Nearest neighbor concept in the study of IVF ICSI/ET treatment effectiveness

Robert Milewski[1], Paweł Malinowski[1], Anna Justyna Milewska[1], Jan Czerniecki[2,3], Piotr Ziniewicz[1], Sławomir Wołczyński[4]

[1] Department of Statistics and Medical Informatics, Medical University of Bialystok

[2] Department of Biology and Pathology of Human Reproduction, Institute of Animal Reproduction and Food Research of Polish Academy of Sciences in Olsztyn

[3] Department of Cytobiochemistry, Institute of Biology, University of Bialystok

[4] Department of Reproduction and Gynecological Endocrinology, Medical University of Bialystok

Abstract. The effectiveness of IVF ICSI/ET infertility treatment depends on many factors. Their identification and classification of individual cases remains a difficult task. This paper presents application of feature selection algorithm MSIMBAF2 and associated kNN classifier to analyze the data set containing results of the infertility treatment process.

Introduction

Infertility is a social problem whose scale is constantly growing. This is probably associated with the upward trend of age of women giving birth to their first child and delaying motherhood to later years of life. But with the age of women the effectiveness of infertility treatment decrease [7]. Increasingly, the only chance to have children become the methods of in vitro fertilization [9–10]. However, their effectiveness also decreases with age of treated women. In women over 40 it fluctuates within 10–15%, that is even 4-fold lower than in younger women [7]. Hence the need for advanced biostatistical methods, which on the one hand would allow to forecast the results of treatment in specific patients, and on the other hand to form the basis for making certain decisions during treatment, leading to increase the probability of success, that is the birth of a healthy child.

There are many statistical methods, referred to the general term "data mining methods", which can be used to predict the effectiveness of infertility treatment. The most important factor, from the data analysis point of view will be the selection of appropriate algorithms for classification and feature selection. Feature selection is quite often used as a preliminary step in data

ISBN 978–83–7431–296–7 ISSN 0860-150X

analysis. It involves reducing the original dimensionality of the data set, by rejecting less important features. In the next step, such prepared data set is subjected to classification in order to generate decision rules. Generated rules allows to predict the target observation class for the new data as well. Since medical data are analyzed, those rules should have high efficiency and resistance to accidental errors and over-fitting. Generated decision rules should be simple to understand for the future user, probably a doctor, who will use them. Like the classifier, feature selection algorithm should take into account various types of features and missing data.

IVF ICSI/ET procedure and database for storing information

The process of IVF ICSI/ET infertility treatment take place according to the established procedure, and consists of several main stages. At each stage, information necessary to continue the treatment, and to subsequent statistical analysis is gathered. First, the personal data of patients is collected. Then, medical information on the treatment history and various tests is collected. This information include, but are not limited to man and woman medical interview, laboratory tests, and a USG image with its description. After passing this stage, if the pair is qualified for further treatment, stimulation protocol is selected and the period of treatment begins. Information about the medicaments used in the subsequent days of the treatment and ovulation stimulation is recorded. At later stage of the treatment many other parameters are recorded: the level of estradiol, endometrial thickness and the amount and size of the developing follicles in both ovaries. The next step concerns the embryology, gathering information about the aspiration of ovarian follicles, the preparation of semen and the ART procedure. In the next stage, information on the developing embryos is collected, until the transfer moment and its implementation. The last part is the final treatment which results in collecting data on the pregnancy, childbirth and also basic information about the newborn.

Infertility treatment specificity requires the collection, storage and continuous analysis of collected information, as well as the ability of rapid access to that information. The Department of Reproduction and Gynecological Endocrinology in the Medical University of Bialystok is using a specially designed application for this purpose which is based on an extensive database [4]. The application provides a statistical module that allows to compare the parameters of patients with average values, and carry out basic statistical analysis in the course of gathering information [8]. It is also shipped with implementation of previously trained neural network, which allows the prediction of treatment efficiency based on the collected data [5].

However, the standard statistical analysis, as well as used neural network technology is still not enough to efficiently and reliably predict the treatment outcome. Hence, the need for exploration and improvement of existing advanced statistical methods which may be effectively used to predict the results of IVF ICSI/ET infertility treatment methods.

Nearest neighbor based feature selection method and classifier

Over the course of many years of research on the feature selection issue, a number of measures and heuristics that can be applied during determining the significance of features process were found. One of the most interesting heuristics notions is the margin. The margin is understood here as the separation extent between observations of different classes. Intuitively, a larger margin induces easier process of classification, and interpretation of the conclusions drawn from it. In [1] the SIMBA, an algorithm was presented which determines the feature importance as a value dependent to a margin generated by it. Used here, the MSIMBAF2 algorithm is a modification and generalization of the following algorithms: MSIMBAF [6], SIMBA [1], Relief [2] and ReliefF [3]. MSIMBAF and MSIMBAF2 algorithms are very similar, therefore most of the equations described latter are the same as the first. The most important difference between them is the modified distance measure (1), and subsequent changes caused by it.

$$\Delta_p(\boldsymbol{x}_1, \boldsymbol{x}_2) = \left\{ \sum_o \left[\alpha(t_o)\phi_{type(o)}(x_{1o}, x_{2o})\right]^p \right\}^{1/p} \tag{1}$$

where: $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ observation; p metric order; $\Delta_p(*,*)$ observation distance measure of order p; $\alpha(*)$ specific function; o index of feature; t_o “hidden” parameters; x_{1*}, x_{2*} value of feature $*$ for given observation; $\phi_{type(o)}(*,*)$ measures of dissimilarity between single values of feature o with given $type(o)$;

$$\phi_{type(o)}(x_{1o}, x_{2o}) = \begin{cases} \phi_{num}(x_{1o}, x_{2o}) & (a) \\ \phi_{ord}(x_{1o}, x_{2o}) & (b) \\ \phi_{cat}(x_{1o}, x_{2o}) & (c) \end{cases} \tag{2}$$

$$\alpha(t_o) = \frac{1}{\pi}\left(\operatorname{arctg}(t_o) + \frac{\pi}{2}\right) = \frac{\operatorname{arctg}(t_o)}{\pi} + \frac{1}{2} \in (0,1); \quad t_0 \in \Re \tag{3}$$

$$\delta_o = \max_{b,d} |x_{bo} - x_{do}|\,; \quad 0 \leq \varepsilon_{o1} \leq \varepsilon_{o2} \leq 1 \tag{4}$$

$$\phi_{num}(x_{1o}, x_{2o}) = \min\left(1, \max\left(0, \frac{|x_{1o} - x_{2o}| - \delta_o \varepsilon_{io}}{\delta_o(\varepsilon_{02} - \varepsilon_{o1})}\right)\right) \tag{5}$$

$$\phi_{ord}(x_{1o}, x_{2o}) = \min\left(1, \max\left(0, \frac{|x_{1o} - x_{2o}| - \delta_o \varepsilon_{o1}}{\delta_o(\varepsilon_{02} - \varepsilon_{o1})}\right)\right) \tag{6}$$

$$\phi_{cat}(x_{1o}, x_{2o}) = \begin{cases} 0 \Leftrightarrow x_{1o} = x_{2o} \\ 1 \Leftrightarrow x_{1o} \neq x_{2o} \end{cases} \tag{7}$$

where: $\phi_{num}(*,*)$, $\phi_{cat}(*,*)$, $\phi_{ord}(*,*)$ measures of dissimilarity between numeric (a), categorical (b) or order (c) feature values; t_o “hidden” o-th feature weight parameter; δ_o value range of o-th feature; ε_{o1} and ε_{o2} cut-off parameters for o-th feature (there are only 2 such parameters for each feature);

Measure (1) resembles Minkowski metric of order of p. It is the main distance function used by algorithm. Index o iterates through all the features of the train data set. $\alpha(t_o)$ is scale factor as well as feature weight. It is a sigmoidal function (3) of internal parameter t_o. Depending on the feature type, algorithm choose suitable dissimilarity measure between values (2). For numerical features measure (2a) is chosen, for ordinal – (2b), for categorical – (2c). These are described in equations (5), (6), (7).

$$m = \sum_{x \in X} \left[\Delta_p(\boldsymbol{x}, miss\,(\boldsymbol{x}, u)) - \Delta_p(\boldsymbol{x}, hit\,(\boldsymbol{x}, u))\right] - \varepsilon \sum_{o} \alpha(t_o) \tag{8}$$

$$dt_o = \frac{\partial \Delta_p(\boldsymbol{x}, miss\,(\boldsymbol{x}, u))}{\partial t_o} - \frac{\partial \Delta_p(\boldsymbol{x}, hit\,(\boldsymbol{x}, u))}{\partial t_o} - \varepsilon \frac{\partial \alpha(t_o)}{\partial t_o} \tag{9}$$

$$t_o = t_o + dt_o \tag{10}$$

$$\frac{\partial \Delta_p(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\partial t_o} = \phi_{type(o)}(x_{1o}, x_{2o}) \left[\frac{\alpha(t_o) \cdot \phi_{type(o)}(x_{1o}, x_{2o})}{\Delta_p(\boldsymbol{x}_1, \boldsymbol{x}_2)}\right]^{p-1} \frac{\partial \alpha(t_o)}{\partial t_o} \tag{11}$$

$$\frac{\partial \alpha(t_o)}{\partial t_o} = \frac{1}{\pi(1 + t_o^2)} \tag{12}$$

where: m margin; $\boldsymbol{x}$ observation; $hit(\boldsymbol{x}, u)$ u-th nearest observation of class same as $\boldsymbol{x}$; $miss(\boldsymbol{x}, u)$ u-th nearest observation of class different from $\boldsymbol{x}$; ε penalty factor; dt_o adjustment for o-th feature

MSIMBAF2, like MSIMBAF utilizes gradient optimization (9), (10) of margin (8) relative to the internal t_o parameter of features weight $\alpha(t_o)$.

ε parameter act as extra weight regulation. It is a kind of punishment level, that makes preferable to the algorithm to search for weights with the lowest possible sum. This factor also causes the gradual diminishing of weights for features that have the same values (or many missing values). Due to different distance measure (1), its (partial) derivative with respect to the o-th feature weight form (11) is changed. Categorical and ordinal feature weights are calculated now in the same way as the numerical ones. The MSIMBAF algorithm updated non-numeric feature weights similar to the ReliefF algorithm [3]. This could lead to the selection of a sub-optimal features subset (Relief has no mechanism against feature redundancy.). With the introduced modification, MSIMBAF2 algorithm can detect redundant ordinal and categorical features as well as numerical ones.

KNN (K Nearest Neighbor) is one of the simplest supervised classification methods to apply. It is based on a simple assumption of similarity of the same class objects. To determine the class of an observation, kNN classifier searches for the specified number of other observations of a known class, most similar to the given one (called neighbors). The dominant class among found neighbors becomes the new observation target class. This determines the division of the feature space into distinct decision areas of (strongly) nonlinear boundary which are dominated by points that have known common class. There are a number of measures that approximate the level of (non-)similarity. For the purpose of IVF data classification the following measure of dissimilarity (distance) was used:

$$\begin{aligned} d_{p'}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \left[\sum_i \left(\frac{|x_{1i} - x_{2i}|}{\delta_i}\right)^{p'}\right]^{1/p'} + \sum_j \phi_{cat}(x_{1j}, x_{2j}) + \\ + \sum_k \frac{|x_{1k} - x_{2k}|}{\delta_k} \end{aligned} \tag{13}$$

where: $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ observation; p' metric order; $d_{p'}(*, *)$ observation distance measure of order p'; x_{1*}, x_{2*} value of feature $*$ for given observation.

In case of missing data the following rules were used:

– When for first and second observation for given feature both values are missing, distance between them is set to 0.
– When numerical value is missing, it is replaced by mean value among this feature values.
– When categorical value is missing, distance is probability that category is different than in compared observation.
– When ordinal value is missing, it is replaced by median value among this feature values.

These rules and the measure (13) are the same as for the MSIMBAF algorithm [6]. The main difference is that now they are used not only in the feature selection algorithm MSIMBAF2 (to find $hit(\boldsymbol{x}, u)$ and $miss(\boldsymbol{x}, u)$ in (8) and (9)), but also in the target kNN classifier.

Analysis method and results

Data set generated using the database [4] was analyzed. Diagram of this set is presented in [Tab. 1]. Each of the 1445 observation corresponds to one cycle of infertility treatment and is described by 150 features. The treatment outcome is dependent feature (pregnancy or no).

Tab. 1. IVF ICSI/ET treatment data scheme

		150 features			
		111 numerical	1 ordinal	37 categorical	1 categorical dependent
1445 observations	486 positive outcome	....	....	....	....
	959 negative outcome	....	....	....	....

To analyze the data set, MSIMBAF2 feature selection algorithm and kNN classifier were used. The parameters of these algorithms are as follows:

- MSIMBAF2 feature selection
  - metric order $p = 2.5$
  - lower cut-off level $\varepsilon_{o1} = 0.1$
  - upper cut–ff level $\varepsilon_{o2} = 0.9$
  - penalty weight factor $\varepsilon = 0.001$
- kNN classifier
  - metric order $p' = 1$

To reduce final result bias, following cross validation procedure was used (in bracket number of observations is given):

1. choose two random subsets of original data set (1445 obs.): validation (481 obs.) and learning (964 obs.)
2. 250 times make:
   2.1. choose two random subsets of learning data set: train (482 obs.) and test (482 obs.)

```
    2.2. train MSIMBAF2 algorithm on train set
    2.3. from feature number $i = 15$ to $i = 149$
        2.3.1. choose $i$ features from train and test set to obtain two “truncated” data sets
        2.3.2. on “truncated” data set learn kNN classification algorithm
        2.3.3. assess the effectiveness of algorithm-learned decision rules on “truncated” data sets: train and test
3. for each feature number from $\langle 15;149 \rangle$ interval search for decision rules with the best accuracy of the 250 runs [Fig. 1].
```

For further analysis subset of 43 attributes and 482 observations (this was “truncated” train data set, on which kNN algorithm presents maximum efficiency [Fig. 1]) was selected, and kNN classification algorithm was train on it. Validation set was also truncated to those 43 features. [Tab. 2] shows trained classifier accuracy on truncated validation set. Because validation set was not used in learning phase, that accuracy [Tab. 2] is expected to be unbiased.

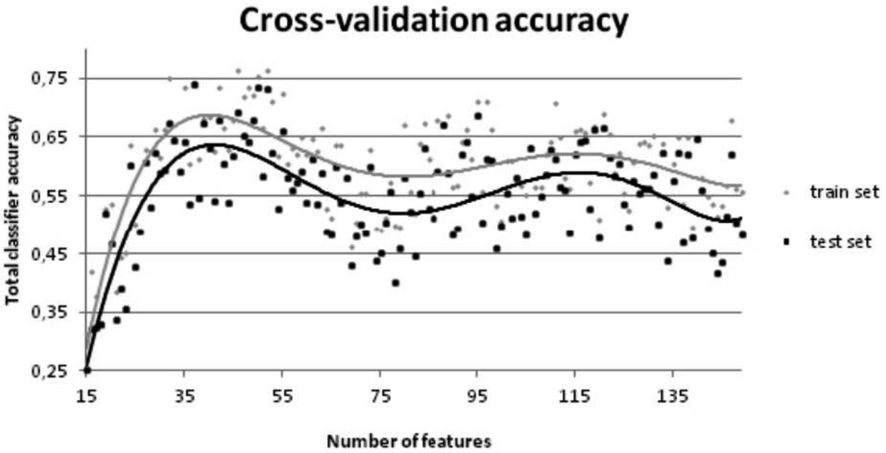

Fig. 1. Train and test set top total classification accuracies from 250 iterations, with approximation lines

[Fig. 2] shows 19 features with the largest weight of the 43 selected. Many of them (endometriosis, male factor, the protocol type of treatment) have been suspected for a long time as having a significant impact on the treatment effectiveness. Most interestingly, some features, previously considered as insignificant, also entered into the prediction model. Their presence probably boosts the predictive power of the model in presence of other selected features and causes higher accuracy achievement.

Tab. 2. Cross-classification accuracy on validate set depending on observed outcome

Total accuracy: 65%		Observed outcome		Total observations
		negative	positive	
Predicted outcome	negative	70% (271 o.)	30% (121 o.)	100% (392 o.)
	positive	54% (48 o.)	46% (41 o.)	100% (89 o.)
Total observations		(319 o.)	(162 o.)	481 o.

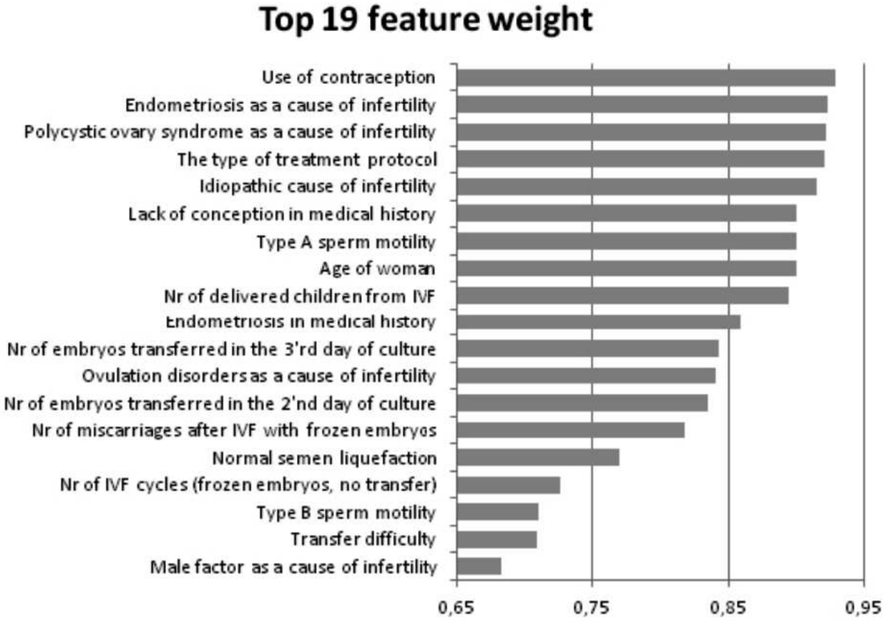

Fig. 2. Top 19 features from 43 according to their presented weights

Conclusions

Cross-validation test results are very promising. It turns out that the reduction in the number of features to 30% (43 of 149) can be performed without the loss of kNN classifier accuracy. For a smaller number of features of the overall accuracy of the classification quickly erodes. The classification error on the validation set has the same magnitude as on the test set, which confirms the resistance of the generated decision rules to bias. Unfortunately, the cross-classification results do not allow to fully predict pregnancy or lack of it. The algorithm properly predicts the absence of pregnancy in 70% of the cases and its presence in only 46%. The above analysis is somewhat consistent with [5], where used the neural network predicted negative cases with much greater accuracy than positive.

Further research should focus on the use of the weighted average prediction accuracy for both positive and negative results, as opposed to the overall accuracy for the cross-validation procedure. Feature selection algorithm also requires further work. With the change of metrics, the margin function is strongly nonlinear due to the weight, so special methods of optimization should be used. It is suspected that the modifications described above can improve the parameters of the developed model to a satisfactory level, enabling it to be used in clinical practice to predict new cases of treatment effectiveness.

REFERENCES

[1] Gilad-Bachrachy R., Navot A., Tishby N., Margin Based Feature Selection – Theory and Algorithms, Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 2004.

[2] Kira K., Rendell L., A practical approach to feature selection, Proceedings 9^{th} International Workshop on Machine Learning, pp. 249–256, 1992.

[3] Kononenko I., Estimating attributes: analysis and extensions of Relief. Bergadano F., De Raedt L. ed, Proceedings European Conference on Machine Learning, 1994.

[4] Milewski R., Jamiołkowski J., Milewska A. J., et al., The system of electronic registration of information about patients treated for infertility with the IVF ICSI/ET method, Studies in Logic, Grammar and Rhetoric, 17 (30), 2009.

[5] Milewski R., Jamiołkowski J., Milewska A. J., et al., Prognozowanie skuteczności procedury IVF ICSI/ET – wśród pacjentek Kliniki Rozrodczości i Endokrynologii Ginekologicznej – z wykorzystaniem sieci neuronowych, Ginekologia Polska, 80 (12), 2009.

[6] Milewski R., Malinowski P., Milewska A. J., et al., Usage of margin-based feature selection algorithm in IVF ICSI/ET data analysis, Studies in Logic, Grammar and Rhetoric, 21 (34), 2010.

[7] Milewski R., Milewska A. J., Domitrz J., Wołczyński S., In vitro fertilization ICSI/ET in women over 40, Przegląd Menopauzalny, 2, 2008.

[8] Milewski R., Milewska A. J., Jamiołkowski J., et al., The statistical module for the system of electronic registration of information about patients treated for infertility using the IVF ICSI/ET method, Studies in Logic, Grammar and Rhetoric, 21 (34), 2010.

[9] Radwan J. (ed.), Niepłodność i rozród wspomagany, Termedia, Poznań, 2005.

[10] Radwan J., Wołczyński S. (ed.), Niepłodność i rozród wspomagany, Termedia, Poznań, 2011.