STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LX**, Number 2, 2015

SEX IDENTIFICATION IN ARCHAEOLOGICAL REMAINS USING DECISION TREE LEARNING

IOAN-GABRIEL MIRCEA, GABRIELA CZIBULA AND MARA-RENATA PETRUŞEL

Abstract. We are approaching in this paper, from a machine learning perspective, the problem of detecting the gender of human skeletal remains from bone measurements. The problem of sex identification of human remains is of major importance for bioarchaeologists, since it provides information regarding the characteristics of past societies. For predicting the gender of human skeletons, an inductive learning method based on decision trees will be used. Computational experiments are performed on publicly available archaeological data sets. The obtained results emphasize the effectiveness of the proposed approach with respect to the similar approaches existing in the literature.

1. Introduction

Detecting the gender of human skeletal remains is very important for studying the gender differences in past populations [5]. This contributes to a better understanding of the social position and attributions of each gender in society. The sex identification task is a very delicate one and is highly influenced by the historical period and the geographic origin of the skeleton.

In this paper we are focusing on the problem of gender detection of human skeletal remains from bone measurements. Most of the approaches existing in the literature regarding the gender detection of human skeletons are using statistical methods or are based on bone measurements and DNA or gene analysis. Few computational intelligence techniques have been investigated for detecting the sex of human skeletons [2, 13]. The previous approach from [13] applies CHAID (*CHi-squared Automatic Interaction Detection*), a tree based technique which uses the Chi-square test [7] to determine the best next split at each node in the tree.

Received by the editors: June 3, 2015.

2010 *Mathematics Subject Classification.* 68T05,68T10.

1998 *CR Categories and Descriptors.* I.2.6 [**Artificial Intelligence**]: Learning – *Induction*; I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems – *Medicine and science* .

Key words and phrases. bioarchaeology, machine learning, decision tree learning.

We propose in this paper an approach based on an optimized ID3 decision tree learning algorithm for solving the sex detection problem. The main contribution of the paper consists in using a Chi-square pre-pruning technique for reducing the overfitting in the learning process. Three case studies will be used for evaluating the performance of our models and a study towards identifying the best feature set for learning is also conducted. The obtained results emphasize that our approach overperforms an existing approach from the literature based on discriminant function analysis. As far as we know, a decision tree based approach for gender detection similar to ours has not been reported in the literature.

In the case of gender classification we are dealing with real-valued features (bone measurements) and a lot of machine learning algorithms could be used, without requiring the transformation of the initial continuous input space into a discrete one. Still, our main motivation behind choosing the *decision tree* learning is that decision trees provide human-readable rules. A decision tree can be easily converted into set of rules and thus, the obtained results can be easily understood by humans and this way, a feed-back from the bioarchaeologists would be simply obtained. We consider this as a main advantage of our tree based approach for gender detection. Certainly, there are limitations of such decision tree based approaches, such as their sensitivity to errors in the training data (e.g outliers, noise, etc) and their tendency to overfittting. The occurrence of errors in the training data may be attenuated when the decision tree enables fuzzy logic in the decision process. The overfitting tendency is reduced in the present approach by use of the Chi-squared pre-pruning technique.

The remainder of the paper is organized as follows. Section 2 outlines the fundamentals of decision tree learning. Section 3 introduces our approach for the detection of gender in human skeletons. Experimental evaluations are given in Section 4 and a comparison to similar work from the literature is presented in Section 5. Section 6 presents the conclusions of the paper and mentions future research directions.

2. Background

In this section we are providing a brief background on decision trees. One of the most commonly employed methods for approximating discrete- valued functions is decision tree learning. In this case, the learned function is defined by a decision tree. Furthermore, the decision tree learning method is able to learn disjunctive statements and is robust to noisy data.

When using decision trees, in order to classify instances, these are sorted starting from the root of the tree, until a leaf node, which will specify the

classification of the instance. All the nodes in the tree point out a test of an attribute of the instance, while every branch downward that node correlates to one of the possible values of this attribute [10]. A well-known algorithm for decision tree learning is the ID3 algorithm, being the basis for other known algorithms for building decision trees such as CART (*Classification and Regression Trees* [4]) and C4.5 [11]. The way the ID3 algorithm learns decision trees is by constructing them top-down, starting with the best attribute at the root of the tree, which is selected after each instance attribute is tested using a statistical test. The measure that is commonly used for deciding the best attribute at a given node in the tree is the *information gain* [12]. After that, a descendent of the root node is built for every available value of this attribute. Also, the training examples will be sorted to the suitable descendant node. The whole process is then repeated and this results in a greedy search, that never returns to reconsider choices already made [10].

It has to be mentioned that a decision tree may be viewed as a set of rules, thus the reasoning process (the way the classification of an instance was decided) is available to the user. This makes the *decision tree* learning reliable.

3. Our approach

This section presents our approach on using *decision trees* for sex identification in human remains from the length of long bones of the arm and leg.

Let us consider a data set consisting of human skeletons. Each skeleton from the data set is labeled as **male** or **female**. Each skeleton is characterized by m numerical features representing different measurements that were performed on it. Usually, the measurements correspond to several significant bones in the body. Therefore, an instance (skeleton) may be viewed as an m-dimensional vector.

The first step in building our inductive learning models is the *data preprocessing* step. Since we are dealing with real-valued features (i.e the bone measurements are real values), the data is discretized. The discretization idea is the one indicated in [10] and presumes that a discrete-valued feature will be defined dynamically to divide the continuous attribute value into a discrete set of intervals. For discretizing a particular feature (bone measurement) we are searching for a threshold t which will divide the continuous attribute space into a discrete one. In our approach we selected two numerical values for discretizing each feature. More exactly, the attribute values that are less than t will be replaced with 1 and the other values are replaced with 2. The most convenient value for the threshold t is the one that will produce the greatest information gain of the considered feature.

3.1. **Building the model.** After the data set is pre-processed as indicated above, the inductive learning model will be built during the training step. The classification process takes place in two phases that indicate the ideas of an inductive learning algorithm: *training* and *testing*. In the training phase the inductive model will be built and further applied for classifying an unseen skeleton as part of the testing phase.

For building the *decision tree* (DT), an optimized variant of the ID3 algorithm [11] is used. When building the tree, an heuristic is used in order to stop splitting a node in the tree if this spilt is considered to be unimportant. More precisely, let us assume that a certain node n is built in the tree (using the ID3 algorithm) and S_n is the set of instances corresponding to it (sorted to node n). If the percentage of instances (skeletons) from the set S_n belonging to one of the two classes (male or female) is less than a given threshold τ, then the node n is considered to be a leaf node and it is labeled with the most common classification of instances from S_n.

Besides the heuristic described above, in order to avoid overfitting, a χ^2 *pruning* [7] is used with the scope of reducing the tree while it is built. This is a form of pre-pruning, in which a statistical test is applied to the data at a particular node in the tree, in order to determine if the distribution of classes in the data is or not statistically significant.

The main idea of performing pruning at a particular node n in the tree, i.e to stop growing the tree below n, is to apply the χ^2 test to verify if the feature X corresponding to the node n is uncorrelated with the decision (of splitting the node). If uncorrelated, we expect that the real number of male and female instances at this node to be close to the expected number of male and female instances at the node. For this, we need a measure of "deviation", defined as in Formula (1). The intuition is the following: for each possible value for the feature X (i.e 1 and 2), we compute the subset S_n^1 (the subset of instances from S_n having the value 1 for the feature X) and S_n^2 (the subset of instances from S_n having the value 2 for the feature X).

$$C = \sum_{v=1,2} \left(\frac{(Real\ count_{S_n^v}^{female} - Expected\ count_{S_n^v}^{female})^2}{Expected\ count_{S_n^v}^{female}} \right) .+$$

$$+ \sum_{v=1,2} \left(\frac{(Real\ count_{S_n^v}^{male} - Expected\ count_{S_n^v}^{male})^2}{Expected\ count_{S_n^v}^{male}} \right). \tag{1}$$

In Formula (1), for a value v of the feature X (1 or 2) and for a given class c (male or female), by $Real\ count_{S_n^v}^{c}$ we denote the number of instances classified with c within the set S_n^v and by $Expected\ count_{S_n^v}^{c}$ we denote the expected number of instances classified with c within the set S_n^v.

Intuitively, the smaller the value of C is, more likely, the feature X at node n is uncorrelated with the splitting decision. Thus, if C is less than a threshold ϵ, a pruning is performed at the node n (we decide to stop build the tree below the node n).

3.2. **Testing.** After the inductive learning model was built as described above, a testing step is performed to evaluate its performance. When a new instance has to be classified, it is sorted down the DT starting from the root node until a leaf node which gives the classification.

First, the confusion matrix for the two possible outcomes (female and male) is computed. Considering that the "female" class is the positive one and the "male" class is the negative one, the following values are computed: *TP* - the number of *true positives* (the number of actual positive instances predicted as positive), *FP* - the number of *false positives* (the number of actual negative instances predicted as positive), *TN* - the number of *true negatives* (the number of actual negative instances predicted as negative) and *FN* - the number of *false negatives* (the number of actual positive instances predicted as negative).

Using the values computed from the confusion matrix, two evaluation measures will be further used to test the performance of the DT model. The *accuracy* (denoted by *Acc*) indicates the percentage of correctly classified instances, i.e $Acc = \frac{TP+TN}{TP+TN+FP+FN}$. The *Area under the ROC curve* measure (denoted by *AUC*) which is considered one of the best evaluation measures used to compare classifiers [9, 6]. The *AUC* measure represents the area under the ROC (Receiver Operating Characteristics) curve.

ROC curves can be constructed for classifiers which, instead of directly providing the class of an instance, return a score which may be transformed into a class label using a threshold. A threshold is applied on the continuous output of the classifier and the ROC curve is actually obtained by varying the decision threshold (over a given range). In such cases, for each threshold different (1-$specificity$, $recall$) pairs are obtained, which are represented on the ROC curve. The *recall* of the classifier is computed as the proportion of actual positive instances which are predicted positive, i.e. $recall = \frac{TP}{TP+FN}$. The *specificity* of the classifier represents the proportion of actual negative instances which are predicted negative, i.e. $specificity = \frac{TN}{TN+FP}$.

In case of classifiers which return directly the class, as our decision tree based approach is, the ROC space has a single point. As presented in [6], the ROC curve is obtained by linking the (1-$specificity$, $recall$) point to the points at (0,0) and (1,1). For the constructed curve, the AUC measure can be computed.

Good classifiers have high *accuracy* and AUC values. Thus, these measures need to be maximized in order to obtain better classifiers.

For evaluating the performance of the DT model, a *leave-one out* cross-validation was used. *Cross-validation* is a well-known technique used for estimating the generalization error of a classifier [15]. In the *leave-one out* cross-validation on a data set with k instances, k-1 instances are used for training and then the obtained model is tested on the instance which was left out. This is repeated k times and the *accuracy* and the AUC measures are computed as described above.

4. Experimental evaluation

This section contains the experimental evaluation of the DT model (described in Section 3) considering three case studies which were performed on two data sets obtained from the literature [1]. The data set from [1] consists of 200 male and 200 female skeletons from the Pretoria Bone and Raymond A. Dart collections. Ten anthropometric measurements were taken from the radius bone and nine measurements from the ulna bone. The skeletal remains represent black South Africans from the 19^{th} and 20^{th} centuries, born between 1863 and 1996. A statistical analysis of the considered data set has proven that , for each anthropometric measurement, the underlying data follows a normal distribution.

In each data set considered for evaluation, the instances (skeletons) within the data sets are labeled as being **male** or **female**.

The experiments are conducted as follows.

In order to study the influence of the features regarding the performance of the classification tasks, experiments were conducted considering different feature subsets. We used correlation based feature selection. For each feature, the Pearson correlation coefficient [14] between the feature and the target classification output (i.e the gender) is computed. We mention that the correlations have been computed before data discretization.

The features were then sorted in the increasing order of their correlation to the output. The performance of the proposed approach was assessed first on the entire set of ordered features and then subsequently on the set obtained by removing the first feature from the previous set.The assessment process ends when the feature set contains only the two most correlated features.

When building the decision tree, two impurity functions are used to measure the heterogeneity of a set of labeled samples. The first impurity function is the *entropy* and is commonly used in building decision trees. The second impurity function we are using is the *misclassification* function. For a set S of instances (consisting of a males and b females), the *entropy* of S is computed

as $-\frac{a}{a+b} \cdot log_2 \frac{a}{a+b} - \frac{b}{a+b} \cdot log_2 \frac{b}{a+b}$ and the *misclassification* of S is computed as $misclassification(S) = \begin{cases} \frac{b}{a+b} & if \quad a > b \\ \frac{a}{a+b} & otherwise \end{cases}$

The decision tree will be built as described in Section 3.1, considering a value of 0.90 for the threshold τ and a value of 0.80 for the threshold ϵ used for the χ^2 *pruning* step.

4.1. **First case study.** The first case study we are considering for evaluation consists of human remains identified by ten radial measurements. Thus, there are 10 features characterizing the instances within the data set. The features represent the following radial measurements [1]: maximum length of the radius (F1), distal breadth (F2), circumference at the midshaft (F3), sagittal diameter at midshaft (minimum diameter) (F4), transverse diameter at midshaft (maximum diameter) (F5), vertical radial head height (F6), minimum head diameter (F7), maximum head diameter (F8), circumference of the radial (F9) and circumference at the tuberosity (F10). The correlations between the features and the target gender are given in Figure 1. We observe that the features are well enough correlated with the output.

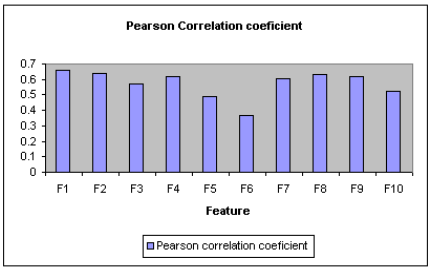

FIGURE 1. Correlations for the features from the first case study

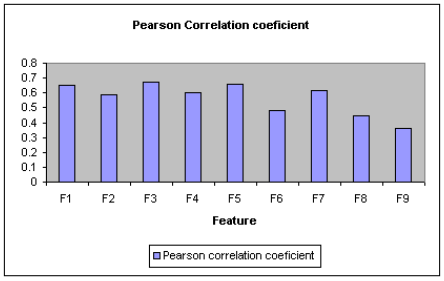

FIGURE 2. Correlations for the features from the second case study

Table 1 presents the results obtained after the experiments performed on the first case study using a *decision tree* constructed with χ^2 pruning. For each experiment we depict the set of features used for classification, the impurity function used for building the tree (*Entropy* and *Misclassification*) and the values obtained for the *Acc* and *AUC* evaluation measures using a *leave-one-out* cross-validation. The best result obtained is marked with bold.

Experiment	Feature set	Entropy		Misclassification	
		Acc	AUC	Acc	AUC
1	$\{6, 5, 10, 3, 7, 9, 4, 8, 2, 1\}$	0.843	0.843	0.835	0.836
2	$\{5, 10, 3, 7, 9, 4, 8, 2, 1\}$	0.843	0.843	0.840	0.841
3	$\{10, 3, 7, 9, 4, 8, 2, 1\}$	0.850	0.850	0.858	0.858
4	$\{3, 7, 9, 4, 8, 2, 1\}$	0.853	0.853	0.853	0.853
5	$\{7, 9, 4, 8, 2, 1\}$	0.853	0.853	0.853	0.853
6	$\{9, 4, 8, 2, 1\}$	0.853	0.853	0.840	0.840
7	$\{4, 8, 2, 1\}$	**0.860**	**0.860**	0.855	0.855
8	$\{8, 2, 1\}$	0.818	0.833	0.803	0.822
9	$\{2, 1\}$	0.825	0.831	0.825	0.832

TABLE 1. Results obtained on the first case study.

4.2. **Second case study.** The second case study consists of nine measurements of the ulna bone from the human skeletons. There are 9 features describing the instances within the data set: maximum length of the ulna (F1), maximum length of the ulna measured using the plumbline geniometer method (F2), anterior-posterior diameter (minimum diameter) (F3), medial-lateral diameter (maximum diameter) (F4), circumference at midshaft (F5), minimum circumference of the ulna (F6), olecranon breadth (F7), minimum olecranon breadth (F8) and height of the olecranon (F9). The correlations between the features and the target gender are given in Figure 2 and show a good correlation with the gender. The results obtained on the second case study are outlined in Table 2 and the best result obtained is highlighted.

Experiment	Feature set	Entropy		Misclassification	
		Acc	AUC	Acc	AUC
1	$\{9, 8, 6, 2, 4, 7, 1, 5, 3\}$	0.868	0.872	0.845	0.847
2	$\{8, 6, 2, 4, 7, 1, 5, 3\}$	0.868	0.872	0.845	0.847
3	$\{6, 2, 4, 7, 1, 5, 3\}$	0.868	0.872	0.843	0.844
4	$\{2, 4, 7, 1, 5, 3\}$	0.875	0.881	0.845	0.847
5	$\{4, 7, 1, 5, 3\}$	**0.878**	**0.885**	0.848	0.849
6	$\{7, 1, 5, 3\}$	0.858	0.858	0.863	0.863
7	$\{1, 5, 3\}$	0.850	0.868	0.855	0.855
8	$\{5, 3\}$	0.838	0.839	0.838	0.839

TABLE 2. Results obtained on the second case study.

4.3. **Third case study.** We considered in this paper, as the third case study, the data set which contains both the radial and ulnar measurements (considered in the first and second case studies). Consequently, in this data set, each

skeleton (instance) will be represented by 19 measurements (features): the first ten are the radial measurements (as in the first case study) and the next nine features represent the ulnar measurements (as in the second case study).

The results obtained on the third case study are given in Table 3. The last line in Table 3 contains the results we have obtained when considering as feature set the union of features which have provided the best results for the first two case studies. One can observe that this set of features provided the best accuracy, using the entropy misclassification function.

Experiment	**Feature set**	**Entropy**		**Misclassification**	
		Acc	**AUC**	**Acc**	**AUC**
1	$\{19, 6, 18, 16, 5, 10, 3, 12, 14, 7, 17, 4, 9, 8, 2, 11, 1, 15, 13\}$	0.880	0.880	0.860	0.860
2	$\{6, 18, 16, 5, 10, 3, 12, 14, 7, 17, 4, 9, 8, 2, 11, 1, 15, 13\}$	0.880	0.880	0.860	0.860
3	$\{18, 16, 5, 10, 3, 12, 14, 7, 17, 4, 9, 8, 2, 11, 1, 15, 13\}$	0.880	0.880	0.873	0.873
4	$\{16, 5, 10, 3, 12, 14, 7, 17, 4, 9, 8, 2, 11, 1, 15, 13\}$	0.880	0.880	0.875	0.875
5	$\{5, 10, 3, 12, 14, 7, 17, 4, 9, 8, 2, 11, 1, 15, 13\}$	0.880	0.880	0.883	0.883
6	$\{10, 3, 12, 14, 7, 17, 4, 9, 8, 2, 11, 1, 15, 13\}$	0.880	0.880	0.883	0.883
7	$\{3, 12, 14, 7, 17, 4, 9, 8, 2, 11, 1, 15, 13\}$	0.880	0.880	0.873	0.873
8	$\{12, 14, 7, 17, 4, 9, 8, 2, 11, 1, 15, 13\}$	0.880	0.880	0.875	0.875
9	$\{14, 7, 17, 4, 9, 8, 2, 11, 1, 15, 13\}$	**0.885**	**0.886**	0.883	0.883
10	$\{7, 17, 4, 9, 8, 2, 11, 1, 15, 13\}$	0.870	0.871	0.880	0.880
11	$\{17, 4, 9, 8, 2, 11, 1, 15, 13\}$	0.870	0.871	0.875	0.875
12	$\{4, 9, 8, 2, 11, 1, 15, 13\}$	0.873	0.873	0.873	0.873
13	$\{9, 8, 2, 11, 1, 15, 13\}$	0.863	0.866	0.845	0.845
14	$\{8, 2, 11, 1, 15, 13\}$	0.863	0.866	0.858	0.859
15	$\{2, 11, 1, 15, 13\}$	0.870	0.874	0.865	0.866
16	$\{11, 1, 15, 13\}$	0.863	0.863	0.865	0.871
17	$\{1, 15, 13\}$	0.865	0.871	0.865	0.871
18	$\{15, 13\}$	0.838	0.839	0.838	0.839
19	$\{4, 8, 2, 1, 14, 17, 11, 15, 13\}$	**0.885**	**0.886**	0.873	0.873

TABLE 3. Results obtained on the third case study.

The fact that there are insignificant fluctuations between the best values obtained in the third case study (both when using entropy and misclassification impurity functions) and the ones obtained when considering as feature set the

best values obtained in the first and second case studies, demonstrates once more the precision of our approach on the sex determination problem using decision tree learning method.

5. Discussion and comparison to related work

Table 4 summarizes the *Acc* and *AUC* values (minimum, maximum, mean and standard deviation) obtained using decision tree learning on the case studies considered for evaluation in Section 4. For each case study, the best values for the *Acc* and *AUC* are highlighted. One can see that the best values, for each case study, are obtained using the *entropy* impurity function.

Case study	Impurity function	Evaluation measure	Mean	Min	Max	Stdev
First	Entropy	Accuracy	**0.844**	**0.818**	**0.860**	0.014
First	Entropy	AUC	**0.847**	**0.831**	**0.860**	0.010
First	Misclassification	Accuracy	0.840	0.803	0.858	0.018
First	Misclassification	AUC	0.843	0.822	0.858	0.012
Second	Entropy	Accuracy	**0.863**	**0.838**	**0.878**	0.013
Second	Entropy	AUC	**0.868**	**0.839**	**0.885**	0.014
Second	Misclassification	Accuracy	0.848	0.838	0.863	0.008
Second	Misclassification	AUC	0.849	0.839	0.863	0.007
Third	Entropy	Accuracy	**0.872**	**0.838**	**0.885**	0.011
Third	Entropy	AUC	**0.873**	**0.839**	**0.886**	0.011
Third	Misclassification	Accuracy	0.868	0.838	0.883	0.013
Third	Misclassification	AUC	0.869	0.839	0.883	0.012

Table 4. Obtained results on the considered case studies.

We also note that the best *Acc* and *AUC* values are obtained in the third case study using the *entropy* impurity function. The variations of the *Acc* and *AUC* values obtained for this experiment are depicted in Figure 3, respectively in Figure 4. The small values obtained for the standard deviation of the *Acc* and *AUC* values (0.011) indicate a good precision of the decision tree model.

Due to the fact that the removal of attributes in the evaluation process does not provoke important fluctuations of *Acc* and *AUC* measurements (Figure 3 and Figure 4) in the results obtained, we can conclude that our tree is well constructed from the beginning. The case in which we test only two attributes is an expected exception of the rule above, because a set consisting of only two attributes is definitely too small to obtain good results.

Most of the approaches existing in the literature for determining the sex of skeletal remains are based on bone measurements, statistical methods or DNA and gene analysis. There is only one approach in the literature that uses

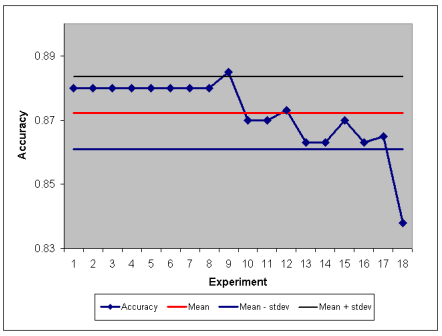

FIGURE 3. *Acc* values obtained on the third case study using the *entropy* impurity function

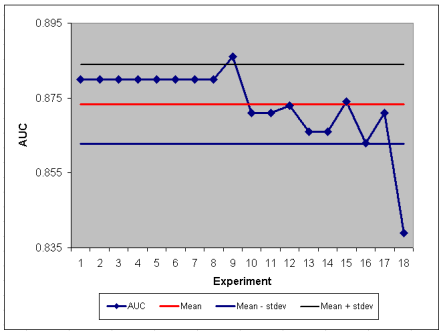

FIGURE 4. *AUC* values obtained for the third case study using the *entropy* impurity function

the same data sets as in our paper, a discriminant analysis method which was introduced in [1]. Five discriminant functions were used in this paper for the data set we have considered in our first case study and four functions were used for the data set considered in our second case study. For estimating the performance of the gender prediction task, only the accuracy is reported in [1], thus we will also use for comparison this evaluation measure.

Table 5 comparatively presents the best *Acc* values (minimum, maximum, mean and standard deviation) reported by our approach and the discriminant analysis method from [1]. The best obtained values are marked with bold. We note that [1] uses the same evaluation method as in our paper, i.e "leave-one-out" cross-validation.

Case study	Classifier	Mean	Min	Max	Stdev
First	Our DT approach	0.844	0.818	0.86	0.014
First	Discriminant functions [1]	0.838	0.81	0.865	0.026
Second	Our DT approach	0.863	0.838	0.878	0.013
Second	Discriminant functions [1]	0.843	0.795	0.875	0.034
Third	**Our DT approach**	**0.872**	**0.838**	**0.885**	**0.011**
Third	Discriminant functions [1]	-	-	-	-

TABLE 5. Comparative results on the considered case studies.

From Table 5 we observe that our DT model is more performant that the discriminant analysis method from [1]. One can easily observe the difference between the maximum value of *accuracy* (0.885) obtained using decision

tree learning method, in comparison with the best result obtained until now in literature using *discriminant functions* [1] (0.875). We mention that our approach achieves best accuracy on the data set containing both radius and ulna measurements using the feature set $\{F14, F7, F17, F4, F9, F8, F2, F11, F1, F15, F13\}$, while the method from [1] uses the entire feature set without any form of feature selection.

Stevenson et al. have approached in [13] the sex prediction problem using CHAID, a type of tree based technique based on the Chi-square test [7] that determines the next best split at each node in the tree. The approach from [13] differs from ours, since the tree is built differently than in our approach. Experiments were performed on 304 remains of Americans, European and African ancestry who died between 1915 and 1955 and accuracies between 85% and 85.5% were obtained. The data set used in [13] is not publicly available, that is why a fair comparison with our approach can not be made. Still, if we look at the obtained accuracies, we observe that our best accuracy exceeds the maximum accuracy reported in [13].

A comparison of our approach to other existing approaches from the literature is hard to be made, since in the existing approaches the experiments are performed on data sets which differ from the one considered in this paper. That is why a comparison that is based only on the obtained accuracies is not relevant, since the data sets used in the experiments are not the same. The good performances of our DT model on the case studies considered in this paper makes us believe that it will perform well when applied on other data sets.

Based on the experimental results we have obtained, we can conclude that decision trees are machine learning models which seem to offer accurate predictions for the gender detection problem. Moreover, when compared to other machine learning models, we consider that decision trees are better alternatives for bioarchaeologists. Decision trees are able to provide a set of rules indicating the way the prediction was made and this would be of great interest for bioarchaeologists.

6. Conclusions and further work

We have proposed in this paper an inductive learning methods for detecting the sex of human remains from bone measurements, which is based on decision trees. The experimental results obtained on three open-source data sets reveal that our approach outperforms similar approaches from the literature.

Further work will be done in order to extend the experimental evaluation of the proposed machine learning based model on real data sets [8] to better investigate their performance. We also plan to investigate the use of *random*

forests [3] and *fuzzy* [16] tree based models, as well as to further consider techniques for feature selection and for data discretization.

Acknowledgments

This work was supported by a grant of the Romanian National Authority for Scientific Research, CNCS–UEFISCDI, project number PN-II-RU-TE-2014-4-0082.

References

[1] I. L. O. Barrier. Sex determination from the bones of the forearm in a modern South African sample. PhD thesis, University of Pretoria, 2007.
[2] S. Bell and R. Jantz. Neural network classification of skeletal remains. In G. Burenhult and J. Arvidsson, editors, *Archaeological Informatics: Pushing The Envelope. CAA2001*, pages 205–212. Archaeopress, Oxford, 2001.
[3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
[4] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone. Classification and Regression Trees. Chapman and Hall, 1984.
[5] M. Faerman, D. Filon, G. Kahila, C. L. Greenblatt, P. Smith, and A. Oppenheim. Sex identification of archaeological human remains based on amplification of the X and Y amelogenin alleles. *Gene*, 167(1-2):327–32, 1995.
[6] T. Fawcett. An introduction to ROC analysis. *Pattern Recogn. Lett.*, 27(8):861–874, 2006.
[7] P. E. Greenwood and M. S. Nikulin. *A Guide to Chi-Squared Testing*. Wiley Series in Probabilities and Statistics. Wiley, 1996.
[8] Institute of Interdisciplinary Research in Bio-Nano-Sciences. http://bionanosci.institute.ubbcluj.ro/.
[9] N. Lavrac, B. Kavsek, P. A. Flach, and L. Todorovski. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188, 2004.
[10] T. M. Mitchell. *Machine learning*. McGraw-Hill, Inc. New York, USA, 1997.
[11] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.
[12] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.
[13] J.C. Stevenson, E.R Mahoney, P.L. Walker, and P.M. Everson. Technical note: prediction of sex based on five skull traits using decision analysis (CHAID). *Am J Phys Anthropol*, 139(3):434 – 441, 2009.
[14] S. Tuffary. *Data Mining and Statistics for Decision Making*. John Wiley and Sons, 2011.
[15] G. Wahba, Y. Lin, and H. Zhang. GACV for support vector machines, or, another way to look at margin-like quantities. *Advances in large margin classifiers*, 298–309, 1998.
[16] L. A. Zadeh. A summary and update of "fuzzy logic". In *2010 IEEE International Conference on Granular Computing, GrC 2010, San Jose, California, USA, 14-16 August 2010*, pages 42–44, 2010.

Department of Computer Science,, Faculty of Mathematics and Computer Science,, Babeş-Bolyai University, Kogălniceanu 1, Cluj-Napoca, 400084, Romania.

E-mail address: {mircea, gabis}@cs.ubbcluj.ro, pmir1335@scs.ubbcluj.ro