

A Review and Analysis on Artificial Intelligence Methods Performance Comparison through Seminal Quality Prediction

Group 17: Chen Lu (A0105483J), Ding Qin (A0105608L)

I. Introduction

According to numerous previous studies, there is an obvious decrease in fertility rates during the past two decades. One possible reason behind this decrease is the decline in semen quality. In the subject literature, several factors have been considered to explain this decline, for example, the increase in the incidence of male reproductive diseases, and the effect of environmental or occupational factors to a certain lifestyle.

Meanwhile, artificial intelligence techniques has become an emerging methodology as decision support systems in medicine nowadays. Gil et al. (2012) paper evaluated and compared the performance of three different artificial intelligence techniques - Decision Trees (DT), Artificial Neural Network (ANN), and Support Vector Machines (SVM), by employing them to predict the seminal quality from the data of the environmental factors and lifestyle.

This project report will describe Gil et al. (2012) paper's contribution, discuss the issues, challenges, and limitations of this study, as well as replicate one experimental plot with R using data provided in the experiment.

II. Issues and Challenge

The main objective of Gil et al. (2012) paper is to compare three different AI classification methods when used on male fertility data sets: Decision Trees (DT), Multilayer Perceptron (MP), and Support Vector Machines (SVM). Therefore, this section will present a brief description of each AI method.

1. Decision Trees (DT)

A Decision tree refers to a classifier in the form of a tree structure, where each node is either a leaf node indicating the value of the target attribute (class) of examples, or a decision node specifying some test to be carried out on a single attribute-value.

For classification and prediction, decision trees can be used as powerful and popular tools. Compared with other AI methods, such as MLP and SVM, the advantage of decision trees is that as the trees represent rules, they can be easily understood and directly stored in database.

DT is a fairly mature technique including several different models, whose differences are in the derivations of formula to determine the attributes that lead to splitting, such as entropy and information gain. However, since most decision trees consist of numerous nodes, under certain circumstances, the great number of nodes may hinder analysis or interpretation of the information.

The classification and regression trees (CART or C&RT) method generates binary decision trees. Although in the real world, the chances of biased binary outcomes are few, however, the binary method will be easier for interpretation and analysis in research. Thus, the study uses the binary method in the DT to interpret the life habits in prediction of male fertility, and attempts to deduce a correlated prediction factor.

Algorithm C4.5 is used in Gil et al. (2012) paper, which is an extension of the basic ID3 algorithm.

The ID3 algorithm can be summarized as follows:

- i. Take all unused attributes and count their entropy concerning test samples;
- ii. Select attribute with minimum entropy (or, equivalently, maximum information gain);
- iii. Create a node with that attribute.

The reason of choosing DT as the first analysing method is that it is similar to the way physicians conducting analysis in their usual procedure, and it help reduce input data as well.

2. ANN - Multilayer Perceptron (MP)

A Multilayer Perceptron consists of multiple layers of neurons, generally three (where each layer fully connected to the next one):

- 1). an input layer that receives external inputs;
- 2). a hidden layer;
- 3). an output layer which generates the classification results.

Except for the input layer, every neuron in other layers is a computational element with a nonlinear activation function. The principle of the network is that when data are presented at the input layer, the network neurons run calculations in the consecutive layers until an output value is obtained at each of the output neurons. This output will indicate the appropriate class for the input data.

3. Support Vector Machines (SVM)

Support Vector Machine (SVM) refers to the generation of the maximal margin classifier to the non-separable case. Radial Basis Functions Kernel (RBF) function has been adopted as a suitable method for this study, as RBF kernel has less numerical difficulties.

Another reason is that, different from the linear kernel, the RBF kernel nonlinearly maps samples into a higher dimensional space; thus, it can handle the case when the relation between class labels and attributes is nonlinear.

Moreover, the number of hyper-parameters will influence the complexity of model selection. Since the polynomial kernel has more hyper-parameters than the RBF kernel, RBF kernel was selected for this study.

4. Semen Parameters

The semen analysis is according to the standardized procedure published by WHO. WHO publication established the following concepts for semen analysis:

- 1). Reference value for the different semen parameters;

2). Nomenclature that describes the deviations from reference values.

The nomenclature includes these terms:

- Normozoospermia: all values are over the lower limit.
- Asthenozoospermia: percentage of progressively motile spermatozoa below the lower reference limit.
- Oligozoospermia: total number of spermatozoa below the lower reference limit.
- Teratozoospermia: percentage of morphologically normal spermatozoa below the lower reference limit.

According to the survey result, it found that almost 38% of the men surveyed had some alteration in semen parameters. The most frequently found seminal alteration was asthenozoospermia in 47.37% of cases, while the least if teratozoospermia with a rate of 2.63%. In Gil et al. (2012) research, these laboratory results of the seminal analysis were used as a control to evaluate the accuracy in the prediction of the diagnosis of the ANN developed.

III. Simulation Results

1. Data Description

The simulation data is extracted from the original data which was collected from the 100 University of Alicante students between the age of 18 and 36 years old. The researchers have done a feature selection process to the original data, and we have directly adopted the processed dataset as our simulation data. There are in total 100 records in the dataset with 9 variables and 1 column of output. The following table has listed out the variable names, descriptions and their data range.

Variable Name	Variable Description	Value (min-max)	Normalized
Season	Season in which the analysis was performed.	1) winter, 2) spring, 3) Summer, 4) fall	(-1, -0.33, 0.33, 1)

Age	Age at the time of analysis	18-36	(0, 1)
ChiDiseases	Childish diseases (ie , chicken pox, measles, mumps, polio)	1) yes, 2) no	(0, 1)
Accident	Accident or serious trauma	1) yes, 2) no	(0, 1)
Surgery	Surgical intervention	1) yes, 2) no	(0, 1)
HighFever	High fevers in the last year	1) less than three months ago, 2) more than three months ago, 3) no	(-1, 0, 1)
FreAlcohol	Frequency of alcohol consumption	1) several times a day, 2) every day, 3) several times a week, 4) once a week, 5) hardly ever or never	(0, 1)
Smoking	Smoking habit	1) never, 2) occasional 3) daily	(-1, 0, 1)
HrSitting	Number of hours spent sitting per day	1 - 16	(0, 1)
Output	Output: Diagnosis (Sperm Concentration)	normal (N), altered (O)	normal (N), altered (O)

Table 1. Experiment Variables

2. Methodology

The aim of this simulation is to generate a Decision Tree Classification model for predicting whether the data object's sperm concentration is at a normal level.

The random forest method in caret package is used to perform the training and testing of the Decision Tree Classification model. Ten-fold cross-validation method is used to

evaluate the performance of generated models without needing any other data than the training data. The ten-fold cross-validation method is embedded in caret's train function. In the paper, classification accuracy, sensitivity, specificity, positive predictive value, negative predictive value and a confusion matrix are generated in order to evaluate the three selected artificial intelligence method. Therefore, in our simulation, these measures are also generated to be compared against the research's results.

3. Experiment R Script

```
fertility_training <- fertility_data  
library(caret)  
rf_model<-train(Output~.,data=fertility_training,method="rf",  
                trControl=trainControl(method="cv",number=10),  
                prox=TRUE,allowParallel=TRUE)  
print(rf_model)  
print(rf_model$finalModel)  
plot(varImp(rf_model), top = 5)  
testdata <- data.frame (Season = 1, Age = 0.65, ChiDiseases = 1, Accident = 0,  
                        Surgery = 0, HighFever = 1, FreAlcohol = 0.6, Smoking = 0, HrSitting = 0.50)  
prediction<- predict(rf_model, newdata = testdata)  
print(prediction)
```

Figure 1. Experiment R script

4. Results and Discussion

The generated confusion matrix for simulation is as follows:

```

Confusion matrix:
      N O class.error
N 84 4  0.04545455
O 10 2  0.83333333

```

Figure 2. Confusion matrix for simulation

The confusion matrix generated by the researchers of this study is as follows:

Actual	MLP		SVM		DT	
	Predicted		Predicted		Predicted	
	P	N	P	N	P	N
P	80	5	83	2	82	3
N	9	6	12	3	13	2

Table 2. Confusion Matrix of the Study

The other measures are listed in the following table:

	Experiment Results	Simulation Results
Classification accuracy (%)	84	86.09091
Sensitivity (%)	96.5	95.45455
Specificity (%)	13.3	16.66667
Positive predictive value (%)	86.3	89.36170
Negative predictive value (%)	40	33.33333

Table 3. Experiment and Simulation Results

From the comparison of results we can see that our results are similar to the results generated by researchers. Moreover, we have also created a new data record to test

out the prediction. The result of the prediction is “Normal”, which should have an accuracy of 86.1%.

Furthermore, we have also identified the top five most important variables in determining the output, i.e. affecting the sperm concentration. These variables are Age, Number of hours spent sitting per day, Season, Frequency of alcohol consumption as well as Accident or serious trauma. The ranking of these variables are shown in the following graph:

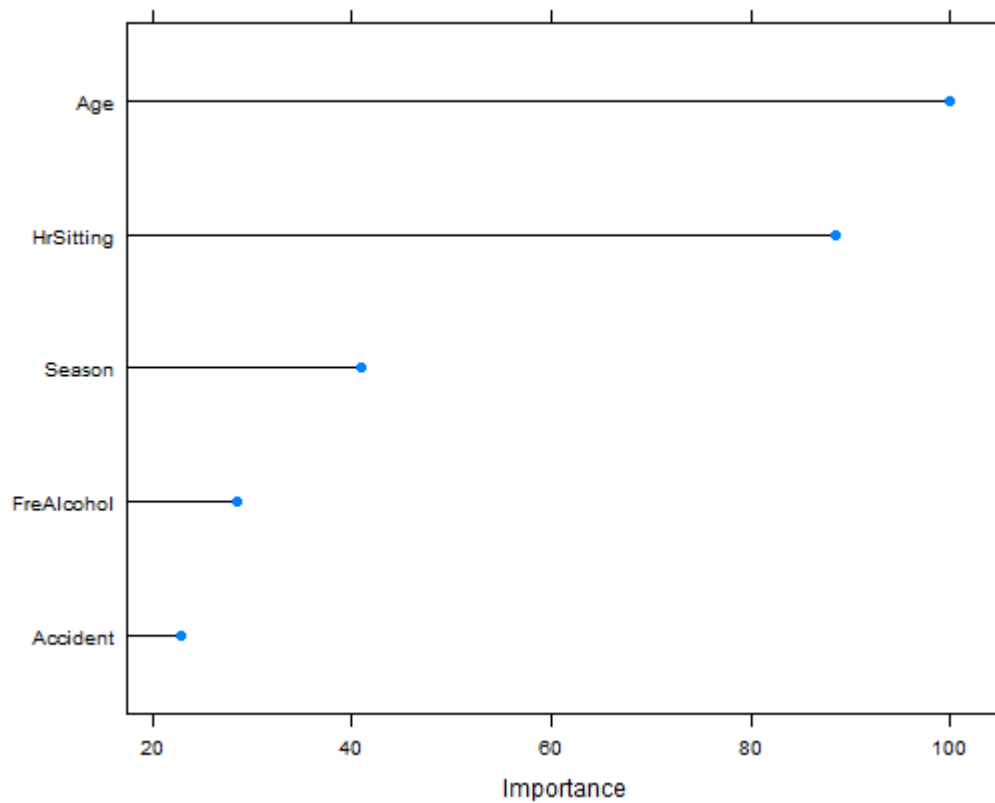


Figure 3. Variables Ranking

IV. Contributions

There are mainly two contributions that this study has made in the field of development of artificial intelligence methods as well as its application in healthcare industry.

Many previous studies have tried to identify the relationship between environmental factors and semen quality using linear statistical techniques, which have shown limitations due to the complexity and diversity of information as well as heterogeneous data types. This study has adopted non-linear statistical techniques to come up with a better solution in tackling the complexity of the problem.

For the applications of the artificial intelligence techniques in the healthcare industry, they might be able to be served as an alternative to the more expensive laboratory tests during the initial stage of fertility tests of couples. Since the study sample mainly include healthy male university students, who represents characteristics of regular semen donor candidates, the developed models may also be used to cut costs in initial semen donor selection processes. These artificial intelligence methods could also be used in next stages of fertility treatment to reduce and avoid painful and expensive examinations that the female partners need to go through.

V. Limitations

There are a few limitations observed in the experiments conducted in the study which might cast shadow on the usefulness and accuracy of the results predicted by the artificial intelligence methods developed in the study.

Firstly, the volunteers are all young healthy students from the University of Alicante in Spain between the age of 18 and 36 years old. The study did not specify to which population group the artificial intelligence methods could be applied. Moreover, the researchers used a non-probability sampling method which might not be representative enough for the target population. For example, if the targeted group is Spanish male between the age of 18 and 36 years old, as suggested by the graph below, the sample selected did not reflect proportion of the age groups according to demographic information in Spain. Moreover, in this specific situation, a sample size of 100 might also be too small to be representative of the entire target population. In future studies related to this topic, the researchers could improve sampling by clearly define the target population, refine reasonable sample size and use probability sampling method to make the study population more representative.

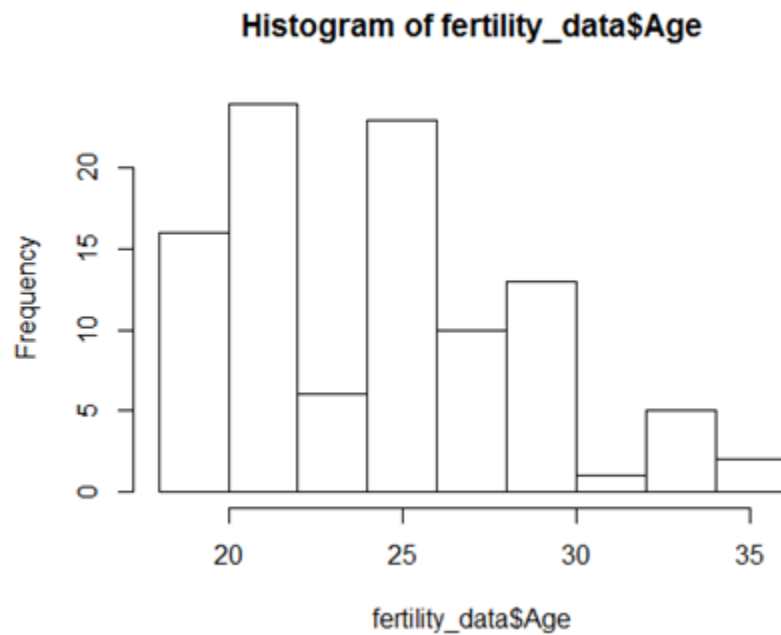


Figure 4. Histogram of fertility_data\$Age

Secondly, for all three artificial intelligence methods used, the specificity and negative predictive values present low percentage. This might be due to the fact that the study sample population has an uneven distribution of normal and altered results. For example, 88% of the samples have normal sperm concentration while only 12% appear to have altered sperm concentration. This imbalanced distribution has affected the performance of the artificial intelligence methods, and further casted doubts on whether the predictions made are accurate. With regard to this limitation, the researchers have suggested that in the future studies they could investigate how imbalanced classes would affect classification performance when developing artificial intelligence methods such as Decision Trees, Multilayer Perceptron and Support Vector Machines. Moreover, the researchers have also suggested machine learning methods as well as data mining to be used in future studies. These methods may provide a more precise correlation between seminal data and the information collected by questionnaires (Gil, Girela, De Juan, Gomez-Torres & Johnsson, 2012).

VI. Conclusion

It was the first time that DT, SVM and MLP are used for studying the effect of environmental factors and life habits on semen quality. Although there are many limitations of the experiment methods, it has nonetheless shown that SVM and MLP have the highest prediction accuracy whereas DT should also be considered for its simplification and easy interpretation. Sperm concentration and motility are shown highly influenced by environmental factors and life habits, while morphology shows more likely link to genetic factors which was not studied in this research. Further studies could be conducted to examine the relationship between genetic factors and seminal parameters using the proposed AI methodology, machine learning or data mining.

Reference

Gil, D., Girela, J. L., De Juan, J., Gomez-Torres, M. J., Johnsson, M., Humanistiska och teologiska fakulteterna. . Kognitionsvetenskap. (2012). Predicting seminal quality with artificial intelligence methods. *Expert Systems with Applications*, 39(16), 12564. doi:10.1016/j.eswa.2012.05.028