guia para utilização de *ensembles*, com ênfase em conjuntos de exemplos da área biomédica. Entretanto, os autores utilizaram apenas o algoritmo Naive Bayes como meta-classificador em sua análise do algoritmo de *stacking*. Com isso em mente, o objetivo deste estudo consiste em avaliar o método de *stacking* em conjuntos de dados biomédicos, utilizando diferentes classificadores tanto de nível-0 como de nível-1, estendendo assim, o trabalho de [Tanwani *et al.* 2009] referente a *stacking*.

O restante deste artigo encontra-se organizado da seguinte forma. Na Seção 2 estão descritos os algoritmos usados para induzir os classificadores e meta-classificadores e os conjuntos de exemplos utilizados, bem como os três experimentos que compõem este estudo. Na Seção 3 os resultados são expostos e discutidos; as conclusões encontram-se na Seção 4.

2. Metodologia Experimental

Como mencionado anteriormente, no estudo de [Tanwani *et al.* 2009] envolvendo *stacking* os autores utilizaram apenas o algoritmo Naive Bayes como meta-classificador. No estudo aqui desenvolvido, o uso de *stacking* é ampliado, comparando o desempenho do uso de três paradigmas diferentes no nível-1: estatístico (Naive Bayes), indução de regras simples (One Rule) e árvore de decisão de um único nível (Decision Stump). O objetivo aqui é tentar verificar se algoritmos de diferentes paradigmas de aprendizado resultam em uma melhor combinação no *stacking*.

Resumidamente, este estudo é composto por três experimentos envolvendo stacking realizados em dezoito conjuntos de exemplos biomédicos utilizando vários algoritmos, que são descritos nas próximas seções.

2.1. Conjuntos de Exemplos

Os conjuntos de exemplos utilizados neste trabalho referem-se ao domínio biomédico selecionados do *UCI Machine Learning Repository* [Frank & Asuncion 2010]. Na Tabela 1 são apresentados os conjuntos de dados, data de publicação, número de exemplos, o número de atributos, o número e a distribuição das classes e se há ou não dados faltantes. Neste trabalho não houve nenhum pré-processamento dos conjuntos de dados. A seguir é fornecida uma breve descrição biológica sobre os conjuntos de exemplos.

- **Breast Cancer**: a partir de atributos com informações clínicas e laboratoriais sobre pacientes, a tarefa é distinguir entre eventos recorrentes e não-recorrentes associados a câncer de mama (Instituto de Oncologia da Iugoslávia).
- **Haberman**: com atributos sobre cirurgias de câncer de mama, o objetivo é predizer se o paciente sobreviveu após a cirurgia durante o estudo.
- **Heart-statlog**: também baseado em aspectos clínicos e laboratoriais, o problema de classificação é prever se há ausência ou presença de doença cardíaca.

- **Hepatitis**: considerando dados clínicos e laboratoriais, a tarefa é predizer se um paciente com hepatite morreu ou sobreviveu durante o período do estudo.
- **Liver Disorders**: predição de uma determinada classe binária baseada em testes sanguíneos relacionados a problemas de fígado. A documentação sobre esta base de dados não deixa claro o que a classe significa. Sabe-se apenas que ela pode assumir 2 valores, cada um representado por um número simplesmente.
- **Pima Indians**: problema com atributos sobre dados clínicos e laboratoriais de mulheres descendentes dos indígenas Pima cuja tarefa de classificação é distinguir entre pacientes testadas positivamente para diabetes e pacientes testadas negativamente para a doença.
- **Promoters**: baseada em sequências de DNA de E. coli, a tarefa consiste em discriminar se uma dada sequência é promotora gênica.
- **Sick**: tarefa de prever se pacientes possuem doença relacionada a tireoide baseando-se em dados clínicos e laboratoriais.
- **Contraceptive Method**: baseada em atributos de natureza sócio-econômica de mulheres da Indonésia, a tarefa é predizer qual o método contraceptivo utilizado.
- **Lung Cancer**: os autores desta base de dados não fornecem informação sobre os atributos utilizados. A tarefa é diferenciar entre 3 tipos de câncer de pulmão.
- **Postoperative**: o problema é determinar para onde pacientes na área pós-operatória de recuperação devem ser encaminhados a seguir, sendo os atributos relacionados a medidas clínicas sobre os pacientes (temperatura corporal, pressão sanguínea).
- **Ann-thyroid**: a tarefa é distinguir entre 3 situações envolvendo hipotireoidismo. Os atributos trazem dados sobre medidas laboratoriais relacionadas a problemas na tireoide.
- **Lymphography**: baseados em achados clínicos e laboratoriais relacionados à linfografia, a tarefa é discriminar condições tumorais da linfa (Instituto de Oncologia da Iugoslavia).
- **Cleveland, Switzerland e Hungarian**: o problema aqui consiste em distinguir entre condições relacionadas a doenças do coração. Os atributos trazem informação sobre achados clínicos e laboratoriais.
- **Dermatology**: O problema é determinar o tipo de doença erimato-esquimatosa a partir de achados clínicos e histopatológicos. **Ecoli**: a tarefa é predizer os sítios de localização celular de proteínas a partir de medidas e pontuações relacionadas a sequências protéicas.

2.2. Algoritmos e Meta-Algoritmos

Nos experimentos realizados foram utilizados 9 algoritmos de aprendizado de máquina da biblioteca Weka [Witten & Frank 2005], com seus parâmetros default, exceto quando mencionado o contrário. Uma breve descrição sobre cada um deles é dada a seguir.

- Naive Bayes usa o método probabilístico para classificação [Rish 2001], assumindo independência entre os atributos.
- Bayes Network utiliza diversos algoritmos de busca e medidas de qualidade, fornecendo estruturas sobre os dados [Chickering *et al.* 2004].
- IBK (*Instanced Based Learner*) classifica segundo os K-vizinhos mais próximos [Aha *et al.* 1991]. Neste estudo foi utilizado K = 3.
- JRip induz regras, sendo uma reimplementação do algoritmo Ripper [Cohen 1995].

- One Rule é um classificador muito simples, que seleciona um atributo para compor a regra com menor taxa de erro [Witten & Frank 2005].
- J48 gera árvores de decisão, sendo uma reimplementação do algoritmo C4.5 [Quinlan 1993].
- Decision Stump constrói uma árvore de decisão de um único nível utilizando entropia [Iba & Langley 1992].
- MLP (*Multi Layer Perceptron*) é uma rede neural multicamada utilizando retropropagação para otimizar os pesos durante o treinamento [Haykin 1998].
- SMO (*Support Vector Machine*) algoritmo de otimização mínima sequencial de John Platt para treinar classificadores de vetores de suporte [Vapnik 1998].

Os experimentos envolvendo *stacking* são identificados como Meta 1, Meta 2 e Meta 3, conforme Tabela 2. Cada um dos 9 algoritmos também teve seu desempenho avaliado individualmente. O desempenho foi medido utilizando validação cruzada com 10 partições (10-*fold cross-validation*).

Tabela 1. Conjuntos de exemplos ordenados por número de classes

Conjunto	Ano	Exemplos	Atributos	Classes	Distribuição de Classes	Faltantes
Breast cancer	1988	286	9	2	(70.28, 29.72)	Sim
Haberman	1999	306	4	2	(73.53, 26.47)	Não
Heart statlog	N/D	270	13	2	(55.56, 44.44)	Não
Hepatitis	1988	155	20	2	(20.65, 79.35)	Sim
Liver disorders	1990	345	7	2	(42.03, 57.97)	Não
Pima indians	1990	768	9	2	(65.10, 34.90)	Não
Promoters	1990	106	58	2	(50.00, 50.00)	Não
Sick	1987	3772	30	2	(93.88, 6.12)	Sim
Contraceptive	1997	1473	9	3	(42.70, 22.61, 34.69)	Não
Lung cancer	1992	32	56	3	(28.13, 40.62, 31.25)	Sim
Postoperative	1993	90	9	3	(71.11, 2.22, 26.67)	Sim
Ann-thyroid	1992	7200	22	3	(2.31, 5.11, 95.58)	Não
Lymphography	1988	148	19	4	(1.35, 54.73, 41.22, 2.70)	Não
Cleveland	1988	303	13	5	(54.46, 45.54, 0.00, 0.00, 0.00)	Sim
Hungarian	1988	294	14	5	(63.95, 34.98, 0.00, 0.00, 0.00)	Sim
Switzerland	1988	123	14	5	(6.50, 39.02, 26.02, 24.39, 4.07)	Sim
Dermatology	1998	366	33	6	(30.60, 16.67, 19.57, 13.39, 14.20, 5.46)	Sim
Ecoli	1996	336	8	8	(42.56, 22.92, 15.48, 10.42, 5.95, 1.49, 0.60, 0.60)	Não

Tabela 2. Algoritmos e Meta Algoritmos em cada Experimento

Experimento	Meta Algoritmo	Algoritmos
Meta 1	Naive Bayes	Bayes Net, IBK, One Rule, JRip, J48, Decision Stump, MLP, SMO
Meta 2	One Rule	Bayes Net, Naive Bayes, IBK, JRip, J48, Decision Stump, MLP, SMO
Meta 3	Decision Stump	Bayes Net, Naive Bayes, IBK, One Rule, JRip, J48, MLP, SMO

3. Resultados e Discussão

Na Tabela 3 estão as medidas AUC (área sob a curva ROC) [Bradley 1997] obtidas para os experimentos Meta 1, Meta 2 e Meta 3 e todos os outros algoritmos utilizados neste trabalho, que obtiveram a seguinte colocação: Meta 1, Bayes Net, Naive Bayes, MLP, J48, IBK, SMO, Meta 3, Jrip, Meta 2, Decision Stump e One Rule. Nesta tabela é possível observar que o classificador obtido por Meta 1 teve um *ranking* melhor do que Meta 2 e Meta 3. Entretanto, os classificadores Meta 2 e Meta 3, em geral, tiveram desempenho pior do que os outros classificadores sozinhos, sendo estes Bayes Net, Naive Bayes, MLP, J48, IBK e SMO.

Para avaliar os resultados obtidos sob o ponto de vista estatístico, foi realizado teste de Friedman [Friedman 1940] com significância α=0.05 para determinar se há diferença significativa entre os valores AUC obtidos. Como a hipótese nula foi rejeitada (ou seja, existe diferença entre os classificadores), foi utilizado o teste *post hoc* descrito em [Benjamini & Hochberg 1995] (menos conservador que o teste de Nemenyi [Nemenyi 1963]) para buscar pares com diferença significativa em uma comparação todos-contra-todos; o teste confirmou que Meta 1 se saiu significativamente melhor do que todos os outros algoritmos bem como Meta 2 e Meta 3, exceto Naive Bayes, Bayes Net e MLP, perante o qual Meta 1 apenas se saiu melhor, mas não de forma significativa.

Os resultados do teste podem ser visualizados na Tabela 4, na qual $\vartriangle$ (▲) indica que o algoritmo da respectiva linha foi melhor (significativamente) do que o algoritmo da respectiva coluna e $\triangledown$ (▼) indica que o algoritmo da respectiva linha foi pior (significativamente) do que o algoritmo da respectiva coluna. Por simetria, o triângulo inferior dessa tabela possui resultados opostos ao triângulo superior e foi omitido por questões de clareza.

Como o desvio padrão pode ser visto como uma medida de estabilidade dos algoritmos a pequenas variações no conjunto de treinamento, o teste de Friedman também foi utilizado para avaliar diferenças significativas entre os desvios padrões de todos os algoritmos, incluindo *stacking*. O teste de Friedman encontrou diferença significativa entre os desvios padrão, entretanto o teste *post hoc* com correção por Benjamini & Hochberg não conseguiu detectar diferenças utilizando α=0.05. Neste sentido, os resultados obtidos permitem afirmar que, para os conjuntos de exemplos avaliados, a utilização de *stacking* não melhora nem piora a estabilidade dos algoritmos.

Tabela 3. Medida AUC(±Desvio Padrão) dos Experimentos

Conjunto	Naive Bayes	BayesNet	IBK	JRip	One Rule	Decision Stump	J48	MLP	SMO	Meta 1	Meta 2	Meta 3
Breast câncer	0.72±0.14	0.71±0.14	0.66±0.13	0.61±0.10	0.54±0.07	0.65±0.13	0.63±0.10	0.62±0.13	0.59±0.08	0.69±0.15	0.57±0.09	0.67±0.11
Haberman	0.67±0.11	0.69±0.09	0.63±0.11	0.62±0.10	0.59±0.08	0.64±0.08	0.58±0.08	0.66±0.13	0.51±0.02	0.72±0.05	0.51±0.07	0.55±0.04
Heart statlog	0.90±0.06	0.91±0.04	0.83±0.07	0.80±0.08	0.71±0.06	0.72±0.07	0.76±0.10	0.85±0.06	0.84±0.06	0.89±0.06	0.79±0.08	0.83±0.08
Hepatitis	0.86±0.11	0.89±0.08	0.79±0.15	0.60±0.15	0.65±0.13	0.67±0.13	0.70±0.20	0.82±0.15	0.75±0.13	0.83±0.12	0.73±0.16	0.73±0.16
Liver disorders	0.65±0.12	0.52±0.03	0.64±0.06	0.64±0.09	0.54±0.07	0.54±0.06	0.67±0.08	0.74±0.07	0.50±0.01	0.75±0.08	0.58±0.06	0.66±0.09
Pima indians	0.82±0.05	0.81±0.05	0.74±0.05	0.72±0.06	0.67±0.06	0.69±0.06	0.75±0.08	0.80±0.04	0.72±0.06	0.83±0.05	0.67±0.04	0.73±0.05
Promoters	0.96±0.11	0.96±0.11	0.92±0.08	0.81±0.13	0.70±0.11	0.71±0.12	0.83±0.14	0.98±0.08	0.93±0.09	0.97±0.08	0.58±0.11	0.90±0.09
Sick	0.93±0.05	0.96±0.02	0.88±0.05	0.94±0.05	0.89±0.04	0.94±0.03	0.95±0.04	0.95±0.03	0.50±0.00	0.99±0.01	0.93±0.03	0.94±0.03
Contraceptive	0.69±0.04	0.70±0.04	0.62±0.04	0.62±0.04	0.58±0.02	0.56±0.02	0.66±0.04	0.70±0.02	0.63±0.03	0.73±0.04	0.59±0.04	0.60±0.02
Lung cancer	0.71±0.32	0.71±0.29	0.68±0.31	0.55±0.15	0.54±0.15	0.55±0.11	0.68±0.23	0.56±0.22	0.63±0.20	0.80±0.19	0.70±0.21	0.57±0.15
Postoperative	0.39±0.22	0.40±0.21	0.31±0.13	0.50±0.00	0.48±0.03	0.46±0.13	0.49±0.02	0.41±0.19	0.47±0.04	0,57±0.18	0.46±0.11	0.43±0.12
Ann-thyroid	0.93±0.02	1.00±0.00	0.74±0.04	0.99±0.01	0.95±0.02	0.99±0.00	0.99±0.00	0.97±0.03	0.59±0.02	1.00±0.00	0.97±0.03	0.99±0.00
Lymph	0.91±0.07	0.91±0.07	0.89±0.10	0.78±0.14	0.77±0.11	0.78±0.11	0.79±0.14	0.91±0.07	0.87±0.08	0.87±0.11	0.80±0.06	0.81±0.07
Cleveland	0.90±0.06	0.91±0.04	0.85±0.07	0.84±0.07	0.72±0.07	0.71±0.07	0.80±0.09	0.89±0.06	0.84±0.08	0.91±0.05	0.79±0.08	0.81±0.07
Hungarian	0.90±0.06	0.91±0.07	0,87±0.06	0.76±0.10	0.75±0.11	0.77±0.11	0.77±0.15	0.89±0.04	0.80±0.11	0.89±0.07	0.83±0.06	0.80±0.08
Switzerland	0.53±0.12	0.54±0.04	0.49±0.09	0.56±0.07	0.53±0.10	0.47±0.06	0.55±0.09	0.54±0.12	0.57±0.09	0.52±0.13	0.50±0.07	0.51±0.04
Dermatology	1.00±0.00	1.00±0.00	0.99±0.01	0.95±0.02	0.67±0.03	0.79±0.02	0.97±0.02	1.00±0.00	0.98±0.01	1.00±0.00	0.82±0.03	0.78±0.01
Ecoli	0.97±0.02	0.97±0.02	0.94±0.03	0.91±0.04	0.75±0.04	0.78±0.02	0.92±0.05	0.96±0.02	0.95±0.02	0.93±0.04	0.81±0.04	0.83±0.02
Ranking Médio AUC	4.00	3.06	6.83	7.64	10.42	9.28	6.56	4.28	7.19	2.58	8.61	7.50
Ranking Médio Desvio Padrão	7.14	5.47	7.31	7.83	6.64	6.08	8.94	5.54	5.14	5.83	6.47	5.33

Tabela 4. Resultados do Teste de Friedman para medida AUC

Algoritmo	Naive Bayes	Bayes Net	IBK	JRip	One Rule	Decison Stump	J48	MLP	SMO	Meta 1	Meta 2	Meta 3
NaiveBayes		∇	▲	▲	▲	▲	∆	∆	▲	∇	▲	▲
Bayes Net			▲	▲	▲	▲	▲	∆	▲	∇	▲	▲
IBK				∆	▲	∆	∇	∇	∆	▼	∆	∆
Jrip					▲	∆	∇	▼	∇	▼	∆	∇
One Rule						∇	▼	▼	▼	▼	∇	▼
Decision Stump							▼	▼	∇	▼	∇	∇
J48								∇	∆	▼	∆	∆
MLP									▲	∇	▲	▲
SMO										▼	∆	∆
Meta 1											▲	▲
Meta 2												∇
Meta 3												

4. Conclusão

Neste estudo a avaliação de *stacking* foi realizada utilizando três paradigmas diferentes para geração do meta-classificador: estatístico, indução de regras simples e árvore de decisão de um único nível, estendendo assim, o trabalho de [Tanwani *et al.* 2009] no qual *stacking* utilizou somente um algoritmo estatístico como meta-classificador. Os resultados aqui obtidos permitem concluir que *stacking* utilizando como meta-classificador um algoritmo estatístico tem um desempenho significativamente superior do que indução de regras ou árvores simples.

Na literatura associada é recomendável a utilização de algoritmos simples para compor o meta-classificador [Dzeroski & Zenko 2002]. Entretanto, o presente estudo indica que o algoritmo utilizado para compor o meta-classificador deve ter um grau de sofisticação tal que o permita representar adequadamente os conceitos oriundos do nível-0, algo que regras contendo um único atributo ou árvores de um único nível claramente não o possuem. Com isso, é possível indicar que estudos futuros utilizem um algoritmo de complexidade controlada para compor o meta-classificador — por exemplo, uma árvore de decisão com níveis decrescentes de poda ou uma rede neural com quantidades crescentes de neurônios, sinapses ou ciclos de treinamento — de forma a identificar, se possível, o nível de complexidade exigido para compor um bom meta-classificador.

Referências

Aha, D. W., Kibler, D. & Albert, M. K. (1991) “Instance based learning algorithms” In Machine Learning, p. 37–66.

Benjamini, Y. & Hochberg, Y. (1995) “Controlling the false discovery rate: a practical and powerful approach to multiple testing”, In Journal of the Royal Statistical Society Series B, v. 57, p. 289–300.

Bradley, A. P. (1997) “The use of the area under the ROC curve in the evaluation of machine learning algorithms”. Pattern Recognition 30(7), 1145–1159.

Chickering, D. M., Heckerman, D. & Meek, C. (2005) “Learning of Bayesian Networks is NP – Hard” In Journal of Machine Learning Research, 5, p 1287–1330.

Cohen, W. W. (1995) “Fast effective rule induction” In Proceedings of Twelfth International Conference on Machine Learning, p. 115–123.

Dzeroski, S. & Zenko B. (2002) “Is combining Classifiers Better than Selecting the Best One?” In Proceedings of the 19th International Conference on Machine Learning, Morgan Kaufmann Publishers, San Francisco.

Frank, A. & Asuncion, A. (2010) “UCI Machine Learning Repository”, http://archive.ics.uci.edu/ml, School of Information and Computer Science, University of California at Irvine, Irvine CA.

Friedman, M. (1940) “A comparison of alternative tests of significance for the problem of m rankings”. The Annals of Mathematical Statistics 11(1), 86–92.

Haykin, S. (1998) Neural networks: a comprehensive foundation, 2nd edition, Pearson Education, London.

Iba, W. & Langley, P. (1992) “Induction of One – Level Decision Trees” In Proceedings of the Ninth International Conference on Machine Learning.

Nemenyi, P. B. (1963) Distribution-free multiple comparisons, PhD. Thesis, Princeton University.

Pereira, M. & Schmitz, A. (2010) "Inteligência Artificial e Geotecnologias Emergentes Aplicadas em Estudos Ecoepidemiológicos de Malária no Município de Bragança-Pará, Brasil, no Período de 2006 a 2008", In Proceedings do X Workshop de Informática Médica, Congresso da Sociedade Brasileira de Computação, p. 1630–1640, Belo Horizonte.

Pollettini, J. T., Tinos, R., Panico, S., Daneluzzi, J. C. & Macedo, A. A. (2009) "Vigilância em atenção básica à saúde a partir do uso de relevance feedback para classificação de pacientes em diferentes níveis de cuidado em saúde", In Proceedings do IX Workshop de Informática Médica, Congresso da Sociedade Brasileira de Computação, p. 1945–1954, Bento Gonçalves.

Quinlan, J. R. (1993) C4.5: programs for machine learning, Morgan Kaufmann, San Francisco.

Rish, I. (2001) “An empirical study of the naive Bayes classifier”, In IJCAI Workshop on Empirical Methods in Artificial Intelligence, p. 41–46.

Seewald, A. K. (2002) “How to make Stacking Better and Faster While Also Taking Care of an Unknown Weakness”, In Proceedings of the 19th International Conference on Machine Learning, p. 554–561, Morgan Kaufmann Publishers, Sydney.

Seewald, A. K. (2002) “Exploring the Parameter State Space of Stacking”, In Proceedings of the 2002 IEEE International Conference of Data Mining (ICDM'02), p. 685–688.

Tanwani, A. K., Afridi, J. Shafiq, M. Z. & Farroq, M. (2009) “Guidelines to Select Machine Learning Scheme for Classification of Biomedical Datasets”, C.Pizzuti, M.D. Ritchie, & M. Giacobini, LNCS 5483, Springer-Verlag Berlin Heidelberg 2009, p. 128–139.

Vapnik, V. N. (1998) Statistical learning theory, Wiley Interscience, USA.

Witten, I. H. & Frank, E. (2005) Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, 2. ed.: Morgan Kaufmann.

Wolpert, D. H. (1992) Stacked Generalization. In *Neural Networks,* p. 241–260.