Sindh Univ. Res. Jour. (Sci. Ser.) Vol.46 (4): 539-546 (2014)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

CLiTA: Discovering Knowledge from Clinical Data

N. A. MAHOTO[++], A. SHAIKH, F. KHUHAWAR*

Department of Software Engineering, Mehran University of Engineering and Technology, Jamshoro

Received on 21st October, 2014 and Revised on 29th November, 2014

Abstract: Healthcare organizations manage patient's disease relevant information in a systemic way, which could be utilized for the purpose of relational analysis. The open issue in the clinical data is its sparseness, which makes it difficult to analyze manually. Data mining techniques are greatly adopted to cope with this problem. This research proposes an approach named as CLiTA (Clinical Treatment Analysis) for transforming raw clinical data into meaningful information. In particular, the aim of this study is to analyze the complaints registered by the patients and their diagnosis carried out by the medical experts. CLiTA exploits well-established data mining techniques (i.e., sequential pattern mining and association analysis) to transform real clinical data into knowledge. The discovered knowledge is evaluated and validated by the medical experts. The experts of the healthcare services may seek and utilize this knowledge for providing the treatment to patients and make the process more effective and purposeful.

Keywords: Knowledge discovery, Healthcare, Pattern analysis, Data mining, Association analysis

1 INTRODUCTION

The current era of modern technology uses various resources and tools to obtain novel and targeted information in several essential areas of scientific research, for instance, healthcare, business, and education (Mahoto, 2013), (Mahoto, Shaikh and Ansari, 2014), (Baralis *et al.,* 2010), (Baralis *et al.,* 2012), (Antonelli *et al.,* 2012), (Antonelli *et al.,* 2013), (Mahoto, Shaikh and Chowdhry, 2015). The desired and new knowledge has been benefited through reputable data mining techniques. Data mining allows us to get the desired information, and is indispensable process of Knowledge Discovery from Databases (KDD). Data mining examines databases and/or sets of evidences of relevant patterns for distinguishing them (Fayyad, Piatetsky-Shapiro and Smyth, 1996).

Data mining has imprudent capability to discover the concealed and expressive patterns in the datasets of medical domain, which are of complicated nature (Antonelli *et al.,* 2013). Data mining techniques with the ability to predict and decide have shown considerable growth in the medical field, particularly in forecasting of several diseases such as cardiovascular, cancer, diabetes and others (Antonelli *et al.,* 2012), (Antonelli *et al.,* 2013), (Nahar *et al.,* 2013).

Medically, one disease may clue to several other linked diseases. For example, heart-block can clue to the incidences of other diseases like hypertension, cardiac-arrest and many others. Data mining can provide benefits to physicians identifying actual treatments and the best practices (Mahoto, Shaikh and Ansari, 2014), (Baralis *et al.,* 2010), (Antonelli *et al.,* 2013). In addition, patients can be facilitated with better and inexpensive healthcare services. Well-established data mining techniques, in particular, association rule mining (Agrawal, Imielinski and Swami, 1993), have increased admiration among medical experts to get insight information and dependency of different physical symptoms (Antonelli *et al.,* 2013).

This study solves the problem of tracing out patterns and correlations from medical data in an efficient way. Epidemiological research has proved successful in definition of patterns. The aims of this research activity is identifying the most frequent complains registered by patients and their diagnostic medical tests carried out by medical experts. Further, most frequent medical pathways actually done by the eye patients. The derived knowledge will help medical experts as well as healthcare personnel to effectively provide treatment procedures. The knowledge can also be beneficial for the care guidelines, since medical science is an evolving field.

The rest of the paper is organized as the following. Section 2 presents the related work carried out in healthcare data. The proposed approach is discussed in Section 3. Section 4 reports experimental results of the proposed approach on clinical data of eye patients as a case study; and finally conclusions are drawn in the Section 5.

[++] Corresponding author. Email: naeemmahoto@gmail.com; Tel.: +92-333-7538991

*Institute of Advanced Research Studies in Chemical Sciences, University of Sindh, Jamshoro, Sindh, Pakistan

2 LITERATURE: DATA MINING TECHNIQUES EXPLOITED IN HEALTHCARE

The use of data mining techniques has been profitable in healthcare domain for the last several years. For example, development of clustering technique to identify clusters and discover information from transactional healthcare dataset (Mahoto, Shaikh and Ansari, 2014). The effectiveness of association rule analysis within dense data and considerable low support values (Smitha and Sundaram, 2012). The association analysis is performed on real data to predict the chances of disease hit area (Smitha and Sundaram, 2012). The study (Nahar *et al.,* 2013) categorizes the data of heart disease on gender basis with the help of different rule mining algorithm, such as Apriori (Agrawal and Srikant, 1994), Predictive Apriori (Scheffer, 2001), and Tertius (Flach and Lachiche, 2001), (Flach, Maraldi and Riguzzi, n.d.). Particularly, the study aims at finding significant risk factors for heart diseases in men and women.

The experiments for detection of tumors in digital mammography were conducted in (Antonie, Zaiane and Coman, 2001) by means of different data mining techniques for finding anomaly. In order to predict the effectiveness of a potential drug for a patient has been worked out. For example, clinical and demographical factors of anti-HIV therapies are presented in (Rosen-Zvi *et al.,* 2008). Heart Disease Prediction System (DSHDPS) (Subbalakshmi, Ramesh and Chinna Rao, 2011) used data mining modeling techniques and Nave Bayes. The DSHDPS applied web-based application, defining questionnaire, to serve as a training tool for training medical students to identify of heart disease patients. The proficient methodology of heart disease and association rules generation is proposed in (Deepika, Shekar and Sujatha, 2011) for predicting of heart attack patterns.

Machine learning approaches have been also remained in focus of the researchers. For example, study in (Wang *et al.,* 2012) searched out heterogeneous temporal clinical event pattern mining, the risks of heart attack during particular drugs taken by the patients are predicted in (Davis *et al.,* 2008). An active way of learning for personalized treatment in homogeneous groups suggesting cluster-analysis based model has been introduced in (Deng, Pineau and Murphy, 2012), which helps for future clinical treatment decisions as an illustration. The uncertainty terms in clinical documents and the query logs of an electronic health record search engine are analyzed in (Hanauer *et al.,* 2013). Based on heterogeneous patient records, similarity measurement is monitored in (Sun *et al.,* 2012), (Wang, Sun and Ebadollahi, 2011), and layout fruitful feedback to experts in terms of distance metric assessment. Neural network based classification approach is also utilized for the diagnosis of breast cancer (Hu *et al.,* 2012).

3 EXPERIMENTAL: CLINICAL TREATMENT ANALYSIS (CLITA) APPROACH

This study focuses on extracting knowledge from clinical data set, which may help clinical staff to improve current guidelines, and understand the possible reasons behind certain adverse condition. To this aim, a novel approach has been proposed referred to as CLiTA (Clinical Treatment Analysis). A big view picture of the proposed approach is depicted in **(Fig. 1)** that discovers knowledge from clinical data set. The details of each block are described in the subsequent sections.

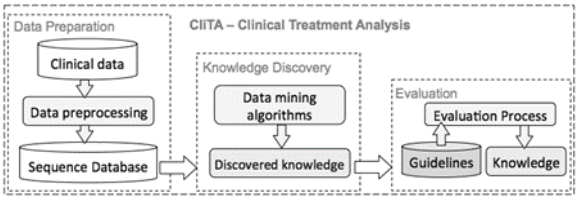

Fig. 1 : Block diagram of Clinical Treatment Analysis (CliTA)

3.1 Data preparation

The clinical dataset (see definition 1) has been collected from Liaquat University Eye Hospital, Hyderabad, Sindh, Pakistan. The six month data has a total of 835 eye records with 656 patients in the year 2010 (January 2010 - June 2010). This data set consists of 391 males and 265 females. Further, the collected dataset is preprocessed to remove unnecessary attributes such as patient identification, address and medical doctors linked with the patient. The data set is then transformed into sequence database (see definition 2). It has been observed that the sequence database contains 14 distinct diagnosis, 6 different complains and 18 different treatments prescriptions that medical experts issued to patients.

Definition 1: *Clinical dataset. Let* $\perp$ *be records of patients* PR*. The clinical dataset* C *corresponding to* $\perp$ *is a set of transactions* C_i*, such that each record of patient* $PR_i \in C$*. Each transaction* $C_I = \{r_1, r_2, r_3, \dots, r_n\}$ *is a set of screening tests or diagnostic tests, complains and prescriptions related to* PR_i*.*

Definition 2: *Sequence database. Let* $\mathcal{P} = \{i_1, i_2, i_3, \dots, i_n\}$ *be the collection of identifiers for patients and* $S = \{s_1, s_2, s_3, \dots, s_n\}$ *be the collection of screening tests or diagnostic tests, complains made by patients and prescriptions given by experts. A sequence database* $\mathcal{D}$ *is set of tuples and each tuple is a pair* (i_{id}, s_i)*, where* $i_{id} \in \mathcal{P}$ *and* $s_i \in S$*.*

3.2 Knowledge discovery

The data mining algorithms used in the study are described in this section. The two algorithms belonging

to prominent techniques of data mining are 1) association rule mining (Agrawal, Imielinski and Swami, 1993), and 2) sequential pattern mining (Agrawal and Srikant, 1995). The CLiTA offers to embed any other data mining algorithm for the knowledge discovery. The considered algorithms, in this study, are described in detail in the following.

Definition 3: *Association rule. Let $\mathcal{X}$ and $\mathcal{Y}$ be two frequent and disjoint items. An association rule, represented as $\mathcal{X} \Rightarrow \mathcal{Y}$, is an implication between $\mathcal{X}$ and $\mathcal{Y}$. The $\mathcal{X}$ is called antecedent and $\mathcal{Y}$ is the consequence of the rule.*

Definition 4: *Closed Association rule. An association rule ($\mathcal{A} \Rightarrow \mathcal{B}$) is said to be closed association rule or simply closed rule, if $\mathcal{A} \cup \mathcal{B}$ is a closed itemset.*

Definition 5: *Association rule support. Let $\mathcal{D}$ be a sequence database, and $\mathcal{X} \Rightarrow \mathcal{Y}$ be an association rule. The support of $\mathcal{X} \Rightarrow \mathcal{Y}$, represented as sup($\mathcal{X} \Rightarrow \mathcal{Y}$) is defined as the frequency of $\mathcal{X} \cup \mathcal{Y}$ in the considered database $\mathcal{D}$.*

Definition 6: *Association rule confidence. Let $\mathcal{D}$ be a sequence database, $\mathcal{X} \Rightarrow \mathcal{Y}$ an association rule. The confidence of the rule$\mathcal{X} \Rightarrow \mathcal{Y}$, represented as conf($\mathcal{X} \Rightarrow \mathcal{Y}$) is given by,*

$$conf(\mathcal{X} \Rightarrow \mathcal{Y}) = \frac{supp(\mathcal{X} \cup \mathcal{Y})}{supp(\mathcal{X})} \quad (1)$$

Where, sup $(\mathcal{X} \cup \mathcal{Y})$ is the frequency of $\mathcal{X}$ and $\mathcal{Y}$ together in the considered database $\mathcal{D}$, and sup$(\mathcal{X})$ is the frequency of item $\mathcal{X}$ in the database.

Definition 7: *Association rule lift. Let $\mathcal{D}$ be a sequence database, $\mathcal{X} \Rightarrow \mathcal{Y}$ an association rule. The lift of the rule $\mathcal{X} \Rightarrow \mathcal{Y}$, represented as conf($\mathcal{X} \Rightarrow \mathcal{Y}$) is given by,*

$$lift(\mathcal{X} \Rightarrow \mathcal{Y}) = \frac{conf(\mathcal{X} \cup \mathcal{Y})}{supp(\mathcal{X})} \quad (2)$$

Where, conf$(\mathcal{X} \cup \mathcal{Y})$ is the confidence of the rule in the database.

The association rule mining (Agrawal, Imielinski and Swami, 1993) is popular data mining technique used for finding associations among collection of items in an information repository or database. Several algorithms are proposed to acquire relationships a set of data (Agrawal, Imielinski and Swami, 1993), (Fournier-Viger, Wu and Tseng, 2012). The problem in association rule (see definition 3) is the number of rules. If there is **d** number of items in a database then number of rules would be $3^d - 2^d + 1$. For instance, if 5 distinct items are available in a database, 212 rules will be generated (Tan, Steinbach and Kumar, 2006). To manage this problem, closed association rule mining concept introduced in (Szathmary, 2006) has been exploited in this study.

3.3 Closed association rule algorithm

The comprehensive form of the correlations, named as closed association rule (see definition 4), are extracted from the considered sequence database.

The closed association rules (Szathmary, 2006) are generated in similar fashion as the conventional association rules are generated in two steps. Firstly frequent closed itemsets are found (see CLOSED-RULES algorithm 1). Secondly, closed rules are generated from the frequent closed itemsets developed in previous step (see RULE-GENERATE algorithm 2). The major difference is that instead of frequent itemsets, closed frequent itemsets are considered in closed rules. The strength of the rules are computed by means of rule support, confidence and lift values (see definitions 5, 6 and 7 respectively).

Algorithm 1 Find closed rules CLOSED-RULES (SD,MINSUPP, MINCONF)

Require: Sequence Database $\mathcal{SD}$

Require: Minimum support $minsupp$

Require: Minimum confidence $minconf$

Ensure: Closed Rules $\mathcal{CR}$

1: **for** each closed frequent k-itemset f_k in $\mathcal{SD}$, $k \geq 2$ and $supp(f_k) \geq minsupp$ **do**

2: $H_1 = \{i | i \in f_k\}\{1 - item\ consequent\ of\ the\ rule\}$

3: **call rule-generate**$(f_k, H_1, minconf)$

4: **end for**

5: **return** $\mathcal{CR}$

Algorithm 2 Generate closed rules RULE-GENERATE $(f_k, H_m, minconf)$

Require: Closed itemsets f_k

Require: m-item consequents of the rule H_m

Require: Minimum confidence $minconf$

Ensure: Closed Rules $\mathcal{CR}$

1: $k = |f_k|\{size\ of\ closed\ frequent\ itemset\}$

2: $m = |H_m|\{size\ of\ rule\ consequent\}$

3: **if** $k > m + 1$ **then**

4: $H_{m+1} = Candidate\ closed\ itemsets\ of\ H_m$

5: **for** each $h_{m+1} \in H_{m+1}$ **do**

6: $conf = \sigma(f_k)/\sigma(f_k - h_{m+1})$

7: **if** $conf \geq minconf$ **then**

8: $output\ the\ rule(f_k - h_{m+1}) \rightarrow h_{m+1}\{Add\ closed\ rule\ in\ \mathcal{CR}\}$

9: **else**

10: $Delete\ h_{m+1}\ from\ H_{m+1}$

11: **end if**

12: **end for**

13: $call\ rule\ -\ generate(f_k, H_{m+1}, minconf)$

14: **end if**

15: **return** $\mathcal{CR}$

Sequential pattern mining (Agrawal and Srikant, 1995) help in understanding the sequences and orders of the items in which they appeared. To understand which diagnostic tests are frequently used in the considered database, BIDE algorithm (Wang and Han, 2004) has been applied.

3.4 BIDE algorithm

The BI-Directional Extension (BIDE) algorithm derives closed frequent sequences from a given sequence database. A sequence $\mathcal{S}$ (see definition 8) is known as closed frequent sequence $\mathcal{S}'$ if there exists no proper super-sequence with the same support as that of $\mathcal{S}$ and it meets the given minimum support criteria. This algorithm mines closed frequent sequences without generating candidates (a sequence that is generated to search out the targeted sequence based on given threshold), contrary to the traditional approaches. The conventional approaches generate large number of candidate sequences. BIDE algorithm, uses bi-directional closure check-up, manages candidate sequences and builds close-packed form of the sequences (i.e., closed frequent sequences). BIDE algorithm is an efficient one and takes less time in generating sequences. However, it has linear property for the size of database in terms of time.

Definition 8: *Sequence. Let* $\mathcal{E} = \{e_1, e_2, e_3, \cdots, e_n\}$ *be the set of events and* $\mathcal{T} = \{t_1, t_2, t_3, \cdots, t_n\}$ *be the corresponding timestamps at which events occurred. A sequence* $\mathcal{S} == \{\{s_1\}, \{s_2\}, \{s_3\}, \cdots, \{s_n\}\}$ *is an ordered temporal relationship among set of events such that* $\mathcal{S}\ \varepsilon\ \mathcal{E}$, *if* $t_1 < t_2 < t_3 < \cdots < t_n$.

Definition 9: *Frequent itemset. Let* $\mathcal{X}$ *be a set of itemsets such that* $\mathcal{X}\ \epsilon\ \tau$, *where* τ *is an enumeration of all itemsets. An item set I is a frequent itemset, if the support of I is higher than a given minsupp (minimum support), where* $I \subseteq \mathcal{X} \subseteq \tau$.

This block exploits closed association rule and BIDE algorithms for discovering knowledge from the clinical data. Nevertheless, several other algorithms can also be applied.

3.5 Evaluation

The evaluation block assess the extracted knowledge in the light of care-guidelines with the help of medical expert. Precisely, mining results are evaluated based on existing care guidelines.

4 EXPERIMENTAL RESULTS

The most frequent medical pathway for the treatment of eye patients are detected and analyzed. Section 4.1 specifically presents the results obtained from the considered dataset using closed association rule algorithm. The results obtained using BIDE algorithm are given in the Section 4.2.

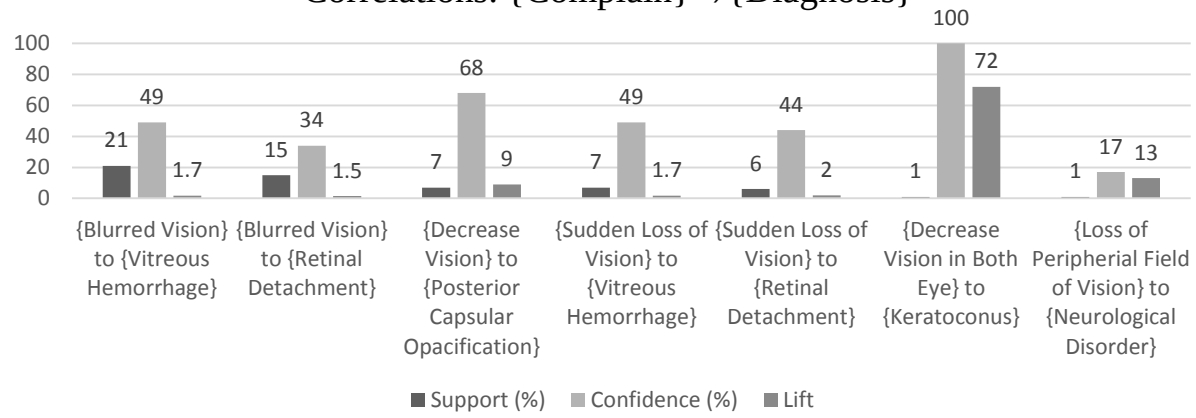

Fig. 2: Frequent Correlations (Complain ⇒ Diagnosis)

4.1 Frequent correlations

The association analysis has been carried out focusing on the complains registered and the diagnosis followed by patients with the discourse of medical

experts. Specially, analysis focuses on correlations between complains and diagnosis. The extracted rules are in the format $< complain > \Rightarrow < diagnosis >$, where antecedent is *complain* registered by patients and consequence is the *diagnosis* performed by medical expert.

(Fig.) and **(Fig.)** presents the most frequent correlation found in the considered clinical dataset. The correlation $Blurred\ Vision \Rightarrow Vitreous\ Hemorrhage$ has (Support 21%, Confidence 49% and Lift 1.7). The lift value higher than 1 reveals that this correlation is positive and strong among the {complain} and {diagnosis}. However, 49% confidence shows that there were 51% of the patients, who didn't follow this correlation. The same complain has been registered but diagnosis has been performed differently in the correlation: $Blurred\ Vision \Rightarrow Retinal\ Detachment$ has (Support 15%, Confidence 34% and Lift 1.5). This presents strong and positive correlation. This different behavior, when similar complaints are registered, medical experts prescribed different diagnostic test due to the fact that medically *Retinal Detachment* is recommended when blood comes from retina. Likewise, *Vitreous Hemorrhage* is medically prescribed when blood doesn't come from retina (WebMD, 2014). Precisely, majority of the patients have been observed complaining vision problem with blood flowing from their eyes. These problems may occur due to environmental pollution, unhealthy food and lack of physical exercises (A.D.A.M., 2014).

Further analysis reveals that the most frequent complains that is **Blurred Vision** has strong association with male gender. For example, $Blurred\ Vision \Rightarrow \{Male\}$ (Support 27%, Confidence 63% and Lift 1.5), whereas the same correlation in female gender has been found as $Blurred\ Vision \Rightarrow \{Female\}$ (Support 16%, Confidence 36% and Lift 0.9). This behavior yields that complain about blurred vision is mostly found in male gender. Similarly, the diagnosis *Retinal Detachment* is positively correlated with female gender. For example, $Retinal\ Detachment \Rightarrow \{Male\}$ (Support 12%, Confidence 57% and Lift 0.96) and $Retinal\ Detachment \Rightarrow \{Female\}$ (Support 9%, Confidence 42% and Lift 1.04).

4.2 Frequent medical pathways

The frequently diagnostic medical tests that are prescribed by medical experts and complains registered by patients are presented in Table 1. Analysis reveals that the complains of *blurred vision* is found most frequently (44%), next *Raised Intraocular Pressure* (23%) and then *Sudden loss of vision* (15%). Whilst, frequent medical pathways, that are followed by patients are *Vitreous Hemorrhage* (29%), *Angle Ocular Glaucoma* (23%) and *Retinal Detachment* (22%). Further, $\{Vitreous\ Hemorrhage\}^2$ is diagnosed twice in (9%) patients. Similarly *Retinal Detachment* is found twice in (5%) patients.

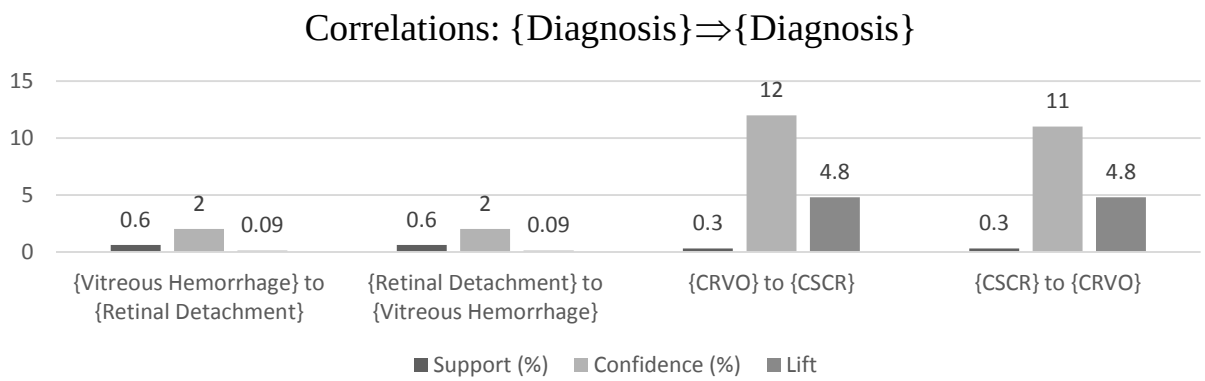

Fig. 3: Frequent Correlations (Diagnosis ⇒ Diagnosis)

Table 1 : Frequent Medical Pathways and Complains

Diagnosis Sequence	**Frequency (%)**
{Vitreous Hemorrhage}	29
{Angle Ocular Glaucoma}	23
{Retinal Detachment}	22
{Vitreous Hemorrhage}x2	9
{Glaucoma}	5
{Retinal Detachment} x2	5
{Diabetic Retinopathy}	3
{Diabetic Retinopathy} x2	2
{Glaucoma} x2	2
Complains	**Frequency (%)**
{Blurred Vision}	44
{Raised Intraocular Pressure}	23
{Sudden Loss of Vision}	15
{Blurred Vision} x2	13
{Decreased Vision}	11
{Sudden Loss of Vision} x2	5
{Blurred Vision} x3	2

Retinal Detachment, also known as *Detached Retina*, happens due to an exceptionally thin or damaged retina allowing eye fluid to enter within it (Medscape, 2014). Retinitis pigmentosa is an eye disease in which the retina is damaged and it is an inherited disorder that can lead to central vision loss (NCBI, 2014).

The frequent complains found in female gender are *blurred vision* (40%), *Raised Intraocular Pressure* (23%) and *Sudden Loss of Vision* (20%). Whilst, the complains registered by male gender varies in frequency, for example, *blurred vision* (47%), *Raised Intraocular Pressure* (22%) and *Sudden Loss of Vision* (12%). Similarly, the diagnostic tests are found as, *Vitreous Hemorrhage* (27% male and 30% female), *Angle Ocular Glaucoma* (22% male, 23% female), and *Retinal Detachment* (21% male, 23% female).

4.3 Performance Evaluation

To verify the feasibility of the CliTA, analysis has been performed on a machine with Pentium 4 at 2.26 GHz core 2 Duo with 4 GB RAM (Random Access Memory). The implementation of the algorithms in java programming language are kindly provided by Phillippe Fournier-Viger (Fournier-Viger, 2014).

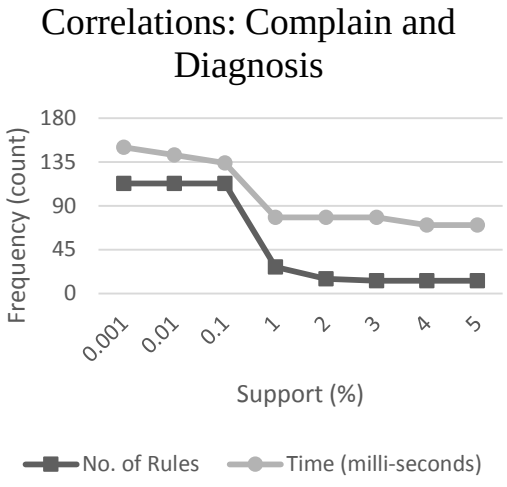

Fig. 2 : No. of Correlations and Timing versus Support

(Fig. 2) shows the timings and number of rules (i.e., correlations) extracted with respect to various support values. Note that the *y-axis* represents the Frequency (count) and *x-axis* represents the support in percentage. The number of rules are inversely proportional to rule support (i.e., *minsupp*) and rule confidence (i.e., *minconf*) values. The higher support and confidence values produce less number of rules and vise-verse. Similarly, more time is needed for lower support values. Thus time and number of rules are inversely proportional to the support value.

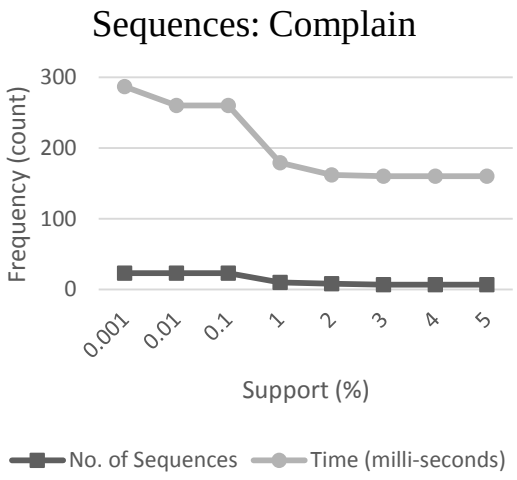

Fig. 3 : No. of sequences and Timing versus Support (Complain)

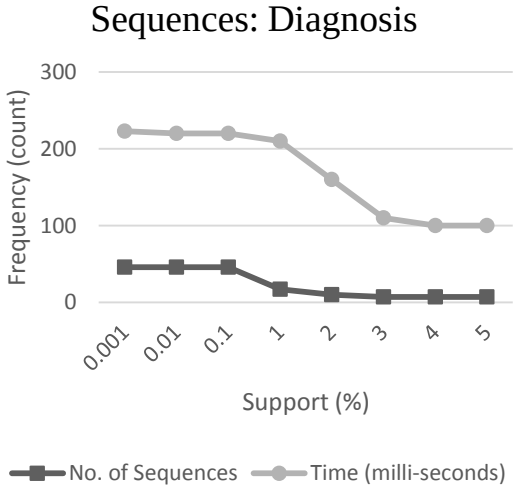

Fig. 4 : No. of sequences and Timing versus Support (Diagnosis)

(Fig. 3) and **(Fig. 4)** illustrate the trends of number of sequences and timings when support values are varying from 0.001% to 5%. Here again the *y-axis* represents Frequency (count) and *x-axis* represent the support in percentage. It is quite understandable that as the support value increases, number of sequences and time is decreased. This trend is due to the fact that lesser support value generates more sequences and vice-versa.

5 CONCLUSION AND FUTURE WORK

To make the most use of large and complex medical databases for deriving productive knowledge in a timely manner, innovative data analysis techniques are required. In this paper, a novel data analysis approach named CliTA (Clinical Treatment Analysis) is proposed for extracting useful information from raw clinical

dataset. The approach makes use of closed association rule mining and BIDE algorithms to determine the useful knowledge. However, several other algorithms can also be injected. The CliTA has been validated based on a real clinical dataset of eye patients. The complaints of patients and their diagnosis by medical experts through association analysis have been in focus of this research. Especially, the correlation between complains and diagnosis has been monitored. Further, most frequent complains and diagnosis were analyzed gender-wise. The discovered information may be utilized for the resources management as well as for updating available care guidelines. Furthermore, the extracted information also indicated the effects of environmental changes that may have caused such problems in the vicinity of LUMHS Hospital Hyderabad, Pakistan. Therefore, authorities should make public awareness regarding these problems in order to improve public health and build healthy societies.

The proposed approach is planned to apply on different clinical datasets (e.g., skin diseases, diabetic, and cancer patients). Moreover, extracting temporal patterns will be considered as future work.

6 ACKNOWLEDGEMENT

The authors would like to thank Dr. Nazir Ahmed Agha (Liaquat University of Medical and Health Sciences (LUMHS) Hospital Hyderabad, Sind, Pakistan) for his support and for serving as domain expert. The authors are also thankful to Mehran Ali Memon and LUMHS personnel who helped in collecting data.

REFERENCES:

A.D.A.M. (2014) *Health Guide*, Available: http://www.nytimes.com/health/guides/symptoms/vision-problems/overview.html.

Agrawal, R., T. Imielinski, and A. Swami, (1993) 'Mining association rules between sets of items in large databases', *SIGMOD Rec.*, vol. 22, no. 2, 207-216, Available: http://doi.acm.org/10.1145/170036.170072

Agrawal, R. and R. Srikant, (1995) 'Mining Sequential Patterns', Proceedings of the Eleventh International Conference on Data Engineering, Washington, DC, USA, 3-14.

Agrawal, R., and R. Srikant, (1994) 'Fast algorithms for mining association rules', Proc. 20th int. conf. very large data bases, VLDB, 487-499.

Antonelli, D., E. Baralis, G. Bruno, T. Cerquitelli, S. Chiusano, and N. Mahoto, (2013) 'Analysis of diabetic patients through their examination history', *Expert Systems with Applications*, vol. 40, 4672-4678, Available: http://www.sciencedirect.com/science/article/pii/S0957417413001206.

Antonelli, D., E. Baralis, G. Bruno, S. Chiusano, N.A. Mahoto, and C. Petrigni, (2012) 'Analysis of diagnostic pathways for colon cancer', *Flexible Services and Manufacturing Journal*, vol. 24, 379-399, Available: http://dx.doi.org/10.1007/s10696-011-9095-2.

Antonie, M.L., O. R. Zaiane, and A. Coman, (2001) 'Application of Data Mining Techniques for Medical Image Classification.', *MDM/KDD*, vol. 2001, 94-101.

Baralis, E., G. Bruno, S. Chiusano, V.C. Domenici, N.A. Mahoto, and C. Petrigni, (2010) 'Analysis of medical pathways by means of frequent closed sequences', Proceedings of the 14th international conference on Knowledge-based and intelligen information and engineering systems: Part III, Berlin, Heidelberg, 418-425.

Baralis, E. M., S. A. Chiusano, N. A. Mahoto, D. Antonelli, G. Bruno, and C. Petrigni, (2012) 'Extraction of medical pathways from electronic patient records', in *Medical Applications of Intelligent Data Analysis: Research Advancements / Rafael Magdalena-Benedito, Emilio Soria, Juan Guerrero Martinez, Juan Gomez-Sanchis, Antonio Jose Serrano-Lopez*, Information Science Reference (IGI Global).

Davis, J., E. Lantz, D. Page, J. Struyf, P. Peissig, H. Vidaillet, and M.Caldwell, (2008) 'Machine learning for personalized medicine: Will this drug give me a heart attack', Proceedings of International Conference on Machine Learning (ICML).

Deepika, N., K. C. Shekar, and D. Sujatha, (2011) 'Association rule for classification of Heart-attack patients', *International Journal of Advanced Engineering Sciences and Technologies (IJAEST)*, vol. 11, no. 2,. 253-257.

Deng, K., J. Pineau, and S.A. Murphy, (2012) 'Active learning for developing personalized treatment', Liu, Y. (Chair), Recent advances in statistical machine learning, San Diego,

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth, (1996) 'From data mining to knowledge discovery in databases', *AI magazine*, vol. 17, no. 3, 37Pp.

Flach, P. A. and N. Lachiche, (2001) 'Confirmation-guided discovery of first-order rules with Tertius', *Machine Learning*, vol. 42, no. 1-2, 61-95.

Flach, P., V. Maraldi, and F. Riguzzi, (n.d) 'Algorithms for Efficiently and Effectively Using Background Knowledge in Tertius', Available: http://www.cs.bris.ac.uk/Publications/Papers/2000629.pdf.

Fournier-Viger, P. (2014) *A Sequential Pattern Mining Framework*, Available: http://www.philippe-fournier-viger.com/spmf/index.php.

Fournier-Viger, P., C.W. Wu,. and V.S. Tseng. (2012) 'Mining top-k association rules', in *Advances in Artificial Intelligence*, Springer.

Hanauer, D.A., Q. Mei, B. Malin, and K. Zheng, (2013) 'Location Bias of Identifiers in Clinical Narratives', AMIA Annual Symposium Proceedings, 560.

Hu, J., F. Wang, J. Sun, R. Sorrentino, and S. Ebadollahi, (2012) 'A healthcare utilization analysis framework for hot spotting and contextual anomaly detection', AMIA Annual Symposium Proceedings, 360.
Mahoto, N.A. (2013) *Data mining techniques for complex application domains*, PhD thesis in information and system engineering, Politecnico di Torino, Italy.

Mahoto, N.A., F.K. Shaikh, and A.Q. Ansari, (2014) 'Exploitation of Clustering Techniques in Transactional Healthcare Data', *Mehran University Research Journal of Engineering and Technology*, vol. 33, no. 1, 77-92.

Mahoto, N.A., F.K. Shaikh, and B. Chowdhry, (2015) 'Innovative architecture to enhance quality of service for laboratory management information systems', In Laboratory Management Information Systems: Current Requirements and Future Perspectives/Anastasius Moumtzoglou, Anastasia Kastania and Stavros Archondakis, 237-251.

Medscape (2014) *Medscape from WebMD Health Professional Network*, Available: http://emedicine.medscape.com/article/1230216-overview.

Nahar, J., T. Imam, K.S. Tickle, and Y.P.P. Chen, (2013) 'Association rule mining to detect factors which contribute to heart disease in males and females', *Expert Systems with Applications*, vol. 40, no. 4, 1086-1093.

NCBI (2014) *U.S. National Library of Medicine (NLM)*, Available: http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0002024/.

Rosen-Zvi, M., A. Altmann, M. Prosperi, E. Aharoni, H. Neuvirth, S, Sch, D. Struck, Y. Peres, F. Incardona, and others (2008) 'Selecting anti-HIV therapies based on a variety of genomic and clinical factors', *Bioinformatics*, vol. 24, no. 13, i399-i406.

Scheffer, T. (2001) 'Finding association rules that trade support optimally against confidence', in *Principles of Data Mining and Knowledge Discovery*, Springer.

Smitha, T. and V. Sundaram, (2012) 'Association models for prediction with Apriori concept', *International Journal of Advances in Engineering \& Technology*, vol. 5, no. 1,354-360.

Subbalakshmi, G., K. Ramesh, and M. C. Rao, (2011) 'Decision support in heart disease prediction system using naive bayes', *Indian Journal of Computer Science and Engineering (IJCSE)*, vol. 2, no. 2, 170-176.

Sun, J., F. Wang, J. Hu, and S. Edabollahi, (2012) 'Supervised patient similarity measure of heterogeneous patient records', *ACM SIGKDD Explorations Newsletter*, vol. 14, no. 1, 16-24.

Szathmary, L. (2006) 'Symbolic Data Mining Methods with the Coron Platform', *These d'informatique, Universit Henri Poincar--Nancy*, vol. 1.

Tan, P.N., M. Steinbach, and V. Kumar, (2006) *Introduction to Data Mining, (2nd Edition)*, 2nd edition, Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Wang, J. and J. Han, (2004) 'BIDE: Efficient mining of frequent closed sequences', Data Engineering, 2004. Proceedings. 20th International Conference on, 79-90.

Wang, F., J. Sun, and S. Ebadollahi, (2011) 'Integrating distance metrics learned from multiple experts and its application in patient similarity assessment', SIAM.

Wang, F., N. Lee, J. Hu, J. Sun, and S. Ebadollahi, (2012) 'Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach', Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 453-461.

WebMD (2014) *Eye Health Center*, Available: http://www.webmd.com/.