INF1771 - Inteligência Artificial - (2018.2)

Professora: Renatha Capua

Trabalho Machine Learning

Dupla:

Rodrigo Pumar Alves de Souza

Bruno Pedrazza

1. Introdução e Dados Analisados

O Dataset que foi utilizado para o trabalho de machine learning foi o Poker Hands do site UCI http://archive.ics.uci.edu/ml/ e foi utilizado o programa WEKA (http://www.cs.waikato.ac.nz/ml/weka/) para tratar os dados.

Os dados consistem em combinações de cartas possíveis num jogo de poker, no total é possível mais que 2,5 milhões de combinações neste jogo. Embora definir se uma combinação de cartas é qual tipo de mão valida no jogo seja trivial programando tradicionalmente, o nosso interesse por jogo e por tentar ver como os algoritmos de machine learning poderiam aprender o jogo secular de poker.

Abaixo listo as Poker Hands (Classificador) e as 5 cartas para cada possível poker hand como exemplo, sendo cada carta um par de atributos, totalizando 10 atributos.

Poker Hands - Classificador	Exemplo - 10 Atributos, 5 pares de naipe e valor (ordenados)	Possible hands	# of combinations	% of hands
Royal Straight Flush (RF)	H,1,H,10,H,11,H,12,H,13	4	480	0.000154%
Straight Flush (SF)	H,2,H,3,H,4,H,5,H,6	36	4,320	0.001385%
Four of a Kind (4)	D,8,C,8,H,8,S,8,H,12	624	74,880	0.024010%
Full House (FH)	S,2,H,2,D,7,H,7,S,7	3,744	449,280	0.144058%
Flush (F)	C,2,C,3,C,4,C,5,C,12	5,108	612,960	0.196540%
Straight (S)	C,9,D,10,H,11,D,12,S,13	10,200	1,224,000	0.392465%
Three of a Kind (3)	S,7,C,11,D,11,S,11,D,13	54,912	6,589,440	2.112845%
Two Pairs (2)	D,1,C,1,H,4,C,4,S,13	123,552	14,826,240	4.753902%
One Pair (1)	H,1,S,1,S,3,H,5,D,9	1,098,240	131,788,800	42.256903%
Only Singles(0)	S,1,S,4,S,6,C,9,C,13	1,302,540	156,304,800	50.117739%
10 Classificadores	**Total**	**2,598,960**	311,875,200	100%
		(2.5 milhões)	311 milhões	

Nesse jogo, todos as classificações precisam saber todos os atributos, excluindo se as 4 primeiras cartas forem de mesmo valor (4 atributos de valor iguais) pois assim nesse caso os outros 6 atributos são irrelevantes.

Outro fato relevante é que quanto mais valiosa a poker hand (no sentido das regras do jogo), são estaticamente mais raras e demandam mais combinações de atributos e correlações que o algoritmo terá que aprender, algo inerente do jogo de poker. Portanto, os algoritmos de aprendizado terão dificuldade de aprender essas jogadas, tanto pela raridade pela complexidade. E é um desafio fazer com esses dados temos acertos quanto mais valiosa é a poker hand.

Contagem Classificadores	25010 (25K)		1Milhão (1M)	
Classificador	Soma	Porcentagem	Soma	Porcentagem
0	12493	49.952%	501209	50.1209%
1	10599	42.379%	422498	42.2498%
2	1206	4.822%	47622	4.7622%
3	513	2.051%	21121	2.1121%
S	93	0.372%	3885	0.3885%
F	54	0.216%	1996	0.1996%
FH	36	0.144%	1424	0.1424%
4	6	0.024%	230	0.0230%
SF	5	0.020%	12	0.0012%
RF	5	0.020%	3	0.0003%
Total	25010	100,00%	1000000	100.00%

2. Implementação e Métodos

Foi utilizado os algoritmos já implementados no WEKA, utilizando 2 tipos de algoritmos de aprendizado diferentes, Naive Bayes, Arvore de decisão C4.5.

2.1. Naives Bayes

Assumimos que os atributos são independentes quando usamos Naive Bayes, oque é relativamente correto, visto que separamos cada carta em dois atributos, sendo naipe e valor, que podem ser considerados independentes.

Foi comparado o uso de estimador de densidade Kernel com distribuição normal, onde o estimador de densidade Kernel e mostrou melhor. Achamos que é porque os atributos são igualmente distribuídos, fazendo com que a distribuição normal fosse pior.

2.2. Decision Tree C4.5 (J48 no WEKA)

A arvore de decisão necessita que os atributos sejam valores categóricos, como o naipe pode ser 4 (H,S,D,C) e o valor 13 (inteiros 1 até 13), esse método funciona.

Foi melhorado as configurações padrões do WEKA mudando parâmetros:

Confiança de 0.25 para 0.5 para diminuir erro por podas, visto que a para classificar as cartas quase todos os atributos são necessários, podar a arvore sem cuidado aumentaria o erro.

Número mínimo de instancias para cada folha de 1 (ao invés do padrão 2), visto que algumas poker hands são muito raras, poucas instancias poderiam acontecer para certas mãos.

3. Metodologia de testes

Foi utilizado o método de 5-fold validation, pois garantiria que os resultados mostrados seriam a medias dos 5 testes e mediria bem casos de overfitting.

Esses métodos foram utilizados para datasets diferentes de 25mil dados e 1 milhão de dados para verificar a escalabilidade do método com o tamanho do dataset, tanto para aumento de acerto para tempo de criar o modelo.

4. Pré-processamento

4.1. Ordenamento das cartas em Valor

Durante os testes, principalmente da arvore de decisão, foi apresentado resultados muito piores que o esperado com os dados sem tratamento.

Na busca de melhoria, estudamos o problema do poker e chegamos a conclusão que embora todas os 10 atributos (5 cartas) fossem importantes para definir o classificador, a ordem delas não importava. Porem, os algoritmos estavam encarando cartas iguais embaralhadas em ordem diferentes como sendo poker hands diferentes, o que não corresponde com a realidade. Como solução, decidimos ordenar as cartas pelos valores num script em C#, de modo agora duas poker hands iguais que antes eram embaralhadas seriam consideradas iguais para o aprendizado de máquina.

Para arvore de decisão foi duplamente útil, pois garantia que os primeiros atributos eram mais previsíveis (menores do caso de valor), sendo assim facilitava na geração de uma boa arvore de decisão.

4.2. Criação de um novo atributo se todos os Naipes eram o mesmo

A análise dos resultados mostrou que todos os algoritmos tinham dificuldade de reconhecer que todos os naipes iguais eram uma variável importante, principalmente devido a raridade dessa poker hands e eles serem nominais.

Foi então adicionado no pré-processamento uma nova variável, que era TRUE caso todos os atributos de naipe fossem o mesmo, e FALSE caso contrário.

5. Resultados e comparações

Naive Bayes 25K padrão - Matriz de Confusão													
Classificadores		Previsto										Falso Negativo	
		0	1	2	3	S	F	FH	4	SF	RF		
Realmente	0	12425	68	0	0	0	0	0	0	0	0	0.54%	68
	1	10557	42	0	0	0	0	0	0	0	0	0.00%	0
	2	1201	5	0	0	0	0	0	0	0	0	0.00%	0
	3	510	3	0	0	0	0	0	0	0	0	0.00%	0
	S	93	0	0	0	0	0	0	0	0	0	0.00%	0
	F	52	2	0	0	0	0	0	0	0	0	0.00%	0
	FH	36	0	0	0	0	0	0	0	0	0	0.00%	0
	4	6	0	0	0	0	0	0	0	0	0	0.00%	0
	SF	5	0	0	0	0	0	0	0	0	0	0.00%	0
	RF	5	0	0	0	0	0	0	0	0	0	TOTAL	68
Falso Positivo		50.08%	19.23%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	TOTAL	Acertos	% Acerto
		12465	10	0	0	0	0	0	0	0	12475	12467	49.8481%
Tempo de build		0.02	segundos										

Naive Bayes 25K ordenado kernel - Matriz de Confusão													
Classificadores		Previsto										Falso Negativo	
		0	1	2	3	S	F	FH	4	SF	RF		
Realmente	0	10728	1765	0	0	0	0	0	0	0	0	14.13%	1765
	1	5809	4723	37	12	18	0	0	0	0	0	0.63%	67
	2	418	709	43	26	10	0	0	0	0	0	2.99%	36
	3	102	287	18	97	9	0	0	0	0	0	1.75%	9
	S	2	70	1	0	8	0	0	0	0	12	12.90%	12
	F	45	9	0	0	0	0	0	0	0	0	0.00%	0
	FH	11	17	3	5	0	0	0	0	0	0	0.00%	0
	4	0	5	0	1	0	0	0	0	0	0	0.00%	0
	SF	1	3	0	0	1	0	0	0	0	0	0.00%	0
	RF	0	3	0	0	0	0	0	0	0	2	TOTAL	1889
Falso Positivo		37.32%	14.53%	21.57%	4.26%	2.17%	0%	0.00%	0.00%	0.00%	TOTAL	Acertos	% Acerto
		6388	1103	22	6	1	0	0	0	0	7520	15601	62.3790%
Tempo de build		0.02	segundos										

Naive Bayes 25K ordenado padrão - Matriz de Confusão													
Classificadores		Predicted										Falso Negativo	
		0	1	2	3	S	F	FH	4	SF	RF		
Realmente	0	9988	2497	0	4	0	0	0	0	4	0	20.05%	2505
	1	5696	4708	47	71	32	0	0	0	37	8	1.84%	195
	2	444	660	27	36	29	0	0	0	10	0	6.22%	75
	3	81	378	7	19	21	0	0	0	7	0	5.46%	28
	S	0	55	9	6	7	0	0	0	4	12	17.20%	16
	F	42	12	0	0	0	0	0	0	0	0	0.00%	0
	FH	7	22	2	3	1	0	0	0	1	0	2.78%	1
	4	0	6	0	0	0	0	0	0	0	0	0.00%	0
	SF	0	2	0	2	1	0	0	0	0	0	0.00%	0
	RF	0	0	0	0	0	0	0	0	0	5	TOTAL	2820
Falso Positivo		38.57%	13.61%	19.57%	7.80%	2.20%	0.00%	0.00%	0.00%	0.00%	TOTAL	Acertos	% Acerto
		6270	1135	18	11	2	0	0	0	0	7436	14754	58.9924%
Tempo de build		0.06	segundos										

Naive Bayes 25K ordenado kernel novo atributo - Matriz de Confusão													
Classificadores		Previsto										Falso Negativo	
		0	1	2	3	S	F	FH	4	SF	RF		
Realmente	0	10728	1765	0	0	0	0	0	0	0	0	14.13%	1765
	1	5809	4724	36	12	18	0	0	0	0	0	0.62%	66
	2	418	709	43	26	10	0	0	0	0	0	2.99%	36
	3	102	287	18	97	9	0	0	0	0	0	1.75%	9
	S	2	71	1	0	8	0	0	0	0	11	11.83%	11
	F	0	1	0	0	1	51	0	0	1	0	1.85%	1
	FH	11	17	3	5	0	0	0	0	0	0	0.00%	0
	4	0	5	0	1	0	0	0	0	0	0	0.00%	0
	SF	0	0	0	2	2	1	0	0	0	0	0.00%	0
	RF	0	0	0	0	0	0	0	0	0	5	TOTAL	1888
Falso Positivo		37.15%	14.38%	21.78%	5.59%	6.25%	1.92%	0.00%	0.00%	0.00%	TOTAL	Acertos	% Acerto
		6342	1090	22	8	3	1	0	0	0	7466	15656	62.5990%
Tempo de build		0.02	segundos										

Naive Bayes 1M ordenado kernel novo atributo - Matriz de Confusão													
Classificadores		Previsto										Falso Negativo	
		0	1	2	3	S	F	FH	4	SF	RF		
Realmente	0	432972	67818	0	0	419	0	0	0	0	0	13.61%	68237
	1	236971	182838	1314	592	783	0	0	0	0	0	0.64%	2689
	2	16971	28304	1806	515	26	0	0	0	0	0	1.14%	541
	3	4604	12381	322	3781	33	0	0	0	0	0	0.16%	33
	S	1	3472	0	0	411	0	0	0	0	1	0.03%	1
	F	0	0	0	0	2	1990	0	0	3	1	0.20%	4
	FH	224	826	115	220	0	0	33	6	0	0	0.42%	6
	4	0	147	15	28	0	0	7	33	0	0	0.00%	0

	SF	0	0	0	0	4	8	0	0	0	0	0.00%	0
	RF	0	0	0	0	0	0	0	0	0	3	TOTAL	71511
Falso Positivo		37.41%	15.26%	12.65%	4.83%	0.36%	0.40%	17.50%	0.00%	0.00%	TOTAL	Acertos	% Acerto
		258771	45130	452	248	6	8	7	0	0	304622	623867	62.3867%
Tempo de build		1.19	segundos										

Árvore de decisão J48 25K padrão - Matriz de Confusão													
Classificadores		Previsto										Falso Negativo	
		0	1	2	3	S	F	FH	4	SF	RF		
Realmente	0	8446	3855	147	33	8	3	0	0	0	1	32.39%	4047
	1	5474	4857	208	55	4	0	0	0	0	1	2.53%	268
	2	447	703	42	13	0	0	0	1	0	0	1.16%	14
	3	188	267	15	42	1	0	0	0	0	0	0.19%	1
	S	53	38	1	0	1	0	0	0	0	0	0.00%	0
	F	39	4	0	0	0	11	0	0	0	0	0.00%	0
	FH	9	22	2	3	0	0	0	0	0	0	0.00%	0
	4	2	3	1	0	0	0	0	0	0	0	0.00%	0
	SF	3	1	0	0	0	1	0	0	0	0	0.00%	0
	RF	3	1	0	0	0	1	0	0	0	0	TOTAL	4330
Falso Positivo		42.40%	10.66%	4.57%	2.05%	0.00%	12.50%	#DIV/0!	0.00%	#DIV/0!	TOTAL	Acertos	% Acerto
		6218	1039	19	3	0	2	0	0	0	7281	13399	53.5746%
Tempo de build		2.9	segundos										

Árvore de decisão J48 25K ordenado padrão - Matriz de Confusão													
Classificadores		Previsto										Falso Negativo	
		0	1	2	3	S	F	FH	4	SF	RF		
Realmente	0	12477	14	0	0	0	2	0	0	0	0	0.13%	16
	1	188	10224	139	3	22	1	2	0	0	0	1.58%	167
	2	0	495	707	2	2	0	0	0	0	0	0.33%	4
	3	0	73	11	425	1	0	1	2	0	0	0.78%	4
	S	0	18	1	0	74	0	0	0	0	0	0.00%	0
	F	54	0	0	0	0	0	0	0	0	0	0.00%	0
	FH	0	5	8	23	0	0	0	0	0	0	0.00%	0

	4	0	0	1	5	0	0	0	0	0	0	0.00%	0
	SF	0	3	0	0	2	0	0	0	0	0	0.00%	0
	RF	0	0	0	0	5	0	0	0	0	0	TOTAL	191
Falso Positivo		1.90%	5.48%	2.42%	6.11%	6.60%	0.00%	0.00%	0.00%	0.00%	TOTAL	Acertos	% Acerto
		242	594	21	28	7	0	0	0	0	892	23907	95.6663%
Tempo de build		0.93	segundos										

Árvore de decisão J48 25K ordenado otimizado - Matriz de Confusão													
Classificadores		Previsto										Falso Negativo	
		0	1	2	3	S	F	FH	4	SF	RF		
Realmente	0	12475	15	0	0	0	3	0	0	0	0	0.14%	18
	1	140	10238	192	8	17	1	2	0	1	0	2.09%	221
	2	0	433	764	3	0	0	6	0	0	0	0.75%	9
	3	0	46	8	448	1	0	7	3	0	0	2.14%	11
	S	0	19	1	0	70	0	0	0	1	2	3.23%	3
	F	54	0	0	0	0	0	0	0	0	0	0.00%	0
	FH	0	2	8	23	0	0	3	0	0	0	0.00%	0
	4	0	0	0	6	0	0	0	0	0	0	0.00%	0
	SF	0	1	0	0	4	0	0	0	0	0	0.00%	0
	RF	0	0	0	0	5	0	0	0	0	0	TOTAL	262
Falso Positivo		1.53%	4.66%	1.75%	5.94%	9.28%	0.00%	0.00%	0.00%	0.00%	TOTAL	Acertos	% Acerto
		194	501	17	29	9	0	0	0	0	750	23998	95.9536%
Tempo de build		0.94	segundos										

Árvore de decisão J48 25K otimizado ordenado novo atributo - Matriz de Confusão													
Classificadores		Previsto										Falso Negativo	
		0	1	2	3	S	F	FH	4	SF	RF		
Realmente	0	12482	11	0	0	0	0	0	0	0	0	0.09%	11
	1	123	10276	181	5	12	0	2	0	0	0	1.89%	200
	2	0	433	761	5	0	0	7	0	0	0	1.00%	12
	3	0	40	8	454	1	0	8	2	0	0	2.14%	11
	S	0	8	0	0	85	0	0	0	0	0	0.00%	0
	F	0	0	0	0	0	51	0	0	3	0	5.56%	3

	FH	0	2	7	24	0	0	3	0	0	0	0.00%	0
	4	0	0	0	6	0	0	0	0	0	0	0.00%	0
	SF	0	0	0	0	0	1	0	0	3	1	20.00%	1
	RF	0	0	0	0	0	0	0	0	0	5	TOTAL	238
Falso Positivo		0.98%	4.48%	1.57%	6.07%	0.00%	1.92%	0.00%	0.00%	0.00%	TOTAL	Acertos	% Acerto
		123	483	15	30	0	1	0	0	0	652	24120	96.4414%
Tempo de build		0.98	segundos										

Árvore de decisão J48 1M otimizado ordenado novo atributo - Matriz de Confusão													
Classificadores		Previsto										Falso Negativo	
		0	1	2	3	S	F	FH	4	SF	RF		
Realmente	0	501209	0	0	0	0	0	0	0	0	0	0.00%	0
	1	0	422498	0	0	0	0	0	0	0	0	0.00%	0
	2	0	0	47622	0	0	0	0	0	0	0	0.00%	0
	3	0	0	0	21121	0	0	0	0	0	0	0.00%	0
	S	0	0	0	0	3885	0	0	0	0	0	0.00%	0
	F	0	0	0	0	0	1994	0	0	2	0	0.10%	2
	FH	0	0	2	2	0	0	1420	0	0	0	0.00%	0
	4	0	0	0	33	0	0	3	194	0	0	0.00%	0
	SF	0	0	0	0	0	4	0	0	8	0	0.00%	0
	RF	0	0	0	0	0	2	0	0	0	1	TOTAL	2
Falso Positivo		0.00%	0.00%	0.00%	0.17%	0.00%	0.30%	0.21%	0.00%	0.00%	TOTAL	Acertos	% Acerto
		0	0	2	35	0	6	3	0	0	46	999952	99.9952%
Tempo de build		157.8	segundos										

5.1. Resultados Agrupados

Método	% Acertos	Tempo de build 5-fold (segundos)	Falso Negativo		Falso Positivo	
Naive Bayes 25K padrão	49.8481%	0.02	68	0.2719%	12475	49.8800%
Naive Bayes 25K ordenado padrão	58.9924%	0.06	2820	11.2755%	7436	29.7321%
Naive Bayes 25K ordenado kernel	62.3790%	0.02	1889	7.5530%	7520	30.0680%
Naive Bayes 25K ordenado kernel novo atributo	62.5990%	0.02	1888	7.5490%	7466	29.8521%
Naive Bayes 1M ordenado kernel novo atributo	62.3867%	1.19	71511	7.1511%	304622	30.4622%
Árvore de decisão J48 25K padrão	53.5746%	2.9	4330	17.3131%	7281	29.1124%
Árvore de decisão J48 25K ordenado padrão	95.6663%	0.93	191	0.7637%	892	3.5666%
Árvore de decisão J48 25K ordenado otimizado	95.9536%	0.94	262	1.0476%	750	2.9988%
Árvore de decisão J48 25K ordenado otimizado novo atributo	96.4414%	0.98	238	0.9516%	652	2.6070%
Árvore de decisão J48 1M ordenado otimizado novo atributo	99.9952%	157.8	2	0.0002%	46	0.0046%

6. Conclusões e Aprendizados

Aprendemos que o pré-processamento é importante para o tratamento de dados. Tivemos os exemplos prático de muito proveito com simples ordenação e criação de um atributo agregador de informação. Para problemas reais isso é útil ao invés de simplesmente tentar ajustar o algoritmo.

No caso de arvore de decisão, a análise de raridade de algumas instancias e o modo de poda da arvore foi essencial para conseguir bons resultados.

O quão bem um algoritmo escala tanto com pré-processamento e com número de instancias também varia, assim como seu tempo para criação de modelo. Portanto, para problemas reais onde o numero de instancias pode ser enorme e o tempo para criação do modelo crítico a análise de tempo de criação do modelo faz se essencial.