RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Cluster Analysis for Anomaly Detection

Sutapat Thiprungsri

Rutgers Business School

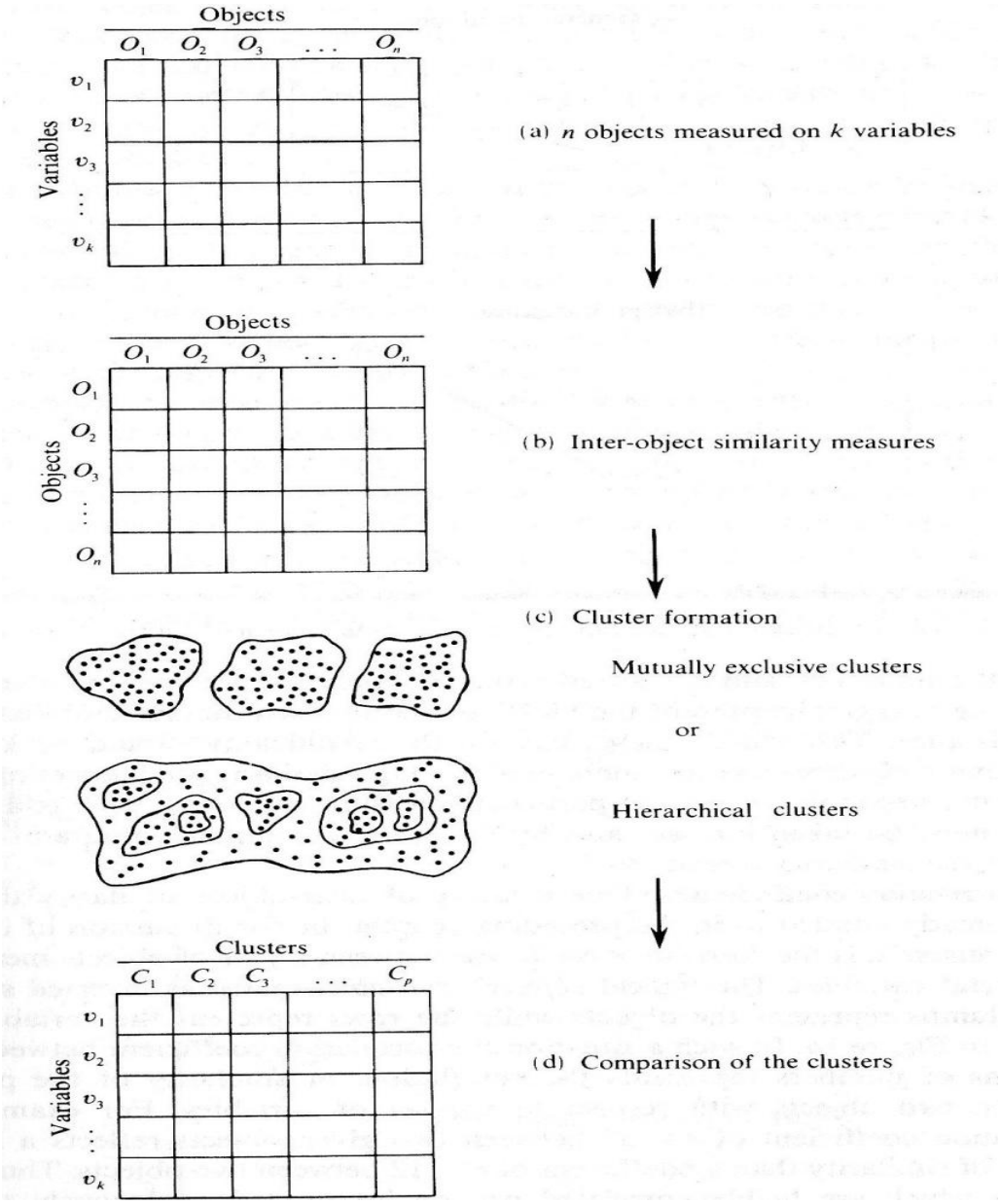
July 31th 2010

Contribution

- To demonstrate that cluster analysis can be used to build a model for anomaly detection in auditing.
- To provide a guideline/example for using cluster analysis in continuous auditing.

Cluster Analysis

- Clustering is an unsupervised learning algorithm.
- Clustering is a useful technique for grouping data points such that points within a single group or cluster are similar, while points in different groups are different.



An Outline of Cluster Analysis Procedure. (Kachigan, 1991)

Cluster Analysis: Application

Marketing

- Cluster analysis is used as the methodologies for understanding of the market segments and buyer behaviors.
  - For example, B. Zafer et al. (2006), Ya-Yueh et al. (2003), Vicki et al. (1992) , Rajendra et al. (1981) , Lewis et al. (2006) , Hua-Cheng et al. (2005)
- Market segmentations using cluster analysis have been examined in many different industries.
  - For instance, finance and banking (Anderson et al, 1976, Calantone et al, 1978), automobile (Kiel et al, 1981), education (Moriarty et al, 1978), consumer product (Sexton, 1974, Schaninger et al, 1980) and high technology industry (Green et al, 1968).

Cluster Analysis for Outlier Detection

- An Outlier is an observation that deviates so much from other observations as to arouse suspicion that it is generated by a different mechanism (Hawkins, 1980)
- Literatures find outliers as a side-product of clustering algorithms (Ester et al, 1996; Zhang et al, 1996; Wang et al. 1997; Agrawal et al. 1998; Hinneburg and Keim 1998; Guha et al, 1998..)
  - Distance-based outliers (Knorr and Ng, 1998, 1999; Ramaswamy et al., 2000)
  - Cluster-based outliers (Knorr and Ng 1999; Jiang et al, 2001, He et al, 2003 ;Duan et al, 2009;)

Research Question:

How can we apply clustering models for detection of abnormal (fraudulent/erroneous) transactions in continuous auditing?

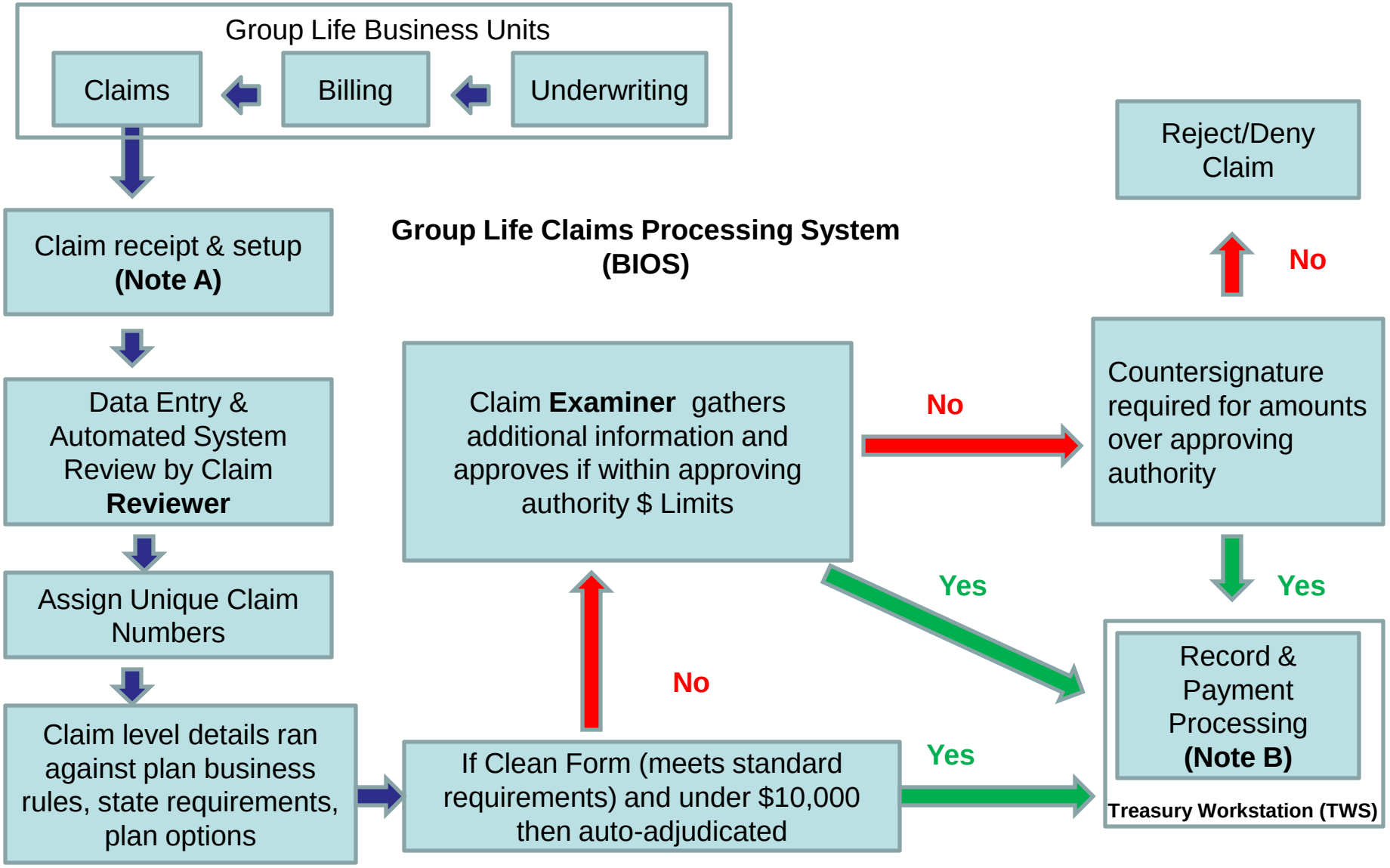
The Setting: Group Life Claim

Purpose

- To detect potential fraud or errors in the group life claims process by using clustering techniques

Data

- Group life claim from a major insurance company from Q1: 2009
- Approximately 184,000 claims processed per year (~40,000 claims per quarter)

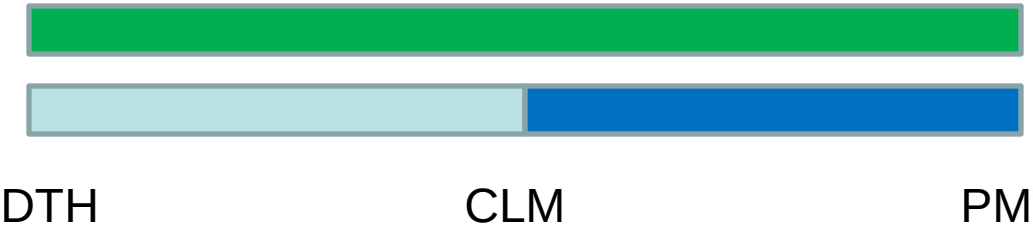


Note A: The Key elements of a claim include Employer's Statement, Beneficiary designation and Enrollment Forms are submitted via the online system . Claimant's statement and death certificate are submitted via paper. All paper documents supporting the claims are imaged.
Note B: Payments are made to beneficiary(s) in one instance but can be made to multiple beneficiaries.

Clustering Procedure

Clustering Algorithm:

- K-mean Clustering



Attributes:

- Percentage: Total interest payment / Total beneficiary payment
  - **N_percentage=(percentage-MEAN)/STD**
- AverageCLM_PMT: Average number of days between the claims received date to payment date (the weighted average is used because a claim could have multiple payment dates)
  - **N_AverageDTH_PMT=(AverageDTH_PMT-MEAN)/STD**
- DTH_CLM: Number of days between the death dates to claim received date.
  - **N_DTH_CLM=(DTH_CLM-MEAN)/STD**
- AverageDTH_PMT: Average number of days between the death dates to the payment dates (the weighted average is used because a claim could have multiple payment dates)
  - **N_AverageDTH_PMT=(AverageDTH_PMT-MEAN)/STD**

Cluster centroids:

Attribute	Full Data (40080)	Cluster# 0 (2523)	1 (54)	2 (84)	3 (222)	4 (295)	5 (31)	6 (768)	7 (36103)
N_AverageDTH_PMT	0	0.6374	**15.177**	**3.5419**	6.9858	0.8778	**10.9006**	2.7806	-0.1937
N_percentage	0	0.2666	**1.8334**	**9.3405**	0.5042	3.4637	**26.6913**	0.3185	-0.1057

Clustered Instances

0	2523 (6%)
1	**54 (0%)**
2	**84 (0%)**
3	222 (1%)
4	295 (1%)
5	**31 (0%)**
6	768 (2%)
7	36103 (90%)

Attributes:

N_AverageDTH_PMT: Normalized Average number of days between the death dates to the payment dates (the weighted average is used because a claim could have multiple payment dates)

N_Percentage: Normalized Total interest payment / Total beneficiary payment

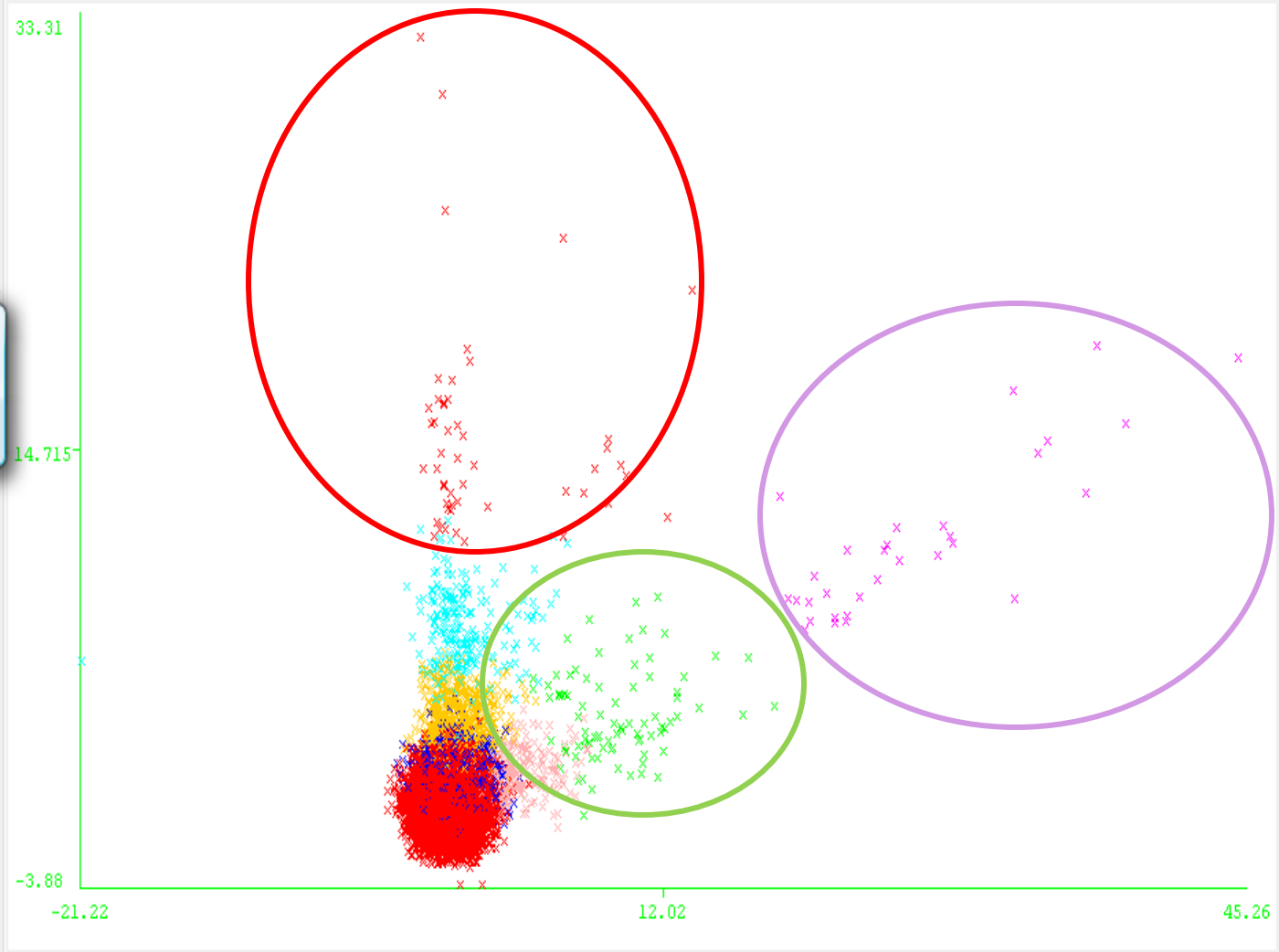
Weka Clusterer Visualize: 00:27:24 - SimpleKMeans (TestSetPayment2)

X: N_percentage (Num) | Y: N_AverageDTH_PMT (Num)

Colour: Cluster (Nom) | Select Instance

Reset | Clear | Open | Save | Jitter

Plot: TestSetPayment2_clustered



Cluster1: 54 claims

Cluster2: 84 claims

Cluster5: 31 claims

Class colour

cluster0 cluster1 cluster2 cluster3 cluster4 cluster5 cluster6 cluster7

Cluster centroids:

Attribute	Cluster# Full	Full Data	0	1	2	3	4	5	6	7	8	9	10	11	12
		(40080)	(510)	(343)	(194)	(98)	(3699)	(30)	(1275)	(741)	(32658)	(286)	(39)	(110)	(97)
N_AverageCLM_PMT		0	3.33	5.85	**1.12**	**0.93**	0.27	**1.08**	1.44	-0.02	-0.26	0.33	**1.28**	**9.81**	**4.04**
N_DTH_CLM		0	0.05	0.29	**5.63**	**9.27**	-0.10	**11.51**	-0.11	0.83	-0.13	2.89	**17.31**	**0.40**	**0.49**
N_AverageDTH_PMT		0	1.24	2.37	**5.64**	**8.93**	0.01	**11.06**	0.40	0.78	-0.21	2.79	**16.50**	**3.80**	**1.90**
N_percentage		0	0.21	0.16	**1.78**	**0.66**	0.11	**26.89**	0.51	0.48	-0.12	1.00	**2.22**	**0.30**	**7.78**

Clustered Instances

0	510 (1%)
1	343 (1%)
2	**194 (0%)**
3	**98 (0%)**
4	3699 (9%)
5	**30 (0%)**
6	1275 (3%)
7	741 (2%)
8	32658 (81%)
9	286 (1%)
10	**39 (0%)**
11	**110 (0%)**
12	**97 (0%)**

Attributes:

N_AverageCLM_PMT: Normalized average number of days between the claim received date to the payment dates (the weighted average is used because a claim could have multiple payment dates)

N_DTH_CLM: Normalized number of days between the death date to the claim dates

N_AverageDTH_PMT: Normalized Average number of days between the death dates to the payment dates (the weighted average is used because a claim could have multiple payment dates)

N_Percentage: Normalized Total interest payment / Total beneficiary payment

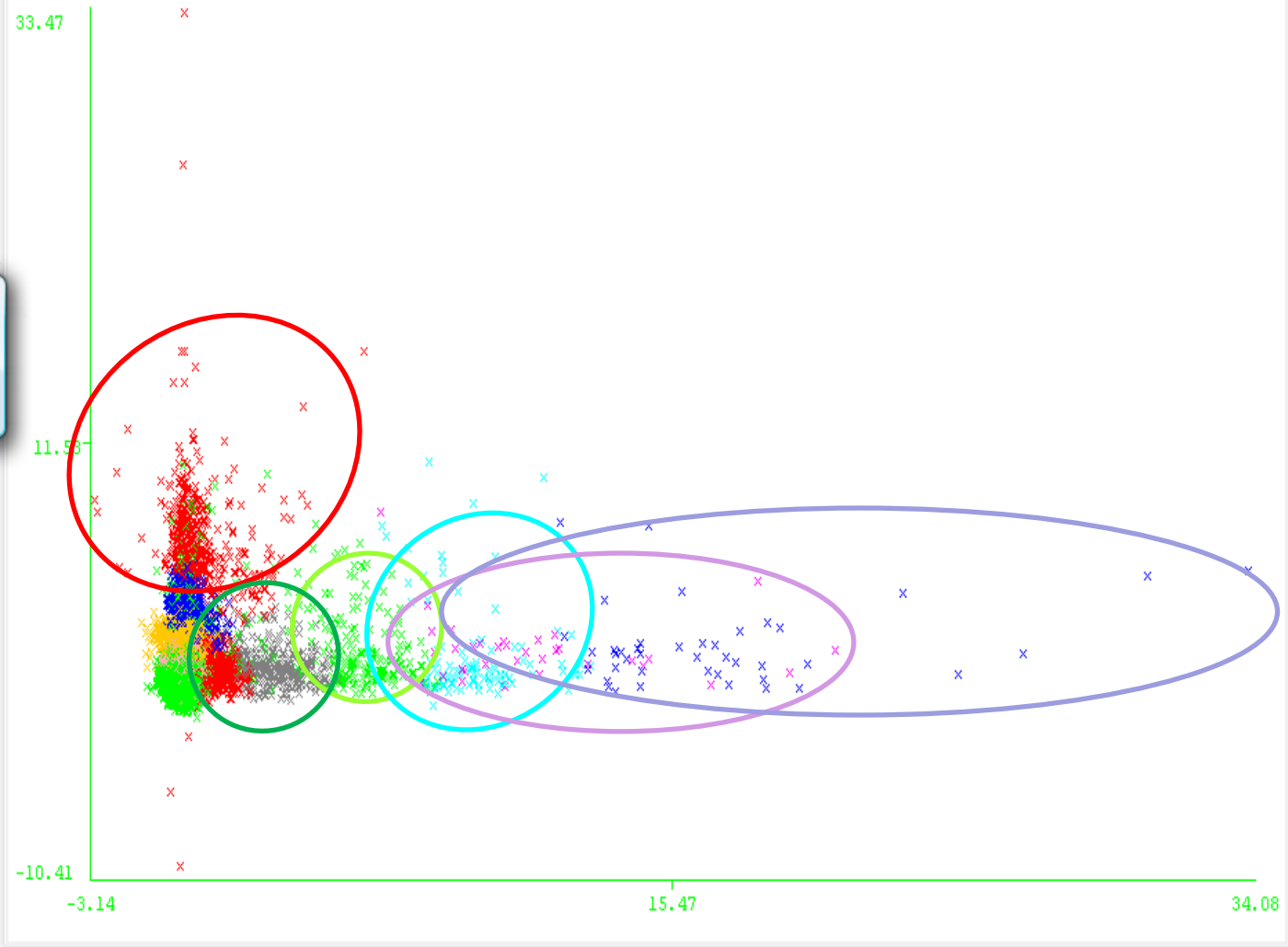
Weka Clusterer Visualize: 15:56:53 - MakeDensityBasedClusterer (TestSetPayment4)

X: N_DTH_CLM (Num) | Y: N_AverageCLM_PMT (Num)

Colour: Cluster (Nom) | Select Instance

Reset | Clear | Open | Save | Jitter

Plot: TestSetPayment4_clustered



Cluster 2:194 claims

Cluster 3:98 claims

Cluster 5:30 claims

Cluster 10:39 claims

Cluster 11:110 claims

Cluster 12:97 claims

Class colour

cluster0 cluster1 cluster2 cluster3 cluster4 cluster5 cluster6 cluster7 cluster8 cluster9 cluster10 cluster11 cluster12

Distance-Based Outliers

- A distance-based outlier in a dataset is a data object having a distance far away from the center of the cluster.
- Probability distribution over the clusters for each observation is calculated.
- The observations which has lower than 0.6 would be identified as possible outliers.

CLM_ID	Cluster0	Cluster1	Cluster2	Cluster3
20808005145	0.00021	0.806916	0.000961	0.191913
20808005307	0	0.114174	0.002238	0.883588
20808005512	0.000075	0.973995	0.000095	0.025835
20808007974	0.96161	0.036733	0.000011	0.001646
…..	…	…	…	..

Distance-Based Outliers

Simple K-mean: 2 attributes

Cluster	Outliers
Cluster 0	154
Cluster 1	0
Cluster 2	6
Cluster 3	9
Cluster 4	22
Cluster 5	2
Cluster 6	36
Cluster 7	96

Simple K-mean: 4 attributes

Cluster	Outliers
Cluster0	31
Cluster1	21
Cluster2	7
Cluster3	2
Cluster4	205
Cluster5	2
Cluster6	49
Cluster7	46
Cluster8	157
Cluster9	11
Cluster10	0
Cluster11	12
Cluster12	4

Results: Distance-based AND Cluster-based outliers

Cluster Analysis	Cluster-Based Outliers	Distance-Based Outliers
Cluster Analysis with 2 Attributes	169	325
Cluster Analysis with 4 Attributes	568	547

- Cluster-based outliers can be used to identify clusters with smaller populations as outliers.
- Distance-based outliers can be used to identify specific observations from clusters as outliers.

Limitations

- Cluster Analysis always generates clusters, regardless of the properties of the data-set. Therefore, the interpretation of the results might not be clear.
- Identification of anomalies will have to be verified.

Future Research

- More attributes related to other aspect of the claims would be used .
- Rule-based selection processes would be incorporated to help in identification of anomalies.