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

A NEW UNSUPERVISED LEARNING BASED APPROACH FOR GENDER DETECTION OF HUMAN ARCHAEOLOGICAL REMAINS

IOAN-GABRIEL MIRCEA, SERGIU-GEORGE LIMBOI,
AND MARA-RENATA PETRUŞEL

Abstract. Detecting the gender of human skeletal remains is an important problem within archaeology, since it is essential for understanding the characteristics of past societies. We approach in this paper, from a machine learning perspective, the problem of sex identification of human skeletal remains from bone measurements. In order to partition a group of skeleton remains according to their gender, different clustering algorithms are considered. Computational experiments carried out on publicly available archaeological data sets show a good performance of the proposed clustering approaches with respect to existing similar approaches from the literature.

1. Introduction

Machine learning (ML) [20] is a challenging field of computational intelligence whose goal is to develop computational systems which are able to improve their performance through experience, by learning some specific domain knowledge.

We approach in this paper, from a machine learning perspective, the problem of sex identification of human skeletal remains from bone measurements. Many studies in the literature propose various approaches for detecting the gender of human skeletons, most of which are based on bone measurements, DNA and gene analysis or different statistical methods.

In this paper we aim at introducing a novel machine learning approach based on clustering for solving the sex detection problem. The case studies used in the experiments for evaluating the performance of our model show that

---

Received by the editors: March 31, 2015.

2010 *Mathematics Subject Classification.* 68T05,62H30.

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, sex determination, machine learning, clustering.

our approach overperforms the existing similar approaches from the literature. To the best of our knowledge, our approach is novel, since there are no similar approaches in the literature.

The rest of the paper is organized as follows. Section 2 presents the problem approached in this paper, highlighting the motivation of our work, as well as different approaches existing in the literature for gender detection of human skeletal remains. The fundamentals of clustering are given in Section 3. Section 4 introduces our novel unsupervised learning based approach for the determination of sex in human skeletons. Experimental evaluations are given in Section 5, while comparisons to related work from the literature are presented in Section 6. Section 7 contains the conclusions of the paper and indicates future research directions.

2. Gender detection of human skeletal remains

Identifying the gender of archaeological human remains is essential for studying the gender differences in past populations [8]. This contributes to a better understanding of the social position and attributions of each gender in society. The sex classification task is a very delicate one and is highly influenced by the historical period and the geographic origin of the skeleton.

Usually the skull and the pelvis measurements are used for determining the gender, but other bones from the body may be used as well. Thus, measurements of the arm and leg bones may be important in the sex detection process. Though, in the case of subadult individuals, sex identification based on bone measurements is not fully accurate.

2.1. **Related work.**

Barrier's idea [3] of elaborating a discriminant function formula for sex determination based on researching forearm bones: radius and ulna, came from the great number of South African skeletal remains. The accuracy obtained was between 76% and 86%, therefore these bones are moderate discriminators for the given assignment.

Wada [25] suggested a distinct discriminant function formula, using for sex determination the radius on 63 instances, both males and females. More precisely, there were 35 correctly classified instances for females (97.2%), respectively 25 for males (92.6%). In order to increase the accuracy of the results, the unclear area of discriminant was diminished. Better results did not fail to appear.

Traditional morphometric analysis cannot solve the sex determination problem for partial and immature skeletal remains. The bests results seem to be provided by a new method built on amplification of the single-copy amelogenin-encoding gene projected by [8]. From a total consisting of 22 skeletal remains, 18 were properly classified. The skeletal remains dated from periods starting from 200 years to 8000 years ago and, among them, young children skeletons were also studied.

There exists another method that deals with sex determination for incomplete or immature skeletal remains suggested by [16]. This method uses preserved DNA and Polymerase Chain Reaction. The advantage of this technique is that the research can be done even on a single unit of DNA or on burnt or charred material.

Anthropologists accept two distinct methods for gender detection, particularly morphological, or non-metrical, and metrical, including geometric morphometrics. Besides these, molecular techniques concentrating on DNA research have been included. Although every single method is confined by a number of constraints [5], they complement each other and produce better results when implemented in conjunction with each other.

Weiss disagrees in [26] with a statistically important bias in sex determination affecting adult skeletons. The causative factor for the bias is (approximately 12%) in favor of male population. In consequence, it is stated the nature of the secondary sex features in the bone. The size of the skull, the femoral head size and the rugosity of the bone illustrate these features. So, when a specimen has medium size, rugosity or development is discovered, there is an inclination to classify it as male [26]. In support of this hypothesis comes the research on adult sex ratios of skeletal populations from various time periods, cultures and geographic areas and its importance illustrates the need of taking into consideration the bias when sex-specific analysis of skeletal material is tried.

Bruzek [7] proposed a visual technique for sex determination that studies only the os coxae, the bone thought to provide the best accuracy for this problem. The approach was based on a formula-based methodology as an alternative to the simple visual remarks. The described changes involve the minimization of observer subjectivity and a better probability of a correct diagnosis with separate fragments of the bone. Tests have been made on a sample of 402 adults of French and Portuguese origins. A correct sexual determination was acquired in 95% of cases.

Another method for gender detection is odontometrics, which was proposed by Vodanovic [24]. The practicability of this method was doubted by the significance of the teeth in cases of poor maintenance of skeletal remains and the better accuracy obtained by adding the odontometric parameters in the procedure that previously was using only craniofacial characteristics. Nevertheless, there was stated as a disadvantage of this method the lack of referent odontometric values for comparison.

A distinct technique for sex determination was suggested by Vanharová and Drozdová in [23]. Analysis of DNA was applied to skeletal remains of children, having results consistent with archeological grave remains and body imposition. The molecular method is not restricted by physical fragmentation and can be applied on immature individuals, in case skeletally based identification is not possible. Although, the molecular method is restricted by some constraints including molecular contamination or molecular preservation.

Other techniques that use analysis of the bones have been attempted with success for skeletal gender detection. Measurements of the bones of the hands and feet (metatarsals,metacarpals, phalanges) were used in [22] and [9] with accuracy rates of 80% and 84-92%. Stature and gender were estimated by use of foot measurements in [27] with accuracies of 95.6% for the right foot measurements, respectively 96.4% for the left foot.

Several computational intelligence techniques have also been investigated for the detecting the sex of human skeletons. Neural networks have been applied in [4] to classify archaeological remains based on osteological measurements. Two archaeological databases were investigated and an accuracy ranging between 81-88 % was obtained. Afrianty et al have used in [1] back-propagation neural networks for gender detection on a data set consisting of remains characterized by pelvic and pattela bones measurements. Accuracies of 86.6%- 98.3% were obtained on a data set generated based on statistical values reported in the literature [2].

Stevenson et al have approached in [21] the sex prediction problem using CHAID (Chi-square automatic interaction detection). CHAID is a type of decision tree technique based on the Chi-square test [11] to determine the best next split at each node in the tree. Experiments are performed on 304 remains of American, European and African ancestry who died between 1915 and 1955 and accuracies between 85% and 85.5% were obtained.

3. BACKGROUND

In this section we are providing a brief background on clustering , the unsupervised machine learning method which will be further used in our approach.

3.1. **Clustering.**

Unsupervised classification, or *clustering* is an important activity within data mining. The goal of clustering is to identify groups (classes or clusters) inside a given data set of instances (objects) [12], and it is considered the most important unsupervised learning problem. The resulting groups of instances are formed such that the instances within each cluster are more similar to one another than the instances belonging to different clusters. An important notion in the clustering process is the *similarity* (or *dissimilarity*) between the instances. The measure used for expressing the dissimilarity between the instances can be any metric or semi-metric function (e.g. *Euclidian distance*, *Manhattan distance*, *Minkovski distance*, etc).

Most clustering algorithms available in the literature [12], [14]are based on two techniques known as *partitional* and *hierarchical* clustering.

In the following, we present a brief overview of the partitional clustering methods which we will use for gender detection.

Given n instances and a number k, a partitioning technique splits the instance set into k distinct and non-empty groups (clusters). The partitioning process is iterative and it stops when a "good" partitioning is obtained. The partitional clustering algorithms try to minimize a criteria (usually defined as a squared error function) and generally they converge to local optima [14].

The *k-means* algorithm divides a set of n objects into k distinct and non-empty clusters[14]. The algorithm starts with k initial centroids, then the clusters and their centroids are iteratively recalculated (each instance is put into the closest centroid) until convergence is reached.

In the *k-medoids* or *PAM* (Partitioning around medoids) clustering algorithm [15], each cluster is characterized by one of the instances within the cluster, and this representative instance is called *medoid*. The *k-medoid* algorithm starts with k initial medoids for the clusters, then the clusters and their medoids are iteratively recalculated (each instance is put into the closest medoid) until convergence is reached. At a given iteration, a medoid of a cluster is replaced with another object from the cluster, if it reduces the total distance of the obtained clustering [15].

Certainly, the initial medoids or centroids impact the performance of the PAM and $k - means$ algorithms. Thus, there is no guarantee for obtaining an optimal solution. Another drawback of the algorithms is that the number of clusters have to be initially specified.

4. OUR APPROACH

This section presents our unsupervised learning approach using *clustering* for determining the gender of human skeletons from the length of long bones.

4.1. Model.

We are considering a data set $\mathcal{AR} = \{ar_1, ar_2, \ldots, ar_n\}$ in which each instance ar_i represents an archaeological remain. Each skeleton is represented by m characteristics (features) $f_1, f_2, \ldots, f_m$ which correspond to different measurements that were performed on it. Usually, the measurements are numerical values and correspond to several significant bones in the body. Thus, an instance ar_i is characterized by an m-dimensional vector, i.e $ar_i = (ar_{i1}, ar_{i2}, \ldots, ar_{im})$ where ar_{ij} $\forall 1 \leq i \leq n$ represents the value of measurement f_j applied to the skeletal remain ar_i.

4.2. **Data pre-processing and feature selection.** The first step in building the machine learning model is the *data pre-processing* step. During this step, an analysis is performed in order to determine those measurements (features) that highly influence the gender classification task. We will use the *information gain* measure which expresses the expected reduction in entropy determined by partitioning the instances according to a given feature [18].

First, we are computing the information gain measure for each feature from the feature set. For computing the information gain of a feature, the values of the feature are first discretized (the feature domain is divided in ten intervals of equal size). Let us consider that $f_{o_1}, f_{o_2}, \ldots, f_{o_m}$ are the features in the increasing order of their information gain. Then, we are trying to remove subsets of features considering the determined order, i.e $\{f_{o_1}\}$, $\{f_{o_1}, f_{o_2}\}$, etc. In order to decide the impact of removing the subset $\{f_{o_1}, \ldots, f_{o_k}\}$ $(1 \leq k \leq m-1)$ of features from the initial feature set, a *decision tree* is built from the set of instances and the currently selected features. For estimating the performance of the constructed decision tree, a k-fold cross-validation is applied on the training data set. Finally, we decide to remove (from the initial feature set) the subset $\{f_{o_1}, \ldots, f_{o_p}\}$ $(1 \leq p \leq m-1)$ of features for which the corresponding decision tree provided a unique maximum average accuracy during the k-fold cross-validation $(k \geq 10)$.

After the feature selection technique is applied, the input data is scaled to [0,1].

4.3. **Building the model.** After the data set is pre-processed as indicated above, the k-means and PAM clustering algorithms will be applied in order to build the unsupervised learning model.

Before applying the partitional clustering algorithms, the number k of clusters have to be identified. For determining the number of clusters, we are using a cluster validity index, the *Dunn index* [19]. This index expresses the quality of a given partition (clustering). The greater the value of this index, the better a partition is, therefore the *Dunn index* should be maximized in order to obtain better partitions. The clustering algorithms are applied considering different values for the number of clusters (starting from two clusters). The number k of clusters which provided the partition with the highest *Dunn index* will be reported as the correct number of clusters, i.e the number of clusters that have to be identified in data.

After the number of clusters was determined, we are using an heuristic method for selecting two initial representative instances (i.e skeletons), which will be used as initial medoids/centroids in the clustering process. The heuristic is described below.

(i) The first representative skeleton is the most "distant" one from all other skeletons from the data set (i.e the skeleton whose average distance from all other skeletons is maximum).

(ii) In order to select the second representative instance we reason as follows. For each remaining instance (that was not selected), we compute the minimum distance (*dist*) from the instance and the already chosen representative instance. As the second representative instance we will choose the skeleton that maximizes *dist*.

The representative instances identified using the heuristic below will be selected as initial medoids/centroids in the clustering process. After the initial medoids/centroids are selected, the PAM/k-means algorithm will be applied and the output partition is reported. We mention that, if a cluster becomes empty during the iterative process, the number of clusters will be decreased.

After the unsupervised learning model was built, a testing step is performed in order to evaluate its performance. Details about the testing step will be given in Section 5.2.1.

5. Experimental evaluation

This section contains the experimental evaluation of the clustering algorithms (described in Section 4) on three case studies which were conducted starting from two data sets obtained from the literature [3]. This data sets were used for determining the gender of South African skeletal remains based on different measurements of the forearm bones, ulna respectively the radius.

5.1. **Case studies.** The data set from [3] 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 were taken from the ulna bone. The skeletal remains represent black South Africans from the 19^{th} and 20^{th} centuries, being born between 1863 and 1996.

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

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 [3]: 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).

As the second case study, nine measurements of the ulna bone are used for 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 previously described data sets were previously used in [3] for the gender identification task. We are considering in this paper as our third case study the data set which contains both radial and ulnar measurements. Consequently, in this data set, each skeleton (instance) will be represented by 19 measurements (features): the first ten are the radial measurements (as for the first case study) and the next nine features represent the ulnar measurements (as for the second case study).

5.2. **Results.** For all the case studies considered for evaluation, the methodology presented in Section 4 will be applied. After the feature selection step, the number of clusters which have to be determined in the data set, as well as the initial centroids/medoids are identified as indicated in Section 4.3. We mention that for all the considered case studies, the number k of desired clusters (heuristically determined using the *Dunn* index) is two. Obviously, two clusters are expected, since the problem we are approaching in this paper is a binary classification problem.

5.2.1. *Evaluation measures.* After two clusters are detected in data using the clustering algorithm (k-means or k-medoids), we identified which is the male cluster and which is the female cluster. Since the labels of the instances within the input data set are known, the cluster which has the number of males greater than the number of females will be considered the male cluster. The remaining cluster will be considered a female cluster.

Since the gender detection task is a binary classification problem, the confusion matrix will be computed. We assume in the following that that the male class is the positive one and the female class is the negative one. Thus, we have to compute the number of true positives (TP - number of males correctly identified), the number of true negatives (TN - number of females correctly identified), the number of false positive (FP - number of misclassified males) and the number of false negative (FN - number of misclassified females).

Two evaluation measures will be further used in order to test the performance of the clustering algorithms for gender identification. These measures are *external evaluation* measures since their computation requires the class label of each instance.

(1) The *accuracy* (denoted by Acc) measures the percentage of instances that are correctly classified, $Acc = \frac{TP+TN}{TP+TN+FP+FN}$.
(2) The *Area under the ROC curve* measure (denoted by AUC) which is considered as one of the best evaluation measures used to compare classifiers [17, 10]. The AUC measure expresses the area under the ROC curve (Receiver Operating Characteristics curve). The ROC curve is obtained by connecting the ($recall$, 1-$specificity$) point to the points at (0,0) and (1,1) [10]. The *recall* of a binary classifier is computed as $recall = \frac{TP}{TP+FN}$ and the *specificity* of the classifier is computed as $specificity = \frac{TN}{TN+FP}$.

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

5.2.2. *First case study.* As mentioned in Section 5.1, there are 10 initial features used for the the gender classification task. First, the information gain for the features is determined. The obtained values are depicted in Figure 1.

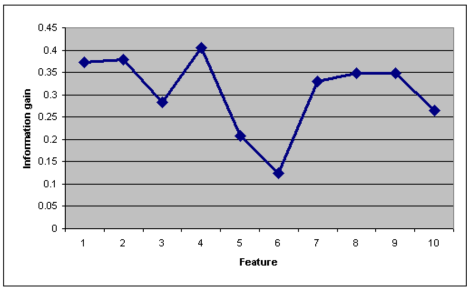

FIGURE 1. Information gain for the features from the first case study

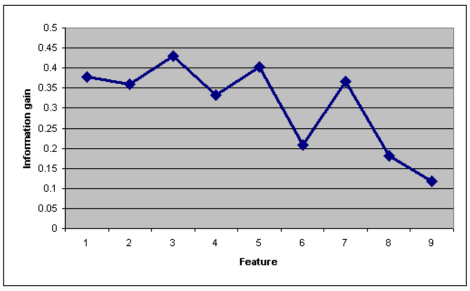

FIGURE 2. Information gain for the features from the second case study

For the first case study, the features in their increasing order of the information gain are $\{F6, F5, F10, F3, F7, F9, F8, F1, F2, F4\}$ (see Figure 1). We applied the feature selection step described in Section 4.2 in order to identify a subset of relevant features. The resulting set of features is $\{F1, F2, F4\}$ (i.e the subset of features which are removed is $\{F6, F5, F10, F3, F7, F9, F8\}$) since it provided a *decision tree* having a maximum accuracy of 86.25% using 10-fold cross validation. Then, the data is scaled to [0,1].

Table 1 presents the values of the evaluation measures (Section 5.2.1) obtained by applying the k-means and PAM algorithms on the first data set, considering different set of features. The best obtained results are marked with bold. We notice that for the set of features identified after the feature selection step, the values for the *Acc* and *AUC* measures are very close (for both algorithms) to the maximum values. We note that when using only the feature set $\{F2\}$, we obtain only one cluster, which is not our purpose. Thus, we do not report in Table 1 the results obtained when using only feature $F2$ for classification.

Feature set	**k-means**		**k-medoids**	
	Acc	**AUC**	**Acc**	**AUC**
$\{F6, F5, F10, F3, F7, F9, F8, F1, F2, F4\}$	0.837	0.837	0.837	0.837
$\{F5, F10, F3, F7, F9, F8, F1, F2, F4\}$	0.837	0.837	0.837	0.837
$\{F10, F3, F7, F9, F8, F1, F2, F4\}$	0.837	0.837	0.832	0.832
$\{F3, F7, F9, F8, F1, F2, F4\}$	0.835	0.835	0.830	0.830
$\{F7, F9, F8, F1, F2, F4\}$	0.847	0.847	0.842	0.842
$\{F9, F8, F1, F2, F4\}$	0.842	0.842	**0.845**	**0.845**
$\{F8, F1, F2, F4\}$	**0.855**	**0.855**	0.835	0.838
$\{F1, F2, F4\}$	0.850	0.851	0.842	0.843
$\{F2, F4\}$	0.822	0.824	0.825	0.826

TABLE 1. Results obtained for the first case study.

Table 4 depicts the minimum, maximum, mean and standard deviation of the *Acc* and *AUC* values obtained on the first case study.

5.2.3. *Second case study.* For the second case study, the features in their increasing order of the information gain are $\{F9, F8, F6, F4, F2, F7, F1, F5, F3\}$ (see Figure 2). We applied the feature selection step described in Section 4.2 in order to identify a subset of relevant features. The resulting set of features is $\{F7, F1, F5, F3\}$ (i.e the subset of features which are removed is $\{F9, F8, F6, F4, F2\}$) since it provided a *decision tree* having a maximum accuracy of 87.25% using 13-fold cross validation. Then, the data is scaled to [0,1].

Table 2 presents the values of the evaluation measures (Section 5.2.1) obtained by applying the k-means and PAM algorithms on the second data set, considering different set of features. For both algorithms, the best obtained results are marked with bold and correspond to the feature set that was identified during the feature selection step described above.

Table 4 illustrates the minimum, maximum, mean and standard deviation of the *Acc* and *AUC* values obtained on the second case study.

Feature set	**k-means**		**k-medoids**	
	Acc	**AUC**	**Acc**	**AUC**
$\{F9, F8, F6, F4, F2, F7, F1, F5, F3\}$	0.865	0.865	0.875	0.875
$\{F8, F6, F4, F2, F7, F1, F5, F3\}$	0.870	0.870	0.875	0.875
$\{F6, F4, F2, F7, F1, F5, F3\}$	0.867	0.867	0.870	0.870
$\{F4, F2, F7, F1, F5, F3\}$	0.872	0.872	0.860	0.860
$\{F2, F7, F1, F5, F3\}$	0.877	0.878	0.880	0.881
$\{F7, F1, F5, F3\}$	**0.877**	**0.878**	**0.882**	**0.884**
$\{F1, F5, F3\}$	0.862	0.862	0.862	0.862
$\{F5, F3\}$	0.840	0.840	0.847	0.847
$\{F3\}$	0.827	0.828	0.835	0.836

TABLE 2. Results obtained for the second case study.

5.2.4. *Third case study.* For the third case study, the features in their increasing order of the information gain are $\{F19, F6, F18, F16, F5, F10, F3, F7, F14, F9, F8, F12, F17, F1, F2, F11, F15, F4, F13\}$. The obtained values for the features information gain are depicted in Figure 3.We applied the feature selection step described in Section 4.2 in order to identify a subset of relevant features. The resulting set of features is $\{F2, F11, F15, F4, F13\}$ (i.e the subset of features which are removed is $\{F19, F6, F18, F16, F5, F10, F3, F7, F14, F9, F8, F12, F17, F1\}$) since it provided a *decision tree* having a maximum accuracy of 88.25% using 15-fold cross validation. Then, the data is scaled to [0,1].

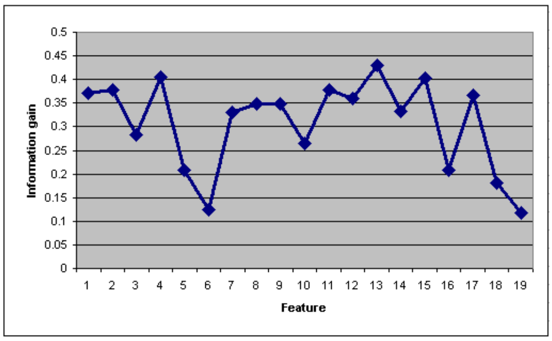

FIGURE 3. Information gain for the features from the third case study

Table 3 presents the values of the evaluation measures (Section 5.2.1) obtained by applying the k-means and PAM algorithms on the third data set,

Feature set	**k-means**		**k-medoids**	
	Acc	**AUC**	**Acc**	**AUC**
$\{F19, F6, F18, F16, F5, F10, F3, F7, F14, F9,$ $F8, F12, F17, F1, F2, F11, F15, F4, F13\}$	0.857	0.858	0.867	0.868
$\{F6, F18, F16, F5, F10, F3, F7, F14, F9, F8, \}$ $F12, F17, F1, F2, F11, F15, F4, F13\}$	0.857	0.858	0.840	0.841
$\{F18, F16, F5, F10, F3, F7, F14, F9, F8, F12, \}$ $F17, F1, F2, F11, F15, F4, F13\}$	0.857	0.858	0.867	0.868
$\{F16, F5, F10, F3, F7, F14, F9, F8, F12, F17, \}$ $F1, F2, F11, F15, F4, F13\}$	0.862	0.863	0.865	0.866
$\{F5, F10, F3, F7, F14, F9, F8, F12, F17, F1, \}$ $F2, F11, F15, F4, F13\}$	0.862	0.863	0.860	0.860
$\{F10, F3, F7, F14, F9, F8, F12, F17, F1, F2, \}$ $F11, F15, F4, F13\}$	0.862	0.863	0.857	0.858
$\{F3, F7, F14, F9, F8, F12, F17, F1, F2, F11, \}$ $F15, F4, F13\}$	0.867	0.868	0.865	0.865
$\{F7, F14, F9, F8, F12, F17, F1, F2, F11, F15, \}$ $F4, F13\}$	0.875	0.877	0.857	0.858
$\{F14, F9, F8, F12, F17, F1, F2, F11, F15, F4\}$ $F13\}$	0.870	0.873	0.872	0.878
$\{F9, F8, F12, F17, F1, F2, F11, F15, F4, F13\}$	0.865	0.865	0.867	0.868
$\{F8, F12, F17, F1, F2, F11, F15, F4, F13\}$	0.880	0.885	0.877	0.885
$\{F12, F17, F1, F2, F11, F15, F4, F13\}$	0.887	0.890	0.870	0.879
$\{F17, F1, F2, F11, F15, F4, F13\}$	0.880	0.882	0.852	0.856
$\{F1, F2, F11, F15, F4, F13\}$	**0.890**	**0.892**	0.882	0.883
$\{F2, F11, F15, F4, F13\}$	0.882	0.884	**0.890**	**0.890**
$\{F11, F15, F4, F13\}$	0.857	0.857	0.860	0.860
$\{F15, F4, F13\}$	0.845	0.845	0.852	0.852
$\{F4, F13\}$	0.860	0.860	0.845	0.845
$\{F13\}$	0.827	0.828	0.835	0.836

TABLE 3. Results obtained for the third case study.

considering different set of features. The best obtained results are marked with bold. We notice that for the k-medoids algorithm, the best results corresponds to the feature set that was identified during the feature selection step described above. For the k-means algorithm, for the set of features identified after the feature selection step, the values for the Acc and AUC measures are very close to the maximum values.

The minimum, maximum, mean and standard deviation of the Acc and AUC values obtained on the third case study are given in Table 4.

Case study	**Algorithm**	**Acc**				**AUC**			
		Mean	Min	Max	Stdev	Mean	Min	Max	Stdev
First	k-means	0.840	0.822	0.855	0.009	0.840	0.824	0.855	0.009
First	k-medoids	0.836	0.825	0.845	0.006	0.836	0.826	0.845	0.006
Second	k-means	0.861	0.827	0.877	0.017	0.862	0.828	0.878	0.017
Second	k-medois	0.865	0.835	0.882	0.015	0.865	0.836	0.884	0.015
Third	k-means	0.865	0.826	0.89	0.015	0.866	0.828	0.892	0.015
Third	k-medoids	0.862	0.835	0.89	0.013	0.863	0.836	0.890	0.014

TABLE 4. Results for all performed experiments.

6. DISCUSSION AND COMPARISON TO RELATED WORK

Analyzing the experimental results presented in Section 5 we observe the following.

For the first and the third case studies, the maximum *Acc* and *AUC* values are obtained using the k-means algorithm. The k-medoids algorithm provided the maximum *Acc* and *AUC* values for the second case study. We notice that, for each case study, the values reported by the two clustering algorithms are sensitively equal. From all the performed experiments, the best values were reported for the k-means algorithm, on the third case study (an *Acc* of 0.89 and an *AUC* of 0.892). Moreover, we observe, for each performed experiment, a small value for the standard deviation, which indicates a good precision of the obtained results.

As we have shown in Section 2.1, most of the approaches existing in the literature for determining the sex of skeletal remains are based on bone measurements, DNA and gene analysis or different statistical methods. As far as we know, there are no approaches that use clustering for learning to identify the gender of human skeletons.

There is only one approach in the literature that uses the same data sets as in our paper, a discriminant analysis method which was introduced in [3]. 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 [3], thus we will also use for comparison this evaluation measure.

Table 5 comparatively presents the values for the *Acc* measures reported by our clustering algorithms (using the feature selection method indicated in Subsection 4.2), as well as the average accuracy value reported by the discriminant analysis method from [3]. We note that in [3] the evaluation is made using the "leave-one-out" cross-validation method.

Case study	Algorithm	Accuracy
First	k-means	**0.850**
First	k-medoids	0.842
First	Discriminant functions [3]	0.838
Second	k-means	0.877
Second	k-medoids	**0.882**
Second	Discriminant functions [3]	0.843
Third	k-means	0.882
Third	k-medoids	**0.890**
Third	Discriminant functions [3]	-

Table 5. Comparative results on the considered case studies.

From Table 5 we observe that our clustering algorithms outperform, for all the considered case studies, the discriminant analysis method from [3] (considering the average accuracy reported). For each case study, the best obtained accuracy is highlighted. We also observe that the maximum value for the accuracy was obtained on the third case study using the k-medoids method. As in the third experiment there was no value for the discriminant functions, in the following graphs, the average of the results obtained in the first and second cases was considered. Figures 6 indicate, for each case study considered in our experiments, the comparative results from Table 5, emphasizing the good performance of our clustering based methods.

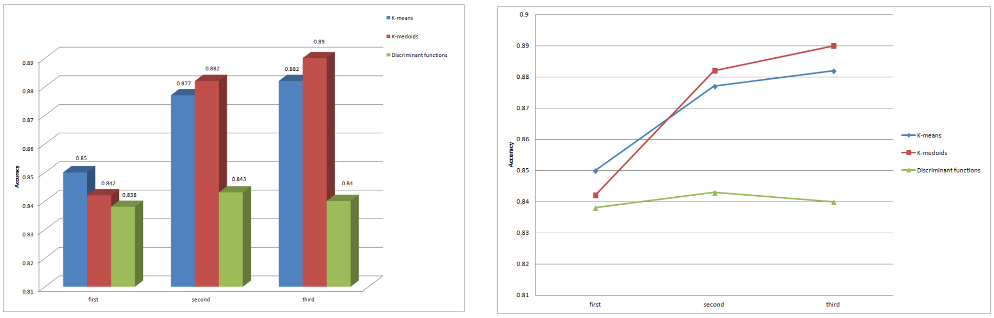

Figure 4. Comparative results.

Considering all case studies, an average accuracy of 0.8705 was obtained. The 95% Confidence Interval [6] for the average value is (0.86, 0.89). Thus, there is a 95% confidence that the accuracy of the partition obtained using clustering exceeds the best value reported in the literature with at least 1.7% and at most 4.7%. Other approaches existing in the literature for gender detection perform experiments on various data sets which are different from the ones we are considering in this paper. Thus, a comparison with these approaches is hard to be conducted.

7. Conclusions and further work

In this paper we proposed an unsupervised machine learning based approach for detecting the gender of human skeletons from bone measurements. The experimental evaluation is performed on two publicly available human skeletal remains data sets and three case studies were conducted in order to test the performance of our technique. The experimental results obtained on two open-source data sets reveal that our approaches outperform the similar approaches from the literature.

Experimental evaluations of the proposed clustering methods on real data sets [13] will be further conducted in order to better test their accuracy. We also plan to investigate the effectiveness of *fuzzy* clustering and decision tree learning [18] for the problem of gender detection of human skeletal remains.

References

[1] I. Afrianty, D. Nasien, M.R.A. Kadir, and H. Haron. Determination of gender from pelvic bones and patella in forensic anthropology: A comparison of classification techniques. In *Artificial Intelligence, Modelling and Simulation (AIMS), 2013 1st International Conference on*, pages 3–7, Dec 2013.

[2] Mitra Akhlaghi, Ardeshir Sheikhazadi, Ali Naghsh, and Gholamali Dorvashi. Identification of sex in iranian population using patella dimensions. *Journal of Forensic and Legal Medicine*, 17(3):150 – 155, 2010.

[3] Isabelle Linda Odile Barrier. Sex determination from the bones of the forearm in a modern South African sample. PhD thesis, University of Pretoria, 2007.

[4] 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.

[5] M.A. Bidmos, V.E. Gibbon, and G. Strkalj. Recent advances in sex identification of human skeletal remains in South Africa. *South African Journal Of Science*, 106, 2010.

[6] LD. Brown, TT. Cat, and A DasGupta. Interval Estimation for a proportion. *Statistical Science*, 16:101 – 133, 2001.

[7] J. Bruzek. A method for visual determination of sex, using the human hip bone. *American Journal Of Physical Anthropology*, 117(2):157–68, Feb 2002.

[8] 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.

[9] A.B. Falsetti. Sex assessment from metacarpals of the human hand. *J Forensic Sci.*, 40(5):774-776, 1995.

[10] Tom Fawcett. An introduction to ROC analysis. *Pattern Recogn. Lett.*, 27(8):861–874, 2006.

[11] Priscilla E. Greenwood and Michael S. Nikulin. *A Guide to Chi-Squared Testing.* Wiley Series in Probabilities and Statistics. Wiley, 1996.

[12] Jiawei Han. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[13] Institute of Interdisciplinary Research in Bio-Nano-Sciences. http://bionanosci.institute.ubbcluj.ro/.

[14] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323, September 1999.

[15] L. Kaufman and P. Rousseeuw. *Clustering by Means of Medoids.* Statistical Data Analysis Based on the L_1Norm and Related Methods, edited by Y. Dodge, North-Holland. pages 405416, 1987.

[16] Jan Kiesslich, Franz Neuhuber, Harald J. Meyer, Max P. Baur, and Jutta Leskovarm. DNA analysis on biological remains from archaeological findings - sex identification and kinship analysis on skeletons from Mitterkirchen, upper Austria. In: *Interpretierte Eisenzeiten. Fallstudien, Methoden, Theorie.*, eds. Raimund Karl - Jutta Leskovar. pages 147–154, 2005.

[17] Nada Lavrac, Branko Kavsek, Peter A. Flach, and Ljupco Todorovski. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188, 2004.

[18] Thomas M. Mitchell. *Machine learning.* McGraw-Hill, Inc. New York, USA, 1997.

[19] Malay K. Pakhira, Sanghamitra Bandyopadhyay, and Ujjwal Maulik. Validity index for crisp and fuzzy clusters. *Pattern Recognition*, 37(3):487 – 501, 2004.

[20] St. Russell and P. Norvig. *Artificial Intelligence - A Modern Approach.* Prentice Hall International Series in Artificial Intelligence. Prentice Hall, 2003.

[21] JC. Stevenson, ER Mahoney, PL. Walker, and PM. Everson. Technical note: prediction of sex based on five skull traits using decision analysis (CHAID). *Am J Phys Anthropol*, 139(3):434 – 441, 2009.

[22] D. Troy Case and Ann H. Ross. Sex determination from hand and foot bone lengths. *J Forensic Sci.*, 52(2):264-70, 2007.

[23] Michaela Vanharová and Eva Drozdová. Sex determination of 4000 years old children and juveniles skeletal remains from Eneolithic burial site Hostice 1 za Hanou (Czech Republic) by ancient DNA analysis. *Anthropological Review*, 71(1):63-70, 2008.

[24] Marin Vodanovic, Zeljko Demo, Vera Njemirovskij, Jadranka Keros, and Hrvoje Brkic. Odontometrics: a useful method for sex determination in an archaeological skeletal population? *Journal of Archaeological Science*, 34:905–913, Sep 2007.

[25] Yo Wada. Discriminant function for sex determination of ancient Iraqis based on radial measurements. *Anthropological Science*, 102:149-158, 1994.

[26] K.M. Weiss. On the systematic bias in skeletal sexing. *American Journal Of Physical Anthropology*, 37(2):239–49, Sep 1972.

[27] G. Zeybek, I. Ergur, and Z. Demiroglu. Stature and gender estimation using foot measurements. *Forensic Sci Int.*, 181(1-3):54.e1-5, 2008.

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

E-mail address: `mircea@cs.ubbcluj.ro, {lsir1296, pmir1335}@scs.ubbcluj.ro`