2019 International Conference on Computing, Power and Communication Technologies (GUCON)
Galgotias University, Greater Noida, UP, India. Sep 27-28, 2019

Analysis of Global Terrorism Dataset using Open Source Data Mining Tools

Mansi Goel
Dept. of CS & IT
Jaypee Institute of Information Technology
Noida, India
goel.mansi23@gmail.com

Nikitasha Sharma
Dept. of CS & IT
Jaypee Institute of Information Technology
Noida, India
nikkisharma827@gmail.com

Mahendra Kumar Gurve
Dept. of CS & IT
Jaypee Institute of Information Technology
Noida, India
manu.manit@gmail.com

***Abstract*— Terrorism is an escalating global agitation over the years despite the development. Since the attacks of 9/11, the discipline of terrorism has seen tremendous extension and acquired attention from divergent communities worldwide. This paper examines the Global Terrorism Database that encompasses statistics on domestic and international terrorist attacks since 1970s and yields its analysis and prediction repercussions using open source data mining tools. we developed the model using orange data mining tool to apply different machine learning algorithms such as SVM, Naive Bayes, Neural Networks and KNN, it further compares their analysis results and uses the most suited results for prediction based on different attributes of dataset.**

Keywords*— Terrorism, Data Mining, Analysis, Prediction, Machine Learning.

I. INTRODUCTION

Terrorism, the indiscriminate violence as a means to create terror among the citizen of world. Terrorism is increasingly seen as the most consequential, disarranged and destructive nuisance of life. Research on terrorism is not science, it encompasses people which are ruined and controlled by the processes. Terrorism is the reverberation of psychopathy and mental-illness. Terrorism is a major issuance in the world, and could occur everywhere in the world. With the expansion of terrorism, law enforcement system demands advanced data analytical tools and techniques to enhance understanding of terror activities.

Data mining techniques and tools allow researchers to analyze different data. Data mining is the process of uncovering the concealed and interested information from the data and constitutes it into some meaningful form [11, 12, 13]. Since the inception of technology and with development in data mining field, lot of commercial and open source tools comes into picture which has made complex data mining procedures simpler. Weka[15], Rapid Miner[16], Knime[17] and Orange[18] are the examples of open source data mining tools.

Classification is the most important and commonly used class prediction technique in data mining.

Different classification techniques have been previously used in various applications like - weather forecasting, financial, health care, medical and business intelligence [14, 12].

This paper aims to examine the Global Terrorism Dataset that encompasses statistics on domestic and international terrorist attacks since 1970s and yield its analysis and prediction repercussions using open source data mining tool.

This research conducts the statistical analysis by summarizing 5 decades of real time global terrorism dataset into some few graphs and figures. This research also predicts crime categories of terror attacks by applying KNN, Naïve Bayes and SVM data mining classification methods and further compares their analysis results and uses the most suited results for prediction based on different attributes of dataset.

The description of the work done as detailed below. Section 2, discusses the past related work done in this research domain. Section 3, illustrates the used dataset and its properties. Section 4 and 5 describe the model building process and methods respectively. In section 6 discusses the experimental results of the classification methods for predicting the crime category. Finally, section 7 covers conclusion and future aspects.

II. RELATED WORK

In the research field, a lot of emphases is upon the use of web mining during analysis of data. The various algorithms like association analysis, classification, clustering, regression prediction, naive bayes, random forest, etc work well in identifying the patterns from the data. Newly invented modern techniques are also able to find the patterns or sequences from both types of data, structured as well as unstructured data. Entity extraction is one such example which enables to recognize particular patterns from data. To maximize or minimize interclass similarity, clustering techniques are used. Association rule mining helps to spot frequently occurring item sets while link analysis techniques can identify similar transactions [1]. The data is visualized from various perspectives and the results are verified by analyzing them with the help of various dependencies [3]. The data set gathered, Global Terrorism Database (GTD) is used which comprises of various sources like CETIS (Centre for Terrorism and Intelligence Studies) and the PGIS (Pinkerton Global Intelligence Service), from where it has been collected.

The strategies employed during gathering and validation of data by improving its quality and comprehensiveness are highlighted [4]. Zequin Shen mentioned that anatomization of social network analysis is a field of study which focuses on the structural linking for its results and it is far beyond the area of sociology [5]. To get precise results, data is filtered. Visualization when combined with analysis provide us the results of whole global network of terrorism [6]. Using effective tools and techniques provides the freedom to focus on the region of interest rather than looking into whole data. Fish eye view technique is used to enlarge the region of interest

978-93-5351-098-5/19/$0.00@2019 Galgotias University

and fractal view focuses more on information reductions [6]. Jialun‟s study based on the reliable data collected in a large scale. On studying about the terrorist network, knowledge about terrorism organizations are gained which help to develop efficient strategies against the terrorism [7]. Daning Hu introduced the web structural mining techniques into the terrorist network analysis field. The results from the analysis provide the empirical implications that will help in business intelligence and security communities to make the nation safer [7].

Terrorism threats have a wide range that spans personal, organizational, and societal levels along with political, economic and social consequences [8]. The first challenge of terrorism is associated with information gathering, searching, management and knowledge creation based on the data. The second aspect of challenges focus on how to trace the evolution of terrorist groups and organizations and how to analyze and predict their activities. The last aspect of challenge is how to give the access to public systems and students about terrorism research [8]. Multimedia has been one alternate option while extracting information about terrorist groups. Algorithm development for future terrorist attacks by simulation of past attacks by real data. Producing scenarios helps enhance performance by generating random networks [9]. Some of the researches also focus on the after effects of people who experienced a terrorist attack. Voxel-based morphometry is an automated method which allows us to explore the structural changes right through the brain [10].

III. DATASET

The data used in this research work is Global Terrorism dataset. Global Terrorism dataset was downloaded from www.kaggle.com. The dataset had 170000 data entries of terrorist attacks all over the world from 1972 to 2016.

There were some missing and null values in the dataset. To make the data clean for further analysis, all rows with missing and duplicate values entries were removed from the dataset.

Used analysis tool doesn't support large dataset so we have considered approximately 35000 terrorist attacks from 2010 to 2016 for analyses using different data mining algorithms. Table 1 shows the considered dataset attributes.

TABLE I. DATA SET ATTRIBUTES

S. No	Attribute	
1	Year	Year of Attack
2	Month	Month of Attack
3	Day	Day of Attack
4	Country	Country Which Has Been Attacked
5	Region	Region of Country Attacked
6	State	State Attacked
7	City	City Attacked
8	Weapon	Weapon Used for Attack
9	Target Type	Target of Attack (e.g. govt, NGOs)

IV. MODEL BUILDING

To meet the objective of research, we developed the data analysis model using open source tools. The tools and techniques used in the analysis of terrorism dataset are listed below:

A. Orange

Orange is an open-source machine learning and data mining tool. The analysis of the dataset is done with inbuilt of machine learning and data mining algorithms like Naive Bayes, KNN (k-nearest neighbors), SVM (Support Vector Machine) and Neural Networks (NN). In orange, a workflow is created where the file widget is selected and data is loaded into it. The attributes are then set as target, feature or meta value. Data File is attached with other widgets by tracing a string between them. In figure-1, different data mining models are applied onto the dataset and their test and score is calculated. This score further helps in prediction and creating confusion matrices. Confusion matrix is used to achieve the representation between actual and predicted data and hence tells amount of data classified correctly.

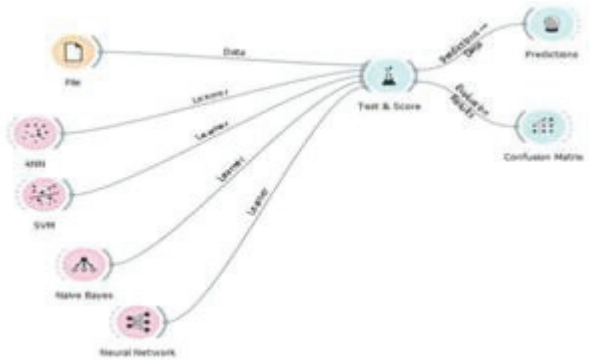

Fig. 1. Model Building using Orange Tool

B. Python

Python is a tool easy to understand and implement machine learning. The libraries of python such as Pandas, Matplotlib, NumPy, Seaborn and Plotly have been used for the statistical analysis dataset. it helps to represent and analyze results graphically.

V. METHODS

The following data mining methods are used to analyze crime dataset:

A. SVM

Support Vector Machines are supervised learning models that analyzed the data used for classification and regression challenges and create a hyperplane or set of hyperplanes in a very high- dimensional space that distinguishes the two or more classes. By using the hyperplanes, a better separation is attained, the higher the margin the smaller the error of the classifier. SVM can convert the non-separable problem to separable problem by using kernel functions. These functions take low dimensional input space, converts to a high dimensional space.

B. Naive Bayes

The Naive Bayes classifier techniques are based on the Bayes Theorem and are part of probabilistic classifier family with naive independent assumptions between the predictors.

Using Bayes Theorem, the conditional probability can be decomposed as:

$$P(C_k|X) = \frac{P(C_k)\,P(X|C_k)}{P(X)} \qquad (1)$$

C. K-Nearest Neighbors

It is used for classification and regression. It is instance-based learning or lazy learning. It requires 3 things feature space, distance metric to compute distance between data and records and the value of k (*k* is a positive integer, commonly small), number of nearest neighbours. The input consists of K nearest training example. The output of KNN classifier is based on class membership. An object is arranged by the larger number of its neighbors, the object is assigned to that class, which is most common among its *k* nearest neighbors. In KNN regression, the output of the object is its property value. The calculated mean of its *k* nearest neighbors is termed as property value.

D. Neural Networks (NN)

Artificial NN is computing system inspired by NN (neural networks). It is a framework for collection of numerous data mining and machine learning algorithms that work together to process provided data inputs. These algorithms work same as the human brain would. The applications of Artificial NN includes the computer vision, machine translation, speech recognition, playing board, social network filtering, medical diagnosis etc.

E. Confusion Matrix (Error Matrix)

It is a matrix which is used to define accuracy and performance of a classification model on a set of testing data for known true values.

- *Recall* (True Positive Rate) - It gives the rate of number of the correct positive predictions to the sum of number of correct positive and negative predictions.

  Recall= TP / (TP + FN)
- *Precision* (Positive Predictive Value) - It says when it predicts True, how often it is correct.
  Precision= TP / (TP + FP)
  where TP, FP and FN stand for True Positive, False Positive and False Negative respectively.
- *F score*- It is also known as F measure and is the weighted average of true positive recall and precision. It measures the test's accuracy.

  F score= 2 (Precision * Recall) / (Precision + Recall)
- *AUC Curve*- Area under ROC curve. ROC curve gathers information about all possible thresholds.

VI. Results and Discussions

We conducted analysis on Global Terrorism Dataset using python script and portrayed the results using graphs.

Figure 2. brief the plot for number of terrorist activities per year and it was found that year 2014 had the most frequent terrorist attacks.

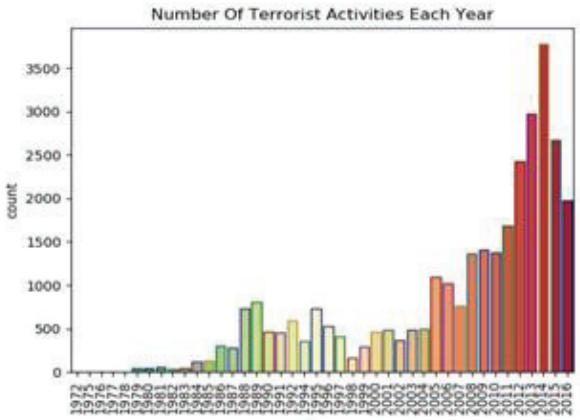

Fig. 2. Terrorist Activities per Year

Figure 3. provides the details of the most frequent targets of the terrorists and it was found out to be private properties and Citizens. It is a really helpful result as it gives an exact idea which sector needs to be more secure to prevent further terrorist attacks.

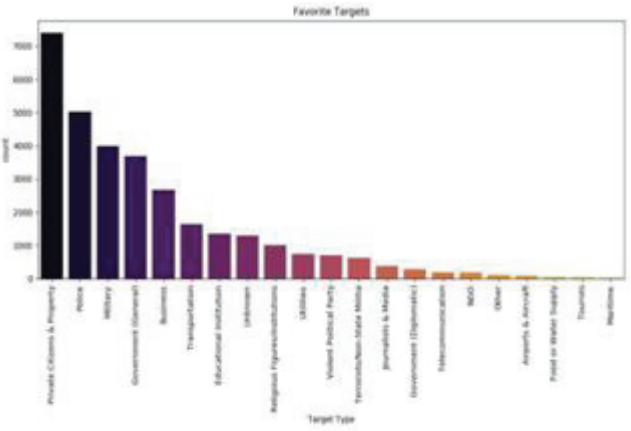

Fig. 3. Most Frequent Targets

An interesting result that caught our eye was that April and May were the months with most terrorist attacks while December being the least. This information can be very helpful for the security agencies to be more alert during April and May. It is depicted in figure 3.

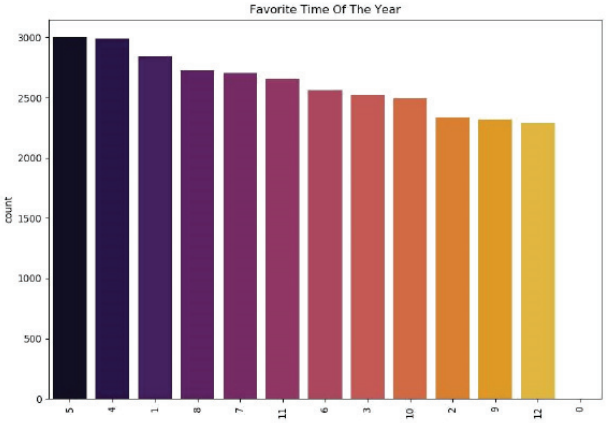

Fig. 4. Most Vulnerable Time

Figure 5. gives the count of attacking methods that were commonly used by terrorists. As it can be seen in figure, bombing/explosion method has the highest number of count and was commonly practiced by the terrorists for attacks.

For prediction purpose, we applied the KNN, SVM, Naive Bayes and Neural Networks data mining methods on dataset using Orange tool. Here we are analyzing the results for two different target values i.e. Attack Type and Target Type by the terrorists. The results provided by all the four algorithms are more or less similar. This verified the obtained results are probably accurate. Now we are discussing the results one by one for each target values.

Table II. gives the evaluation results of prediction by different methods, when target value was set to be „Attack Type". As it can be seen in the figure, neural network has the highest precision value.

TABLE II. Comparative Analysis of Algorithms

Method	AUC	CA	F1	Precision	Recall
kNN	0.512	0.274	0.285	0.444	0.274
SVM	0.456	0.100	0.072	0.066	0.100
Neural Network	0.595	0.490	0.337	0.333	0.490
Naïve Bayes	0.597	0.491	0.339	0.413	0.491

After doing comparative analysis of different algorithms on precision and recall value when target value set to be Attack Type.

Now we are comparing their produced confusion matrices. Table 3 to 6 gives comparative results of confusion matrix produced by these algorithms.

TABLE III. Confusion Matrix using KNN

Actual/ Predicted	Armed Assualt	Assassination	Bombing	Facility/ Infrastr	Hijacking	Hostage
Armed Assualt	**1878**	0	0	0	0	**2887**
Assassination	**971**	0	0	0	0	**1125**
Bombing	**3930**	0	0	0	0	**4338**
Facility/ Infrastr	556	0	0	0	0	695
Hijacking	19	0	0	0	0	35
Hostage	23	0	0	0	0	28
Hostage	467	0	0	0	0	1063
Unarmed Assualt	70	0	0	0	0	90

TABLE IV. Confusion Matrix using SVM

Actual/ Predicted	Armed Assualt	Assassi-nation	Bombing	Facility/ Infrastr	Hijacking	Hostage
Armed Assualt	924	0	**3744**	95	0	0
Assassi-nation	401	1	**1676**	17	0	0
Bombing	1002	0	**7206**	54	0	0
Facility/ Infrastr	13	0	816	297	0	0
Hijacking	8	0	45	1	0	0
Hostage	8	0	41	2	0	0
Hostage	8	0	41	2	0	0
Unarmed Assualt	12	0	123	17	0	0

TABLE V. Confusion Matrix using Naïve Bayes

Actual/ Predicted	Armed Assualt	Assassi-nation	Bombing	Facility/ Infrastr	Hijacking	Hostage
Armed Assualt	28	10	**7986**	97	0	0
Assassination	1	109	**3077**	18	0	0
Bombing	25	75	**14982**	60	0	0
Facility/ Infrastr	23	12	1191	306	0	0
Hijacking	0	0	68	1	0	0
Hostage	5	0	76	2	0	0
Hostage	2	2	2241	6	0	0
Unarmed Assualt	6	4	147	25	0	0

TABLE VI. Confusion Matrix using Neural networks

Actual/ Predicted	Armed Assualt	Assassination	Bombing	Facility/ Infrastr	Hijacking	Hostage
Armed Assualt	28	15	**7986**	97	0	0
Assassi-nation	1	122	3077	18	0	0
Bombing	25	87	14982	60	0	0
Facility/ Infrastr	23	14	1191	306	0	0
Hijacking	0	1	68	1	0	0
Hostage	5	1	76	2	0	0
Hostage	2	16	2241	6	0	0
Unarmed Assualt	6	7	147	25	0	0

Table VII depicts the prediction results in a more elaborated and descriptive way for each used algorithm.

TABLE VII. Comparison of Prediction Values

Attack Type	City	kNN	SVM	Naïve Bayes	Neural Network
Bombing/ Explo.	Herat	Bombing/Explo.	Armed Assualt	Bombing/Explo.	Bombing/ Explo.
Bombing/ Explo.	Kabul	Bombing/Explo.	Armed Assualt	Bombing/Explo.	Bombing/ Explo.
Bombing/ Explo.	Kabul	Bombing/Explo.	Armed Assualt	Bombing/Explo.	Bombing/ Explo.
Bombing/ Explo.	Kabul	Bombing/Explo.	Armed Assualt	Bombing/Explo.	Bombing /Explo.
Bombing/ Explo.	Angaston	Bombing/Explo.	Armed Assualt	Bombing/Explo.	Bombing/ Explo.
Bombing/ Explo.	Kabul	Bombing/Explo.	Armed Assualt	Bombing/Explo.	Bombing/ Explo.
Bombing/ Explo.	Unknown	Bombing/Explo.	Armed Assualt	Bombing/Explo.	Bombing/ Explo.
Bombing/ Explo.	Kabul	Bombing/Explo.	Armed Assualt	Bombing/Explo.	Bombing/ Explo.
Bombing/ Explo.	Unknown	Bombing/Explo.	Armed Assualt	Bombing/Explo.	Bombing/ Explo.

Table VIII shows the evaluation results of prediction by applied methods, when target value was set to be "Target Type". As it can be seen in the table, neural networks have the highest precision value.

TABLE VIII. Comparative Analysis of Algorithms

Method	AUC	CA	F1	Precision	Recall
kNN	0.513	0.148	0.113	0.143	0.148
SVM	0.549	0.035	0.017	0.026	0.035
Neural Network	0.604	0.249	0.157	0.178	0.249
Naïve Bayes	0.604	0.249	0.157	0.158	0.249

After doing comparative analysis of different algorithms on precision and recall value when target value set to be Attack Type. Now we are comparing their produced confusion matrices. Table IX to XII gives comparative results of confusion matrix produced by these algorithms.

TABLE IX. Confusion Matrix using KNN

Actual/ Predicted	Airport/ Aircraft	Business	Education	Food/ Water	Journal
Airport/ Aircraft	0	39	0	0	0
Business	0	**1336**	0	0	0
Education	0	**936**	0	0	0
Food/ Water	0	**27**	0	0	0
Journal	0	156	0	0	0
Maritime	0	6	0	0	0
Military	0	1614	0	0	0

TABLE X. Confusion Matrix using SVM

Actual/ Predicted	Airport/ Aircraft	Business	Education	Food/ Water	Journal
Airport/ Aircraft	0	30	0	0	0
Business	0	**381**	0	0	0
Education	0	**128**	0	0	0
Food/ Water	0	**9**	0	0	0
Journal	0	136	0	0	0
Maritime	0	25	0	0	0
Military	0	1439	0	0	0

TABLE XI. Confusion Matrix using Naïve Bayes

Actual/ Predicted	Airport/ Aircraft	Business	Education	Food/ Water	Journal
Airport/ Aircraft	0	1	0	0	2
Business	0	**236**	0	0	4
Education	0	**7**	0	0	0
Food/ Water	0	**2**	0	0	0
Journal	0	9	0	0	0
Maritime	0	0	0	0	0
Military	0	10	0	0	0

TABLE XII. Confusion Matrix using Neural Networks

Actual/ Predicted	Airport/ Aircraft	Business	Education	Food/ Water	Journal
Airport/ Aircraft	0	0	0	0	3
Business	0	**218**	0	0	70
Education	0	**7**	0	0	11
Food/ Water	0	**2**	0	0	2
Journal	0	7	0	0	25
Maritime	0	0	0	0	1
Military	0	9	0	0	193

Table XIII depicts the prediction results in a more elaborated way.

TABLE XIII. Comparison of Prediction Values

Target Type	City	kNN	SVM	Naïve Bayes	Neural Network
Business	Herat	Religious	Business	Police	Police
Utilities	Kabul	Religious	Business	Police	Police
Business	Kabul	Religious	Business	Police	Police
Business	Kabul	Religious	Business	Police	Police
Private Citizens	Kabul	Religious	Business	Police	Police
Business	Kabul	Religious	Business	Police	Police
Private Citizens	Kabul	Religious	Business	Police	Police
Violent Political	Angaston	Religious	Business	Police	Police
Private Citizens	Kabul	Religious	Business	Police	Police
Government	Kabul	Religious	Business	Police	Police

VII. Conclusion

In this research, the main objective was to analyze the Global Terrorism Dataset, and produce some beneficial and interesting results. The analysis and evaluation are done using the Orange data mining tool. Orange facilitates to analyse Global Terrorism Dataset using different data mining methods and compare their results. The benefit of this tool is that, according to the attributes chosen, the best classifier can be found to produce more accurate results for future references and it can help to achieve global security. Being a paper of national importance and catering to global problems, this paper can be further used on a larger scale to look into the future of terrorism. The data can be drilled down to look specifically for India particularly, or could be rolled up and viewed on a global level as well. The data set can be diced or sliced too, based on what kind of information needs to be extracted. This paper can be eventually expanded and worked upon.

References

[1] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin & M. Chau,"Crime data mining: a general framework and some examples." computer, 37(4), 50-56, Source: doi.ieeecomputersociety.org(May, 2004).

[2] S. Ressler, "Social Network Analysis as an Approach to Combat Terrorism: Past, Present, and Future Research." *Homeland Security Affairs* 2, Article 8 (July 2006).

[3] G. LaFree, "The global terrorism database: Accomplishments and challenges", Perspectives on Terrorism, 4(1) (2010).

[4] X. Cao (Master University of Data Science, NewYork University), "Global Terrorism Analysis: An Interactive Tool for Visual Analysis of Global Terrorism", February 2017.

[5] Z. Shen, K. L. Ma & T. Eliassi-Rad, "Visual analysis of large heterogeneous social networks by semantic and structural abstraction", IEEE transactions on visualization and computer graphics, 12(6), 1427-1439, (2006).

[6] C. C. Yang, N. Liu & M. Sageman, "Analysing the terrorist social networks with visualization tools", In International Conference on Intelligence and Security Informatics (pp. 331-342). Springer, Berlin, Heidelberg (2006, May).

[7] J. Qin, J. J. Xu, D. Hu, M. Sageman & H. Chen, "Analysing terrorist networks: A case study of the global Salafi jihad network", In International Conference on Intelligence and Security Informatics (pp. 287-304). Springer, Berlin, Heidelberg (2005, May).

[8] E. Reid, J. Qin, W. Chung, J. Xu, Y. Zhou, R. Schumacher & H. Chen, "Terrorism knowledge discovery paper: A knowledge discovery approach to addressing the threats of terrorism". In

International Conference on Intelligence and Security Informatics (pp. 125-145). Springer, Berlin, Heidelberg (2004, June).
[9] C. Weinstein, W. Campbell, B. Delaney, & G. O'Leary, "Modelling and detection techniques for counter-terror social network analysis and intent recognition". In Aerospace conference, 2009 IEEE (pp. 1-16). IEEE.
[10] H. Yamasue, K. Kasai, A. Iwanami, T. Ohtani, H. Yamada, O. Abe, N. Kuroki, R. Fukuda, M. Tochigi, S. Furukawa, M. Sadamatsu, T. Sasaki, S. Aoki, K.Ohtomo,N. Asukai, and N. Kato "Voxel-based analysis of MRI reveals anteriorcingulate gray-matter volume reduction inposttraumatic stress disorder due to terrorism".
[11] Kochar B, and Chhillar R (2012). An Effective Data Warehousing System for RFID using Novel Data Cleaning, Data Transformation and Loading Techniques. Arab Journal of Information Technology, vol 9(3), 208–216
[12] Santhi P, and Bhaskaran V. M (2010). Performance of clustering algorithms in healthcare database, International Journal for Advances in Computer Science, vol 2(1), 26–31.
[13] Wahbeh A H, Al-Radaideh Q A, et al., (2011). A comparison study between data mining tools over some classification methods, International Journal of Advanced Computer Science and Applications, Special Issue, 18–26.
[14] Li G, and Wang Y (2012). A privacy-preserving classification method based on singular value decompo-sition, Arab Journal of Information Technology, vol 9(6), 529–534.
[15] https://www.cs.waikato.ac.nz/ml/weka/
[16] https://rapidminer.com/
[17] https://www.knime.com/
[18] https://orange.biolab.si