COPYRIGHT © 2011 IJCIT, ISSN 2078-5828 (PRINT), ISSN 2218-5224 (ONLINE), VOLUME 01, ISSUE 02, MANUSCRIPT CODE: 110116

Classifying the Practitioner's Behavior in Medical Informatics by Using Data Mining

Kamal Uddin Sarker, Nazmun Nessa Moon *and* Samsuddin Ahmed

Abstract— Most information in medical sector is still only available in text format without a standard predefined format, especially in Bangladesh and the amount of this data is increasing. Text mining is an essential part to dig out knowledge from plain text. The aim is to transform data into information. However for the, efficient support of biomedical researchers or efficient support of end users, facets of computer science alone are insufficient; the next step consists of making the information both usable and useful. This paper addresses the effectiveness of data mining techniques in analyzing and retrieving unknown behavior patterns from gigabytes of data collected in the Hospital. We developed a new relationship known as *Behavior Rules* (BR) and used moderate of Neural Segmentation [1] to detect Practitioners behaviors. The results obtained from this study demonstrate the potential value of data mining in health, by detecting patterns in the ordering of pathology services and by classifying the general Practitioners into groups reflecting the nature and style of their practices.

Index Terms —Behavior Rules, Neural Segmentation, Practitioner's behavior, Support and Confidence System (SCS), Textual-data-mining.

1 INTRODUCTION

DATA mining technique in the field of medical science is being progressed, in order to cope with rapidly increases information system [3]. However document for text mining techniques in the area of clinical information systems and medical documentation are rare [4]. The broad application of sophisticated medical document systems amasses large amount of medical documents, which must be reviewed, observed and analyzed by human experts [4]. All essential documents of patients' records contain at least a certain amount of data which has been entered in *free-text* fields and has long been in the focus of research. As these databases grow larger, with gigabytes sizes becoming quite common, they are overwhelming the traditional query and report-based methods of analysis [5]. Data mining is the data driven extraction of information from such large database, a process of automated presentation of patterns, rules, and functions to a knowledgeable user for reviewing and examination [1]. With the steady rise in health-care costs, specially in Bangladesh where a huge number of pharmaceutical company who spend marketing cost more than several times of production cost of a medicine. We consider that it is a thread for the general people of the country as well as doctors' tendency to prescribe unnecessary medicine or physiological tests. It is demand of time to control these costs. Though this kind of activities available in modern country they can enjoy proper treatment and health insurance facilities. From paper [1], we informed that the Australian Health Insurance Commission (HIC) has collected detailed claims information and has established a homogenous claims database; this has been done trough the administration of various programs (Australia's Medicare, Pharmaceutical benefits scheme, Child care and rebate scheme, Medibank private, and Fraud and inappropriate practice prevention) are great inspiration for our work.

- ***Kamal Uddin Sarker*** *is with the Department of Computer Science and Engineering, Northern University Bangladesh, Dhanmondi, Dhaka-1209, Bangladesh. E-mail: ku_sarker@yahoo.com.*
- ***Nazmun Nessa Moon*** *is with the Department of Computer Science and Engineering, State University of Bangladesh, Dhanmondi, Dhaka-1205, Bangladesh. E-mail: moon_ruet@yahoo.com.*
- ***Samsuddin Ahmed*** *is with the Department of Computer Science and Engineering, State University of Bangladesh, Dhanmondi, Dhaka-1205, Bangladesh. E-mail:suacsecu@gmail.com.*

2 BACKGROUND

Bangladesh is a country of third world, pharmaceutical is one of the important sectors in the field of economic of this country. An internal competition is important factor for the existence in the market for an industry. Marketing is common and important phenomenon for the future of the industries but now-a-days it acts as an influence and folk are being suffered for overloading price of medicine as well as pathological test. Most of practitioners like to advice unnecessary test and medicine for more benefited from the company or percentage of test. It is painful that Medical Promotion Officer (representative) of the companies stands in front of medical and observe the prescription. That means industries are busy with promotional activities than increasing quality of drug. The people have to pay more money for their simple problem in health. Inappropriate practice deals with issues such as requesting or providing services which are unreasonable, unnecessary or excessive (e.g., indiscriminate ordering of cholesterol test) [1]. Typically this type of analysis on human experts, but these experts are both expensive and scared. HIC Company has supported the process with a neural network; though this approach can only be used on a subset of database is mentioned in paper [1] also. We preferred to divide subset to subset by using concept of *Divide and Conquer* algorithm for better classifica-

tion of data base.

3 PREPROCESSING

It was a hard work to collect raw information from the prescription where exactly no well defined structure for the practitioner. In paper [2] mentioned that for additional assistance in finding topological relations between anatomic structures including regional indicators or diseases such as tumor. Given this requirements, there was a demand for domain specific database for anatomic structures and pathologic expressions. It was a hard job to represent in a format. We tried to represent within common format for reducing number of different words. We felt three series problems: synonyms, medical dialects and abbreviation which also mentioned in paper [10].

4 TEXT MINING

Text mining sometimes called *textual-data-mining* is generally defined as a knowledge intensive process in which end users interact with a collection of textual information by using analytic tools; typical text mining tasks include categorization, clustering, concept extraction, production of taxonomies, sentiment analysis, document summarization and entity relationship model [6]. Consequently, Text mining stands for multidisciplinary field involving information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, machine learning and data mining.

Contrary to structured information, textual information is characterized by its inherently unstructured and fuzzy nature [1], being language and domain dependent as well as consisting of sentences and sub-sentences [7]. Text represents factual information in a complex, rich, and opaque manner which makes it difficult to be analyzed by standard statistical data mining method: relying on human analysis results in either huge works load or the analysis of only a tiny function of the data base [8].

Text mining in medical documentation: Text mining approaches can be useful for many different application scenarios such as pattern search, text classification and knowledge recovery for medical diagnosis [2], analyzing and retrieving unknown behavior from health insurance information system [1], also well known application for genome database for protein sequence (special field in bioinformatics). We have applied mining technique in another field of medical informatics system. The major challenge of biomedical text mining over the years is to make such systems useful to biomedical researchers. Certainly, this will require enhanced access of full text, better understanding of the feature space of biomedical text, enhanced methods for measuring the usefulness of systems to end users and continued cooperation the biomedical research community to ensure that their needs and requirements are appropriate addresses [9].

To extract information used a limited structure of sentences and sentence of not more than 16 characters. A short (won defined) word is used for a long term. If we get a unique computerized prescribed system in Bangladesh it was easy to apply our algorithm for extraction of information. We have found a few words with less frequency and considered rare case so neglected from database.

5 BEHAVIOR RULES

From doctor's prescription we can identify the information of (*identifiers*) i) Doctor's qualification ii) Patient's information: primary problem with symptoms iii) Issued medicine and diagnosis texts and later one is important for our work. It is easy to put out by an expert manually for a few amount but not possible for an area. Also factors are biasing or Fairless available here. Paper [1] described the mining of association rules relates to finding intersection patterns to recover knowledge from Health Insurance Information System (HIIS). For our research: *Support* and *Confidence System* (SCS) a new term to identify Practitioner's Behavior (BH). Where:

Support is the number or fraction of the transactions that contain a given item set.

Confidence measures the frequency that items in a multi-item set.

Consider Minimum Support S_{min} and Minimum Confidence C_{min}.

The mining of behavior rules relates to finding intra-transaction patterns and can be defined as follows: given a database of transactions, where each transaction represents a set of items, generates all behavior such that the presence of some specific items S_{items} in a transaction implies the presence of other items O_{items} .The behavior rule S_{items} => O_{items} with hold S_{min} and C_{min}.

If the user-specified S_{min} >= C_{min} of the practitioner we can consider positive situation.

From mentioned (iii) identifiers Reasonable Support (R_S), Unreasonable Support (U_S), Reasonable confidence (R_C) and Unreasonable confidence (U_C) used to identify primary situation (+ve or -ve).

If U_C is increases then the situation is going to negative.

$$\text{Satisfaction} \propto R_S \quad \ldots \quad \ldots\ldots\ldots \quad \ldots\ldots\ldots \quad \ldots\ldots (1)$$

$$\text{Satisfaction} \propto \frac{1}{U_S} \quad \ldots\ldots\ldots \quad \ldots\ldots\ldots\ldots\ldots \quad \ldots\ldots (2)$$

$$\text{Satisfaction} \propto R_C \quad \ldots \quad \ldots\ldots\ldots \quad \ldots\ldots\ldots \quad \ldots\ldots (3)$$

$$\text{Satisfaction} \propto \frac{1}{U_C} \quad \ldots\ldots\ldots \quad \ldots\ldots\ldots\ldots\ldots \quad \ldots\ldots (4)$$

In combine we can write:

$$\text{Satisfaction} \propto \frac{R_S R_C}{U_S U_C} = S\frac{R_S R_C}{U_S U_C} \quad \ldots \quad \ldots\ldots\ldots\ldots \quad (5)$$

where S is the satisfaction constant. If all element contain 1 unit or $R_S R_C = U_S U_C$ satisfaction is optimal.

Actually so many parameters including users characte-

ristics depend on the satisfaction, so we can not remove the satisfaction constant from the equation 5 with a suitable value. Actually value of S depends on physical objects, mental health as well as background of the people. We can categories practitioner by the values of equation. Actually we preferred heuristic solution except exact solution for this kind of data mining.

6 DATABASE SEGMENTATION

We have created two separate databases: i) General Practitioner Database (GPD) and ii) Pathological Database (PD). From GPD we have found the common medicine for most of the patient which is rare or totally absent to another practitioner for same type of diseases; knowledge that may be unnecessary which is also for vice versa. Same knowledge is recovered from tests. Other side from PD a great amount clinic reports having no symptom of desired diseases, unnecessary tests.

Figure-1, the simplest representation of database for diagnostic centre, we stored the information is retracted from diagnostic center. It contains near about 1 thousand records and 20 attributes including some plain text, operated by relational database and queried to observe phenomenon of the service.

Figure-2, the Common Practitioners database contains 22 attributes with approximately 1 thousand records, which correspond to active general practitioners during one month period. Which assess the Quality of Service (QoS) i,e benefited by the organization. Additional descriptive elements include data such as age or sex of the physician.

We have statistical analyzed the following tables for month October-2009, for two same type practitioners in a clinic in Dhaka city basis on Pathological test.

Same way we calculated the table for advised medicine through out the month which are not given here. But special case they used a common medicine as a foreign vitamin to the most of the patients.

TABLE: 1
GENERAL PRACTITIONER-X

Symptom	Total test	S_{ITEMS}	O_{items}	S_{ITEMS}%	O_{items}%
S1	650	136	514	21	79
S2	578	201	377	35	65
S3	635	107	528	17	83
S4	624	210	414	34	66

TABLE: 2
GENERAL PRACTITIONER-Y

Symptom	Total test	S_{ITEMS}	O_{items}	S_{ITEMS}%	O_{items}%
S1	540	123	417	23	77
S2	609	187	422	31	69
S3	595	196	399	33	67
S4	608	231	377	38	62

TABLE -3
INTERSECTION OF TABLE-1 AND TABLE-2

Symptom	Total test	Common S_{ITEMS}%	Unexpected Test%
S1	100	13	87
S2	100	17	83
S3	100	24	76
S4	100	19	81

Our sample database::

Id	Age	Benefit	Test-1	Test-2	…	Service		Found
						Me	Test	

Fig. 1. Record in diagnostic database

Id	Total Service	Patient	Cost	…..	Sex	Advices	
						Medicine	Test

Fig. 2. Common Practitioner Database

Both practitioners' advice more than 50% patients which is meaning less (special consideration by expert). Same way we found tendency of writing medicine from individual companies (e.g. say he/she advice to take 5 kinds of medicine these from several companies). We assume that threshold_of_benefits_from_company >> threshold_of_user_satisfaction .Which is strongly ignored by the modern society.

Neural Segmentation: Neural Segmentation is a pattern detection algorithm, in which the base technology is a self organizing feature map [11]. Self organization feature mapping, also known as topological feature maps or loosely preserve topology of the multi-dimension space in the two dimensional map [1]. That is similar prototypes near each others.

A self-organizing feature map consists of a two-dimensional array of units; each unit is connected to n input nodes, and contains n dimensional vector w_{ij} where (ij) identifies the unit at location of the array. A neuron computes the Euclidean distance between the input vector x and stored weight vector w_{ij} . The neuron with the minimum distance is declared the "winner" and the input vector is assigned to this neuron. In addition each of the weight vectors is modified as follows. New weight vector=w_{ij}+LR+*NF*(x-w_{ij})
Where,

- LR is the learning rate, a linearly decreasing scalar which changes after each term.
- NF is the neighborhood function, a Guassian distribution function in map space, centered on the winning neuron.

Key success in this analysis is the presentation of behavioral data. Since self organizing data (practitioner's will) is a form of clustering, must be taken into properly balancing inputs. Where balancing inputs implies that each equally-

important aspect of the problem has the same number of vector elements. This overcame by an extensive study, and carefully reviewed the variance of each of the inputs. The output of the algorithm is a two dimensional array of segments, each one described by both its behavior.

7 EXPERIMENT

Description for experiment Behavior Rules:

The behavior rules applied to extract knowledge from the database of general practitioner database and statistical analysis given among table1,2,3.

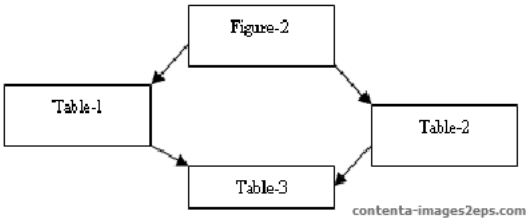

Fig. 3. Data Extraction

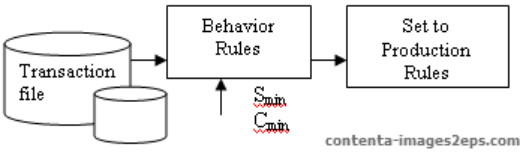

Fig. 4. Application of Fig. 3.

Also short description given below the table with clues, how can any one retrace desired information for next steps? There are three major steps involved when applying this rules:

- Data preprocess (collection raw data, convert to standard format, if need use data manipulation, take input to computer)
- Application to behavior rules (classify data with several dimension, and calculate statistical analysis)
- Post process or analysis of result (knowledge recovery from the table and apply data for decision making)

In this study, the interesting attributes are practitioner's comments though it is difficult to represent in our format. At assigned treatment the most important factor that doctors are not like mention diseases name directly or list of symptoms, is a most important factor for knowledge. By applying technique of Figure-3 we identify characteristics and finally decision making by Figure-4. The inputs include the transaction file (Figure-4), and a name file contains a text description for each code used in the transaction file. Minimum Confidence C_{min} and Minimum Support S_{min} values were specified during the experimentation.

Description for experiment Neural Segmentation:

Neural segmentation was applied to the pathological database services like [1]. The aim was to examine the ordering profiles of General Practitioners (GPs), and determine segments impotence of tests that the GPs ordered. The GPD was used for obtaining data describing the nature of the practice as well as the identification of the selection and frequency of tests. Among the 100 tests for each items (x-ray, blood sugar, urine etc.) the different result rectify the importance of the tests.

Vectors	Test-1	Test-2	Test-3	……	Test –m
V1	1	.5	0	0	0
V2	1	.3	.02	0	1
V3	0	0	1	0	.2
….. ….. …..					
Vn	0	1	.3	0	0

Fig. 5. Input Vector for Neural Network (PD)

Where more than 5 thousand records we categorized firstly) i) Desired Found (DF) and ii) Desired Not Found (DNF), the clinic rectifies that:

Percentage Resolution=DF / DNF%
=37 / (100-37)%
=58.73%

Where our goal 100% but not possible; we considered limitation of a person to imagine perfect with 10% error. But the resolution is not acceptable.

The database was used for creating the input vectors to the segmentation algorithm which mentioned [1]. If 20 test per iteration for 5000 records, a rotation of the records aggregation by GP is required. For 5000 vectors, we considered the number of tests performed by each physician was scaled from 0 to 1 with respect to the total number of tests. The final format is given to Figure-5. Where n=5000 and m=20.

The system ran by presenting vectors vi (for i=0 to 5000) for 250 iterations. The parameters used here:

- Learning rate (LR) with an iteration value 0.6 to 0.05; the learning rate was decreased linearly after each iteration.
- Neighborhood function (NF) with the width of the Gaussians distribution function varied from the square root of the number of nodes to 0.1.

8 RESULTS

In fact, we described our result in the previous sections with consequently theory and experiments. Here we mentioned some important issues, when we increased number rules the satisfaction was decreased for example we apply without rules Boolean conditions satisfied or not, result is mentioned at table-3, unexpected rate is around four times than expected test, but when we applied conditions (rules) the gap is increased, for five rules gap is near about twice time than previous. Another observation the most frequently medical test occurs 38.5% for all iteration and 30% fee for reference-practitioner. The tendency of claimed another issue where if test-1 is claimed with test-2, there was a 85.8% chance that a test-1 also be claimed. This finding may be indicative of different ordering habits for similar clinical

situation. Another case, if test-3 was ordered with test-4, it was relevant with a 65% chance that test-1 also be ordered in 2.5% of cases. This rules arises questions on whether this testing is a screening nature.

Noise in Data and way of reduction:

Several combination of items with medical services centers several non meaning full or not important for my experience rules and noise in the results, which observed by us. These are complicated to extract and analyze. More than 30% raw data are ignored during input to a computer program. It can be reduced if we can assign a standard format for the practitioners or they prescribed by software base which impossible to think in Bangladesh now-a-days. May be it will possible after some days. But for the researcher of a developed country can easily handle this from software base prescription.

Sometimes it was case of ability of a practitioner to advice something (medicine, test) but it was too negligible for a qualified doctor. Otherwise from statistical analysis it not need essential to consider because huge amount of data automatically reduce error. We considered if data more than thousand for an experiment, say important mistake in the point part (see tables, % portion) is ignored where $S_{items}\% << O_{items}\%$.

For a few case where intersection of both *practitioner-x* and *practitioner-y* is equal to 0 (zero), in general case it was sensitive for health sector we have considered also 100% common in both case a great acceptable case in health sector.

Raw data collection is a hard job from the clinic or doctors' chamber which we performed by the help of some technicians or employer of the organization who are not friendly and delivered incomplete information, this also less than 2% which we used in database.

We applied our algorithms (for both behavior rules and neural segmentation) separately and several times for output, it is true that result was not always same due to the lack of technical support in our normal single machine (Pentium Processor, home configuration). But we think theoretical knowledge is appropriate and possible to apply by threading for parallel system or distributed system for better performance.

9 FUTURE WORK

It is a new wing of data mining application for equation-5, there is a great scope to calculate Satisfaction constant S by using statistical analysis in the real database or mathematical induction. A better algorithm is always acceptable in the field of computer science by reducing time complexity, so this is a window for every researcher for text mining. It also open a new horizon to develop software based practice and government can enjoy optimal tax while user benefited by low price treatment. A multidimensional approach is applicable here; a researcher can implement the neural function with Multiple Choice Multiple Dimension Knapsack (MMKP) Problem, we also going to implement this algorithm for next work by MMKP concept. In addition we like to mention graphical technique can be applied by Convex Hull technique for vector classifying. The calculation or *gap-rate* which is mentioned in paragraph viii is another part to create equation.

10 CONCLUSION

We have addressed the effectiveness of two data mining techniques. We have performed from two views and created link of knowledge for decision making. We have shown that data mining algorithm can be used successfully on large, real customer data with acceptable execution time. In addition we like to mention that, from this algorithm organization can take specific action for better health service. In particular, among the results obtained we can mention the following:

- The study provided a classification of general practitioners in to groups of various sizes which reflect their nature and style of practice. By normalization we can create subgroups for better efficient health service.
- It also provide a relationship among patient, practitioner, diagnostic clinic and pharmaceutical company for development of health treatment with affordable cost from realizing morality of a human being as well as his/her responsibilities for the society.

From this experience, we can realize the importance of large database, and uses of database to extract knowledge (meaning full information) with reasonable effort. After all we have created equation (5) for satisfaction of a user.

REFERENCES

[1] Andreas Holzingers, Regina Geierhofer, Felix Mödritscher and Roland Tatzl on "Semantic Information in Medical Information System: Utilization of Text Mining Techniques to Analyze Medical Diagnoses", Journal of Universal Computer Science, vol.. 14, no. 22 (2008), 3781-3795.

[2] Marisa S. Viveros, John P. Nearhos and Michael J. Rothman on "Applying Data mining Techniques to a Health Insurance Information System", Proceeding of the 22 VLDB conference Mumbai, India, 1996.

[3] Hall, A. Walton, G.: "Information Overloaded within the health care system: A literature review ", Health information and libraries journal, 21, (2004)102-108.

[4] Holzinger, A., Geierhofer, R. and Errath, M.: "Semantic Information in Medical Information System – from data and information to knowledge: Facing Information Overloaded , Proceeding of I-MEDIA '07 and i-semantics '07, Graz, 2007a, 323-330.

[5] C. Matheus, G. Piatetsky-Shapiro, D. McNeill. Selecting and Reporting what is interesting. In *Advances in knowledge Discoveryand Data Mining.* AAAI Press, 1996.

[6] Feldman, R. and Sanger, J.: "The text mining hand book: Advance approaches in analyzing unstructured data", Cambridge University press, Cambridge 2007.

Kamal Uddin Sarker has been serving as a senior lecturer in the department of Computer Science & Engineering, Northern University Bangladesh (NUB). He joined NUB in May 2006 as a Lecturer.
He has research publications in different international journals/conferences like ISESCO, ICECE and University Journals in the field of Bioinformatics and Computatiuonal Biology. In addition he is a member of IEB (M-25042).

Nazmun Nessa Moon has been serving as a lecturer with the department of Computer Science & Engineering, State University of Bangladesh (SUB). She joined SUB in October 2010 as a Lecturer.
She served as a lecturer with the department of Computer Science & Engineering, Darul Ihsan University from September '07 to September '10. At that time, she was a course co ordinator and club-moderator of the Computer & Communication Club (C3). Currently, she continues her M.Sc Thesis work in the field of bio informatics in BUET.

Samsuddin Ahmed has been serving as a lecturer in the department of Computer Science & Engineering, State University of Bangladesh (SUB). He joined at SUB in September 2010 as a Lecturer.
He completed his graduation from University of Chittagong in Computer Science and Engineering. His undergrade thesis was "Handling Uncertainty in Spatial Feature Extraction". He is now working on Data and Image mining techniques.