2018 Sixth International Conference on Enterprise Systems

Big Data and its Usage in Systems of Early Warning of Traffic Accident Risks

Marián Lamr
Department of informatics
Technical university of Liberec
Liberec, Czech Republic
marian.lamr@tul.cz

***Abstract*—This article deals with the usage of big data in systems facilitating the warning of heightened risks of a traffic accident occurring. The article describes the possibilities of using traffic accident data collected by the Czech police and the data regarding the intensity of traffic on roads and motorways, which is provided by the Road and Motorway Directorate of the Czech Republic.**

Keywords—Big data, warning, system, data mining, car traffic, accident

I. INTRODUCTION

At the present time many organisations and enterprises collect a large amount of data. However, this digital data is not always efficiently used. In some cases the public administration collects a large amount of data with a specific purpose in mind, e.g. the generation of various statistics, yet this data could also be used for much more interesting purposes - for example predictions based on the collected data [1], [2], [3].

Even though in recent times the traffic accident fatality occurrence rate has been decreasing in Europe, the amount of traffic accidents has not; neither has the amount of accidents with light injuries [4]. However, there is a large number of countries where countless traffic accidents occur, entailing a large amount of traffic accident fatalities per 100,000 population per year. According to the WHO statistics from the year 2013, the international average is 17.3 victims per 100,000 population per year, the European average is 9.3 victims per 100,000 population. There are however also such countries as Libya (73.4 victims per 100,000 population per year), Thailand (36.2 victims per 100,000 population per year) and many others, where a very large number of traffic accident fatalities occurs [5], [6]. For this very reason, it is necessary to keep inventing new methods and approaches which could increase road safety. For example, one possibility is to use traffic accident data or traffic intensity data. In the Czech Republic for example, the traffic accident data is collected by the Czech police and the traffic intensity data is collected by the Road and Motorway Directorate of the Czech Republic [7], [8].

The objective of this article is to present the possibilities of using public data (which has originally been collected with a different purpose in mind) to predict the risks of a traffic accident occurring, in real-time. These principles can be generalized and be taken advantage of to use an early warning system in the Industry 4.0 environment.

II. EARLY WARNING OF HEIGHTENED TRAFFIC ACCIDENT OCCURRENCE RISKS SYSTEM

As has been mentioned earlier, the objective of this article is to prove that public big data can also be used for different purposes than has been originally intended. The traffic accident and traffic intensity data can e.g. be used to predict the risks of a traffic accident occurring, in real time and location. As an example, it is possible to use the traffic accident data in an early warning system, which would (using a special device inside of the vehicle) warn the driver in cases where the situation on the road at the time displayed a high degree of similarity to the prediction models based on past data.

As a component of this system, the traffic accident and traffic intensity data would be used to generate the prediction models. The concept of the Early warning of heightened traffic accident occurrence risks system can be described in one sentence as: a system predicting the risk of a traffic accident occurring in real time and location, and under other specific conditions describing the given situation. For the predictions, models generated based on data mining algorithms and past traffic accident data are used. More information about this system can be found in [9], [10].

The concept of the Early warning of heightened traffic accident occurrence risks system consists of two primary components. The user component of the system evaluates the situation in real-time and appropriately notifies the driver about the heightened risk of a traffic accident occurring. The control component of the system manages the collection, processing and distribution of the traffic accident data. A second task of the control component is the generation and distribution of the prediction models [9], [10], [11].

An important function of the control component of the system is to search for and evaluate clusters of traffic accident, which are also generated based on their GPS coordinates. The concept presumes that in each cluster generated using the appropriate cluster analysis algorithms, association rules amongst other attributes of the traffic accident record will be subsequently discovered. To search for clusters in geographical data, the cluster analysis based on density appears to be the most appropriate (e.g. DBSCAN, OPTICS, DENCLUE) [12], [13] [14]. The main advantage of these algorithms is that they can find clusters of any shape. Association rules adding in the detection of specific clusters from the perspective of further parameters, such as e.g. the cause of the accident or the weather conditions at the time, can then be found using e.g. the APRIORI algorithm [16]. A big advantage in using the DBSCAN and OPTICS algorithms is that they are able to

2572-6609/18/$31.00 ©2018 IEEE
DOI 10.1109/ES.2018.00031

detect "noise", which could entail accidents which do not belong into any cluster. Even though such algorithms as DBSCAN or OPTICS can more or less elegantly create clusters of traffic accidents of various shapes, their main issue lies in searching for clusters with differing density of traffic accident occurrence. For this reason we have decided to couple the traffic accident data with the traffic intensity data.

III. TRAFFIC ACCIDENT DATA

Traffic accident data is collected in a large number of countries. It is usually set up in such a way, that in case a traffic accident occurs, the police forces create a record consisting of several attributes describing the aspects of the accident. E.g. the Czech police has been populating the traffic accident database since the year 2007, which as of now contains more than 700,00 records and 46 attributes [8]. The traffic accident data is published by the police at the site www.jdvm.cz. Each traffic accident record can be downloaded as a PDF file. This format is not, however, appropriate for further processing of the data or for generating deeper analysis or data mining. To search for hidden correlations in the data, it is necessary to have a separate database, or to package the data in a structured form, for example in .csv files [15]. The Czech police does not provide a singular API which would facilitate the download of the traffic accident data. At the moment, the only way to transfer the data into its own database is through custom scripts written in the cURL language. In this manner we transfer the data from the police servers into our own internal database which is updated regularly.

Traffic accident data is also available in other countries. For example, the traffic accident data for the UK is freely available on the server data.gov.uk under the OGL license (Open Government License). The data can be downloaded in the CSV format. These files provide detailed road safety data about the circumstances of personal injury road accidents in GB from 1979, the types (including Make and Model) of vehicles involved and the consequential casualties [17]. All the data variables are coded rather than containing textual strings. The lookup tables are available in the "Additional resources" section towards the bottom of the table [17]. The structure of a traffic accident record is displayed on Fig. 1.

The detailed analysis of traffic accident data and the research of the possibility of using the individual attributes is a very important component of the Early warning of heightened traffic accident occurrence risks system. The necessary preparation and transformation of this data into a form appropriate for the creation of models has a large influence on the quality of the results gathered during the search for clusters of traffic accidents according to the location, and also on the generation of models serving the purpose of predicting the risk of traffic accident occurrence in real time and location. The data preparation phase is undoubtedly very time intensive, as is the case with all data mining projects.

ID	2100070013
Municipality	Praha
Region	Hlavní město Praha
Accident date	01.01.2007
Time	19:00:00
Date	Monday
Road type	Local road - other parking lot etc.
Road number	0
Accident culpability	Vehicle driver
Alcohol level of culprit	Yes, alcohol level lower than 0,99‰ (2)
People killed (number)	0
People severely injured (number)	0
People slightly inujred (number)	0
Accident type	Collision with a parked vehicle
Collision type of moving vehicles	Not applicable
Obstacle type	Not applicable
Main cause of accident	Improper turning or handling
Road surface type	Asphalt
Road surface condition	Wet surface
Road technical condition	Good, no defects
Weather conditions	Unhindered
Visibility	At night, artificial lightning, not hindered by weath
Visibile distance	Good
Lanes	None of the above
Accident location within the road	None of the above
Traffic management	None
Local right of way adjustments	No local adjustment
Specific objects at the location	Parking lot with exit to the road
Direction conditions	Straight section
Accident location	Outside the road
Crossing road type	Not specified
Drift	None
Driving or parking direction	Vehicle moving
Number of involved vehicles	4
Vehicle type	Passenger vehicle
Vehicle brand	ŠKODA
Manufacturing year	98
Vlastník vozidla	Private owned
Total damage (100 Kč)	450
Damage on vehicle (100 Kč)	130
Vehicle state after accident	No fire
Liquid leakage	None
Driver category	Driver licence B
Driver state	Under the influence of alcohol
Driver distraction	Driver not distracted
Lng	14.515
Lat	50.022

Fig. 1. The structure of a traffic accident record

IV. TRAFFIC INTENSITY DATA

As has been mentioned earlier, the control component of the system manages the search for clusters of traffic accidents using such algorithms as is DBDCAN or OPTICS. Using the algorithms on the data from e.g. the entirety of the Czech Republic is neither efficient nor well executable. Thus, the input parameters for the cluster searching algorithms must be appropriately set for various areas of search. To ensure that the algorithms search for relevant clusters of traffic accidents, it is necessary to separate locations by their traffic density. However, the data contains traffic accident records from all sorts of locations, such as big cities, motorways, first class roads or even backwater regional roads etc. When projected onto a map, a first glance is enough to tell us that that the character and shape of the clusters is completely different in e.g. motorways than it is in rural areas, and so forth. We can partially separate the locations with different density using only the attributes available in the traffic accident dataset provided by the Czech police. However, better results can be achieved during the search for traffic accident clusters, when the data matrix is coupled with the traffic intensity data, which is collected

once every few years by the Road and Motorway Directorate of the Czech Republic.

Again, the traffic intensity data is freely available online [7]. The Road and Motorway Directorate of the Czech Republic divides the individual roads into specific sections, due to the fact that traffic intensity differs even within the scope of a single road. This data is publicly available in the form of an interactive map which enables us to look up detailed data about any given sector, as gathered by the traffic census. More than 20 attributes describing the traffic intensity in any given sector are available.

Each traffic accident record needs to be coupled with a traffic intensity record. However, the coupling of the traffic accident data and traffic intensity data matrices carries one large obstacle with it. There are no common attributes which would allow these two data sets to be coupled by. We have however managed to overcome this obstacle - through custom scripts designed by us, we can query about each traffic accident at the server containing traffic intensity data. Each traffic accident record contains the GPS coordinates. For each traffic accident, a query is sent to the server asking whether there is any data about the traffic intensity in the area of the location given by us. If there is, the data is downloaded and assigned to the given traffic accident record. In the opposite case, the radius of the given traffic accident area is enlarged. In this manner we have managed to couple the traffic accident record data matrix with the traffic intensity data matrix. As an example, Figure 2 displays a part of the traffic density data for two traffic accidents.

ID	21013	21019
number_of_counted_sector	I.25	I.02
initial_sector_ULS	1242A01	1242A24
end_sector_ULS	1242A09	1242A24
start_delimitation_SU	1852	203
end_delimitation_SU	250	100
length_of_counted_sector_(m)	4516	5027
light_freight_vehicles	5922	2461
freight_vehicle_trailer_extensions	3580	4280
medium_freight_vehicles_curbweight_35_10t	5151	1140
medium_freight_vehicles_curbweight_35_10tTra	296	504
heavy_freight_vehicles_curbweightover10t	1232	415
heavy_freight_vehicles_curbweightover10tTraile	316	644
Buses	644	34
Buses_articulated	0	0
Tractors_without_trailers	0	0
Tractors_with_trailers	0	0
heavy_motor_vehicles_total	17141	9478
personal_and_freight_vehicles_without_trailers_	61859	24039
two_wheeled_vehicles	360	338
all_motor_vehicles_total	79360	33855
heavy_freight_vehicles	16533	13811
summer sunday intensity compared to the year-r	0.590000	0.000000
summer workday intensity compared to the year-	0.910000	0.000000
Cyclists (bicycle/day)	0	0
start_of_counted_sector_description	Chodov	Zbraslav
end_of_counted_sector_description	Pruhonice	Lochkov
number_of_administrative_unit	CZ010	CZ010
road_designation	D1	D0
road_class_code	1	2

Fig. 2. The structure of the traffic intensity data (of 2 accidents)

Figure 3 shows the histogram of the traffic density for the data from the "Královehradecký" and "Liberecký" regions regarding every road type. The graph makes the distribution of this quantity apparent. The largest group is formed by the traffic accidents with the traffic density of fewer than 10,000 vehicles per 24 hours. The largest group of traffic accidents with the traffic density of fewer than 5,000 vehicles per 24 hours is formed by the accidents occurring on 2nd and 3rd class roads. Further analysis of the data shows that e.g. 1st class roads fall within the density range of between 5,000 to 15,000 vehicles per 24 hours. An interesting category is formed by the accidents occurring on local roads - this category consists primarily of city roads. A significant portion of all traffic accidents occur exactly on these roads, with no regard to traffic density. However, this category shares the biggest portion of traffic accidents in locations with high traffic density.

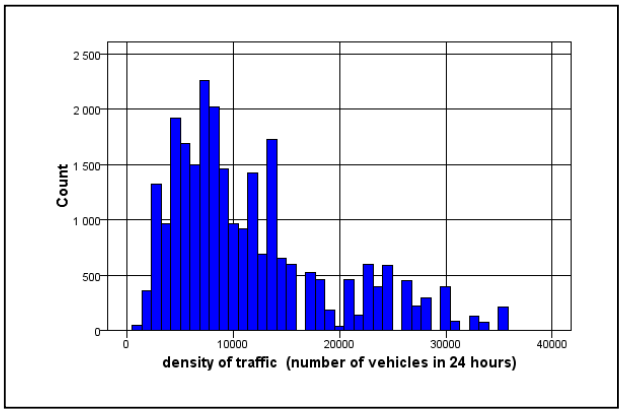

Fig. 3. Traffic density histogram

V. The Optimisation of Accident Cluster Detection Using Traffic Intensity Data

To showcase the optimisation of cluster detection using traffic intensity data, we have chosen the data from the Liberec city which is situated in the North of the Czech Republic. The city has more than 100,000 residents and it is the fifth largest city in the Czech Republic. The data matrix contains more than 9484 traffic accident records. Each cluster has its own colour. For clarity, the map intentionally omits accidents that do not belong to any cluster (noise).

The case of when the DBSCAN algorithm input parameters are exactly specified during the search for clusters can be seen on the accident cluster map displayed on Figure 4. The input parameters entail the following values: Epsilon = 20 meters, MinPts = 15 accidents.

Figure 5 shows a portion of the accident cluster map, which were detected using the DBSCAN algorithm, this time however while the input parameters have been dynamically altered using the traffic intensity data.

It is very apparent from figure 5, that when the input parameters of the algorithm are dynamically altered, the algorithm detects more clusters, and that is even in locations for which the input parameters have not been optimally specified.

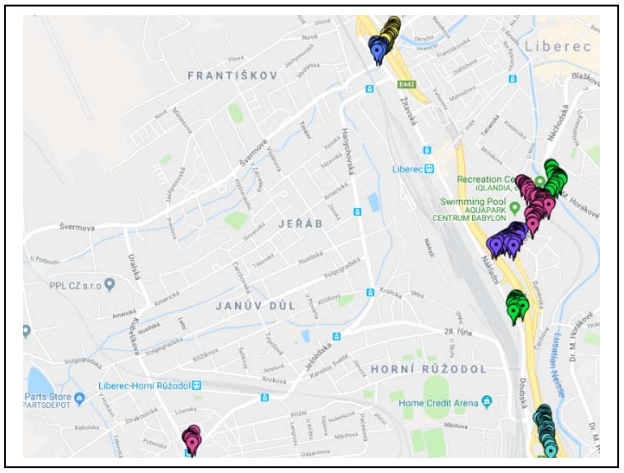

Fig. 4. Accident clusters discovered without the optimisation of the DBSCAN algorithm input parameters

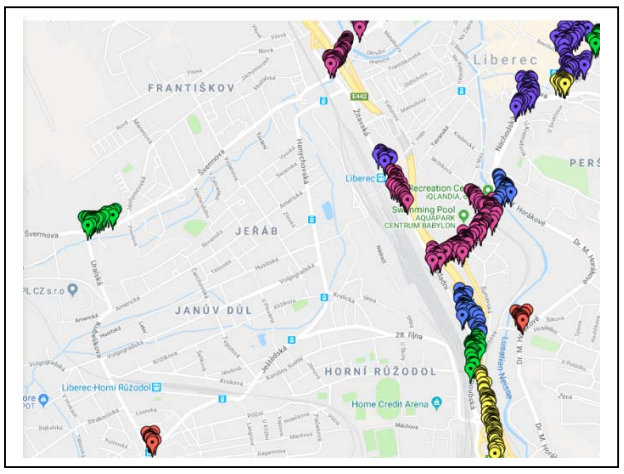

Fig. 5. Accident clusters discovered without the optimisation of the DBSCAN algorithm input parameters

VI. Conclusion

Our research proves that the state collected data (originally for statistical purposes) can serve the purpose of creating predictions regarding traffic safety. Thanks to the traffic accident data collected by the Czech police, it is possible to search for traffic accident clusters which can be used in e.g. the Early warning of heightened traffic accident occurrence risks system. Thanks to algorithms such as APRIORI, it possible to detect traffic accident clusters carrying specific attributes. Coupling the traffic density data with the traffic accident data can increase the accuracy in searching for relevant accident clusters, due to the fact that the algorithms appropriate for searching for clusters in geographical data tend to be sensitive to locations with varying density of traffic accident occurrence.

Acknowledgment

The current work is supported by the SGS project called „Prediction of extraordinary events using data mining techniques“ (number 21228), from Technical University of Liberec.

References

[1] S. Yin and O. Kaynak, “Big Data for Modern Industry: Challenges and Trends [Point of View]”, *Proceedings of the IEEE*, vol. 103, no. 2, pp. 143-146, 2015.

[2] Y. Sun, H. Song, A. J. Jara, and R. Bie, “Internet of Things and Big Data Analytics for Smart and Connected Communities”, *IEEE Access*, vol. 4, pp. 766-773, 2016.

[3] F. Provost and T. Fawcett, “Data Science and its Relationship to Big Data and Data-Driven Decision Making”, *Big Data*, vol. 1, no. 1, pp. 51-59, 2013.

[4] “Annual Accident Report 2017”, Statistics – accidents data - European Commission, 2017. [Online]. Available: https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/statistics/dacota/asr2017.pdf. [Accessed: 31-Jul.-2018].

[5] “Road Safety Annual Report”, The International Traffic Safety Data and Analysis Group, 2018. [Online]. Available: https://www.itf-oecd.org/sites/default/files/docs/irtad-road-safety-annual-report-2018_2.pdf. [Accessed: 31-Jul.-2018].

[6] “Global status report on road safety 2015”, World Health Organization, 2015. [Online]. Available: http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/. [Accessed: 31-Jul.-2018].

[7] “Prezentace výsledků sčítání dopravy 2010”, Ředitelství silnic a dálnic ČR, 2011. [Online]. Available: http://scitani2010.rsd.cz/pages/informations/default.aspx. [Accessed: 25-Sep.-2016].

[8] “Jednotná dopravní vektorová mapa”, Jednotná dopravní vektorová mapa, 2006. [Online]. Available: http://www.jdvm.cz/. [Accessed: 31-Jul.-2018].

[9] M. Lamr and J. Skrbek, “Searching for Traffic Accident Clusters to Increase Road Traffic Safety”, in IDIMT-2016 Information Technology, Society and Economy - Strategic Cross-Influence - 24th Interdisciplinary Information Management Talks, 2016, pp. 425-432.

[10] M. Lamr and J. Skrbek, “A Systems Approach to Designing a Traffic Collision Avoidance Early Warning System”, Journal of Systemics, Cybernetics and Informatics, vol. 15, no. Volume 15 - Number 3, pp. 55-59, 2017.

[11] M. Lamr and J. Skrbek, “Searching for Traffic Accident Clusters to Increase Road Traffic Safety”, in IDIMT-2016 Information Technology, Society and Economy - Strategic Cross-Influence - 24th Interdisciplinary Information Management Talks, 2016, pp. 425-432.

[12] M. Ester, H. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996, pp. 1-6.

[13] M. Ankerst, M. M. Breunig, H. -P. Kriegel, and J. Sander, “OPTICS: Ordering Points To Identify the Clustering Structure”, in Proceedings of the 1999 ACM SIGMOD international conference on Management of dat, 1999.

[14] S. Brecheisen, H. -P. Kriegel, and M. Pfeifle, “Efficient Density-Based Clustering of Complex Objects”, in Fourth IEEE International Conference on Data Mining (ICDM'04), 2004, pp. 43-50.

[15] M. Lamr and J. Skrbek, “Traffic Data and Possibilities of their Utilization for Safer Traffic”, in Proceedings of the International Conference: Liberec Informatics Forum 2016, 2016, pp. 61-73.

[16] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules”, in Proceeding VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases, 1994, pp. 487-499.

[17] “Road Safety Data”, Data.gov.uk, 2015. [Online]. Available: https://data.gov.uk/dataset/road-accidents-safety-data. [Accessed: 20-Sep.-2016].