Analysis of Health Care Data Using Different Data Mining Techniques

Anjana Gosain
Dept. of Computer Science
USIT, GGSIPU
Delhi, India
anjana_gosain@yahoo.com

Amit Kumar
Dept. of Information Technology
USIT, GGSIPU
Delhi, India
amitkrranjan@gmail.com

Abstract—**Data mining is an interesting field of research whose major objective is to acquire knowledge from large amounts of data. With advances in heath care related research, there is a wealth of data available. However, there is a lack of effective analytical tools to discover hidden and meaningful patterns and trends in data, which is essential for any research.**

In recent years, human immune-deficiency virus (HIV) related illnesses have become a threat to the modern world. Researchers all over world, including India, are trying hard to find suitable answer to this and this led to lots of research in the field. Therefore, a tool which can process data in meaningful way is the need the time.

In this study, we briefly examine the potential use of classification based data mining techniques such as decision tree and association rule to massive volume of health care data. Further we developed a prototype/approach that is specially designed to monitor the patients receiving antiretroviral therapy (ART). As monitoring of individual is not a difficult task however deriving inferences from a large cohort and then use this information for future guidelines need this kind of prototype/approach. We expect, this would have great impact in current management and future strategies against HIV.

Keywords- decision tree, association rule, human immunedefiancy virus, Antiretroviral therapy

I. INTRODUCTION

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner [1].

There are several data mining techniques, such as, Decision Trees, Artificial Neural Networks (ANNs), Clustering, Naive Bayes, Association Rules, Time series etc[2] but decision tree is one of the more powerful technique that produce interpretable result and thus widely used in clinical purpose[3].

Hence, decision tree may be considered mining knowledge from large amounts of data since it involves knowledge extraction, as well as data/pattern analysis in tree diagrams [4].

Considering decision trees, classification is the most common data mining task and it consists of examining the features of a newly presented object in order to assign it to one of a predefined set of classes. Classification deals with discrete outcomes, estimation deals with continuously-valued outcomes.

In this study, our objective are to: (1) present an evaluation of techniques such as decision tree and association rules to predict the occurrence of route of transmission based on treatment history of HIV patients. (2) demonstrate that data mining method can yield valuable new knowledge and pattern related to the HIV patient; (3) assesses the utilization of healthcare resources and demonstrate the socioeconomic, demographic and medical histories of patient.

Organization of paper is as follows; in section II we give a brief explanation of the data mining techniques. In section III we give a brief explanation of the data mining life cycle. In section IV we give a brief explanation of the case study of HIV patient. In section V we analyze of HIV patients using decision tree and association rules. Result and conclusion are discussed in Section VI and section VII.

II. DATA MINING TECHNIQUES

A. *Decision Trees*

The decision tree is probably the most popular data mining technique. The most common data mining task for a decision tree is classification [5].

The principle idea of a decision tree is to split data recursively into subsets so that each subset contains more or less homogeneous states of target variable (predictable attribute). At each split in the tree, all input attributes are evaluated for their impact on the predictable attribute. When this recursive process is completed, a decision tree is formed.

There are a few advantages of using decision trees over using other data mining algorithms, for example, decision trees are quick to build and easy to interpret. Each path from the root to a leaf forms a rule. Prediction based on decision trees is explainable and efficient.

Microsoft Decision Trees is a hybrid decision tree algorithm developed by Microsoft research. It supports classification and regression tasks. One of the unique features of Microsoft Decision Trees is that it can also be applied for association analysis [6].

B. *Association Rules*

Association rules are one of the major techniques of data mining and it is perhaps the most common form of local-pattern discovery in unsupervised learning systems. It is a form of data mining that most closely resembles the process that most people think about when they try to understand the data-mining process; namely, "mining" for gold through a vast database. The gold in this case would be a rule that is interesting, that tells you something about your database that we didn't already know and probably weren't able to explicitly articulate. These methodologies retrieve all possible interesting patterns in the database. This is strength in the sense that it leaves no stone unturned [7].

III. Data Mining Life Cycle

This model describes the process of a Data Mining project's life cycle. It includes six different phases and each phase consists of generic task and process instances [6]. Crisp-DM does not describe a particular data mining technique; rather it focuses on the process of a data mining project's life cycle [2].

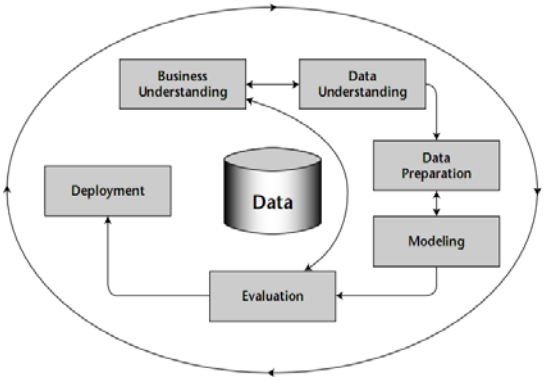

Figure 1. CRISP-DM Reference Model.

A. *Business Understanding*

This initial phase focuses on understanding the objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.

B. *Data Understanding*

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

C. *Data Preparation*

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

D. *Modeling*

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values.

E. *Evaluation*

At this stage we have built models that appear to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives.

F. *Deployment*

Creation of the model is generally not the end of the research. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process.

IV. Case Study Description

Significant research efforts have been undertaken in investigating the association between HIV and ART. We have considered HIV database (HIVDB) which is one of the deadly diseases all over the world including India.

The following tests and procedures may be used:

- Physical exam and history
- Complete blood count (CBC)
- Lymphadenopathy
- Stages of infection
- Blood chemistry studies
- Chest x-ray
- Stage

A. *Sources of Analytical Data*

ART Clinic, All India Institute of Medical Sciences (AIIMS), New Delhi.

B. *Methods of Data Collection*

We developed software named ART System in All India Institute of Medical Sciences and patient data are stored through the ART system and further we collect it for research work.

We derived a dataset from HIV Database (HIVDB) that included 1054 enrolled patients out of which we have considered only 672 unique patient because rest of patients are defaulter.

C. *Data Cleaning*

Real world data, like data acquired from ART Clinic, AIIMS, tend to be incomplete, noisy and inconsistent. Data cleaning routines attempt to fill on missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

D. *Missing Values*

Most datasets contain missing values. There are a number of causes for missing data. Many methods were applied to solve this issue depending on the importance of the missing value and its relation to the search domain.

- Fill in the missing value manually
- Use a global constant to fill in the missing value
- Replace the missing values with the most popular
- Value (constant).

AMIT-PC.ARTClinic - dbo.ARTBase1

BWt	Ht	BMI	Marital
7	0.620000...	18	2
65	0	0	1
48	1.649999...	17	1
44	1.559999...	18	1
41	1.659999...	14	1
69	1.730000...	23	1
60	0	0	1
20	0	0	2
60	0	0	1
10	1	10	2
70	1.730000...	23	1

Figure 2. Missing Values

E. *Noisy Data*

Noise is a random error or variance in a measured variable. Many techniques were used to smooth out the data and remove the noise.

F. *Data Integration*

Data Mining often requires data integration, the merging of data from multiple data sources into one coherent data store. These sources include in our case ART database. Careful integration of the data from multiple sources helped reducing and avoiding redundancies and inconsistencies in the resulting data set. This helped improving the accuracy and speed of the subsequent mining process.

G. *Data Selection*

Selecting fields of data of special interest for the search domain is the best way to obtain results relevant to the search criteria. In this paper HIV care was the aim, so data concerning the diagnosis of HIV are carefully selected from the overall data sets, and mining techniques were applied to these specific data groups in order to reduce the interesting patterns reached to the ones that represent an interest for the domain.

Main Focus on following data

- Age
- Sex
- Marital Status
- Route of transmission
- Body Weight
- HAART Regime
- TLC
- DLC
- Hemoglobin

V. Data Mining within Business Intelligence

In health care application, Microsoft business Intelligence SQL Server 2008 analysis service offers a strategic approach to finding useful relationships in large data sets. In contrast to more traditional statistical methods, we do not necessarily need to know what we are looking for when we start. We can explore our data, fitting different models and investigating different relationships, until we find useful information.

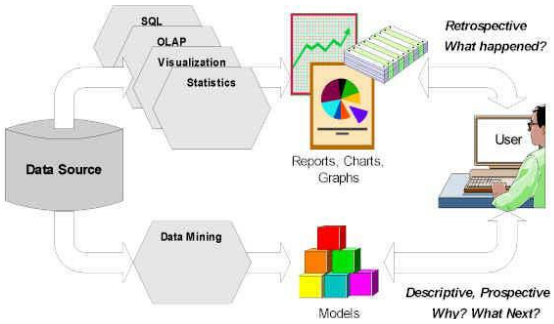

Figure 3. Microsoft Business Intelligence Process

A. *Creating the Data Source to use*

In our case a Microsoft SQL Server 2008 database is used and data is stored into different tables and logical table i.e. view, ODBC was used to link the data source by using the SQL Server Analysis Service.

B. *Viewing Data*

After establishing connection by analysis service to the data source, we view the data from the database in a tabular form by linking the data source to a "table" output.

ARTCIno	Age	Gender	Preg	BreastFeed	BMI	Marital	Edu	Route	Routetrans
596	1	2	0	0	18	2	1		2
597	40	1			0	1	6		1
598	29	1			17	1	5		1
599	24	2	0	0	18	1	2		1
6	45	1			14	1	5		1
60	55	1			23	1	6		1
600	30	1			0	1	6		1
601	8	2	0	0	0	2	1		2
602	26	1			0	1	2		1
603	3	1			10	2	1		2

Figure 4. View Data in Tabular Form

This enables us to assure that the link is successfully built and let us take a look on the form of data read by the software

to detect any loss, inconsistency or noise. Following database diagram represents our HIVDB data warehouse.

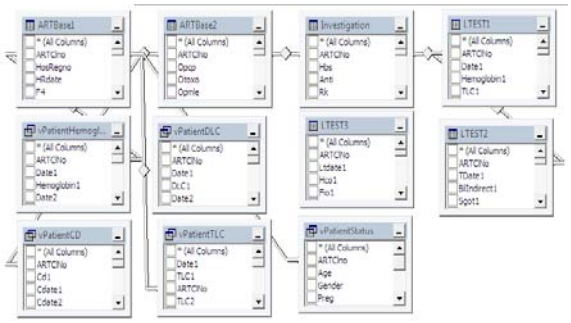

Figure 5. Relational Database HIV

C. *Manipulating Data*

Using the data mining technique we were focus on specific fields that allows us to explore the data deeper, by selecting and filtering some fields as input, output fields and predictive fields.

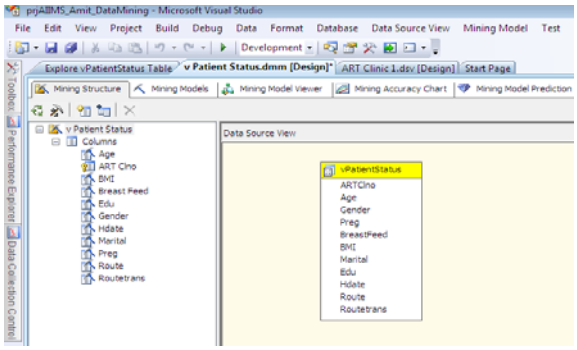

Figure 6. Field operation

D. *Data Evalution*

After applying the data mining techniques comes the job of identifying the obtained results, in form of interesting patterns representing knowledge depending on interestingness measures. These measures are essential for the efficient discovery of patterns of value to the given user. Such measures can be used after the data mining step in order to rank the discovered patterns according to their interestingness, filtering out the uninteresting ones. More importantly, such measures can be used to guide and constrain the discovery process, improving the search efficiency by pruning away subsets of the pattern space that do not satisfy pre-specified interestingness constraints.

In diagram given below we can see that only 779 patient are using HAART regimes and rest of the patient are defaulter or not eligible.

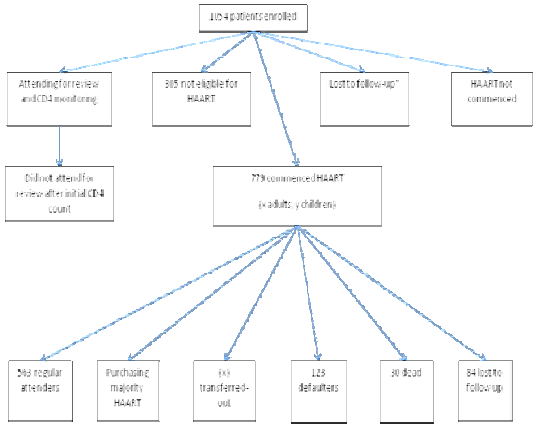

Figure 7. Flow diagram showing disposition of patients enrolled at 1 year

VI. Results and Discussion

In this study, we analyze data on Route of transmission in the HIVDB and investigate their association with patient's history of ART. Decision tree is first applied to the database for identifying patterns with relatively high support and confidence. Further Association rule is applied to predict the changes which are occurring frequently among people. This will be very beneficial to remove deadly disease. On this basis of this study government can take initiative to conduct awareness programme.

A. *Case I : Decision Trees*

In HIVDB we have used certain code for route of transmission and code of that is given in table1.

TABLE I. CODE FOR ROUTE OF TRANSMISSION

Type of transmission	Value
Heterosexual	1
Mother-to-Child	2
Blood Transfusion	3
IVD	4
Surgical Instrument	5
Unsafe Injection	6
Professional needle stick Injury	7
Homosexual	8
Unknown	9

Decision trees given below clearly represnts the age of people who are suffering from this deadly disease.

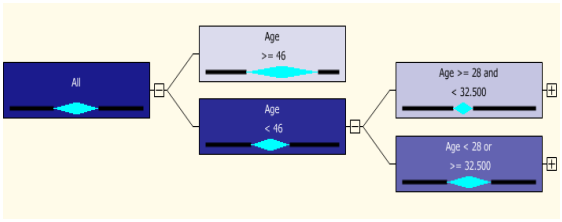

Figure 8. Decision trees

We identified different result by using decision tree techniques. Each level clearly depicts the case of route of transmission. There are 55 HIV case found on age 46 and above and route of transmission found are due to blood transfusion and there are 413 case found on age 45 below and route of transmission found are heterosexual or Mother –to-child. This result shown in table II.

TABLE II. CONCLUSION OF ROUTE OF TRANSMISSION

Level	Outcome
1	All Existing Cases: 468

	Missing Cases: 0 Routetrans = 1.547
2-1	Age >= 46 Existing Cases: 55 Missing Cases: 0 Routetrans = 2.255
2-2	Age < 46 Existing Cases: 413 Missing Cases: 0 Routetrans = 1.447-0.030*(Age-30.722)
2-2-3	Age >= 28 and < 32.500 Existing Cases: 119 Missing Cases: 0 Routetrans = 1.164+0.093*(Age-30.252)
2-2-4	Age < 28 or >= 32.500 and < 46 Existing Cases: 294 Missing Cases: 0 Routetrans = 1.577-0.031*(Age-30.912)

Now we have considered the gender of the people who are suffered from HIV and we have used certain code for gender in ART system given in table IV. From graph shown below we could see that about 500 male have been found HIV case but only 172 female patients have HIV. So it is advisable to focus on male. Indian government should take initiative and conduct awareness program among male.

TABLE IV: CODE FOR GENDER

Attribute	Value
Male	1
Female	2
Eunuch	3

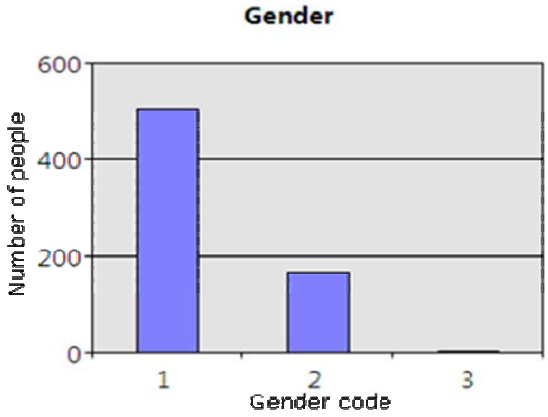

Figure 9. HIV Distribution among gender

Similarly now, we have considered the pregnancy case and for this we also used certain code for pregnancy in ART system given in table V. No such case of HIV found in female who is pregnant so chances of mother –to-child transmission is minimal. missing represents male.

TABLE V: CODE FOR PREGNANCY

Attribute	Value
Yes	1
No	0

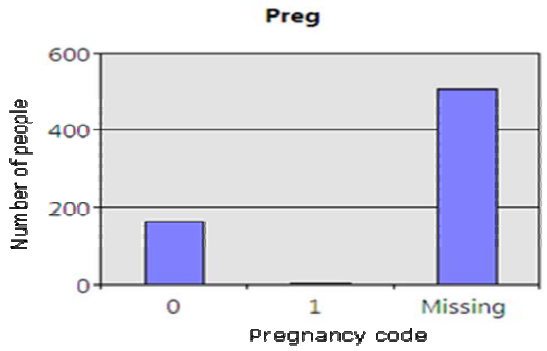

Figure 10. HIV Distribution in female

Most importantly to find out the education of people who are suffering from HIV and for this we have considered the education case and for this we also used certain code for education in ART system given in table VI. Most of the cases of HIV are found in either illiterate or less educated people.

TABLE VI: CODE FOR EDUCATION

Attribute	Value
Illiterate	1
Primary school or literate	2
Middle school completion	3
high school certificate	4
Intermediate or higher secondary	5
Bachelor degree	6
Professional, Postgraduate & above	7

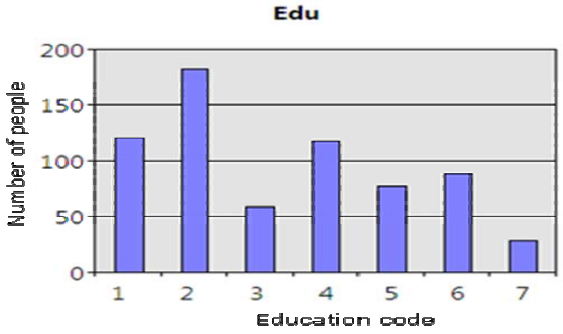

Figure 11. HIV Distribution also depend upon their education

Indian government should take initiative to educate the people so that this deadly disease could completely remove from our country. HIV is not a curable disease and it removed only through awareness program.

B. Case II : Association Rule

Microsoft Association Rules were generated with the default parameters settings [8]. The figure given below represents the classification matrix for the association rules. All of the instances from the *RouteOfTransmission* class have been misclassified. Good results were yielded for the *Education* and *Age* classes, i.e. most of the predictions were correct.

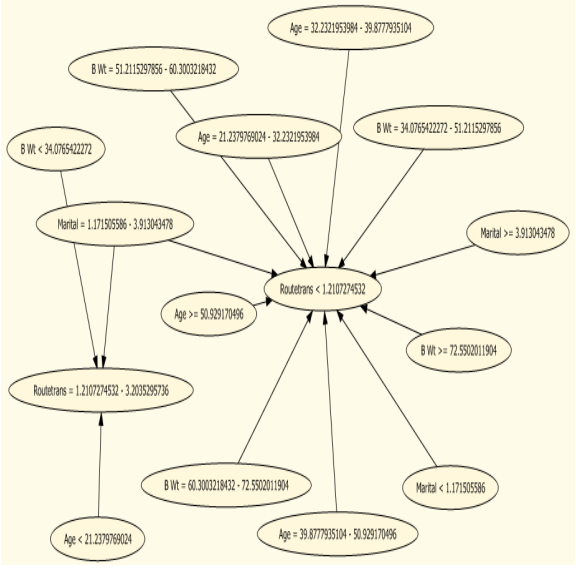

Figure 12. association rule dependency network

Support	Size	Itemset
388	1	Routetrans < 1.2107274532
370	1	Marital < 1.171505586
334	2	Marital < 1.171505586, Routetrans < 1.2107274532
161	1	Age = 21.2379769024 - 32.2321953984
160	1	B Wt = 51.2115297856 - 60.3003218432
155	1	B Wt = 34.0765422272 - 51.2115297856
149	2	Age = 21.2379769024 - 32.2321953984, Routetrans < 1.2107274532
148	1	Age = 32.2321953984 - 39.8777935104
147	2	B Wt = 51.2115297856 - 60.3003218432, Routetrans < 1.2107274532
139	2	B Wt = 34.0765422272 - 51.2115297856, Routetrans < 1.2107274532
138	2	Age = 32.2321953984 - 39.8777935104, Marital < 1.171505586
137	2	Age = 32.2321953984 - 39.8777935104, Routetrans < 1.2107274532
135	2	B Wt = 34.0765422272 - 51.2115297856, Marital < 1.171505586
134	2	B Wt = 51.2115297856 - 60.3003218432, Marital < 1.171505586
127	3	Age = 32.2321953984 - 39.8777935104, Marital < 1.171505586, Routetrans < 1.2107274532
124	3	B Wt = 34.0765422272 - 51.2115297856, Marital < 1.171505586, Routetrans < 1.2107274532
124	2	Age = 21.2379769024 - 32.2321953984, Marital < 1.171505586
124	3	B Wt = 51.2115297856 - 60.3003218432, Marital < 1.171505586, Routetrans < 1.2107274532
117	3	Age = 21.2379769024 - 32.2321953984, Marital < 1.171505586, Routetrans < 1.2107274532
85	1	Age = 39.8777935104 - 50.929170496

Figure 13. Rule is generated using association rule for pattern analysis

VII. CONCLUSION

1. Decision Trees and association rule are very important data mining technique, in the process of knowledge discovery in the medical field, especially in the domains where available data have many limitations like inconsistent and missing values.

2. Identifying predictors of infrequent events is an important but difficult task in studying clinical databases. We have addressed these challenges by using association rule to find interesting pattern involving rare events and classification trees to establish high order association among those events.

3. Using better quality of data influences the whole process of knowledge discovery, takes less time in cleaning and integration, and assures better results from the mining process.

In future, data mining techniques, when combined with traditional statistical analyses, provide powerful new methods to elucidate and provide insight into undiscovered knowledge embedded in clinical data. However, it is important to recognize that the successful application of data mining techniques requires that domain experts, in this case healthcare professionals, work closely with the data mining expert to develop analyses that are relevant to clinical decision making.

ACKNOWLEDGMENT

I am very grateful to Prof and Head Dr S.K.Sharma Department of Medicine, All India Institutes of medical Sciences (AIIMS), New Delhi, India to permit me to develop an ART system. My special thanks to Dr Vijay Hadda, SR, AIIMS for his invaluable suggestion.

REFERENCES

[1] David Hand, Heikki Mannila, and Padhraic Smyth, Principles of Data Mining,MIT Press, Cambridge, MA, 2001.

[2] Z. Tang and J. MacLennan., “Data Mining with SQL Server 2008”,USA 2008.

[3] M. Berry and S. Gordon, "Data Mining Techniques: For Marketing, Sales, and Customer Support", May 1997.

[4] Han, J. and Kamber, M., "Data Mining Concepts and Techniques", 2001.

[5] Tang Z., MacLennan J., Data Mining with SQL Server 2005, USA, Wiley Publishing Inc., 2005.

[6] M. Negnevitsky, "Artificial Intelligence, A Guide to Intelligent Systems", England: Pearson Education Limitd, 2002.

[7] Mehmed Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons, 2003.

[8] Jiang Y. et. al., The potential of Computer-Aided Diagnosis (CAD) to reduce variability in radiologists' interpretation of mammograms, Academin Radiology, 10 (8), Elsevier, 2003.

[9] www.sqlserverdatamining.com.

[10] www.sqlservercentral.com.