DE GRUYTER OPEN

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 39 (52) 2014

DOI: 10.2478/slgr-2014-0044

The Use of Data Mining Methods to Predict the Result of Infertility Treatment Using the IVF ET Method

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

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

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

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

Abstract. The IVF ET method is a scientifically recognized infertility treatment method. The problem, however, is this method's unsatisfactory efficiency. This calls for a more thorough analysis of the information available in the treatment process, in order to detect the factors that have an effect on the results, as well as to effectively predict result of treatment. Classical statistical methods have proven to be inadequate in this issue. Only the use of modern methods of data mining gives hope for a more effective analysis of the collected data. This work provides an overview of the new methods used for the analysis of data on infertility treatment, and formulates a proposal for further directions for research into increasing the efficiency of the predicted result of the treatment process.

Introduction

Infertility is a disease affecting an increasing number of married couples around the world seeking to have offspring. Often the only effective way to deal with this situation is to turn to Assisted Reproductive Technologies (ART), including, inter alia, In Vitro Fertilization with embryo transfer (IVF-ET). The effectiveness of this type of treatment has increased in recent years by almost 10%, but is still fits in only about 40% of cases (Milewski et al., 2013). In addition to the implementation of more and better technological solutions, effective predictors of treatment results, which allow the treatment process to be customized for each case in or-

ISBN 978–83–7431–428–2 ISSN 0860-150X

der to maximize the chance of a positive result, also have an impact on raising the effectiveness of treatment. Traditional statistical methods are not able to cope with this challenge. Hence the need to appeal to much more advanced methods like data-mining, or even methods of artificial intelligence.

Methods for the Analysis of Data Obtained from the IVF ET Treatment Process

Artificial Neural Networks are one of the most popular and most commonly used artificial intelligence methods in medicine. An artificial neural network approach was formed as an attempt to transfer the operation of the human brain using simple mathematical tools and appropriate algorithms. A single neuron in a typical network performs a weighted sum of the inputs, which is transformed by a special kind of activate function, and the result is sent to the output of the neuron. Through appropriately controlled connection strength (input weights) between individual neurons, such a network may reflect the nature of the phenomena studied. A breakthrough idea, as far as the application of ANNs in artificial intelligence, was the discovery of the mechanism of the reverse learning network (Rumelhart et al., 1986). It allows for weight correction of inner network connections between layers on the basis of a comparison of the anticipated (the output of the network) and the correct result.

The described neural network implementation has been applied to result prediction of infertility treatment with the IVF method (Milewski et al., 2009a). The parsed data set, which has been harvested using specialized software, spans a period of over three years of research, including more than 1,000 cycles of fertility treatment (Milewski et al., 2009b). The method of analysis used combined a reduction in dimension along with training on the target network. The original data set was used to train a large amount of neural networks, of which the best was selected. After the identification of the most significant factors (42 out of over 150 features), the training phase of selected neural network models was repeated. Finally, the selected network consisted of 3 layers with 45 (input), 14 (hidden layer) and 1 (network output) neuron, respectively. The trained network was very effective in identifying negative cases (no pregnancy), in which the correct result was obtained almost 90% of the time, which is a significant result. Of much lower interest was the ability to properly classify a positive result – slightly less than 50%.

Self-Organizing Feature Maps, also called Kohonen Neural Networks are related to ANNs (Kohonen, 1982; Milewska et al., 2013). A self-organizing Kohonen network approach in the analysis of data on infertility treatment was proposed by Siristatidis (2011). Such a network adapts topologically to selected data patterns (Kohonen, 1988), building a map of data patterns. The trained network has to demonstrate highly efficient prediction capabilities (in the authors' assumption of more than 75%). In the selection of variables (analysed features), the authors followed a review of literature. The work by Siristatidis (2011) considered in this review deserves a special mention because it provides an overview of the many methods and their application to predicting results of treatment by IVF.

Feature selection is a very important phase of the analysis because it allows you to designate the most relevant factors in the context of a particular result. Milewski et al. (2010) and Milewski et al. (2011) apply feature selection to the analysis of a data set, used also in a previous Milewski et al. (2009a) publication, but supplemented by a further 3 years of research. After the initial selection of cases, it includes more than 1,400 treatment cycles. The data set was treated by the feature selection with the original methods of SIMBAF and MSIMBAF2, which are interesting generalization methods of SIMBA, among others, on non-numeric data. Starting from the following form of generalized function of distance is:

$$\Delta(\boldsymbol{x}_1, \boldsymbol{x}_2) = \begin{cases} \left\{ \sum_i \left[\alpha(t_i) \phi_{num}(x_{1i}, x_{2i}) \right]^p \right\}^{1/p} + \\ + \sum_j \alpha(t_j) \phi_{cat}(x_{1j}, x_{2j}) + \sum_k \alpha(t_k) \phi_{ord}(x_{1k}, x_{2k}) & * \\ \left\{ \sum_o \left[\alpha(t_o) \phi_{type(o)}(x_{1o}, x_{2o}) \right]^p \right\}^{1/p} & ** \end{cases}$$

where: * – SIMBAF algorithm; ** – MSIMBAF2 algorithm; $\Delta(x_1, x_2)$ – a measure of the distance between observations; i, j, k, o – leading indexes after the appropriate features; $\phi_*(x_{1*}, x_{2*})$ – a measure of separation for individual features, x_{**} – the value of the specified feature of a particular observation; $\alpha(t_*)$ – feature weight function

The algorithm optimizes the so-called margin – a value that is a measure of separation between observations belonging to different classes:

$$m = \begin{cases} \sum_{x \in X} [\Delta(\boldsymbol{x}, miss(\boldsymbol{x}, u)) - \Delta(\boldsymbol{x}, hit(\boldsymbol{x}, u))] & * \\ \sum_{x \in X} [\Delta(\boldsymbol{x}, miss(\boldsymbol{x}, u)) - \Delta(\boldsymbol{x}, hit(\boldsymbol{x}, u))] - \varepsilon \sum_{o} \alpha(t_o) & ** \end{cases}$$

where: m – margin; $miss(\boldsymbol{x}, u)$ – u'th neighbor of a different class; $hit(\boldsymbol{x}, u)$ – u'th neighbor of the same class

MSIMBAF2 completes the margin with a LASSO type normalization stage, which task is to additionally lower unnecessary weights. Optimization by the hidden t parameters (with a simple gradient method) allows a set of features that specify the largest margin to be found.

Milewski et al. (2010) focused on the more in-depth analysis of previously selected most important factors. A significant subset of the eventually selected features was suspected for a long time to be of important impact on the results of infertility treatment, as directly related to the occurrence of infertility, the treatment process, or the factors related to the causes of infertility. The above set of features includes, among others: the age, the type of treatment protocol (protocol types described by Milewski et al. (2009b)), causes of infertility (fallopian tube factor, polycystic ovary syndrome, endometriosis, male factor), the number of embryos transferred on the second and third days of culture, expanding the number of blastocysts on the fifth day of culture, semen parameters and preparation, hyperprolactinemia in medical history, which has been proven to have a known impact on the results of the treatment before. There were, however, also features such as mucus during ovulation, the type of anesthesia at the puncture of ovarian follicles, pain, temperature increase and blood spotting during ovulation, whose impact on the results so far was not seriously considered. There are no known mechanisms that explain why these features are considered essential.

A development of the ideas described by Milewski et al. (2010) is present in another article by Milewski et al. (2011). It takes the following approach to the classification of cases of infertility treatment using the k nearest neighbors. The classifier is assisted by the preceding feature selection phase using the previously mentioned MSIMBAF algorithm. The whole has been inscribed in the loop of the cross validation in‘order to obtain unbiased results. With a validation set, 65% of the observations have been correctly classified, including 70% of the negative cases and 46% of the positive ones. At the same time, the dimension set was reduced to 30% (43 features with 149).

Milewski et al. (2012) and Malinowski et al. (2013) focus on using stronger classifiers: a Random Forest and the SVM across the data set with-

out any dimension reduction phase. The idea of Random Forest, originally proposed by Breiman (2001), involves the construction of a series of decision trees, which then collectively classify the observations by the vote majority. Each of the trees is built using a random subset of features and observations. A completely different methodology is applied by the SVM (Boser et al., 1992). This method builds a linear classifier in the transformed (and often highly dimensional) features space. Because the selection of such a classifier is not clear, the additional assumptions of the optimal solution are imposed here, including the maximum separation of observations. In both articles, single data completion methods were used, including:

- completion with the median, the fashion, the dominant
- completion as above, but in a sample is referred to the nearest neighborhood of the observation (Templ et al., 2013)
- completion based on the Random Forest algorithm (Breiman, 2001; Stekhoven et al., 2012).

Using the relevant procedures of cross validation, all possible combinations of algorithms were tested in order to find the best of them. As a result of the analysis of Milewski et al. (2012), the Random Forest classifier in conjunction with the relevant data completion procedure, obtained a significant result of 79% compatibility prediction (performance prediction for a negative result – 88%, for the result of the positive – 66%). Noteworthy is the fact that this result was obtained on a validation set, so this is potentially a bias-free classification quality estimator. Interestingly, the SVM classifier could not get satisfactory accuracy, usually classifying all cases as a lack of pregnancy.

The above-mentioned results served as a hint for Malinowski et al. (2013). This time the unattended version of data completion algorithms with a preselected classification algorithm – Random Forest – was used. This time, the purpose of the study was to predict the treatment result as early as possible. With the aim of reaching this goal, the original data set was divided into 4 subsets, corresponding to specific phases of treatment. The work did not produce qualitatively significant results. However, some aspects are worth notice:

- There is already a possibility of prediction at the beginning of treatment with a precision of up to 65%, including a similar result prediction for the success or failure of treatment, but:
- Information from anamnesis is insufficient for predicting the success of treatment (50% of the classifier relevant response).

Conclusions and Suggestions for Further Methods of Analysis

Modern data mining methods allow even more accurate analysis of information, significantly more accurate than the traditional statistical methods. As one can see, their usage allows for even more accurate prediction of such complex issues as infertility treatment with methods of assisted reproduction. Reflecting on the information presented in the previous chapter allows one to draw the following conclusions:

- The high effectiveness of classification methods based on an optimization approach (ANN) or a randomized one (RF) with a slight superiority of the last, should be noticed.
- Effective selection of features allows one to increase the accuracy of the classification.
- We have found a small SVM classifier performance that has been noticed in Uyar's et al. work (2009), which may be the result of a large number of discrete features.
- Prediction of a negative treatment result is much simpler to obtain than prediction of a positive one.

One of the simpler ways of raising algorithm awareness regarding the positive cases is to give them greater weight. Virtually all of these algorithms allow one to perform such a process. Generalizing the probabilistic approach to data analysis methods with methods based on the immediate vicinity has recently been introduced. Examples of this that should be mentioned include the RKNN classification method and related RKNN-FS (Li et al., 2011) feature selection method. An interesting idea, therefore, becomes to connect the MSIMBAF2 algorithm with the aforementioned to find an even better feature selection method and a classifier. Target feature selection algorithms and classifiers would be surrounded by an appropriate procedure of cross-validation in order to obtain as many unbiased results as possible and maximum generalization capability. There is no doubt that the efforts described above should continue, increasing step-by-step abilities of considered predictor models, which in clinical practice undoubtedly can be translated into generated treatment effectiveness.

REFERENCES

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In D. Haussler (Ed.), *5th Annual ACM Workshop on Computational Learning Theory* (pp. 144–152). Pittsburgh, PA, USA: ACM Press.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.

Kohonen, T. (1982). Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics*, 43, 59–69.

Kohonen, T. (1988). *Self-organization and associative memory* (2nd ed.). NY, USA: Springer-Verlag.

Li, S., Harner, E. J., & Adjeroh, D. A. (2011). Random KNN feature selection – a fast and stable alternative to Random Forests. *BMC Bioinformatics*. 12, 450. DOI:10.1186/1471-2105-12-450.

Malinowski, P., Milewski, R., Ziniewicz, P., Milewska, A. J., Czerniecki, J., & Wołczyński, S. (2013). Classification issue in the IVF ICSI/ET data analysis: early treatment outcome prognosis. *Studies in Logic, Grammar and Rhetoric. Logical, Statistical and Computer Methods in Medicine*, 33(45), 103–115.

Milewska, A. J., Jankowska, D., Cwalina, U., Więsak, T., Citko, D., Morgan, A., & Milewski, R. (2013). Analyzing Outcomes of Intrauterine Insemination Treatment by Application of Cluster Analysis or Kohonen Neural Networks. *Studies in Logic, Grammar and Rhetoric. Logical, Statistical and Computer Methods in Medicine*, 35(48), 7–25.

Milewski, R., Jamiołkowski, J., Milewska, A. J., Domitrz, J., Szamatowicz, J., & Wołczyński, S. (2009a). Prognozowanie skuteczności procedury IVF ICSI/ET – wśród pacjentek Kliniki Rozrodczości i Endokrynologii Ginekologicznej – z wykorzystaniem sieci neuronowych. *Ginekologia Polska*, 80(12), 900–906.

Milewski, R., Jamiołkowski, J., Milewska, A. J., Domitrz, J., & Wołczyński, S. (2009b). 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), 225–239.

Milewski, R., Malinowski, P., Milewska, A. J., Czerniecki, J., Ziniewicz, P., & Wołczyński, S. (2011). Nearest neighbor concept in the study of IVF ICSI/ET treatment effectiveness. *Studies in Logic, Grammar and Rhetoric. Logical, Statistical and Computer Methods in Medicine*, 25(38), 49–57.

Milewski, R., Malinowski, P., Milewska, A. J., Ziniewicz, P., Czerniecki, J., Pierzyński, P., & Wołczyński, S. (2012). Classification issue in the IVF ICSI/ET data analysis. *Studies in Logic, Grammar and Rhetoric. Logical, Statistical and Computer Methods in Medicine*, 29(42), 75–85.

Milewski, R., Malinowski, P., Milewska, A. J., Ziniewicz, P., & Wołczyński, S. (2010). The usage of margin-based feature selection algorithm in IVF ICSI/ET data analysis. *Studies in Logic, Grammar and Rhetoric. Logical, Statistical and Computer Methods in Medicine*, 21(34), 35–46.

Milewski, R., Milewska, A. J., Czerniecki, J., Leśniewska, M., & Wołczyński, S. (2013). Analysis of the demographic profile of patients treated for infertility using assisted reproductive techniques in 2005–2010. *Ginekologia Polska*. 84(7), 609–614.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536.

Siristatidis, Ch., Pouliakis, A., Chrelias, Ch., & Kassanos, D. (2011). Artificial Intelligence in IVF: A Need. *Systems Biology in Reproductive Medicine*, 57(4), 179–185.

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest – non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112–118.

Templ, M., Alfons, A., Kowarik, A., & Prantner, B. (2013). VIM: Visualization and Imputation of Missing Values. R package version 3.0.3.1. Retrieved from http://cran.r-project.org/package=VIM.

Uyar, A., Bener, A., Ciray, H. N., & Bahceci, M. (2009). A frequency based encoding technique for transformation of categorical variables in mixed IVF dataset. *31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 6214–6217). Minneapolis, USA.