Pre-processing and Induction Methods for IDA of Small and Incomplete Medical Data for Cystic fibrosis Patients

Michaela Black\(^a\), Le Roy Dowey\(^b\), Ray Hickey\(^a\), David Thurnham\(^b\), Hugh McGlynn\(^b\), Lawrence McGrath\(^c\), Lorraine Martin\(^d\), Stuart Elborn\(^d\)

\(^a\) Faculty of Informatics, University of Ulster at Coleraine, Cromore Road, Co. Londonderry, N Ireland.
\(^b\) NICHE, School of Biomedical Sciences, University of Ulster, Cromore Road, Coleraine, BT52 1SA.
\(^c\) Department of Therapeutics and Pharmacology, Whitla Medical Building, Queens University, Belfast, BT9 7BL.
\(^d\) Adult Cystic Fibrosis Centre, Belfast City Hospital, Lisburn Road, Belfast, BT9 7AB.

Abstract

Intelligent Data Analysis (IDA) has many benefits to offer the analysis of medical data. Much hidden information exists within real world data. This represents concepts or hypotheses not initially expected or perceived by the domain expert. However a lot of the medical data available is not always of the standard expected for IDA. Many of these issues relate to the quality and scale of the data. The data set that was used in this analysis was obtained from patients attending the Adult Cystic Fibrosis Unit of the Belfast City Hospital. In this work we analyse such issues in relation to the small patient data set and present some initial pre-processing, a follow-up tree induction technique and the outcomes. Initial findings of this work highlights that the key issues affecting clinical outcome in CF patients are even more important in a subgroup infected by \textit{Burkholderia cepacia} (BC). Due to the effect that BC has on the CF community world wide, it is vitally important that we utilise all tools available.

Introduction

This work is based on data from the Bioinformatics Centre at University of Ulster Coleraine. The data records relate to adult cystic fibrosis (CF) sufferers in Northern Ireland. Cystic fibrosis (CF) occurs in the European Caucasian population at a rate of approximately 1 in 2000. These patients suffer from chronic infection. This chronic disease state usually leads to decreased lung function, poor nutritional status, elevated immune function and raised oxidative stress.

The type of bacterium that the host is infected with seems to affect morbidity and mortality. Infection of CF patients by \textit{Burkholderia cepacia} (BC) has been shown to lead to clinical decline. The occurrence of BC has increased greatly since it was first isolated in CF patients. Approximately 30% of adult CF patients in Northern Ireland are infected.

In general the fat soluble micronutrient status of patients with cystic fibrosis is significantly lower than normal subjects and carotenoids may play a protective role in immune function [Winklhofer-Roob et al 1995]. Their airway surface fluid is of high viscosity thus giving better conditions for bacterial infection e.g.: \textit{Ps aerguinosa} & \textit{B. cepacia}.

Pulmonary disease is responsible for over 90% of death in CF patients. CF patients with \textit{B. cepacia} have an increased mortality rate compared to other bacterial infections [Muhdi et al 1996]. Approximately 30% of adult CF patients in Northern Ireland are infected with \textit{B. cepacia}. Caloric intake has been shown to extend the life span of CF patients [Corey et al 1988].

The aim of this study was to examine markers of three specific areas: oxidative stress, nutrition and disease progression in patients with and without \textit{B. cepacia} infection. Fourteen CF patients with \textit{B. cepacia} (range 16-30 yr, mean 24.9) and twenty-six non-\textit{B. cepacia} CF patients (range 16-38 yr, mean 26) who were attending the Belfast City Hospital Adult Cystic Fibrosis Unit were enrolled for the study. The three specific areas of interest were measured by the attributes shown below.

Health/Nutritional status: \textit{BMI}; Plasma fat-soluble vitamins (FSV); micronutrients; Total protein; cholesterol levels.

Disease progression and severity: plasma \textit{α-1-anti-chymotrypsin} (ACT); neutrophil-elastase-\textit{α-1-anti-trypsin} complex (NEAAT); urinary neopterin: creatinine ratio (N:C); markers of lung function (FEV\textsubscript{1} and FVC); hospital severity scaling.

Oxidative stress: plasma levels of malondialdehyde (MDA); protein thiols (S-H); protein carbonyls (C=O).
1. Data Issues

Scales of data available for IDA can range a very wide spectrum. When looking at automated collection of data we see tetra-bytes of data as in protein sequences. At the other end of the spectrum we have small groups of patients with a disease where the data is collected by hand. For this work we started with 40 records in total based on 40 patients. Some of this data may also be the result of collecting some samples from the patients and carrying out tests on them as opposed to a direct reading. These issues give rise to a number of problems for IDA.

The main issue lies with the scale of the data. This small number of records may prohibit the learning process from attaining a complete and stable knowledge structure. We aim to learn some hidden knowledge within the data that does not appear obvious to the domain expert. To achieve stability in this learning process we would require enough data for the information measurements to stabilise. This initial small data set may prove to be a challenge.

The next issue is related to the number of attributes. Initially there are 32 attributes in the data set. Too many attributes can confuse the learner. This can be catastrophic to the IDA process. During the induction process attributes are compared and tested, for their information gain. Many attributes in this selection process can generate confusion and in some cases inhibit the performance of the algorithm. Within the induction process attributes are selected for their informativeness based on calculations like information gain ratio [Quinlan 1993]. With the inclusion of too many attributes this becomes a more difficult task. With all of the 32 attributes competing for inclusion, the calculations can be skewed. The main aim would be to reduce as many attributes as possible with the aid of the experts with the domain knowledge. However in this case the main aim was to identify if a correlation existed between the beta-carotene in the patients body and the level of their disease, but it was not restricted to this. Any other hidden knowledge within the data would prove beneficial.

This leads us onto the final issue of missing values. For various reasons certain readings or tests may not be available for some attributes. The quality analysis of the original data showed some fields with a high percentage of missing values. For these fields their inclusion in the process becomes somewhat questionable. Again the significance of interest in these fields will determine what corrective issue is taken. Techniques such as probability and frequency based calculations can be used in an aim to fill the missing values [Quinlan 1991]. The technique used here was to try and reduce the number of fields that had a high missing ratio. The justification for this was that we were not willing to compromise on the accuracy, as this was medical patient data analysis. We were able to reduce the number of fields from 32 to 16.

Initially removing attributes was an iterative process beginning by targeting those attributes with the highest missing value ratio. Iteratively each attribute was taken in turn and removed, with the affect of this on the learning process observed. If the exclusion of the attribute had no effect it was removed permanently. Within this data set this iterative process proved successful with the extraction of all attribute with missing values having no effect on the overall learning process. This reduced the data set from 32 attributes to 19. There was of course one exception to this and that was the beta-carotene attribute. Due to the high interest in this attribute it was retained. There only existed 2 records with this value missing and these were removed, leaving 38 records in total out of initially 40. Again probabilistic methods could have been used to fill these, [Horn et al 1997] if more data was available. Other methods to reduce the number of attributes could be applied which measure the relevance of the attributes [Wang 1997].

2. Data Set

Following the pre-processing stage we initiated the induction phase with the fields specified below. Additional fields also used are the determining class, which reflects if the patient has or has not cepacia, and additional attributes for the sex and age of the patient.

Class: C/non-C: 1= cepacia, 2= non-cepacia
Sex: 1= male, 2= female
Age: continuous

<table>
<thead>
<tr>
<th>1 Nutrition</th>
<th>2 Morbidity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retinol</td>
<td>Severity Scale</td>
</tr>
<tr>
<td>Lutein</td>
<td>Av. ACT</td>
</tr>
<tr>
<td>γ−Tocopherol</td>
<td>NE/AAT</td>
</tr>
<tr>
<td>α−Tocopherol</td>
<td></td>
</tr>
<tr>
<td>β−Carotene</td>
<td></td>
</tr>
<tr>
<td>Total protein (Tot ptm)</td>
<td>3 Oxidative stress</td>
</tr>
<tr>
<td>Total cholesterol (Tchol)</td>
<td>Protein Thios</td>
</tr>
<tr>
<td>γ−Tocopherol: cholesterol</td>
<td></td>
</tr>
<tr>
<td>α−Tocopherol: cholesterol</td>
<td></td>
</tr>
</tbody>
</table>

3. Induction Algorithms for IDA

The initial objective was to induce any correlation between the patient attributes available and the disease level of the patient i.e. cepacia or non-cepacia. As there was no specific initial hypothesis identified this was a clear induction task. The aim was to induce some information from within the attributes in determining the disease level. The inductive learning process will search among possible hypotheses that sufficiently well explain the training instances presented. Induction of decision trees is one of the widely used approaches and presents the induced classifier as a tree representation [Quinlan 1986].

With the pre-processing stage we reduced the attributes from 32 to 16 with the class being disease level with two values. Initial preparation of the data explained above was carried out using SPSS system called 'Clementine'.
On completion of the pre-processing stage we have now got our training data comprised of 38 individual instances. This will be presented to a tree induction algorithm to produce a classification system for future unseen cases [Quinlan 1986]. The induction algorithm C5 was used [http://www.rulequest.com/]. This is an induction system which uses trees to induce knowledge also offers the facility to convert this induced tree structure into a set of rules.

We used the expert settings within the C5 package. These allow variations to the level of pruning to be applied to the tree. Our aim was to produce generalised knowledge that was not just specific to the training data. Pruning is a mechanism to prevent the knowledge induced over-fitting the training data [Niblett 1986]. We were also able to specify the minimum number of instances we wanted to be covered by each leaf of the tree.

4. Results and discussion

We have applied pre-processing techniques and an induction algorithm to this small data set and have successfully found strong correlations between the attributes and the disease level class.

We reduced the number of the attributes from 32 to 16 without hindering the learner task. We also removed some records with missing data to aid the inclusion of the beta-carotene attribute. The induction algorithm produced the following classifier.

![Induced classifier with >= 3 instances per leaf](image)

Figure 1: Induced classifier with >= 3 instances per leaf

The discrete level for the total protein in the patients body has been selected as 82.46. All patients above this threshold are at disease level 1. It further refines those patients whose reading is equal or below this threshold. Three important attributes have now also been discretised by the induction algorithm. Total cholesterol has been split at a threshold level of 2.71. Any patients less than or equal to this are at disease level 1.

A rule is deduced from the tree by reading from the root to a leaf and the class at that leaf is that of the rule. The first rule: $\text{Tot ptn} = \leq 82.46$ and $\text{Tchol} = \leq 2.71 \rightarrow 1$ can be read as, ‘If a patient’s total protein is less than or equal to 82.46 and the total cholesterol is less than or equal to 2.71 then this patient is at disease level 1’.

An attribute relevance can be read from its level within the tree. The root is referred to as position zero and is selected because it is the most informative in determining the class.

The numbers in brackets at each leaf of the tree represent the number of examples in the training data covered by that rule and a confidence coefficient for that rule. All but one of the rules has a coefficient of 1. The exception to this is 0.923, which is still at an acceptable level.

If we increase the number of required instances per leaf, i.e. the number of instances to support the rules we see a slight change in the induced classifier. Figure 3 shows two attributes have swapped levels within the tree i.e. Age and Av. ACT. The discrete level selected for Age remains the same, however that selected for Av.ACT changes.

![Induced classifier with >= 4 instances per leaf](image)

Figure 3: Induced classifier with >= 4 instances per leaf

Note the confidence for two rules in figure 3 is reduced below any from the classifier in figure 2. However this just re-emphasises that with too little data, stability is difficult to achieve. It must be noted that the same attributes were selected for both classifiers and the two top attributes remain the same, as does the value for their binary split.

The use of IDA in relation to medical data has been applied successfully during these analyses. We can suggest from
these results that the measurement of total protein is not acting as an indicator of nutrition but rather as an indicator of disease state. Total protein would increase in patients with an infectious load owing to raised acute phase proteins including ACT.

Our results for total cholesterol are in keeping with the findings of Corey (1988) whose landmark paper highlighted the importance of dietary fat intake in the survival of CF patients. Without an adequate caloric intake it is probably more difficult for CF patients to recover from, or reduce the occurrence of, acute exacerbations. The intake of dietary fats is also important for absorption of FSVs, which are known to have antioxidant properties. Depressed antioxidant status contributes to increased oxidative stress in CF patients [Lepage et al 1996]. This data analysis technique has also identified ACT and age as important factors in both the BC and non-BC subgroups of CF patients. Unfortunately at this stage of the work the instability in the findings due to the small data set delays patients. Unfortunately at this stage of the work the instability in the findings due to the small data set delays drawing any further conclusions about these fields.

5. Boosting

We also applied this data to the boosting mechanism implemented with See5 [Quinlan, 1996]. The boosting mechanism maintains a weight for each instance. The number relates directly to the instance influences on the classifiers learned. There are x number of trials iterated and x classifiers generated. For this experiment the default of 10 trials was selected. The final classifier aggregates the learned classifiers by a voting mechanism. This vote is based on the individual classifiers accuracy. This final classifier can not be used to give us a visual understanding of the correlation between the attributes and the class. However 100% accuracy was achieved on this training data. More data will allow us to test the stability of this accuracy.

6. Future Work

This work is an initial step with this data. The follow up work will be aimed at carrying out similar experimentation with more data. Our initial aim was to highlight strong correlation within the data between the attributes and a disease class. Initially it was hoped that beta-carotene would result as a strong factor. Unfortunately this is not the case. However strong hidden correlations were identified. The lack of data starves the induction learner and inhibits it from completing its task. Thus it can result in this instability and affects the decision of which tree to select. The tree has not been allowed to fully expanded. More data will benefit this problem. It will also allow us to test the accuracy of our classifier against that of the final classifier from the boosting trials to reinforce the accuracy. Other techniques like cross validation could also be tried.

The development of a larger database using clinical data from CF centres in UK and Ireland should reinforce the findings of this analysis. Use of larger data sets may also help to elucidate the relative importance of ACT and age for BC and non-BC subgroups of CF patients. The importance of other nutritional components such as FSVs may also be evaluated using this IDA method.

Recent analysis of the data has shown that certain fields that were blank, are actually so because of a low reading that is not interpretable by the machine. In such cases the machine substitutes the value with a ‘ND’ value, which was then entered into the database as a blank value. Research into the background of the machine and these low readings will allow these values to be updated and then included into the induction process.

We have been looking directly for automated generation of hypotheses that link available attributes to the cepacia infection. Future analysis will also adapt this to look for hypotheses that relate the cepacia infection with other attributes in determining new classes such lung function.

References


