1 Introduction

Medical science as well as medical diagnosis and treatment have progressed rapidly in recent years. In addition, medical data that need to be processed have been increased in number and variety, and have become increasingly complex. Accordingly, they should be treated in an integrated manner. Therefore, we proposed a framework, referred to as Cyber Integrated Medical Infrastructure (CIMI), which provides an integrated management of clinical data on computer networks and consists of a database with a visualizer, a knowledge base, an inference component for diagnoses, and a data analysis (learning or data mining) component. All of these are connected with each other in the network (Abe et al., 2007). In addition, the generated and collected data can be used for medical training of students or physicians as well as medical diagnoses. The medical data are usually collected during clinical examinations or inspections, because it is rather easy and requires only reasonable costs to collect medical data during such events. However, such data usually lack several parts and it is difficult to collect perfect data sets. Therefore, in the International Research and Educational Institute for Integrated Medical Science (IREIIMS) project, we decided to collect complete medical data sets not in clinical examinations but by employing a planned medical data collection scheme. As a project, we have collected medical data of more than 3500 subjects until 2009. Physicians have assigned health levels\(^1\) to all the data. Thus each subject’s medical data are related to their health situation and relationships between health situation and medical data can be modelled after collecting sizable sets of medical data. Usually data mining techniques were applied to in the modelling process. Medical data are analyzed or data mined to build a knowledge base for predicting all possible diseases and to support medical diagnosis.

In medical modelling, it is important to generate generalized models that can explain most of the phenomena. However, it would be more important to generate a model that can explain rare or less known

\(^1\) A definition of health level will appear in the following section.
phenomena. Previously C4.5 (Quinlan, 1993) was applied to analyze the medical data and to show the possibility of a chance discovery (Ohsawa and McBurney, 2003) procedure (Abe, Kogure and Hagita, 2003). In this procedure, data or relationships in a gray zone\(^2\) can be regarded as a chance, which is a novel or a rare event/situation that can be conceived either as an opportunity of being healthy or a risk of becoming ill. Actually, a “chance” is a key factor in chance discovery. A chance is defined as “a novel or rare event/situation that can be conceived either as an opportunity or a risk in the future” (Ohsawa and McBurney, 2003). That is, it can also be regarded that a chance is an alarm like an inflation of money supply that will change the middle or long term economic situation (Japan, in 1990).

For the generation of novel or rare rules, Suzuki proposed a method to generate an exception rule in a probabilistic manner (Suzuki, 1999). Joshi proposed PN-rule Learning (Agarwal and Joshi, 2000), (Joshi, 2001). PN-rule Learning is a two-phase rule induction which involves a P-phase that covers most of the possible examples with high support and an N-phase that removes negative examples from those covered in the P-phase. Then it can discover rules that predict the presence of a target class (P-rule) as well as rules that predict absence of a target class (N-rule). In fact, N-rules are generated to improve over-fitting learning of P-rules.

In medical application, Tsumoto regards chances as temporal trends that predict risky events and periodicities in temporal sequences (Tsumoto, 2003). He applied probabilistic induction to discover rare risky events. Indeed, since a chance has a probabilistic aspect, the probabilistic method is a useful way to obtain abnormal or rare knowledge. Thus various researchers such as Tsumoto (Tsumoto, 2004), Ohsawa (Ohsawa et al., 2004), Suzuki (Suzuki et al., 2003), and Ohsaki (Ohsaki et al., 2004) showed the importance of data mining of medical data to discover rare or unknown (interesting) rules for a medical diagnosis. Though the importance of data mining technique in medical data analysis could be shown, the result was not easy to be understood by general physicians. Therefore, it is necessary to develop an understandable interface for general physicians.

In this chapter, first an interactive interface is proposed to show results obtained with C4.5 on a web browser, where the user can check the correspondence data on decision trees. In addition a web-based interface is proposed that can show computationally estimated health levels of medical data, which are not assigned by physicians. Rules for the health level determination are learned from the collected clinical data to which health levels have been assigned by physicians. The interface can be used for the support of medical diagnoses. Consequently the merit of using the web-based interface for managing and analyzing large-scale medical data is demonstrated. Finally, the possibility of using chance discovery approaches in the interface to discover hidden or rare but very important relationships is considered.

In fact, powerful interactive data mining tools such as Clementine have been developed, where interactivity was supported for the purpose of the gradual reconstruction of decision trees to obtain better decisions or to control users’ focus on decision trees. Of course, the construction of better decision trees is important in the data mining process. However the main purpose is to discover rare or exceptional relationships. An additional objective is to discover human related errors (e.g. mis-writing) in the data. Therefore, the interactivity is mainly developed to obtain an effective interface for discovering rare or exceptional data. No other data mining tools except KeyGraph (Ohsawa, Benson and Yachida, 1998) explicitly focussed on this type of objective. KeyGraph can suggest rare or hidden events by its visualization, but cannot suggest data that might have errors. In order to support physicians, Ho proposed an integrated medical data mining tool where he combined appropriate data mining methods, active participation of physicians with a visual interface D2MS (Ho et al., 2003), and effective exploitation of domain knowledge (Ho et al., 2007). His aim to support physicians will be the same as ours. However, the main difference between Ho’s and the system

\(^2\) A zone where an important or necessary phenomenon exists and which is slightly different from the typical or standardized one.
described in this chapter is the attempt to build a portable and dynamic interface where physicians can check original data as well as results from data mining, and medical diagnosis based on these results. In the proposed system, portability is also focused. It is sometimes ridiculous to prepare a peculiar system (including hardware and OS) or software to conduct medical diagnosis. If the system only requires a general software which can be obtained easily, physicians will be easily to use the system. In addition, by the portability, it is possible to conduct medical diagnosis everywhere on portable computers via (secured) networks. It is not always necessary to store all data on individual computers. For instance, a server type computer can store all data and medical systems to conduct an online medical diagnosis.

In fact, this chapter does not focus on visualization with rich graphical patterns. The proposed system offers a portable and rather simple interface which functions on the Web-based browser such as Firefox. In addition, the web-based interface is automatically built after medical the data analysis. It is not necessary to prepare any fixed frameworks, formats, and interfaces for a diagnosis software, which is usually difficult to prepare.

Section 2 describes Cyber Integrated Medical Infrastructure. Section 3 illustrates the feature of medical data. Section 4 analyzes the medical data using data mining. Section 5 discusses a health level determination by the analyzed clinical data. Section 6 proposes an interactive medical diagnosis support system utilizing a web-based interface. Section 7 shows a simple operation example of the proposed system. Section 8 is a conclusion which provides an overview of the possibilities of chance discovery process for medical diagnosis support.

2 Cyber Integrated Medical Infrastructure

Recently, for medical research, integration and cooperation among the various fields has been advocated. Therefore, we propose a framework called Cyber Integrated Medical Infrastructure (CIMI) which is a framework of integrated management of clinical data on computer networks. Figure 1 is an image of CIMI. As shown, CIMI consists of a database, a knowledge base, and an inference and learning part which are connected to each other in the network. First, a brief summary of functions in CIMI is shown below;

- **Database+knowledge base**
  Various types of data are stored in the database to be analyzed by machine learning techniques. The analyzed results are regarded as medical knowledge and stored in the knowledge base to be used in computational inferences to support medical diagnosis. A medical data viewer shown in the left top of Figure 1 has been developed, which will not be dealt with in this chapter.

- **Training part**
  By using the database of patient in CIMI, actual trainings are conducted. A network-based software was developed for medical trainings in IREIIMS. By accessing the database and the network, participants study a total of 24 clinical case once a week. Currently, the course is supervised by several tutors who aid the learning (Kira et al., 2007). In the future, the generated knowledge base can be also used for the medical training.

- **Analysis part**
  By using C4.5, this part analyze medical data to generate medical knowledge. As shown in this chapter, results are also displayed in web-based interactive interface. Where the user can check peculiar
patient’s medical data from the generated decision tree and correct the data (A correction part is to be developed). Details are illustrated in the followings.

- **Diagnosis support part**

  By the similar interface as above, estimated health levels are calculated and displayed on the web-based interface. The user can check the patient’s medical data on the same interface, if estimated health level is different from the actual health level. Details are illustrated in the followings.

![Figure 1: Cyber Integrated Medical Infrastructure (CIMI) framework](image)

Thus, in the CIMI framework, medical, clinical and the other information (e.g. personal information, interview) is analyzed or data mined to discover relationships between the medical, clinical and other data and all possible diseases. Although, CIMI includes various types of data, as a first step we mainly deal with clinical data. The clinical data includes liver, pancreas, and kidney test results, tumor markers, and blood test results. In addition, clinical interviews, Ryodouraku results and plethysnographic analysis (Eastern medicine) data are added. Thus, the database contains data from both Western and Eastern medicine. Usually, Western and Eastern medical research is conducted separately. One of our aims is to discover relationships between Western and Eastern medical treatment. However, in this chapter, since we focus on an interface techniques for medical diagnosis support, such relationships is not discussed.

### 3 Features of the Clinical Data

In this section, we describe the features of the clinical data collected for CIMI. Although, as shown in the previous section, various types of data such as Ryodoraku results will be stored in the database, we are mainly dealing with clinical data in this chapter.
3.1 Clinical data

We have intentionally collected various types of clinical data, such as those obtained in blood and urine tests. From certain persons, the data were collected more than once. More than 130 items are included in the medical data of each person. Item sets in the data are, for instance, β2-microglobulin (serum), SCC antigen, γ-seminoprotein, TPA, TK activity, NSE, Ferritin, BFP (serum), Total protein, albumin, γ-GTP, ZTT, serum protein fraction-α1-globulin, ALP Triglyceride, LDL Cholesterol, HDL Cholesterol, urine sugar, HbA1c, White blood cell count, Platelet, Cl, Cell immunity T Cell number, and ACP. In addition, to each medical data, health levels are assigned by physicians considering with the medical data and clinical interviews.

Health levels that express the health status of patients are defined based on Tumor stage (Kobayashi and Kawakubo, 1994) and modified by Matsuoka. Originally, health levels consist of 5 levels (I–V). Persons at level I and II can be regarded as being healthy, but those at levels III, IV, and V can possibly develop cancer. In (Kobayashi and Kawakubo, 1994), level III is defined as the stage before the shift to preclinical cancer, level IV is defined as conventional stage 0 cancer (G0), and level V is defined as conventional stages 1–4 cancer (G1–G4). Categorization of the health levels is shown in Figure 2 (“%” represents a standard distribution ratio of persons in the level).

For more detailed analysis, Matsuoka defined more detailed categorization, which are I, II, III, IVa, VIb, VIc, Va, and Vb. Since in our case health levels VI and V contain many data, health levels VI and V are categorized more.

<table>
<thead>
<tr>
<th>Health Level</th>
<th>Health Condition</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Excellent</td>
<td>0</td>
</tr>
<tr>
<td>II</td>
<td>Good</td>
<td>10</td>
</tr>
<tr>
<td>III</td>
<td>Fair</td>
<td>60</td>
</tr>
<tr>
<td>IV</td>
<td>Needs an improvement in lifestyle</td>
<td>25</td>
</tr>
<tr>
<td>V</td>
<td>Needs a precise examination and therapy</td>
<td>5</td>
</tr>
</tbody>
</table>

Figure 2: Health levels (standard)

Table 1: Health levels.

We have collected more than 3500 data with health level assignments. Medical data are mainly collected from the same subject group as that described in (Abe et al., 2007). Thus we collected data from office
workers (aged 40 to 50 years old) but not from students. Therefore, despite of the new categorization, there still exists a unbalanced health level distribution (Table 1). For instance, half of them belong to the health level IVb (not so healthy). Accordingly the imbalance of the data should still influence the analyses. In the following sections, descriptions of 1,2,3,4, and 5 are used for health levels instead of I, II, III, IV, and V.

3.2 Data categorization

For the detailed analysis, we categorize the above mentioned items into 6 categories as shown below. The categorization is decided according to effects or functions to our health status, for instance, cancer and metabolism.

1) peculiar tumor markers (21 items):
2) inflammatory tumor markers (11 items):
3) liver, pancreas, and kidney test data (28 items):
4) metabolic function test data (16 items):
5) blood and immunity test data (12 items):
6) others (47 items):

These categories include such items as shown below:

1) β2-microglobulin (serum), SCC antigen, γ-seminoprotein, TPA
2) TK activity, NSE, Ferritin, BFP (serum)
3) Total protein, albumin, γ-GTP, ZTT, Serum protein fraction-α1-globulin, ALP
4) Triglyceride, LDL Cholesterol, HDL cholesterol
5) Urine sugar, HbA1c, White blood cell count, Platelet
6) Cl, Cell immunity T Cell number, ACP

Although we can analyze the medical data without the categorization, this categorization plays a significant role in medical data analysis from the viewpoint of chance discovery as well as the better medical model generation.

4 Clinical Data Analysis

In this section, we illustrate a simple data mining of medial data. For the database in CIMI, medical data includes patients’ clinical and personal data. Currently, more than 130 items including tumor markers, are included in the clinical data. As shown above, these items are categorized into 6 categories. In addition, health levels are assigned by physicians to the clinical data for each person.

As pointed out above, medical science as well as clinical diagnosis and treatment has progressed rapidly in recent years. Accordingly, medical data that should be analyzed has been increasing. Of course, such data are also intentionally collected for medical research purposes. We have also intentionally collected various types of clinical data that are obtained during blood and urine tests and so on. Previously, such data was (manually) analyzed by physicians. However, due to the enormous increase in the amount, complexity and variety of the data, it is rather difficult, even for specialized physicians, to handle them. Therefore, automatic analysis is necessary to deal with this problem. In the field of artificial intelligence, the data
mining technique is widely applied to analyze numerous and complex data to discover the features of the data. We applied C4.5 (Quinlan, 1993), which is a machine learning tool that generates a decision tree, to the medical data of around 1200 individuals\textsuperscript{3}. We apply C4.5 because it generates a decision tree which can be regarded as a set of logical formulae and does not require any background knowledge. Since we try to discover novel, hidden, or rare relationships as well as well-known relationships, it is rather difficult to prepare background knowledge and sometimes improper background knowledge leads to incorrect results. In addition, we applied a statistical method to the data and compared the results. Statistical methods cannot generate “logical formulae,” but can generate images of data relationships and the results are more intuitive and understandable than logical formulae. On the other hand, the generated decision tree actually provides us a certain image for data relationships. However, it can also be regarded as a set of logical formulae representing the relationships between clinical data and health levels.

In fact, when we apply C4.5 to the clinical data, a decision tree can be obtained. For instance, a decision tree such as below (part of the generated decision tree) can be obtained.

\begin{verbatim}
ICTP <= 5.8
| CYFRA <= 2.1
| | TK Activity <= 5
| | | Leucocyte image:Neutro <= 40
| | | | SCC Antigen <= 0.9 : 4a........
\end{verbatim}

The shape of the generated decision tree has a tree or hierarchical pattern, and does not seem like a logical formula. However, it contains logical interpretations. For instance, the above decision tree can be interpreted as the following formula (if-then style).

If ICTP $\leq 5.8$, CYFRA $\leq 2.1$, ......, and SCC Antigen $\leq 0.9$ then health level = 4a.

Thus, if sets of the generated decision tree is confirmed by physicians, we can use them as sets of rules for a medical diagnosis support. However, the most serious problem is that since ordinal physicians are not familiar with the decision tree, it is difficult for them to interpret the meaning of decision trees. Actually, more complicated relationships than those shown above can be obtained according to data complexity. Therefore, it is necessary to prepare an understandable interface for ordinal physicians. Even for researchers who are familiar with decision tree, it would be better to provide more understandable interface to conduct chance discovery (Ohsawa and McBurney, 2003) as well as medical diagnosis.

5 Health Level Determination by the Analyzed Clinical Data

As shown above, a medical rule set can be generated by decision tree learning of medical data labeled with health levels. For a medical diagnosis support, we applied C4.5 to the medical data from the reason shown above. First, we simply apply C4.5 to the data and evaluate the quality of generated rules.

When we apply rule sets generated by C4.5 to the same medical data sets that C4.5 is applied to determine health levels, we can obtain a good result, where 93.53% of data are correctly determined, and 6.67% of data are incorrectly determined. It is an acceptable result. Actually from 2463 medical data sets we extracted 1500 medical data sets to generate rules. However, when we apply the above rule sets to the other medical data sets (963 data sets), we obtain rather insufficient result, where 380 (39.46%) of data are correctly determined, and 583 (60.54%) of data are incorrectly determined. In fact, part of the medical data is collected from the same subjects on different dates. Therefore, there could be a very small gap between

\textsuperscript{3}Some data are taken from the same person on different days.
learned data and health level determined data. However, a decision tree generated from 963 data sets is quite different from that generated from 1500 data sets. Even an item on the root (top) of a decision tree is different. According to our experience, an item on the top of decision tree is usually the same, when we learn similar data. A reason might be that since the data set contains various types of data and the number of items is quite large (around 140), C4.5 might generate rather specialized (over-fitting) models for the given medical data.

It will be necessary to analyze the original medical data as well as generated relationships (rules). In fact, we introduced various types of techniques to reduce negative influences to results (Abe et al., 2008a), but it was not easy to improve the estimation accuracy. Therefore, it is also necessary to analyze original data by referring to learnt results. For that, it would be necessary to introduce such an interface that can check relationships between the original data and analyzed results interactively and in detail.

6 Medical Diagnosis Support System

6.1 Clinical data analysis interface

In Section 4, we briefly illustrated data mining of clinical data, where we applied C4.5 to obtain relationships between health levels and clinical data. At the end of Section 4, we pointed out that for those who are not familiar with decision tree, it is not easy to understand or comprehend decision tree. In addition, it is necessary to check relationships between the original data and analyzed results interactively and in detail. For the reason, we sometimes need to check which patients are on each trees and leaves. Therefore, we have developed an interface that can show who are on each trees and leaves. Results by C4.5 can be displayed on web browsers and leaves have link anchors that can show various information on data in the decision tree. The system’s interface is shown in Figure 3. If we click link anchors on leaves, another browser will be coming up that has more detailed information (the center browser in Figure 3). In the second (center) browser, we can check a range of each items relating to the health level. For instance, for health level 5a, the range for TK Activity is less than 5.0. We can also check who are on the leaf that has a certain health level (e.g. “5a”) in the top of the second browser. In addition, we can check individual data by following links in the top of the second browser. Thus we can review results by C4.5 in the organic view and we can check each value relating to certain health levels in detail.

The other merit of using the interface is that we can confirm or check the clinical data. For instance, in Figure 4, differently classified data are shown in different colour (red: 2012,2006-06-01,3). In the case, a datum that has health level 3 is on the leaf of health level 4b. If we encounter differently classified data, by checking the data on the browser, we can confirm whether the health levels are correct or not. When the health level is incorrect, we can check individual data to change the health level. On the other hand, when the health level is correct, we should check the clinical data. This procedure sometimes functions as a correction of error in data recording or writing. In addition, when both data and health level are correct, we might be able to discover hidden factors that have serious effects on patients health. For instance, for the data of “2012,2006-06-01,3” we can see several out-ranged data, if we check each data, we can determine if he/she is in health level 3 or 4b. If he/she should be categorized to 3, data that do not appear on decision trees might give positive effects on patient, which we can call chances and we can add new rule as discovered in this process. On the other hand, if a person who has health level 5a is classified to 4a, and if he/she should

---

4 Of course, the other factors also affect the health level. In fact, in the case Figure 3, the range of CA72-4 is more than 4.0. If the range of the data is outside of the standard value range, the item is shown in red color. The standard value is obtained by referring to the data library of Mitsubishi Chemical Medience Corporation (http://www.medience.co.jp/).
be categorized to 5a, data that do not appear on decision trees might give negative effects on patient, which we can call chances and we can add new rule as discovered in this process. Of course, in the process, we can discover wrong classifications. Thus we can check the data from the various viewpoint as well as analyzing the clinical data on the web interface.

The above discovery or correction can be recorded to the original data and knowledge base. We have not developed such an editor, but in the future, such an editor can be developed to conduct interactive data and knowledge management.

6.2 Medical Diagnosis Support Interface

In the previous section, we illustrated the developed web interface for analysis of clinical data. The same interface can be used for a medical diagnosis support, because in the above application, all clinical data are checked to determine decision trees that can explain them. The same procedure can be applied to determine health levels of clinical data. Thus for a medical diagnosis support, we mean that the system can estimate patients’ health levels by applying results of known patients’ data analyses by C4.5. Since we are planning to develop an integrated medical diagnosis system involving a data analyzer and a decision maker in medical cases.

Figure 5 shows the first browser that shows estimated health levels of each patients. To generate health levels, we applied the results of the 1200 individuals’ clinical data analysis by C4.5. As pointed out above, we can adopt the rules shown in the decision trees for the estimation. In this framework, we directly use the decision trees for the estimation. That is, health levels are calculated by applying each data to the decision
As shown in Figure 5, a list of estimated health levels for each patient are shown in the left window. In addition, the applied decision trees are shown in the right window. When we click an ID information, a tree that contains the ID’s data appears (moves) in the top of the right window. Figure 6 shows the interactive procedure on a patient “1138, 2007-01-22.” At the top of the right window of the left browser, “... 105: 5a(0/0/4)” appears which includes a link to the patient “1138, 2007-01-22.” By clicking “5a(0/0/4),” the middle browser appears.

The continuing procedure is the same as that shown above. We can check the data on leaves to confirm or change health levels.

For Figure 6, we can review a certain person whose health level is estimated as 5a. In the right window of the right browser, we can check the selected data to confirm if he/she is in the health level of 5a. Similarly, we can view all the estimated health levels at a glance and check doubtful or uncertain person’s data. The proposed system offers a flexible interface to conduct medical diagnosis and data compilation. The most important point of the proposed system is that all windows (interfaces) are automatically generated according to the data set and the analyzed results.

7 Operations on the Medical Diagnosis Support System

7.1 Health level determination evaluation

In the previous section, insufficient result were obtained when the generated rule sets were applied to a data set that was different from the data used to generate the rule sets in the first place. From the 2500 data
sets, 1500 data sets were used for rule generation and another 1000 data for rule evaluation. Usually, such an evaluation is conducted as shown in Section 5. However, for a better understanding and achieving of the best analysis, an interactive interface was adopted to evaluate the process of health level determination. When C4.5 is applied to 1500 medical data sets, a decision tree shown as Figure 7 is generated. At the top of the figure (actually in the middle of the full decision tree), “EBV-EBNA > 20 : 4a(0/2/1)” is shown, which means that two items are incorrectly determined and one is assigned to a new health level (the health level has not yet been assigned.). This link should be checked, because it contains two incorrect health level determinations. As shown in Figure 8, the individual whose ID is 1223 is assigned to health level 5a and the individual whose ID is 1155 is assigned to health level 4b. Actually, they are assigned to health level 4a by physicians. Health level 4b is close to 4a, both of which belong to the not-so-healthy category and roughly belong to the same category (health level 4). Thus the assignment will be accepted. In fact, the individual had rather abnormal values in several items such as NSE (6.2), EBV-VCA-IgG (160.0), and EBV-EBNA (40.0). However, certain problems in determining the health levels should arise with health level 5a. Health level 4 and 5 are different. Therefore, the data of the patient with ID: 1223 were reviewed in detail.

For the review of these data, the interactive interface was utilized. By clicking at the ID (in this case, 1223) in the second browser, the third browser was obtained as shown in Figure 9, which presents the medical data for that ID. In the browser, the abnormally high values can be found as written in red and abnormally low values written in blue (in BW print, light gray). For instance, in this case, Apolipoprotein A-I is 192.0, Apolipoprotein B is 107.0, and blood sugar is 120. That means that he/she suffers from rather serious Lipid metabolism abnormalities or Arteriosclerosis. In fact, with the proposed interface, it is rather easy to focus on such abnormal data. Why did rules extracted from C4.5 determine his/her health level as
4a instead of 5a? When the rule set was applied to his/her data, neither Apolipoprotein A-I, Apolipoprotein B nor blood sugar were used to determine the health level. Apolipoprotein B is not in the path after “TK activity $\leq 5.4$.” The patient’s TK activity is 5.4. Of course, this will not cause a serious problem. Even if we follow the path after “TK activity $> 5.4$,” neither of the above factors is used to determine the health level. In fact, Apolipoprotein B appears in the decision tree. Accordingly, the rule set was applied that was only generated from the metabolic function test data including Apolipoprotein A-I, Apolipoprotein B, and blood sugar. In this case, a health level of a subject (ID: 1223) become 5a. Due to an imbalance in the influence of attributes, sometimes generated rules (models) are influenced by such powerful factors. In (Abe et al., 2007), an integrated data mining approach has been proposed to remove influential factors which might disturb an appropriate model generation. Thus, in this application, the influence of the TK activity (tumor marker) should be removed in this process, but still a correct health level determination cannot be made.

7.2 Model generation and chance discovery

As shown above, from the current data set, it will be difficult to model relationships between medical data and health levels only by applying C4.5. In fact, it is necessary to collect more data, especially data of health level 3 etc. which are rather few, to model more proper relationships between medical data and health levels. In addition, one of the serious weak points of the decision tree approach is the inability to generate multiple models in a simple way. That is, the top of decision tree has only a set of two alternatives. For the other analysis, the Clementine system was applied to perform a basket analysis (Agrawal et al., 1993). A 3-item-sets basket analysis for reasons of computational limitations, we could not perform more than 4-item-sets basket analyses.
sets consisting of “Immunosuppressive acidic protein and ALP Type 2 isoenzyme,” “Immunosuppressive acidic protein and Apolipoprotein E,” “Alpha1-Globulin and Apolipoprotein E,” etc. are key factors for determining health levels. For the health level 4a, “ALP Type 5 isoenzyme and Immunosuppressive acidic protein,” “White blood cell counts: Eosino and Vitamin A,” “Pepsinogen I and HDL Cholesterol,” etc. are key factors for determining health levels. Thus categorization information can also be obtained from the result. Then relationships between medical data and health levels can be modeled by applying both C4.5 and basket analysis, where basket analysis can support the categorized modeling for C4.5. In fact for health level 5a, significant or effective combinations of category are, for instance, “inflammatory tumor markers and liver, pancreas, and kidney test data,” and “metabolic function test data and liver, pancreas, and kidney test data.” These combinations seem to control rather unhealthy health status. However, as shown above, there still exist exceptions such as the case of subject (ID: 1223).

For the chance discovery based modeling, it would be necessary to analyze such exceptional cases. And it would be necessary to generate missing or hidden paths or new rules to explain events that cannot be explained by standard models. For such explanations, usually abduction\(^6\) can be applied. If a part of

---

\(^6\) For abduction, Peirce wrote that abduction is an operation for adopting an explanatory hypothesis, which is subject to certain conditions, and that in pure abduction, there can never be justification for accepting the hypothesis, other than through interrogation (Peirce, 1955). In Peirce’s sense, abduction is a type of discovery reasoning. In AI, if we make use of the framework of logic, a type of justification will be made by generating and testing candidate hypotheses to find acceptable hypotheses. Therefore, in simple
In the above case, a new model should be built to explain the situation. That is, if models cannot properly determine health levels, a new model that can explain an exception should be added to the current model. After that, additional models can be considered during health level determination and be compared with the standard model generated for instance by C4.5. Such additional models can then be regarded as chances in the context of chance discovery which do not frequently appear but are very significant for determining exceptional cases.

In fact, these types of exceptions are not regarded as “exception” when we check data personally, because we can usually focus on abnormal data easily. In addition, for medical diagnosis, such cases are reported as abnormal case for health. However, if computers view such data based on own generalized models to determine health levels, it is not easy to focus on such exceptions. From the computational viewpoint, the preparation of such exceptional models is very important to perform a proper health level determination. In that case, standard models can be given up and exceptional models can be applied to determine health levels. That is, a computational procedure should jump to rare or novel models when necessary.

To develop the mechanism to determine exact health levels is not the main aim of this chapter, though of course it is very important for medical diagnosis. However, in the future, it is necessary to improve the estimation quality or develop such a mechanism to determine exact health levels. In general, if we can collect large amount of clinical data to analyze and tuneup the data mining engine, the estimation quality will be improved. Furthermore, as shown in the above, the estimation quality can be improved by using the developed interface as the procedure of chance discovery. In fact, in (Abe et al., 2010), we pointed out several relationships that could improve the health level determination accuracy. Those relationships were discovered by using the proposed interface.

8 Conclusion

In this chapter, we proposed an interactive interface for a medical diagnosis support. First, we showed a clinical data analysis interface where individual data are placed on decision trees and we can review individual data by following links on a web browser. The web-based interface is automatically built during clinical data analysis. Thus it is possible to analyze the original medical data in detail as well as the generated relationships. It has been pointed out above that by using the interface, individual data can be checked to change or confirm health levels of subjects. By using the interface, exceptional data for determined health levels can be easily discovered. Such exceptions can sometimes be regarded as chances.

In words, abduction can be formalized as a logical explanation of an observation. For computational abduction, Theorist (Poole et al., 1987) and Abductive Logic Programming (ALP) (Kakas et al., 1992) are developed.
Second, in addition to computational medical diagnosis support, the same interface has been used to conduct medical diagnosis based on generated relationships between health levels and medical data. These were generated relationships regarded as medical rules for health level determination. As shown above, the web-based interface is automatically built after medical data analysis. It is not necessary to prepare any fixed frameworks, formats, and interfaces for a diagnosis software, which is usually difficult to prepare. Since it uses the same interface as a clinical data analysis interface, it can also change or confirm health levels. In addition, it was emphasised that during interactions, it might be possible to discover hidden factors that might have negative or positive effects on patients health.

For discovered rare rule sets, the necessity of generating a missing or hidden path or a new rule was emphasized. The purpose of this rule generation is to obtain rare or novel models. In Section 7.1, the major merits and features of this interface were shown as given below:

- Each application uses an interactive interface that is dynamically generated during data analyses.
- Only browsers such as Firefox are required to perform health level checks and medical diagnoses.
- Even if health levels are differently recorded, when the subjects health levels are properly determined, it is easy to correct them. The reason for this is that differently recorded health levels are displayed in red colour and therefore the link can be followed to check the original data and to confirm if their health levels are correct or not.
- Exceptional (abnormal or normal) factors and the missing rules can be discovered with an interactive interface. For instance, necessary rules which are not considered during health level determination can be discovered by focusing on values that are abnormal for their health levels. As shown above, it is easy to detect abnormal values by their colour.

For the same reason, it is not difficult to discover exceptions, because exceptions usually contain abnormal data for their health levels.

- Without the proposed interface, it is rather difficult to focus on necessary and meaningful data sets, because more than 100 items are contained in each individual data set and more than 2500 data have been collected. By using the interface, it is easy to determine which data should be checked.

It is also highlighted that it is possible to use abduction for the generation of missing or hidden paths. In this chapter, a concrete strategy to generate missing or hidden paths or new rules was not proposed. In the future it is necessary to propose a strategy to suggest missing or hidden paths or new rules. An abductive discovery approach will be one of the computational solutions.

According to a physician’s evaluation of the described system, though it is very useful to discover hidden, rare or unusual data and relationships, it is a system for rather professional physicians, it is rather difficult to obtain relationships between subjects’ status, and it will necessary to change the order of items to classify similar or related items. Actually, the current system is designed for rather professional person, but in the next version, it will be necessary to install the above pointed out function for the effective and intuitive analysis.

For an interactive interface, it is planned to develop an intuitive visualization of decision trees. As mentioned above, the output from C4.5 is not easy to understand for those who are not familiar with it. In the future, it is necessary to develop understandable interfaces for those who are not familiar with the C4.5 tool. This work will also be discussed from the viewpoint of chance discovery.
Acknowledgement

This research was conducted when the authors worked in the IREIIMS, Tokyo Women’s Medical University. This research was supported in part by the Program for Promoting the Establishment of Strategic Research Centers, Special Coordination Funds for promoting Science and Technology, Ministry of Education, Culture, Sports, Science and Technology (Japan). We thank Dr. Yoshiaki Umeda (NTT-AT) for supporting to develop analysis tools.

References


