Outliers in COVID 19 data based on rule representation - the analysis of LOF algorithm

{ Agnieszka Nowak-Brzezińska, Czesław Horyń } University of Silesian in Katowice

Aim of research

Our work aims to detect unusual **hidden** data (outliers) in complex data, which are decision rules. Intelligent expert applications will soon be able to diagnose disturbing symptoms of the disease quickly. These types of applications operate based on decision rules. **Applications:**

1. network traffic,
2. potential intrusion detection,
3. fraud detection,
4. in medical applications: patient monitoring to detect critical, possibly life-threatening situations.

Summary

Figure 3, confirms that in each case, with 1%, 5%, and 10% of the outliers detected, **the cluster quality assessment indicators improved, after removing them from the dataset.** We applied the LOF algorithm to complex data, such as rules in knowledge bases.

	Phase I – before removing outliers					k	Phase II – after removing ouliers			
	Silhouette	Dunn	Davies-Bouldin	CPCC		k	Silhouette	Dunn	Davies-Bouldin	CPCC
1%	0.312414	0.4	0.483487	0.604610		7	0.316253	0.4	0.470753	0.607284
5%	0.312414	0.4	0.483487	0.604610		300	0.320450	0.4	0.475842	0.635482
10%	0.312414	0.4	0.483487	0.604610		3	0.325561	0.47	0.479455	0.611951

Figure 3: Compare the quality of cluster with oulier and after their removal

In the description of the rules to a large extent, apart from quantitative features, there are **qualitative features.** When using this algorithm for real data, we have to execute it many times, changing the algorithm's parameter (degree of the neighbourhood). **Consultation with a domain expert confirms, that all the rules indicated as deviations are in some sense atypical.**

Outliers in Rules

According to the definition given by **Grubbs in 1969:** *"An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs".* Today outliers are known to have two important characteristics: outliers are different from the norm concerning their features, and they are rare in a dataset compared to normal instances. **Unusual rules are not errors but have got unusual features and differs from other rules.** They need to be discovered by knowledge engineers and discussed with knowledge experts, who may extend the domain to which an unusual rule belongs.

Experiments phase I

We can see that the optimal index values, were obtained for **2 clusters**, and the **centroid** method.

id	Silhouette	Dunn	Davies-Bouldin	CPCC	Clusters	Clustering method
1	0.256469	0.1	1.255498	0.627978	2	Ward.D
2	0.133531	0.1	1.817726	0.627978	3	Ward.D
...						
48	0.123899	0.111111	2.029511	0.627978	49	Ward.D
49	0.124642	0.111111	2.017838	0.627978	50	Ward.D
1	0.149606	0.1	1.875984	0.518124	2	Ward.D2
2	0.189705	0.1	1.540459	0.518124	3	Ward.D2
...						
48	0.151667	0.125	2.104539	0.518124	49	Ward.D2
49	0.147501	0.125	2.099490	0.518124	50	Ward.D2
1	0.255129	0.4	0.577142	0.431075	2	single
2	0.224813	0.4	0.541113	0.431075	3	single
...						
48	-0.230618	0.2	0.788247	0.431075	49	single
49	-0.230997	0.2	0.794906	0.431075	50	single
1	0.194951	0.1	1.376539	0.534238	2	complete
2	0.079259	0.1	2.262753	0.534238	3	complete
...						
48	0.039924	0.166667	2.030654	0.534238	49	complete
49	0.047490	0.166667	2.040525	0.534238	50	complete
1	0.283532	0.2	1.018698	0.747380	2	mcquitty
2	0.226740	0.2	1.488808	0.747380	3	mcquitty
...						
	0.056885	0.125	1.774643	0.747380	49	mcquitty
	0.058056	0.125	1.833024	0.747380	50	mcquitty
1	0.251177	0.2	0.518758	0.518853	2	median
2	0.158096	0.2	0.530860	0.518853	3	median
...						
	-0.288508	0.1	1.209505	0.518853	49	median
	-0.295951	0.1	1.198937	0.518853	50	median
1	**0.312414**	**0.4**	**0.483487**	**0.604610**	2	centroid
2	0.208589	0.2	0.512787	0.604610	3	centroid
...						
	-0.273642	0.2	0.895063	0.604610	49	centroid
	-0.275366	0.2	0.891933	0.604610	50	centroid

Table 1: Phase I - looking for the optimal values of clustering quality

Table 1 shows a repetition for 49 different values of the number of groups, and 8 different methods of combining the values of 4 selected measures of cluster quality assessment: **Dunn, Davies-Bouldin, silhouette and CPCC measures.**

Research methodology

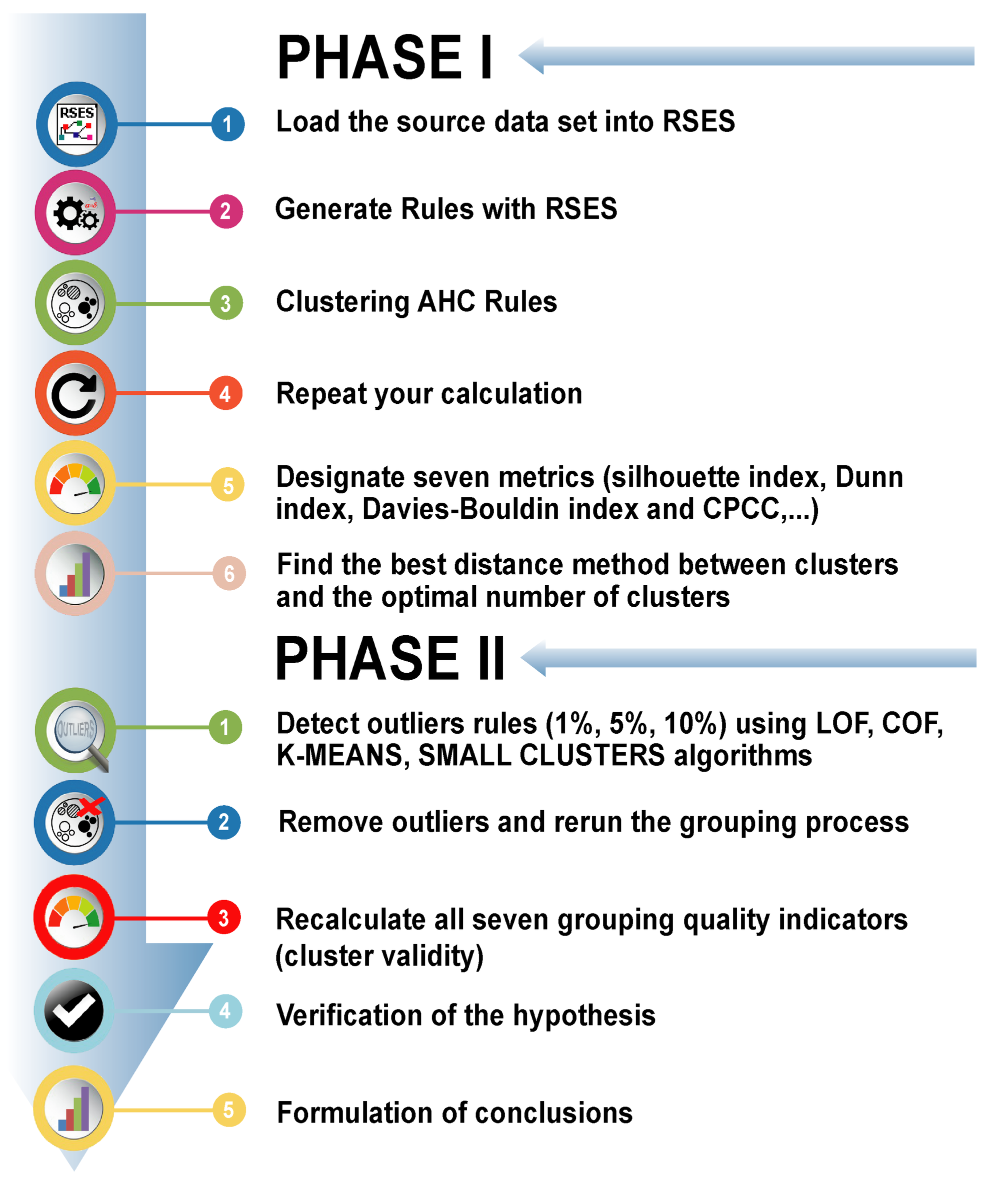

Figure 1: Research methodology. **Dataset:** *Covid 19 Case Surveillance Public Use Data* have qualitative characteristics provided by CDC (data.cdc.gov)

Phase I - rules induction and clustering:

1. Loading source dataset into RapidMiner Studio and sampling using RapidMiner,
2. Loading the source dataset into RSES,
3. Generating rules with RSES using LEM2 algorithm,
4. Rule clustering. Find the value of as many metrics and indicate the optimal result.

Phase II - we want to learn about unusual rules, which constitute 1 % , 5 %, 10 % of all rules in the knowledge base:

1. Removing outlier rules from the knowledge base and returning the clustering,
2. Recalculating all four grouping quality indicators,
3. Verifying how many cases the quality of the cluster improved. Analysing the results of the studies.

Experiments phase II

	LOF outliers id
1%	359, 519, 190, 322, 2394, 177, 231, 2356, 276, 218, 196, 948, 2302, 176, 261, 337, 1866, 1015, 2440, 766, 353, 1833, 1902, 362, 2279
5%	2489, 485, 165, 242, 504, 105, 479, 471, 483, 2419, 102, 103, 243, 79, 475, 377, 128, 52, 129, 2173, 2364, 164, 537, 80, 2135, 2358, 51, 172, 24, 100, 63, 244, 2347, 359, 2377, 252, 40, 2176, 2367, 499, 2155, 2156, 2154, 1330, 1251, 2487, 1354, 1329, 2344, 512, 484, 2411, 202, 2465, 2475, 1208, 526, 251, 246, 358, 2325, 2368, 2164, 2157, 2159, 2175, 468, 1209, 2212, 362, 2393, 2151, 2491, 352, 152, 2482, 173, 2404, 1116, 1152, 71, 1346, 490, 203, 521, 2392, 2455, 1244, 167, 2275, 1299, 346, 1118, 2328, 1312, 1131, 1227, 1194, 1141, 525, 2149, 2169, 1243, 133, 498, 2178, 488, 2397, 1185, 727, 1305, 1262, 126, 354, 1275, 85, 1231, 1229, 1335, 1268, 1255, 1267, 106, 469, 1170
10%	2231, 359, [illegible]82, 2167, 1017, 1358, 519, 2160, 231, 2105, 212, 1270, 2188, 1144, 461, 2394, 85, 70, 1206, 136, 2134, 190, 189, 332, 1337, 350, 218, 33, 346, 2425, 171, 25, 340, 2177, 134, 2033, 40, 322, 2224, 480, 192, 337, 1013, 2120, 221, 1369, 108, 250, 42, 8, 2393, 1946, 1914, 1204, 1322, 196, 948, 5, 2228, 1071, 2423, 2021, 6, 2166, 1290, 1242, 90, 1366, 525, 86, 2029, 456, 1010, 1166, 425, 18, 1356, 304, 524, 138, 2101, 1115, 1288, 109, 2041, 1849, 591, 1314, 2119, 2366, 78, 1162, 179, 1263, 32, 254, 1937, 536, 2410, 21, 642, 2110, 1866, 826, 1153, 1364, 2418, 295, 2093, 881, 919, 903, 819, 529, 947, 936, 284, 143, 713, 1150, 1015, 1075, 1916, 1961, 2133, 2323, 1833, 1902, 1847, 2190, 1229, 1268, 2072, 2014, 1345, 2331, 1734, 1671, 1516, 2100, 916, 1977, 2010, 2053, 407, 2126, 1714, 1703, 463, 841, 865, 2036, 2031, 1921, 549, 1825, 1851, 1874, 1048, 481, 1363, 1987, 1378, 2027, 806, 1442, 1417, 2091, 1568, 370, 1462, 2248, 2270, 2303, 1544, 450, 438, 1821, 1892, 663, 733, 1470, 620, 652, 1505, 2244, 1643, 651, 1373, 394, 1912, 398, 490, 1396, 554, 1498, 1635, 215, 558, 1447, 1283, 1333, 1618, 1404, 1596, 1398, 382, 2311, 77, 1604, 771, 548, 556, 801, 1523, 1482, 2443, 57, 477, 1611, 1627, 1674, 1606, 1381, 2173, 419, 1009, 1471, 584, 623, 632, 1426, 1571, 193, 300, 2489, 121, 547, 567, 587, 624, 846, 1474, 837, 1901, 391, 2312, 1936, 1481, 1070

Figure 2: Phase II - lists of detected ouliers for 1 %, 5 % and 10 % case

The second phase results, in which we are looking for outliers in the data in the optimal structure of rule groups established in the first phase. For this purpose, **we choose 1%, 5%, and 10% of the most unusual rules using the LOF method.** We are looking for cases that, after eliminating unusual rules, will improve the evaluation of the cluster quality. **The most unusual rule, indicated by all trials, is Rule No. 359** When we look at it closely, it becomes clear, why it is unusual. Besides indicating the age group and gender, all other features (attributes) are undefined (value unknown). The rule is shown to the field expert for evaluation.

References

Nowak-Brzezińska A., Horyń C. **"Exploration of Outliers in If-Then Rule-Based Knowledge Bases"** Entropy, 22 (10) (2020) https://doi.org/10.3390/e22101096

Future Research

Future research will focus on two issues: research on **improving the effectiveness of the inference process** in rule-based knowledge bases, after the elimination of outlier rules, and further **search for algorithms, that will detect unusual rules even more effectively.**

Contact Information

Web https://us.edu.pl/instytut/ii/
Email czeslaw.horyn@us.edu.pl
Phone +48-32-3689757