PUC

ISSN 0103-9741

Monografias em Ciência da Computação
n° 38/09

Estudo sobre Algoritmos de Classificação para o Referenciamento de Gestantes de Alto-risco

Ingrid Oliveira de Nunes
Dárlinton Barbosa Feres Carvalho
Carlos José Pereira de Lucena

Departamento de Informática

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO
RUA MARQUÊS DE SÃO VICENTE, 225 - CEP 22451-900
RIO DE JANEIRO - BRASIL

Monografias em Ciência da Computação, No. 38/09 ISSN: 0103-9741
Editor: Prof. Carlos José Pereira de Lucena Dezembro, 2009

Estudo sobre Algoritmos de Classificação para o Referenciamento de Gestantes de Alto-risco [1]

Ingrid Oliveira de Nunes, Dárlinton Barbosa Feres Carvalho, Carlos José Pereira de Lucena

{ionunes,dcarvalho,lucena}@inf.puc-rio.br

Resumo. O Sistema Unificado de Assistência Pré-natal (SUAP) é um sistema em desenvolvimento que visa dar suporte ao atendimento pré-natal. Uma de suas funcionalidades apóia o processo de referenciamento de gestantes de alto-risco através da indicação da unidade para a qual a gestante deve ser referanciada de acordo com sua complicação e localização. Ambiciona-se fazer uso de casos históricos e algoritmos de aprendizado de máquina para tal funcionalidade. Dessa forma, visa-se neste trabalho fazer um estudo exploratório de algoritmos de aprendizado de máquina e ferramentas que possam ser incorporadas no SUAP para resolver o problema. Dado que o sistema ainda não possui um conjunto de dados representativo, foram utilizados *datasets* com características similares. Concluiu-se que a ferramenta Weka é apropriada para a incorporação ao SUAP, e pode ser facilmente parametrizada para a escolha do algoritmo que seja melhor adequado ao problema em questão.

Palavras-chave: Sistemas de Saúde, Atendimento Pré-natal, Aprendizado de Máquina, Algoritmos, Processo de Referenciamento.

Abstract. The Prenatal Care Unified System (SUAP) is a system under development whose aim is to support the prenatal care. One of its functionalities is to support the referral process of high-risk pregnancy by the indication of which unit a pregnant must be referred to according to her complications and location. Our goal is to use historical cases and machine learning algorithms for such functionality. Thus, in this work we aim at making an exploratory study of machine learning algorithms and tools that may be incorporated in the SUAP to solve our problem. Given that the system does not have a representative dataset yet, we used datasets with similar characteristics. We concluded that the Weka tool is appropriate to be incorporated in the SUAP, and it can be easily parametrized to choose an algorithm that is more adequate to the target problem.

Keywords: Healthcare Systems, Prenatal Care, Machine Learning, Algorithms and Referral Process.

[1] Trabalho patrocinado pelo Ministério de Ciência e Tecnologia da Presidência da República Federativa do Brasil e FINEP.

Responsável por publicações:
Rosane Teles Lins Castilho
Assessoria de Biblioteca, Documentação e Informação
PUC-Rio Departamento de Informática
Rua Marquês de São Vicente, 225 - Gávea
22451-900 Rio de Janeiro RJ Brasil
Tel. +55 21 3527-1516 Fax: +55 21 3527-1530
E-mail: bib-di@inf.puc-rio.br
Web site: http://bib-di.inf.puc-rio.br/techreports/

Contents

1 Introdução **1**

2 Aprendizado no Sistema Unificado de Assistência Pré-natal (SUAP) – Referenciamento de Gestantes de Alto-risco **1**
2.1 Definição do Problema 3
2.2 Objetivos 3

3 *Datasets* **3**
3.1 *Post-Operative Patient Data Set* 3
3.1.1 Atributos do *dataset* 3
3.1.2 Estado da Arte 4
3.2 *Contraceptive Method Choice Data Set* 5
3.2.1 Atributos do *dataset* 5
3.2.2 Estado da Arte 6
3.3 *Adult Data Set* 6
3.3.1 Atributos do *dataset* 6
3.3.2 Estado da Arte 7

4 Metodologia **7**
4.1 Algoritmos 8
4.2 Configuração 8
4.3 Avaliação 9

5 Resultados **9**

6 Conclusão **11**

References **11**

1 Introdução

O SUAP (Carvalho, Choren, Carvalho, Lucena, Condack & de Sá 2009, de Sa, Carvalho, Moraes, Stein, dos Santos, Carvalho & Lucena 2009) é um sistema de software que faz parte do projeto de pesquisa AGENTESGRA financiado pela Financiadora de Estudos e Projetos (FINEP), em desenvolvimento pelo Laboratório de Engenharia de Software (LES) da PUC-Rio em associação com ginecologistas e obstetras do Hospital Universitário Antônio Pedro (HUAP). Ele tem por objetivo apoiar a decisão médica principalmente em problemas relacionados com a hipertensão na gravidez. Além da informatização do acompanhamento pré-natal – que hoje é feito essencialmente de forma manual – o sistema oferece funcionalidades que apresentam um comportamento autônomo, pró-ativo e "inteligente", e para isso faz uso das abstrações de agentes de software (Jennings 2001).

Uma diferença entre a assistência pré-natal e o atendimento de pacientes é que no primeiro não existe doença a ser tratada, parte-se de uma gestante saudável que deve ser acompanhada durante a gravidez para evitar possíveis complicações. Por exemplo, uma gestante que apresente edemas e pressão alta potencialmente pode evoluir para um quadro de eclâmpsia. Assim, identificada uma gravidez de risco, a gestante deve receber um tratamento adequado. O sistema de saúde pública brasileiro possui diversas unidades onde a gestante pode ser atendida. Esse atendimento ocorre inicialmente em unidades primárias, que possuem os recursos necessários para fazer a realização de um pré-natal de baixo risco. Caso uma situação de risco seja identificada, ela deve ser referenciada a uma unidade secundária que ofereça mais condições para tratamento. As unidades diferenciam-se pela localização (por questões de distribuição das gestantes e importante também para as mesmas, visto que as gestantes de baixa renda não possuem recursos para deslocamento) e recursos (tanto materiais como humanos). Assim, de acordo com a situação da gestante, ela deve ser referenciada para uma unidade específica que seja mais adequada.

Neste sentido, uma das funcionalidades do SUAP serve para apoiar a decisão no processo de referenciamento de gestantes. O sistema possui dados sobre as unidades de atendimento e casos históricos. Com base no (in)sucesso de casos históricos, nos quais gestantes foram referenciadas e devidamente tratadas, ele deve sugerir para qual unidade novos casos devem ser referenciados. Assim, neste artigo apresentamos uma solução para o aprendizado de unidades de referenciamento de gestantes de alto-risco. Visto que SUAP ainda encontra-se em fase de desenvolvimento e não existe um *corpus* que possa ser utilizado neste estudo, escolheu-se um problema que seja análogo ao que deve ser resolvido, para que a solução possa ser posteriormente incorporada ao sistema.

O restante deste artigo está organizado como segue. Na Seção 2, detalha-se o problema de referenciamento de gestantes de alto-risco. A Seção 2.2 apresenta os objetivos deste trabalho. A Seção 3 descreve os conjuntos de dados utilizados para o aprendizado de máquina. A Seção 4 apresenta o algoritmo escolhido para a resolução do problema, e os resultados são apresentados na Seção 5. O artigo é concluído na Seção 6.

2 Aprendizado no SUAP – Referenciamento de Gestantes de Alto-risco

Uma gravidez de baixo risco pode ser acompanhada em unidades de saúde que possuam recursos suficientes para coletar os seguintes dados: (i) peso, (ii) altura uterina, (iii) pressão,

(iv) batimentos cardiofetais e (v) presença de edemas. Esses dados são coletados em cada consulta do atendimento pré-natal. Dessa forma, estas unidades, ditas primárias, geralmente não desfrutam de recursos suficientes para lidar com complicações durante a gestação.

Quando não-conformidades são detectadas a partir desses dados, ou através de algum exame, e a gravidez é classificada como de alto risco, e a gestante é referenciada para unidades que possuam uma melhor infra-estrutura, em termos de equipamentos e recursos humanos. Esse referenciamento leva em consideração principalmente o local de residência da gestante, a qual é encaminhada para uma unidade secundária dentro do seu município. Por exemplo, a cidade de Niterói, localizada no estado do Rio de Janeiro, possui dois hospitais, entre eles o HUAP, que atende gestações de alto risco.

Entretanto, podem ocorrer situações especiais a serem consideradas no referenciamento. Em primeiro lugar, considerando-se que existe mais de unidade secundária dentro de um mesmo município, é desejável se constatar qual das unidades é mais apropriada para atender um certo tipo de complicação. Em segundo lugar, em alguns casos a gestante é referenciada para uma unidade secundária, mas quando os médicos desta unidade a avaliam, verificam que o caso pode ser atendido pela unidade primária e a gestante é contra-referenciada. Assim, o ideal é que nesses casos já se conheçam as situações que não precisam de referenciamento. Em terceiro lugar, existem certas complicações que necessitam de recursos mais adequados. Um exemplo é que na cidade do Rio de Janeiro existe um hospital que possui uma infra-estrutura para atender recém-nascidos que nascem com problemas. Dessa forma, mesmo que se tenha uma gestante de Niterói com uma complicação deste tipo, ela será melhor atendida no Instituto Fernandes Figueira da cidade do Rio de Janeiro. Por fim, existem casos em que a gestante reside nos limiares do município, então talvez uma unidade secundária de um município vizinho seja mais viável para ela.

A Figura 1 ilustra uma visão canônica do problema de referenciamento de gestantes de alto-risco. Com base nos casos históricos do atendimento pré-natal, ambiciona-se generalizar as exceções à regra geral (localização) relacionadas com as questões previamente mencionadas.

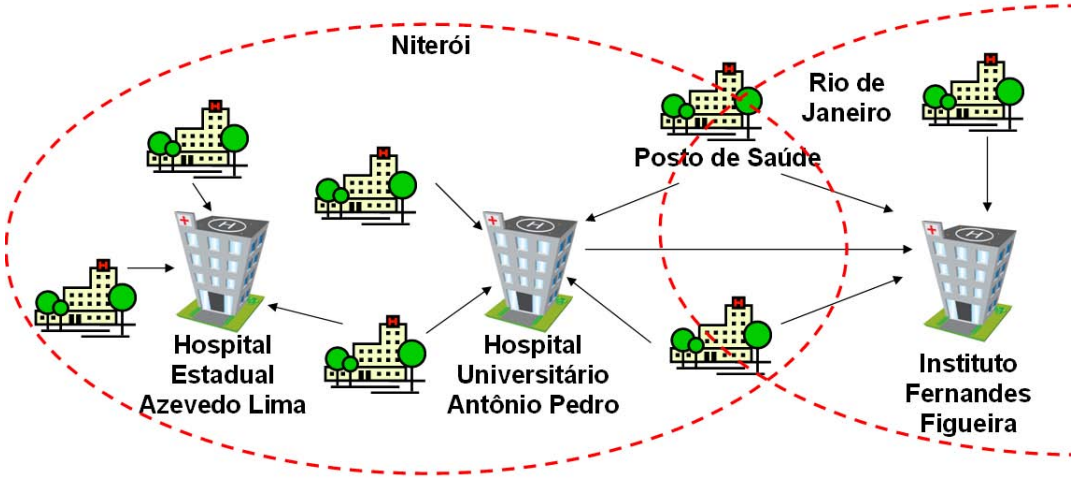

Figure 1: Visão Canônica do Problema de Referenciamento

Na próxima seção, o problema de referenciamento é descrito de forma mais precisa,

através do detalhamento de seus atributos e o tipo de problema de aprendizado de máquina.

2.1 Definição do Problema

O referenciamento de gestantes de alto-risco no SUAP é um problema de classificação, no qual as classes-alvo são os hospitais de referenciamento. A idéia é ter essas informações por cada estado brasileiro de forma independente.

Os atributos são dados relacionados com a gestante. Eles incluem a localização dela (bairro e cidade), além de informações sobre se ela é hipertensa, diabética, HIV positiva, entre outros. Essas informações são derivadas dos dados coletados no atendimento pré-natal e dos resultados dos exames.

Entretanto, visto que o SUAP ainda está em fase de desenvolvimento e não há um *corpus* para realizar o aprendizado de máquina, buscou-se problemas análogos no UCI Machine Learning Repository, a fim de se produzir uma solução que possa ser posteriormente utilizada no SUAP.

2.2 Objetivos

O objetivo deste trabalho é prover uma solução para o problema de referenciamento de gestantes de alto risco do SUAP. Visto que ainda não existe um *corpus* para este problema, serão escolhidos *corpus* que representem problemas de classificação, e que possuam tanto variáveis numéricas como categóricas, tal qual o problema-alvo.

Assim, visa-se a elaboração de um *framework* ou a busca de alguma ferramenta que dê suporte para *datasets* com estas características, para a posterior incorporação ao SUAP.

3 *Datasets*

Foram escolhidos três *datasets* do UCI Machine Learning Repository, ambos da área de saúde, com os pré-requisitos estabelecidos (problema de classificação, com variáveis numéricas e categóricas) para o desenvolvimento do *framework* ou testes da ferramenta selecionada. Um deles é pequeno (90 instâncias), para a realização dos primeiros testes, um médio (1473) e outro grande (48842). Estes *datasets* são descritos nas próximas seções.

3.1 *Post-Operative Patient Data Set*

A tarefa de classificação do conjunto de dados *Post-Operative Patient*[2] (Pacientes em Pós-operatório) é determinar para onde pacientes na área de recuperação pós-operatória devem ser enviados a seguir. Visto que hipotermia é uma preocupação significante depois da cirurgia, os atributos correspondem às medidas da temperatura corporal. Informações sobre o *dataset* são apresentadas na Tabela 1.

3.1.1 Atributos do *dataset*

- L-CORE (temperatura interna do paciente em Celsius): alta (> 37), média (≥ 36 e ≤ 37), baixa (< 36).

[2]Disponível em `http://archive.ics.uci.edu/ml/datasets/Post-Operative+Patient`

Características do *dataset*:	Atributos multi-variados
Número de instâncias:	90
Características dos atributos:	Categóricos, Inteiros
Número de atributos:	8
Tarefas associadas:	Classificação
Valores incompletos?	Sim
Área:	Saúde

Table 1: Informações sobre o *Post-Operative Patient Data Set.*

- L-SURF (temperatura superficial do paciente em Celsius): alta (> 36.5), média (≥ 36.5 e ≤ 35), baixa (< 35).
- L-O2 (saturação do oxigênio em %): excelente (≥ 98), boa (≥ 90 e < 98), razoável (≥ 80 e < 90), fraca (< 80).
- L-BP (última medida da pressão sangüínea): alta ($> 130/90$), média ($\leq 130/90$ e $\geq 90/70$), baixa ($< 90/70$).
- SURF-STBL (estabilidade da temperatura superficial do paciente): estável, meio-estável, instável.
- CORE-STBL (estabilidade da temperatura central do paciente): estável, meio-estável, instável.
- BP-STBL (estabilidade da pressão sangüínea do paciente): estável, meio-estável, instável.
- COMFORT (percepção de conforto do paciente na baixa): medido como um inteiro entre 0 e 20.
- decisão ADM-DECS (decisão de baixa): I (paciente enviado a Unidade de Tratamento Intensivo), S (paciente preparado para ir para casa), A (paciente enviado ao andar geral do hospital).

3.1.2 Estado da Arte

O trabalho de (Budihardjo, Grzymala-Busse & Woolery 1991) foi o primeiro a resolver este *dataset* e atingiu um resultado de 48% de acurácia com a metodologia LERS (LEM2). Este *dataset* também foi utilizado em outros trabalhos (Owen 1999, Kontkanen, Lahtinen, Myllymäki & Tirri 2000), mas com alterações. Por exemplo, em (Owen 1999) a variável de decisão foi reduzida apenas a valores binários, através de exclusão das instâncias com a classe menos freqüente, no total 2 exemplos com a classe I, e das instâncias com valores incompletos, também 2 exemplos. Neste caso, a taxa de acerto alcançada com o método proposto foi de 62.8%, mas também foi reportado que utilizando uma regressão logística global conseguiu-se 69.8% de acertos. O trabalho (Kontkanen et al. 2000) também utiliza este *dataset* na versão completa e atinge uma acurácia de 71.1%.

3.2 *Contraceptive Method Choice Data Set*

O conjunto de dados *Contraceptive Method Choice*[3] (Escolha do Método Contraceptivo) é um subconjunto da Pesquisa Nacional da Prevalência Contraceptiva da Indonésia de 1987. Os exemplos são mulheres casadas que não estavam grávidas ou não sabiam se estavam na época da entrevista.

O problema é predizer o método atual escolhido para contracepção (não uso, método de longo-prazo, método de curto-prazo) de uma mulher, baseado nas suas características demográficas e sócio-econômicas. Informações sobre o *dataset* são apresentadas na Tabela 2.

Características do *dataset*:	Atributos multi-variados
Número de instâncias:	1473
Características dos atributos:	Categóricos, Inteiros
Número de atributos:	9
Tarefas associadas:	Classificação
Valores incompletos?	Não
Área:	Saúde

Table 2: Informações sobre o *Contraceptive Method Choice Data Set*.

3.2.1 Atributos do *dataset*

- Idade da esposa: numérico.
- Educação da esposa: categórico (1=baixa, 2, 3, 4=alta).
- Educação do esposo: categórico (1=baixa, 2, 3, 4=alta).
- Número de filhos já nascidos: numérico.
- Religião da esposa: binário (0=Não-islâmica, 1=Islâmica).
- Esposa está trabalhando? binário (0=Sim, 1=Não).
- Ocupação do esposo: categórico (1, 2, 3, 4).
- Índice do padrão de vida: categórico (1=baixo, 2, 3, 4=alto).
- Exposição média: binária (0=Boa, 1=Não boa).
- Método contraceptivo utilizado (atributo classe): 1=Não usa, 2=Longo prazo, 3=Curto prazo.

[3]Disponível em `http://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice`

3.2.2 Estado da Arte

O primeiro trabalho a utilizar este *dataset* comparou uma série de algoritmos de classificação (Lim, Loh, Shih & Algorithms 1999). O melhor resultado obtido executando com validação cruzada de 10 alcançou taxa de acerto de 57%, e na média os algoritmos obtiveram 43%. Os melhores resultados reportados na literatura obtém acurácia de 69.79% também com validação cruzada de 10 (Ray & Page 2005).

3.3 *Adult Data Set*

O objetivo do conjunto de dados *Adult*[4] (Adulto) é predizer se as receitas de um indivíduo excede $50.000/ano baseados nos dados do censo. O conjunto de dados também é conhecido como "Census Income" (Receitas do Censo). Informações sobre o *dataset* são apresentadas na Tabela 3.

Características do *dataset*:	Atributos multi-variados
Número de instâncias:	48842
Características dos atributos:	Categóricos, Inteiros
Número de atributos:	14
Tarefas associadas:	Classificação
Valores incompletos?	Sim
Área:	Social

Table 3: Informações sobre o *Adult Data Set*.

3.3.1 Atributos do *dataset*

- Idade: contínuo.
- Classe de Trabalho: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- fnlwgt: contínuo.
- Educação: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- Educação-num: contínuo.
- Estado Civil: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- Ocupação: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- Relacionamento: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

[4]Disponível em `http://archive.ics.uci.edu/ml/datasets/Adult`

- Raça: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- Sexo: Feminino, Masculino.
- Capital-ganho: contínuo.
- Capital-perda: contínuo.
- Horas por semana: contínuo.
- País Nativo: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

3.3.2 Estado da Arte

O primeiro trabalho a utilizar este *dataset* comparou uma série de algoritmos de classificação (Kohavi 1996). Os melhores resultados reportados foram com o uso de C4.5 (84.46+-0.30), Naive-Bayes (83.88+-0.30) e NBTree (85.90+-0.28). A variação nos resultados obtidos correspondem à execução dos algoritmos com a remoção de instâncias com valores desconhecidos. Entretanto, em (Caruana & Niculescu-Mizil 2004), foi reportado o resultado de 90.74% com Boosted stumps (BST-STMP).

4 Metodologia

Para exercitar os algoritmos de aprendizagem de máquina nos *datasets* escolhidos, ao invés de implementar os algoritmos desde o princípio, optou-se por fazer o uso do aplicativo Weka[5]. Este software é uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados. Os algoritmos podem tanto ser aplicados diretamente a um conjunto de dados (*dataset*) através da interface da aplicação, como ser chamados a partir de código Java. O Weka contém ferramentas para o pre-processamento de dados, classificação, regressão, clusterização, regras de associação e visualização. Ele também é apropriado para o desenvolvimento de novos esquemas de aprendizado de máquina. O principal motivo para a escolha deste software é o fato dele ser implementado em Java, além dele prover boas implementações dos algoritmos de aprendizado. Como o SUAP também é implementado nesta linguagem de programação, o Weka pode ser integrado ao sistema.

O Weka possui uma interface gráfica que permite carregar um conjunto de dados, mostrar estatísticas a respeito dos mesmo, escolher e configurar o algoritmo a ser utilizado, e, obviamente, executar o método escolhido. Diversos formatos de arquivos para entrada de dados podem ser utilizados. O formato utilizado foi o `arff`, próprio do Weka, que permite uma descrição dos dados através de anotações no início do arquivo, e logo a seguir os exemplos são listados um por linha com os atributos separados por vírgula.

[5] http://www.cs.waikato.ac.nz/ml/weka/

4.1 Algoritmos

A escolha dos algoritmos utilizados como estudo de caso neste trabalho foi realizada de modo a exercitar as principais técnicas relacionadas com a tarefa de classificação. Vale dizer que foram avaliados outras métodos disponíveis no Weka, mas não são reportados por não terem obtido resultados significativos. Para se determinar um nível inferior de desempenho foi utilizado o modelo `0-R`, que consiste em classificar qualquer amostra com apenas de um único valor, ou seja, a classe mais comum (moda). Os outros algoritmos utilizados são descritos a seguir.

A técnica de *Support Vector Machines* (SVM) foi exercitada através da implementação por *Sequential Minimal Optimization* (SMO) de John Platt (Platt 1998, Keerthi, Shevade, Bhattacharyya & Murthy 2001) para treinamento de um classificador por vetores de suporte. Esta implementação normaliza os dados da instância, ajusta valores ausentes e transforma atributos nominais em binários. Problemas multi-classes, como os que são considerados neste trabalho, são resolvidos usando classificação *pairwise* (Hastie & Tibshirani 1998). Foram avaliadas diversas funções de *kernel* disponíveis no sistema, com alteração de alguns parâmetros de configuração para tentar obter melhorar os resultados.

Modelos neuronais foram avaliados através de um classificador com retroalimentação, o *Multilayer Perceptron*. A rede de neurônios foi construída automaticamente, com os valores padrão de configuração. A camada escondida é composta por $(atributos + classes)/2$, a taxa de aprendizado é de 0.3, o momentum 0.2 e os atributos são normalizados. Os nós na rede são todos do tipo sigmóide.

Para comparar com outras estratégias de aprendizado de máquina, também foram utilizadas técnicas baseadas em árvore de decisão e classificadores bayesianos. Os algoritmos escolhidos foram um classificador *naïve bayes* (John & Langley 1995) e uma implementação do *minimal cost-complexity pruning* (Breiman, Friedman, Stone & Olshen 1984) para criar uma árvore de classificação (C&RT).

4.2 Configuração

A configuração de execução dos algoritmos é apresentada através dos esquemas utilizados no Weka.

- `0-R`: `weka.classifiers.rules.ZeroR`.
- SMO_1: `weka.classifiers.functions.SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"`.
- SMO_2: `weka.classifiers.functions.SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.RBFKernel -C 250007 -G 0.01"`.
- SMO_3: `weka.classifiers.functions.SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -M -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 2.0"`.
- `Multilayer Perceptron`: `weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a`.

- Naïve Bayes: `weka.classifiers.bayes.NaiveBayes`.
- C&RT: `weka.classifiers.trees.SimpleCart -S 1 -M 2.0 -N 5 -C 1.0`.

4.3 Avaliação

Para avaliar os experimentos deste trabalho, utilizamos a técnica de *cross-validation* (Geisser 1993). *Cross-validation* é uma técnica para medir como os resultados de uma análise estatística vão ser generalizados para um conjunto de dados independente. Ela é principalmente usada em configurações onde o objetivo é a predição, e alguém deseja estimar o quão correto um modelo preditivo irá ser executado na prática. Uma rodada do *cross-validation* envolve o particionamento de uma amostra de dados em subconjuntos complementares, executando a análise de um subconjunto (chamado de conjunto de treinamento), e validando a análise em outro subconjunto (chamado de conjunto de validação ou teste). Para reduzir a variabilidade, múltiplas rodadas do *cross-validation* são executadas usando diferentes partições, e os resultados de validação são a média das rodadas. Nos *datasets* deste trabalho, utilizou-se 10-fold *cross-validation*.

5 Resultados

Nesta seção, são apresentados os resultados obtidos através da execução dos algoritmos conforme metodologia descrita na Seção 4. Os *datasets* utilizados foram apresentados na Seção 3. Na Tabela 4 estão a acurácia obtida pelos algoritmos na resolução dos *datasets* que servem como estudo de caso neste trabalho. Para o método `SMO` é reportado o melhor desempenho de todas as configurações avaliadas, sendo que o número em subescrito ao lado do resultado corresponde a configuração utilizada.

Algoritmo	**Post-operative**[1]	**Post-operative**[2]	**Contraceptive Method**	**Adult**
`0-R`	**71.1111%**	**72.093%**	42.702%	75.919%
`SMO`	**71.1111%**$_2$	**72.093%**$_2$	50.4413%$_3$	84.9022%$_1$
`Multilayer Perceptron`	58.8889%	55.814%	52.3422%	82.8936%
`Naïve Bayes`	67.7778%	**72.093%**	50.7807%	83.428%
`C&RT`	70%	70.9302%	55.1935%	86.1091%
`Melhor da literatura`	**71.1%**	69.8%	**69.76%**	**90.74%**

Table 4: Acurácia dos métodos avaliados com validação cruzada de 10.

O *dataset* `Post-operative`[1] é o mesmo da definição original, apresentado na Seção 3.1, e disponível no respositório da UCI com 90 instâncias. Já o `Post-operative`[2] é uma versão reduzida com apenas 86 instâncias, em que foram removidos dois exemplos com valores incompletos e dois exemplos da classe "*I*",e por isso possui apenas duas classes para classificação. Os resultados obtidos neste *dataset* ilustram que se não houver um conjuntos de dados adequado os métodos não conseguem construir bons modelos de classificação.

Há um grande viés nos exemplos deste dataset, sendo que a moda de uma das classes é aproximadamente 70%, e parece não haver diferença significativa nos atributos que possibilite aos algoritmos estudados construir um modelo melhor para identificar outras classes. Na versão original, `Post-operative`[1], a classe "*I*" possui apenas dois exemplos. Mesmo assim, os resultados obtidos são bem superiores aos que encontramos no primeiro trabalho para este problema (48%). Acredita-se que isso ocorreu visto que o trabalho que reportou esse resultado é relativamente antigo (Budihardjo et al. 1991), e muitas das técnicas implementadas pelo Weka, até mesmo usadas com sua forma padrão, já incorporam técnicas mais avançadas. Esta precisão, conforme observado em (Owen 1999), é menor do que simplesmente dizer que todos serão hospitalizados, embora o método possa ter conseguido uma separação útil dos pacientes em grupos. Além disso, através das técnicas executadas, observou-se resultados bem diferentes como, por exemplo, 58.8889% do `Multilayer Perceptron` e o $\texttt{SMO}_3$ com 61.1111% no `Post-operative`[1]. Isso provavelmente ocorreu devido ao tamanho do tamanho do *dataset*, pois sendo ele muito pequeno, os algoritmos ficam muito sensíveis a alterações. Assim, demonstra-se a necessidade de um *dataset* representativo no SUAP antes de adicionar a funcionalidade de referenciamento ao sistema.

Os resultados com o *dataset* `Contraceptive Method`, apresentado na Seção 3.2, ilustram o comportamento dos métodos em um *dataset* de médio porte. O desempenho do `SMO` foi bastante instável, obtendo acurácia de $48.201\%_1$ e $42.3625\%_2$, o que não aconteceu no *dataset* anterior em que os resultados não variaram mais do que 2% do melhor. Os resultados obtidos ficaram bem abaixo do melhor encontrado na literatura, mas também vale dizer que em uma comparação de uma série de algoritmos de classificação (Lim et al. 1999) a melhor taxa de acerto foi de 57%, e na média os algoritmos obtiveram 43%. Neste *dataset*, também foi possível avaliar questões de desempenho dos métodos em relação ao tempo de execução. Na Tabela 5 estão os tempos de execução aproximados para se ter uma noção da ordem de grandeza de computação requerida pelos métodos.

Algoritmo	**Contraceptive Method** (s)	**Adult** (s)
`0-R`	> 0.01	0.02
$\texttt{SMO}_1$	2.5	4116.31
$\texttt{SMO}_2$	5.2	–
$\texttt{SMO}_3$	18.6	> 129600
`Multilayer Perceptron`	21.82	4717.21
`Naïve Bayes`	> 0.01	0.23
`C&RT`	1.63	173.88

Table 5: Tempo de execução para construção de um modelo de classificação.

O último *dataset* avaliado foi o `Adult`, apresentado na Seção 3.3. Trata-se de um *dataset* de grande porte (48842 instâncias), em que pode-se perceber claramente a diferença de desempenho entre os métodos, principalmente em relação ao tempo de execução necessário (vide Tabela 5). O `SMO` mostrou-se uma técnica promissora, mas a configuração para tentar melhores resultados inviabiliza sua execução por consumir um tempo de execução superior ao tempo estabelecido para a execução deste trabalho. Após um dia e meio construindo apenas um modelo, sua execução teve quer ser interrompida. Outros problemas também foram observados como falta de memória durante a execução. Entretanto, o Weka permite parametrizar a aplicação indicando o tamanho máximo de memória que o sistema pode

alocar. Fazendo-se uso deste recurso, o problema de memória foi contornado.

6 Conclusão

Este trabalho visou uma exploração do uso de algoritmos de aprendizado de máquina para a sua utilização no Sistema Unificado de Assistência Pré-natal (SUAP). O SUAP é um sistema que visa suportar o atendimento pré-natal do sistema de saúde público brasileiro. Uma de suas funcionalidade é auxiliar na indicação de unidades secundárias de atendimento para gestações de alto risco.

Dada a atual não existência de um conjunto de dados do SUAP, três *datasets* de diferentes portes, com propriedades similares ao problema que deve ser resolvido, foram escolhidos para permitir a exploração dos algoritmos. Utilizou-se a ferramenta Weka para a execução de tais testes. Os resultados foram satisfatórios.

O Weka mostrou-se um aplicativo extremamente poderoso para a execução dos algoritmos de aprendizado de máquina. A quantidade de algoritmos disponíveis é bastante grande, e é possível a realização de uma grande variedade de parametrizações para os mesmos. Além disso, certas funções podem ser estendidas e implementadas pelo desenvolvedor e utilizadas nos algoritmos.

Conclui-se que a ferramenta Weka é apropriada para ser incorporada no SUAP, visto que: (i) é implementada em Java; (ii) pode ser facilmente incorporada no sistema; (iii) disponibiliza os principais algoritmos de aprendizado de máquina; e (iv) permite uma fácil parametrização dos algoritmos para a obtenção de resultados satisfatórios. Entretanto, é fundamental a existência de um conjunto de dados grande para que se obtenham bons resultados.

References

Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. (1984), *Classification and Regression Trees*, Chapman & Hall/CRC.

Budihardjo, A., Grzymala-Busse, J. W. & Woolery, L. (1991), Program lers_lb 2.5 as a tool for knowledge acquisition in nursing, *in* 'Proc. of the 4th Int. Conf. on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems', pp. 735–740.

Caruana, R. & Niculescu-Mizil, A. (2004), An empirical evaluation of supervised learning for roc area, *in* 'First Workshop of ROC Analysis in AI (ROCAI'04)'.

Carvalho, G., Choren, R., Carvalho, C., Lucena, C., Condack, J. & de Sá, R. M. (2009), Pré-natal digital: um ambiente colaborativo para discussão de casos clínicos em obstetrícia, *in* 'IX Workshop de Informática Médica (WIM) no CSBC 2009', Bento Gonçalves.

de Sa, R. M., Carvalho, C., Moraes, V., Stein, E., dos Santos, T. V., Carvalho, G. & Lucena, C. (2009), Community of obstetrics practice and knowledge exchange: A useful tool for collaboration between obstetricians in an emerging country, *in* 'XIX FIGO World Congress of Gynecology & Obstetrics (FIGO 2009)', Cape Town.

Geisser, S. (1993), *Predictive Inference*, Chapman and Hall.

Hastie, T. & Tibshirani, R. (1998), Classification by pairwise coupling, *in* 'NIPS '97: Proceedings of the 1997 conference on Advances in neural information processing systems 10', MIT Press, Cambridge, MA, USA, pp. 507–513.

Jennings, N. R. (2001), 'An agent-based approach for building complex software systems', *Commun. ACM* **44**(4), 35–41.

John, G. H. & Langley, P. (1995), Estimating continuous distributions in bayesian classifiers, pp. 338–345.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C. & Murthy, K. R. K. (2001), 'Improvements to platt's smo algorithm for svm classifier design', *Neural Comput.* **13**(3), 637–649.

Kohavi, R. (1996), Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid, *in* 'Proceedings of the Second International Conference on Knowledge Discovery and Data Mining'.

Kontkanen, P., Lahtinen, J., Myllymäki, P. & Tirri, H. (2000), Unsupervised bayesian visualization of high-dimensional data, *in* 'KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, New York, NY, USA, pp. 325–329.

Lim, T.-S., Loh, W.-Y., Shih, Y.-S. & Algorithms, N. C. (1999), 'A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms'.

Owen, A. B. (1999), 'Tubular neighbors for regression and classification'.

Platt, J. (1998), Machines using sequential minimal optimization, *in* B. Schoelkopf, C. Burges & A. Smola, eds, 'Advances in Kernel Methods - Support Vector Learning', MIT Press.

Ray, S. & Page, D. (2005), Generalized skewing for functions with continuous and nominal attributes, *in* 'ICML '05: Proceedings of the 22nd international conference on Machine learning', ACM, New York, NY, USA, pp. 705–712.