Clustering The Driving Features Based On Data Streams

Rizwana Kalsoom and Zahid Halim
Faculty of Computer Science and Engineering
Ghulam Ishaq Khan Institute of Engineering Sciences and Technology
Topi, Pakistan
E-mail: gcs1205@giki.edu.pk, zahid.halim@giki.edu.pk

***Abstract* — This paper presents an innovative idea for the classification of individual drivers. The classification is based on each driver's driving features like, ratio of indicators to turns, number of brakes, number of time horn used, average gear, average speed, maximum speed and gear. K-means and hierarchical clustering is used to separate out the slow, normal and fast driving styles based on recorded data. Experimental result shows that k-means outperformed hierarchical clustering for recorded multi-attribute data.**

Keywords — Clustering, driver profiling, data stream, road safety, k-means clustering.

I. INTRODUCTION

Fast moving vehicles plying on the roads has substantially increased the hazards of accidents resulting in loss of precious human lives besides causing tremendous burden on the economy. Study carried out by World Health Organization (WHO) reveals road accidents to be among the top ten causes of deaths around the world [1]. Most of the traffic accidents are caused by the carelessness of the driver so it is necessary to predict the dynamic behavior of the driver.

Selection of a specific clustering is more complicated than classification because there is no clear accuracy for clustering for a specific data. So in order to reduce the risk of unsuitable clustering selection, we have chosen two different clustering algorithms these are effective for categorical values. In this paper we applied different clustering method to group the driving data and analyze driving state of the driver and their results are compared to find out the most appropriate and effective clustering for this data. Clustering will categories the driving behavior in following states:

Slow: When driver is driving car in a very slow manner applying brakes or moving on a road which is highly congested.

Normal: Driving with an average speed and acceleration with brakes.

Fast: Driving with high acceleration and with high number of horns.

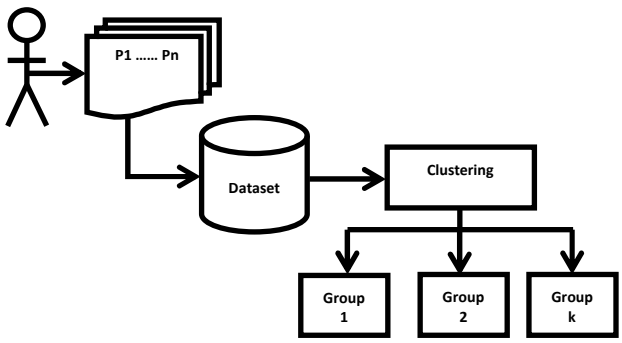


Fig.1. Driver profiling Based on Clustering

The whole system is shown in Fig. 1 which shows that first driving data is be collected from number of users and then clustering algorithm is be applied on that dataset to group the data according to the driver's behavior.

II. RELATED WORK

An early example of understanding driver behavior was cerebellum model articulation controller (CMAC), because every action is controlled by cerebellum, introduced by Albus. Their main focus was only on the brake and gas pedal pressure. Performance of CMAC can be analyzed by the potential of feature extraction for driver's behavior identification [2]. Also in 2009 Gaussion mixture model (GMM) was used to identify and the driving features that may be efficiently and effectively used to profile each driver. Later extracted features from accelerator and brake pedal pressure were used as input to fuzzy neural network to identify the driver [3]. Every driver has its own style for driving that's why speed profile analysis is also used for the behavioral identification. Speed and other driving steps like acceleration distribution over specific time interval were determined and compared with other speed profile. Results show that each driver has different speed intensity [4].

In 2009 Yi Lu Murphey has analyzed the driving state using jerk variations, like high acceleration/deceleration and also proposed an algorithm which classifies the driving style by using statistical information from the jerk profile and the road way type and traffic congestion level prediction [5]. Recently an android based CarSafe application is developed for driver safety that uses dual cameras and other embedded sensors on smart phone and fuses all information [6]. Also a mobile crash prediction system was developed which takes as input different attribute of driver profile and gives result like fit, unfit and partially fit [7]. Another research is consisted on the facial expression analysis for predicting unsafe driving behavior. Using bottom-up approach, the movement of 22 facial features was analyzed. Using the collected videos and driving simulator data, a dataset to build the computational models was constructed. These features proved more useful in predicting the accidents three to four seconds prior [8].

In another instance [9] using empirical data a framework methodology is presented for profiling driver behavior along multiple dimensions of behavior and risk to the driver and other road users. Results in [9] shows that 90% of the drivers have variability in driving profile regarding speed, acceleration, and braking performance between different roads than the same road environment. Also research shows that people have different speed graph in different scenarios and speed graph can be used to detect the driver behavior and other psycho-physiological states [4].

III. PROCEDURE

Our main objective is to partition the driving data of different drivers into different groups based upon their similarity. Clustering is an easiest and more efficient scientific technique to group the data by searching hidden patterns on the data set. In this study we have applied k-means and hierarchical clustering to group the driving data for driving profiling.

A. K-means Clustering Algorithm

K-means is a partition based clustering algorithm and it is the most simple and unsupervised method for grouping. K-means partitions the "n" objects into k clusters in which each object belongs to the cluster with the nearest mean [10].Following are the main steps of k-means algorithm [11]:

1. Initial cluster seeds are chosen at random. These represent the temporary means of the clusters.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Fig. 2 shows the steps of k-means algorithm in the form of flow chart.

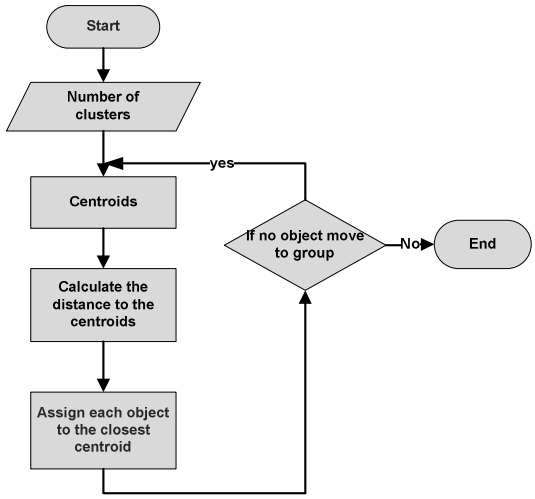


Fig.2. K-means Clustering Algorithm

The k-means clustering algorithm converges to local minimum. Before the k-means algorithm converges, calculations of distance and cluster centers are done while loops are executed a number of times. The number of iterations varies, depending on the initial starting cluster centers [11], [12].

B. Hierarchical Clustering

Hierarchical clustering is another category of clustering; it groups the data by creating a tree or dendrogram [13], [14]. There are two different approaches in hierarchical clustering; agglomerative and divisive. Most commonly used one is agglomerative. Following are the main steps of agglomerative hierarchical clustering:

1. Find the similarity or dissimilarity between every pair of objects in the data set.
2. Compute the proximity matrix.
3. Merge the closest clusters.
4. Update the proximity between new and original clusters.
5. Repeat step 3 and 4 and Stop when only one clusters remains.

C. Silhouette value and plot

Silhouette plot is a graphical aid for representing clustered data, to get the idea that how well separated the clusters are? Silhouette plot provide a measure over view that how much a cluster is close to its neighbor cluster [15].

Silhouette value lies between +1 and -1, maximum positive value shows that point is well separated and at a very distant from its neighbor clusters and 0 indicates that its lies approximate between two clusters and while negative value indicates that probably it is assigned to a wrong cluster.

D. *Dendrogram*

Dendrogram is a U-shaped tree structured graph which is used for the representation of hierarchical clustering. Height of u shape tells the distance between two objects those are clustered together. Leaf represents the original data points. Height is basically known as cophenetic distance between two objects [13]. It shows the correlation between distance matrix and linkage of the tree. Larger the value of correlation the more precisely data is clustered. Few other studies and representations of clustering can be found in [16-20].

IV. DESCRIPTION OF DATA SETS

Studies have shown that simulators allow researchers to the data three times faster than in traditional driving, that's why most of the research in this field is based on the simulator data [8, 21-23]. We have used 2011 driving simulator and total 5 minutes dynamic data for each driver using 10 second time window for observation. The data streams includes number of left and right turns, number of left and right indicators, number of brakes, horns and gear change with speed. A snapshot of the data is listed in Appendix-I.

Following attributes are selected as signature for each driver and used for the experiments.

1. Ratio of No. of Left indicator to the No. of left turns.
2. Ratio of No. of right indicator to the No. of left turns.
3. Number of brakes
4. Number of horns
5. Number of reverse gear
6. Average Gear
7. Maximum Gear
8. Average Speed
9. Maximum Speed

V. EXPERIMENTAL RESULTS

In this section, k-means and hierarchical clustering is performed on the driving data and results are compared with each other. The experiments by k-means and hierarchical clustering were carried out in MATLAB R2009a.

We have applied k-means and hierarchical clustering on the data set having 70 samples. First we have applied k-means clustering on the dataset with different number of clusters.

TABLE I. K-MEANS CLUSTERING COMPARISON OF TOTAL AND AVERAGE SILHOUETTE VALUE OF DIFFERENT NO OF CLUSTERS

S.no	No. of clusters	Total Distance	Average Silhouette value
Exp. 1	2	1435.82	0.4379
Exp. 2	3	1183.02	0.3769
Exp. 3	4	1001.01	0.3709
Exp. 4	5	915.09	0.2964
Exp. 5	6	846.23	0.2675
Exp. 6	7	783.23	0.3052
Exp. 7	8	765.47	0.2723

Table I shows that by increasing number of clusters sum of distances is decreasing and average distance between each point with in the clusters is also decreasing.

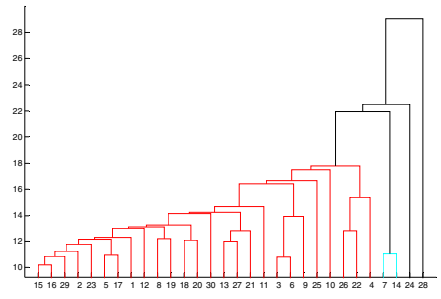
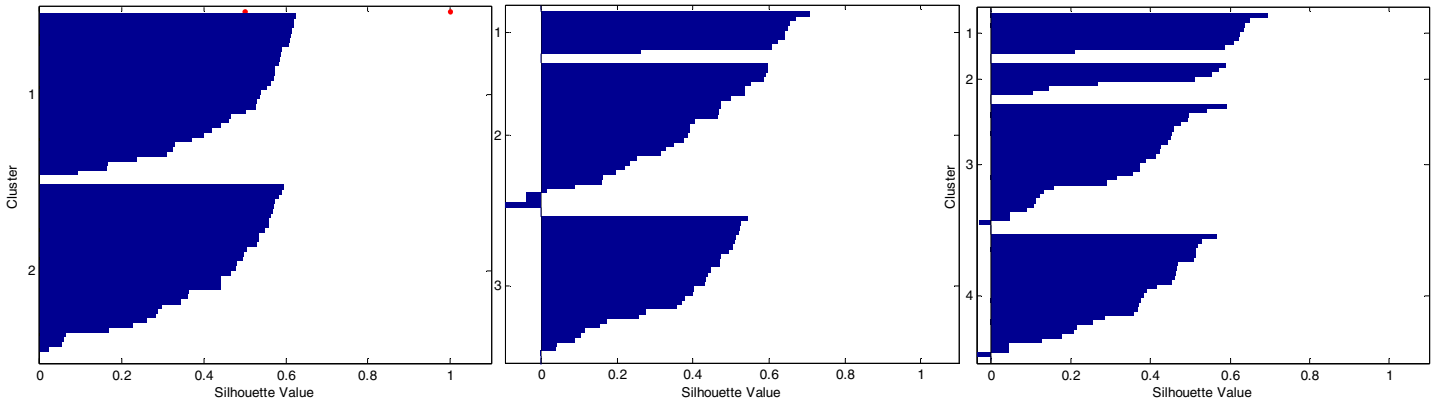


Fig.3. Dendrogram of data

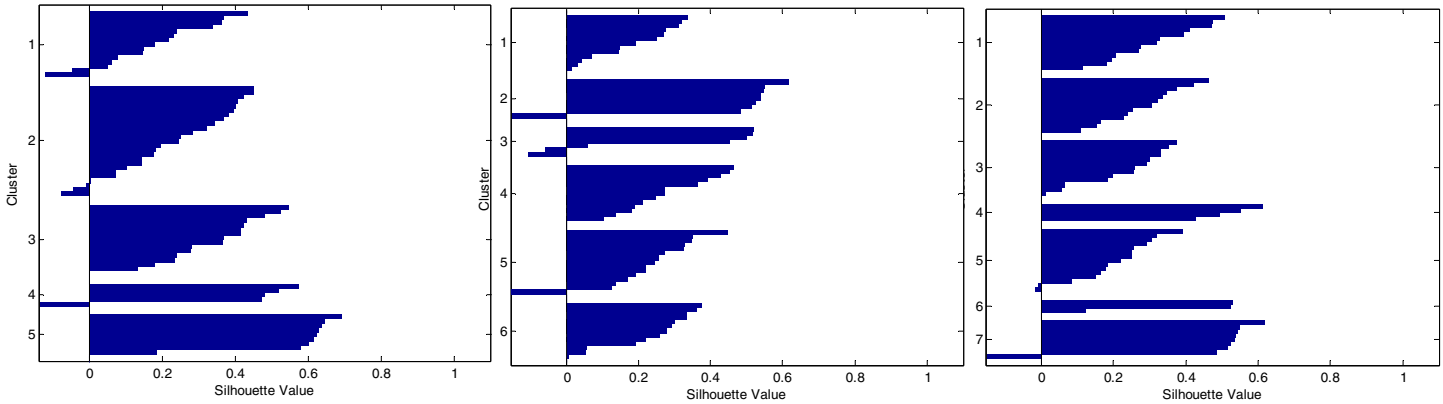
As shown in Fig. 3 dendrogram vertical axis is representing the normalized distance between two clusters that are merged together so greater the height greater will be the difference. While horizontal axis represents the label of patterns that are arranged together in a cluster at each stage. But in this case Fig. 3 represents that most of the data is grouped in a single cluster at the last stages of hierarchical clustering.

Fig. 4 displays the silhouette plot of k-means having different number of clusters. Fig. 4(a) shows the plot having 2 clusters, it is clear from plot that there is no negative silhouette value and whole data is grouped properly between two clusters. In Fig. 4(b) and 4(c) it is clear that when this data is grouped in 3 or 4 clusters then there are only 2 points which are not correctly grouped.

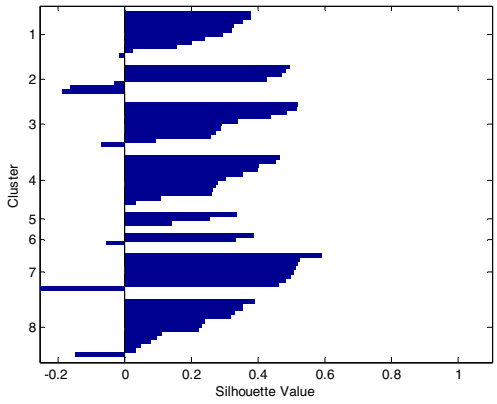
While in case of figure 4(d), 4(e) and 4(g) there are more negative silhouette values which depicts that data cannot be well separated in 5, 6 and 8 clusters correctly. While in case of 7 clusters data is grouped better then in 5, 6 and 8 clusters.



a) K-means having 2 clusters b) K-means having 3 clusters c) K-means having 4 clusters



d) K-means having 5 clusters e) K-means having 6 clusters f) K-means having 7 clusters



g) K-means having 6 clusters

Fig 4: Silhouette plot of K-means clustering having different number of clusters

CONCLUSION

In this paper we have clustered the driving behavior. Experimental result shows that k-means clustering group the data correctly according to similar attributes while hierarchical clustering performance is not good as compare to k-means clustered data. In future we will plan to analyze the driver driving steps using this clustered data in detail to predict the accident. The recorded data can be used in other domain like automatic generation of car racing games tracks [24, 25], improving vehicle design [26] and road safety analysis [27]. As an extension to the current clustering the extracted clusters can be represented using sonification [28] approaches and can be an add-on feature of modern vehicles to inform/alarm the driver for different states of driving.

REFERENCES

[1] Hu G, Baker T, Baker SP, "Comparing Road Traffic Mortality Rates from Police-Reported Data and Death Registration Data In China", Bulletin of the World Health Organization, vol. 89(1). pp. 41-45, 2011.

[2] Abdul Wahab, Toh Guang Wen and Norhaslinda Kamaruddin, "Understanding Driver Behavior Using Multi-Dimensional CMAC", 6th International Conference on Information, Communications & Signal Processing, pp. 1-5, 2007.

[3] Abdul Wahab A, Quek C, Tan CK, Takeda K, "Driving Profile Modeling and Recognition Based on Soft Computing Approach. IEEE Transactions on Network, vol. 20/4. pp. 563-582, April 2009.

[4] Rygula A, "Driving Style Identification Method Based on Speed Graph Analysis", International Conference on Biometrics and Kansei Engineering, pp. 76-79, 2009.

[5] Yi Lu Murphey, Robert Milton, Leonidas Kiliaris, "Driver's Style Classification Using Jerk Analysis", IEEE Workshop on Computational Intelligence in Vehicles and Vehicular Systems. CIVVS '09, pp. 23 – 28, 2009.

[6] Chuang-Wen You, Martha Montes-de-Oca, Thomas J. Bao1, Nicholas D. Lane, Giuseppe Cardone4, Lorenzo Torresani1, and Andrew T. Campbell, " CarSafe: A Driver Safety App that Detects Dangerous Driving Behavior using Dual-Cameras on Smartphones" , ACM, 2009.

[7] Garima R. Singh and Snehlata S. Dongre, "Crash Prediction System for Mobile Device on Android by Using Data Stream Mining Techniques", Sixth Asia Modeling Symposium, 2012.

[8] Maria E. Jabon, Jeremy N. Bailenson, Emmanuel Pontikakis, Leila TakayamaFacial. Expression Analysis for Predicting Unsafe Driving Behavior car driving simulator, IEEE Pervasive Computing, Vol. 10, pp.84-95, 2011.

[9] Ellison, A.B., Greaves, S.P. & Daniels, R, "Profiling Drivers' Risky Behaviour Towards All Road Users", Australasian College of Road Safety Conference, Sydney, 9-10 August 2012.

[10] Shalove Agarwal, Shashank Yadav and Kanchan Singh, "K-Means Versus K-means ++ Clustering Technique", Students Conference on Engineering and Systems (SCES)", pp.1-6, 2012.

[11] Shi Na, Liu Xumin, Guan yong, "Research on k-means Clustering Algorithm", Third International Symposium on Intelligent Information Technology and Security Informatics, pp. 63 – 67, 2010.

[12] Sariel Har-Peled and Soham Mazumdar, "On coresets for k-means and k-median clustering", Proceedings of the thirty-sixth annual ACM symposium on Theory of computing, ,. pp. 291–300, New York, NY, USA, 2004.

[13] Morteza Jalalat-evakilkandi, Abdolreza Mirzaei, "A New Hierarchical-Clustering Combination scheme Based on Scatter Matrices and Nearest Neighbor Criterion", 5th International Symposium on Telecommunications (IST), pp.904-908, Dec. 2010.

[14] David M. Blei, "Hierarchical clustering", Princeton University, 2008.

[15] Peter J. ROUSSEEUW, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis", Journal of Computational and Applied Mathematics, pp. 53-65, 1987.

[16] Ali, M. H., Sundus, A., Qaiser, W., Ahmed, Z., & Halim, Z, "Applicative implementation of D-stream clustering algorithm for the real-time data of telecom sector", IEEE International Conference on Computer Networks and Information Technology (ICCNIT), pp. 293-297, 2011.

[17] Yu, D., & Zhang, "AClusterTree: integration of cluster representation and nearest-neighbor search for large data sets with high dimensions", IEEE Transactions on Knowledge and Data Engineering, , vol. 15, 1316-1337, . 2003.

[18] Aniq, M., Halim, Z., & Baig, R, |MST and SFMST based Clustering", First National Conference on Security, Computing, & Communication , p. 68, May 2008.

[19] Antoniadis, A., Brossat, X., Cugliari, J., & Poggi, J. M, "Clustering functional data using wavelets", International Journal of Wavelets, Multiresolution and Information Processing, vol. 11/1, 13 February 2013.

[20] De Smet, W., & Moens, M. F, "Representations for multi-document event clustering. Data Mining and Knowledge Discovery, vol. 26, issue 3, pp.533-558, 20 june, 2012.

[21] Huazhong Ning, Wei Xuy, Yue Zhou, Yihong Gongy and Thomas Huang, "Temporal Difference Learning to Detect Unsafe System States ", Proceedings of the International Conference on Pattern Recognition, pp. 1431-1434, 2008.

[22] Huazhong Ning, Xu W, Zhou Y, Gong Y, Huang TS, "A General Framework to Detect Unsafe System States From Multisensor Data Stream", IEEE Transactions on Intelligent Transportation Systems pp. 4-15, 2009.

[23] HuazhongNing, Xu W, Zhou Y, Gong Y, Huang TS, "Temporal difference learning to detect unsafe system states", Proceedings of the International Conference on Pattern Recognition, pp. 1-4, 2008.

[24] Halim, Z., Baig, A. R., & Mujtaba, H," Measuring entertainment and automatic generation of entertaining games. International Journal of Information Technology, Communications and Convergence, vol. 1(1), 92-107, 2012.

[25] Halim, Z., Baig, A. R., & Hasan, M, "Evolutionary Search For Entertainment In Computer Games", Intelligent Automation & Soft Computing, vol. 18(1), pp. 33-47, 2012.

[26] Choi, Seibum B., and J. Karl Hedrick, "An observer-based controller design method for improving air/fuel characteristics of spark ignition engines." IEEE Transactions on Control Systems Technology, pp. 325-334, 1998.

[27] Mohan, Dinesh. "Road safety in less-motorized environments: future concerns", International Journal of Epidemiology, vol. 31.3. pp. 527-532, 2002.

[28] Halim, Z., Baig, R., & Bashir, S, " Sonification, a Novel Approach Towards Data Mining", IEEE International Conference on Emerging Technologies, 2006. ICET'06., pp. 548-553, 2006.

Appendix-I

s.no	left	right	clutch	Gear 1	Gear 2	Gear 3	Gear 4	Gear 5	reverse Gear	break	Horn	left indic.	right indic.	Speed Km/h
1	0	0	1	1	1	0	0	0	0	0	0	0	0	35
2	0	0	1	0	0	1	0	0	0	0	0	0	0	45
3	0	0	1	0	1	0	0	0	0	1	0	0	0	20
4	0	0	1	0	0	1	0	0	0	0	0	0	0	50
5	0	0	1	0	0	0	1	0	0	0	0	0	0	65
6	0	0	0	0	0	0	1	0	0	0	0	0	0	58
7	0	0	1	0	1	1	0	0	0	1	0	0	0	20
8	0	0	1	0	0	1	0	0	0	0	0	0	0	40
9	0	0	0	0	0	1	0	0	0	0	0	0	0	48
10	0	0	0	0	0	1	0	0	0	0	0	0	0	50
11	0	0	1	0	1	0	0	0	0	1	0	0	0	25
12	0	0	1	0	0	1	0	0	0	0	0	0	0	45
13	1	0	1	0	1	0	0	0	0	0	0	1	0	38
14	1	0	0	0	1	0	0	0	0	0	0	1	0	5
15	0	0	0	0	1	0	0	0	0	0	0	0	0	20
16	0	0	0	0	1	0	0	0	0	0	0	0	0	25
17	0	0	0	0	1	0	0	0	0	0	0	0	0	10
18	0	0	0	0	1	0	0	0	0	0	1	0	0	32
19	0	0	0	0	1	0	0	0	0	0	1	0	0	35
20	0	0	0	0	1	0	0	0	0	1	0	0	0	20
21	0	0	0	0	1	0	0	0	0	0	0	0	0	35
22	0	0	0	0	1	0	0	0	0	1	0	0	0	10
23	0	0	1	0	0	1	0	0	0	0	1	0	0	40
24	0	0	1	0	1	0	0	0	0	0	0	0	0	20
25	0	1	0	0	1	0	0	0	0	0	0	0	1	10
26	0	0	1	0	0	1	0	0	0	0	0	0	0	40
27	0	0	1	0	1	0	0	0	0	1	0	0	0	20
28	0	0	1	0	0	1	0	0	0	0	1	0	0	30
29	0	0	0	0	0	1	0	0	0	0	1	0	0	40
30	0	0	1	0	1	0	0	0	0	1	0	0	0	5