Intelligent Data Analysis:
A Case Study of the Diagnostic Sperm Processing

Mariza Ferro
Unioeste – State University of West Paraná, LABI - Bioinformatics Laboratory,
Foz do Iguaçu Campus, Computer Science Course
PO Box 961 CEP 85870-900 , Foz do Iguaçu, PR - Brazil
Tel.: 45 5752727   Fax: 45 5752733
mariza98@dcc.unioeste-foz.br
Huei Diana Lee
Unioeste – State University of West Paraná, LABI - Bioinformatics Laboratory,
Foz do Iguaçu Campus, Computer Science Course
PO Box 961 CEP 85870-900 , Foz do Iguaçu, PR - Brazil
Tel.: 45 5752727   Fax: 45 5752733
huei@dcc.unioeste-foz.br
Sandro C. Esteves
Androfert - Masculine Infertility Reference Center, Campinas, SP - Brazil
s.esteves@androfert.com.br

Topic Area:
Artificial Intelligence

Key Words:
Intelligent Data Analysis, Machine Learning, Sperm Processing, Knowledge Extraction, Constructive Induction, Medical Databases.

Abstract

This work presents a case study of a human reproduction data set, more specifically a sperm processing data set, in which is applied Intelligent Data Analysis using Machine Learning methods, an Artificial Intelligence sub-area. The main objective of this work is to extract knowledge, possibly an unrealized knowledge, that can be useful and interesting to the specialist. The extracted knowledge can support the specialist during the decision making process. To achieve our objective we apply a knowledge-constructive induction method proposed by a former work using the inductive algorithm *See5*.

1. Introduction

In the real world, a huge data volume has being stored in the data bases and a great gap stays between these data generation and understanding. Extracting manually useful and interesting knowledge from the data is becoming more difficult every day. Medical data, gotten from for example hospitals, clinics and laboratories, are not different. Covering this gap between generation, understanding and effective use of the data becomes crucial to medicine. One of the ways of performing Intelligent Data Analysis - IDA - in medicine is to use Machine Learning - ML, which has being used in a variety of medical domain [1].

Machine Learning searches for computational methods to assist knowledge extraction from data bases and also construction of predictors that help out the decision making process.

In this work, we use Symbolic ML methods to extract knowledge from a real world data set on sperm processing. The objective is to predict which assisted reproduction method to use without performing a sperm processing given only the data provided by a conventional sperm analysis thus reducing the total cost of the process.

This work is organized as follows: Section 2 gives some background on Machine Learning, Section 3 presents Constructive Induction, Section 4 shows the experimental setup while Section 5 presents the experiments. Section 6 shows the results and Section 7 presents the discussion. Finally, some conclusions are given in Section 8.

2. Machine Learning

Machine Learning can be defined as an Artificial Intelligence - IA - subarea that searches for computational methods related to the automatic acquisition of new knowledge, new abilities and new forms of organizing the existent knowledge.

Before using ML methods to extract knowledge or to construct predictors, it is necessary to define the paradigm and the strategy. The paradigm establishes the way that knowledge is acquired. Some examples of paradigms are: connectionism, genetic, statistic, instance-based and symbolic [2]. In the symbolic paradigm, systems learn constructing symbolic representations of a concept through the analysis of positive and negative examples of the concept. The most commonly used symbolic representations are the decision trees and rules. The induction of rules started with the simple translation of decision trees and evolving to methods that employ, for example, generalization [3,4].

Learning strategies are classified according to the complexity degree, ié, the concept learning difficulty. In any learning process, the learner uses the accumulated knowledge to obtain new knowledge. This new knowledge is then remembered to further use. Learning a new concept can occur in different ways [2]: habit, instruction, deduction, analogy and induction. The last learning form requires the greatest inference complexity. It is characterized by starting reasoning from the specific to general. It is also the logic inference form that allows to obtain general conclusions from particular examples.

There are two forms of inductive learning. The first one is called Unsupervised Learning or Learning by Observation and Discovery. In this form the learner analysis given objects and try to determine if there is any subset that can be grouped in certain useful classes (the objects have not been preclassified by a teacher). In the second learning form, named Supervised Learning, the learner seeks to develop a concept description from examples that have been preclassified by the teacher.

In this work we focus on Symbolic Inductive Supervised Machine Learning.

3. Constructive Induction

Features, individually inadequate to the description of a concept, can sometimes be conveniently combined into new features that can show to be highly representative to the concept description. The process of constructing new features is known as Feature Construction, Constructive Learning or Constructive Induction - CI.

Constructive Induction methods can be grouped according to the information that is used to search for the best representation space [5]:

- data-driven constructive induction, based on analysis of the training data;
- hypothesis-driven constructive induction, based on analysis of inductive hypothesis. In this approach, useful concepts in the rules can be extracted and used to define new features;
- knowledge-driven constructive induction, based on domain knowledge provided by an expert;
- multi-strategy constructive induction, based on two or more of the other methods.

The Constructive Induction process can be guided and controlled by the user/expert or can be automatically conducted by the learning system. In this work, we focus on Constructive Induction guided by user/expert using thus the knowledge-driven approach.

4. Experimental Setup

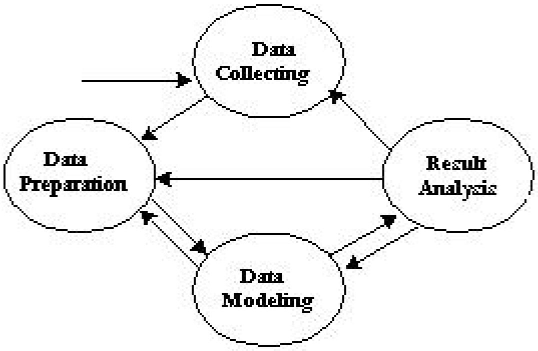


Figure 1 – Methodology Stages

The applied methodology was compounded by basically four distinct stages: Data Collecting, Data Preparation, Data Modeling and Result Analysis.

The studied data set contains 278 cases described by 28 features. Each case information was obtained by the sperm analysis and the sperm processing results from the same patient. This unprocessed data was collected directly from Androfert (http://www.androfert.com.br), a human reproduction reference center. After the collecting stage, data was prepared and cleaned, ié the input file for the next stage was created. In the next stage we used the induction algorithm See5 to generate knowledge in the form of decision rules. In this stage, divided into 3 steps (described in details in Section 5), we applied knowledge-driven constructive induction [6,7]. Finally, during the last stage the obtained results were analyzed with the aid of the domain specialist. The Data Preparation, Data Modeling and Data Analysis stages can be resumed during the complete process.

4.1. Algorithm

Extracting knowledge, in this specific case in the medical domain, implies that the results should be easily interpretable by the specialists from other domains besides computing. Yet, these results should also be transparent which means that it should be possible to prove them easily.

The algorithm *See5* [8] presents these characteristics. Results can be given in the form of rules *If* <> *then* or in the form of decision trees. This makes the knowledge quite intelligible and easy to be proved.

This tool is an upgraded version of the C4.5 algorithm [9] for the induction of decision trees. C4.5 uses a split decision criteria based on the entropy, ié, a greatness that measures disorder in physic objects as in information. The split criteria seeks to minimize the entropy. In C4.5 each leaf node has the following form:

<Ci> (N) or <Ci> (N/E)

where Ci is the name of the class, N is the fraction sum of the instances that reach that node, E is the number of instances that belong to other classes then Ci, but were classified as belonging to Ci [10].

4.2. Data Set

The case study involved a real world data set containing 278 cases described by 28 features that contain all the results obtained by a sperm analysis and a sperm processing from the same patient such as total concentration of spermatozoon per collected milliliter and percentage of alive spermatozoon.

The sperm processing is a quite important stage to the assisted reproduction [11,12]. It allows to quantify the semen quality (diagnostic processing) and still recover the possible larger quantity of spermatozoon (therapeutic processing). This permits the decision of which assisted reproduction technique - ART - to use in the treatment. There are mainly three assisted reproduction methods: IUI - Intrauterine Insemination, IVF - In Vitro Fertilization and ICSI - Intracitoplasmic Sperm Injection. However, the sperm processing rises the costs up to 80% compared to the cost of a sperm analysis [7].

Cases were classified into three classes of clinic interest to the assisted reproduction according to the number of recovered spermatozoon (*n*) after the diagnostic sperm processing. These three classes are:

- class 1: $n < 1.0x10^6$ (ICSI);
- class 2: $1 <= n < 5x10^6$ (FIV) and
- class 3: $n >= 5x10^6$ (IUI).

Table 1 summarizes the sperm processing data set in four columns:

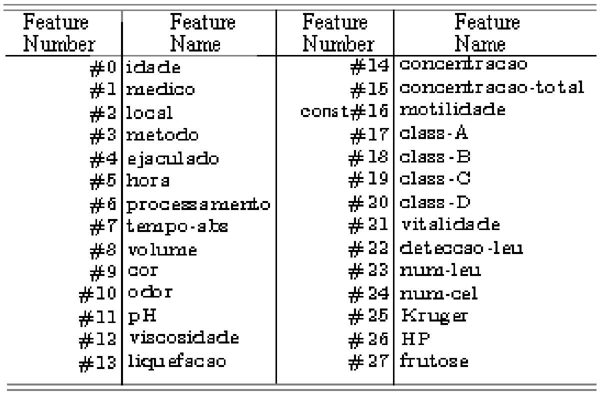
- # Instances: number of instances;
- # Features (cont, nom): total number of features and the number of continuous and nominal features;
- Class (%): class and each class distribution and
- Error CM: majority class error.

Table 1 - Data Set Summary Description

# Instances	# Features (cont, nom)	Class (%)	Error CM
278	28 (17,11)	1 (40.65%) 2 (22.30%) 3 (37.05%)	59.4% on value 3

Table 2 shows the sperm processing data set features initially considered for the analysis. Note that this set of features contains only one constructed feature (#16 motilidade) that is given by class-A+class-B+class-C.

Table 2 - Data Set Features



Feature Number	Feature Name	Feature Number	Feature Name
#0	idade	#14	concentracao
#1	medico	#15	concentracao-total
#2	local	const#16	motilidade
#3	metodo	#17	class-A
#4	ejaculado	#18	class-B
#5	hora	#19	class-C
#6	processamento	#20	class-D
#7	tempo-abs	#21	vitalidade
#8	volume	#22	deteccao-leu
#9	cor	#23	num-leu
#10	odor	#24	num-cel
#11	pH	#25	Kruger
#12	viscosidade	#26	HP
#13	liquefacao	#27	frutose

5. Experiments

The experiments were divided into three steps, based on a previous work [7]. For all the nine subset of features, all the 278 cases were considered, generating thus nine variations of the original data set. Rules were generated using *See5* default setup values.

The first step considers three sets of features: 1) All - all the features (primitives and one constructed), 2) Primitives - only the primitive features (feature #16 is not considered) and 3) Constructed - all features except the primitive features that are used to construct feature #16.

To the second step we considered again three different sets of features also using the idea of primitive and constructed features: 1) No f7 - all features are considered except a subset of seven features (f7 = {medico, local, metodo, ejaculado, deteccao-leu, num-leu, frutose}) that are considered not important by the specialist, 2) all features except f7 and class-A, class-B and class-C and 3) all features except f7 and motilidade. The objective of the second step is to verify if the accuracy would be better without the set f7 of features.

Finally, in the third step, again using the same idea as in the first step, experiments were performed considering three sets of features: 1) all features plus mot-prog (class-A + class-B), a new constructed feature not present at the original set of features, 2) all features plus mot-prog except f7 and 3) all features plus mot-prog except f7, class-A, class-B and motilidade.

6. Results

Table 3 presents the results summary in eight columns:

- Step: indicates the experiment step;
- Feature Set: shows the subset of considered features;
- # F: shows the number of features in the subset;
- Error: indicates the apparent error, ié, training and testing using the same data set;
- Error (10CV): shows the error using 10 k-fold cross-validation[1];
- Std Dv: indicates the standard deviation;
- MC Error: indicates the majority error and
- Imp.%: shows the improvement rate which is given by 100-(Error x 100/MC Error).

Table 3 - Results Summary

Step	Feature Set	# F	Error	Error (10 CV)	Std Dv	MC Error	Imp. %
Step 1							
	All	28	17.6	42.1	2.3	59.4	29.12
	Primitives (no motilidade)	27	17.6	39.6	2.6	59.4	**33.33***
	Constructed (no class-A, class-B and class-C)	25	20.5	41	3.4	59.4	30.98
Step 2							
	No f7	21	10.1	41.7	2.6	59.4	**29.80***
	No f7 and class-A, class-B and class-C	18	10.8	46.5	2.6	59.4	21.72
	No f7 And motilidade	20	10.1	44.2	3.6	59.4	25.59
Step 3							
	With mot_prog	29	17.6	43.6	2.2	59.4	26.60
	With mot_prog And no f7	22	8.3	42.5	2.5	59.4	**28.45***
	With mot_prog and no f7, class-A, class-B and motilidade	19	18.7	45.4	2.2	59.4	23.57

The best result, considering the improvement percentage, in each step is shown in boldface and marked with a *.

7. Discussion

Each best result of the three steps was considered for discussion with the specialist. Thus, for step 1 the data set (1) **Primitives (no motilidade)** was selected, for step 2 the data set (2) **No f7** and for step 3 the data set (3) **With mot_prog and no f7** was selected.

Table 4 summarizes the confusion matrix information for these three set of rules.

Looking to Table 4 it can be seen that the most confusing class for all them is the class 2. This is an explainable behavior because class 2 represents a range in which it is not well defined which assisted reproduction technique - ART - to use, since the cost to perform an IVF is near the cost to perform an ICSI.

Table 4 - Confusion Matrix Summary

Set of Rules	Error	Conf. Class (%)	Class (%)	Class (%)
(1)	17.6%	2 40.3%	1 6.2%	3 4.9%
(2)	10.1%	2 27.4%	1 3.5%	3 6.8%
(3)	8.3%	2 8.3%	1 4.4%	3 5.8%

On the other hand, all the three sets of rules show that the decision between classes 1 and 3 is well defined. The specialist considered this fact interesting because the ARTs use are evolving to the use of only IUI and ICSI.

For the set of rules (1) **Primitives (no motilidade)**, eight (1, 2, 6, 7, 11, 12, 13 and 14) of the 14 rules were considered good, one (9) was considered very good, but this rule covered only two instances. Rule 2, considered as good rule, covering 80 instances, is shown below:

If concentracao_total <= 95.2 and class_A <= 17
then class = 1 [0.854]

For the set of rules (2) **No f7**, nine (2, 3, 4, 7, 15, 20, 21, 22 and 24) of the 29 rules were considered good by the specialist. Two (5 and 6) rules were considered not regular. It may be interesting to investigate more about them in the future. Three (10, 11 and 12) rules were found very interesting by the specialist because they considered not conventional features such as tempo_abs (abstinence time before the sperm was collected), Kruger

[1] 10 k-fold cross-validation is performed dividing data into ten subsets. The induction algorithm is trained and tested ten times; each time it is tested using a subset and trained using the complete data set except the testing subset. The accuracy is the mean value of the accuracies in the each one of the ten subsets.

(spermatozoa morphology) and num_leu (number of found leukocytes) although they are from class 2 . For example, consider rule 10, covering 6 instances:

If tempo_abs <= 2 and viscosidade = normal
and class-A > 6 and vitalidade > 60
and HP <= 76
then class = 2 [0.857]

For the set of rules (3) **With mot_prog and no f7**, 10 (2, 8, 9, 10, 11, 12, 13, 15, 18 and 19) of the 31 rules were considered good and two (14 and 21) were considered very good, as shown bellow (rule 21, covering 24 instances):

If processamento > 25 and liquefacao = completa
and concentracao_total > 95.2
and class-A >6 and num_leu <= 5.2
then class = 3 [0.923]

8. Conclusions and Future Work

In this work, we describe a case study using a real world data set. Performing case studies with the aid of a specialist, using real world data sets, allows the use of Machine Learning methods through two points of view: construction of predictors and knowledge extraction.

In the medical area, statistical tests are conventionally employed to analyze data. The case study presented in this work is a continuation of a previous work and permitted the specialist to perform data analysis in a different way applying Intelligent Data Analysis using Machine Learning methods.

The specialist found interesting rules, such as the rules that use not conventional features to decide which class a new case should be assigned to. These features will be taken in account for future work. Also as future work, we intend to apply quality measures, such as rule interestingness to the rules and submit these results to the specialist evaluation.

Also, as the data set grows up, this work will be continued and a IDA module will be integrated to the Androfert's data management system to help the specialist in the decision making process.

Acknowledgments: We are grateful to Prof. Dr. Maria Carolina Monard (LABIC – ICMC/USP-SC) for the valuable aid during the definition and performing of the experiments. We also would like to thank Unioeste and Paraná Tecnologia for financial and material support. Thanks also to ITAI - Technology and Automation in Informatics Institute. This research was partially supported by IEL - Euvaldo Lodi Institute.

References

[1] N. Lavrac, "Data Mining in Medicine: Selected Techniques and Applications", J. Stefan Institute, Ljubljana, Slovenia, 1998, in <http://citeseer.nj.nec.com/cs> april/2001.

[2] M. C. Monard, G. E. A. P. A Batista, S. Kawamoto, and J. B.Pugliesi, "Uma Introdução ao Aprendizado Simbólico de Máquina por Exemplos", Notas do ICMC, *ICMC - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo*, São Carlos, SP, Outubro 1997.

[3] J. R. Quinlan, "Generating Production Rules From Decision Trees", In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, Italy, 1987.

[4] J. R. Quinlan, "Simplifying Decision Trees", *International Journal of Man-Machine Studies,* 1987.

[5] E. Bloedorn and R. S. Michalski, "Data-Driven Constructive Induction", *IEEE Intelligent Systems*. Vol. 13 no. 2, pp. 30-37, 1998.

[6] H. D. Lee, M. C. Monard, J. A. Baranauskas, "A Practical Approach for Knowledge-Driven Constructive Induction", *in Proceedings of ASAI - Argentine Symposium on Artificial Intelligence - 29th JAIIO*, Tandil, Argentine, 2000, pp. 71-85.

[7] H. D. Lee, M. C. Monard, S.C. Esteves, "Indução Construtiva Guiada pelo Conhecimento: Um Estudo de Caso do Processamento de Sêmen Diagnóstico", *in Proceedings of SBIA – Simpósio Brasileiro de Inteligência Artificial*, Atibaia, SP, Brasil, 2000, pp. 157-166.

[8] Rulequest-Research. "Data mining tools See5 and C5.0", <http://www.rulequest.com/see5-info.html>.

[9] J. R. Quinlan, "C4.5: *Programs for Machine Learning"*, Morgan Kaufmann. San Francisco, CA, 1993.

[10] J. A. Baranauskas,. M.C. Monard, "An Unified Overview of Six Supervised Symbolic Machine Learning Inducers", Technical Report 103, *ICMC - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo*, São Carlos, SP, <ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/RT_103.zip> Fev 2000.

[11] S. C. Esteves, F. C. Bento, "Sperm processing for assisted reproductive technology (ART): effects of pentoxifyline on recovery of motile sperm in asthenozoospermic men", FertilSteril, 70: 5196, 1998.

[12] S. C. Esteves, R. K. Sharma, A. J. Jr. Thomas, A. Agarwal, "Improvement in motion characteristics and acrosome status in cryopreserved human spermatozoa by swim-up processing before freezing", HumReprod, 15: 2173-2179, 2000.