IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Special Issue-14, May 2019 E-ISSN: 2347-2693

An Efficient Cluster Analysis of Cyber Crime Records using R

Mir Abdul Samim Ansari[1*], Gopal K.Shyam[2]

[1,2] Schools of Computing and Information Technology, Reva-University, Bangalore, INDIA

Corresponding Author: Samimnakhwan@gmail.com (+91-7348908890)

DOI: https://doi.org/10.26438/ijcse/v7si14.141145 | Available online at: www.ijcseonline.org

***Abstract*—** Cluster evaluation divides the records into groups which can be meaningful and beneficial. It's also used as a start line for different functions of information summarization. This paper speak some very fundamental algorithms like k-means, Fuzzy C-method, Hierarchical clustering to give you clusters, and use R information mining device. The outcomes are examined at the datasets specifically on-line news popularity, Cyber Crime information Set information evaluation. All datasets became analyzed with specific clustering algorithms and the figures we're displaying the running of them in R information mining tool. Each set of rules has its specialty and antithetical conduct.

***Keywords*—** *K-means algorithm, Fuzzy C-method algorithm, Hierarchical clustering algorithm, R tool.*

I. Introduction

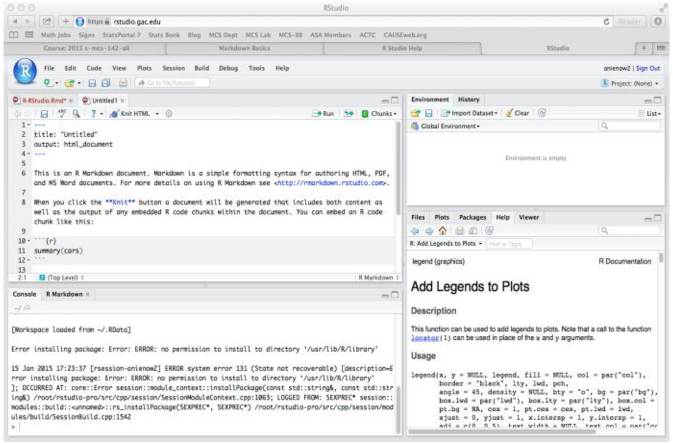
Cluster evaluation divides information into meaning full organizations (clusters) which proportion commonplace characteristics i.e. equal cluster are similar to every apart from those in different clusters. It is observe of routinely locating training. An internet web page specially information articles which might be flooded within the net need to be grouped. The clustering of those specific organizations is a leap forward closer to the automation manner, which calls for many fields, together with net search engines like Google, net robots and records evaluation.

Any new net page is going thru several levels along with records acquisition, pre-processing, characteristic extraction, Category and publishes processing into the database. Cluster evaluation may be appeared as a shape of the type which creates a labelling of items with elegance labels. But it derives those labels most effective from the information. Records mining functionalities are the Characterization and discrimination, mining common styles, affiliation, correlation, category and prediction, cluster evaluation, outlier evaluation and evolution evaluation [1]. Clustering is a vibrant approach. The answer isn't always unique and it firmly relies upon the analysts' selections. Clustering continually offers groups or clusters, even supposing there may be no predefined shape. Even as making use of cluster evaluation we're thinking of that the groups exist. However this hypothesis can be fake. The final results of clustering must in no way be generalized [9].

II. R Tool

R is free software program surroundings for statistical computing and snap-shots. It compiles and runs on an extensive type of UNIX platforms, home windows and Mac OS [12]. R is public area software program broadly speaking used for statistical evaluation and picture strategies [10]. A middle set of applications is covered with the set-up of R, with extra than 7,801 extra applications (as of January 2016[update]) to be had on the complete R Archive community (CRAN), Bio conductor, Omega-hat, Git Hub, and different repositories [14].

It should include important findings discussed briefly. Wherever necessary, elaborate on the tables and figures without repeating their contents. Interpret the findings in view of the results obtained in this and in past studies on this topic. State the conclusions in a few sentences at the end of the paper. However, valid colored photographs can also be published. R device affords an extensive elegance of statistical that consists of classical statistical exams, linear and nonlinear modelling, category, time-collection evaluation, clustering and numerous graphical features [13]. R makes use of collections of applications to carry out specific capabilities [11]. CRAN venture perspectives offer severa applications to extraordinary customers consistent with their flavour. R bundle incorporate exceptional features for records mining processes. This paper compares diverse clustering algorithms on datasets the use of R that allows you to be beneficial for researchers running on scientific records and organic records as nicely. For this IDE, R Studio is used refer the under discern [1].



~~~
---
title: "Untitled"
output: html_document
---

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF,
and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as
well as the output of any embedded R code chunks within the document. You can embed an R code
chunk like this:

```{r}
summary(cars)
```

~~~

```
[Workspace loaded from ~/.RData]

Error installing package: Error: ERROR: no permission to install to directory '/usr/lib/R/library'

15 Jan 2015 17:23:37 [rsession-anienow2] ERROR system error 131 (State not recoverable) [description=E
rror installing package: Error: ERROR: no permission to install to directory '/usr/lib/R/library'
]; OCCURRED AT: core::Error session::module_context::installPackage(const std::string&, const std::str
ing&) /root/rstudio-pro/src/cpp/session/SessionModuleContext.cpp:1063; LOGGED FROM: SEXPREC* session::
modules::build::<unnamed>::rs_installPackage(SEXPREC*, SEXPREC*) /root/rstudio-pro/src/cpp/session/mod
ules/build/SessionBuild.cpp:1542
>
```

legend {graphics} R Documentation

Add Legends to Plots

Description

This function can be used to add legends to plots. Note that a call to the function locator(1) can be used in place of the x and y arguments.

Usage

```
legend(x, y = NULL, legend, fill = NULL, col = par("col"),
       border = "black", lty, lwd, pch,
       angle = 45, density = NULL, bty = "o", bg = par("bg"),
       box.lwd = par("lwd"), box.lty = par("lty"), box.col =
       pt.bg = NA, cex = 1, pt.cex = cex, pt.lwd = lwd,
       xjust = 0, yjust = 1, x.intersp = 1, y.intersp = 1,
       adj = c(0, 0.5), text.width = NULL, text.col = par("co
```

Fig 1*: R-tool Studio*

III. LITERATURE SURVEY

Arit Thammano [1] describes the maximum famous clustering set of rules due to its performance and advanced performance. However, the overall performance of k-means algorithm relies upon closely on the choice of preliminary centroids. This paper proposes an extension to the unique k-means algorithm permitting it to clear up category troubles. First, the entropy idea is hired to evolve the conventional k-means algorithm for use as a category method. Then, to enhance the overall performance of k-means algorithm, a brand new scheme to choose the preliminary cluster facilities is proposed. The proposed fashions are examined on seven benchmark records units from the UCI device studying repository. Records category is one of the essential issues in statistics mining. Type, as defined, is a system of locating a version that describes and distinguishes records instructions, for the reason of being capable of use the version to expect the magnificence of items which elegance label is unknown. There are numerous category strategies which have been used so far including decision tree, neural networks, aid vector machines, and Bayesian networks. This paper makes a specialty of a sort of type version this is primarily based on k-means clustering set of rules. K-means is the maximum famous clustering algorithm. It's miles very green and really smooth to put into effect. Except getting used as a clustering method, k-means has additionally been tailored for records type.

Ying zhao, George karypis [2] describe a quick and record clustering algorithms play an essential function in presenting intuitive navigation and surfing mechanisms through organizing huge quantities of records right into a small variety of significant clusters. Specially, clustering algorithms that construct significant hierarchies out of big record collections are perfect equipment for his or her interactive visualization and exploration as they offer records-perspectives which are constant, predictable, and at exceptional stages of granularity. This paper makes a speciality of record clustering algorithms that construct such hierarchical answers and (i) gives a complete observe of partition and agglomerative algorithms that use special criterion features and merging schemes. (ii) offers a brand new elegance of clustering algorithms referred to as limited agglomerative algorithms, which integrate capabilities from each partition and agglomerative techniques that lets in them to lessen the early-level mistakes made through agglomerative strategies and consequently enhance the excellent of clustering answers.

Chun-Nan Hsu, Han-Shen Huang , Bo-Hou Yang [3] describe the expectancy-Maximization (EM) algorithm is one of the maximum famous algorithms for records mining from incomplete statistics. But, while implemented to huge records units with a huge percentage of lacking records, the EM algorithm may additionally converge slowly. The triple soar extrapolation approach can successfully boost up the EM algorithm through extensively decreasing the quantity of iterations required for EM to converge.

IV. RELATED WORK

- **Record Pre-Processing**

Records pre-processing [20] is a record mining method that includes remodelling uncooked records into a comprehensible layout. Regularly the records is unstructured, inconsistent, has lacking values, and absence in positive behaviour or tendencies that offers many mistakes. Consequently, it desires to be wiped clean, incorporated, converted, and therefore decreased. Cleansing fills within the lacking values and eliminates noise. Integration takes the information cubes or chunks collectively the usage of a couple of databases. Transformation makes use of normalization and aggregates the records and discount enables in reducing the extent of records maintaining comparable analytical outcomes. The records set as noted above is taken from a UCI Repository: groups and Crime dataset.

It has overall 1994 times and 128 attributes like populace, race, and age. The attributes are actual and of multivariate traits. These records turned into first transformed into CSV document the use of JSON document from the internet site using Python. For naming conference, authentic records turned into assumed as “grimy records” and the records without a lacking values as “wiped clean records”. For smooth records, elimination of lacking values became needed to get the correct crime records set. To begin with, columns that had these missing values or sparse values have been deleted as undefined values could have a terrible effect at the accuracy of the version. For grimy information, the lacking records of a characteristic were transformed into median cost of that characteristic. For predicting characteristic, in step with “Capita Violent Crimes”, a brand new column known as “excessive Crime” became created that had a fee “1” for in line with “Capita Violent Crime” more than 0.1 and “0” in any other case. The edge of 0.1 turned into determined upon guide evaluation of records through view-thru procedure. All of the capabilities needed to be expected the usage of this target characteristic excessive crime.

- **Crime sample evaluation**

The records mining is information reading strategies that used to investigate crime records ahead saved from numerous assets to discover styles and traits in crimes. In extra, it may be implemented to expand performance in fixing the crimes faster and additionally may be implemented to routinely recommend the crimes. Crime preclusion and revealing end up an essential fashion in crime and a completely hard to clear up crimes. Numerous researches have observed numerous strategies to clear up the crimes that used too many programs. Such research can assist to hurry up the technique of fixing crime and assist the large records are very tough and complicated.

Goals

- Crime prevention and detection emerge as a critical fashion in crime and a totally hard to clear up crimes.
- The records used for evaluation require the accuracy and sufficiency.
- This proposed gadget focuses on visitors Violation and Border manipulate, Violent Crime, the Narcotics, and Cyber Crime.
- Problems and demanding situations on crime are records series and Integration, Crime sample, overall performance, Visualization.

V. PROPOSED WORK

After pre-processing we can use various clustering techniques

K-nearest acquaintances algorithm

The k-Nearest neighbour’s algorithm (k-NN) is a non-parametric method utilized for class and regression. In each instances, the center includes the k closest schooling exemplar within the feature area. In k-NN regression, the k-NN algorithm is used for estimating non-stop variables. One such set of rules makes use of a weighted common of the k nearest acquaintances, weighted by means of the inverse in their distance. This set of rules works as follows:

1. Compute the Euclidean distance from the question instance to the classified examples.
2. Order the classified examples by means of growing distance.

K-NN Algorithm

Input: Crime Data, Watermark Data
Output: Modified Crime Observation Data

```
Add the Crime Profiles (P).
Add the Crime Observation Data (O).
Enter watermark content (W).
Convert the watermark data to bytes and find the length of watermark data (L).
Sort the Crime Observation Data (O) Crime wise.
I=0
For Each Crime’s Observation Set in (O)
Alter the Observation Data’s third value such that OD(3) = 301 + W(I)
Change the OD(1) position = OD(1) position + W(I)
I=I+1
If I>=L Then
Break
End If
Next
Output the New Crime Data Set.
```

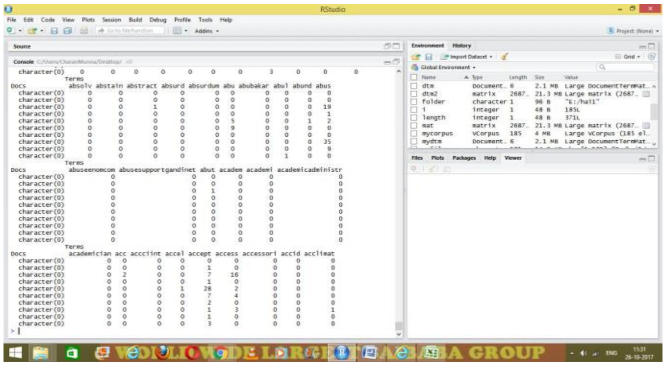
Input: Modified Crime Observation Data

Output: Crime Observation Data, Extracted Watermarked Data

```
Select the Crime Data Set (where Watermark Data Embedded) (P).
I=0;
For Each Crime’s Observation Set in (O)
    W(I) = Observation Data’s third value - 301
    Change the OD(3) value= OD(3) value -301
    If I=0 Then
        L= W(I)
    End If
    I=I+1
    If I >L Then
        Break
    End If
Next
Convert the watermark bytes to data.
Check the KNN Property.
Output Watermark Data.
```

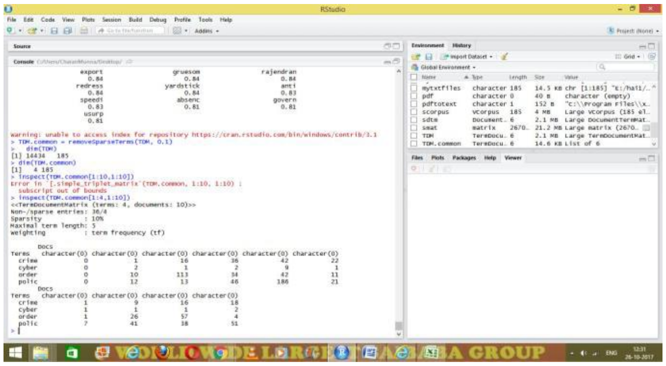
VI. RESULT EVALUATION

Now we will take input records set and we will take two hundred papers charters information dataset



```
  character(0)    0      0      0      0      0      0      0      3      0      0      0
                Terms
Docs            absolv abstain abstract absurd absurdum abu abubakar abul abund abus
  character(0)       0       0        0      0        0   0         0    0     0    0
  character(0)       0       0        0      0        0   0         0    0     0    0
  character(0)       0       0        1      0        0   0         0    0     0   19
  character(0)       0       0        0      0        0   0         0    0     0    1
  character(0)       0       0        0      0        0   5         0    0     1    2
  character(0)       0       0        0      0        0   9         0    0     0    0
  character(0)       0       0        0      0        0   0         0    0     0    0
  character(0)       0       0        0      0        0   0         0    0     0   35
  character(0)       0       0        0      0        0   0         0    0     0    9
  character(0)       0       0        0      0        0   0         0    1     0    0
                Terms
Docs            abuseenomcom abusesupportgandinet abut academ academi academicadministr
  character(0)             0                    0    0      0       0                  0
  character(0)             0                    0    0      0       0                  0
  character(0)             0                    0    1      0       0                  0
  character(0)             0                    0    0      0       0                  0
  character(0)             0                    0    0      0       0                  0
  character(0)             0                    0    0      0       0                  0
  character(0)             0                    0    0      0       0                  0
  character(0)             0                    0    0      0       0                  0
  character(0)             0                    0    0      0       0                  0
  character(0)             0                    0    0      0       0                  0
                Terms
Docs            academician acc accciint accel accept access accessori accid acclimat
  character(0)            0   0        0     0      0      0         0     0        0
  character(0)            0   0        0     0      1      0         0     0        0
  character(0)            0   2        0     0      7     16         0     0        0
  character(0)            0   0        0     0      1      0         0     0        0
  character(0)            0   0        0     1     28      2         0     0        0
  character(0)            0   0        0     0      7      4         0     0        0
  character(0)            0   0        0     0      2      0         0     0        0
  character(0)            0   0        0     0      1      3         0     0        1
  character(0)            0   0        0     0      1      0         0     0        0
  character(0)            0   0        0     0      3      0         0     0        0
> 
```

After records pre-processing each of those phrases happened greater than 150 instances.

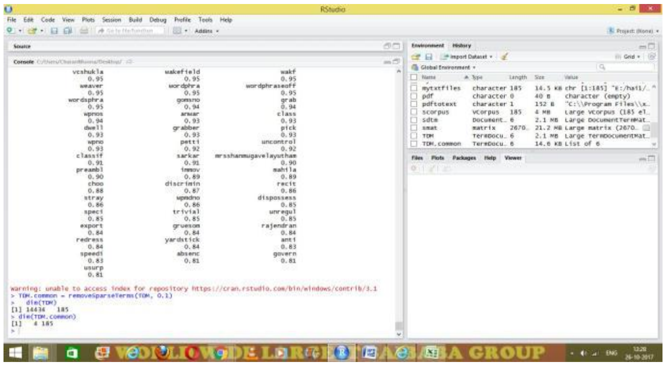


```
           export                  grueson                rajendran
             0.84                     0.84                     0.84
          redress                yardstick                     anti
             0.84                     0.84                     0.83
           speedi                   absenc                   govern
             0.83                     0.81                     0.81
            usurp
             0.81

Warning: unable to access index for repository https://cran.rstudio.com/bin/windows/contrib/3.1
> TDM.common = removeSparseTerms(TDM, 0.1)
>   dim(TDM)
[1] 14434   185
> dim(TDM.common)
[1]   4 185
> inspect(TDM.common[1:10,1:10])
Error in `[.simple_triplet_matrix`(TDM.common, 1:10, 1:10) : 
  subscript out of bounds
> inspect(TDM.common[1:4,1:10])
<<TermDocumentMatrix (terms: 4, documents: 10)>>
Non-/sparse entries: 36/4
Sparsity           : 10%
Maximal term length: 5
Weighting          : term frequency (tf)

       Docs
Terms   character(0) character(0) character(0) character(0) character(0) character(0)
  crime            0            1           16           36           42           22
  cyber            0            2            1            2            9            1
  order            0           10          113           34           42           11
  polic            0           12           13           46          186           21
       Docs
Terms   character(0) character(0) character(0) character(0)
  crime            1            9           16           18
  cyber            1            1            1            2
  order            1           26           57            4
  polic            7           41           38           51
> 
```

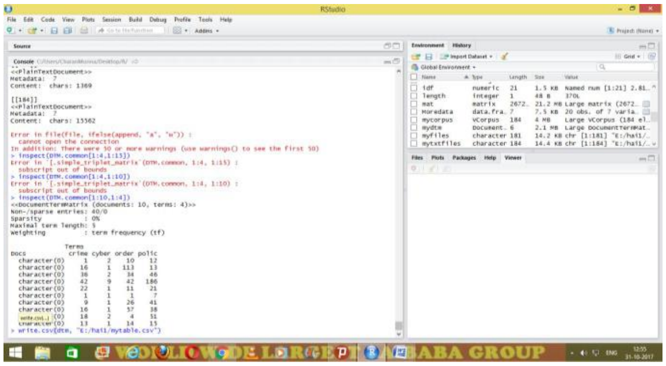
From the 14434 phrases that we began with, we're now left with a TDM which considers on four generally happening phrases.



```
          vcshukla             wakefield                  wakf
              0.95                  0.95                  0.95
            weaver             wordphra          wordphraseoff
              0.95                  0.95                  0.95
          wordsphra               gomsno                  grab
              0.95                  0.94                  0.94
             wpnos                 anwar                 class
              0.94                  0.93                  0.93
             dwell               grabber                  pick
              0.93                  0.93                  0.93
              wpno                 petti             uncontrol
              0.93                  0.92                  0.92
           classif                sarkar mrsshanmugavelayutham
              0.91                  0.91                  0.90
           preambl                 immov                mahila
              0.90                  0.89                  0.89
              choo              discrimin                 recit
              0.88                  0.87                  0.86
             stray                wpmdno             dispossess
              0.86                  0.86                  0.85
             speci               trivial                unregul
              0.85                  0.85                  0.85
            export               grueson              rajendran
              0.84                  0.84                  0.84
           redress             yardstick                  anti
              0.84                  0.84                  0.83
            speedi                absenc                govern
              0.83                  0.81                  0.81
             usurp
              0.81

Warning: unable to access index for repository https://cran.rstudio.com/bin/windows/contrib/3.1
> TDM.common = removeSparseTerms(TDM, 0.1)
>   dim(TDM)
[1] 14434   185
> dim(TDM.common)
[1]   4 185
> 
```

Time period frequency



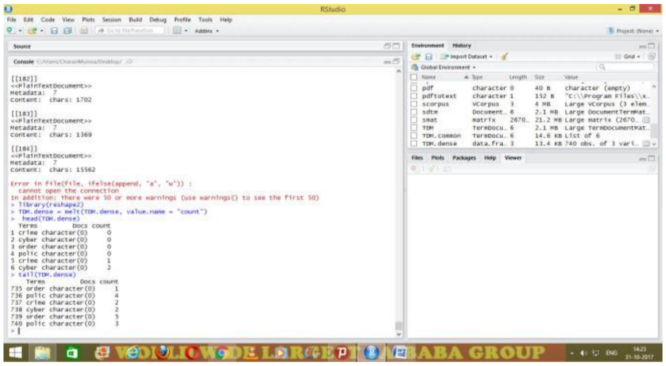
```
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 1369

[[184]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 15562

Error in file(file, ifelse(append, "a", "w")) : 
  cannot open the connection
In addition: There were 50 or more warnings (use warnings() to see the first 50)
> inspect(DTM.common[1:4,1:15])
Error in `[.simple_triplet_matrix`(DTM.common, 1:4, 1:15) : 
  subscript out of bounds
> inspect(DTM.common[1:4,1:10])
Error in `[.simple_triplet_matrix`(DTM.common, 1:4, 1:10) : 
  subscript out of bounds
> inspect(DTM.common[1:10,1:4])
<<DocumentTermMatrix (documents: 10, terms: 4)>>
Non-/sparse entries: 40/0
Sparsity           : 0%
Maximal term length: 5
Weighting          : term frequency (tf)

              Terms
Docs            crime cyber order polic
  character(0)      1     2    10    12
  character(0)     16     1   113    13
  character(0)     36     2    34    46
  character(0)     42     9    42   186
  character(0)     22     1    11    21
  character(0)      1     1     1     7
  character(0)      9     1    26    41
  character(0)     16     1    57    38
  write.csv(...) (0)     18     2     4    51
  character(0)     13     1    14    15
> write.csv(dtm, "E:/hai1/mytable.csv")
```

So, because it seems the sparse illustration became clearly losing area! (This can normally no longer be real although: it'll best observe for a matrix inclusive of simply the not unusual phrases). Besides, we want the records as an ordinary matrix with a view to produce the visualisation. The following step is to transform it right into a tidy format.



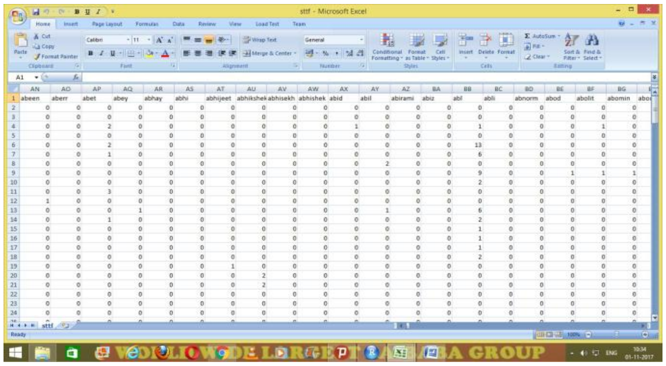
```
[[182]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 1702

[[183]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 1369

[[184]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 15562

Error in file(file, ifelse(append, "a", "w")) : 
  cannot open the connection
In addition: There were 50 or more warnings (use warnings() to see the first 50)
> library(reshape2)
> TDM.dense = melt(TDM.dense, value.name = "count")
>  head(TDM.dense)
  Terms         Docs count
1 crime character(0)     0
2 cyber character(0)     0
3 order character(0)     0
4 polic character(0)     0
5 crime character(0)     1
6 cyber character(0)     2
> tail(TDM.dense)
    Terms         Docs count
735 order character(0)     1
736 polic character(0)     4
737 crime character(0)     2
738 cyber character(0)     2
739 order character(0)     5
740 polic character(0)     3
> 
```

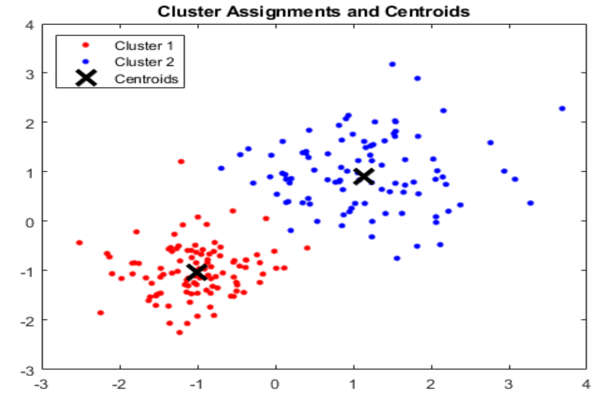
186 files and terms are



sttf - Microsoft Excel

	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG
1	abeen	aberr	abet	abey	abhay	abhi	abhijeet	abhikshek	abhisekh	abhishek	abid	abil	abirami	abiz	abl	abli	abnorm	abod	abolit	abomin
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	2	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	2	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0	0
7	0	0	1	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	1	1	1
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
11	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	1	0	0	0	0	0	0	0	1	0	0	6	0	0	0	0	0
14	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
19	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Graph representation for the above clusters



Cyber-attack techniques have been improved dramatically over time, especially in the past few years. Criminals have also adapted the advancements of computer technology to further their own illegal activities. Certain precautionary measures should be taken by all of us while using the internet which will assist in challenging this major threat Cyber Crime. There is a need to conduct research analysis of cyber-crimes to find out a best approach to protect sensitive data and take appropriate action against the cyber-attack.

CONCLUSION

Crime are characterised which exchange over the years and boom constantly. The converting and growing of crime cause the troubles of knowledge the crime conduct, crime predicting, specific detection, and handling huge volumes of information acquired from numerous assets. Studies pursuits have attempted to clear up those problems. Within the crime research tactics, enter information may be very important to apply in schooling procedure and trying out method. The schooling method is used to perform the crime version and the trying out system is used to validate the set of rules. The troubles of crime sample are regarding with locating and predicting the hidden crime. The proposed technique offers protection for the crime information in the course of outsourcing. Clustering and category is made at the crime facts. at the same time as classifying the crime statistics, watermark content material is brought for the reason of protection. The watermark content material is used for verifying the type facts. Primarily based on clustering and category, the records may be categorized and saved secured way. Additionally the crime statistics is been break up as according to the crime ratio.

REFERENCES

[1]. J. Han and M. Kamber, Data Mining: Concepts and Techniques, second ed. Morgan Kaufmann, 2006.

[2]. C.C. Aggarwal and P.S. Yu, “Finding Generalized Projected Clusters in High Dimensional Spaces,” Proc. 26th ACM SIGMOD Int‟l Conf. Management of Data, pp. 70-81, 2000.

[3]. K. Kailing, H.-P. Kriegel, P. Kro ̈ger, and S. Wanka, “Ranking Interesting Subspaces for Clustering High Dimensional Data,” Proc.Seventh European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD), pp. 241-252, 2003.

[4]. K. Kailing, H.-P. Kriegel, and P. Kro ̈ger, “Density-Connected Subspace Clustering for High- Dimensional Data,” Proc. Fourth SIAM Int‟l Conf. Data Mining (SDM), pp. 246-257, 2004.

[5]. E. Mu ̈ ller, S. Gu ̈nnemann, I. Assent, and T. Seidl, “Evaluating Clustering in Subspace Projections of High Dimensional Data,” Proc. VLDB Endowment, vol. 2, pp. 1270-1281, 2009.

[6]. E. Agirre, D. Martı ́nez, O.L. de Lacalle, and A. Soroa, “Two Graph-Based Algorithms for State-of-the-Art WSD,”Proc. Conf.Empirical Methods in Natural Language Processing (EMNLP), pp. 585- 593, 2006.

[7]. K. Ning, H. Ng, S. Srihari, H. Leong, and A. Nesvizhskii, “Examination of the Relationship between Essential Genes in PPI Network and Hub Proteins in Reverse Nearest Neighbor Topology,” BMC Bioinformatics,vol. 11, pp. 1-14, 2010.

[8]. D. Arthur and S. Vassilvitskii, “K-Means++: The Advantages of Careful Seeding,”Proc. 18th Ann. ACM-SIAM Symp. Discrete Algorithms (SODA),pp. 1027-1035, 2007.

[9]. I.S. Dhillon, Y. Guan, and B. Kulis, “Kernel k Means: Spectral Clustering and Normalized Cuts,”Proc. 10th ACM SIGKDD Int‟lConf. Knowledge Discovery and Data Mining,pp. 551- 556, 2004.

[10]. T.N. Tran, R. Wehrens, and L.M.C. Buydens, “Knn Density-Based Clustering for High Dimensional Multispectral Images,”Proc.Second GRSS/ISPRS Joint Workshop Remote Sensing and Data Fusion over Urban Areas,pp. 147-151, 2003.

Authors Profile

Mr Mir Abdul Samim Ansari Received BE in Information Science and Engineering from VTU, Bangaluru. Currently he is a student of Master of Data Engineering and Cloud Computing in Reva University, Bangaluru, India. His field of interest is to work in Data Analysis, Cloud Computing and IOT where he can apply knowledge and skills to develop himself.

Dr. Gopal Krishna Shyam received BE and M.Tech and Ph.D in Computer science and engineering from VTU, Belagavi. His research interest includes Cloud Computing, Grid Computing, High performance computing etc. He has published about 10 papers in highly reputed National/ International conferences like IEEE, Elsevier etc. and 5 papers in journals with high impact factor like Elsevier Journal on Network and Computer Applications and International Journal of Cloud computing (INDERSCINCE). His research articles on Cloud computing co-authored by Dr Sunilkumar S.Manvi have been cited by several researchers. He is a lifetime member of CSI and is actively involved in motivating students/faculties to join CSI/IEEE/ACM societies.