Chapter 5

A Data Mining and Knowledge Unearthing Concept of Designing DSS for Prevalent Neonatal Disease Diagnosis

5.1. Introduction

On chapter 4, the process of designing a decision support system for the neonatal disease diagnosis has been discussed. It is the initial stage of development of such a system which will predict the disease, on the basis of signs and symptoms provided by the patient. Rule based approach using chaining reasoning have been applied. When the diagnosis result was tallied with the domain expert and specialist, we found the optimum accuracy. There the major challenge was to building up the knowledge base. Huge number of clinical data has been processed every day. In this respect, all the data are not relevant and giving actual information. Thus, there is a need of proper management of the data bases and knowledgebase. In this chapter we just initialized the process of data mining on the neonatal database and integrating that with decision support system.

This study is focusing on the neonatal diseases. As we mentioned in the previous chapters that the neonates constitute a large population group, and also vulnerable or special risk group. The risk is related with growth, development, disease pattern and survival. Thus by improving health status of neonates, proper diagnosis of the neonatal diseases is essential. Particularly Health care facilities have at their disposal of vast amount of data. Different analysis of available data on a given problem can lead to more efficient decision-making, which needs extracting relevant knowledge from this data and act upon it in a timely manner. But due to the scarcity of the neonate experts in rural and remote areas increases the mortality rate of neonates [1].

Even sometimes it is very critical to extract and analysis the real data from huge databases. Data mining is one of method to classify and analyze clinical data [2] for this concern. Our study of designing decision support system for prevalent disease diagnosis and management identifies the need of mining the clinical data.

5.1.1. Needs of the Present Study

A few hospitals worldwide are in the advanced stage of developing clinical data repositories for their basic needs of documentation on systems. Very few hospitals are

§ This chapter is based on the publication made by the author entitled “Designing Decision Support System for Prevalent Disease Diagnosis and Management for Neonates: A Data Mining and Knowledge Unearthing Concept”, Proc. Nat. Conf. NCMicroCom-2010, February, 2010.
fully capitalizing on the benefits offered by knowledge base systems; physicians order entry, clinical decision support system, data mining of medical data and rule engines. It has been estimated that the amount of information in the world doubles every twenty months and the size of the number of databases increasing even. Developing new decision support system out of those data is cumbersome if proper data mining concept is not applying properly. Data Mining aims at discovering knowledge out of data and presenting it in a form that is easily comprehensible to humans. Considering our study, majority of the neonates in developing countries are born and cared for in rural homes but the available information is mostly hospital based [3]. Neonates not only constitute a large population group, but also vulnerable or special risk group. The risk is related with growth, development, disease pattern and survival. From the commonly accepted indices, it is evident that mortality rates in this age group are higher than adult population especially in developing countries. Thus by improving health status of neonates, we contribute to the health of the general population. These considerations have led to the formulation of special health services for children all over the world. A study shows that the incidence of the diseases namely, Neonatal Septicemia, HIE, Metabolic Disturbances, Neonatal jaundice etc. are quite high in the North eastern parts of India specially in Tarai region [4]. This has been discussed on chapter 3.

If neonates suffers from these diseases and proper management is not available in proper time the mortality and morbidity will increase considerably. To overcome the problems mentioned above proper treatment planning is required. And for proper treatment planning the basic requirement is Neonatologists. But the pediatric expert distribution in rural region is not at all satisfactory [5]. Not even that, for disease diagnose and treatment planning one has to go through plenty of conditions, sign and symptoms and critical case history. To mitigate the scarcity of the domain expertise, an expert system for prevalent disease diagnosis and management for neonates may help considerably using data mining and data unearthing concepts.

An estimated two-thirds of childhood deaths occur in infancy, and, in turn, two-thirds of infant deaths occur in the first month of life. In addition to these four million neonatal deaths, primarily due to serious bacterial infections, birth asphyxia, and complications of pre-maturity and intrauterine growth restriction, an estimated 3.9 million pregnancies end in stillbirth[6]. These deaths, at least half, could have been avoided if proper interventions were taken up. Research suggests that integration of maternal and neonatal health care services is key to the health benefits for mothers and their newborns. This is particularly so if a given intervention has potential direct health benefits for both mothers and newborns.

5.1.2. Related Studies

Several studies have demonstrated the value of data mining and knowledge discovery techniques have been used to automate the development of rules that detect clinical conditions by interpreting data generated from huge database which have been used to
analyze and structure narrative patient reports in order to provide data for applications, such as automated encoding, decision support, patient management, quality assurance, outcomes analysis, and clinical research [7-10]. A study on Refractive errors in schools going children identifies the need of mining the clinical data shows how knowledge unearthing and data mining process have been used to development of rules and design the knowledge base of clinical data [11]. Expert System technology has been applied in many medical disciplines also for new born babies whom we call neonates [12-17]. To the best of our knowledge, no such comprehensive work has been reported for neonates for such prevalent disease diagnosis and management.

This chapter presents a rule-based, object-oriented expert system for prevalent disease diagnosis and management of neonates using data mining concept. Firstly, the database was designed on the basis of signs and symptoms provided by the patient. Then we applied data mining clustering approach for classification.

The chapter is summarized as below:

In section 5.2, the common causes of neonatal deaths have been pointed out. Section 5.3 explains the knowledge engineering, knowledge accusation and data mining technique applied, Section 5.4 presents the research methodology for system design and implementation. Section 5.5 describes the performance of the expert system. In the last section, i.e. section 5.6 is about the conclusion of this study.

5.2. Common Causes of Neonatal Deaths

In respect to this study we need to know about the problems with the neonates in the study area. Thus again there is need of discussing common causes of neonatal deaths. In India most common causes of neonatal deaths are systematic bacterial sepsis, birth asphyxia, congenital malformation, pre-maturity, Hyperbilirubinemia and others. Few common diseases of neonates are Neonatal Sepsis, Hypothermia, Birth Asphyxia, HIE, Metabolic Disorders, Inborn Errors of Metabolism, Respiratory Disorders, Neonatal Seizure, Hyperbilirubinemia, Prematurely, Hemorrhagic Disease of New Born and others. Few common disease pattern concerning neonatal health problems and diagnosis criteria are tabulated below in Table 5.1. This also to note that single most important determinate is LBW i.e. birth weight less than 2500 gms.

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>Neonatal Health Problems</th>
<th>Diagnostic Criteria</th>
</tr>
</thead>
</table>
| 1.      | Birth asphyxia           | (i) Mild: At 1 minute after birth, no cry, or the breath was absent or slow, weak or gasping.  
<pre><code>      |                          | (ii) Severe: At 5 minutes after birth, the breath was |
</code></pre>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(iii) Indirect: In the absence of direct observations by VHWs about newborn’s condition at 1 and 5 minutes, presence of following two: (a) baby did not cry on its own so the care provider had to make efforts to make the baby cry; (b) color of the umbilical cord was green or yellow.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td><strong>Neonatal Sepsis</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Septicemia, meningitis or pneumonia diagnosed clinically): Simultaneous presence of any two of the following six criteria any time during 0-28 days: (i) Baby which cried well at birth, it’s cry became weak or abnormal, or stopped crying; or baby who earlier sucked or licked well, stopped sucking or mother feels that sucking became weak or reduced; or baby who was earlier conscious and alert, became drawsy or unconscious. (ii) Skin temperature &gt;99°F or &lt;95°F (iii) Sepsis in skin or umbilicus (iv) Diarrhea or persistent vomiting or distension of abdomen (v) Grunt or severe chest indrawing. (vi) Respiratory rate (RR) 60 or more per minute even on counting twice.</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td><strong>Preterm</strong></td>
<td>Less than 8 months and 14 days (37 weeks) of gestation counted from the onset of the last menstrual period as per the history given by the mother.</td>
</tr>
<tr>
<td>4.</td>
<td><strong>Low birth weight</strong></td>
<td>Weight less than 2500 g.</td>
</tr>
<tr>
<td></td>
<td><strong>Delayed breastfeeding</strong></td>
<td>Due to traditional practice, breastfeeding not started in first 24 hours after birth, but baby licked/sucked the sweetened water.</td>
</tr>
<tr>
<td>---</td>
<td>----------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| 6. | **Problems in breastfeeding** | Presence of any one of the following:  
(i) Baby did not suck breast for more than continuous 8 hours even when offered.  
(ii) –Mother unable to breast feed, or –baby fed on extracted breast milk, or goat, or cow milk, or bottle, or --sweetened water beyond 3 days, or –inadequate breast milk evidenced by continuous crying of baby and failure to gain weight. |
| 7. | **Diarrhea** | Watery, liquid motions 3 or more, or > 9 motions of normal consistency in 24 hours; or mucus or blood in liquid stool. |
| 8. | **Hemorrhage** | bleeding from mouth, anus, eyes, nose or in skin or in urine any time or vaginal bleeding after first week. |
| 9. | **Conjunctivitis** | Mother complained of excessive discharge from the eyes of baby and on examination, eyes were red, and purulent discharge or dried pus |
| 10 | **Skin Infection** | 
(i) *Pyoderma*: pus, ulcer, boil, pustule in skin.  
(ii) *Intertrigo*: excoriation with moist, cracked skin at skin folds. |
<p>| 11. | <strong>Abnormal Jaundice</strong> | Skin or eyes yellow on the first day or yellowness persisted at 3 weeks, or when yellowness associated with sepsis. |
| 12. | <strong>Meconium Aspiration:</strong> | History of difficult delivery or presence of birth asphyxia and respiratory distress (RR 60 or more; or severe indrawing of lower chest) started in first 24 hours after birth. |
| 13. | <strong>Hyaline Membrane</strong> | Respiratory distress started within 6 hours after |</p>
<table>
<thead>
<tr>
<th>Disease (HMD)</th>
<th>birth in preterms baby.</th>
</tr>
</thead>
<tbody>
<tr>
<td>14. <strong>Pneumonia</strong></td>
<td>RR 60 or more, persistent even when counted twice (Increased RR when associated with other signs symptoms of sepsis was included in neonatal sepsis).</td>
</tr>
<tr>
<td>15. <strong>Upper Respiratory Infection (URI)</strong></td>
<td>- Cough or nasal discharge present for three days or more without respiratory distress or increased RR.</td>
</tr>
<tr>
<td>16. <strong>Hypothermia</strong></td>
<td>Auxilliary temperature &lt;95ºF.</td>
</tr>
<tr>
<td>17. <strong>Umbilical Sepsis</strong></td>
<td>Pus discharge from umbilicus.</td>
</tr>
<tr>
<td>18. <strong>Tetanus</strong></td>
<td>Baby which earlier sucked well, stopped taking feeds from 4th day or more; and appearance of seizures, spasm and trismus.</td>
</tr>
<tr>
<td>19. <strong>Convulsive Disorder</strong></td>
<td>Seizures but baby conscious, alert and feeds well between seizures (excludes tetanus, asphyxia, sepsis)</td>
</tr>
<tr>
<td>20. <strong>Unexplained fever</strong></td>
<td>Axillary temperature &gt;99ºF without any attributable cause.</td>
</tr>
<tr>
<td>21. <strong>Failure to Gain Weight</strong></td>
<td>Total weight gain during 0-28 days &lt;300 g.</td>
</tr>
</tbody>
</table>

*Source: Indian Pediatrics, 38: 952-965, 2001.*

### 5.3. Knowledge Engineering Process

#### 5.3.1. Knowledge Accusation and Data Mining Techniques Applied

In this chapter we applied the same knowledge engineering processes that have been discussed in chapter 4. But for developing an actual expert system and to have the highest performance, knowledge acquisition plays one of the most vital roles during the development of the system. Here the system’s knowledge had been acquired through three main sources namely, (a) Medical Experts, (b) Published literature, (c) Real observation at the hospitals and (d) Text Books [18,19,20].

Three Experts in Pediatrics, each of them having 12-33 years of experience were the first source of the main knowledge. The main experts are from Dept. of Pediatrics, North Bengal Medical College and Hospital and others from said department and other
hospitals were consulted through structured interviews. To record the knowledge extracted from the domain experts, forms were prepared in consultation with them. The experts were asked to give their judgments for different sets of possible real observation and cases. In this phase, it was decided to involve multiple experts in the knowledge acquisition process. Thereafter, knowledge was gathered from various research publications including books and workshop reports published by Indian Pediatrics [3], some other journals and text books [18, 19, 20].

For real pragmatic cases, frequent visits were conducted at neonatal unit of Dept of Pediatric Medicine of North Bengal Medical College & Hospitals, Siliguri, West Bengal, India. All real life physical observations were recorded and incorporated in the knowledge base of the system which afterwards creates a repository of knowledge.

5.3.2. Knowledge Repository

The unstructured knowledge as acquired from the above three sources of knowledge was then structured by the knowledge engineer. The knowledge was then represented in an Object Oriented form for later implementation. An Object Oriented Approach to KR (Knowledge Representation) schemes is more structured than other well known schemes and improves consistency, understandability, maintainability and modifiability of the knowledge base [21].

In the knowledge repository, the knowledge in the system is stored in as group of objects. Each group is represented by a class with its attributes. A class defines the general properties of structure of a group of objects. Attributes describe the object’s important characteristics. The knowledge library class serves as a database.

5.3.3. Clustering

Afterwards data mining technique-Clustering has applied. This Cluster detection consists of building models that finds data records similar to each other. This is inherently undirected data mining, since the goal is to find previously unknown similarities in the data. Clustering data may be considered a very good way to start any analysis on the data. Self-similar clusters can provide the starting point for knowing what is in the data and for figuring out how to best make use of it.

To identify and discover patterns, Clustering is one of the best tools in data mining. Clustering helps in recognizing exciting distributions inside the data [26]. When there is no class to be predicted, then Clustering techniques may be applied. Here all the instances are divided into groups. Data clustering identifies the sparse and the crowded places. In this way it discovers the overall distribution patterns of the dataset. Clustering is considered one of the most important unsupervised learning techniques. In most of the other problems in similar type, it deals with finding a structure in a collection data which is not labeled.
Clustering is the process of organizing objects into groups whose members are similar in some way. Hence, a cluster is a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. The reasons for making clustering are as follows:

i. **Detection of Patterns:**

   The new born babies may have several patterns of disease. Even there may of different patterns of sign and symptoms. Thus Pattern detection is one of the vital reasons.

ii. **Simple and Robust:**

   Clustering techniques are simple to find the similar kind of classes. Hence always gives robust result, if the proper information is given.

iii. **Data Concept Construction:**

   In the knowledge base designing process, data concept creation is important. It helps to build similar categorical database. Hence searching process is easy and simple.

iv. **Unsupervised Learning:**

   Clustering techniques apply when there is no class to be predicted. The instances are divided into natural groups. A mechanism causes some instances to bear a strong resemblance to each other than they do to the remaining instances.

Figure 5.1 shows the process of clustering:

![Clustering Process](image_url)
In respect of our study, various information related to neonatal disease were collected from North Bengal Medical College during field surveys and from experts. Both the experts and medical persons are accustomed with various medical terms like “APGAR”, “HIE” etc. which have been incorporated as attribute values. Then the clustering process started for categorizing the similar group class.

5.3.4 Knowledge Unearthing Concept

Knowledge unearthing is a concept of data mining only. Data mining strategies can be broadly classified as either supervised or unsupervised. Supervised learning builds models by using input attributes to predict output attribute values. Many supervised data mining algorithms only permit a single output attribute. Whereas other supervised learning tools allow us to specify one or several output attributes. Output attributes are known as dependent variables as their outcome depends on the values of one or more input attributes. Input attributes are referred to as independent variables.

When learning is unsupervised, an output attribute does not exist. Therefore all attributes used for model building are independent variables. Supervised learning strategies can be further labeled according to whether output attributes are discrete or categorical, as well as by whether models are designed to determine a current condition or predict future outcome. In this section we examine three supervised learning strategies, take a closer look at unsupervised clustering, and introduce a strategy for discovering associations among retail items sold in catalogs and stores. Thus knowledge unearthing process is really helpful for the followings:

- Text and graphical based data may be used applicable.
- Very useful in data mining as a assisting tools.
- Help simplify data complexity.
- Classification.
- Detect hidden pattern in data. Etc.

5.4. Research Methodology for System Design and Implementation

Data Mining often requires data integration, the merging of data from multiple data sources into one coherent data store. These sources include in our case knowledge repository, flat files, and data entry values. Equivalent real-world entities from multiple data sources must be matched up, for example, disease_pattern in one database must be matched up with cause_of_disease in another database. Careful integration of the data from multiple sources helped reducing and avoiding redundancies and inconsistencies in the resulting data set. This helped improving the accuracy and speed of the subsequent mining process.
The system is designed to aid the decision-making process for identification of common diseases in neonates. It also selects the controls measure taken/ management for confirmation of disease. A system flow chart of the system is shown on figure 5.2. Basically three phases are there in the expert system. During the first phase, preliminary identification of the sick neonate in the field can be done on the basis of neonatal information observed in the field and clustering them from knowledge repository. This identification is related with complain/ sign-symptoms of the sick baby. This identification is further confirmed with the knowledge related to characteristics such as Birth Weight, Apgar score, and Gestational Age etc. The system uses domain knowledge in this phase. The system asks for the inputs related to the sick neonate history and sign & symptoms for which he/she is suffering. Firstly the system asks for history related with mother and child, like Mothers age, LMP, DOD (for calculating term/preterm/post term baby), Parity(no of children), Mothers Blood Group, History of hereditary diseases, weight of the baby( to find out LBW / VLBW/ ELBW / NORMAL or finding out weather the bay is AGA(appropriate for gestational age/ SGA-small for gestational age )/LGA(large for gestational age) etc.

After that it asks for the primary signs such as lethargy, refusal to suck poor cry, poor weight gain, Hypothermia, Sclerema, excessive jaundice, Bleeding, GI disorder, seizure, sluggish neonatal reflex etc. The production of rules of inference are used a good level of accuracy is achieved in the resulting identification. The forward chaining of reasoning is performed with the rule based knowledge.
For inferences in the system “IF {I} AND {C} THEN {D} CF” rules have been used, where {I} is the information of the sick baby, {C} is complaint or symptoms and {D} is the differential diagnosis. Here CF is the certainty factor which often call confidence factor as associated with the rule. The confidence factor CF attached to the most of the rule represents actual confidence of that rule. The value assigned to CF ranges from 0 to 100. This knowledge is taken from expert doctors after few interactions. Multiple experts in the pediatric domain actually have suggested the confidence factor (CF) at the time of knowledge acquisition.

The format chosen for the definition of rules allows flexibility in structuring the knowledge [22,23]. A predecessor of any rule may be a composite of number of clause connected through logical operation OR and AND. In addition to the “if-then” structures, common to all rule-based systems, knowledge representation supports a new type of concept—the criterion—that enables a more compact and natural way of expressing rules of the form: “if at least $n$ conditions are satisfied then...”[24].

It was implemented in an object-oriented environment by using Level 5 Object for Microsoft Windows (release 3.0 or higher) by the Information Builders Inc, USA. It has an integrated array of powerful tools like GUI development, Forms and Display.
builders and also has got capability to chain more than one knowledge base together and also have clustering facility by writing code. The facilities provided by the Level 5 Expert System Cell have been discussed on the chapter 4. Considering that facilities we developed this knowledge based decision support system where data mining and knowledge unearthing process has been applied efficiently.

5.5. Performance and Evaluation of the Expert System

In this study, the system had been analyzed for 68 real field cases in different neonatal disease category from the database repository. After applying the data mining techniques comes the job of identifying the obtained results, in form of interesting patterns representing knowledge depending on attention-grabbing measures. Such measures can be used after the data mining step in order to rank the discovered patterns according to their interestingness, filtering out the uninteresting ones. More importantly, such measures can be used to guide and constrain the discovery process, improving the search efficiency by pruning away subsets of the pattern space that do not satisfy pre-specified interestingness constraints.

The introduction of the expert system caused no additional clinical work; the only minor change in working practice required was for the auxiliaries, who normally perform the sampling, to note the mother’s name and hospital identification number before analysis. The system was designed to be simple to use, and the users’ quick acceptance of the system endorsed this. The multiple sticky labels for patient notes eliminated the time consuming and error prone process of transcription, and this helped significantly with user acceptability. For trail run, some randomized cases taken from the set of real field case used for evaluation.

It was found that 90% accuracy or certainty of the system, of course after being verified by the expert pediatricians of North Bengal Medical College & Hospital, Siliguri, and Darjeeling. Case studies are presented in Appendix B as representative examples. These examples have been analyzed by the domain experts. After applying knowledge unearthing concepts the decision of system were matched with expert opinion. We received an accuracy of 90% for accurate diagnosis of the neonatal disease by the system.

5.6. Conclusion

This chapter presents an integrated environment for the development of diagnosis-oriented knowledge based systems applied to diagnose and manage the common neonatal diseases in line with various medical domains [17]. Data mining approach for knowledge repository has successfully shaped. Out of several sign and symptoms mining approach really helped to find the useful knowledge to be processed for reasoning and development of decision support system.

Based on the work done, the following conclusions were drawn:
1. This system provides graphical support for every phase of the development cycle, from knowledge acquisition and addition to reasoning explanation and knowledge-base validation.

2. Level 5, is very suitable as a mining engine with its interface and manipulating modules that allow data exploration, manipulation and exploration of any interesting knowledge patterns.

3. Using the same data sets with different mining techniques such as clustering and comparing results of each technique in order to construct a full view of the resulted patterns and levels of accuracy of each technique may be very useful for this application.

4. Data mining technique is very useful in the process of knowledge discovery in the medical field, especially in the domains where available data have many limitations like inconsistent and missing values.

5. It is an interactive Object Oriented Rule-Based Expert System developed to meet the needs of such places where is scarcity of the medical expert for diagnose of neonatal disease and give proper clinical control management, specially in hospital, and primary health centres at remote villages.

6. Appropriate Graphical User Interface components incorporated in the system a facility to select and deselect multiple options from the easy to use menu. It needs almost no training for its user. Very easy and commonly used linguistic variable inputs and output, which are in natural language and commonly used terms add advantages to a less-trained person seeking expert system guidance for proper neonatal disease diagnosis and treatment plan practices. [Appendix C]

7. The extensive system testing and validation process in conjunction with the overall structured project methodology, has greatly contributed to the expert system reliability and user-acceptance.
References


