BABEŞ-BOLYAI UNIVERSITY
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Machine learning applied in bioarchaeology

Abstract of the PhD Thesis

PhD student: Mircea Ioan-Gabriel
Scientific supervisor: Prof. Dr. Czibula Gabriela

2016

Keywords: ***machine learning, bioarchaeology, self-organizing maps, decision trees, reinforcement learning, association rule mining, fuzzy systems.***

Content of the thesis

List of publications

Publications in ISI Web of Knowledge

Publications in ISI Science Citation Index Expanded

1. Gabriela Czibula, Istvan Gergely Czibula, Adela Sîrbu, **Ioan-Gabriel Mircea**. A novel approach to adaptive relational association rule mining. *Applied Soft Computing journal*, Vol. 36, pp. 519-533, 2015. (**IF = 2.810**)

2. Gabriela Czibula, Vlad-Sebastian Ionescu, Diana-Lucia Miholca and **Ioan-Gabriel Mircea**. Machine learning-based approaches for predicting stature from archaeological skeletal remains using long bone lengths. *Journal of Archaeological Science*, Vol. 69, pp. 8599, 2016. (**IF = 2.196**)

Publications in ISI Conference Proceedings Citation Index Expanded

1. **Ioan-Gabriel Mircea** and Gabriela Czibula, and Maria-Iuliana Bocicor. *A Q-learning approach for aligning protein sequences.* Proceedings of the 11th International conference on Intelligent Computer Communication and Processing, Cluj-Napoca, Romania, pp. 51-58, 2015.

2. Vlad-Sebastian Ionescu and **Ioan-Gabriel Mircea** and Diana-Lucia Miholca and Gabriela Czibula. *Instance Based Learning Approaches for predicting the height of human skeletons.* Proceedings of the 11th International conference on Intelligent Computer Communication and Processing, Cluj-Napoca, Romania, pp. 309-316, 2015.

3. Diana Miholca, Gabriela Czibula, **Ioan-Gabriel Mircea**, and Istvan-Gergely Czibula. Machine learning approaches for gender detection of human skeletal remains. *18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, under review, 2016.

4. **I.G. Mircea** A fuzzy decision tree based method for skeletal sex determination. In *Proceedings of the Symposium on Applied Computational Intelligence, Timisoara, Romania*, pp. 447-452, 2016

5. Maria-Iuliana Bocicor, **Ioan-Gabriel Mircea**, and Gabriela Czibula. A novel reinforcement learning based approach to multiple sequence alignment. *Proceedings of the 7th International Workshop on Soft Computing Applications* , to be published, 2016.

Papers published in international journals and proceedings of international conferences

1. **Ioan-Gabriel Mircea** and Maria-Iuliana Bocicor and Andra Dîncu. On reinforcement learning based multiple sequence alignment. *Studia Universitatis Babes-Bolyai Series Informatica*, LIX(2):50–65, 2014. (**indexed Mathematical Reviews**)

2. **Ioan-Gabriel Mircea** and Sergiu Limboi and Mara-Renata Petruşel. A New Unsupervised Learning Based Approach for Gender Detection of Human Archaeological Remains. *Studia Universitatis Babes-Bolyai Series Informatica*, LX(2):5–20, 2015. (**indexed Mathematical Reviews**)

3. **Ioan-Gabriel Mircea** and Gabriela Czibula and Mara-Renata Petruşel. Sex identification in archaeological remains using decision trees. *Studia Universitatis Babes-Bolyai Series Informatica*, LX(2):91–103, 2015. (**indexed Mathematical Reviews**)

4. **Ioan-Gabriel Mircea**. An Evaluation of Basic Techniques and Methods Used in Skin Color Detection. *Studia Universitatis Babes-Bolyai Series Informatica*, LVII(3):24-34, 2012. (**indexed Mathematical Reviews**)

The purpose of the present doctoral thesis is, first and foremost, to attenuate the gap between two fields of research that may seem, at a first glance, highly unrelated. One the one hand, machine learning represents an ambitious field in computer science whose purpose is to build algorithms that enable computing machines to learn, to use past experiences in order to better solve various problems. On the other hand, bioarchaeology is a challenging area of research which enables the use of biological knowledge and techniques in the study of ancient societies based on the archaeological materials unearthed on site. The two research fields, while each intrinsically exotic and intriguing, initially seem highly unrelated.

During the course of the doctoral studies, we have come to realize that even if the worlds of machine learning and bioarchaeology are surely different, governed by different customs and philosophies towards research, they are intricately connected and this connection will be more and more visible in the future as both lines of research will evolve into a state in which the collaboration between researchers of both domains will be inevitable.

The current thesis is a pleading towards interdisciplinarity. Its goal is to employ machine learning techniques to solve practical problems from the field of bioarchaeology, in an attempt to emphasize the interdependencies revealed by the current state of the art of the research currently conducted in both areas. It has become clearer and clearer that, in a research environment such as bioarchaeology, in which the amount of gathered raw data is increasing substantially as modern protocols require the collection of additional information in situ, be it related to ancient biology, forensic anthropology etc. that the need for a digital means of collection, representation and analysis for larger and larger amounts of data will eventually grow to be dire.

However, an increasing source of training data material is beneficial to the field of machine learning, which, as stated before, like a true branch of computer science, is governed by more peculiar laws than the humanities field of bioarchaeology. This antagonism that turns out to actually be useful to the research process was one of the reasons bioarchaeology has become the topic of interest for the present doctoral thesis. It is proof that interdisciplinarity may occur when what seems to be an issue to one field of research favors the research in the other field instead.

Having decided that bioarchaeology is to be the topic of the doctoral research, the next step was to highlight the most interesting problems raised in the field of bioarchaeology and whether a machine learning solution to those problems may be approached. In this respect, the Institute on Bio Nano Sciences has offered great help and expertise. As a result of the discussions with researchers from the institute, the first research topic was established as follows.

Since genetical studies have become more and more common in archaeology and even more frequent in forensic anthropology, the results of genetical tests on the osteological material obtained from the archaeological site, usually provided in the form of nucleotide sequences, may reveal highly interesting information about the individual they are describing or the population the individual was part of. A notable example that highlights the way molecular biology aids the archaeological research is an article [SKW$^+$10] that used genetic dissimilarities between species of turkey in the south of the United States of America to prove that it was the native populations who have first domesticated the famous American bird before the arrival of Columbus.

In order to achieve such compelling results with scientific accuracy, the analysis of the genetic material needs to be performed as thoroughly as possible and this is difficult in the context of a large amount of data to be analysed. Since dissimilarities (or similarities for that matter) can be easily found when two or more nucleotide sequences are aligned in such a way that their natural structure reveals exactly the positions or the regions where the differ. Usually these differences occur as a result of a mutation which leads to the apparition of new species with different abilities or aspect. Therefore, the process of sequence alignment is highly important to the accurate determination of dissimilarities in the genetic material extracted from archaeological sites.

With that in mind, we have tried to find if the problem of aligning multiple nucleotide sequences as accurately as possible can be solved by using machine learning techniques. As it a biological problem rather than an archaeological one, several approaches have already been attempted, mostly using genetic algorithm approaches. Since multiple sequence alignment is a well-known NP-complete problem form the field of biology, a solution that was searching the problem space in an intelligent manner seemed the way to go. A machine learning technique involving reinforcement learning was preferred to the more popular GA approaches in the literature as it seemed that the problem candidate solutions could be formalized as permutations of more simple pair alignments. In this way the first topic of research was established in the form of developing a reinforcement learning-based algorithm for the multiple sequence alignment problem (commonly abbreviated as MSA).

However, molecular biology is not the only aspect of great importance in bioarchaeology. The analysis of the osteological material present on site is also very important because it leads to the discovery of the basic biological identification attributes. These attributes have been inherited in the field of bioarchaeology from forensic anthropology where they are they primary identifiers in the case of missing persons. The indicators are : sex, age, race, stature, weight. In many of the cases encountered in practice they accurate determination of these basic biological indicators is a very demanding job. The field of bioarchaeology aggravates the problem as it appears in forensic anthropology because the studied skeletal remains are often very old,eroded or incomplete and there are no clues with respect to the identity of the individual under study. The second research direction followed in the doctoral research for the present paper was to develop custom machine learning mechanisms for sex or age determination or for stature estimation. Several machine learning methods were employed in order to achieve such tasks, as previous attempts used either classic mathematical regression methods or standard classification techniques.

The purpose of this direction of research was, on the one hand to try and solve the aforementioned classification or regression tasks as accurately as possible since properly establishing sex, age, race or stature based only on osteological information (sometimes enriched with additional forms of information extracted in situ -molecular, fungal, historical, etc.) is a primordial task in bioarchaeology on which several other more complex forms of analysis are based. Most of the methods employed in the literature still use rather classical approaches based on mathematical findings first develop half a century ago or on arbitrarily established protocols of classification that may be error prone. It seems that the most honest approaches towards these kinds of tasks are the ones based directly on the osteological information, on actual measurements rather than discrete scores. The method utilised to obtain proper solutions should be able to handle a significant level of noisiness in data and to treat the uncertain cases in such a way that presumptions should not be imposed categorically in the decisional process as such an approach may lead to a precocious false classification.

On the other hand, the idea was to employ a wide range of machine learning techniques that felt appropriate for tackling the given problems. The proposed versions of machine learning techniques needed to be adjusted to fit the problem requirements and therefore they are very particular implementations, non-standard variants of the traditional machine learning algorithms. Therefore, unifying the two goals presented above, the achieved results were expected to be novel original forms of machine learning algorithms customized for fundamental bioarchaeological tasks but that could be further generalized to fit other purposes as well. They would materialize in actual tools for bioarchaeologists, tools that will enable either a better visualization of data or a better form of abstraction that will improve subsequent analysis.

The thesis presents the main aspects of the research conducted during the past three years and therefore, in order to highlight the progress in each of the three main directions of study briefly discussed so far, it is was structured in the following manner starting with the current introduction section aimed at providing context and motivation to the entire research process behind the doctoral thesis thus presented.

Following the introduction, the first chapter, entitled **Background**, presents the necessary concepts and methods employed in the process of solving the problems that arise in each of the three major directions the present doctoral thesis attempts to tackle. Moreover, for each of the three major areas of interest a clear literature review is conducted in order to establish the state of the art approaches and techniques in each area.

The next three chapters propose each of the three novel approaches for solving the bioarchaeological problems introduced in the previous sections. The first problem, Multiple Sequence Alignment (MSA) is presented in the chapter entitled **A novel approach to Multiple Sequence Alignment**. The motivation for choosing this subject as the first research direction in the current doctoral thesis was presented in the introduction, but the actual solution for this complex problem using machine learning approaches is detailed in chapter two. A reinforcement learning based solution is proposed, based on an initial refactoring of the states inside the problem space in a permutation format that enables the employment of a q-learning approach towards finding the best sequence of pairwise dynamic programming alignments that would lead to the best overall alignment of nucleotide sequences.

The next chapter regards the sex determination problem and is entitled in a generic fashion **Novel machine learning approaches in bioarchaeology**. This is a rather substantial chapter whose goal is to convince that machine learning are truly efficient when employed in solving fundamental bioarchaeological tasks such as the classification task of sex determination from osteological measurements. The proposed methods are based on classical machine learning constructs tailored or adapted to fit the needs required by the research domain. First, a very visually compelling method of Self-Organizing Maps is developed and employed for the sex determination problem. The SOMs can be employed to a certain degree in the classification but, due to their pregnantly visual aspect are also useful to the data analysis process as a means of visualizing huge amounts of data in such a way that highlights certain similarities or dissimilarities. As it is a classification problem, the sex discrimination task can be regarded also as a clustering problem and therefore clustering techniques are employed as an alternative. Finally, the sex determination problem is tackled using approaches based on decision trees. Initially traditional decision trees were employed but as a subsequent analysis of the statistical shape of the osteological material with respect to sex revealed certain useful patterns a novel fuzzy decision tree variant based on these findings was conceived leading to competitive accuracy rates.

The third main direction of research is presented in the subsequent chapter - **A novel approach to adaptive relational association rule mining**. It is presented as a separate direction of study as it promises a new perspective in the sex determination process apart from the ones presented in the previous chapter. Relations between osteological measurements are employed to enrich the classification process with additional information.

The conclusions together with ideas for continuing the research in this exciting interdisciplinary field are presented in the final chapter, entitled **Conclusions**. An interesting method for combining the classification powers of the later two approaches is discussed as a method that may lead to good accuracy rates in the future.

We would like to highlight that the research process for the current doctoral thesis has lead to several original contributions to the domain of bioarchaeology which will be briefly emphasized in the following paragraphs.

The main contributions of this thesis, separated on the three main topics of interest, are as follows.

The first contribution consists of a novel reinforcement learning based approach for Multiple Sequence Alignment (Chapter 2) [BMC16] and a thorough experimental evaluation and comparison to other approaches from the literature [BMC16, MBD14]. We also provide a detailed study on the effect of parameter variations and on using the Needleman-Wunsch with affine-gap penalty algorithm [MBD14]. The performance of the reinforcement learning based model for multiple sequence alignment was also tested on protein sequences in [MCB15].

The second original contribution consists of machine learning based models for sex determination based on archaeological osteological material and are contained in Chapter 3. The proposed machine learning based approaches for the archaeological *sex determination* problem are: *self organizing maps*, *clustering* and *decision trees* [CMMM16], [MLs15], [MCs15]. A thorough experimental evaluation and comparison to other approaches from the literature is also conducted [CMMM16, MLs15, MCs15].

The third original contribution regards the employment of the Adaptive Relational Association Rule Mining algorithm and are contained in Chapter 4. The novel concept of Adaptive Relational Association Rule Mining is an original contribution that was part of the research process related to the current doctoral study and the impact it yields on the field of bioarchaeology is presented in the original paper [CCSiM15]. Several computational experiments that test the performance of the newly introduced concept of Adaptive Relational Association Rule Mining in the field of bioarcheaology are also provided in Chapter 4.

In conclusion, it must be stated that, after a research conducted over the span of merely three years, the results obtained along the way represent an important step in the right direction of adapting machine learning algorithms to the needs of the bioarchaeological research domain. The research work performed for the achievement of the present doctoral thesis has enabled the author to enrich their knowledge in the fields of both biology and archaeology while developing several original machine learning algorithms tailored to solve problems of utmost importance in the bioarchaeological field such as the multiple sequence alignment problem or the sex and stature determination based on the length of the osteological material collected on the archaeological site.

The informational needs of modern archaeology have been increasing drastically over the past few years and this phenomenon was revealed clearly throughout the research process for the present doctoral thesis. Three years ago, the main disadvantage when trying to solve a bioarchaeological problem from a machine learning perspective was the significant lack of digital information regarding either osteological measurements or DNA extraction results. However, during the past few years, several consistent data sets containing osteological data such as the Terry collection provided by the Smithsonian Museum and more and more DNA sample data sets concerning individuals from archaeological sites were available online, enabling researcher such as ourselves with the substantial information needed by machine learning algorithm to better adapt to solving bioarchaeological problems.

As it was stated in the introduction, the problems that occur in the field of bioarchaeology, even the ones fundamental to the bioarchaeological studies such as the ones tackled throughout the current doctoral research process, usually present a high degree of complexity thus hinting machine learning approaches. Given the sustained increase of data regarding osteological and anthropological measurements collected from individuals of several archaeological sites that was visible in the past few years and since at least the molecular biology studies alone require cross-referencing DNA material from a vast range of archaeological sites at a global scale in order to perform adequately, the digitalization of bioarchaeological data seems inevitable in the foreseeable future. It is more likely that more and more data will be uploaded only leaving archaeologists with a huge amount of information to be analysed which would at a certain point become difficult to be handled by human intelligence. Artificial intelligence mechanisms such as the ones developed as a result of the studies conducted for the present doctoral thesis, may prove to be the tools much needed by researchers in the field of bioarchaeology.

We dare to say that future researchers in the field of bioarchaeology will need to possess at least basic knowledge of data mining tools and how they can be used to solve bioarchaeological matters. Also, these tools must be provided by research teams with a predominant computer science background that will need to posses above average knowledge of the problems that arise in the field of bioarchaeology even if they are related to molecular biology or to forensics, anthropology or osteology in order to adapt the machine learning mechanisms to the particularities of the problems proposed by the field of bioarchaeological research.

The status quo concerning the interdisciplinary relations between the aforementioned two fields of research was not clear from the beginning of the research process but it began to reveal itself as the research progressed over the past few years. We consider the pace of development in this respect to be significantly high, leading us to believe that in the following decade both fields will be drastically shaped into forms that will require their co-dependence. However, co-dependence has a slightly pejorative connotation, as in the case of machine learning and bioarchaeology it seems to be more the case of a symbiosis rather than co-dependence due to a reason stated in the introduction of the present doctoral thesis: from the bioarchaeological perspective the huge amount of data gathered on an archaeological site is indeed a challenge, but it is utterly required for machine learning mechanisms to be trained on in order to perform optimally; the use of machine learning tools by bioarchaeological researchers during the research process will enable them to have a more global and analytical view upon the archaeological site and the underlying populations under study but it will in fact produce different forms of significant information that can enrich the machine learning process itself over time. It is clear that the relationship between these two fields of study that are at a first glance highly unrelated is strong and complex by it is not one of co-dependance and rather one of mutual complementarity and support. This,we consider, is an aspect that strongly suggests that the future shapes of the research process in both fields will be strongly influenced by one another.

In this context, we consider the present doctoral research to be a small hearthstone for future interdisciplinary research in the field of bioarchaeology. Its primal purpose was to address several important issues in the field of bioarchaeology from a machine learning perspective, proving that the two research fields are not as unrelated as it may seem at first sight. The work conducted along the way has revealed both the mutual advantages but also the challenges researchers from both fields are subjected to in the process. The mutual advantages will become apparent only when proper tools are provided by computer science-related research teams and used at a significant scale in the field research of bioarchaeology. The current status of facts highlights the challenges that need to be overcome by researchers from both fields in order to help create those proper tools. One major challenge was the interdisciplinary dialog between researchers from the two different fields as each domain has its custom rules, language, its idiosyncratic particularities that seem at least at first rather incommode to the other field. We tend to think that to some extent a certain barrier was breached in the communication aspect during the course of the present doctoral study as good communication relations were established with a team of bioarchaeologists from ICIBNS and as well as with Professor David Hunt from the Smithsonian Museum not only in the aspect of simple data acquisition but also in the instruction of the members of a research team focused mainly on developing machine learning algorithms into the fields of molecular biology, forensic anthropology or archaeology.

Another important aspect concerning the interdisciplinary dialog was the relation to the reviewers from publications in either the computer science or archaeological domain. In the beginning of the research process, we have targeted mainly publications concerned with applied machine learning techniques perhaps in view of a certain communication barrier we have anticipated from the complementary field of bioarchaeology. The conversation with several reviewers has lead us to the conclusion that the proposed approaches were more interdisciplinary than foreseen and therefore in later submission we directed our attention towards publications from the bioarchaeological realm as well. In many cases it was an inspired choice as it seems, as in fact we dared to presume that the bioarchaeological researchers feel the need for computerized tools that would aid them in the research process, especially in the data analysis process which is fundamental to the entire bioarchaeological research.

The fact that several of the papers proposed during the research for the present doctoral thesis were published in reviews such as the Journal of Archaeological Science is a clear indicator that the bioarchaeological science community is ready to accept machine learning-based approaches to problems in their field since in this respect there is very much to be done. The osteological research is in some area still based on numerical solutions derived from formulae dating from over half a century ago. There is obviously room for experimentation with new approaches in the area and this was one substantial trend of research in the current doctoral studies. It is extremely interesting what happened after new machine-learning based methods were proposed for solving classification or regression tasks from the bioarchaeological field : not only new formulae that can be used in practice by bioarchaeologists were obtained, but some of the machine learning algorithms developed in the process exhibited clearly original forms.

For example, the fuzzy decision tree approach used to improve the accuracy of the sex determination protocols employed by bioarchaeologists drew advantage from the almost perfectly normal distributions of the measurements of a certain bone for both sexes that both differentiated between the clear cases of male and female individuals but allowed a more flexible differentiation method in the more complex cases when indecisive information was provided. A new machine learning-based method, tailored perfectly on the bioarchaeological requirements of the problem was therefore introduced in an original form. This decision mechanism is considered one of the bases for future improvement with respect to the findings resulted from the current doctoral research, as we plan to extend it into a hybrid from that takes advantage of another method that was employed for the first time to solve problems of a bioarchaeological nature in the course of the current research : the Fuzzy Relational Association Rules. A future direction of study that builds on the knowledge accumulated in the process is to base the creation of the Fuzzy Decision Tree not only on the information gain provided by a single attribute used in classification at a time but to first mine significant Fuzzy Relational Association Rules from the data set and then used the most representative such rules that transpire from the data in building a far more improved decision tree that will not only use individual attributes in the decision process but the relations between the attributes as well.

It is therefore encouraging to see that the field of bioarchaeology shapes machine learning mechanisms into new forms that are not particular to rather than mostly inspired by bioarchaeology and that can eventually be generalized into forms that may be applied to other domains as well since the hybrid fuzzy decision tree variant present in the previous paragraph is not in its essence related to the bioarchaeological realm of research, it being very easily applicable to problems from various other domains provided the fuzzy approach that was adopted is applicable to the provided data. Also, the formulae for stature estimation obtained using machine learning techniques achieve improved accuracy rates when tested in practice reshaping the classical regression formulae computed in the past by researchers from the bioarcaheology field and most of the proposed approaches behave rather well in practice with average accuracy rates exceeding 85%. In this manner, it is the computer science domain that helps shape the bioarchaeological field as the tools provided by computational approaches could be utilized on a broader and broader scale in practice, on the field, by bioarchaeological researchers.

Bibliography

[BMC16] Maria-Iuliana Bocicor, Ioan-Gabriel Mircea, and Gabriela Czibula. A novel reinforcement learning based approach to multiple sequence alignment. In *Proceedings of the 7th International Workshop on Soft Computing Applications*, 2016.

[CCSiM15] G. Czibula, I.G. Czibula, A. Sîrbu, and I.G. Mircea. A novel approach to adaptive relational association rule mining. *Applied Soft Computing*, 36:519–533, 2015.

[CMMM16] G. Czibula, I.G. Mircea, D.L. Miholca, and R. Moldovan. Machine learning approaches for gender detection of human skeletal remains. page under review, Timisoara, Romania, 2016. IEEE Computer Science.

[MBD14] Ioan-Gabriel Mircea, Maria-Iuliana Bocicor, and Andra Dîncu. On reinforcement learning based multiple sequence alignment. *Studia Universitatis Babes-Bolyai Series Informatica*, LIX(2):50–65, 2014.

[MCB15] I.G. Mircea, G. Czibula, and M.I. Bocicor. A q-learning approach for aligning protein sequences. In *Proceedings of the 11th International conference on Intelligent Computer Communication and Processing, Cluj-Napoca, Romania*, page to appear, 2015.

[MCs15] Ioan-Gabriel Mircea, Gabriela Czibula, and Mara-Renata Petruşel. Sex identification in archaeological remains using decision trees. *Studia Universitatis Babes-Bolyai Series Informatica*, LX(2):91–103, 2015.

[MLs15] Ioan-Gabriel Mircea, Sergiu Limboi, and Mara-Renata Petruşel. A new unsupervised learning based approach for gender detection of human archaeological remains. *Studia Universitatis Babes-Bolyai Series Informatica*, LX(2):5–20, 2015.

[SKW$^+$10] Camilla F. Speller, Brian M. Kemp, Scott D. Wyatt, Cara Monroe, William D. Lipe, Ursula M. Arndt, and Dongya Y. Yang. Ancient mitochondrial DNA analysis reveals complexity of indigenous North American turkey domestication. *Proceedings of the National Academy of Sciences*, 107(7):2807–2812, February 2010.