Volume 8, No. 5, May-June 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

Fuzzy Based DBSCAN Text Mining Technique for Malware Detection

Surabhi Thapa
Department of Computer Science & Engineering,
Guru Nanak Dev University, Regional Campus,
Jalandhar, India

Neena Madan
Department of Computer Science & Engineering,
Guru Nanak Dev University, Regional Campus,
Jalandhar, India

***Abstract*:** Text mining is an approach to perceive or find interesting and useful relationships and patterns in considerable amount of text. The Density based spatial clustering of applications with noise (DBSCAN) clustering is a vital method in data mining. DBSCAN was projected to adopt density-reachability and density attachment for handling the randomly shaped clusters and noise. It has been perceived that in existing work has introduced method persists until the density-linked cluster is completely discovered. Then, a new obscure element is extracted and processed, leading to the finding of a further cluster or noise. Therefore this paper has proposed fuzzy based DBSCAN text mining technique for malware detection that enhances ambiguity and functioning of the clusters formed. The experimental outcomes bring about the suggested method that clearly shows the fact that suggested method outperforms over the existing methods.

Keywords: clustering, data mining, DBSCAN, fuzzy based DBSCAN, malware, text mining.

I. INTRODUCTION

In present century data is growing vastly. Data is present in different forms like graphs, images, records, documents etc. To analyze these different forms it is necessary to obtain useful details. 'Data mining' is a mechanism that inputs the data and gives knowledge as output or a significant method in recognition of legitimate, unorthodox, potentially functional, and ultimately fathomable patterns in data. It also is the operation through which useful and important information is obtained from large heaps of data [1]. It is also explained as finding the links in a large relational database established on the different depth of angles [2]. Text mining is an operation to obtain interesting and significant patterns to examine erudition from data in text form[3].It is the procedure for analyzing a class of documents to understand the matter and meaning of the details they contain. Furthermore it is the way of finding correct, potentially useful and ultimately fathomable knowledge from large text data set [4].Clustering is the mechanism of merging a stack of objects (usually illustrated as elements in a multidimensional space) into categories of identical objects. Cluster analyzing is a very vital tool in data analysis. Cluster analyzing can be utilized as an individual data mining mechanism to access prescience into the data roll out, or as a pre step for alternative data mining rules operating on the detected clusters [5]. Malware categorization has been an exacting problem in the recent past and several researchers have strived to solve this problem using various mechanisms. It is security threat which can break machine functioning while not knowing user's data and it's hard to notice its behavior. Clustering algorithm based on density is capable of classifying clean malware and malicious malware with promising results in basis of precision, recall value and accuracy. Density based clustering algorithms have a wide application in data mining. They exert a local criterion to group objects: clusters are regarded as areas in data space where the objects are crammed, and which are divided by fields of low object density (noise). Out of the density based clustering algorithms DBSCAN is very popular. DBSCAN as it is most common and it discovers clusters starting from the evaluated density dispensation of corresponding nodes. A cluster, which is a fragment of the elements of the text base, complies with two properties:

1. All elements within the cluster are cooperatively density-connected.
2. If an element is density-connected to any element of the cluster, it is fragment of the cluster as well.

DBSCAN needs two stipulations: (eps) and the minimum quantity of elements needed in forming a cluster ($minPts$).This element-neighborhood is extracted, and if it contains enough elements, a cluster is started. Otherwise, it is noise. The neighborhood of element is also fragment of the cluster whose compact part is the element itself. Hence, all elements are added that are within the neighborhood, as with their own neighborhood when they are also compact. This mechanism persists until the cluster that is density-connected is completely found. Then, a new isolated element is extracted and processed, leading to the finding of more cluster or noise. Therefore this paper has proposed fuzzy based DBSCAN text mining procedure for malware detection thus improve the ambiguity and performance of the clusters formed.

II. RELATED WORK

S.M. Junaid et.al 2014[1] describes major current clustering procedures utilized in the mechanism of data mining. Different algorithms of these techniques are considered in this paper. The unearthing of useful and important information from large heaps of data happens to be objective of data mining. Many areas, like bioinformatics, image analysis, pattern identification, data mining, machine learning, use data scrutiny whose general method is clustering which is a key mechanism of data mining. This paper evaluates four major clustering procedures namely hierarchical, partitioned, density based and grid based. So this paper provides a quick review for clustering techniques.

S. Agarwal et.al 2013[2] focused on Data mining concepts and techniques which is an area of junction of statistics and

computer science used in uncovering patterns in the information bank. The retrieval of useful particulars from the dataset and modification into fathomable structure for future use happens to be data mining's objective. Different processes and mechanisms are used to operate data mining successfully.

R.Talib et.al 2016[3] presents a brief synopsis of text mining procedures that help to improve it. Specific patterns and sequences are applied for obtaining useful information by eliminating irrelevant details for predictive analysis. Selection and use of right techniques and tools as per the domain help the text mining procedure be easy and efficient. Domain knowledge integration, varying concepts granularity, multilingual texfinement, and natural language operation ambiguity are major issues and challenges that arise during text mining method.

G.G.Sundarkumar et.al 2013[4] suggests a static analysis method to perceive Malware based on (application programming interface) API call progression using text and data mining in tandem. We examined the dataset obtainable at CSMINING group. In the paper, they used 10-fold cross validation method for examining the procedures. They perceived that (support vector machine)SVM and (one class SVM)OCSVM achieved 100% sensitivity after balancing the dataset.

A. Joshi et.al 2013[5] describes different clustering mechanisms in data mining. Settling identical data into groups is an operation of clustering which is maintained as the most vital learning method like every other issue of this kind; it deals with discovering a structure in a collection of anonymous data. This paper examines six types of clustering methods- k-Means, hierarchical, DBSCAN, (ordering points to identify clustering structure) OPTICS, (stastical information grid)STING.

Y. Zhang et.al 2010[6] presented a system Reverse examination for spotting risky executable (Radux) for automatically spotting malicious code using the collected dataset of the harmless and malicious code. This system rests on fuzzy conjecture based on behavior hidden in malicious code. Decompile method is used to specify structural and behavioral traits of binary code, which generates more conceptual descriptions of malware. The suggested procedure can obtain the fuzzy subsets and its membership function in an automatic way with the (fuzzy neural network) GD-FNN learning algorithm.

P.Rai et.al 2010[7] review of divergent clustering procedures in data mining is provided. Clustering is vital for data mining and data analysis applications. The capability to discover highly correlated regions of objects when their number becomes very large is highly desirable, as data sets grow and their properties and data interrelationships change. While, it is notable that any clustering "is a division of the objects into categories based on a set of rules – it is neither true nor false".

S.R.Pande et.al 2012[8] explained that the Clustering is of vital use in data mining and data analysis applications. The capability to discover highly correlated regions of objects when their number becomes very large is highly desirable, as data sets grow and their properties and data interrelationships change. This paper reported the mechanism of clustering from the data mining element of view. This gave the properties of a "good" clustering mechanism and then methods used to find meaningful partitioning.

P. Mishra et.al 2012[9] reviewed the various data mining applications. The data depository is utilized in the noteworthy business value through enhancement of effectiveness of decision-making. In an undetermined and contentious business conditions, the value of planned information systems such as these are easily identified however in today's business environment, efficiency or speed is not the only key for competitiveness. Such enormous quantity of data has basically changed science and engineering, modifying many disciplines from data-poor to increasingly data-rich, and looking for new, data-intensive methods to lead research in science and engineering. To examine this huge data and drawing fruitful conclusions, it needs the special mechanisms called data mining tools. This paper gives overview of the data mining systems and some of its applications in different field. This review would be helpful to researchers to emphasize on the various issues of data mining.

M.Sukanya et.al 2012[10] discussed about text mining with its various techniques which can be used. Currently, the reserved information is increasing enormously day by day. This is indefinite form so we cannot obtain the needed information so text mining mechanisms are used. Some data mining methods are used to gather the useful details from textual records, such as classification, clustering, visualization and information extraction. Graphical visualization is used to provide better understandable information for mining the documents. An application in the field such as identifying news stories, junk emails, and examining of the market is studied.

A.K. Mann et.al 2013[11] focused on clustering and its different methods in data mining. Gathering details through a huge data set and modifying into a fathomable form for further use is aim of the data mining procedure. Clustering is vital in data examining and data mining applications. It categorizes classes of objects in such manner that objects in the identical group seem identical to one another than those in other groups (clusters). Algorithms such as hierarchical, partitioning, grid and density based algorithms can do it.

R. Kaur et.al 2013[12] presented the clustering mechanisms and algorithms used for it. Clustering can be contemplated as the most vital unsupervised learning method; so, like every other issue, it deals with discovering a structure in grouping of anonymous data. "Categorizing objects into groups whose members are identical in some way" is what clustering is. Therefore, a cluster is a set of objects which is "identical" between them and is "non-identical" to the objects affiliated to other clusters.

A.Smiti et.al 2013[13] presents an efficient clustering method, named "Soft DBSCAN" that merges DBSCAN and fuzzy set theory. Their new procedure is stimulated by fuzzy C Means in the form of using the fuzzy membership functions. The conclusions of their procedure show that it is efficient not only in handling noises, opposite to fuzzy C Means, but also, able to allocate one data element into different clusters. Simulative experiments are performed on distinct datasets, through different evaluation's basis, which highlight the soft DBSCAN's efficacy and cluster validity to check the better quality of clustering conclusions.

D. Sharma et.al 2014[14] introduced cluster mechanisms in data mining. Finding meaningful patterns and trends from

the data concealed in repositories is the main objective of data mining operation. For data mining and data analysis application, clustering is vital. It is a mechanism of categorizingcollection of objects that belong to the identical class. Forspotting natural groups in large heaps of data, Regulations, such as statistics, software engineering, biology, psychology and other social sciences are there in cluster analysis. These data sets are continuously becoming larger, and their dimensionality averts easy inspection and validation of the conclusions. Different mechanisms of clustering are farthest first, simple K-means, filtered, hierarchical, etc.

S.H. Ganesh et.al 2015[15] presents the current works carried out on EDM and examine their merits and drawbacks. This paper also concentrates on the progressive outcomes of the different data mining practices and methods applied in the surveyed articles, and thereby proposing the researchers on the upcoming works on EDM. In addition, an experiment was also carried out to assess, certain categorization and clustering algorithms to perceive the most reliable algorithms for future researches.

R.R. Tated et.al 2015[16] focused on the concept of text mining, text mining process, mechanisms used in text mining also presenting some real world implementation of text mining. In addition, brief discussion of text mining benefits and limitations has been presented.

K. Shah et.al 2015[17] focused on data mining mechanism for dynamic inspection of malwares by using dynamic inspection of executable and based on mining methods. API calls induced by samples during execution are used as parameter of experimentation.

A. Malhotra et.al 2016[18] introduced novel algorithm to categorize malwares in polymorphic or metamorphic malwares and clean or normal malwares. The perspective is to generate pydasm report. The instruction sets will be retrieved from the report via text mining and text preprocessing will be done for different procedures like comment removal, function extraction etc.

A.Shalaginov et.al 2016[19] explored application of neuro-fuzzy for multinomial classification of malware families and categories. They collected a novel dataset containing 400k samples for static malware analysis. Neuro-fuzzy performs well considering convolution of the issue and non-linearity of the data.

A. Malhotra et.al 2016[20] analyzed various literatures related to mobile malware detection. A fastidious study of the items related to mobile malware and the approach used for the detection of malware is done. Some suggested methods and type of approaches used in those methods are also summarized.

III. PROPOSED METHODOLOGY

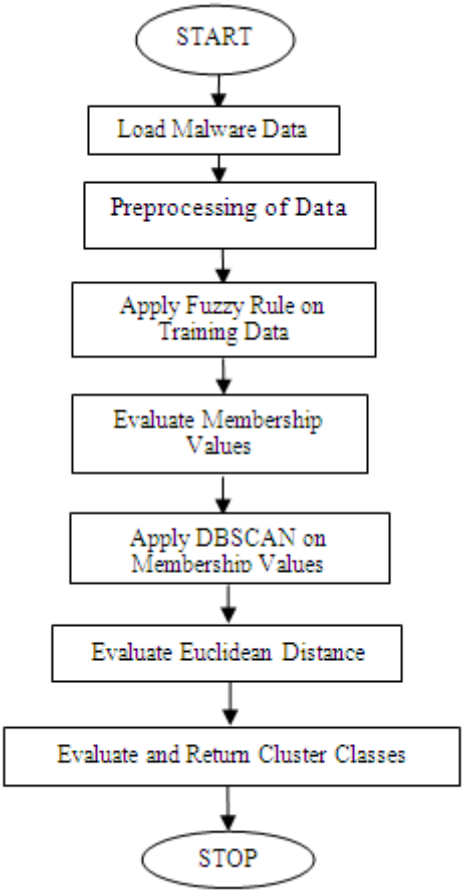

Fig.1 Flowchart of Proposed Methodology

Above figure 1 displays the flowchart of proposed methodology and steps of proposed methodology is given below:

1. Load Malware data in MATLAB.
2. In this step preprocessing of data is done.
3. Apply fuzzy rules on training data.
4. Now in this step we will evaluate membership values of given training data.
5. After calculating membership values of given training data, apply DBSCAN on membership values.
6. Evaluate Euclidean distance of given dataset.
7. After applying all these above steps we will evaluate and return cluster classes.

IV. IMPLEMENTATION AND PROPOSED ALGORITHM

The augmentation of the classic DBSCAN algorithm we propose, named fuzzy based DBSCAN, is obtained by considering crisp the distance, as seen in the classic approach, and by introduction of an approximate value of the desired properties of the neighborhood of an element $Minp_s$. It's done by substituting the numeric value $Minp_s$with a soft constraint described by a non-decreasing membership function on the domain of the positive integers. This soft constraint specifies the approximate number of elements that is required in the neighborhood of an element for generating a fuzzy core of a cluster. Let us define the piecewise linear membership function as follows:

$$\mu minP(y)\begin{cases}1, & if y \geq M_f p_{s_{Max}} \\ \dfrac{y - M_f p_{s_{Min}}}{M_f p_{s_{max}} - M_f p_{s_{Min}}}, & if M_f p_{s_{Min}} < y < M_f p_{s_{Max}} \\ 0, & if y \leq M_f p_{s_{Min}}\end{cases} \quad (1)$$

This membership function gives the value 1 when the number y of modules in the neighborhood of an element becomes greater than $M_f p_{s_{max}}$, a value 0 when y is below $M_f p_{s_{Min}}$ and intermediate values when y is in between $M_f p_{s_{Min}}$ and $M_f p_{s_{max}}$. Since users may find it diffi cult to specify the two values $M_f p_{s_{Min}}$ and $M_f p_{s_{max}}$, when they are not aware of the total items of objects involved in the process, they can specify two percentage values, $\%M_f p_{s_{Min}}$ and $\%M_f p_{s_{Max}}$ which are then converted into $M_f p_{s_{Min}}$ and $M_f p_{s_{Max}}$ as follows:

$M_f p_{s_{Min}} = \%M_f p_{s_{Min}} * N'$ and $M_f p_{s_{Max}} = \%M_f p_{s_{Min}} * N'$, in which N' is the total items of objects and n returns the closest integer to 'n'.

Let us redefine the fuzzy procedure .Given a set P' of N' objects represented by N' elements in the n-dimensional domain $Q^n p_{t1}, p_{t2}, \ldots p_{tn}$, where p_a has the coordinates $y_{i1}, y_{i2}, \ldots . y_{in}$. Given an element $p_t \in P'$, if y element $p_a \exists$ in the neighborhood of element p_t, i.e. with $\|p_a - p_t\| < \in$, s.t. $\mu_{min} P'(y) > 0$ the p_t is a fuzzy core element with membership degree to the fuzzy core given by $fuzzycore(p_t) = \mu_{min} P'(y)$ If two fuzzy core elements p_a, p_b ($fuzzycore(p_a) > 0$ and $fuzzycore(p_b) > 0$) $\exists$ with $a \neq b$ s.t. $\|p_a - p_t\| < \in$, then they define a cluster c_l, $p_a, p_b \in c_l$, and are fuzzy core elements of c_l, i.e., $p_a, p_b \in fuzzycore(c_l)$ with membership degree $fuzzycorec_l(p_a)$ and $fuzzycorec_l(p_b)$. An element p_t in a cluster that isn't a fuzzy core element is a boundary or border element if it satisfies the following: Given $p_t \nexists fuzzycore(c_l)$ if $\exists p_a \in fuzzycore(c_l)$, i.e., with membership degree $fuzzycore_{c_l}(p_a) > 0$, s.t. $\|p_a - p_t\| < \in$, then p_t gets a membership degree to c_l defined as: $\mu_{c_l}(p_t) = max_{p_a} \in fuzzycore(c_l) fuzzycorec_l(p_a)$ finally, element p_t that are not segment of a cluster are considered noise: $\forall c_l$ if $p_a \in fuzzycorec_l$ s.t. $\|p_a - p_t\| < \in$, then p_t is noise. The fuzzy procedure is sketched in Algorithms 1 and 2.In the fuzzy version, an element is marked as $NOISE$ if its neighborhood size seems less than or equal to $Minp_{s_{min}}$ otherwise it will be a fuzzy core element with a given membership value. Once the element is recognized as fuzzy core element the procedure $expandclusterfuzzycore$ is called (Algorithm 2). As seen in the classical $DBSCAN$, this procedure is committed in discovering all the reachable elements from p_t and to mark them as core or border elements. In the origin the commission of the element p_t is crisp while we introduce a fuzzy assignment (line 1) modeled by the fuzzy function $\mu_{min_l} P'()$. The identical function is engaged when a new fuzzy core element is identified (line 8). Firstly, we verify the density around a stated element p_t w.r.t. $Minp_{s_{min}}$ and then, if the element verifies the soft constraint, we add the element to the fuzzy core of cluster c_l with its associated membership value. Contrary to past, line 10 is only devoted to identify border elements not yet assigned to any cluster.

Algorithm1. $Approx\ Fuzzy\ Core\ DBSCAN(B, \in, Minp_{s_{min}}, Minp_{s_{max}})$

Require: P' : dataset of elements

Require: $\in$: the maximum distance around an element defining the element neighborhood

Require: $Minp_{s_{min}}, Minp_{s_{max}}$: soft constraint interval for the density around an element to be considered a core element to a degree

1. $c_l = 0$
2. $Clusters = \emptyset$
3. for all $p_t \in P'$ s.t. p_t is unvisited do
4. mark p_t as visited
5. neighbors p_s = regionQuery($p_t, \in$)
6. if($sizeof(neighborsp_s) \leq Minp_{s_{min}}$) then
7. mark p_t as NOISE
8. else
9. c_l = next cluster
10. $Clusters = ClustersexpandClusterFuzzyCore(p_t, neighborsp_s, c_l, \in, Minp_{s_{min}}, Minp_{s_{max}})$
11. end if
12. end for
13. return $Clusters$

Algorithm2.

$ExpandClusterFuzzyCore\ (p_t, neighborsp_s, c_l, \in, Minp_{s_{min}}, Minp_{s_{max}})$

Require: p_t: the element just marked as visited

Require: $neighborsp_s$: the elements in the neighborhood of p_t

Require: c_l: the actual cluster

Require: $\in$ the distance around an element to compute its density

Require: $Minp_{s_{min}}, Minp_{s_{max}}$: soft constraint interval for the density around an element to be considered a core element

1.add p_t to c_l with membership $Fuzzycore(p_t) = \mu Minp_t(|neighborsPts|)$

2. for all $p_t \in neighborsp_s$ do
3. if p_t is not visited then
4. mark p_t as visited
5. $neighborsp_s = regionQuery(p_t, \in)$
6. if $sizeof(neighborsp_s) > Minp_{s_{min}}$ then
7. $neighbors\ p_s = neighborsp_s \cup neighborsp_s$
8. add p_t to c_l with membership $Fuzzycore(p_t) = \mu Minp_t(|neighborsp_s)$
9. end if
10. if p_t is not yet member of any cluster then
11. add p_t to c_l (as border element)
12. end if
13. end if
14. end for
15. return c_l

V. EXPERIMENTAL RESULTS

The proposed fuzzy based DBSCAN text mining technique for malware detection is represented and executed in the MATLAB 2013.We have implemented seven standard parameters in assessing the accuracy, ambiguity and performance of our proposed system: accuracy (ACC), area under cover (AUC), relative absolute error (RAE), error rate(ER),FP rate, TP rate and F-measure.

The metrics used are composed as follows:

1. **ACC:** Accuracy is known as correctly classified instances.

$$\text{ACC}=\frac{Totalnumberofcorrectpredictions}{(Totalnumberofpredictions\quad)}$$

2. **AUC:** It is described as coverage of cases.
3. **RAE:** Relative error is the ratio of absolute error to magnitude of the exact value.

$$\text{RAE}=\frac{absoluteerror}{magnitudeoft\ \ heexactvalue\ .}$$

4. **ER:** Error rate is described as $1-accuracy$.

$$\text{ER}=\frac{FP+FN}{All}$$

5. **FP Rate:** FP Rate is positive instances which are incorrectly classified.

$$\text{FP Rate}=\frac{FP}{FP+TN}$$

6. **TP Rate:** TP Rate is positive instances which are correctly classified.

$$\text{TP Rate}=\frac{TP}{TP+FN}$$

7. **F-measure:** F-measure is a calculation of test's accuracy. F-Measure or F1 score contains both precision and recall. It is generally use to check the precision and reliability.

$$\text{F\_Measure}=2*\frac{\text{P}*\text{R}}{\text{P}+\text{R}}$$

5.1SNAPSHOT OF EXISTING TECHNIQUE

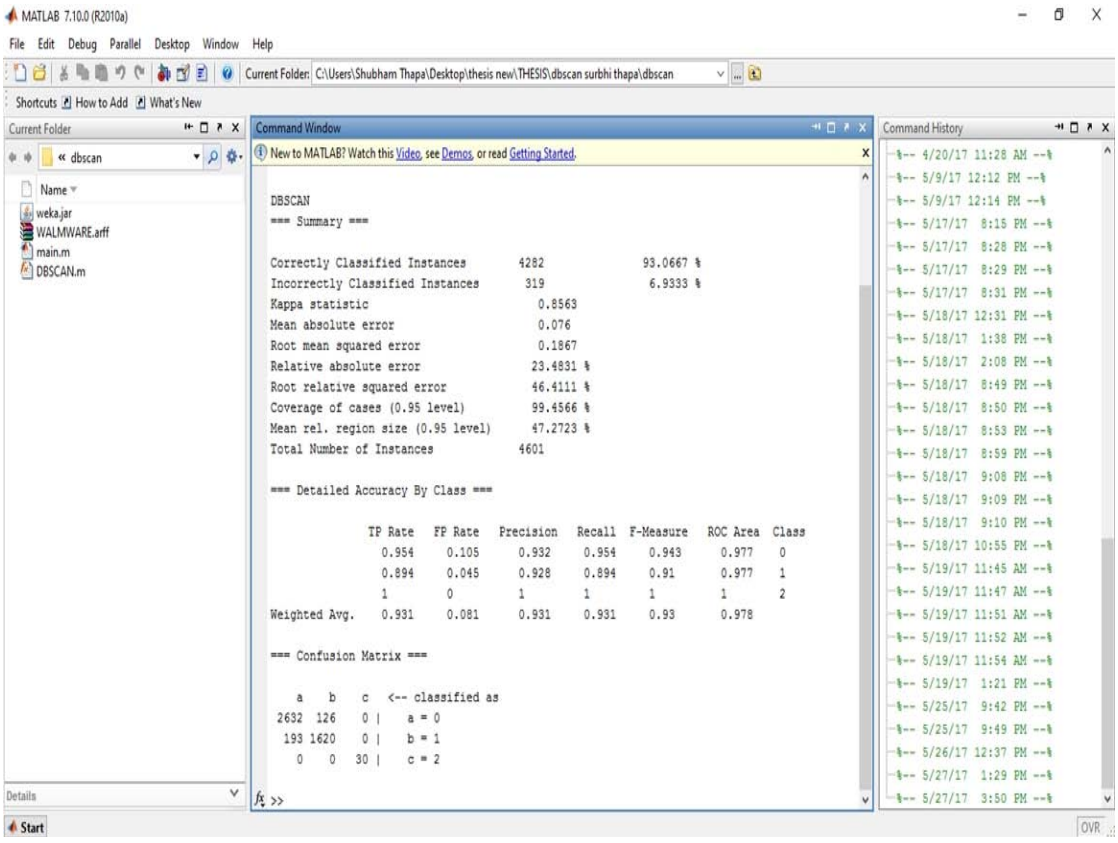

Fig.2 Snapshot of DBSCAN

Above figure 2 is snapshot of text mining technique for malware detection using DBSCAN. By applying DBSCAN on malware data we will get ACC93.0667%, F-measure 0.93%, TP rate0.931%, FP rate0.081%, AUC99.4566%, RAE23.4831% and ER 6.9333%.

5.2SNAPSHOT OF PROPOSED TECHNIQUE

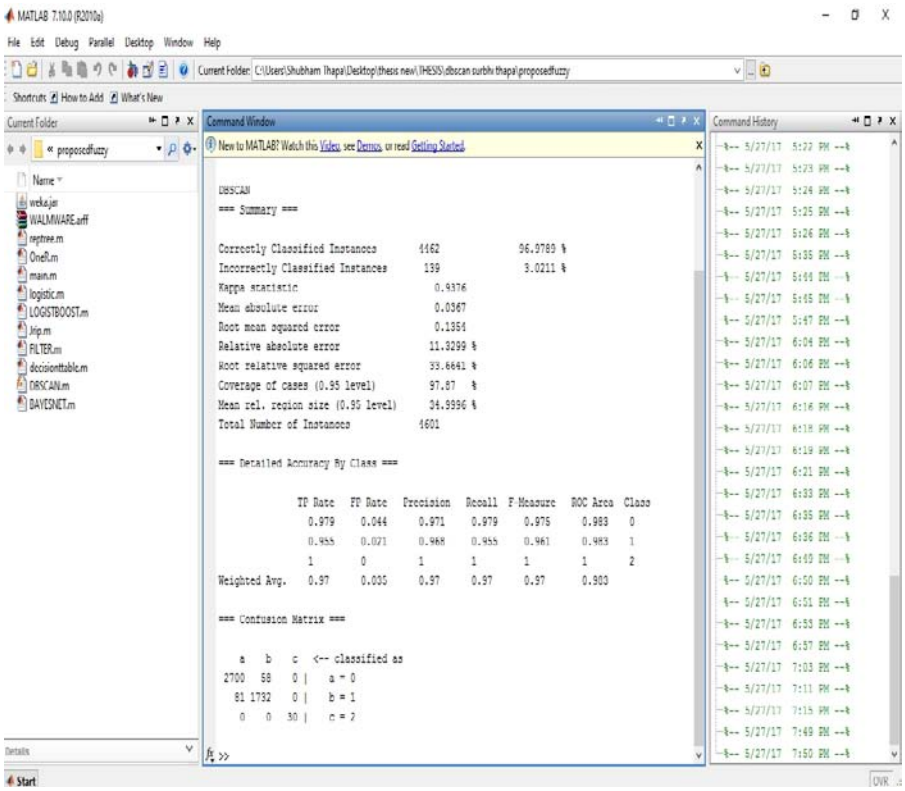

Fig.3 Snapshot of fuzzy based DBSCAN

As we can derive from above figure 3 that by applying fuzzy based DBSCAN on given dataset there is improvement in accuracy and other parameters. By applying fuzzy based DBSCAN on malware data we will get ACC 96.9789%, F-measure 0.97%, TP rate 0.97%, FP rate0.035%, AUC 97.87%, RAE 11.3299% and ER3.0211%.

VI. RESULT IN TABULAR

Table 1: Comparison of ACC, AUC, RAE and ER

	ACC	AUC	RAE	ER
Existing	93.0667	99.4566	23.4831	6.9333
Proposed	96.9789	97.87	11.3299	3.0211

Table 1 represents the comparison between existing and proposed method. This shows that in proposed method there is huge difference in ACC, AUC, RAE, and ER and in proposed technique ACC is more.

Table 2: Comparison of FP Rate, TP Rate and F-measure

	FP Rate	TP Rate	F-measure
Existing	0.081	0.931	0.93
Proposed	0.035	0.97	0.97

Table 2 represents the comparison between FP Rate, TP Rate and F-measure of existing and proposed method. This presents that in proposed method F-measure and TP Rate is higher and FP Rate is lower.

VII. ANALYSIS OF RESULTS

In figure 4 graphs of both the techniques older and new are compared. After comparison it is found that proposed algorithm will have greater ACC, and lower AUC, RAE, and ER.It represents that algorithm suggested here is superior to existing one.

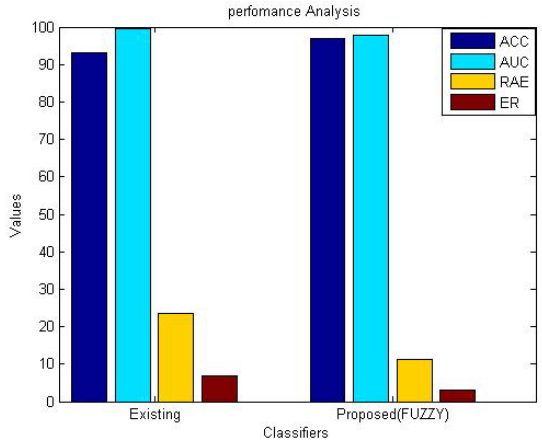

Fig.4: Comparison of ACC, AUC, RAE and ER

In figure 5graphs of both the techniques older and new are compared. After comparison it reveals that proposed algorithm will have greater F-measure, TP rate and lower FP rate. Thus proposed technique is comparatively better.

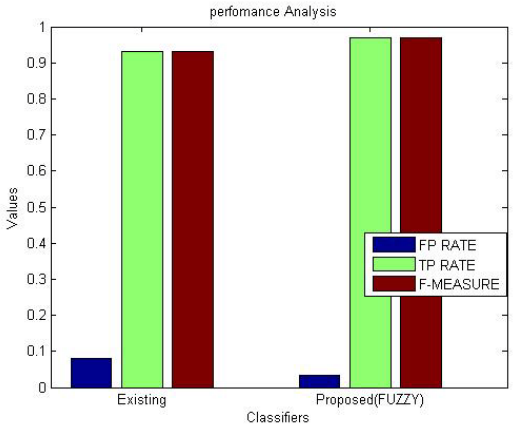

Fig.5:Comparison of FP Rate, TP Rate and F-measure

VIII. CONCLUSION

Data mining is a field of intersection of statistics and computer science used to discover patterns in the information bank. The objective of the data mining procedure is to extricate the particulars from the collection of data and modify it into a fathomable structure for future use. Different processes and approaches operate data mining successfully. Text mining is the method of determining and identifies the patterns from text and also called knowledge discovery in text bases, the method of finding interesting and relevant relations and patterns. It has been noticed that in existing work has introduced method persists until the density-connected cluster is entirely found. Then, a new obscure element is extracted and processed, leading to the detection of a further cluster or noise. The DBSCAN clustering is vital method for data mining. DBSCAN was proposed to adopt density-reach ability and density connectivity for handling the randomly shaped clusters and noise. Therefore this thesis work has proposed to fuzzy based DBSCAN text mining technique for malware detection thus improves the ambiguity and functionality of the clusters formed. The proposed technique is designed and executed in the MATLAB 2013 a by using data analysis toolbox and weka. The simulation outcome shows that by applying hybrid pattern based text mining technique for malware detection using fuzzy based DBSCAN by using various parameters i.e. ACC, F-measure, TP rate, FP rate, AUC, RAE and ER. This proposed method shows better results than the present technique.

REFERENCES

[1] S.M. Junaid, K.V. Bhosle, "Overview of Clustering Techniques," International Journal of Advanced Research in Computer Science and Software Engineering, Vol.4, Issue 11, Nov 2014.

[2] S. Agarwal, "Data Mining Concepts and Techniques," International Conference on Machine Intelligence and Research Advancement, pp. 203-207, Dec 2013.

[3] R. Talib, M.K. Hanify, S. Ayeshaz, and F.Fatima, "Techniques, Applications and Issues," International Journal of Advanced Computer Science and Applications (IJACSA), Vol.7, Issue 11, Nov 2016.

[4] G.G. Sundarkumar, V.Ravi,"Malware Detection by Text and Data Mining,"IEEE Computational Intelligence and Computing Research (ICCIC),pp. 1-6, Dec 2013.

[5] A. Joshi, R. Kaur, "A Review: Comparative Study of Various Clustering Techniques in Data Mining," International Journal of Advanced Research in Computer Science and Software Engineering, Vol.3, Issue 3, March 2013.

[6] Y. Zhang, J. Pang, F. Yue, J. Cui,"Fuzzy Neural Network for Malware Detect," International Conference on Intelligent System Design and Engineering Application,pp. 780-783, Oct 2010.

[7] P. Rai, S. Singh,"A Survey of Clustering Techniques," International Journal of Computer Applications, Vol.7,Issue 12, Oct 2010.

[8] S.R. Pande, Ms. S.S. Sambare, V.M. Thakre,"Data Clustering Using Data Mining Techniques," International Journal of Advanced Research in Computer and Communication Engineering, Vol.1, Issue 8, Oct 2012.

[9] P. Mishra, N. Padhy, R.Panigrahi, "The survey of data mining applications and feature scope," Asian Journal of Computer Science and Information Technology, Vol.2, Issue 3, June 2012.

[10] M. Sukanyal, S. Biruntha, "Techniques on Text Mining," IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), pp. 269-271, Aug 2012.

[11] N. Kaur, A.K. Mann, "Survey Paper on Clustering Techniques," International Journal of Science, Engineering and Technology Research (IJSETR), Vol.2, Issue 4, April 2013.

[12] R. Kaur, G.S.Bhathal, "A Survey of Clustering Techniques," International Journal of Advanced Research in Computer Science and Software Engineering, Vol.3, Issue 5, May 2013.

[13] A. Smiti and Z.Eloudi, "Soft DBSCAN: Improving DBSCAN Clustering method using fuzzy set theory,"6th International Conference on Human System Interactions (HSI),pp. 380-385,June 2013.

[14] D. Sharma,"A Review on Clustering Techniques in Data Mining," International Journal of Advanced Research in Computer Science and Software Engineering, Vol.4, Issue 5, May 2014.

[15] S.H. Ganesh, A.J. Christy,"Applications of Educational Data Mining: A Survey," International Conference on Innovations in Information Embedded and Communication Systems (ICIIECS), pp. 1-6, March 2015.
[16] R.R. Tated, M. M. Ghonge, "A Survey on Text Mining-techniques and application," International Journal of Research in Advent Technology, Special Issue, 1st International Conference on Advent Trends in Engineering, Science and Technology "ICATEST 2015", Vol.1, pp. 380-385, March 2015.
[17] K. Shah, D.K. Singh, "A Survey on Data Mining approaches for Dynamic Analysis of Malwares," Green Computing and Internet of Things (ICGCIoT), pp. 495-499, Oct 2015.
[18] A. Malhotra, K. Bajaj, "A hybrid pattern based text mining approach for malware detection using DBSCAN,"Special Issue Redset (CSIT), Vol.4, Issue 2, pp 141–149, Dec 2016.
[19] A. Shalaginov, L.S Grini, K.Franke, "Understanding Neuro-Fuzzy on a class of multinomial malware detection problems,"Neural Networks (IJCNN), pp. 684-691, July 2016.
[20] A. Malhotra, K. Bajaj,"A Survey on Various Malware Detection Techniques on Mobile Platform," International Journal of Computer Applications, Vol.139, Issue 5, April 2016.