

Comparação de técnicas de aprendizado de máquina utilizando o WEKA

Amilton Fontoura de Camargo Junior¹, Kauane Larisse de Oliveira Benitis¹

¹Curso de Engenharia de Computação – Universidade Tecnológica Federal do Paraná (UTFPR)
CEP – 86.300-000 – Cornélio Procopio – PR – Brasil

amiltonjunior2007@gmail.com, kauane@alunos.utfpr.edu.br

Abstract. *This paper presents a comparison study of supervised and unsupervised machine learning techniques using 10 different data sets in the open source WEKA tool for data mining and a new machine learning proposal.*

Resumo. *Esse trabalho apresenta um estudo de comparação de técnicas de aprendizado de máquinas supervisionado e não supervisionado, utilizando 10 diferentes conjuntos de dados na ferramenta WEKA de código aberto, para mineração de dados e uma nova proposta de aprendizado de máquina.*

1. Introdução

Inteligência artificial é um termo abrangente que engloba o Aprendizado de Máquina (AM), que é o conjunto de técnicas e ferramentas que permitem que o computador “pense” criando algoritmos matemáticos baseados em dados acumulados. Também sob esse termo genérico está o Aprendizado Profundo (AP), um subconjunto do aprendizado de máquina que utiliza modelos de redes neurais para executar tarefas como reconhecimento de imagem e processamento de linguagem.

A inteligência artificial propicia diversas melhorias em áreas distintas. Ao medir a biometria no esporte, os dados podem ajudar a avaliar de que modo o tempo de atividade dos atletas afeta a probabilidade de lesões. Ajudam os agricultores a saber quando regar as plantações para obter colheitas excelentes, permitem que os meteorologistas calculem o degelo [Mannila 1996]. Cidades inteligentes utilizam dados para o gerenciamento de energia, profissionais de saúde recorrem à inteligência artificial para detectar doenças, realizar sequenciamento genômico e acompanhar tratamentos.

2. Técnicas e Ferramentas Utilizadas

2.1. Introdução

O estudo das técnicas de aprendizado de máquina foi dividido em 5 partes. A primeira parte consistiu em:

- Selecionar 10 conjuntos de dados;
- Fazer uma tabela com as informações dos conjuntos de dados;
- Descrever sucintamente o conjunto de dados;
- Construir um programa que realizava a divisão do conjunto de dados;

A segunda parte do estudo foi para realizar experimentos e analisar resultados aplicando as técnicas de aprendizado supervisionado, vistos na sequência do trabalho.

A aplicação da técnica não supervisionada deu-se na terceira etapa, onde foi utilizado o algoritmo k-means.

A proposta da quarta parte do estudo foi descrever uma nova estratégia de aprendizado, de forma a selecionar amostras mais representativas e informativas ao processo de aprendizado do classificador.

2.2. O ambiente WEKA

O WEKA é um software de código aberto implementado por um grupo de aprendizado de máquina da Universidade de Waikato, que atualmente fornece algoritmos que podem ser facilmente aplicados a grandes conjuntos de dados. WEKA implementa várias classificações de aprendizado de máquina, algoritmos de regressão e agrupamento, juntamente com uma série de ferramentas de visualização. Hoje [Mallios et al. 2011] em dia, é aceito como um dos ambiente mais poderosos e adequados para a mineração de dados. Sua interface principal possibilita que o usuário escolha uma das quatro aplicações disponíveis, isto é, o Explorador, o Experimentador, o Fluxo de Conhecimento e uma Interface de Linha de Comando.

Ao longo desse estudo, todos os dados analisados estavam no formato ARFF (Arquivo de Relação Formato de Atributo), que é o formato de arquivo que o WEKA utiliza. O conjunto de dados escolhidos para o desenvolvimento desse estudo deu-se de maneira aleatória, porém variando entre os conjuntos o número de amostras, classes e atributos.

2.3. Técnicas

A comparação realizada e apresentada nesse trabalho engloba seis tipos de técnicas de aprendizagem, descritas a seguir.

2.4. Divisão

A divisão consistiu em separar o conjunto de dados de maneira a obter uma parte para teste e outra parte para treinamento, podendo ser escolhida a porcentagem para cada fim.

Split - Dados

Split de Dados

Arquivo de entrada:

Arquivos de saída

Treinamento:

Teste:

Divisão:

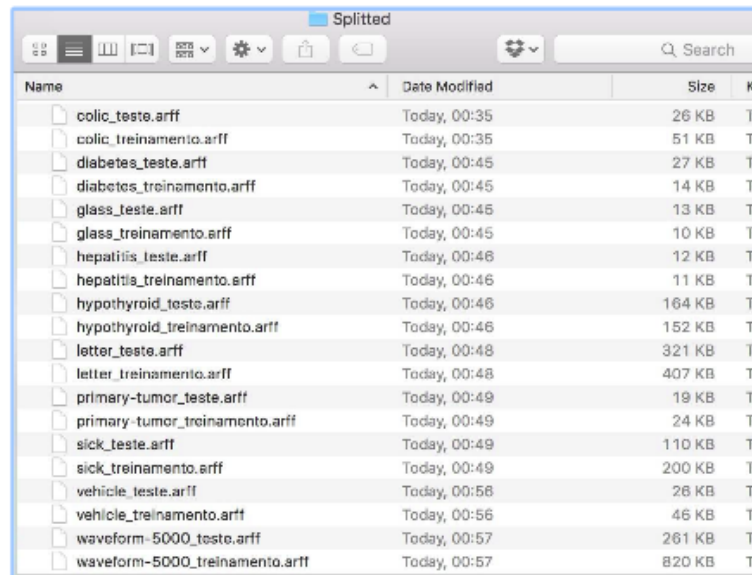
Treinamento: 20 % e Teste: 80 %

☒ Escolher as amostras aleatoriamente

Figura 1. Interface do programa para divisão do conjunto de dados

A interface do programa utilizado para realizar a divisão do conjunto pode ser visualizada na Figura 1. Os conjuntos selecionados na divisão, eram particionados em

doid subconjuntos, um para teste e outro para treinamento, ambos eram salvos na mesmo local do conjunto original como pode ser visto na figura 2.



Name	Date Modified	Size	Kind
colic_teste.arff	Today, 00:35	26 KB	Te
colic_treinamento.arff	Today, 00:35	51 KB	Te
diabetes_teste.arff	Today, 00:45	27 KB	Te
diabetes_treinamento.arff	Today, 00:45	14 KB	Te
glass_teste.arff	Today, 00:45	13 KB	Te
glass_treinamento.arff	Today, 00:45	10 KB	Te
hepatitis_teste.arff	Today, 00:46	12 KB	Te
hepatitis_treinamento.arff	Today, 00:46	11 KB	Te
hypothyroid_teste.arff	Today, 00:46	164 KB	Te
hypothyroid_treinamento.arff	Today, 00:46	152 KB	Te
letter_teste.arff	Today, 00:48	321 KB	Te
letter_treinamento.arff	Today, 00:48	407 KB	Te
primary-tumor_teste.arff	Today, 00:49	19 KB	Te
primary-tumor_treinamento.arff	Today, 00:49	24 KB	Te
sick_teste.arff	Today, 00:49	110 KB	Te
sick_treinamento.arff	Today, 00:49	200 KB	Te
vehicle_teste.arff	Today, 00:56	28 KB	Te
vehicle_treinamento.arff	Today, 00:56	46 KB	Te
waveform-5000_teste.arff	Today, 00:57	261 KB	Te
waveform-5000_treinamento.arff	Today, 00:57	820 KB	Te

Figura 2. Conjuntos de Treinamento e Testes

2.4.1. Validação Cruzada

A validação cruzada é um processo de aprendizagem supervisionada em mineração de dados onde o conjunto é dividido em duas partes, sendo uma para teste e outra para treinamento.

Numa primeira etapa um algoritmo de indução de conhecimento é aplicado à base de treinamento. Na segunda etapa o modelo obtido é aplicado ao fragmento da base de dados que foi dividido como teste. Como a base de testes é previamente rotulada torna-se possível medir a taxa de acerto do modelo.

2.4.2. Naive Bayes

Naive Bayes é um classificador bayesiano e produz regras probabilísticas. Quando um novo item de dados é apresentado, categoriza-o, apresentando um porcentagem de probabilidade em direção as possíveis categorias de classes [Mallios et al. 2011]. A classificação é realizado quando a regra de Bayes bem conhecida é aplicada em cada atributo do modelo e a probabilidade sobre uma variável de classe independente é calculado.

2.4.3. Árvore de Decisão - J48

Entre as várias técnicas de aprendizado de máquina, a árvore de decisão pode ser caracterizada como uma das mais usadas. Representa o mapeamento dos atributos e consiste em

nó que ligam duas ou mais subárvores. O nó calcula um resultado específico baseado no valor da instância e cada possível resultado está ligado com uma das subárvores. O algoritmo J48, uma implementação do C4.5 versão 8, é um método eficiente para estimação e classificação de dados fuzzy [Mallios et al. 2011].

2.4.4. Rede Neural - MultiLayer Perceptron

A Rede Neural é um sistema adaptativo que muda sua estrutura baseado em informações externas e internas através de uma rede durante uma fase de aprendizado inicial. Em termos mais práticos, as redes neurais são estatísticas não lineares de modelagem de dados. Podem ser usadas em modelos complexos de relacionamentos entre entradas e saídas ou para encontrar padrões de dados [Mallios et al. 2011].

2.4.5. k-Vizinhos Mais Próximos - IBk

Uma das formas mais simples de algoritmos de classificação é a implementação dos Vizinhos Mais Próximos. Tais esquemas de aprendizado são delineados como algoritmos de aprendizagem estatística e são gerados simplesmente por dados armazenados. Para que a classificação seja uma métrica de distância é escolhido um dado qualquer e comparado com itens de dados "memorizados" totalmente prontos [Rezende 2003]. O novo item é atribuído à classe que é mais comum entre seus k-vizinhos mais próximos classificados. O número de vizinhos mais próximos(k) podem ser atribuídos de maneira manual ou determinados automaticamente usando validação cruzada.

2.4.6. kMeans

O k-means é um algoritmo de mineração de dados não supervisionado, ou seja, fornece uma classificação de informações de acordo com os próprios dados [Mannila 1996]. Esta classificação é baseada em análise e comparações entre valores numéricos dos dados. Desta maneira, o algoritmo automaticamente vai fornecer uma classificação sem a necessidade de nenhuma supervisão humana, isto é, sem nenhuma pré-classificação já existente.

O algoritmo vai analisar todos os dados e criar classificações, de modo que uma classe(cluster) será indicada para dizer quais linhas pertencem a esta classe [Alsharif and Philip 2016]. Conforme a quantidade de classes desejadas, este número deve ser passado para o algoritmo.

2.5. Conjunto de Dados

Foram escolhidos 10 conjunto de dados de maneira aleatória, procurando selecionar conjuntos em que diversificassem o número de amostras, números de classes e números de atributos.

O conjunto de dados selecionados pode ser visualizado na tabela 1, juntamente com suas características utilizadas para aplicar as técnicas de aprendizado de máquina.

Tabela 1. Conjunto de Dados

Nome do Conjunto	Número de Amostras	Número de Classes	Número de Atributos
Colic	63983	368	23
Dermatology	32417	366	35
Diabetes	37419	768	9
Glass	17823	214	10
Hepatitis	17135	155	20
Hypothyroid	310897	3772	30
Iris	7486	150	4
Primary Tumor	34090	339	18
Vehicle	63838	4	18
Tae	4120	151	6

2.5.1. Descrição dos Conjunto de Dados

A seguir serão descritos de maneira breve cada conjunto de dados selecionado para realizar o estudo atual.

Colic

Esse conjunto é uma base de dados de cólica de cavalo, onde foram separados 300 instâncias para treinamento e 68 para testes. Apresenta tipos de dados contínuos, discretos e nominais e 28 atributos.

Dermatology

Este banco de dados contém 34 atributos, dos quais 33 são valores lineares e 1 nominal. Todos eles compartilham características de eristemas (clínicas e histopatológicas).

Diabetes

Base de dados de diabetes onde foi usado 576 intâncias para treinamento e 192 para testes por meio do algoritmo ADAP para servir de variável diagnóstica de pacientes que apresentam sinais de diabetes.

Glass

Glass é uma base de dados de 214 intanciãs onde foi empregado o algoritmo do vizinho mais próximo e análise discriminante para determinar se o vidro era do tipo flutuante, onde classifica-se por exemplo se o vidro em questão pertence a uma janela de casa ou de carro.

Hepatitis

Este banco de dados contem características apresentadas em portadores de hepatite como método computarizado para realizar diagnóstido de hepatite. As informações apresentadas nas 155 instâncias estão relacionadas com 32 indivíduos que já entraram em óbito e 123 vivos.

Hypothyroid

Base de dados compostas por 3272 instâncias cujos atributos permitem identificar

a hipotireóide primária, compensada, negativo e hipotireóide conforme os atributos.

Iris

O conjunto de dados contem 150 instancias onde 3 classes de 50 instancias se refere a um tipo da planta Iris.

Primary Tumor

Primary Tumor contem 339 instâncias e 18 atributos onde um dos atributos é a classe de localização de tumor.

Vehicle

Conjunto de dados que usa um conjunto de recursos da silhueta para classificar conforme os quatro tipos de veículos do banco.

Tae

Base de dados para avaliação de assistente de ensino onde os dados consistem em avaliações de desempenho de ensino em três semestres regulares e dois meses de verão caracterizados em baixo, medio e alto para formar a variável de classe

3. Análise e Resultados

Aqui serão apresentados os resultados obtidos nos experimentos realizados com o WEKA, mostrando o custo de cada conjunto de dados, acurácia de cada conjunto de dados, média e desvio padrão de custo de cada técnica por conjunto, média e desvio padrão de custo de conjunto considerando todas as técnicas, acurácia de cada técnicas por conjunto, acurácia de conjunto considerando todas as técnicas.

Esse dados serão mostrados no trabalho por meio de gráficos e tabelas no decorrer do trabalho, assim como uma análise sobre os mesmos.

Tabela 2. Média e Desvio Padrão dos Conjuntos de Dados

Nome do Conjunto	Média de Acurácia	DP de Acurácia	Média de Custo	DP de Custo
Colic	79,11	15,53	0,71	1,94
Dermatology	72,35	34,63	2,97	8,32
Diabetes	77,55	14,52	0,13	0,20
Glass	81,09	27,11	0,05	0,13
Hepatitis	85,73	8,86	0,05	0,11
Hypothyroid	96,74	3,58	4,66	7,78
Iris	73,67	33,43	0,01	0,03
Primary Tumor	50,33	31,95	0,96	1,72
Vehicle	58,81	32,03	0,43	0,73
Tae	58,39	26,36	0,02	0,03

Na tabela 2 pode-se observar que o conjunto Hypothyroid obteve melhor média de acurácia e menor desvio padrão, no entanto, também foi o que apresentou maior custo. Também na tabela 2 verificou-se que a base dados Primary Tumor teve a menor média de acurácia, assim como não houve média inferior à 50% em nenhum outro conjunto.

3.1. Acurácia das técnicas

Os resultados das acurácias obtidos de cada técnica utilizada no estudo mostrados na tabela 3 estão apresentado a seguir.

Tabela 3. Acurácia								
Conjunto	CV	NB	J48	MLP	IBk1	IBk2	kMeans1	kMeans2
Colic	63,04	77,99	85,33	80,43	63,04	63,04	100,00	100,00
Dermatology	30,60	97,27	93,99	96,17	30,60	30,60	99,80	99,80
Diabetes	65,10	76,30	73,83	75,39	65,10	65,10	99,80	99,80
Glass	35,51	48,60	66,82	67,76	35,51	35,51	99,50	99,50
Hepatitis	79,35	84,52	83,87	80,00	79,35	79,35	99,70	99,70
Hypothyroid	92,29	95,28	99,58	94,17	92,29	92,29	100,00	100
Iris	33,33	96,00	96,00	97,33	33,33	33,33	100,00	100,00
Primary Tumor	24,78	50,15	39,82	38,35	24,78	24,78	100,00	100,00
Vehicle	25,65	44,80	72,46	81,68	25,65	25,65	94,60	100,00
Tae	34,44	54,30	59,60	54,30	34,44	34,44	97,80	97,80
Média	48,41	72,52	77,13	76,56	48,41	48,41	99,12	99,66
Desvio Padrão	24,40	21,19	18,54	18,90	24,40	24,40	1,72	0,68

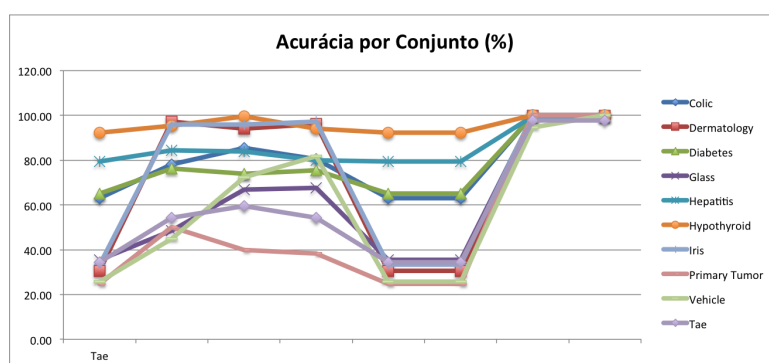


Figura 3. Gráfico da Acurácia de Cada Conjunto de Dados

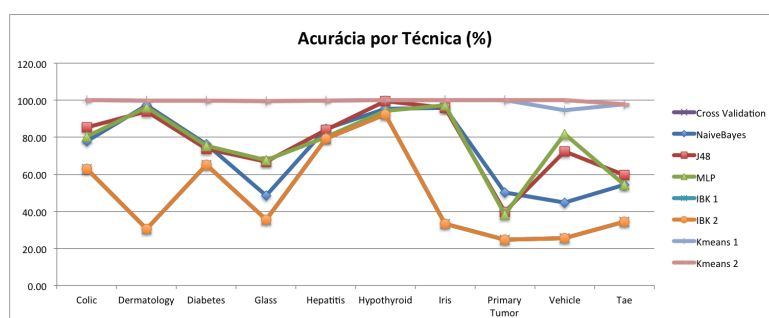


Figura 4. Gráfico da Acurácia das Técnicas de Aprendizado

A tabela 3 juntamente com as figuras 3 e figura 4 mostraram que independente da técnica de aprendizado de máquina utilizada, a base de dados Hypothyroid foi a que apresentou os melhores valores. Das técnicas supervisionadas, as que mostraram melhor acurácia foram a J48 e a MultiLayer Perceptron, e, as que demonstraram menor eficiência no aprendizado foram a Cross Validation e IBk.

3.2. Custos das técnicas

Os custos obtidos de cada técnica utilizada no estudo mostrados na tabela 4 estão dispostos na sequência.

Tabela 4. Custo								
Conjunto	CV	NB	J48	MLP	IBk1	IBk2	kMeans1	kMeans2
Colic	0,00	0,01	0,01	5,51	0,00	0,00	0,11	0,07
Dermatology	0,00	0,00	0,00	23,56	0,00	0,00	0,11	0,11
Diabetes	0,00	0,00	0,02	0,57	0,00	0,00	0,27	0,15
Glass	0,00	0,00	0,01	0,38	0,00	0,00	0,01	0,02
Hepatitis	0,00	0,00	0,00	0,32	0,00	0,00	0,01	0,02
Hypothyroid	0,00	0,01	0,02	21,92	0,00	0,00	7,78	7,51
Iris	0,00	0,00	0,00	0,09	0,00	0,00	0,01	0,01
Primary Tumor	0,00	0,00	3,74	3,76	0,00	0,00	0,11	0,06
Vehicle	0,00	0,00	0,10	2,12	0,00	0,00	0,60	0,62
Tae	0,00	0,00	0,00	0,10	0,00	0,00	0,10	0,10
Média	0,00	0,00	0,39	5,83	0,00	0,00	0,91	0,86
Desvio Padrão	0,00	0,00	1,18	9,10	0,00	0,00	2,42	2,34

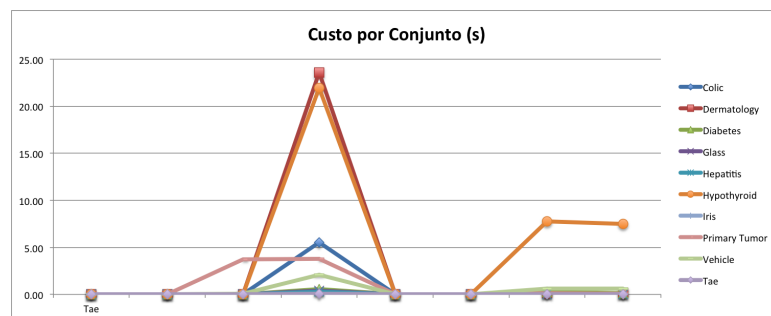


Figura 5. Gráfico do Custo de Cada Conjunto de Dados

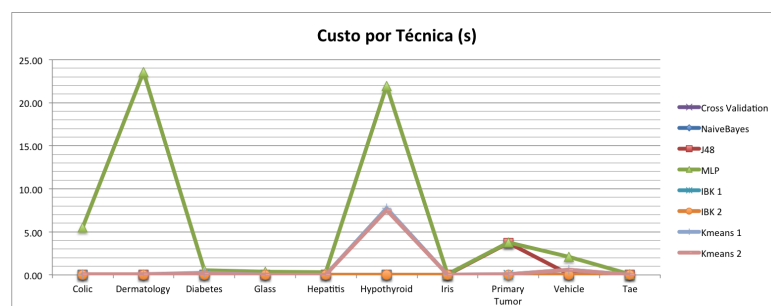


Figura 6. Gráfico do Custo das Técnicas de Aprendizado

A tabela 4 juntamente com os gráficos da figura 5 e figura 6 mostram que praticamente todas as técnicas são rápidas, exceto MultiLayer Perceptron e kMeans.

3.3. Split de Dados

Como visto em 2.4, foi usado um software para realizar a divisão do conjunto para realização de teste e treinamento. O programa permite selecionar a porcentagem que se deseja para teste e para treinamento. Ao selecionar aleatoriamente, o conjunto é separado de forma aleatória, caso contrário, a divisão é realizada de maneira a coletar-se informações de maneira sequencial.

O conjunto selecionado para estudo do Split deu-se pelo fato de o conjunto apresentar o maior número de amostras, maior acurácia apresentada na tabela 3 e um dos maiores custos apresentados na tabela 4.

Os dados referentes ao split do conjunto de dados *Hypothyroid* estão divididos 80% para treinamento e 20% para teste.

A seguir, será mostrado as matrizes de confusão das técnicas J48, NayveBayes, MLP, e IBk, onde, coincidentemente, as acurácias de treinamento e teste corresponderam a 99,80% e 99,47% e o custo das técnicas não extrapolaram 0,04 segundos.

Técnica: J48 - Training					
Conjunto: hypothyroid.arff		Classe: Class	Amostras: 3013		
↓ Real	Escolhido →	negative	compensated_hypothyroid	primary_hypothyroid	secondary_hypothyroid
negative		91.84%	0.07%	0.00%	0.00%
compensated_hypothyroid		0.00%	5.54%	0.00%	0.00%
primary_hypothyroid		0.03%	0.07%	2.42%	0.00%
secondary_hypothyroid		0.03%	0.00%	0.00%	0.00%

Figura 7. Matriz de Confusão J48

Técnica: J48 - Test					
Conjunto: hypothyroid.arff		Classe: Class	Amostras: 755		
↓ Real	Escolhido →	negative	compensated_hypothyroid	primary_hypothyroid	secondary_hypothyroid
negative		93.77%	0.00%	0.00%	0.00%
compensated_hypothyroid		0.00%	3.58%	0.00%	0.00%
primary_hypothyroid		0.26%	0.13%	2.12%	0.00%
secondary_hypothyroid		0.13%	0.00%	0.00%	0.00%

Figura 8. Matriz de Confusão J48

Técnica: NaiveBayes - Training					
Conjunto: hypothyroid.arff		Classe: Class	Amostras: 3013		
↓ Real	Escolhido →	negative	compensated_hypothyroid	primary_hypothyroid	secondary_hypothyroid
negative		91.11%	0.50%	0.27%	0.03%
compensated_hypothyroid		3.62%	1.86%	0.07%	0.00%
primary_hypothyroid		0.27%	0.20%	2.06%	0.00%
secondary_hypothyroid		0.00%	0.00%	0.00%	0.03%

Figura 9. Matriz de Confusão NayveBayes

Técnica: NaiveBayes - Test					
Conjunto: hypothyroid.arff		Classe: Class	Amostras: 755		
↓ Real	Escolhido →	negative	compensated_hypothyroid	primary_hypothyroid	secondary_hypothyroid
negative		93.11%	0.13%	0.07%	0.07%
compensated_hypothyroid		0.46%	1.72%	0.00%	0.00%
primary_hypothyroid		0.00%	0.26%	2.25%	0.00%
secondary_hypothyroid		0.13%	0.00%	0.00%	0.00%

Figura 10. Matriz de Confusão NayveBayes

Técnica: MLP - Training				
Conjunto: hypothyroid.arff	Classe: Class	Amostras: 3013		
↓ Real Escolhido →	negative	compensated_hypothyroid	primary_hypothyroid	secondary_hypothyroid
negative	91.54%	0.27%	0.10%	0.00%
compensated_hypothyroid	3.09%	1.99%	0.46%	0.00%
primary_hypothyroid	0.07%	0.07%	2.39%	0.00%
secondary_hypothyroid	0.00%	0.00%	0.03%	0.00%

Figura 11. Matriz de Confusão MLP

Técnica: MLP - Test				
Conjunto: hypothyroid.arff	Classe: Class	Amostras: 755		
↓ Real Escolhido →	negative	compensated_hypothyroid	primary_hypothyroid	secondary_hypothyroid
negative	92.19%	1.32%	0.26%	0.00%
compensated_hypothyroid	2.25%	0.93%	0.40%	0.00%
primary_hypothyroid	0.00%	0.13%	2.38%	0.00%
secondary_hypothyroid	0.00%	0.00%	0.13%	0.00%

Figura 12. Matriz de Confusão MLP

Técnica: IBk k = 2 - Test				
Conjunto: hypothyroid.arff	Classe: Class	Amostras: 755		
↓ Real Escolhido →	negative	compensated_hypothyroid	primary_hypothyroid	secondary_hypothyroid
negative	93.38%	0.40%	0.00%	0.00%
compensated_hypothyroid	3.31%	0.26%	0.00%	0.00%
primary_hypothyroid	1.46%	0.53%	0.53%	0.00%
secondary_hypothyroid	0.13%	0.00%	0.00%	0.00%

Figura 13. Matriz de Confusão IBk = 2

Técnica: IBk k = 2 - Training				
Conjunto: hypothyroid.arff	Classe: Class	Amostras: 3013		
↓ Real Escolhido →	negative	compensated_hypothyroid	primary_hypothyroid	secondary_hypothyroid
negative	91.90%	0.00%	0.00%	0.00%
compensated_hypothyroid	4.38%	1.16%	0.00%	0.00%
primary_hypothyroid	0.80%	0.27%	1.46%	0.00%
secondary_hypothyroid	0.00%	0.03%	0.00%	0.00%

Figura 14. Matriz de Confusão IBk = 2

Técnica: IBk k = 5 - Training				
Conjunto: hypothyroid.arff	Classe: Class	Amostras: 3013		
↓ Real Escolhido →	negative	compensated_hypothyroid	primary_hypothyroid	secondary_hypothyroid
negative	91.77%	0.13%	0.00%	0.00%
compensated_hypothyroid	4.61%	0.90%	0.03%	0.00%
primary_hypothyroid	1.16%	0.17%	1.19%	0.00%
secondary_hypothyroid	0.03%	0.00%	0.00%	0.00%

Figura 15. Matriz de Confusão IBk = 5

Técnica: IBk k = 5 - Test				
Conjunto: hypothyroid.arff	Classe: Class	Amostras: 755		
↓ Real Escolhido →	negative	compensated_hypothyroid	primary_hypothyroid	secondary_hypothyroid
negative	93.25%	0.53%	0.00%	0.00%
compensated_hypothyroid	3.05%	0.53%	0.00%	0.00%
primary_hypothyroid	1.59%	0.00%	0.93%	0.00%
secondary_hypothyroid	0.13%	0.00%	0.00%	0.00%

Figura 16. Matriz de Confusão IBk = 5

4. Trabalho Correlato

Nessa seção será apresentado o trabalho *Lexicographic preferences for predictive modeling of human decision making: A new machine learning method with an application in accounting*[Bräuning et al. 2016] onde aparece uma nova estratégia de aprendizado.

O algoritmo para expansão dos pares de escolha mais prováveis pode ser visualizado na figura 19.

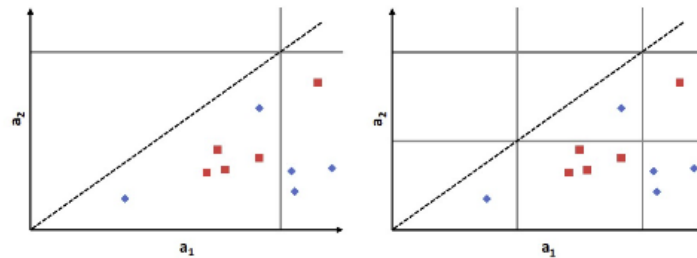


Figura 17. Agrupamento/Expansão dos Atributos de Decisão

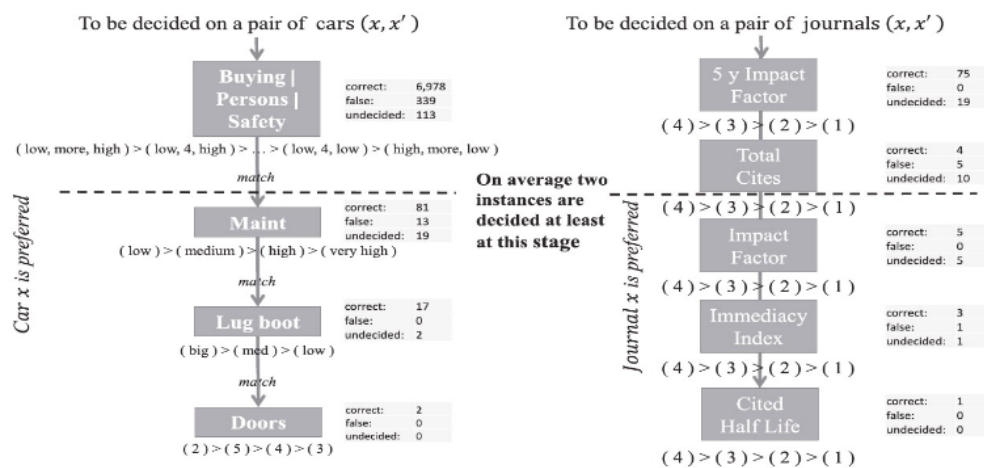


Figura 18. Decisão

De acordo com os autores, uma estratégia de tomada de decisão humana é a heurística "Escolher a Melhor", sem integrar todas as informações, apenas de forma "competitiva" escolher a aparentemente melhor. Outra estratégia é o modelo lexicográfico, que é comparar duas opções, verificando um critério entre eles que difere e usando o mesmo como fator de decisão, sem considerar os demais.

Ao invés de comparar cada atributo separadamente, faz a comparação de todos eles em simultâneo, aproximando-se da heurística humana. Outra diferença proposta pelos autores é um método para discretizar os domínios de atributos numéricos, para análise par a par, facilitando a aplicação da heurística proposta.

Em um estudo de caso, o artigo investiga um problema importante e altamente complexo de decisão do mundo real com relação à elaboração de relatórios financeiros sobre modelo de pensões profissionais (contribuinte/não-contribuinte, segurado/não-segurado, etc.)

Da mesma forma que outras decisões de negócios, a administração da empresa tem que considerar muitos fatores (atributos) ao decidir sobre um método, e pode ser uma boa alternativa aplicar uma heurística lexicográfica.

Algorithm 1: LPLL.

Input : training data T , set of attributes V , maximal grouping size g_{\max} , maximum partition size r_{\max}
Output: LP list l

$l \leftarrow \emptyset, V' \leftarrow V, \mathcal{I}' \leftarrow \{1, \dots, n\}$
while $\mathcal{T} \neq \emptyset$ and $V' \neq \emptyset$ **do**
 $l' \leftarrow \emptyset, CR \leftarrow 0, CP \leftarrow 0$
 for $\mathcal{I} \subseteq \mathcal{I}', |\mathcal{I}| \leq g_{\max}$ **do**
 discretize $A_i, i \in \mathcal{I}$ into at most r_{\max} bins
 determine $\sqsupset_{\mathcal{I}}$ on $\mathcal{D}(V_{\mathcal{I}})$ maximally consistent with T
 compute $CR(\sqsupset_{\mathcal{I}}, T)$ and $CP(\sqsupset_{\mathcal{I}}, T)$
 if $CR(\sqsupset_{\mathcal{I}}, T) = CR$ && $CP < CP(\sqsupset_{\mathcal{I}}, T)$ **then**
 $CP \leftarrow CP(\sqsupset_{\mathcal{I}}, T)$
 $l' \leftarrow \mathcal{I}$
 if $CR(\sqsupset_{\mathcal{I}}, T) > CR$ **then**
 $CR \leftarrow CR(\sqsupset_{\mathcal{I}}, T)$
 $CP \leftarrow CP(\sqsupset_{\mathcal{I}}, T)$
 $l' \leftarrow \mathcal{I}$
 reverse discretization for $A_i, i \in \mathcal{I}$
 discretize $A_i, i \in l'$ into at most r_{\max} bins
 $\mathcal{I}' \leftarrow \mathcal{I}' \setminus l'$
 $V' \leftarrow V' \setminus V_{l'}$
 remove every $(\vec{x}, \vec{x}') \in \mathcal{T}$ decided by $\sqsupset_{l'}$
 add item $(V_{l'}, \sqsupset_{l'})$ to l

Figura 19. Algoritmo LPLL

Possui Lista de Preferências Lexicográficas (LPs) com as alternativas a serem consideradas, onde o algoritmo faz o aprendizado destas listas para, então, tomar as decisões.

Um exemplo pode ser visto na figura 18 onde ocorre a decisão entre um par de carros e um par de artigos, considerando sempre dois atributos.

O Agrupamento dos atributos de decisão pode ser observado na figura 17 onde à esquerda existe apenas uma divisão e, na direita, duas divisões, onde cada divisão possui elementos "puros", sendo eles do tipo vermelho (negativo) ou do tipo azul (positivo).

$$x \in X = D(V) = D(A_1) \times \dots \times D(A_n) \quad (1)$$

Onde V são as alternativas e D os domínios dos atributos.

$$\tau = (x_1^*, x_i)_{(i=1)}^N \quad (2)$$

5. Conclusão

O estudo permitiu verificar que o aprendizado supervisionado possui maior acurácia, por terem especificadas as classes, assim como observar que o aprendizado não-supervisionado é melhor quando não se há muita informação a respeito do conjunto de dados a ser analisado.

Após empregar a divisão para treinamento e teste dos conjuntos de dados, todas as técnicas obtiveram resultados superiores a 90%, ou seja, aumento de precisão nos dados e maior confiabilidade.

Mais atributos/dimensões resultam numa melhor classificação, porém deve-se atentar à maneira com que seleciona-se tais dados. Técnicas de aprendizado ativo não foram citadas, mas podem fazer muita diferença quando os conjuntos de dados são muito grandes.

Se o conjunto de dados é muito grande, o computador que está executando o experimento deve conter muita memória RAM! As técnicas de aprendizado facilitam na quantidade das classificações, mas têm um grande custo computacional.

Referências

- Alsharif, A. H. and Philip, N. (2016). Data mining technique for the enhanced smoking cessation management system (smoke mind). In *2016 International Conference on Engineering MIS (ICEMIS)*, pages 1–1.
- Bräuning, M., Hüllermeier, E., Keller, T., and Glaum, M. (2016). Lexicographic preferences for predictive modeling of human decision making: A new machine learning method with an application in accounting. *European Journal of Operational Research*.
- Mallios, N., Papageorgiou, E., and Samarinas, M. (2011). Comparison of machine learning techniques using the weka environment for prostate cancer therapy plan. In *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2011 20th IEEE International Workshops on*, pages 151–155.
- Mannila, H. (1996). Data mining: machine learning, statistics, and databases. In *Scientific and Statistical Database Systems, 1996. Proceedings., Eighth International Conference on*, pages 2–9.
- Rezende, S. O. (2003). *Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda.