December 2005

Applications of data mining techniques in assisted reproductive technology

Joseph Davis
University of Sydney

Peter Illingworth
University of Sydney

A. Salam
University of Sydney

Applications of Data Mining Techniques in Assisted Reproductive Technology

Dr M Abdus Salam
A/Prof Peter Illingworth
A/Prof Joseph Davis
The University of Sydney

School of Information Technologies and
Western Clinical School, Westmead Hospital
The University of Sydney
Sydney, Australia
Email: msalam@it.usyd.edu.au

Faculty of Medicine
The University of Sydney
Director of Fertility Medicine
Westmead Hospital
Sydney, Australia
Email: petri@westgate.wh.usyd.edu.au

School of Information Technologies
Knowledge Management Research Group
The University of Sydney
Sydney, Australia
Email: jdavis@it.usyd.edu.au

Abstract

In this paper we present a clinically modified info-fuzzy network (IFN) algorithm and a modified composite association rule algorithm for the analysis of discrete valued clinical data obtained at the Westmead Fertility Clinic. The clinically modified IFN (CMIFN) algorithm takes into account the clinical significance of an attribute in relation to a specified target attribute as well as its statistical significance. This gives us the flexibility for exploring the data set according to the clinical question and hypothesis. The MCAR algorithm gives the flexibility of selecting the composition of the "if" node. It also recursively incorporates all relevant attributes. The results show that the MCAR algorithm is marginally better. On the other hand we note that the CMIFN has the ability to produce negative associations as well as positive and nil associations.

Keywords

Medical informatics, data mining

INTRODUCTION

The traditional research on 'evidence-based medicine' is based on the classical works by Cochran (1977), Cochran and Cox (1992) and Snedecor and Cochran (1989) on sampling techniques, statistical methods and experimental designs. Although in other areas of interest, sampling is used more for its advantages such as reduced cost, greater speed, greater scope and greater accuracy, in medicine it is a necessity. Unlike national census where almost every individual is included in the study or survey, in medicine, it is impossible to include every patient in any study. That means the validity of the results very much depends on the statistical properties of the data. Such statistical analysis has limited scope in medicine where it is used for testing a hypophysis such as the comparative efficacy of one treatment over another.

Data mining techniques are widely used in the context of enterprise and business data but are relatively unexploited in the medical and clinical domain. In recent years there has been a great deal of interest in the use of data mining techniques in medical domain with wide ranging applications such as genetic analysis, decision support systems, clinical management and diagnostic problems. Pechenizkiy et. al., (2004) presented an evaluation and comparison of several data mining strategies that apply feature transformation for subsequent classification, and to consider their application to medical diagnostics. Azuaje et. al., (1999) discuss improving clinical decision support system based on data fusion methods. Lenic et. al., (2004) have addressed the hard question of harnessing the subjective and at times conflicting opinions drawn from human expertise in computer based decision support systems. Lavrac et. al., (1997) surveyed the big picture about the data analysis in

medicine. They discussed the use of rule induction, if-then rules, rough sets, association rules, ripple down rules, learning of classification and regression trees, inductive logic programming, discovery of concept hierarchies and constructive induction, case-based reasoning, instances-based learning, neural networks, Bayesian classifier, etc. Jurisica et. al., (1998) proposed a case-based reasoning system to suggest possible modifications to an IVF (In Vitro Fertilization) treatment plan in order to improve overall success rate. This system is named *TA3* $_{IVF}$ system. A practical use of data mining for the prediction of a disease process using statistical methods are demonstrated at a website by Tigrani and John (2005). Lavrac (1999) gave a detailed overview of some of the techniques used in the analysis of the medical data. Brault et. al, (2002) showed the power of data mining in determining interesting patterns from a large data warehouse of a specific disease progression. The medical domain also poses unique ethical issues in data mining in terms of confidentiality as discussed by Berman (2002).

There is a greater need to design algorithms to meet the varying demands of the specific requirements in dealing with medical data. Current success rate of fertility treatments (% of success) is in the low twenties. Our objective is to use data mining techniques to enhance our understanding of the efficacy of various treatment cycles and regimes. For this purpose in this paper we propose and clinically modify Information Fuzzy Network methodology presented by Maimon and Last (2000) that takes into account the clinical significance as well as statistical values. Most techniques such as IFN and Bayesian networks utilize the power of statistical manipulation, which is widely applicable in the business and other non-medical domains. There is a specific requirement in medicine to incorporate the clinical knowledge and expertise in the decision support systems. The results should also be clinically plausible. In order to achieve that we modify the Composite Association Rules algorithm of Ye and Keane (1997) to include conjunctions as well as disjunctions. This allows us to explore various combinations possible in medical context. We use a data set to compare these two approaches and then apply both the algorithms on two test data sets to compute the respective accuracies.

In this paper we have focused on two important aspects of ovulation induction method of fertility treatment. In this method of treatment the ovarian follicles are induced with follicle stimulating hormones (FSH). As a result, in the middle of the menstrual cycle, the number of mature follicles obtained are generally higher than the number of follicles normally obtained under natural process. Under natural cycle, one or two follicles will mature to release oocytes but under the treatment, it is not uncommon to find more than ten such follicles. We also find the size of the artificially stimulated follicles are much larger than the ones produced through natural cycles. These are not only the main determining factors in the treatment outcome but also the ones that can be modified by careful choice of the type of hormones, dosage, and the rate by which these hormones are administered. That is why in this paper we focus on these factors as they are more important from clinical point of view. There are several other variables in the data sets that may be worthy of further investigation.

IUI, DI AND OI DATA SETS

The datasets were obtained from the Westmead Fertility Clinic, Westmead Hospital, Australia. IUI (Intra Uterine Insemination), DI (Donor Insemination) and OI (Ovulation Induction) are different treatment cycles of similar nature. Each record refers to an incident relating to a couple treated with artificial insemination. These datasets consist of 104 attributes, IUI has 1,597 records, DI has 615 records and OI has 762 records. The attributes are about the history of pregnancy of the couples, the present situation of the couple, the normal situation of the couples, the laboratory examination results, and the treatment details. For example, the hormone type used in the treatment, the dose of the hormone used, and the results from the treatment in pregnancy.

We prepared, cleaned and checked data quality in terms of reliability, accuracy, relevance, completeness, consistency, precision, etc. If we find that a datum is absolutely wrong for example someone being reported to have a body temperature of 100 C, then we treat it as missing value. We exclude the attributes which have large number of missing values because they do not have the data to process, for example, one attribute has only two records which have the data. We do not make any attempt to estimate any missing values, simply because we do not have any mechanism for finding either default or average value nor can we use any interpolation methods in this domain.

CLINICALLY MODIFIED INFORMATION FUZZY NETWORK (CMIFN)

The main idea of this algorithm is using the information sharing between the candidate input attribute and the target to create the network. The information sharing is represented by the conditional entropy. There are two main processes in building such a network. First of all, the selection of layer which is based on conditional mutual information and clinical significance and relevance with the target attribute. Secondly every node of the

layer formed by the discrete values is split if the likelihood ratio is greater than a minimum significant value. The main steps of the algorithm are shown below.

Algorithm

Step 1 Define the minimum significant level of likelihood ratio and clinical significance of each attribute *i*' and target attribute *i*, $C_s(i',i)$,

Step 2 Calculate unconditional probability (apriori) of each value of the target attribute by

$P(V_{ij}) = O_{ij} / n$

where

O_{ij} – number of occurrences of the value j of a target attribute *i* in the relation

n – number of complete tuples in the relation

Step 2.1 Calculate unconditional entropy of the target attribute by

$$H(A_i) = - \sum P(V_{ij}) * \log P(V_{ij})$$

Step 2.2 Select the first attribute to be the first hidden layer in the network by using clinical significance.

Step 2.3 Calculate the conditional mutual information of the candidate input attribute and the target attribute, given the node, by

$$MI\ (A_i\ ;\ A_{i'}\ |\ z\ ) = \sum_{j=0}^{Mi-1} \sum_{j'=0}^{Mi'-1} P(V_{ij}\ ;\ V_{i'j'}\ ;\ z) * \log\ (P(V^{ij}_{\ i'j'} |\ z\ ) / (P(V_{ij}\ |\ z) * P(V_{i'j'}\ |\ z)))$$

where

$P(V_{ij};\ V_{i'j'}\ ;\ z)$ - joint probability of a value *j* of the target attribute *i* , a value *j'* of the candidate input attribute *i'* and the node *z*

$P(V^{\ ij}_{\ i'j'}\ |\ z)$ -conditional probability of a value *j'* of the candidate input attribute *i'* and a value *j* of the target attribute *i* given node *z*

$P(V_{i'j'}\ |\ z)$ - conditional probability of a value *j'* of the candidate input attribute *i'* given node *z*

$P(V_{ij}\ |\ z)$ - conditional probability of a value *j* of the target attribute *i* given node *z*

Step 2.4 Calculate the likelihood ratio statistic of the candidate input attribute and the target attribute, given the node *z* by

$$G^2(A_{i'};\ A_i\ |\ z) = 2 * (ln2) * E(z) * MI\ (A_{i'};\ A_i\ |\ z)$$

where $E(z)$ is the number of tuples in the node *z*

Step 2.5 – If the likelihood ratio is significant and the clinical significance is greater than $C_s(i',i)$, mark the node as "split" and increment the conditional mutual information of the candidate input attribute and the target attribute, given the final hidden layer of nodes; else mark the node as "unsplit"

Step 2.6 – Find the candidate input attribute maximizing the conditional mutual information ("the best candidate attribute")

Step 2.7 – If the maximum conditional mutual information is greater than zero:

-Make the best candidate attribute an input attribute

Step 2.8 – Calculate the connection weights linking the unsplit nodes and the nodes of the final layer to the target nodes by

$$w_z^{\ ij} = P(V_{ij}\ ;\ z) * \log\ (\ P(V_{ij}\ |\ z\ ) / P(V_{i'j'})\ )$$

Step 3 – Select a value *j* maximizing the conditional probability of the target attribute *i* at the node *z* ($P(V_{ij}\ |\ z\ )$) and make it the predicted value of the target attribute *i* at the node *z*

For example the following diagram shows a single layer CMIFN.

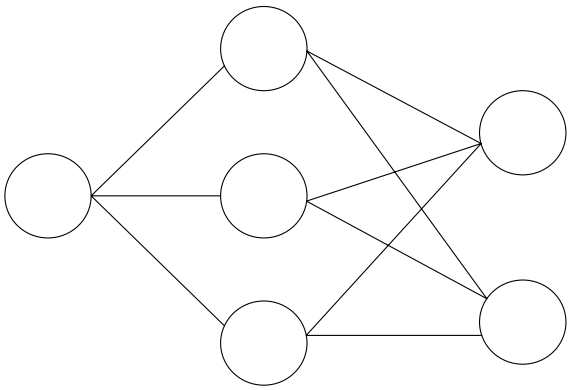

Figure 1: CMIFN for number of follicles of size 10-15mm and pregnancy outcome

MIXED COMPOSITE ASSOCIATION RULES ALGORITHM (MCAR)

The association rules algorithm were initially proposed by Agrawal et. al. (1993) and later modified by Srikant and Agrawal (1995). Since then a variety of techniques have been developed. Maimon and Last (20000) presented an algorithm to generate disjunction composite association rules. Here we propose a variation to that algorithm that is capable of generating mixed association rules, that is, both conjunction and disjunction if node and a single target node. This algorithm is purposely designed to suit the needs of the analysis of clinical data where human expert knowledge plays a significant role. It takes the user input to determine if every new item added to the composite list is included as conjunction or disjunction with the rest of the atomic items. Major steps of this algorithm in addition to the usual association rules algorithm are shown below.

Let I $=\{a_1, a_2, \ldots a_k\}$ be the set of k items excluding the target item a_t and Ω an empty set of "if" node then we proceed as follows.

Algorithm

Step 1. Select minimum support m.
Step 2. Select minimum confidence c.
Step 3. Repeat the followings for all $j \in 1$-k.
Step 4. Select $a_j \in$ I
Step 5. x=1
Step 6. If x=1 then calculate the simple association rule $a_j \rightarrow a_t$ provided the support is >= m and confidence >= c
else add a_j to Ω as conjunction or disjunction depending on user preference.
Step 7. If x>1 then calculate the simple association rule $\Omega \rightarrow a_t$ provided the support is >= m and confidence >= c
Step 8. x++
Step 9. j++
Step 10. Compile all the rules.

EXPERIMENT AND COMPARISON OF CMIFN AND MCAR ALGORITHMS

We applied the algorithms on the IUI data as experimental data and then used the OI and DI data sets as test data. The clinical question we were exploring is the correlation between the number and size of follicles, and the final outcome, that is, the pregnancy or no pregnancy. It is an important question as much effort goes into achieving a certain number and size of the follicles.

The MCAR algorithm recursively generates a large number of rules from simple to composite as they are applied on discrete data values. The results are "fuzzified" according to the desired range that gives meaningful support and confidence. It is a recursive algorithm that tests various combinations.

First of all we compare the results obtained by single-layered CMIFN and simple MCAR algorithms. Some of the examples are as follows.

Follicle Size 10 – 15 mm

CMIFN Single-layered

If #follicle size 10-15 = 0-4
then pregnant = yes
weight 0.018893254

If #follicle size 10-15 = 5-9
then pregnant = yes
weight 0.01355831

If #follicle size 10-15 = 10-13
then pregnant = yes
weight 0.00705802

The above results indicate that as the number of follicles of sizes between 10 and 15 mm increases the likelihood of a successful pregnancy decreases. This result is clinically significant as we try to induce the follicles using hormones.

Simple MCAR

If #follicles of size 10-15mm = 0-4
then pregnant = yes
Support **28%**, Confidence **52%**

If #follicles of size 10-15mm = 5-9
then pregnant = no
Support **4%**, Confidence **41%**

If #follicles of size 10-15mm = 10-15
then pregnant = yes
Support **0.7%**, Confidence **33%**

These results are consistent with the results obtained from CMIFN algorithm.

Follicle Size > 15mm

CMIFN Single-layered

If # follicle size larger than 15 = 0-1
then pregnant = no
Weight 0.00874

If # follicle size larger than 15 = 2-3
then pregnant = yes
Weight 0.030031893

If # follicle size larger than 15 = 4-5
then pregnant = no
Weight 0.02730611

The above results indicate that the larger the size of the follicles the lesser the chances of successful pregnancies. Again it is clinically very significant as we determine how much and how long the hormone therapy be applied to achieve an optimal result.

Simple MCAR

If #follicles of size >15mm >3
then pregnant = no

Support **10%**, Confidence **100%**

This result is a very strong indication of the results obtained using CMIFN algorithm.

Next we compare the results of multi-layered CMIFN and complex MCAR algorithms as follows.

Complex IF node

CMIFN Multi-layered

If #follicles of size 10-15mm <=2 ***and***
#follicles of size >15mm <=2
then pregnant = yes
Weight 0.030241994

previously successful pregnancy *and*
Gonadotrophin F *and*
total number of follicles <=3
then pregnant = no
Weight -0.026050873

Complex MCAR

If #follicles of size 10-15mm <=2 ***and***
#follicles of size >15mm <=2
then pregnant = yes
Support **12%**, Confidence **100%**

If (#follicles of size 10-15mm >2 ***or***
#follicles of size >15mm>2) *and*
Clomiphene
then pregnant = no
Support **9%**, Confidence **100%**

The above example demonstrates the power and diversity of both algorithms. The results obtained by MCAR algorithm gives us very strong indications for future fertility treatments. We used the OI and DI data sets to test these algorithms and found that the accuracy for CMIFN when applied to OI data was 78% and for DI data was 86% whereas the accuracy of the MCAR algorithm for OI data was found to be 88% and for DI data 92%.

CONCLUSION AND FUTURE WORK

We have demonstrated that the association rules algorithm we present here is promising in two ways. Firstly it incorporates both the disjunction and conjunction as compared to IFN algorithm. Secondly, the results show that the association rules algorithm is marginally better.

We see that both the algorithms are particularly suitable for exploring medical data in search of answers for complex clinical questions. The CMIFN has a unique ability of determining negative associations as well as positive or nil association but is limited to conjunctions only.

The data sets we have used in our experiments are relatively small but the results are very promising. Most of the results obtained thus far are in line with pathophysiological explanations. These results can potentially improve the outcome and categorise patients into groups most suitable for particular treatment cycles. In future we intend to test these algorithms on large data sets collected from a cross section of Australasian fertility clinics.

REFERENCES

Agrawal, R., Imielinski, T. and Swami, A. (1993) "Mining association rules between sets of items in large databases.", *Proceedings of ACM SIGMOD Conference*, 207-216.

Azuaje, F.,Dubitzky, W., Black, N. and Adamson, K. Improving clinical decision support through case-based data fusion, *IEEE Transactions on Biomedical Engineering*, 46(10), 1181 – 1185.

Berman, J. J. (2002) Confidentiality issues for medical data miners, *Artificial Intelligence in Medicine*, 26, 25-36.Breault, J. L., Goodall, C. R. and Fos, P. J. (2002) Data mining a diabetic data warehouse, *Artificial Intelligence in Medicine*, 26, p. 37-54.

Cochran, W. G. (1977) *Sampling Techniques*, Wiley, New York.

Cochran, W. G. and Cox, G. M. (1992) *Experimental Designs*, Wiley, New York.

Jurisica, I., et al., (1998) Case-based reasoning in IVF: prediction and knowledge mining. *Artificial Intelligence in Medicine*, 12(1), 1-2.

Lavrac, N. (1999) Selected techniques for data mining in medicine, *Artificial Intelligence in Medicine*, 16(1), 3-23.

Lenic, M., Povalej, P.,Zorman, M. and Kokol, P. "Multiple opinions for medical decision support", *Proceedings of 17th IEEE Symposium on Computer Based Systems*, Bethesda, MD, USA, 230-235.

Maimon, O. and Last, M. (2000) *Knowledge Discovery and Data Mining, the Info-Fuzzy Network (IFN) methodology*, Kluwer, Norwell, MA.

Pechenizkiy, M., Tsymbal, A. and Puuronen, S. **"**PCA-based feature transformation for classification: issues in medical diagnostics" in *Proceedings of 17th IEEE Symposium on Computer Based Medical Systems*, Bethesda, MD, USA, 535-540.

Snedecor, G. W. and Cochran, W. G. (1989) *Statistical Methods*, Ames: Iowa State University Press, USA.

Srikant, R. and Agrawal, R. (1995) "Mining generalised association rules.", *Proceedings of 21st Int. Conference on VLDB.*

Tigrani, V. and John, G. H. (2005) Data Mining And Statistics In Medicine: An Application In Prostate Cancer Detection, URL. http:// citeseer.nj.nec.com.

Ye X. and Keane, J. A. (1997) "Mining Composite Items in Association Rules", *Proceedings of IEEE International conference on Systems, Man and Cybernetics*, 1367-1372.

COPYRIGHT

The following copyright statement with appropriate authors' names must be included at the end of the paper