International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

http://ijpsat.ijsht-journals.org
Vol. 9 No. 1 June 2018, pp. 116-125

Credit Card Transaction Validity Detection Using Density Based Clustering

Elliana Gautama

Faculty of Information Technology
Perbanas Institute
Jakarta, Indonesia

The use of credit cards is increasing every year, causing abuse of credit card transactions is also increasing. The perpetrators of crime with credit card fraud are also increasingly sophisticated and have a wide network. This causes a loss for all parties. To detect credit card misuse in electronic transactions, the author tries to make research using Density Based Clustering approach (DBSCAN) to detect whether credit card number is valid or not. DBSCAN generally starts by setting the starting point randomly, then finding all the dots around in Eps the starting point distance, if the number of dots around is greater than or equal to MinPts then the cluster is formed, using density distributions from the dots in the database, DBSCAN can categorize the points into separate groups indicating different classes. It is hoped that by using DBSCAN Method in the decision making process can know the existence of misuse in the use of transaction with credit card. DBSCAN method in this research is used to help determine credit card problem in electronic transaction by using parameter of minimum point (minpts) and epsilon (eps). So it can be concluded whether the credit card is valid or not.

Keywords* - *Credit Card; Fraud; Density Based Clustering.

I. Introduction

Along with the growing use of credit card, credit card abuse was also increased. The perpetrators of the crime with ATM card or credit card mode is more modern and has a wide network, even news from media information that this network has come to foreign countries. As well as techniques and equipment and raw materials of counterfeit card makers in this network have been exchanging information and mutually buying and selling raw materials for counterfeiting.

The perpetrators of Credit Card offenses have card-making machines. The encoding machine dumps the data on the card's magnetic stripe according to the data recorded on the original card. This card is often used to make ID card identification, membership card, and others. The raw materials can be purchased from abroad or from the bank in the country which then printed in original or resemble. There are several alternative allegations in the theft of victim data: The data and numbers originally obtained by means of Skimming means electronically recording data on magnetic stripe skimming.

Figure 1. Parts of Credit Card

Corresponding Author: Elliana Gautama

Credit cards as a means for transactions certainly equipped with various identification tools. The front of the credit card consists of the bank name as the issuer, the card number (Fig.1), the card logo, the cardholder's name and the credit card's validity period. The back of the credit card is equipped with a magnetic tape called magstripe. The magnetic tape consists of magnetic bars in milli size arranged parallel along the card. Each merchant has an electronic verification tool that can identify whether the credit card is still valid or not. Something that is the benchmark verifier of credit cards is the combination of numbers behind it.

To detect credit card misuse in electronic transactions, the author tries to make a research using Clustering Based Density Approach, to detect whether the credit card used is valid or not.

The most efficient clustering algorithm for determining clusters in data with different densities is the density-based clustering algorithm [1]. DBSCAN is one example of a pioneering development of grouping techniques based on density or commonly known as density-based clustering.

This study discusses the implementation of DBSCAN Methods in the decision making process to determine the existence of abuse in the use of transactions with credit cards. DBSCAN method in this research is used to help determine credit card problem in an electronic transaction by using parameter of minimum point (minpts) and epsilon (eps). The process of determining the value of parameters is trial and error, meaning that the determination of parameter values should be tested several times to get a certain number of clusters.

II. Density Based Spatial clustering Of Application With Noise (DBSCAN)

Density-Based Spatial Clustering of Application with Noise (DBSCAN) is a clustering method that builds areas based on density connected (density connected). DBSCAN is a type of partition clustering where high density areas are considered clusters whereas low density or incompatible clusters are considered noise [9].

Advantages of this algorithm include [8] :

a. Does not require to know the number of clusters.
b. Can find cluster.
c. Able to cope with noise.

DBSCAN generally starts with a random starting point. Then find all the dots around in Eps the starting point distance. If the number of dots around is greater than or equal to MinPts then the cluster is formed.

DBSCAN is an algorithm designed by et.al ester in 1996 to identify groups in large spatial data sets by looking at the local density of database elements, using only one input parameter. DBSCAN can also determine whether information is classified as noise or outlier. In addition, DBSCAN's work process is fast and excellent for a variety of database sizes - almost linear [9].

By using density distributions from points in the database, DBSCAN can categorize them into separate groups indicating different classes.

The concept of the DBSCAN algorithm is as follows [9]:

a. Epsilon (eps). Epsilon is the distance that determines the closeness of an object with other objects around it. Objects that meet the epsilon value are said to have closeness to an observed object.
b. Minimum number of objects (minpts). Minpts determines cluster formation. A cluster will be formed if the number of adjacent objects has exceeded the minimum value (minpts) given.
c. The algorithm then names objects that are located close to the given epsilon value called core objects whereas those objects that are closely related to core objects are called border objects. Beyond that, objects that are not core or border objects are an outlier or an anomaly.

Density-Based Spatial Clustering of Application with Noise (DBSCAN) is a clustering method that builds areas based on density connected (density connected). DBSCAN is a type of partition clustering where high density areas are considered clusters whereas low density or incompatible clusters are considered noise [9].

In this Algorithm are known some terms as follows:

a. Core: The center point in the cluster is based on the density where there are a number of points that must be within Eps (radius or threshold value), MinPts
(At least point in the cluster) specified by the user.

b. Border: The point that becomes the boundary in the area of the central point (core).

c. Noise: The point that can not be reached by the core and not a border.

$$\text{noise} = \{x \in X \mid \forall i : x \notin C_i\} \tag{1}$$

d. Direct affordable density: A point is said to be an affordable point directly if the point is connected directly to the center point (core).

$$x \in N_{Eps}(y) \wedge |N_{Eps}(y)| \geq MinPts \quad (2)$$

Where N_{Eps} is the point around of the radius of Eps, MinPts is the minimum point in the cluster.

e. Affordable density: A point is said to be an affordable point when the point is connected indirectly to the core.

f. Connected density: A dot is said to be connected to each other by another point.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise), the algorithm grows areas of high density into clusters and finds clusters in random form in a spatial database containing noise. DBSCAN defines the cluster as the maximum set of density-connected density points. All objects that do not fit into any cluster are considered noise.

The basic principles of the density based clustering method are as follows:

A neighborhood located within a radius ϴ is called the ∈ -neighorhood of the data object.

If the ∈ -neighorhoodof an object contains at least a minimum number, the MinPts of an object, the object is called a core object.

An object p is the reachable density of the q object with respect to and MinPts in a set of objects D if there is an object chain p1, p2, ..., pn, where p1 = q and pn = p, where pi + 1 density reachable directly Of pi with respect to and MinPts, for 1 in, pi member D.

An object p is density connected to object q with respect to and MinPts in a set of object D if there is an object o member D where the two p and q are the reachable density of o with respect to and MinPts.

DBSCAN finds clusters by:

DBSCAN traces the clusters by examining ∈ *-neighborhood*of each point in the database. If the ∈ *-neighborhood*of point p contains more than MinPts, a new cluster with p as the core object is created.

Then DBSCAN iteratively collects directly the reachable density objects of the core object, which may involve the merging of some reachable cluster densities.

III. Analysis Of Test Result Using Density Based Clustering

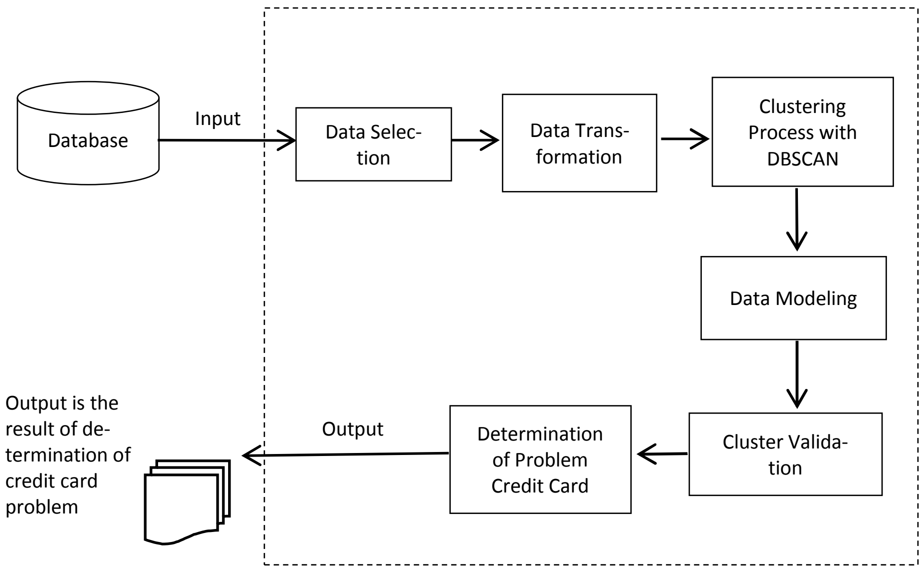


Figure 2. Stages of Research

3.1 DBSCAN Algorithm

The algorithm of DBSCAN is:

a. Select the starting point r randomly

d. If the point that meets Eps over MinPts then point r is the core and the cluster is formed by following the following conditions:
∀ x, y : if x ∈ C(cluster) and y affordable density from x then y ∈ C. (Maximality)
(3)
∀ x, y ∈ C: x connected density from y. (Connectivity)(4)

e. Repeat steps 3 - 4 until all the points are processed

f. If r is the border point and no density point is reached to r, then the process is continued to another point. [6]

3.2 Research Result

Accuracy in detecting anomalies depends on two parameters ie epsilon and minimum point. Therefore in the detection test anomalies were first performed on epsilon trials and minimum point. Trial epsilon and minimum point is used to obtain the epsilon value and the minimum point corresponding to the process of detecting the validity of credit cards. Trial is performed on 1000 credit card data from German Credit Card Fraud - UCI machine learning

b. Initialize input parameters: minpts and eps

c. Calculate Eps or any affordable density distance to r using Euclidean distance.

repository. There are 319 anomalous data in the group.

Dataset information used:
Relation: German Credit
Instances: 1000
Number of Attributes: 21 (checking_status, duration, credit_history, purpose, credit_amount, savings_status, employment, installment_commitment, personal_status, other_parties, residence_since, property_magnitude, age, other_payment_plans, housing, existing_credits, job, num_dependents, own_telephone, foreign_worker, class)

Testing Scenario:
The scenarios used are:
Selection of 2 DBSCAN parameters is Min Pts and Eps used heuristik. Compare some values of both parameters with the number of clusters generated.
By default use MinPts = 0.9 and Eps = 6
Number of iterations: 5
Within cluster sum of squared errors: 5665.9976202840735

Table 1. Initial starting points (random)

Cluster	**0**	**1**
checking_status	no checking	0
duration	36	24
credit_history	critical/other existing credit	critical/other existing credit
purpose	new car	used car
credit_amount	7855	6615
savings_status	<100	<100
employment	1<=X<4	unemployed
installment_commitment	4	2
personal_status	female div/dep/mar	male single
other_parties	none	none
residence_since	2	4
property_magnitude	real estate	no known property
age	25	75
other_payment_plans	stores	none
housing	own	For free
existing_credits	2	2
job	skilled	highqualif/selfemp/mgmt

num_dependents	1	1
own_telephone	yes	yes
foreign_worker	yes	yes
class	bad	good

Table 2. Final cluster centroids

Attribute	Full Data	Cluster 0	Cluster 1
	(1000.0)	(643.0)	(357.0)
checking_status	no checking	no checking	<0
duration	20.903	19.9285	22.6583
credit_history	existing paid	existing paid	existing paid
purpose	radio/tv	radio/tv	new car
credit_amount	3271.258	2924.7869	3895.2941
savings_status	<100	<100	<100
employment	1<=X<4	1<=X<4	>=7
installment_commitment	2.973	2.9611	2.9944
personal_status	male single	male single	male single
other_parties	none	none	none
residence_since	2.845	2.5599	3.3585
property_magnitude	car	car	no known property
age	35.546	33.2364	39.7059
other_payment_plans	none	none	none
housing	own	own	own
existing_credits	1.407	1.3701	1.4734
job	skilled	skilled	skilled
num_dependents	1.155	1.1011	1.2521
own_telephone	none	none	yes
foreign_worker	yes	yes	yes
class	good	good	good

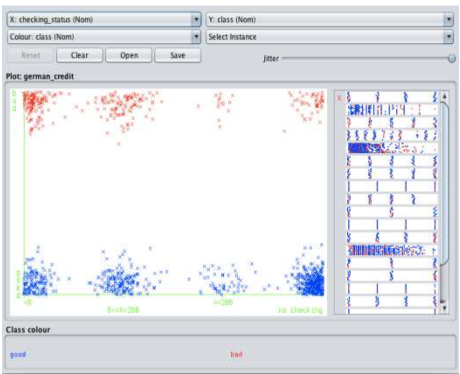
Table 3. Fitted estimators (with ML estimates of variance)

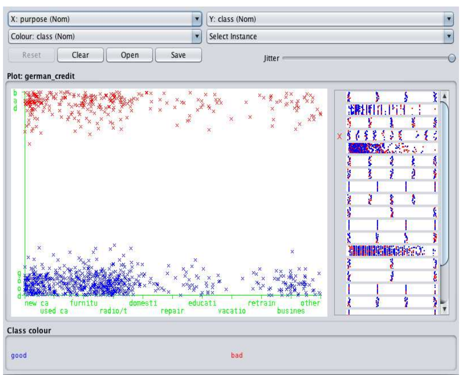
	Cluster: 0	**Cluster: 1**
Attribute	Prior probability: 0.6427	Prior probability: 0.3573
checking_status	Discrete Estimator. Counts = 109 179 44 315 (Total = 647)	Discrete Estimator. Counts = 167 92 21 81 (Total = 361)
duration	Normal Distribution. Mean = 19.9285 StdDev = 11.6676	Normal Distribution. Mean = 22.6583 StdDev = 12.5274
credit_history	Discrete Estimator. Counts = 26 27 364 58 173 (Total = 648)	Discrete Estimator. Counts = 16 24 168 32 122 (Total = 362)

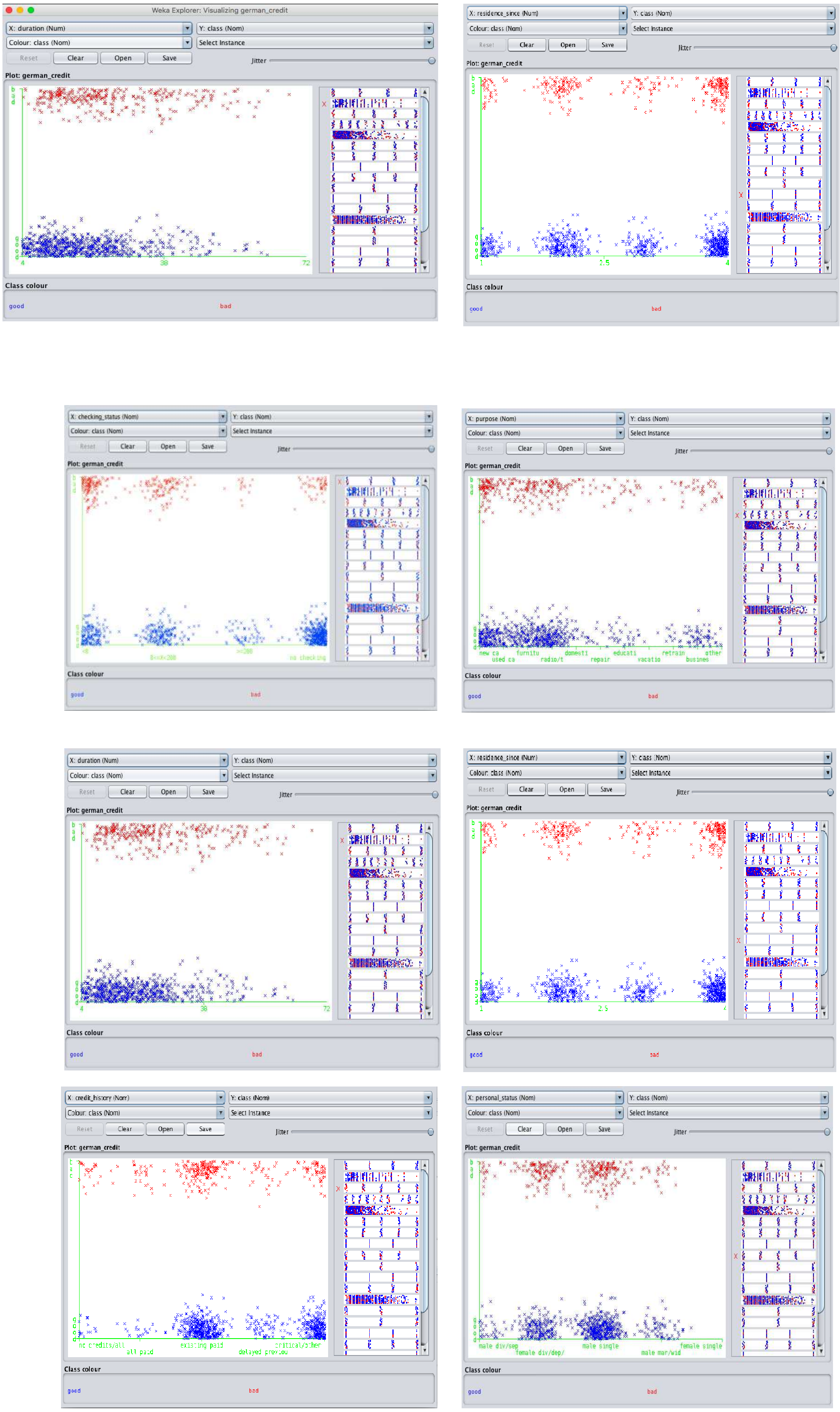
purpose	Discrete Estimator. Counts = 99 53 121 250 11 15 26 1 7 65 6 (Total = 654)	Discrete Estimator. Counts = 137 52 62 32 3 9 26 1 4 34 8 (Total = 368)
credit_amount	Normal Distribution. Mean = 2924.7869 StdDev = 2498.4577	Normal Distribution. Mean = 3895.2941 StdDev = 3232.309
savings_status	Discrete Estimator. Counts = 382 72 49 32 113 (Total = 648)	Discrete Estimator. Counts = 223 33 16 18 72 (Total = 362)
employment	Discrete Estimator. Counts = 26 119 294 117 92 (Total = 648)	Discrete Estimator. Counts = 38 55 47 59 163 (Total = 362)
installment_commitment	Normal Distribution. Mean = 2.9611 StdDev = 1.107	Normal Distribution. Mean = 2.9944 StdDev = 1.1376
personal_status	Discrete Estimator. Counts = 32 216 325 74 1 (Total = 648)	Discrete Estimator. Counts = 20 96 225 20 1 (Total = 362)
other_parties	Discrete Estimator. Counts = 580 25 41 (Total = 646)	Discrete Estimator. Counts = 329 18 13 (Total = 360)
residence_since	Normal Distribution. Mean = 2.5599 StdDev = 1.0644	Normal Distribution. Mean = 3.3585 StdDev = 0.9789
property_magnitude	Discrete Estimator. Counts = 199 142 275 31 (Total = 647)	Discrete Estimator. Counts = 85 92 59 125 (Total = 361)
age	Normal Distribution. Mean = 33.2364 StdDev = 10.3077	Normal Distribution. Mean = 39.7059 StdDev = 11.9928
other_payment_plans	Discrete Estimator. Counts = 82 35 529 (Total = 646)	Discrete Estimator. Counts = 59 14 287 (Total = 360)

housing	Discrete Estimator. Counts = 114 514 18 (Total = 646)	Discrete Estimator. Counts = 67 201 92 (Total = 360)
existing_credits	Normal Distribution. Mean = 1.3701 StdDev = 0.5319	Normal Distribution. Mean = 1.4734 StdDev = 0.646
job	Discrete Estimator. Counts = 14 143 427 63 (Total = 647)	Discrete Estimator. Counts = 10 59 205 87 (Total = 361)
num_dependents	Normal Distribution. Mean = 1.1011 StdDev = 0.3014	Normal Distribution. Mean = 1.2521 StdDev = 0.4342
own_telephone	Discrete Estimator. Counts = 481 164 (Total = 645)	Discrete Estimator. Counts = 117 242 (Total = 359)
foreign_worker	Discrete Estimator. Counts = 619 26 (Total = 645)	Discrete Estimator. Counts = 346 13 (Total = 359)
class	Discrete Estimator. Counts = 478 167 (Total = 645)	Discrete Estimator. Counts = 224 135 (Total = 359)

Time taken to build model (full training data) : 0.18 seconds
Clustered Instances
0 681 (68%)
1 319 (32%)
The results are as follows:
Cluster 0 there are 681 instances is good transaction (Valid)
Cluster 1 there are 319 instances is bad transaction (Fraud)







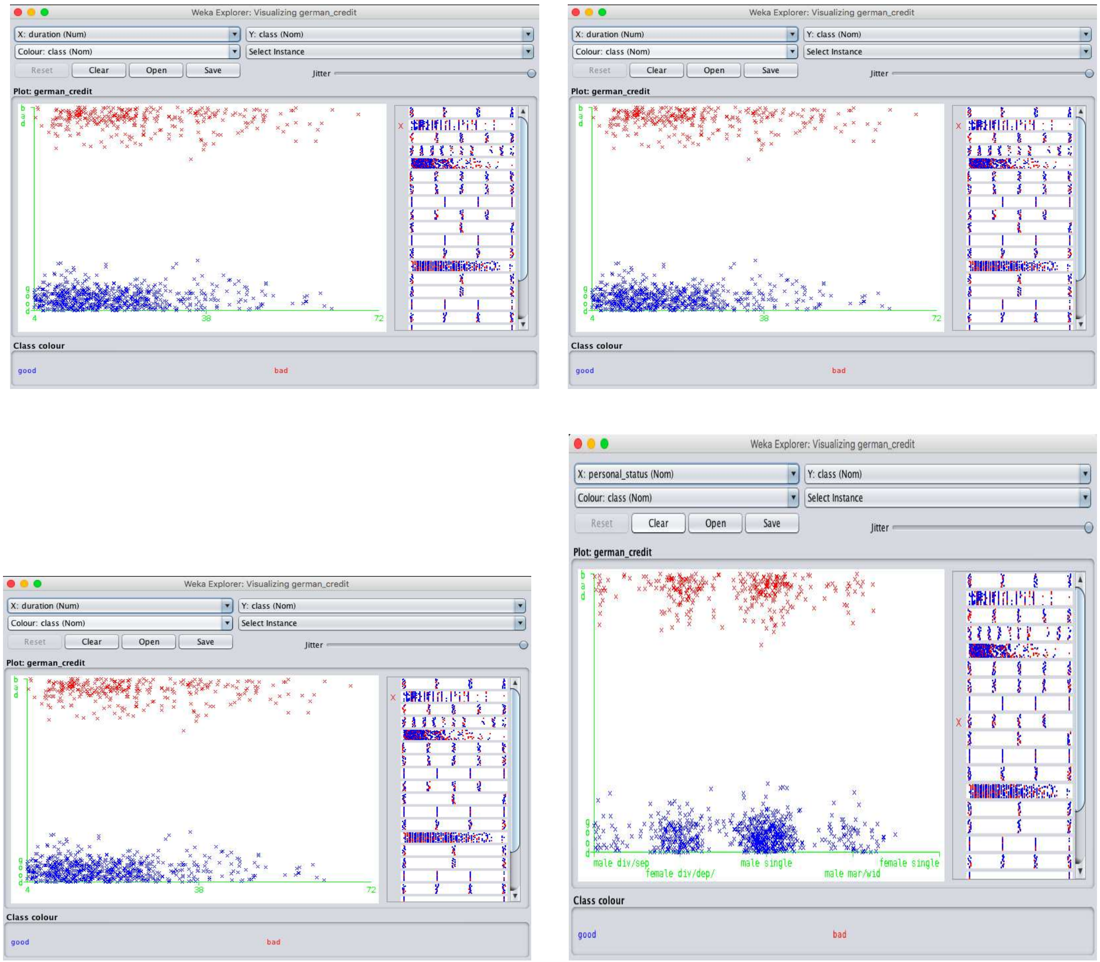


Figure 3. Density Based Clustering Visualization Based on Attribute

IV. Conclusion

Based on this research, it can be concluded that the result of analysis of some attributes on the dataset used to detect the validity of credit card transactions using density-based clustering (DBSCAN) obtained 2 clusters, cluster 0 for valid transaction data (good) and cluster 1 for data / transaction fraud is invalid (bad).

V. Recommendation

In this study used 21 attributes, when building applications it is advisable to reduce some attributes so that the results obtained can be better and more accurate.

References

[1] Matheus CJ, Chan PK, Piatetsky-Shapiro G. "Systems for Knowledge Discovery in Databases". *IEEE Transactions on Knowledge and Data Engineering*. 1993; 5(6): 903-913.

[2] Tan, P. "Introduction to Data Mining". Boston: Pearson Education, 2006.

[3] B. Santosa. (2007). Data Mining. Teknik Pemanfaatan Data untuk Keperluan Bisnis, First Edition ed. Yogyakarta: GrahaIlmu.

[4] Ester, M., Kriegel, H.P., Sander, J. and Xu, X. (1996) A Density Based Algorithm for Discover Clusters in Large Spatial Datasets with Noise. *Proceedings of International Conference on Knowledge Discovery and Data Mining*, 226-231.

[5] Guha, S., Rastogi, R. and Shim. K, (2001) Cure: An Efficient Algorithm for Large Databses, Information Systems, 26, 35-58.

[6] Anindya Santika Devi, N. M., Gede Darma Putra, I. K., &Sukarsa, I. (2015). Implementasi Metode Clustering DBSCAN pada Proses Pengambilan Keputusan. *Lontar Komputer*, 6(3).

[7] Maimon, O., &Rokach, L. (Eds.). (2005). Data mining and knowledge discovery handbook (Vol. 2). New York: Springer.

[8] Mumtaz, K. and Duraiswamy, K., (2010). An analysis on density based clustering of multi dimensional spatial data.*Indian Journal of Computer Science and Engineering*, 1(1), pp.8-12.

[9] Nagpal, P. B., & Mann, P. A. (2011). Comparative study of density based clustering algorithms. *International Journal of Computer Applications*, 27(11), 44-47.

[10] Tan, M. Steinbach, and V. Kumar, (2005) Introduction to Data Mining, (First Edition). Boston, MA, USA: Addison-Wesley Longman Publishing CO., Inc.

[11] Yan, B., and Deng, G. (2015). Improved Clustering Algorithm Based on Density- Isoline. *Open Journal of Statistics*, 5(04), 303-310.

[12] Kurniawan, M.F., Nugraha B.C., Putri I., & Hasyim A. "Implementasi Agoritma Density Based Spatial Clustering of Applications With Noise (DBSCAN) Pada-Konsumsi Daya Listrik Rumah Tangga".

[13] Zakariya, Achmad Zainuddin, Arif Djunaidy, Renny-Pradina Kusumawardani. (2012). "PembuatanAplikasi Pendeteksi Anomali PadaPola Konsumsi Listrik Pelanggan Kota Surabaya Menggunakan ". *JurnalTeknikPomits* Vol.1. pp 1-5.

[14] Devi, Ni Made Anindya Santika, I KetutGede Darma Putra, I Made Sukarsa. (2015). "Implementasi Metode Clustering DBSCAN pada Proses Pengambilan Keputusan". *Lontar Komputer* Vol. 6, No.3.

[15] http://www.akki.or.id/index.php/cr "Credit Card Growth" Asosiasi Kartu Kredit Indonesia, Tanggalakses 1 November 2016.