31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

A Frequency Based Encoding Technique for Transformation of Categorical Variables in Mixed IVF Dataset

Asli Uyar, Ayse Bener, H. Nadir Ciray and Mustafa Bahceci

Abstract— **Implantation prediction of in-vitro fertilization (IVF) embryos is critical for the success of the treatment. In this study, Support Vector Machine (SVM) method has been used on an original IVF dataset for classification of embryos according to implantation potentials. The dataset we analyzed includes both categorical and continuous feature values. Transformation of categorical variables into numeric attributes is an important pre-processing stage for SVM affecting the performance of the classification. We have proposed a frequency based encoding technique for transformation of categorical variables. Experimental results revealed that, the proposed technique significantly improved the performance of IVF implantation prediction in terms of Area Under ROC curve (0.712±0.032) compared to common binary encoding and expert judgement based transformation methods (0.676±0.033 and 0.696 ± 0.024, respectively).**

I. INTRODUCTION

In-vitro fertilization (IVF) [1] is a common infertility treatment method during which female germ cells (oocytes) are inseminated by sperm under laboratory conditions. Fertilized oocytes are cultured between 2-6 days in special medical equipments and embryonic growth is observed and recorded by embryologists. Finally, selected embryo(s) are transferred into the woman's womb. Selection of the embryos with highest reproductive viabilities and the decision of number of embryos to be transferred is crucial for achieving successful pregnancies. Predicting implantation potentials of individual embryos may expedite and enhance expert judgement for these critical decisions.

The factors affecting the implantation outcome of embryos are related to patient characteristics and embryo morphological variables [2]. Hence, various patient and embryo related features should be considered as input to a machine learning system for implantation prediction. We have constructed an IVF dataset consisting of data feature vectors for each embryo. Input data features include both continuous (e.g. age, hormone levels etc.) and categorical (infertility factor, treatment protocol etc.) variables.

Analysis and pre-processing of mixed datasets including a combination of continuous and categorical variables has been an important research interest in the last decade [3] [4] [5] [6]. In this study, we analyze the performance of implantation prediction on mixed IVF dataset using Support Vector Machine (SVM) method. SVM is basically a distance based binary classifier constructing the optimal separating hyperplane in the input feature space and performance of distance based classifiers depends on accurate transformation of categorical variables into numeric data. A new frequency based encoding scheme has been proposed in this study and compared to binary encoding and expert judgement transformation methods.

II. RELATED WORK

Binary encoding maps categorical variables to higher dimensional features representing equal Euclidean distances between categories and has been applied as a common pre-processing stage for SVM applications [7] [8].

Johannson, *et al.*, deal with visualization of mixed datasets and propose interactive quantization of categorical variables that incorporates information about relationships among continuous variables as well as makes use of the domain knowledge of the data analyst [9]. A Simple Correspondence Analysis (SCA) has been applied based on the frequencies of categories in the dataset.

A frequency based encoding scheme has previously been proposed as a data transformation technique for car injury prediction [10]. The proposed method represents the categorical code of a particular variable with a numerical value derived from its relative frequency between injury and non-injury outcomes as defined below:

$$v_{ik} = P(injury)_{ik}/(P(injury)_{ik} + P(non-injury)_{ik})$$

where, v_{ik} is the new numerical value of categorical variable v_i originally in code k. The authors reported that the proposed frequency based technique outperformed the traditional 1-to-N binary coding.

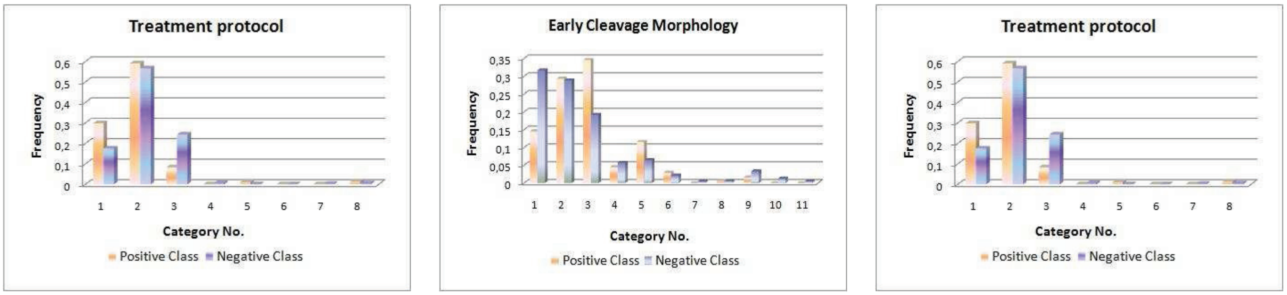
Consequently, dealing with mixed data types within a single classifier is an important pre-processing stage affecting the performance of the learning algorithms. Each real world application domain requires detailed analysis of the data type conversion methods in order to state the best fitting model.

III. PROBLEM STATEMENT

In this study, we analyze IVF procedure and related database of IVF Unit of German Hospital in Istanbul. Dataset features and data types are given in Table I. The features have been selected depending on experiences of senior embryologists in [11] and related studies in the literature [12]. The IVF dataset includes 2453 fresh, non-donor in-vitro human embryos transferred in day 2 or day 3 after Intra-Cytoplasmic Sperm Injection (ICSI) and each embryo data vector is represented by 12 feature values. There are two classes of

A. Uyar and A. Bener is with Bogazici University, 34342 Bebek Istanbul, Turkey, asli.uyar, bener@boun.edu.tr
H.N. Ciray and M. Bahceci is with the IVF unit of German Hospital, Istanbul, Turkey, nadirc, mbahceci@superonline.com

978-1-4244-3296-7/09/$25.00 ©2009 IEEE



(a) Treatment Protocol (b) Early Cleavage Morphology (c) Infertility Factor

Fig. 1. Distribution of categories for each categorical variable among both positive and negative implantation classes

TABLE I
SELECTED DATASET FEATURES FOR EACH EMBRYO FEATURE VECTOR

Dataset Features	Data Type
Patient Characteristics	
Woman age	Numerical
Infertility factor	Categorical
Treatment protocol	Categorical
Follicular stimulating hormone dosage	Numerical
Peak Estradiol level	Numerical
Embryo Related Data	
Early cleavage morphology	Categorical
Early cleavage time	Numerical
Transfer day	Categorical
Number of cells	Numerical
Nucleus characteristics	Numerical
Fragmentation rate	Numerical
Equality of blastomeres	Numerical

embryos labeled as 1 and -1 indicating that implantation was successful or not-successful, respectively.

The dataset includes three categorical variables: infertility factor, treatment protocol and early cleavage morphology with 14, 8 and 11 categories, respectively. Fig. 1 represents the distribution of the categorical variables among both positive and negative implantation classes.

Transformation of categorical variables into numeric values is an important pre-processing step for SVM classifier. Defining the most proper method for transformation is critical for achieving better prediction results. The aim of transformation is to accurately represent the relative distances between categorical codes in multidimensional input space.

Previously, we have applied binary encoding technique as a routine transformation process, however, contrary to the previous work [13] our SVM results had poor performance. Therefore we tackled the efficiency of the common binary encoding method. In this study, we propose a self-learning frequency based encoding technique and compared this new approach to binary encoding and expert judgement of ordering categorical variables.

IV. METHODOLOGY

A. Support Vector Machines

Given a set of training data pairs $(x_i, y_i),\ y_i \in \{+1, -1\}$, the aim of SVM classifier is to estimate a decision function by constructing the optimal separating hyperplane in the feature space [14]. The key idea of SVM is to map the original input space into a higher dimensional feature space using kernel functions. Final decision function is in the form:

$$f(x) = \left(\sum_i \alpha_i y_i K(x_i \cdot x) + b \right) \tag{1}$$

where $K(x_i \cdot x)$ is the Kernel transformation. SVM requires each data sample to be represented as a feature vector of real numbers [7]. Therefore, categorical features should be converted into numeric values prior to classification. After transformation of categorical variables, the input data were normalized to 0 mean and standard deviation of 1.

In this study, Gaussian kernel has been used because of the superior performance (data not shown here). The hyperparameters of SVM with Gaussian kernel were optimized using a grid search method: *cost* and *gamma* are searched in the ranges [2^{-5}, 2^{15}], and [2^{-15}, 2^{3}], respectively, using cross validation on the training set. The optimum parameters with the highest area under the Receiver Operating Characteristic (ROC) curve (AUC) has been used on the test set.

B. Performance Evaluation

The performances of classifiers were compared using ROC analysis. The ROC curve plots the True Positive (TP) rate (i.e. a measure of accuracy for correctly detecting the embryos that will implant) vs. False Positive (FP) rate, (i.e. erroneous positive implantation prediction) by adjusting the decision threshold of classification and enables comparison of techniques using a single performance measure that is the AUC. The AUC represents the most informative and objective measure within a benchmarking context [15] especially in case of imbalanced class distributions [16]. The dataset used in this study has an imbalanced nature consisting of 89% not-implanted and 11% implanted embryos. Hence, SVM parameter optimization and comparison of data type transformation schemes have been performed according to AUC measure. The ROC curve of SVM is obtained by adjusting the classification threshold in the range of minimum and maximum distance outputs, where the default threshold was 0. The statistical tests on experimental results have been performed using paired t-test with P = 0.05.

V. TRANSFORMATION OF CATEGORICAL VARIABLES

The aim of data type transformation is to preserve the information content of the original dataset while adapting

TABLE II

EXAMPLE TRANSFORMATION OF 'TREATMENT PROTOCOL' FEATURE INCLUDING 8 CATEGORIES

Original category code	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Binary encoding	00000001	00000010	00000100	00001000	00010000	00100000	01000000	10000000
Frequency based transformation	0.123	0.024	-0.16	-0.006	0.0094	0	-0.0031	0.0086
Expert judgement	3	3	3	2	1	1	4	2

the input data to a particular analysis tool.

A. Binary Encoding

For a particular categorical variable including N categories, each category is represented by a sequence of N bits. The i^{th} bit corresponding to original category is set to 1 and the others are set to 0. For example, the treatment protocol feature in IVF dataset includes eight categories. When binary encoding is applied, the categories 1,2...8 correspond to 00000001, 00000010... 10000000 respectively. In this case the Euclidean distance between each category is equal, however, this may not be the actual case. Also, the input dimensionality is increased by adding dummy variables that may yield to "curse of dimensionality" in learning phase. The input dimension of our dataset is increased to 42 from initial 12 features after binary encoding.

B. Proposed Frequency Based Encoding Technique

The literature presents various solutions of binary encoding, frequency based and expert judgement approaches for transformation of categorical variables as mentioned before. However, comparative analysis of these methods is limited and also, to the best of our knowledge, there is not a generalized frequency based encoding scheme. In this study, we propose a new frequency based transformation method for continuous numerical representation of categorical variables in mixed IVF data. The new numerical values are derived from the relative frequencies of categorical codes among both positive and negative implantation classes as defined below:

$$v_{ik} = P(C_p)_{ik} - P(C_n)_{ik}$$

where,

- v_{ik} is the new numerical value of categorical variable v_i originally in code k;
- $P(C_p)_{ik}$ is the frequency of categorical code k in positive implantation class C_p, and
- $P(C_n)_{ik}$ is the frequency of categorical code k in negative implantation class C_n.

The basic idea behind this transformation is to reflect the effect of categorical code on implantation outcome. The frequency of any categorical code in positive class is assumed to have positive effect while the occurrence in negative class is considered as negative effect. Hence, the new numerical value of a categorical code is defined as the difference between frequencies in positive and negative classes in the range of [-1,1]. Again for treatment protocols, the categories 1,2,3...8 correspond to 0.123, 0.024, -0.16...0.0086 as a result of frequency transformation as shown in Table III. The frequency based encoding has the advantage of self-learning from the training set and therefore supposed to minimize the bias of transformation. This method also has the advantage of preserving the original number of features [5].

C. Expert Judgement

This approach transforms the categories manually, making use of the domain knowledge and experience of medical specialists. The senior embryologists in IVF unit of German Hospital were asked to assign a numerical value to each category representing the relative predictor effect of that category on implantation outcome. They have assigned numerical values from the set of 1,2,3,4, where the greater numbers represent more predictor effect. For the treatment protocols, the manually assigned values are shown in Table II. The same strategy has also been applied to early cleavage morphology and infertility factor variables. This approach may be useful for reflecting user control, however may also insert bias to the original data distribution.

VI. EXPERIMENTS AND RESULTS

A. Training and Testing Strategy

Two-thirds of the dataset was randomly selected for establishing a predictor model and the remaining one-third was utilized for testing. This random splitting has been performed using stratification principle in order to ensure that the proportions of positive and negative classes remain the same in both training and test sets as in the original dataset. The model parameters were optimized on the 2/3 dataset by 10 fold cross validation. After selecting the best parameters, the trained model was tested on the separate 1/3 dataset. The random two-thirds, one-third partitioning of dataset into training and test sets has been repeated 10 times in order to overcome sampling bias. The presented results are the mean and average of these 10 repetitions.

B. Results

The average ROC curves of SVM classification of embryos using three different transformation schemes have been represented in Fig. 2. For clarity, the results have also been shown in Table III in terms of AUC, TP rate, FP rate and accuracy. Statistical tests on the results reveal that, the proposed frequency based encoding technique significantly improves the performance of classification in AUC measure compared to binary encoding scheme (0.712±0.032 and 0.676±0.033 respectively) with p-value = 0.027. The proposed method also increase accuracy and reduce FP rates as desired while slightly decreasing the TP rate. However, these differences are not significant (p-value $>$ 0.05).

An interesting result is that, each of the three transformation methods utilized in the experiments dominates

TABLE III

COMPARISON OF TRANSFORMATION METHODS FOR CATEGORICAL VARIABLES

Transformation Method	**AUC**	**TP Rate** (%)	**FP Rate** (%)	**Accuracy** (%)
Binary encoding	0.676 ± 0.033	67.9 ± 4.0	37.8 ± 4.3	62.7 ± 3.7
Frequency based encoding	0.712 ± 0.032	65.6 ± 4.9	32.5 ± 7.9	67.3 ± 6.8
Expert judgement	0.696 ± 0.024	69.8 ± 8.2	36.9 ± 6.8	63.7 ± 5.5

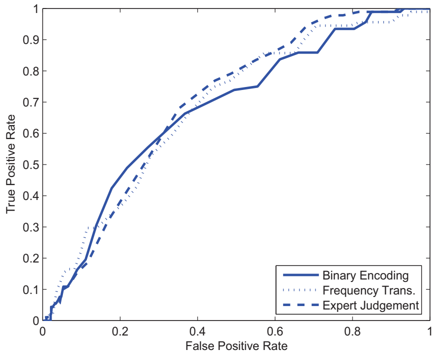


Fig. 2. Demonstration of mean ROC curves for transformation methods

in different parts of ROC area. This may yield to further analysis to combine the three methods for better classification performance in IVF domain.

In machine learning applications, it is crucial to deal with possible biases arising from sampling procedure and training-testing strategies. In order to overcome sampling bias, we have applied stratified 10 fold cross validation. In terms of construct validity, our observations are well translated into clear and widely accepted measures such as AUC, TP rate, FP rate, and accuracy in order to combine our methodology with its theory behind. Statistical validity is also investigated by conducting t-tests. The data comes from a single source challenging the external validity of the results. However, in this domain there are no public datasets nor different labs are willing to share their data.

VII. CONCLUSIONS AND FUTURE WORKS

Each real world application of standard machine learning algorithms require careful pre-processing of input data. Better representation of underlying information content of datasets provides better recognition performance. This is crucial for providing reliable decision support to domain experts especially in medical applications.

Most of the medical datasets include mixed categorical and numerical attributes. This study has examined the effect of categorical variables in SVM classification and presented a comparative analysis of three methods for transformation of categorical variables into numeric values in mixed IVF data. Experimental results have shown that classification after frequency transformation significantly improved the performance of SVM based implantation prediction. We aimed to question the efficiency of traditional binary encoding method and we would conclude that our results represent what we really intend to see.

Future work includes replication of the presented study in different medical datasets to generalize our results. The frequency based transformation method may be modified to enhance the current results by incorporating expert knowledge.

VIII. ACKNOWLEDGMENTS

This study is supported by Bogazici University, (BAP) under grant number 09A104D.

REFERENCES

[1] P. C. Steptoe and R. G. Edwards, "Birth after re-implantation of a human embryo," *Lancet*, vol. 2, p. 366, 1978.

[2] T. Baczkowski, R. Kurzawa, and W. Glabowski, "Methods of embryo scoring in in vitro fertilization," *Reproductive Biology*, vol. 4, no. 1, pp. 5–22, 2004.

[3] R. Brouwer, "A hybrid neural network with fuzzy rules for categorical and numeric input," *International Journal of Intelligent Systems*, vol. 19, pp. 979–1001, 2004.

[4] N. Rogovschi, M. Lebbah, and Y. Bennani, "Probabilistic mixed topological map for categorical and continuous data," in *Seventh Int. Conf. on Machine Learning and Applications*, 2008.

[5] C. Orsenigo and C. Vercellis, "Predicting HIV protease-cleavable peptides by discrete support vector machines," in *EvoBIO*, 2007.

[6] T. Ninomiya, "Clustering observations using fuzzy similarities between ordered categorical data," in *Systems, Man and Cybernetics, IEEE Int. Conf. on*, 2005.

[7] R. Damasevicius, "Optimization of SVM parameters for promoter recognition in DNA sequences," in *International Conference, 20th EURO Mini Conference,Continuous Optimization and Knowledge-Based Technologies (EurOPT-2008)*, 2008.

[8] T. Jung and D. Polani, "Sequential learning with LS-SVM for large-scale data sets," in *ICANN*, 2006.

[9] S. Johansson, M. Jern, and J. Johansson, "Interactive quantification of categorical variables in mixed data sets," in *Proc. of IEEE Int. Conf. on Information Visualisation, IV08*, 2008, pp. 3–10.

[10] C. Nukoolkit, H. Chen, and D. Brown, "A data transformation technique for car injury prediction," in *Proceedings of the 39th Annual ACM-SE Conference*, 2001.

[11] H. N. Ciray, S. Tosun, O. Hacifazlioglu, A. Mesut, and M. Bahceci, "Prolonged duration of transfer does not affect outcome in cycles with good embryo quality," *Fertil. Steril.*, 2007.

[12] D. A. Morales, E. Bengoetxea, B. Larranaga, M. Garcia, Y. Franco, M. Fresnada, and M. Merino, "Bayesian classification for the selection of in vitro human embryos using morphological and clinical data," *Computer Methods and Programs in Biomedicine*, vol. 90, pp. 104–116, 2008.

[13] A. Uyar and F. Grgen, "Arrhythmia classification using serial fusion of support vector machines and logistic regression," in *IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, 2007.

[14] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[15] S. Lessmann, B. Baesens, C. Mues, and S. Pietsch, "Benchmarking classification models for software defect prediction: A proposed framework and novel findings," *IEEE Trans. on Software Engineering*, vol. 34, pp. 485–496, 2008.

[16] A. M. Maloof, "Learning when data sets are imbalanced and when costs are unequal and unknown," *Workshop on Learning from Imbalanced Data Sets*, 2003.