Application of Data Visualization and Machine Learning Algorithms for Better Decision Making

Olta Llaha [1], Azir Aliu [2]

[1,2] *South East European University, Tetovo, North Macedonia*

Abstract

Nowadays, huge and accessible data is an ever increasing field of study. The change in technology is, in turn, increasing its degree of interactivity, configuring several scenarios of great complexity in which data is understood on the basis of our interaction with it at different levels. Data visualization involves presenting data in graphical or pictorial form which makes the information easy to take in. It helps to explain facts and determine courses of action. Criminology is an interesting application where data visualization plays an important role in terms of prediction and analysis. Crime analysis plays an important role in devising solutions to crime problems and formulating crime prevention strategies. The purpose of this paper is to evaluate the performance of machine learning algorithms, which can be used for analyzing data collected of the past crimes. We identified the most appropriate machine learning algorithm to analyze the collected data from sources specialized in crime prevention. This study helps the institutions against crime to better predict and classify it.

Keywords

Data Visualization, Machine learning, Decision making

1. Introduction

Crime is a complex social phenomenon that has grown due to major changes in society. Law enforcement agencies need to learn the factors that lead to an increase in crime tendency. This study focuses on crime prevention, which is an important component of an overall strategy to reduce crime and to strengthen public safety. Decision Making in crime prevention has attracted a great concern and attention. Decision making is very important in crime prevention in order to decide accurate actions and law enforcement strategies. Law enforcement agencies face a large volume of data that needs to be processed and turned into useful information. Data visualization approach has been exposed to be a proactive decision-making concept in preventing and predicting crime. By processing criminal data, law enforcement agencies can use models that may be important in the crime prevention process. The Database Management System is designed for case management and overall crime counting and not for data analysis. The addition of machine learning algorithms to the database management system enables a system that functions as a criminology expert. Crime analysis can produce a superior result by integrating machine learning algorithms into Decision Support System (DSS). The DSS is required in crime analysis because it has the capability to improve the quality of decision making for crime prevention.

Figure 1 shows the Conceptual framework of DSS in crime prevention. In the system, the use of machine learning algorithms is also noted, which will make it possible to create models and as an output will determine knowledge about crimes, such as crime trends, crime location, etc.

Proceedings of RTA-CSIT 2023, April 26–27, 2023 Tirana, Albania
EMAIL: ol29064@seeu.edu.mk (A. 1);
ORCID: 0000-0002-7607-7413 (A. 1);

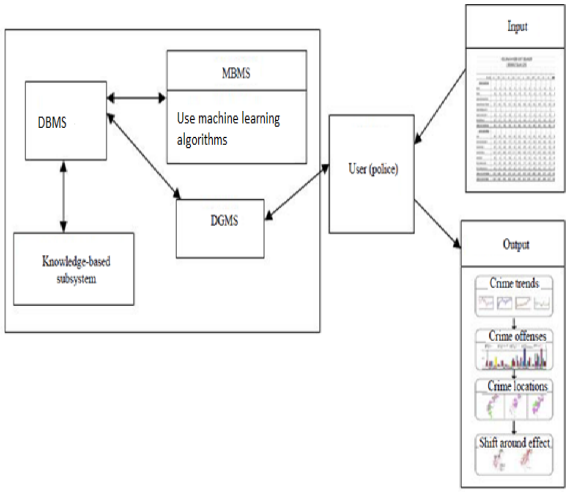

Figure 1: Conceptual framework of DSS in crime prevention [1]

2. Data visualization and machine learning for decision making

The primary purpose for data visualization is to assist people with processing large amounts of information. Data volumes are large and human cognitive capacities to remember and understand data are limited [2]. Data visualization should be made to simplify visualization as much as possible to help people make more effective decisions. Put simply, data visualization is a method of producing an output so that all problems and solutions can be clearly seen by the domain experts [3]. Data visualization plays an important role in terms of prediction and analysis.

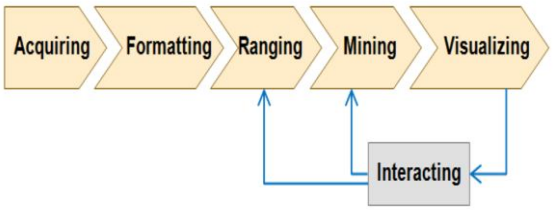

Figure 2. Data visualization process [4]

Figure 2 presents the process of data visualization which starts from acquiring the data then formatting it. By visualizing and interacting with the data we can predict the stages in which crime can happen so that decisions to prevent it are made on time.

Decision support systems (DSS) are defined as interactive application systems which are intended to help decision makers utilize data and models in order to identify problems, solve problems and make decisions. Decision support systems (DSS) are classically designed to serve the management level of organizations [5]. They help managers, in this case law enforcement agencies make decisions that are semi structured, unique or rapidly changing and are not easily specified in advance. DSS use sophisticated analysis and modeling tools. Data visualization and machine learning extend the possibilities for decision support by discovering patterns and relationships hidden in data and therefore enabling the inductive approach of data analysis [6].

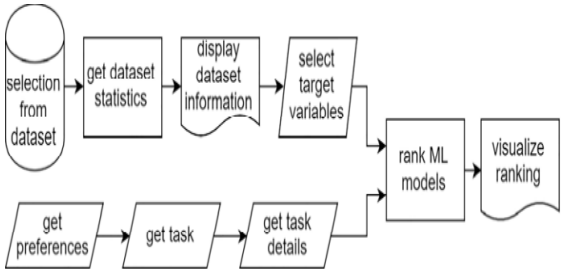

Figure 3. The Process Model of Data visualization and machine learning on DSS (General system operation) [7]

The process of DSS in general is shown in fig. 3. DSS supports the ability to import various training datasets and test ML models in .csv, .xls and .json formats.

To display the relationship between users and the system, a diagram of cases was compiled (fig. 4). The main unary scenarios of user interaction with the system are: selecting a data set, viewing data statistics, updating the task, viewing models rating and visualization.

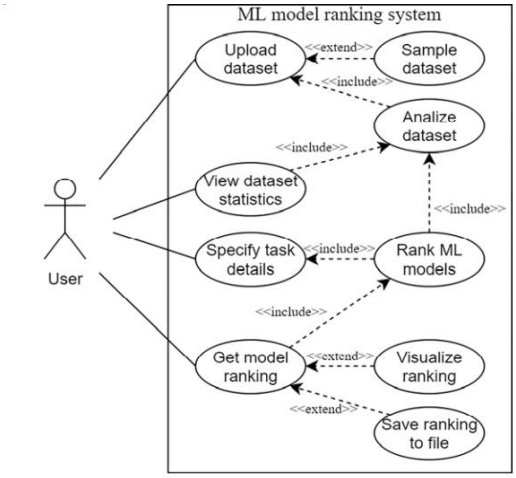

Figure 4. DSS use cases diagram [7]

The paper introduces data visualization and machine learning based on decision support systems. It is designed to enable the using of decision support systems by just having a basic level of knowledge on data visualization and machine learning algorithms.

3. Methodology

In this study we compare data visualization techniques and machine learning algorithms to discover the most suitable method or algorithm for our data. The machine learning algorithms are compared by applying them to the WEKA [8] environment. The implemented algorithms are: EM, COBWEB, DBCSCAN, Hierarchical Cluster, Make Density Based Cluster, K-Mean, Farthest First, Filtered Cluster. Algorithms have been applied to these data to determine their effectiveness in crime prediction and prevention. The data analyzed is extracted from the database of law enforcement agencies. The collected data is stored into database for further process. The number of instances or records is 90. There are 90 records because the data is very sensitive and we could not get more data from the law enforcement agencies, for this reason. This is also a limitation of this article. Reduced number of data due to their sensitivity.

Table 1

Dataset details

The name of the dataset	Number of examples	Number of input attributes	Number of possible classes	Total number of attributes	Values that are missing
Crime Dataset	90	6	2	7	0

The data relates to areas where crimes occur and to the information about the perpetrators. Some of the features we have considered are: the area where the crime occurred (urban or rural), age (from 17 to 55 years old), employment status (whether employed or not), gender, education (middle school, high school, university), civil status (whether married, single, or divorced) and whether the person who committed the crime was previously convicted or not. Crime dataset is in csv format.

3.1. Clustering

Clustering is a partitioning of data into groups of similar objects. Presenting data from a few of these groups certainly loses some details, but it achieves simplicity. It models the data according to his groupings. From a machine learning perspective, groups correspond to hidden patterns, group search is learned without supervision and the final system presents a data concept. Clustering is the main subject of active research in various fields such as statistics, pattern recognition and machine learning.

1. Hierarchical clustering methods

The method will create a hierarchical decomposition of a given set of data objects. Based on how the hierarchical decomposition is formed, we can classify hierarchical methods.

Agglomerative Approach is also known as Button-up Approach [9]. Here we begin with every object that constitutes a separate group. It continues to fuse objects or groups close together.

Divisive Approach is also known as the Top-Down Approach [9]. We begin with all the objects in the same cluster. This method is rigid, i.e., it can never be undone once a fusion or division is completed.

2. Partitioning based Methods

Partition methods move the instances from one group to another, starting from an initial partition. The partition algorithm divides data into many subsets. One of the most commonly used algorithms is EM (Expectation – Maximization) [10]. This algorithm tends to work with isolated and compact groups. The basic idea is to find a clustering structure that minimizes a certain error criterion, which measures the "distance" of each instance to its representative value.

K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid is at the minimum [11]. The less variation we have within clusters, the more homogeneous the data points are within the same cluster.

3. Density Based Methods

The basic idea of "density-based" methods is that for every instance of a group of zones near a given radius must contain a minimum number of instances. These methods identify the clusters and the distribution of their parameters. The

algorithms produce clusters in a determined location based on the high density of data set participants. The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm detects arbitrary groups and forms and is efficient for large databases. This algorithm is based on this intuitive notion of "clusters" and "noise" [11]. The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points.

4. Based Model Methods

These methods use a hypothesized model based on probability distribution. Model-based clustering methods find characteristic descriptions for each cluster, with each cluster representing a concept or class. The COBWEB algorithm yields a clustering dendrogram called classification tree that characterizes each cluster with a probabilistic description [12]. The algorithm assumes that all attributes are independent. It causes us to achieve a high predictability of the values of the nominal variables, given a set.

This paper uses some data visualization techniques such as charts and graphs.

a. Charts

What is the easiest method to display how one or more data sets develop? It is a chart, of course. Charts have a variety of forms, such as bar and line charts, which may show relationships between items over time. Pie charts can show how the elements or portions relate together within a whole.

A line chart is created by connecting data points within a data series using line segments. Line charts are frequently employed for showing trends in data that vary continuously over a period of time or range [13].

b. Graphs

The use of graphs provides a general means to transform the data and their relationships into an abstract view for showing complex relationships and improving data comprehension. Meanwhile, graphs can also be adjusted flexibly to answer specific questions based on the distinctive characteristics of the data [14]. We demonstrate the effectiveness of graph-based representations by applying them to our data.

4. Experimental results

To conduct this study we used WEKA software based on the approach and familiarity with its use. The WEKA software package has different programs for different techniques and algorithms. WEKA is a collection of machine learning algorithms and contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Table 2 presents a comparison of the results of the algorithms applied to our data in WEKA.

Table 2

Comparison of the results of the algorithms applied in WEKA

Algorithm	No. of groups	Clustered Instances	No. of iterations	Time to build a model	Log Likelihood	Non Clustered Instances
		0 52 (58%)				
EM	3	1 35 (39%)		0.5 s	-6.69022	0
		2 3 (3%)				
		3 1 (1%)				
		5 1 (1%)				
		6 1 (1%)				
		7 1 (1%)				
		8 1 (1%)				
COBWEB	102	10 1 (1%)		0.03 s		0
		11 2 (2%)				
		13 2 (2%)				
		14 1 (1%)				
		15 7 (8%)				
		19 1 (1%)				
		20 1 (1%)				
		22 1 (1%)				
		23 1 (1%)				
		25 5 (6%)				
		...				
		0: 12(34%)				
DBSCAN	3	1: 14(40%)		0.03 s		55
		2: 9 (26%)				
Hierarchical Cluster	2	0: 73(81%) 1: 17(19%)		0.02 s		0
Make Density Based Cluster	2	0: 75(83%) 1: 15(17%)	3	0.02 s	-7.43104	0
K-Mean	2	0: 75(83%) 1: 15(17%)	3	0 s		0
Farthest First	2	0: 69(77%) 1:21(23%)		0.02s		0
Filtered Cluster	2	0: 75(83%) 1: 15(17%)	3	0.02s		0

In this paper we used some algorithms (Table 2) and among them is K-mean algorithm. This algorithm provides clear results which are easy to interpret. Model construction is done by modifying the parameter values and this algorithm groups the crime data in less time to build the model. The K-mean algorithm was applied to these data. The visualization of this algorithm is shown in Figure 5.

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 133.52267774699902
Missing values globally replaced with mean/mode

Cluster centroids:
                                     Cluster#
Attribute                Full Data          0           1
                            (90)         (75)        (15)
=============================================================
Mosha                     33.1889      31.8267          40
Gjinia                          M            M           M
Statusi i punesimit     pa pune      pa pune     I/e punesuar
Vendi i krimit           rurale       rurale       urbane
Statusi civil         i/e martuar  i/e martuar  i/e martuar
Arsimimi                I mesem      I mesem      I larte



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       75 ( 83%)
1       15 ( 17%)
```

Figure 5. Result of K-means algorithm

The number of clusters is two (0 and 1) and the instances are grouped according to this scheme: Cluster 0 has 75 instances or 83% cluster, while cluster 1 has 15 instances or 17% cluster. The number of iterations is 3.

According to the results of the K-mean algorithm, the persons who commit the most crimes are jobless and with secondary education.

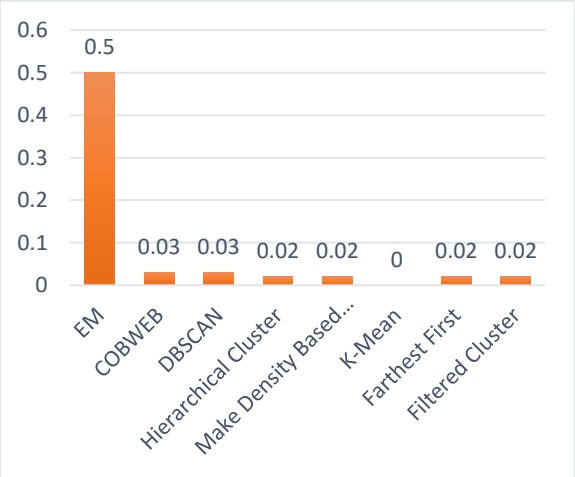

Figure 6. The time to build a model

The implementation of this algorithm has clustered the data in the least amount of time to build a model, exactly 0 seconds. This is shown in Figure 6.

Comparing this time with the time of other algorithms for the same number of instances, this algorithm has the shortest time, so it is faster.

5. Conclusions

This study presents a contribution in the field of data visualization, with a focus on crime data. This involves designing and creating a data set, which will be used in data visualization and machine learning using various visualization techniques and machine learning algorithms. This is about doing data visualization for those who want to have knowledge of the data, interpret it and take decisions from the data.

The results of the experiments performed in this study indicate that data visualization is applicable in the field of criminology. The K-mean algorithm aggregates the data in less time to build a model, compared to other algorithms. This algorithm shows promising results for the crime prevention problem because the accuracy rate is high in our experiments. The k-mean clustering algorithm is easy to interpret and simple to implement.

Decision making is very important in crime prevention in order to take accurate actions and build law enforcement strategies. Through our data analysis law enforcement agencies can create strategies, operating in areas where most crimes occur or for the perpetrators, their features (from our study were those who were unemployed and with secondary education). Data visualization techniques and machine learning contribute to predicting the likelihood of a crime occurring and as a result to prevent it.

6. References

[1] Maizura, Noor & Ab Hamid, Siti Haslini & Mohemad, Rosmayati & Jalil, Masita & Hitam, Muhammad. (2015). A Review on a Classification Framework for Supporting Decision Making in Crime Prevention. Journal of Artificial Intelligence. 8. 17-34. 10.3923/jai.2015.17.34.

[2] Mohd, Maseri & Abdullah, Embong & Mohamad Zain, Jasni. (2010). A Framework of Dashboard System for Higher Education Using Graph-Based Visualization Technique. 87. 55-69. 10.1007/978-3-642-14292-5_7.

[3] Donohoe, David & Costello, Eamon. (2020). Data Visualisation Literacy in Higher Education: An Exploratory Study of Understanding of a Learning Dashboard Tool. International Journal of Emerging Technologies in Learning (iJET). 15. 115. 10.3991/ijet.v15i17.15041.

[4] (Cho, Wonhee & Lim, Yoojin & Lee, Hwangro & Varma, Mohan & Lee, Moonsoo & Choi, Eunmi. (2014). Big Data Analysis

with Interactive Visualization using R packages. 10.1145/2640087.2644168.)
[5] Jantke, K.P., Memmel, M., Rostanin, O., Thalheim, B., & Tschiedel, B. (2003). Decision Support By Learning-On-Demand. CAiSE Workshops.
[6] Khademolqorani, Shakiba & Zeinal Hamadani, Ali. (2013). An Adjusted Decision Support System through Data Mining and Multiple Criteria Decision Making. Procedia - Social and Behavioral Sciences. 73. 388–395. 10.1016/j.sbspro.2013.02.066.
[7] Rudnichenko, N., Vychuzhanin, V., Petrov, I., & Shibaev, D. (2020). Decision support system for the machine learning methods selection in big data mining. International Workshop on Computer Modeling and Intelligent Systems.
[8] Frank, Eibe & Hall, Mark & Holmes, Geoffrey & Kirkby, Richard & Pfahringer, Bernhard & Witten, Ian & Trigg, Len. (2010). Weka-A Machine Learning Workbench for Data Mining. 10.1007/978-0-387-09823-4_66
[9] Z. Abdullah, A. R. Hamdan, Hierarchical Clustering Algorithms in Data Mining, World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering, Vol: 9, No: 10, 2015.
[10] Mohammed J. Zaki, Wagner Meira Jr. 2014, Data Mining and Analysis Fondamental Concepts and Algorithms, ISBN: 978-0-521-76633-3
[11] Amelio, Alessia & Tagarelli, Andrea. (2017). Data Mining: Clustering. 10.1016/B978-0-12-809633-8.20489-5.
[12] Pankaj SAxena & Sushma Lehri, 2017, Analysis of Various Clustering Algorithms of Data Mining on Health Informatics, International Journal of Computer & Communication Technology ISSN (PRINT): 0975 -7449, Volume-6, Issue-2, 2017.
[13] Gandhi, Parul & Pruthi, Jyoti. (2020). Data Visualization Techniques: Traditional Data to Big Data. 10.1007/978-981-15-2282-6_4.
[14] Fisher, Christian & Andersen, Fredrik & Darbyshire, Cole. (2021). Overview of Data Visualization.