Comparando técnicas de aprendizagem de máquina para Dimorfismo Sexual

Felipe C. Farias, Carmelo J. A. Bastos Filho
Escola Politécnica, Universidade de Pernambuco
fcf@ecomp.poli.br , carmelofilho@poli.br

Flávio R. S. Oliveira
Instituto Federal de Pernambuco
flavio.oliveira@paulista.ifpe.edu.br

Camila P. B. Barbosa, Luiz M. P. Sanches , Rosane C. S. Galvão
Universidade Federal de Pernambuco
camila_bacelar1@hotmail.com ,
luiz.sanches.ufpe@gmail.com ,
rosanecs.galvao@gmail.com

***Resumo*— A identificação do sexo a partir de ossos humanos é uma tarefa complexa e por vezes imprecisa, embora seja uma rotina importante no trabalho de cientistas forenses e legistas. Esta identificação é realizada comumente segundo observação visual e depende da experiência do especialista. O presente trabalho investigou o emprego de diversas técnicas de Aprendizagem de Máquina para identificação do sexo de sujeitos a partir de medidas da pelve humana. Os resultados foram promissores e sugerem que seria viável utilizar esta abordagem para complementar e apoiar a decisão de especialistas.**

Palavras-chave — dimorfismo sexual; inteligência computacional; aprendizagem de máquina;

I. Introdução

A Inteligência Computacional [1] é uma área que estuda técnicas e ferramentas que são inspiradas em comportamentos inteligentes. Dentro desta grande área, existe uma subárea chamada Aprendizagem de Máquina, que contempla um conjunto de algoritmos que permitem o computador aprender a realizar determinadas tarefas. Existem diversas técnicas que podem ser usadas para classificação, por exemplo: Árvores de Decisão, o *Random Forest*, os K-Vizinhos mais próximos, Redes Neurais Artificiais e Máquinas de vetor de suporte. O dimorfismo sexual trata-se da identificação do gênero de uma pessoa, utilizando partes do esqueleto, constituindo-se em uma área de estudos bastante importante na antropologia e na ciência forense. Existe uma grande imprecisão na classificação de gênero pelo fato desta análise ser, em sua grande maioria, visual e com alto grau de subjetividade.

Neste trabalho, foi proposto o uso de diversas técnicas inteligentes para o dimorfismo sexual, com o objetivo de torná-lo uma tarefa mais precisa, visto que nesse caso, não estaríamos sujeitos a subjetividade causada pela análise visual. Para isso, foram mensuradas algumas regiões do osso da pelve, pelo fato dessa estrutura anatômica ser um dos segmentos do esqueleto que apresenta maior diferenciação entre os gêneros. O restante deste artigo está organizado conforme a seguir: na seção 2 o referencial teórico é apresentado; na seção 3, a base de dados, a configuração experimental e os detalhes da análise estatística realizada são detalhados; na seção 4 os resultados experimentais são descritos e na seção 5 as conclusões são apresentadas.

II. Referencial Teórico

A. *Dimorfismo Sexual em Humanos*

A determinação do sexo é de grande importância para o levantamento do perfil biológico de restos humanos não identificados nas ciências forenses [2]. O dimorfismo sexual trata das diferenças de características de partes do corpo humano existentes para o sexo masculino e feminino. Para realizar este tipo de identificação, pode-se utilizar, por exemplo, medições de dentes específicos [3]. A pelve humana é uma das regiões morfológicas mais complexas do esqueleto humano, que também pode ser utilizada para o dimorfismo sexual do sujeito a ser identificado [4], [5]. Devido ao fato desta região ser de identificação bastante complexa, o uso de técnicas inteligentes pode ser útil, notadamente para auxiliar os peritos que trabalham na área forense de identificação do sexo a partir de partes do corpo humano.

B. *Árvores de Decisão*

Para realizar a tarefa de classificação, pode-se utilizar uma técnica chamada de árvores de decisão [6]. Esta técnica funciona através da aplicação de sucessivas regras de decisões simples, começando pela raiz da árvore e passando por cada nó até chegar nas folhas da mesma. Normalmente cada nó representa uma pergunta sobre determinado atributo e cada folha representa a escolha de uma classe. A montagem da árvore de decisão pode ser feita conforme diversos algoritmos [6] a partir de uma distribuição de probabilidade encontrada durante o treinamento da técnica. A técnica de treinamento utilizada neste trabalho foi a C4.5 [7], através da implementação *open-source* J48 existente na ferramenta Weka [8]. A principal vantagem das árvores de decisão é sua capacidade de explicar como as respostas são geradas. Em domínios como a área médica, esta é uma característica diferencial.

C. Random Forest

Uma abordagem derivada das árvores de decisão é conhecida como *Random Forest* (RF) [9]. Esta técnica funciona a partir do conceito de comitê de classificadores, composto por um conjunto de árvores de decisão. Várias árvores de decisão são treinadas com conjuntos de amostras diferentes dos dados. O particionamento dessas amostras para este comitê de classificadores é denominado *bootstrapping*. Para realizar a classificação, é feita uma votação baseando-se nas respostas de cada árvore, que individualmente aplicam uma decisão sobre o mesmo dado. Este mecanismo de votação é conhecido como *bagging*. Os mecanismos utilizados pelo *RandomForest*, são úteis para reduzir a chance ter-se *overfitting* e *underfitting*, que são problemas relacionados a perda do poder de generalização e a falta do treinamento necessário para o classificador, respectivamente.

D. K-Vizinhos Mais Próximos (KNN)

Os K Vizinhos mais próximos [10] é uma técnica bastante simples que utiliza alguma métrica de distância, normalmente euclidiana, de uma amostra que se deseja saber a classe para todos os pontos conhecidos previamente. Após o cálculo de todas as distâncias, a resposta será a classe com a maior frequência dentre as classes dos elementos que têm as K menores distâncias.

E. Redes Neurais Artificiais

As Redes Neurais Artificiais [1], [11], [12] são estruturas computacionais de alta complexidade, utilizadas amplamente na ciência e tecnologia com aplicações em várias áreas de pesquisa, como na química, robótica, aeroespacial, biologia e física. São úteis em tarefas onde é necessário haver: reconhecimento de padrões, previsões, aproximação de funções, controle de sistemas dinâmicos, etc. Baseadas no funcionamento do cérebro humano, as RNAs são formadas por neurônios (unidades de processamento) interconectados através de sinapses (pesos), que são as estruturas responsáveis pelo aprendizado da técnica. A estrutura de uma RNA é composta por uma camada de entrada, que é responsável por entregar as variáveis do problema para dentro do algoritmo, uma ou mais camadas intermediárias, responsáveis pelo capacidade de aproximar funções não-lineares, e uma camada de saída, que indica a qual classe pertence determinada amostra de dados. As RNAs que apresentam estas camadas, são conhecidas como *Multilayer Perceptron* (MLP). Existem diversos algoritmos de treinamento para as MLPs, o *backpropagation* [13] é um dos mais utilizados devido a sua simplicidade e boas taxas de classificação para diversos problemas. Da mesma forma que existe um comitê de árvores de decisão, pode-se ter um comitê de MLPs realizando a classificação do problema.

F. Máquina de Vetor de Suporte (SVM)

A Máquina de Vetor de Suporte [14] é um algoritmo bastante utilizado na área de reconhecimento de padrões, baseado no aprendizado estatístico [15] que tem por objetivo minimizar o erro de treinamento e diminuir a complexidade estrutural do modelo [16]. É um método não-paramétrico, que armazena instâncias específicas durante o treniamento, para a criação dos seus vetores de suporte. Caso a distribuição de probabilidade da população seja conhecida, pode-se obter o erro de generalização, conhecido como Risco Funcional. Geralmente nos problemas reais, a população não é conhecida, tem-se apenas uma amostra da população, onde é possível obter o Risco Empírico. Ainda segundo [15], no aprendizado supervisionado, a SVM consegue obter a capacidade máxima de generalização quando é minimizado o Risco Funcional, que é equivalente a minimizar o Risco Empírico e a complexidade do modelo, conhecida como Risco Estrutural. As SVMs possibilitam a criação de classificadores de margens máximas, ou seja, que cria fronteiras de decisão com a maior distância possível entre classes diferentes, isso faz com que o algoritmo tenha um bom poder de generalização. A SVM pode utilizar *kernels* que transformam o espaço de características com a finalidade de facilitar a separação através dos hiperplanos da SVM. Uma das técnicas de treinamento da SVM é o *Sequential Minimal Optimization* (SMO) [17]

A escolha dessas técnicas aconteceu baseada na grande utilização das mesmas em diversos trabalhos da medicina, e em vários outros campos de estudo que tratam de problemas de classificação.

III. Metodologia

A. Base de Dados

Para realizar este estudo, foram obtidas 22 medidas em 75 ossos de pelves humanas, com o sexo conhecido *a priori*, sendo 32 peças femininas e 43 masculinas. Na figura 1, os principais ossos da pelve podem ser vistos. As medidas foram obtidas com paquímetros digitais, de modo a diminuir o erro de mensuração. Estas medidas foram realizadas no Laboratório de Anatomia do Centro Acadêmico de Vitória de Santo Antão da Universidade Federal de Pernambuco, sob a supervisão de profissionais da área, membros da equipe e estão dispostas na Tabela I.

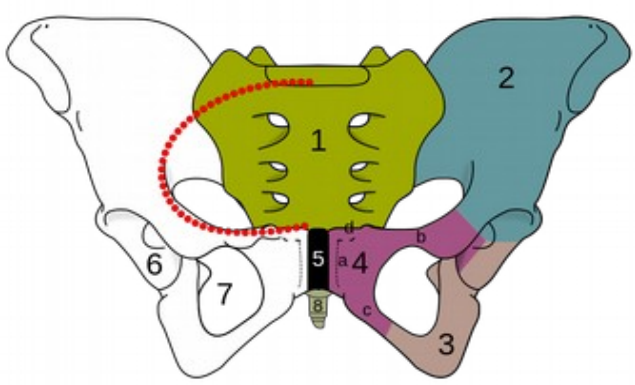

Fig. 1: Ossos da Pelve - (1) Sacro, (2) Ílio, (3) Ísquio, (4) Púbis, (5) Sínfise Púbica, (6) Acetábulo, (7) Forame Obturado, (8) Cóccix.

TABLE I. MEDIDAS DA PELVE UTILIZADAS NA BASE DE DADOS

Atributo	*Medida*
1	Altura total da pelve
2	Largura total do ílio
3	Comprimento da crista ilíaca
4	Espessura margem da Crista ilíaca
5	Comprimento do púbis
6	Altura púbica
7	Largura do púbis
8	Espessura da sínfise púbica
9	Comprimento ramo isquio-púbico inferior
10	Espessura ramo isquio-púbico inferior
11	Diâmetro transverso do acetábulo
12	Largura Máxima do tuber Isquiatico
13	Extensão da espinha isquiática
14	Largura da incisura isquiática maior
15	Profundidade máxima da incisura isquiática
16	Comprimento da superfície auricular
17	Altura da superfície auricular
18	Distância entre a crista ilíaca e a superfície auricular
19	Depressão (cavidade) paragleinodal
20	Forma da Superfície Auricular
21	Comprimento Forame obturado
22	Largura Forame obturado

As simulações dos classificadores foram realizadas no Weka [8] e utilizaram as 22 medidas para decidir o sexo do sujeito dos quais as peças foram retiradas.

Foram realizados vários testes preliminares com as técnicas: MLP, Árvores de Decisão, *Random Forest* e SVM. Depois da busca de bons parâmetros de configuração das técnicas, foi utilizada a técnica de validação cruzada *10-fold*, e repetida 30 vezes, totalizando 300 experimentos por cada arquitetura de classificador. As arquiteturas utilizadas foram: MLP com 20 neurônios na camada escondida, taxa de aprendizado = 0,1 e momento = 0,1; Apenas um vizinho mais próximo; SVM com *kernel* linear; SMO com *kernel* polinomial; *Random Forest* com 100 árvores; comitê com 100 MLPs. Por fim, foram realizados testes estatísticos de Wilcoxon para comparar as performances das técnicas na classificação do sexo humano através das medidas apresentadas.

IV. RESULTADOS EXPERIMENTAIS

Os resultados na Tabela II abaixo, evidenciam que as técnicas tais como SMO, SVM e MLP obtiveram melhor precisão e menor desvio padrão nos experimentos realizados. Esta observação possivelmente está relacionada ao fato que estas técnicas são eficientes na modelagem de dados com comportamento não-linear.

TABLE II. RESULTADOS DAS CLASSIFICAÇÕES

Classificador	**Precisão Média (%)**	**Desvio Padrão**
SMO	91,73	0,10
SVM	91,70	0,10
Comitê 100 MLPs	91,58	0,10
MLP	90,31	0,11
RF	90,23	0,11
KNN	88,93	0,12
J48	78,45	0,13

A Tabela III abaixo, exibe os resultados do teste de Wilcoxon. Observa-se que grande parte dos elementos (o) os classificadores foram considerados estatisticamente equivalentes. As árvores de decisão treinadas com o algoritmo J48 foram consideradas inferiores aos demais (-). O mesmo acontece com o algoritmo KNN que apresentou desempenho inferior às técnicas mais avançadas.

TABLE III. TESTES ESTATÍSTICOS DE WILCOXON

	J48	**SMO**	**RF**	**KNN**	**MLP**	**Comitê 100 MLPs**	**SVM**
J48	o	-	-	-	-	-	-
SMO	+	o	o	+	o	o	o
RF	+	o	o	o	o	o	o
KNN	+	-	o	o	o	-	-
MLP	+	o	o	o	o	o	o
Comitê 100 MLPs	+	o	o	+	o	o	o
SVM	+	o	o	+	o	o	o

V. CONCLUSÃO

Este trabalho apresentou um estudo comparativo do emprego de técnicas de aprendizagem de máquina para dimorfismo sexual em humanos.

Os resultados dos experimentos preliminares sugerem que este pode ser um mecanismo válido para apoiar especialistas na identificação do sexo a partir de medidas da pelve humana.

Os trabalhos futuros envolvem duas frentes de atuação: (i) aprofundar o estudo dos mecanismos de classificação, inclusive, verificando a possibilidade de seleção paramétrica e (ii) construir protótipo de sistema baseado em tecnologias móveis. Através da seleção paramétrica, pode-se obter uma melhor precisão nos classificadores, ou ao menos, pode-se permitir manter níveis de precisão aceitáveis necessitando de

menos medidas. Ao disponibilizar os modelos de classificação em dispositivo móvel, será possível realizar testes piloto com cientistas forenses e legistas, auxiliando-os na solução de problemas reais de identificação.

REFERÊNCIAS

[1] S. J. Russell e P. Norvig, *Artificial Intelligence: A Modern Approach*, 2º ed. Pearson Education, 2003.

[2] Y. Bilge, P. S. Kedici, Y. D. Alakoç, K. Ü. Ülküer, e Y. Y. İlkyaz, "The identification of a dismembered human body: a multidisciplinary approach", *Forensic Sci. Int.*, vol. 137, nº 2–3, p. 141–146, nov. 2003.

[3] S. Gupta, "Establishment of Sexual Dimorphism in North Indian Population by Odontometric Study of Permanent Maxillary Canine", *J. Forensic Res.*, vol. 05, nº 02, p. 2–5, 2014.

[4] M. W. Purcell e D. J. Wescott, "Sexual dimorphism of the femur: Biomechanical influence of pelvic morphology", in *AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY*, 2014, vol. 153, p. 212.

[5] K. M. Brown, "Selective pressures in the human bony pelvis: Decoupling sexual dimorphism in the anterior and posterior spaces", *Am. J. Phys. Anthropol.*, vol. 440, p. n/a–n/a, 2015.

[6] J. R. Quinlan, "Induction of Decision Trees", p. 81–106, 2007.

[7] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[8] M. Hall, H. National, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, e I. H. Witten, "The WEKA Data Mining Software : An Update", *SIGKDD Explor.*, vol. 11, nº 1, p. 10–18, 2009.

[9] L. Breiman, "Random Forrest", *Mach. Learn.*, p. 1–33, 2001.

[10] T. Cover e P. Hart, "Nearest neighbor pattern classification", *IEEE Trans. Inf. Theory*, vol. 13, nº 1, 1967.

[11] M. J. da S. Valença, *Aplicando Redes Neurais: Um Guia Completo*. Olinda: Livro Rápido, 2005.

[12] S. S. HAYKIN, *Redes Neurais - 2ed.* BOOKMAN COMPANHIA ED, 2001.

[13] P. J. Werbos, "Backpropagation through time: what it does and how to do it", *Proc. IEEE*, vol. 78, nº 10, p. 1550–1560, 1990.

[14] C. J. Burges, "A tutorial on support vector machines for pattern recognition", *Data Min. Knowl. Discov.*, vol. 2, nº 2, p. 121–167, 1998.

[15] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[16] M. VALENÇA, *FUNDAMENTOS DAS REDES NEURAIS*. MEUSER VALENÇA, 2010.

[17] J. C. Platt, "Fast Training of Support Vector Machines Using Sequential Minimal Optimization", *Adv. kernel methods*, p. 185 – 208, 1998.