Differential Diagnosis of Neonatal Disease: A Data Mining Model

9.1. Introduction

Data mining basically refers to information elicitation from data warehouse. Since its birth in the year 1993, data mining techniques are being deployed in various disciplines including medical domain. Data mining and knowledge discovery techniques are being used after its birth in the year 1993 [1] in connection with a business. It is now equally being applied to interpret huge clinical data base(s) all over the world in order to provide data for applications such as automated encoding, decision support, quality assurance, patient management, outcome analysis, and clinical research [2][3][4]. There are different algorithms available for classification problems: decision tree, naïve nays, support vector machine, and feed forward neural networks. Decision tree approach has been found suitable for this purpose since decision tree construction can make use of both symbolic or nominal and real-valued attributes [5] – a characteristic of medical domain.

This chapter intends to report some results of differential diagnosis of neonatal disease applying two important data mining algorithms, namely, ID3 and C4.5. We use data base from our earlier study. It is also planned to find relative importance of different disease parameters for differential diagnosis. For the purpose, we use rough set theory having the provision of reduct and core. Results are compared for ID3 and C4.5 without/with reduct and core. The relevant issues related to model development and validation such as algorithm settings, overfiting, confusion matrix etc., have been discussed here.

Out of the different phases of the development of a child, neonatal phase is considered to be a critical one. They are also vulnerable or special risk group; the risk is related with growth, development, disease and survival. In rural and remote areas of the world, the mortality and morbidity are still significant in number. One of the prime reasons of such high mortality and morbidity is prevalent diseases. There are a number of neonatal diseases and a number of parameters involved; a multi-criteria decision making system (MCDMS). For a typical MCDMS, it is useful to take advantage of

†† This chapter is based on the publication made by the author entitled “A Data Mining Model for Differential Diagnosis of Neonatal Disease”, IFRSA’s International Journal of Computing, Vol. 1, Issue 2, pp 143-150, April 2011.
any automated system for rational decision making. Moreover, the relevant data bases created by experienced medical practitioners and researchers should be mined for information elicitation. Moreover, for a domain which mostly vulnerable, like pediatric domain, where uncertainties are involved [6], we are to use a suitable theory for managing such uncertainties. In the recent years, rough set theory of soft computing paradigm has found its applications in various disciplines including medical domain. At the same time, data mining techniques are being deployed to the data bases for useful information elicitation.

In the research period we have experienced few things related to neonatal disease. These are various signs and symptoms of diseases, medical problems, working knowledge with several doctors and medical test reports of several neonates having problems or suffering from diseases. The specialist or the doctor will use all that evidence to arrive at diagnosis for the baby and also informs what is wrong with the neonate. It was noticed that most of the time it works, but other times it may be misdiagnosed or even doctor may fail the diagnosis. In the study area, this misjudging and misdiagnosed frequency rate is very high. Thus there is a need to know how doctors diagnose a child disease and what one can do to confirm before he or she has arrived at the right answer. Doctors basically use information drawn from this description of sign symptoms, medical tests report, knowledge of medicine, experience and additional inputs. Domain specialist or doctors will then make a list of all the possible diagnoses that could explain what is medically wrong with patient. Then, one by one, using that same information, he will begin to narrow down the list by finding clues that don't fit. This process of elimination is called "differential diagnosis." Finally doctor will be left with one diagnosis, and that's the one he gives the patients.

If these procedures use data mining modeling, then there is a high chance of improvements for the disease diagnosing procedure. This has been statistically proved on the study. Medical domain involves a number of inexactnesses. The nature of inexactness can be vagueness, uncertain, partial truth, imprecision. The sources of inexactness can be logical as well as physical type [6]. The typical logical sources are: lack of adequate data, inconsistency of data, inherent human fuzzy concepts, matching of similar rather than identical situations, differing (expert) opinions, imprecision in measurements, lack of available theory to describe a particular situation. Concentrating on the pediatric domain, the typical physical sources are: problem domain itself, child, parents/guardians, doctors, laboratory tests/technicians, symptoms, non-availability of laboratory results. That’s why dealing with the pediatric problems is very sensitive during rational decision making. Moreover, over the years thousands data have been accumulated in different disciplines on medical domain and the process of accumulation is continuing all over the world. For knowledge discovery from such huge data bases, data mining tools are being deployed. On the other hand, soft computing techniques are being deployed which are capable to deal with imprecision and uncertainty especially needed in ill-defined problem areas. Data
Data Mining Model for Differential Diagnosis

mining basically refers to information elicitation and soft computing is meant for extracting intelligent behavior. Data mining offers exact output within the error bounds estimated. On the other hand, soft computing offers approximate output but treated as intelligent one. So, it is tempting to use the merits of both the paradigms synergistically in a complementary way leading to knowledge discovery in databases.

This work in the chapter reports the results of a study where some coupling with data mining and rough set theory is proposed for differential neonatal disease diagnosis. The experimental set up was planned to extract the relative importance of neonatal disease attributes; dimensionality reduction; finding core attributes, if any; classification prediction using data mining approach.

The chapter summary is as follows:

Section 9.2 intends to discuss data mining as well as soft computing perspectives with special reference to ID3 and C4.5, and rough set theory. Section 9.3 presents the problem description and experimental setup. Section 9.4 presents the results of the study. In the last section, i.e. section 9.5 presents the discussions and conclusion of the study.

9.2. Data Mining and Soft Computing Paradigms

9.2.1. Data Mining: Decision Tree

The decision tree algorithm is a common and one of the most popular algorithms used in data mining because it is easy to understand how it makes predictions. The goal is to create a model that predicts the value of a target variable (dependent) based on several input variables (independent). A tree can be ‘learned’ by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions [7]. Data comes in records of the form:

\[(X, Y) = (x_1, x_2, x_3, \ldots, x_n, Y)\]  

The dependent variable, Y, is the target variable that we are trying to understand, classify or generalize. The vector X is composed of the input variables \(x_1, x_2, x_3, \ldots, x_n\), that are used as input for the model. Through discussion regarding this is elaborated in chapter 8.

In data mining, there are different categories of trees available in the literature. We use classification tree analysis, as because our predicted outcome is the class to which the data belongs.
9.2.2. **ID3 Algorithm** [8]

ID3 (Interactive Dichotomizer 3) is an algorithm used to generate a decision tree invented by Ross Quinlan. ID3 can be thought of as an inductive inference procedure for machine learning or rule acquisition. At any point we examine the feature that provides the greatest gain in information or the greatest decrease in entropy. Entropy is defined as $-p \log_2 p$, where $p$ is the probability which is determined on the basis of frequency of occurrence. The ID3 algorithm can briefly be stated as follows:

- Consider all unused attributes and count their entropy;
- Choose that attribute for which entropy is minimum, or information gain is maximum to serve as the root node of the decision tree;
- Build the next level of the decision tree providing the greatest information gain;
- Repeat step 1 through step 3. Continue the procedure until all subpopulations are of a single class and the system entropy is zero.

In chapter 8 we have broadly discussed about the ID3 algorithms with its all implementation details.

9.2.3. **C4.5 Algorithm** [9]

The C4.5 algorithm is based on information gain which is again based on the concept of entropy of information theory. For a random variable $X$ with $N$ outcomes \{ $x_i : i = 1,2,3,4,\ldots,N$ \}, the Shannon entropy, a measure of uncertainty and denoted by $H(X)$, is defined as:

$$ H(X) = - \sum_{i=1}^{N} P(x_i) \log_2 P(x_i) \quad \text{(2)} $$

where $P(x_i)$ is the probability mass function of outcome $x_i$. This algorithm C4.5 uses a set of training data $S ( S = s_1, s_2, s_3, \ldots, s_n )$ of classified samples for building decision trees. Each sample $s_i$ will be having a set of attributes or features for classification along with a class attribute. Based on the normalized information gain (difference in entropy), at each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The attribute with the highest normalized information gain is chosen to make the decision. It then uses recursion on the smaller sub lists for further building and completion of decision tree. The details of C4.5 tree has been discussed on chapter 7.
9.2.4. Soft Computing and Rough Set Theory

Soft computing is a consortium of methodologies which works synergistically and provides, in one form or another, flexible information processing capability for handling real life ambiguous situations [10]. The aim of soft computing is to exploit the tolerance for uncertainty, imprecision, partial truth and approximate reasoning in order to achieve robustness, tractability, low cost solution, and close resemblance with human like decision making. Soft computing paradigm includes Fuzzy Sets (FS), Rough Sets (RS), Artificial Neural Networks (ANN), Genetic Algorithms (GSs), Genetic Programming (GP), Support Vector Machines (SVM), Swarm Optimization (SO), Ant Colony Optimization (ACO), Memetic Algorithms (MA) and others.

Rough sets theory was first presented by Pawlak in the 1980’s [11]. Rough set is a formal approximation of a crisp set in terms of a pair of sets which give lower approximation with positive region and upper approximation with negative region. In between there a boundary. Let there be an information system \( I = (U, A) \) (attribute -value system), where \( U \) be the universe of discourse and is a non-empty set of finite objects; \( A \) is a non-empty finite set of attributes. With any \( P \subseteq A \), there is an associated equivalence relation \( \text{IND}(P) \). The relation \( \text{IND}(P) \) is called \( P \)-indiscernibility relation. Let \( X \subseteq U \) be a target set we wish to represent using attribute subset \( P \). Now, the target set \( X \) can be approximated using only the information contained within \( P \) by constructing \( P \)-lower (\( P_X \)) and \( P \)-upper (\( \bar{P}_X \)) approximation of \( X \). The tuple \((P_X, \bar{P}_X)\) is called a rough set. The accuracy of the rough-set representation of the set \( X \) can be given [11] by the following:

\[
\chi_P(x) = \frac{|P_X|}{|\bar{P}_X|}
\]

Rough set theory is an intelligent technique for managing uncertainties that is used for the discovery of data dependencies, to reduce redundancies, to evaluate the importance of attributes, to discover patterns in data, and to classify objects. There are several useful features of rough sets such as (i) extraction of rules from data sets in the form of if-then rules; (ii) it requires no external parameters unlike other intelligent techniques except the data itself; (iii) it can predict whether the data is complete or not. The computation of reduct and core using rough set theory is an important feature. For the purpose of the current study we have taken these rough set details again, which has already been discussed in chapter 8.

9.3. Problem Description and Experimental Setup

9.3.1 Problem Description

There are different pediatric age groups: Neonates (0 – 4 weeks), Infants (4 weeks – 1 year), Toddler (1 year – 3 years), Pre-school (3 years – 5 years), School-going (5 years – 10 years) and Adolescence (10 years – 18 years). Out of these age groups,
neonates belong to a special risk group. The risk is related with the growth, development, disease, and survival. Obviously, mortality and morbidity heavily depends on disease diagnosis and management. We consider thirteen differential disease states (objects). Each object has 14 attributes (sign and symptoms). In addition, each object belongs to a decision class. A summary of the data used[12] in the present study is shown in 9.1.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Values (occurrences)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth_Term_Status</td>
<td>Term (52); Pre_Term (22), Post_Term (21)</td>
</tr>
<tr>
<td>Birth_Weight_Status</td>
<td>Normal(33); LBW(40); VLBW(2); ELBW(20)</td>
</tr>
<tr>
<td>Age_in_Hours&gt;72</td>
<td>Y(48); N(47)</td>
</tr>
<tr>
<td>Lethargy</td>
<td>Y(62); N(37)</td>
</tr>
<tr>
<td>Refusal_to_Suck</td>
<td>Y(43); N(52)</td>
</tr>
<tr>
<td>Poor_Cry</td>
<td>Y(50); N(45)</td>
</tr>
<tr>
<td>Poor_Weight_gain</td>
<td>Y(73); N(22)</td>
</tr>
<tr>
<td>Hypothalmia</td>
<td>Y(63); N(32)</td>
</tr>
<tr>
<td>Sclerema</td>
<td>Y(51); N(44)</td>
</tr>
<tr>
<td>Excessive_Jaundice</td>
<td>Y(53); N(42)</td>
</tr>
<tr>
<td>Bleeding</td>
<td>Y(57); N(38)</td>
</tr>
<tr>
<td>GI_Disorder</td>
<td>Y(77); N(18)</td>
</tr>
<tr>
<td>Seizure</td>
<td>Y(47); N(48)</td>
</tr>
<tr>
<td>Sluggish_Neonatal_Reflex</td>
<td>Y(27); N(68)</td>
</tr>
<tr>
<td>Disease_differential</td>
<td>HIE_III(16), No_Disease(8), MD_Hypocalcemia(9), Septicimia(37), Hypo_Thalimia(2), Hemorrhage(1), Others(5), Jaundice(2), MD_Hypothermia(3), Jaundice_BA(3), MD_Hypoglycimia(3), HIE_II(4), Seizure_Disorder(1)</td>
</tr>
</tbody>
</table>

* The above table also contains the following information:
  Number of Independent Attributes = 14 (INPUT)
  Dependent Attribute = Disease_differential (OUTPUT)
  Number of Instances = 95,
  Missing Values = Nil.

From the above table 9.1, it is to be noted that the amount of data what is available to this project is not equally distributed over the different decision classes.
9.3.2. Experimental Setup

For the experimental setup we have used the following steps:

Step I. Find the reducts and core using rough set theory by different algorithms and compare.

Step II. Explore the attribute dependency.

Step III. Classify the instances with ID3 and C 4.5 using the two test options
(i) on training data and
(ii) by folding method.

Three sets of data were tested: (i) with original 14 attributes;
(ii) with attributes from reducts; and
(iii) with core attributes.

9.4. Results

9.4.1. Reduct and Core

Reduct is a subset of attributes which can, by itself, fully characterize the knowledge in the database. However, for a problem domain, there might be more than one such reduct. If there are more than one reduct, some attributes might be common to all such reducts; those attributes are called core. The core attribute(s) are indispensable for an information system. We had used initially three algorithms [13][14] namely, Exhaustive algorithm, Genetic algorithm, and Dynamic reducts for finding reduct and core with the help of RSES 2.2 [15], a software tool that provides the means for analysis of tabular data sets with the use of various methods, in particular those based on Rough Set Theory. For a comparative study, we attempted other algorithms [16][17] with the help of ROSETTA[24] software tool. Comparative results are shown in the table 9.2.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Methods / Algorithms</th>
<th>No. of Reducts</th>
<th>Length of Reducts</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSES</td>
<td>Exhaustive</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Genetic</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Dynamic</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>ROSETTA</td>
<td>Genetic</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Johnson (with approx. solutions)</td>
<td>1</td>
<td>10</td>
</tr>
</tbody>
</table>
9.4.2 Attribute Dependency

One of the most important aspects of predictive analysis is the discovery of attribute dependencies. This essentially means that one has to discover which attributes are strongly related to which other attributes. These strong relationships need further investigation, and that will ultimately be of use in predictive modeling. Table 9.3 presents the results of such dependencies. It is observed that the reducts generated by all the algorithms exclude ‘Excessive_Jaundice’, ‘Sclerema’, and ‘GI_Disorder’. Moreover, it is evident from the table 9.3 that all algorithms suggest { X, Bleeding } as core attributes; so we concentrate on { X, Bleeding } for further study.

Table 9.3. Core Attributes.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Methods/Algorithms</th>
<th>Core Attributes (8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSES</td>
<td>Exhaustive</td>
<td>X, Bleeding</td>
</tr>
<tr>
<td></td>
<td>Genetic</td>
<td>X, Bleeding</td>
</tr>
<tr>
<td></td>
<td>Dynamic</td>
<td>X, Bleeding</td>
</tr>
</tbody>
</table>

Let X = { Birth_Term_Status, Birth_Weight_Status, Lathery, Refusal_to_Suck, Poor_Cry, Hypothalmia, Sluggish_neonatal Reflex }

9.4.3 Applying Data Mining Tools

It is to be stated that originally we had 14 attributes (sign and symptoms). After applying rough set theory, we find that all algorithms do not include ‘Excessive_Jaundice’, ‘Sclerema’, and ‘GI_Disorder’ into their reducts. So we consider rest 11 attributes for further study. Further we consider 8 core attributes for performance evaluation. We now apply ID3 and C4.5 algorithms using WEKA (Version 3-6-2)[18], an open source, java-enabled, platform independent data mining software. It has different features and capabilities implementing different algorithms, algorithm settings, evaluating model quality, experimentation for the data miners, integrating data bases of different formats etc. The characteristics of the model are discussed in the next section.
9.4.3.1. Algorithm Settings

Not only the selection of the right algorithm is important but also the proper settings of the parameters from data mining expertise, knowledge of the available algorithms, and often experimentation to determine which algorithm best fits the problem with suitable values of the parameters are equally important. Algorithm settings allow users to exert finer control over the algorithm to attain better results during the build process. Decision tree models can be extremely accurate on the build data if allowed to overfit the build data. This occurs by allowing the algorithm to build deeper trees with rules specific to even individual cases. Hence overfit models give very good accuracy with the build data, but do not generalize well on new data, resulting in decreased predictive accuracy [19]. To avoid overfitting as well as to control tree size, one has to apply pruning techniques and/or stopping criteria for decision tree algorithms. At the same time, goodness of a node split is determined by the information gain. So, in order for our model to generalize well it must not be built around the training data too closely [20]. Different pruning techniques have been proposed along with different splitting criteria, it has been found that there is not much variation in terms of performance [21][22][23]. All these issues have been taken in consideration during the present study.

Now, we present our results in table 9.4 with analysis. It is observed that the predicted accuracy of ID3 is better than that of C4.5 when tested with self data. Moreover, if we reject ‘Excessive_Jaundice’, ‘Sclerema’, and ‘GI_Disorder’, the prediction accuracy does not change substantially. It indicates that these three parameters are not that significant in rational decision making. Now, if the system is operated with 10-folding for testing, the performance of ID3 degrades substantially. C4.5 prediction accuracy shows much better performance in this situation. The results may be interpreted in the way that ID3 works on unpruned tree structure but C4.5 works with unpruned/pruned structure. As we are aware of data mining concepts till now, recapitulate the whole thing we can say, data mining is all about automating the process of searching for patterns in the data. In this respect we, basically interested on which patterns we should intersected for selection for the prediction. Even there was a question, that how it has to be exploited. In our C4.5 algorithm setting we set pruning mechanism so that the model does not lead to overfitting. From the results it is evident that ID3 model tested on self data leads to an overfit model and that leads to substantial degradation on predictive accuracy when tested with 10-folds.

9.4.3.2. Needs of 10 folds Cross Validation

Using 10 folds Cross Validation we have used 90% of full data for training rest and 10% for testing in each fold test. This is a compromise practically motivated by: - 90% is not too far from full 100%, which means that cross-validation produces a fair
estimation of test performance when the training model with 100% is tested against another unseen test set. We have not used any other folding method because, say having 5 folds only means the folds are 80% trained, which can be shown to have great effect on the robustness (train » test) of the produced model. Again having more than 10 say, 20 poses two problems: not only is it computationally more demanding, there is an increasing problem with small datasets, i.e. in a dataset of 100 instances, each test fold in 20-fold Cross Validation would have only 5 instances, and it is then increasingly likely that target classes with few instances are tested and not trained i.e. that all the instances wind up in the test fold producing guaranteed zero classification accuracy or that they all end up in train folds in which case that part of the training resource does not 'pay off'.

Although it must be said that Weka enforcing 'stratified' which mean class-balanced cross validation does not suffer this problem other than with classes having only one instance in the full training data. Though 10 is not a definite number therefore, but dependent on this singleton problem and the total number of training instances itself. If the data set is larger, the fewer folds are needed to produce a best model.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Tested on Self</th>
<th>Tested on 10 Folds</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID3 prediction accuracy (%)</td>
<td>71 71 68</td>
<td>21 22 21</td>
</tr>
<tr>
<td>C4.5 prediction accuracy (%)</td>
<td>66 64 63</td>
<td>48 47 43</td>
</tr>
<tr>
<td>Number of Objects</td>
<td>14 11 8</td>
<td>14 11 8</td>
</tr>
</tbody>
</table>

The above comparative study for predicting the accuracy between J48 and C4.5 algorithm has been shown on Appendix E graphically.

9.5. Conclusion and Discussion

With ever increasing medical knowledge as well as ever increasing population, the physicians are sometimes confused and overloaded during decision making. It is then suggested to consult the specialists to resolve confusion. But, however, this facility of consulting specialists might not always be available or not timely available. This is certainly a major problem of medical domain. Moreover, from pediatric domain, especially neonates come to the physicians with incomplete or ambiguous information.
about patient's history. The situation becomes more acute when they come from lower socio-economic background. These ambiguous and/or incomplete inputs introduce substantial amount of uncertainty in a decision making system.

For managing such uncertainties, a number of methods are deployed such as fuzzy sets, ANNs, rough set theory, which are also dependent on attribute types. We deploy rough set theory for managing uncertainties as well as knowledge extraction. This helps finding reducts and core; thereby reducing the search space. Next, we deploy data mining tools ID3 and C4.5 for classification on the original as well as reduct and core data. This helps understanding the functions and power of this coupled scheme for this problem domain, at least. But, however, this type of coupling scheme may result better performance. Note that this coupling scheme produces somewhat 48% prediction accuracy which seems to be less as expected.

The possible reasons may be as follows:

(i) The amount of data that is available to this project is not equally distributed over the different classes. Some classes are highly populated and some are really less populated with one or two cases. It will certainly be our efforts in future to have a data base with more or less equal distribution over the different classes.

(ii) One may deploy other algorithms. At the same time algorithms settings with varying parameters need to be tested for improved performance.

This study ensures us that with the use of conventional analysis along with data mining and statistical studies in patient data can improve better disease diagnosing capacity with good accuracy rate. Even using data mining techniques data quality and standard of data, diagnosing plans and treatment procedures and decreases of treatment timings must be improved no doubt. On the other hand, the quality of predicting the disease is usually based upon providing the indices. Mostly, measures have to be taken by physicians in the clinics, and it is their acceptance of the results which is required to ensure that they are put into action. Therefore, understandable results in a clear and adequate presentation are essential. At this juncture this particular study definitely used as a helpful tool for the medical practitioner in real time application. Additionally this work is concerning the statistical consolidation of the results primarily, support for the final conversion as well as the development of an adapted automated data mining process.

The present investigation provides a decision support system for differential diagnosis of neonatal diseases as well as a methodology for coupling rough set theory with data mining tools. Lastly, as this is differential diagnosis, the results might be accepted as first order inference. The next higher order performance is achieved with the results of laboratory tests.
Chapter 9

References


