An Efficient Approach of Providing Rattle Knowledge as A Service for Non Expert Data Miners

Shivanand ganaagi[1],
2nd Year M.Tech. Student Dept. of CSE,
BTL Institute of Technology & Management
Bangalore-562125, Karnataka, India

S Basavaraj Patil[2]
[2]Prof HOD. Dept. of CSE
[2]BTL Institute of Technology & Management
Bangalore-562125, Karnataka, India

Abstract- **The process of data mining is to extract information from a data set and transform it into an understandable structure. Due to the massive amount of information embedded in huge data warehouse maintained in several domains, the extraction of meaning full patterns is no longer feasible. the issue turns to more obligatory for developing several tool in data mining software is to build a resourceful predictive model for handling the more amount of information in a more efficient manner , Data mining mainly contrast with excessive collection of data that inflicts huge rigors computational constraint. Theses out coming challenges lead to emerge of powerful data mining technologies data mining tools are exemplified and also contras and also they are very expensive, in this paper it is proposed to design a knowledge based tool known as Application tool.**

Key words – Application(GUI)

I. INTRODUCTION

The domain of data mining and discovery of knowledge in various research fields such as Pattern Recognition, Information Retrieval, Medicine, Image Processing, Spatial Data Extraction, Business and Education has been tremendously increased over the certain span of time. Data Mining highly endeavors to originate, analyze, extract and implement fundamental induction process that facilitates the mining of meaningful information and useful patterns from the huge dumped unstructured data. This Data mining paradigm mainly uses complex algorithms and mathematical analysis to derive exact patterns and trends that subsists in data. The main aspire of data mining technique is to build an effective predictive and descriptive model of an enormous amount of data. Several real world data mining problems involves numerous conflicting measures of performance or intention in which it is need to be optimized simultaneously. The most distinct features of data mining is that it deals with huge and complex datasets in which its volume varies from gigabytes to even terabytes. This requires the data mining operations algorithms to be robust, stable and scalable along with the ability to cooperate with different research domains. Hence the various data mining tasks plays a crucial role in each and every aspect of information extraction and this in turn leads to the emergence of several data mining tools. From a pragmatic perspective, the graphical interface used in the tools tends to be more efficient, user friendly and easier to operate in which they are highly preferred by researchers. this concept is the minimum spanning tree.

This paper describes the previous work done on the data mining services, This paper use the tool called Application with various data mining services, and builds the own application based on the Application with more efficient data mining algorithm, and paper will also leads in deploying the tool into the cloud.

II . RELATED WORK

Data mining, often also called knowledge discovery in databases (KDD), is the process of extracting (unknown) patterns from data. There exists a variety of different data mining methods and algorithms, which commonly involve the following classes of tasks: Inferring rudimentary rules, statistical modeling, constructing decision trees, constructing rules, mining association rules, In our work we refer to data mining as part of the application that is represented by a business process. Data Mining on event log data in order to construct processes (Process Mining) and data mining on business processes, By providing an efficient tool called Application[2], The paper include in building an application (tool) the provide the data mining services in more efficient way as that provided by the Application, and also paper will also lead in deploying the tool into the Cloud where all the common user can easily extract the tool, The algorithm that have been used in the tool are like k -Means Hierarical, Bi Cluster. The tool (the R Analytical Tool To Learn Easily) is a graphical data mining application built upon the statistical language R. An understanding of R is not required in order to use Application. However, a basic introduction to Application is simple to use, quick to deploy, and allows us to rapidly work through the modelling phase of a data mining project. R, on the other hand, provides a very powerful language for performing data mining, the application uses the Gnome graphical user interface and runs under various operating systems, including GNU/Linux, Macintosh OS/X, and MS/Windows. The application is build by using the architecture of data mining services Before starting with the architecture description, we must say that our service has been designed as a complete service, functioning autonomously, this means, it does not require any other component or service to work, although in the future, it could be orchestrated with other services in order to offer a more powerful functionality. For this reason our service has been designed following the SOA principles and implemented by means of Web Services.In order to explain the service architecture using a reference framework, we used that proposed by Arsanjani [2].

Fig. 1 depicts an adaptation of Arsanjani's architecture for our service. The architecture of our service is divided in five layers. Data layer (first layer) gathers the Data Mining Service Repository and other data sources which store data to be processed by the service. The data access is based on a wrapper which mediates between calls from client application components to the data sources by transforming incoming requests into message format that is understandable to the Enterprise Components The second layer, called Enterprise Components, gathers the components that are responsible for realizing functionality and maintaining the QoS of the exposed services in the third layer. This currently consists of fourWeb Services: one for wrapping datamining algorithms and the preprocessing tasks, another for visualizing the results of the obtained model, another for validating the xml data file sent to the service and transforming it to the formatwhich the datamining algorithm requires and the fourth for connecting to and querying the repository. The communication among these Web Services is based on an XSD schema defined for this purpose as a consequence of the lack of standards for exchanging data and knowledge as Podpecan et al. stated in [54]. Although there are some advances in this direction, for instance of the atributes which the data set must have, the preprocessing tasks to be carried out and the selection of algorithms and their settings in order to answer a specific business problem, whereas the end-user (human being or machine) only needs to indicate where the data is

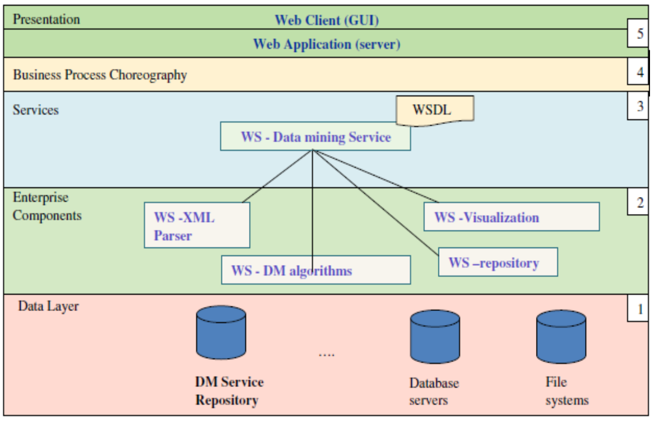


Fig. 1.1 System Architecture

Operating mode

The service works as follow. It offers a set of templates that specify the data which must be sent to the service in order to obtain certain patterns or models that give response to the users' questions. These templates contain the definition of the attributes as well as the mining algorithms which are suitable for obtaining the patterns. Since one of the difficulties which data miners face is the selection of parameters and how these affect the result, the parameters of the algorithms are established by the service itself, by making a previous analysis of the data and/or using other mining algorithms.

The definition of these templates is made from a rigorous experimentation in each business domain. Thus, the user interface, which is built to use this service, must allow the end-user to send the data file or indicate where it is stored and next, invoke the corresponding method get Non Advanced Results or get Advanced Results (see Section 4.2) and finally process and show the results obtained. As the service offers the possibility of invoking the template again and changing the parameters, the interface must contemplate this functionality

Predictive Model Markup Language (PMML) for describing various data mining and statistical models, and ExpML language for sharing machine learning information [70], the majority are unsupported by the general community. Moreover, there is no common and generally accepted XML-based language for describing tabular and other types of data and most data mining algorithms still use old style data formats like csv, tab, or arff [74]. In our current implementation, the DM algorithms' Web Service wraps four data mining algorithms: SimpleKmeans [28], Xmeans [51] and EM [28] from Weka [74] and the implementation of Apriori (association rule miner) developed by Borgelt [4]. It presents its results in the proprietary format of the algorithm. The WS-Visualization offers different kinds of graphs such as histograms, spider and pie charts for graphically showing clustering results and a 3D-graph for visualizing association rules [75]. All of these components have been programmed in java except for the visualization module which also uses the graphical capabilities

The third layer exposes the services which can be consumed by a client application or software whichwants to include this functionality, for example, a Learning Content Management System (LCMS) as we

show in our case study. This service can be discovered or be statically bound and then invoked, or possibly, choreographed into a composite service. The service is described in WSDL (see Fig. 2).

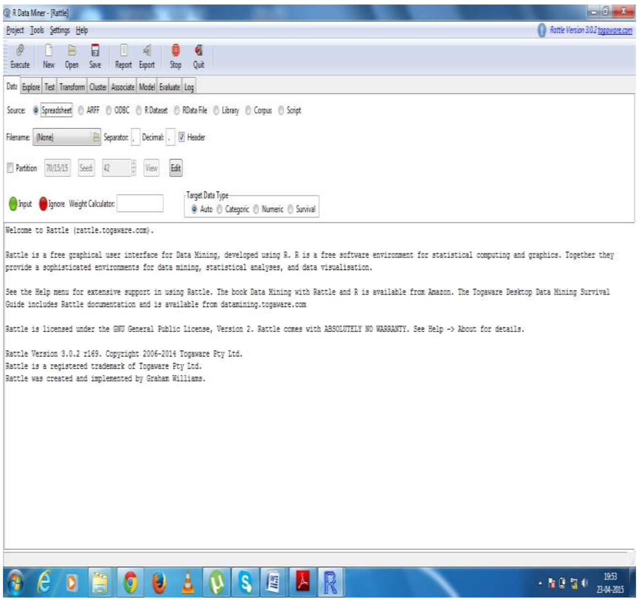
Pihur et al. developed an automatic method that was making use of a set of validation indexes to place a group of clustering algorithms for a given clustering task. This method automatically selects the best clustering algorithm by simultaneously testing multiple validation methods [8].

The validation of the clustering process is based on the comprehension of biological information related to the problem. The use of this kind of information helps to define “natural groups,” and thus, it helps to find meaningful clusters in problems where that knowledge is valid. However, this type of methods can be too focused on the problem at hand, which can finally lead to an analysis that is only valid for a specific.

III. EXISTING SYSTEM

This section describes some of the Data mining services in which the results can analyzed in more efficient manner with use of tool(Application).

APPLICATION (R Analytical Tool to Learn Easily) is a tool [9] in data mining that gives an uncomplicated and logical interface. It is built on top of the open source and free statistical language R with the help of Gnome graphical interface. This interface takes the user through the basic step of data mining. It turns out to be user friendly software by illustrating the R code that is used to achieve this. It uses the R Statistical Software through a graphical user interface. The software contains a Log Code tab, which may replicates the R code for any activity by GUI, which can be copied and posted. The software permits the dataset to be partitioned into training, validation and testing. The dataset can also be viewed and edited by the user. This software [10] also has the option for scoring an external data file. Application is compatible with GNU/Linux, Macintosh OS X and MS/Windows. It presents statistical and visual summaries of data, which transforms their data into forms that is readily modeled, support for building both unsupervised and supervised models from the data, it represents the performance of models graphically and scored new datasets. The Application software is used in Australia and other countries for business, government, research, statistical analysis, model generation and for teaching data mining. Whilst the tool itself may be sufficient for all of a user's needs, it also provides a more sophisticated processing



Source: ● Spreadsheet ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library ○ Corpus ○ Script

Filename: (None) Separator: , Decimal: . ☑ Header

☐ Partition 70/15/15 Seed: 42 View Edit

Input Ignore Weight Calculator:

Target Data Type ● Auto ○ Categoric ○ Numeric ○ Survival

```
Welcome to Rattle (rattle.togaware.com).

Rattle is a free graphical user interface for Data Mining, developed using R. R is a free software environment for statistical computing and graphics. Together they
provide a sophisticated environments for data mining, statistical analyses, and data visualisation.

See the Help menu for extensive support in using Rattle. The book Data Mining with Rattle and R is available from Amazon. The Togaware Desktop Data Mining Survival
Guide includes Rattle documentation and is available from datamining.togaware.com

Rattle is licensed under the GNU General Public License, Version 2. Rattle comes with ABSOLUTELY NO WARRANTY. See Help -> About for details.

Rattle Version 3.0.2 r169. Copyright 2006-2014 Togaware Pty Ltd.
Rattle is a registered trademark of Togaware Pty Ltd.
Rattle was created and implemented by Graham Williams.
```

Fig.1.2 Snapshot of Application

The figure 3.2 expalin about the existing data mining application, the application provides data mining services such as K means and clustering, the results of the application will be graphical or it may be in the text fom where the results can be easily analyzed and understood by the end-users

IV. PROPOSED SYSTEM

This section describes an approach of a application which is more advanced then the Application tool, the New application shows the result in graphical representation, and the proposed application contains more data mining algorithms, Application is simple to use, quick to deploy, and allows us to rapidly work through the modelling phase of a data mining project. R, on the other hand, provides a very powerful language for performing data mining, and when we need to fine tune our data mining projects we can always migrate from Application to R simply by taking tool underlying commands and deploying them within the R console.

Tool uses the Gnome graphical user interface and runs under various operating systems, including GNU/Linux, Macintosh OS/X, and MS/Windows, The application is build by using the librarires such as Gtk2, RGtk2 tcltk, utils. The application also provide additional features where we can load N number of data set and can view the result in any of the form, it also provides 1.Load a Dataset;

2. Select variables and entites for exploring and mining;

3. Explore the data;

4. Transform the data into training and test datasets;.

The application is built with Menus and Buttons, many function are provided by the menus and tool bar buttons. A project is a packaging of a dataset, variable selections, explorations, Clusters and models built from the data. Application allows projects to be saved for later resumption of the work or for sharing the data mining project with other users. A project is typically saved to a file with the .Application extension (although in reality it is just a standard .RData file. At a later time you can load a project into Application to restore the data, models, and other displayed information relating to the project, and resume your data mining from that point. You can also share these project files with other Application users, which is quite useful for data mining teams. You can rename the files, keeping the .Application extension, without impact is calculated. Using the MED the average Intracluster index is obtained to determine the cluster solution in the hadoop environment. The Name Node in the hadoop environment stores the location of the given data. The Data Node process the dMED and IC-av functions parallel in the hadoop environment which produces the results in an efficient timely manner.

Tools Menu and Toolbar.

It is important to understand the user interface paradigm used within Rattle. Basically, we will specify within each Tab window what it is we want to happen, and then click the Execute button to have the actions performed. Pressing the F5 function key and selecting the menu item Execute under the Tools menu have the same effect.

The Export button is available to export various objects and entities from Rattle. Details are available together with the specific sections in the following. The nature of the export depends on which tab is active, and within the tab, which option is active. For example, if the Model tab is on display then Export will save the current model as PMML. Export is not yet implemented for all tabs and options, The existing system is not provided the security standards as the proposed that I'm building is provide with the login function, where the expert Data miners and non expert data miners can tend to be relaxed for their data set

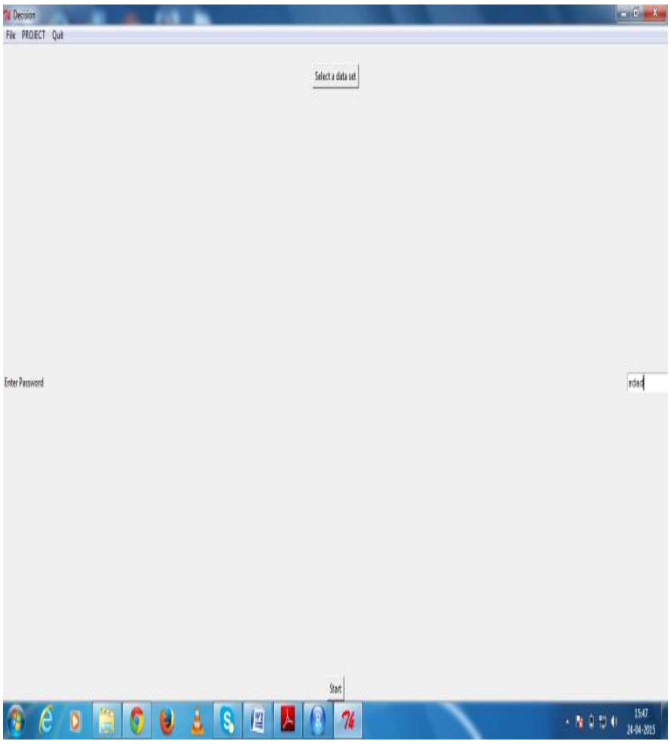
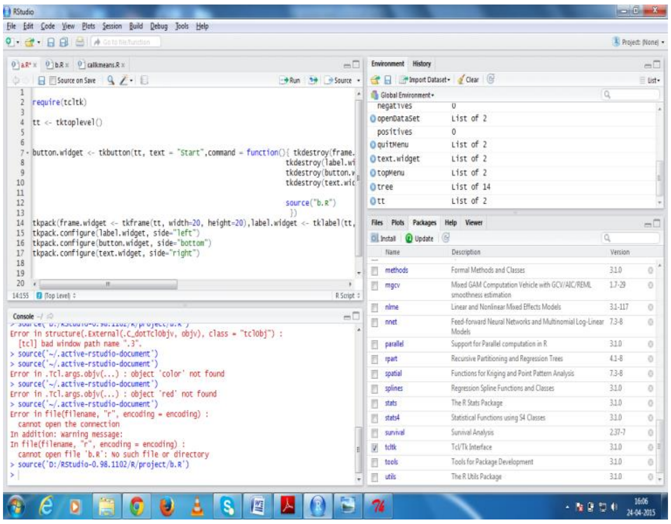


Fig 1.3 login functionality of the Rattle

Loading DataThe Data tab is the starting point for Rattle, and is where we can load a specific dataset into Rattle. Rattle is able to load data from various sources. Support is directly included in Rattle for comma separated data files (.csv files as might be exported by a spreadsheet), tab separated files (.txt, which are also commonly exported from spreadsheets), the common data mining dataset format used by Weka (.arff files), and from an ODBC connection (thus allowing connection to an enormous collection of data sources including MS/Excel, MS/Access, SQL Server, Oracle, IBM DB2, Teradata, MySQL, and Postgress). Underneath, R is very flexible in where it obtains its data from, and data from almost any source can be loaded. Consequently, Rattle is able to access this same variety of sources. It does, however, require the loading of the data into the R console and then within Rattle loading it as an R Dataset. All kinds of data can be loaded directly into R—including loading data directly from CSV and TXT files, MS/Excel spreadsheets, MS/Access databases, SAS, SPSS, Minitab, Oracle, MySQL, and SQLite.   Rattle uses what is called the Cairo device for displaying any plots. If the Cairo device is not available within your installation then Rattle resorts to the default window device for the operating system (x11 for Linux and window for MS/Windows). The Cairo device has a number of advantages, one being that the device can be encapsulated within other windows, as is done with Rattle to provide various operating system independent functionality. The Save button of the plot window (on the Cairo device) allows you to save the graphics to a file in one of the supported formats: pdf, png (good for vector images and text), and jpg (good for colorful images) A popup will request the filename to save to. The default is to save as PDF format, saving to a file with the filename extension of .pdf. You can also save the data into an window where that data can be extracted for further use. The Cairo device is been use rigoursly where the user can easily interact with the data.

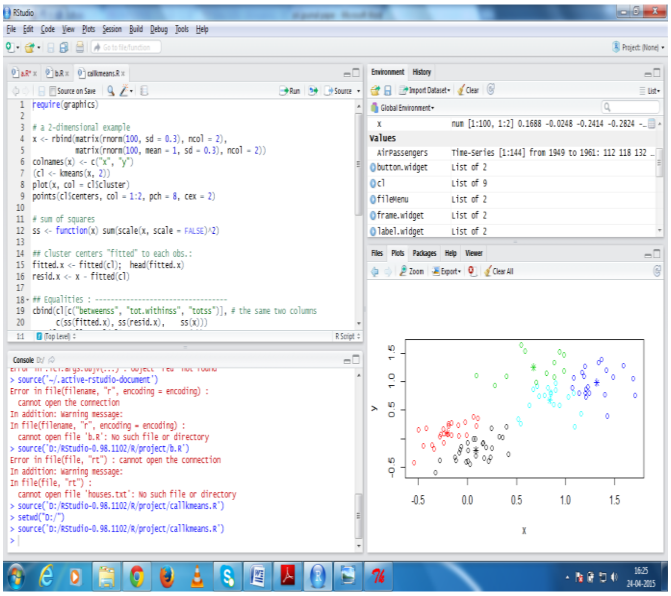


```
require(tcltk)

tt <- tktoplevel()


button.widget <- tkbutton(tt, text = "Start",command = function(){ tkdestroy(frame.
                                                              tkdestroy(label.wi
                                                              tkdestroy(button.v
                                                              tkdestroy(text.wic

                                                              source("b.R")
                                                               })
tkpack(frame.widget <- tkframe(tt, width=20, height=20),label.widget <- tklabel(tt,
tkpack.configure(label.widget, side="left")
tkpack.configure(button.widget, side="bottom")
tkpack.configure(text.widget, side="right")
```

```
Error in structure(.External(.C_dotTclObjv, objv), class = "tclObj") :
  [tcl] bad window path name ".3".
> source('~/.active-rstudio-document')
> source('~/.active-rstudio-document')
Error in .Tcl.args.objv(...) : object 'color' not found
> source('~/.active-rstudio-document')
Error in .Tcl.args.objv(...) : object 'red' not found
> source('~/.active-rstudio-document')
Error in file(filename, "r", encoding = encoding) :
  cannot open the connection
In addition: Warning message:
In file(filename, "r", encoding = encoding) :
  cannot open file 'b.R': No such file or directory
> source('D:/RStudio-0.98.1102/R/project/b.R')
>
```

Fig. 1.4 show the code used in R Studio.

The above figure describes about the code, by using the R studio, the code has been liked with one another, so that the application is build starting with the login functionality, the code also explains about the algorithm which I have used lake K means and hierarchal algorithm, where the result are presented in the graphical form which can be easily understood by users.



```
require(graphics)

# a 2-dimensional example
x <- rbind(matrix(rnorm(100, sd = 0.3), ncol = 2),
           matrix(rnorm(100, mean = 1, sd = 0.3), ncol = 2))
colnames(x) <- c("x", "y")
(cl <- kmeans(x, 2))
plot(x, col = cl$cluster)
points(cl$centers, col = 1:2, pch = 8, cex = 2)

# sum of squares
ss <- function(x) sum(scale(x, scale = FALSE)^2)

## cluster centers "fitted" to each obs.:
fitted.x <- fitted(cl);  head(fitted.x)
resid.x <- x - fitted(cl)

## Equalities : ----------------------------------
cbind(cl[c("betweenss", "tot.withinss", "totss")], # the same two columns
      c(ss(fitted.x), ss(resid.x),    ss(x)))
```

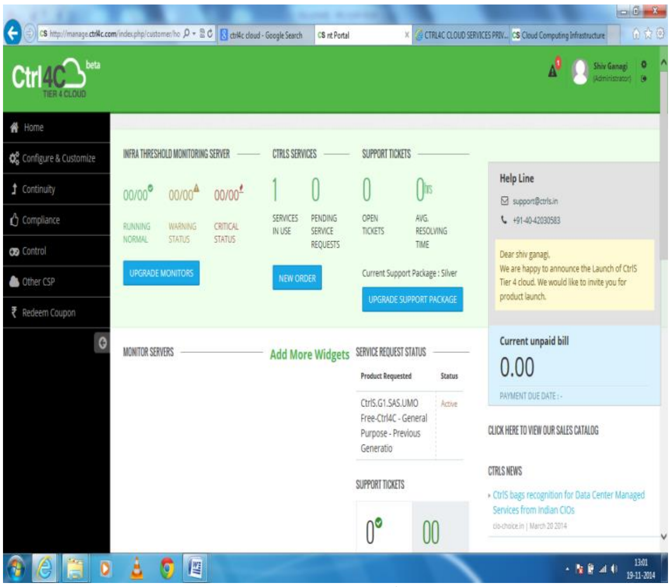
```
> source('~/.active-rstudio-document')
Error in file(filename, "r", encoding = encoding) :
  cannot open the connection
In addition: Warning message:
In file(filename, "r", encoding = encoding) :
  cannot open file 'b.R': No such file or directory
> source('D:/RStudio-0.98.1102/R/project/b.R')
Error in file(file, "rt") : cannot open the connection
In addition: Warning message:
In file(file, "rt") :
  cannot open file 'houses.txt': No such file or directory
> source('D:/RStudio-0.98.1102/R/project/callkmeans.R')
> setwd("D:/")
> source('D:/RStudio-0.98.1102/R/project/callkmeans.R')
>
```

Fig. 1.5 Result in K means Algorithm

Clustering is the process of partitioning a group of data points into a small number of clusters. For instance, the items in a supermarket are clustered in categories (butter, cheese and milk are grouped in dairy products). Of course this is a qualitative kind of partitioning. A quantitative approach would be to measure certain features of the products, say percentage of milk and others, and products with high percentage of milk would be grouped together. In general, we have n data points xi,i=1...nthat have to be partitioned in k clusters. The goal is to assign a cluster to each data point. K-means is a clustering method that aims to find the positions μi,i=1...k of the clusters that minimize the distance from the data points to the cluster. K-means clustering solves

Arg minc∑i=1k∑x∈cid(x,μi)=argminc∑i=1k∑x∈ci‖ x−μi‖ 22

where ci is the set of points that belong to cluster i. The K-means clustering uses the square of the Euclidean distance d(x,μi)=‖ x−μi‖ 22. This problem is not trivial (in fact it is NP-hard), so the K-means algorithm only hopes to find the global minimum, possibly getting stuck in a different solution.
In my Future work I will be implementing various Data mining services such as Apriori, c4,5, Naïve Bayes. I will be using the cloud computing techonology where I will be deploying the application into the cloud where the user can easily access the application, the cloud which I have used is the Ctl4c



00/00 RUNNING NORMAL
00/00 WARNING STATUS
00/00 CRITICAL STATUS
1 SERVICES IN USE
0 PENDING SERVICE REQUESTS
0 OPEN TICKETS
0HRS AVG. RESOLVING TIME

Current Support Package : Silver

support@ctrls.in
+91-40-42030583

Dear shiv ganagi,
We are happy to announce the Launch of CtrlS Tier 4 cloud. We would like to invite you for product launch.

Current unpaid bill
0.00
PAYMENT DUE DATE : -

Product Requested	Status
CtrlS.G1.SAS.UMO Free-Ctrl4C - General Purpose - Previous Generatio	Active

SUPPORT TICKETS
0 00

Ctrls bags recognition for Data Center Managed Services from Indian CIOs
cio-choice.in \| March 20 2014

13:01
19-11-2014

Fig. 1.5 Home page of Ctrl4c.

The above figure explains about the home ctrl4c, before logging into home page , I will be purchasing ctrl4c cloud where, the ctrl4c will be providing one month version based on the purchase, the cloud which i have purchased provides 50gb disk space, and memory about 2048mb, it also provide IP address for virtualization and security purpose, we need to access on function button called configure and customize, there it provide option or tell about the space and memory which I will be using, it also provide option for deployment.

V. CONCLUSION

The delivery of data mining as a service is an emergent necessity, above all for small to medium range organizations which are the most constrained by the high cost of data mining software and the availability of expert data miners to use this software. Until now, the tools deployed as Bi-a-as-Service in the cloud are conceived more for license cost-saving than as a product which can be used directly by end-users without data mining knowledge.
The application is build to provide the data mining services, where the result will be in the form of graphical or text, where the user can easily go through the result, and also the application will be deployed into cloud, so that users can easily use the application

REFRENCES

C. Borgelt, Efficient implementations of Apriori and Eclat, First Workshop of Frequent Item Set Mining Implementations, Melbourne, 2003.

[1] P. Brezany, I. Janciak, A. Woehrer, A.M. Tjoa, GridMiner: a framework for knowledge discovery on the Grid — from a vision to design and implementation, Cracow Grid Workshop, Cracow, December 12–15, 2004.

[2] F.Castro, A. Vellido, A. Nebot, F.Múgica,Applyingdataminingtechniques toe-Learning problems, in: J. Kacprzyk (Ed.), Studies in Computational Intelligence, Springer-Verlag, 2007, pp. 183–221.

[3] K. Channabasavaiah, K. Holley, E.M. Tuggle, Migrating to a service-oriented architecture, IBM DeveloperWorks Retrieved April, 2011 from https://www.ibm.com/ developerworks/library/ws-migratesoa/2004.

[4] M.C. Chen, A.L. Chiu, H.H. Chang, Mining changes in customer behavior in retail marketing, Expert Systems with Applications 28 (2005) 773–781.

[5] Y. Chen, S. Spangler, J. Kreulen, S. Boyer, T. Griffin, A. Alba, A. Behal, B. He, L. Kato, A. Lelescu, C. Kieliszewski, X. Wu, L. Zhang, SIMPLE: a strategic information mining platform for licensing and execution, Proc. of the 2009 IEEE international Conference on Data Mining Workshops, IEEE Computer Society, Washington, DC, 2009, pp. 270–275.

[6] X. Cheng, H. Liu, Personalized services research based on web data mining technology, Second International Symposium on Computational Intelligence and Design, Changsha, 2009.

[7] K. Chine, Scientific computing environments in the age of virtualization, toward a universal platform for the Cloud, Proc. of the 2009 IEEE International Workshop on Open Source Software for Scientific Computation (OSSC), 2009, pp. 44–48.

[8] M. Cocea, S. Weibelzahl, Cross-system validation of engagement prediction from log files, Proc. of the Second European Conference on Technology Enhanced Learning Sustaining TEL: from Innovation to Learning and Practice, 2007, pp. 14–25.

[9] M. Colan, Service-Oriented Architecture expands the vision of Web services Part 2, IBM DeveloperWorks, 2004 Retrieved April, 2011 from http://www. ibm.com/developerworks/webservices/library/ws-soaintro2/.

[10] F. Curbera, Y. Goland, J. Klein, F. Leymann, D. Roller, S. Thatte, S.Weerawarana, Business Process Execution Language forWeb Service (BPEL4WS) 1.0 Retrieved April, 2011 from http://www.ibm.com/developerworks/library/ws-bpel August 2002.