Lab Semantic Data Web

Outlier Detection on Financial RDF Data

Mentor

Christiane Engels

Group students

Zuhair Almhithawi, Nayef Roqaya, Berivan Ekmez

Lab Project Report

Summer semester / 2016

ENTERPRISE INFORMATION SYSTEMS

universitätbonn

Table of Contents

1 Introduction:

1.1 Motivation:

The World Wide Web has enabled the creation of a global information space comprising linked documents. Linked Data provides a publishing paradigm in which not only documents, but also data, can be a first class citizen of the Web, thereby enabling the extension of the Web with a global data space based on open standards[1].

We use the Linked Open Data Cloud for retrieving additional information on e.g. demographics or economics to enrich the data sets at hand before analyzing them. The project aims to detect outliers on financial data and to compare the results. Certain data mining methods such as Chauvenet and K-means can be applied to these separate datasets (Financial data) to efficiently detect anomalies/outliers and these datasets were enriched from external sources like DBPedia to have accurate findings.

1.2 Project Purpose:

Applying different methods to detect outliers and anomalies in financial RDF data and compare the results.

1.3 Definitions:

1.3.1 What is an outlier/anomaly?

An anomaly is something that deviates from what is standard, normal, or expected **[2].**

1.3.2 Anomaly detection in data-mining

In data mining, anomaly detection (outlier detection) is the identification of items, events or observations, which do not conform to an expected pattern or other items in a dataset [3].

1.3.3 Outliers in financial data:

One aspect of analyzing financial data is finding unusual values,i.e. outliers or anomalies. These may indicate:

- errors in the data
- irregular behavior (corruption, fraud, . . .)
- regions of special interest that e.g. require more subsidies or a better handling of those

2 Project Time plan:

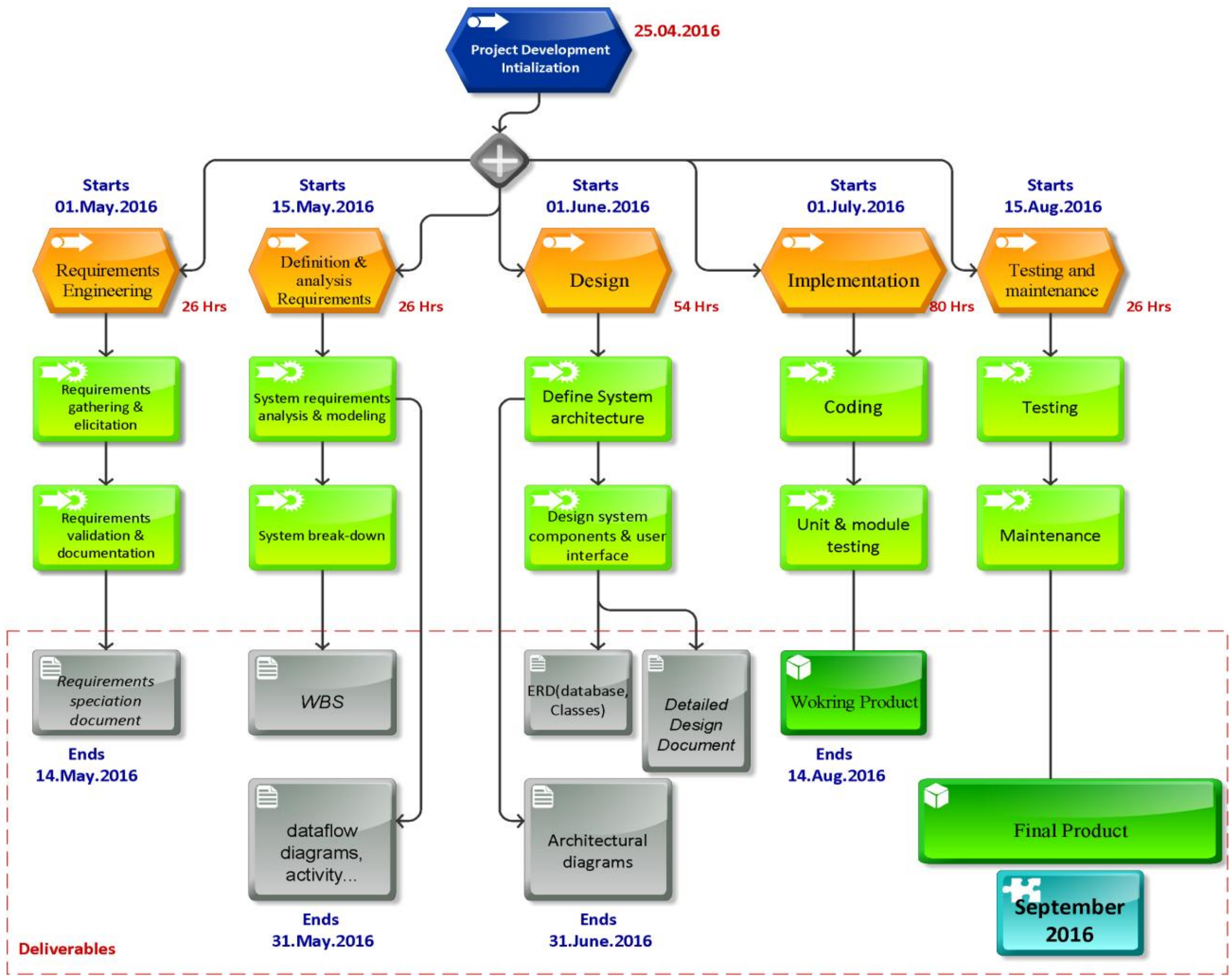


Figure1:Time plan

3 System Requirements:

3.1 Functional requirements:

ID	Name	Comments
FRQ - 1	The user shall import his/her RDF datasets	Datasets may be in XML/RDF format
FRQ - 2	The system shall Link local datasets to external dataset cloud	External dataset will be chosen manually the user in financial field
FRQ –3	The system shall provide reports on analyzed data	Reports may include: Comparison for DM methods Results visualization Explanations

3.2 Technical Requirements:

- User interface: cross browser interface that runs on most web browsers (e.g. IE, Chrome, Firefox, Safari)
- The system shall be implemented using Java SE environment
- Front end languages and technologies: HTML5/CSS3, JavaScript, JQuery library, AngularJS, Google Maps API
- Restful API for web services
- Tools to be used: Fuseki, Rapid miner and Apache Jena

3.2.1 FUSEKI[1]:

Fuseki is an HTTP interface to RDF data. It supports SPARQL for querying and updating. The project is a sub-project of Jena and is developed as servlet. Fuseki can also be run stand-alone server as it ships preconfigured with the Jetty web server.

[1] http://www.szabadsolyom.hu/go/doc/dictionary-of-modern-fuseki-korean-style.pdf

3.2.2 Apache Jena:

A free and open source Java framework for building Semantic Web and Linked Data applications (Jena.apache.org, 2016).

3.3 Use case:

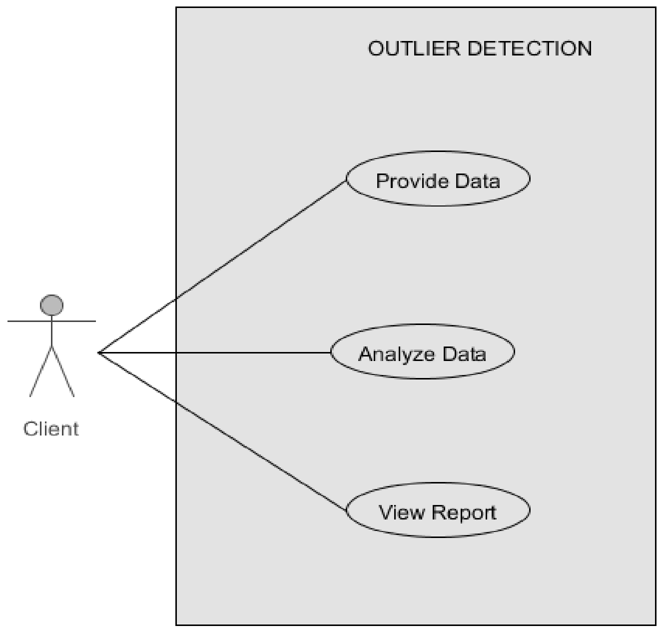


Figure 1 :Use case Diagram Level 0

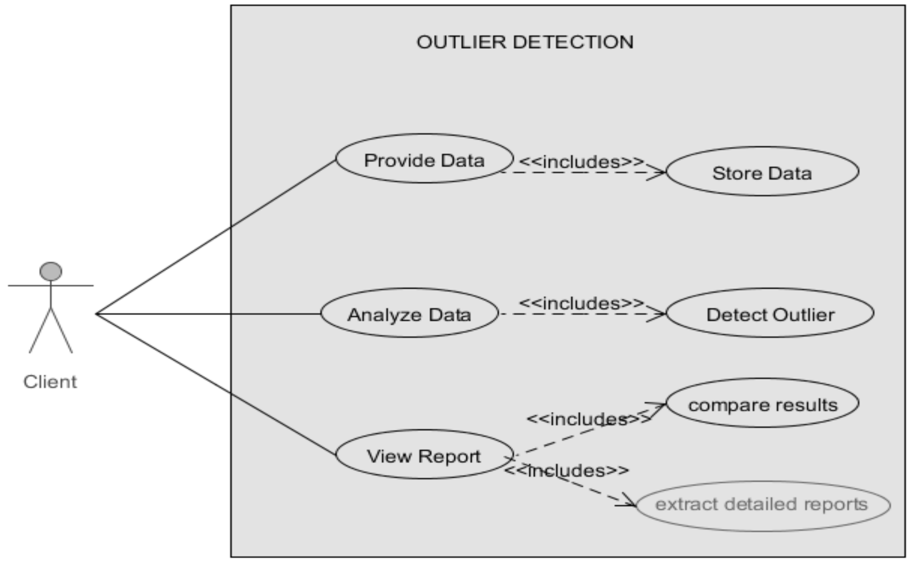


Figure 2:Use case Level 1

3.4 Non- functional requirements:

ID	Name	Comments
NRQ - 1	Usability	The system must provide an intuitive and easy-to-use interface.
NRQ - 2	Capacity	The ability of the system to handle transactional volumes is a very important characteristic for the system
NRQ – 3	Performance	The system should consider the processed transactions per second and response time to user input and performance of the implemented DM methods
NRQ – 4	Open source	Source code made available with

		a license in which the copyright holder provides the rights to study, change, and distribute the software to anyone and for any purpose.
NRQ – 5	Quality	Quality of the outlier and detection methods for giving precise result.

4 System architecture:

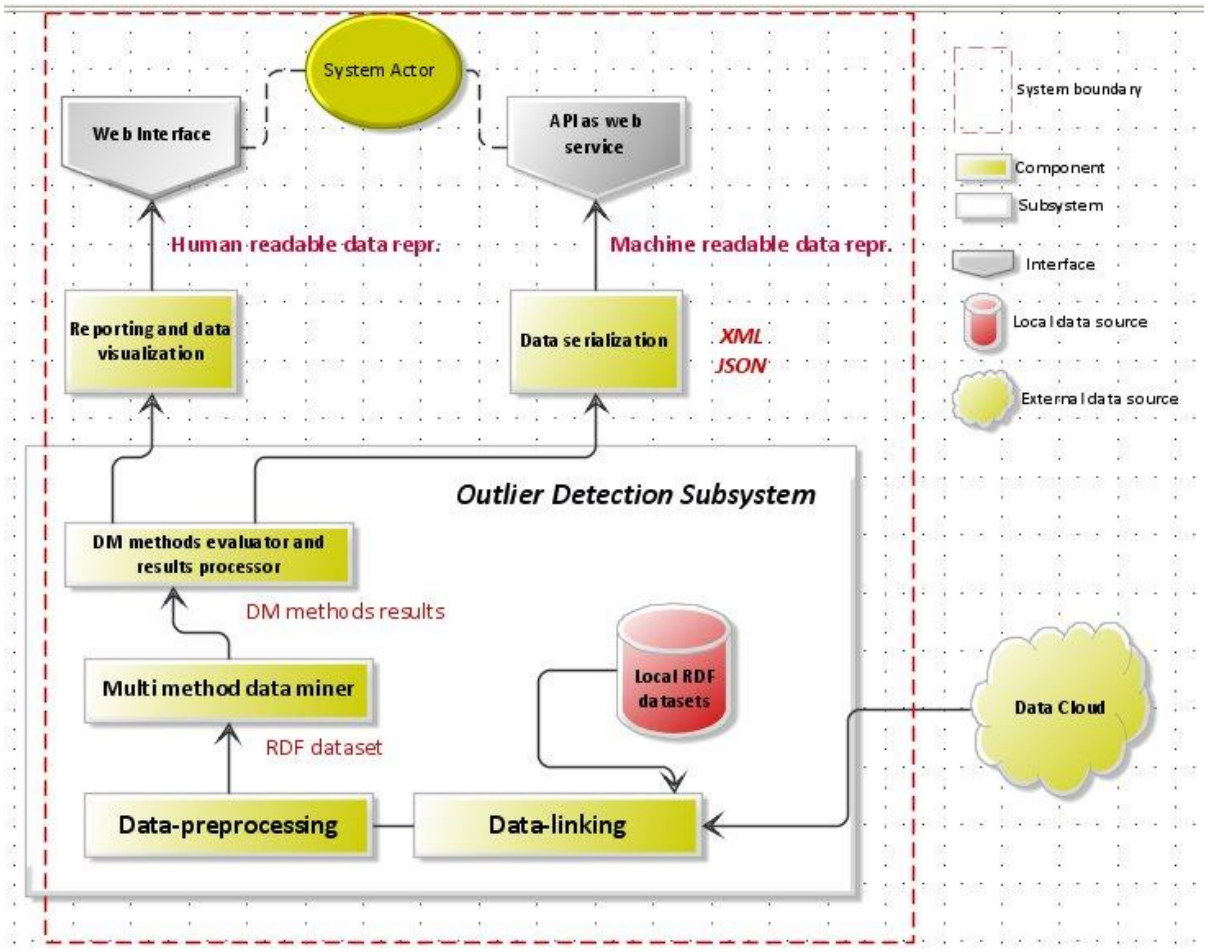


Figure 3: System Architecture

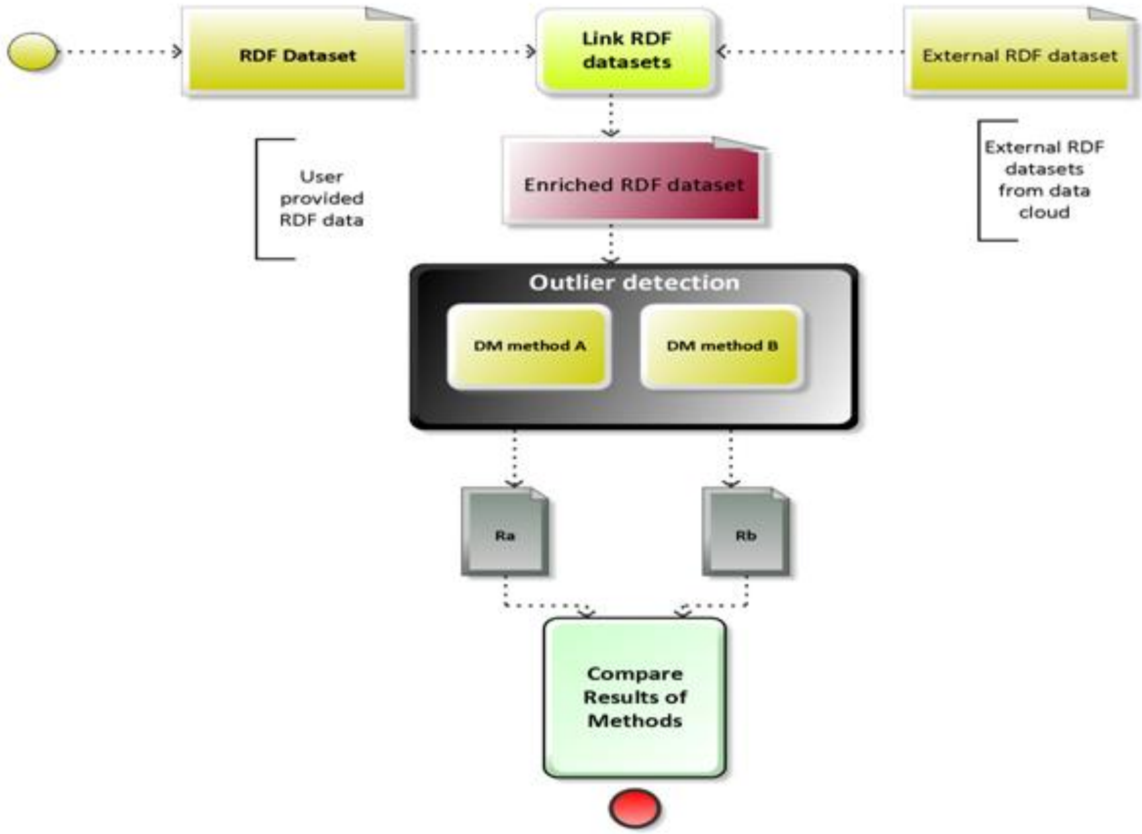


Figure 4 : System Architecture

Component ID	CP-DL
Component Name	Data-linking module
Component core functionality	Links a local RDF dataset to a DBpedia and enrich it with additional predicate-value pairs
Dependencies	Apache Jena Framework, DBpedia HTTP Endpoint
Input	Local RDF Dataset
Output	Enriched local RDF Dataset

Component ID	CP-DP
Component Name	Data-preprocessing module
Component core functionality	Applying sub-population on an RDF Dataset
Dependencies	---
Input	RDF Dataset
Output	Sub datasets from the input dataset

Component ID	CP-MMD
Component Name	Multi-method Data Mining Module
Component core functionality	Applies multiple data mining methods on sub-population RDF dataset
Dependencies	-----
Input	RDF Dataset
Output	List of outliers

Component ID	CP-ENP
Component Name	DM methods evaluating and results processing module
Component core functionality	Compares applied DM methods and processes the resulted output from component CP-MMD
Dependencies	Component CP-MMD
Input	Output of component CP-MMD
Output	Statistical comparison

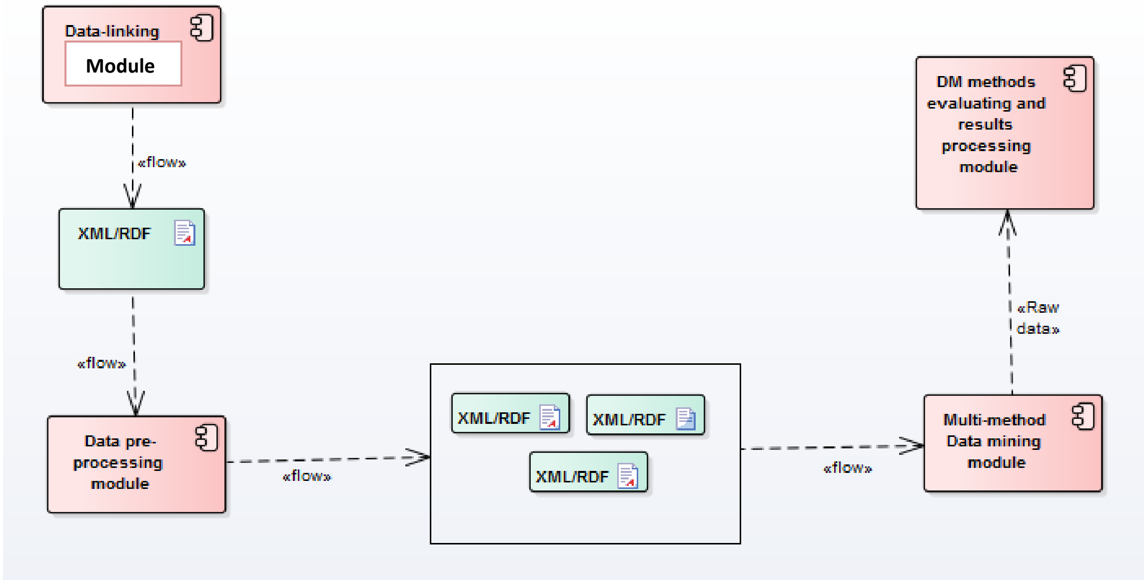


Figure 5 : DFD diagram of system components

4.1 Subpopulation:

A subset of a population that shares one or more additional properties is called a subpopulation. For example, if the population is all German people, a subpopulation is all German males; if the population is all Ph.D. students in the world, a subpopulation is all Ph.D. in Germany. By contrast, a sample is a subset of a population that is not chosen to share any additional property. Before starting subpopulation, there are some concepts that have to be clear because this concept will play a vital role to have a correct subpopulation. In the RDF file, we have many triples that are attached to every observations .Each triple is formed by subject, predicate, and object. For instance, (Germany - located in - Europe = S:P:O).*Located in* is a property and *Europe* is a value of this property then *located in* = *Europe* is a constraint .

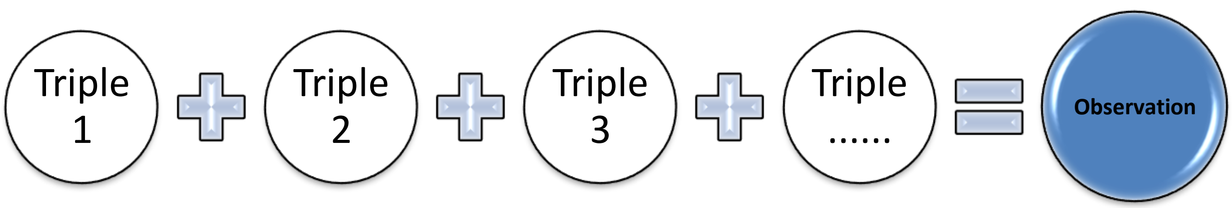


Figure 6: Observation structure

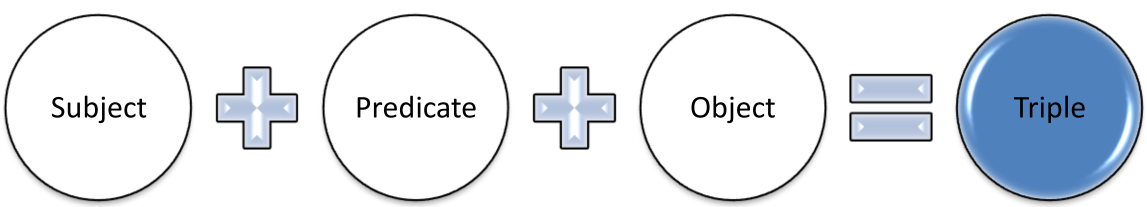


Figure 7: Triple structure

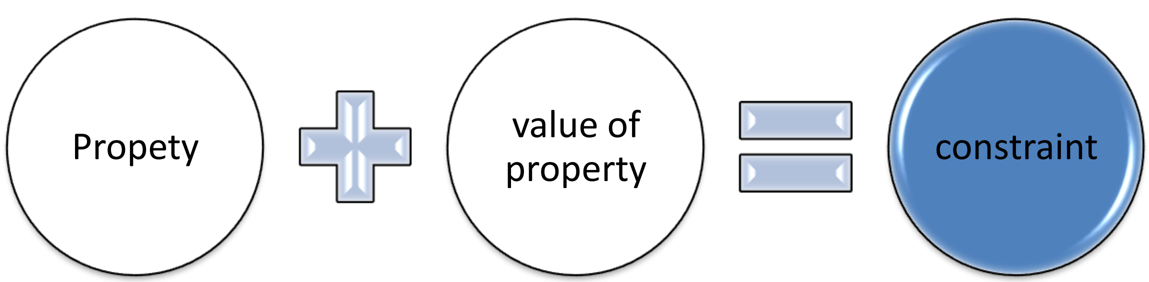


Figure 8: Structure constraint

There are three different types of constraints:

- **Class constraints**: A class constraint on class C applied to an instance set limits it to instances which belong to this class.
- **Property constraints**: A property constraint p limits the instances to those connected to an arbitrary object (instance or data value) by means of p.
- **Property value constraints**: A property value constraint is defined by a property **p** and a value **v** which can be either an instance or a data value. It limits the instances to those which are connected to a value **v** by means of **p**, and we are using only this approach[2].

4.1.1 Pre-subpopulation:

1. Use parser to extract the triples from data set file (for instance file.ttl).
2. Format these triples in this formula: "S: P: O".

4.1.2 Subpopulation [1] :

1. The root node includes all triples (instances) that are retrieved from the RDF parser.
2. For the set of instances in the root node, we start to compute a histogram which represents the distribution of values in the subpopulation.
3. Starting with the root node, this approach manages a queue of all not yet extended nodes and thus extends the lattice in a breadth-first-manner.
4. When processing a node from this queue, we create its child nodes, each having an additional constraint compared to the parent node.
5. The additional constraints are those from the set of possible constraints which are not yet used in the parent node.
6. If a node for the resulting set of constraints already exists in the lattice, we do not consider the new node further.
7. We determine the instances which adhere to this new set of constraints and compute the histogram of the value distribution [4].

[2] http://www.heikopaulheim.com/docs/iswc_2014.pdf

8. In each step of subpopulation (each level) , we have to make correlation lattice and finding the sharing bins between the paths in the lattice.
9. We prune subpopulations which only contain a low number of instances or maybe no instances at all, we consider the instance reduction ratio. For instance, the change ratio in the number of instances of the new node compared to its parent node. If the additional constraint leads to a reduction of less than 1%, our approach prunes the node. Too low KL divergence and for not reducing the instance set further, we prune the node. Additionally if the number of instances in the bin is low , we prune the bin from lattice .

$$\text{divergence}(parent, child) = \left| \frac{|child|}{|parent|} \cdot \sum_{i=1}^{B} \ln\left(\frac{h_{\text{parent}}(i)}{h_{\text{child}}(i)}\right) h_{\text{parent}}(i) \right|$$

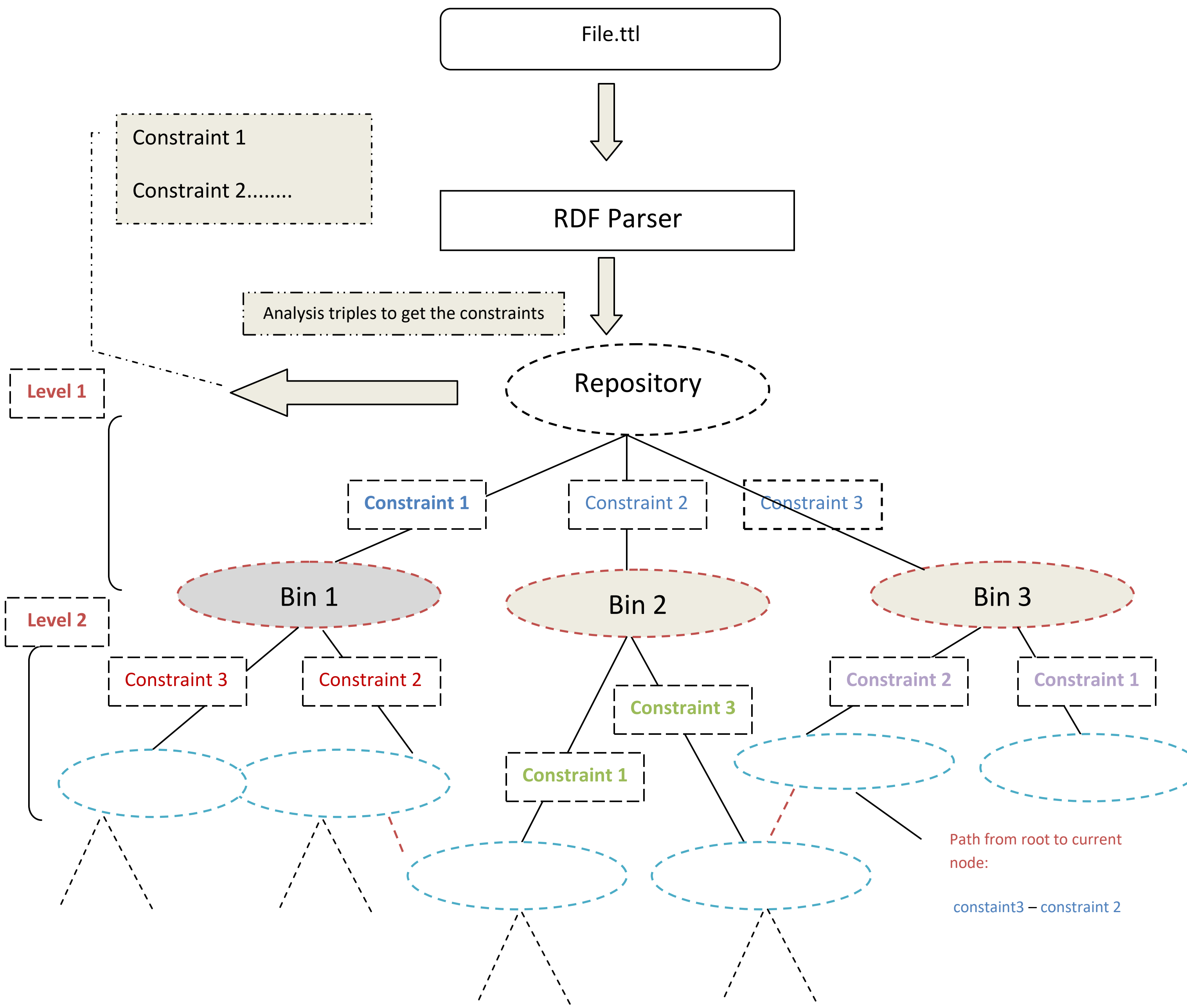


Figure 9 : Sub population - process

- It is so important to analysis the triples and concludes the values of each property especially the properties that have many different values in a data set .For instance (budget Phase= approved, budget Phase = draft ...Etc).
- Each property with its values gives constraint and value of recursive levels related to a number of properties with considering the repeating in a list of properties and constraints.
- Number of levels in subpopulation is increasing gradually then the bins that are created by subpopulation will be increased, but the rule of pruning sometimes decrease the number of the bins.

4.1.3 Affect pruning on subpopulation – histogram:

Number of the bins in each level will increase when we move from one level to another.

Before pruning:

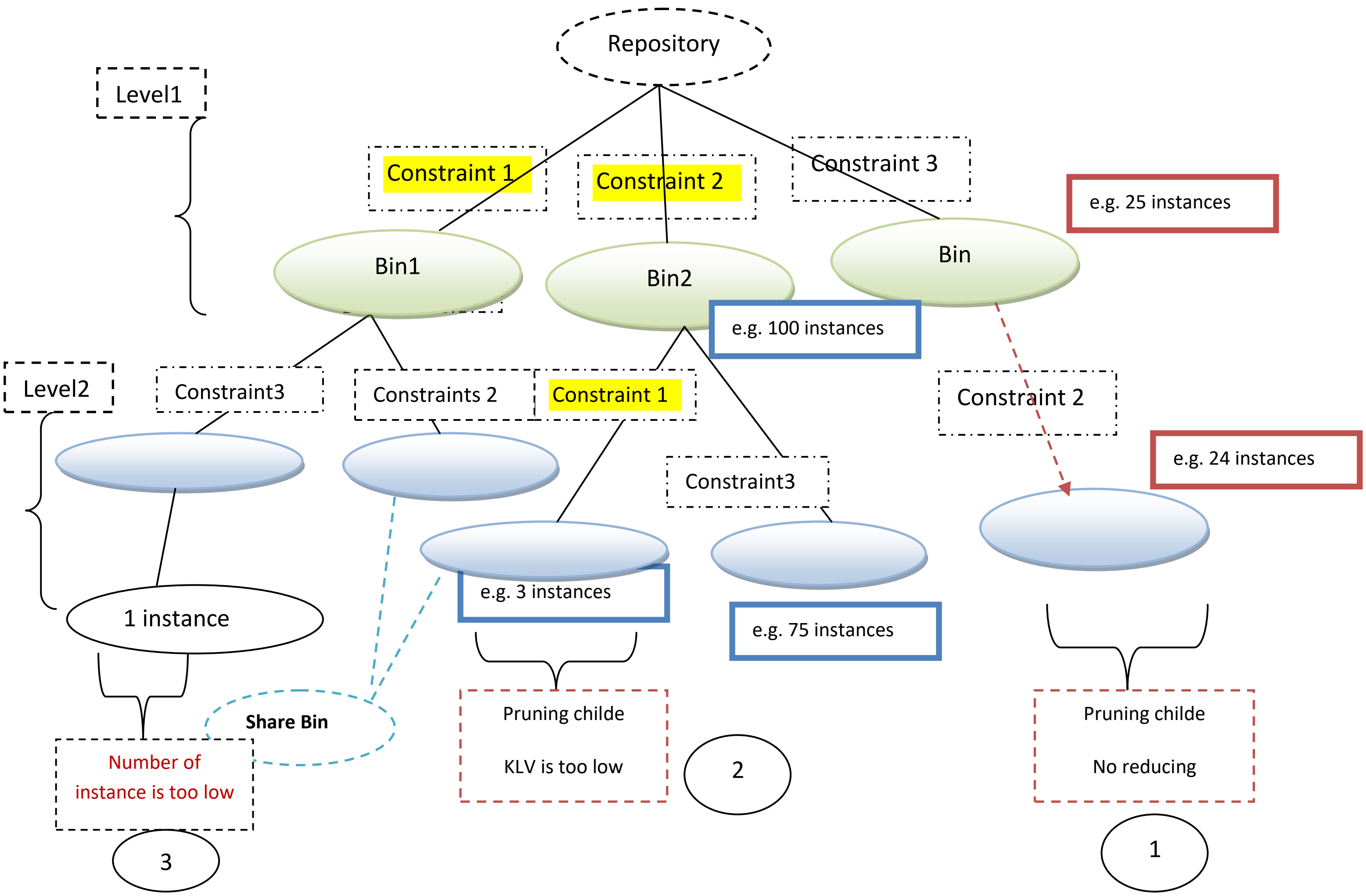


Figure 10: Pruning - subpopulation

Before pruning we can note from the graph above many important points:

Some vertices in the lattice have the same constraint (property=value) but these constraints has a different sequence.

- For each pair parent-child, we calculate the KLV value and not reducing rule between the parent and child.

After pruning:

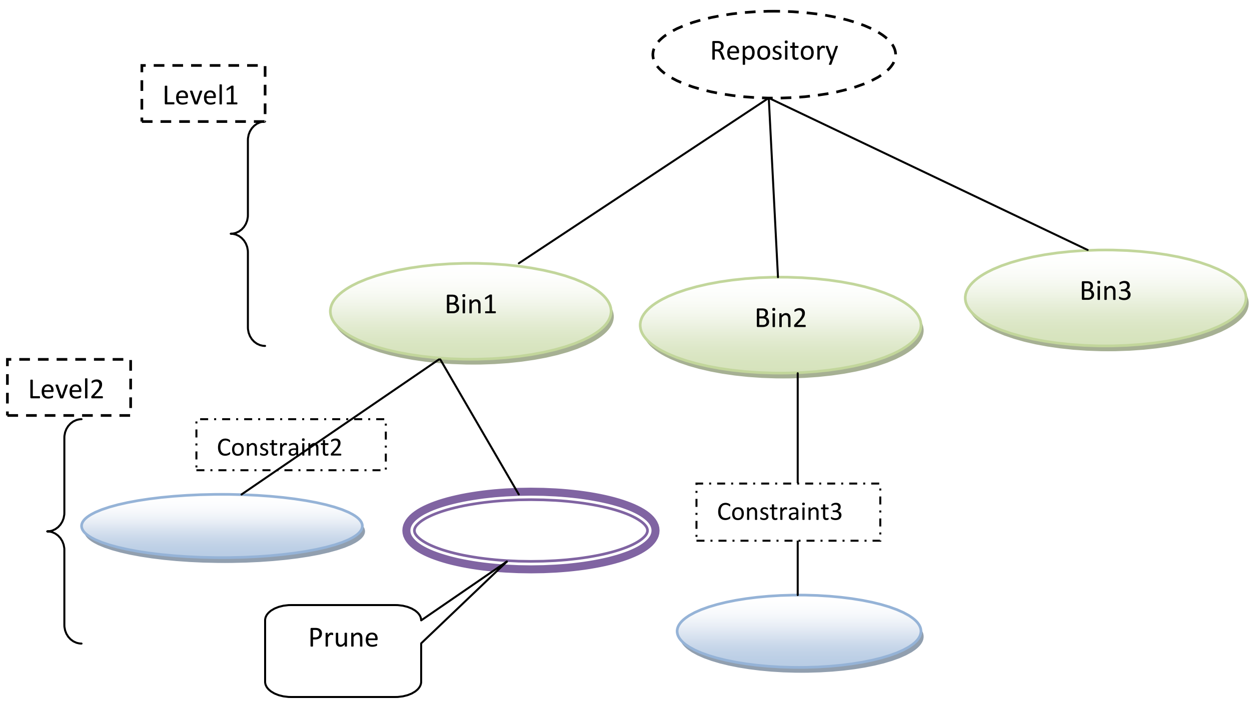


Figure 11: Lattice after pruning

We can note that the number of bins in the second level was decreased as result to apply pruning rules between the parent bins in the leve1 and children bins in the level 2 .

4.2 Outlier detection

After finishing subpopulation stage, we perform outlier detection on all non-pruned nodes of the lattice and store the resulting outliers together with the set of constraints which led to the corresponding instance set. Values may not only be detected as outliers when they are wrong but also if they are natural outliers in the considered dataset.

We chose k-means method and Chauvenet method because each method relate to different types of approach. K-means is a clustering method that is sensitive to noise and outliers comparing with Chauvenet is a mathematical method that depends on statistical approach. That will give us a chance to compare two different methods that have different approaches.

4.2.1 K-means algorithm for outlier detection:

An outlier is an object that differs from most other objects significantly. Therefore it can be considered as an anomaly. For outlier detection, only the distance to the appropriate centroid of the normal cluster is calculated. If the distance between an object and the centroid is larger than a predefined threshold dmax, the object is treated as an outlier.

In cluster analysis, a fundamental problem is to determine the best estimate of the number of clusters, which has a deterministic effect on the clustering results. However, a limitation in current applications is that no convincingly acceptable solution to the best-number-of clusters problem is available due to high complexity of real datasets [5].

Unfortunately, the researchers did not follow a specific way to solve specifying number of clusters in k-means algorithm. Some of them suggested following test results of algorithm for different number of cluster then choosing the best result and others suggest analyzing the data then deciding the number of clusters. For instance, if we want to cluster population in low, medium and high category, the number of cluster has to be K but this approach is not precise and cannot be applied for all cases [6].

K-Means clustering intends to partition n objects into k clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly k different clusters of greatest possible distinction. The best number of clusters k leading to the greatest separation (distance) is not known as a priori and must be computed from the data. The Algorithm includes these steps :

1. Clusters the data into k groups where k is predefined.
2. Select k points at random as cluster centers.
3. Assign objects to their closest cluster center according to the Euclidean distance function.
4. Calculate the centric or mean of all objects in each cluster.
5. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

K-Means is relatively an efficient method. However, we need to specify the number of clusters, in advance and the final results are sensitive to initialization and often terminates at a local optimum. Unfortunately there is no global theoretical method to find the optimal number of clusters. A practical approach is to compare the outcomes of multiple runs with different k and choose the best one based on a predefined criterion. In general, a large k probably decreases the error but increases the risk of over fitting.

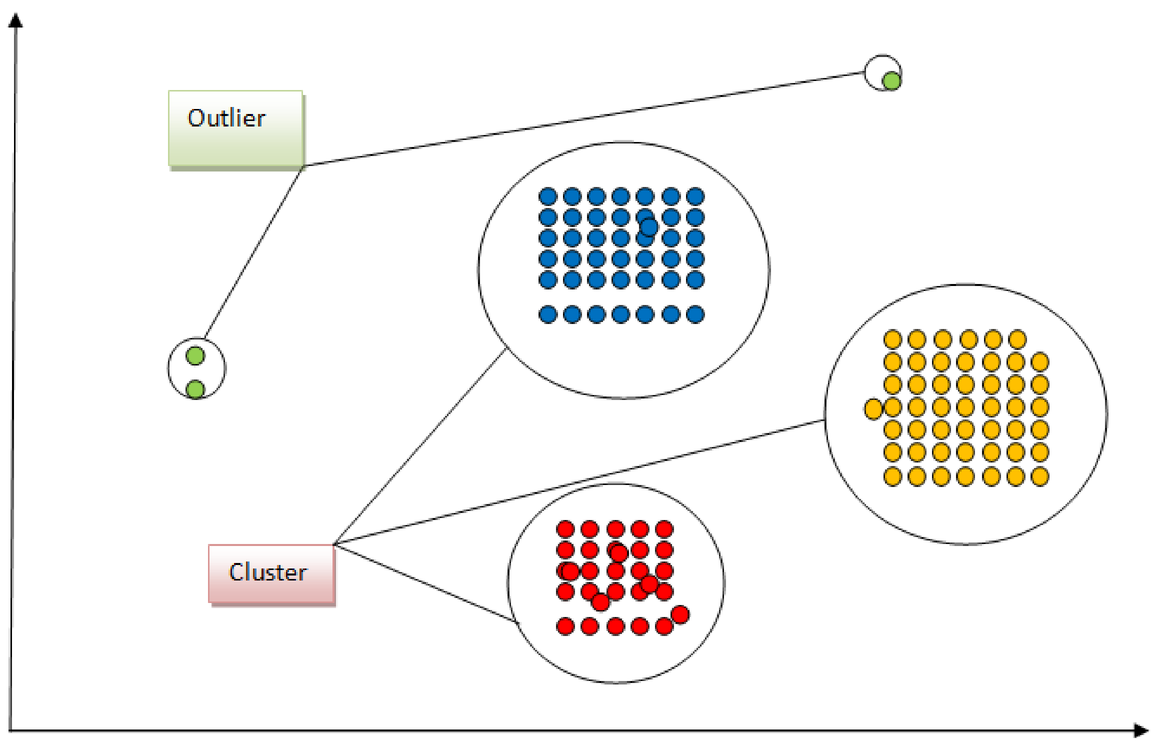


Figure 12: K-means clusters

4.2.1.1 Cons of K-means algorithm:

- K value not known precisely and that affect the results of outlier if the user did not apply the cluster for different k value and choose the best results.
- Sensitive to outliers and noise .
- Sensitive to initial points and local optimal, and there is no unique solution for a certain K value that play role in specifying outlier values[7].

4.2.1.2 Pros of K-means algorithm:

- Practical, works well even some assumptions are broken especially with a big data set that have different values
- Simple, easy to implement;
- Easy to interpret the clustering results and that help in finding the correct outlier cluster
- Fast and efficient in terms of computational cost, typically O (K*n*d).

4.2.2 Chauvenet's Criterion

The idea behind Chauvenet's criterion is to find a probability band, centered on the mean of a normal distribution that should reasonably contain all n samples of a data set. By doing this, any data points from the n samples that lie outside this probability band can be considered to be outliers, removed from the data set, and a new mean and standard deviation based on the remaining values and new sample size can be calculated.
This identification of the outliers will be achieved by finding the number of standard deviations that correspond to the bounds of the probability band around the mean (D_{max}) and comparing that value to the absolute value of the difference between the suspected outliers and the mean divided by the sample standard deviation.

Steps Chauvenet's Criterion[3]**:**

- Step 1: Calculate the sample mean
- Step 2: Find the sample standard deviation.
- Step 3: We start by calculating the mean and standard deviation of the sample, $\bar{x}$ and S. We then calculate the standardized deviation from the mean for all data values as:

$$\tau i = |Xi - \bar{x}| / s$$

- Step 4: Compare the values you got in Step 4 with a table of Chauvenet's criterion values to see if you can reject each data point.

4.2.2.1 Properties Chauvenet's Criterion:

A long established method based on probability theory that is widely used in government, universities and industry for outlier detection. It has the disadvantage that it assumes that data are from a normal distribution - always a very questionable assumption! If this is assumed, then outliers are identified based on the mean and standard deviation of the data.

4.3 Semantic Web role:

The subpopulation method can be applied on many types of constrains and by choosing the right constrains on which the subpopulation is applied, we can have more consistent outcomes of it. In our RDF dataset we have financial data about observations recorded in certain location and that location might be a city or even a country. By deploying liked data principles we can fetch additional information (that will be finally considered as constraints) for the observation location (like area and population) and then we apply the subpopulation method on these new constraints. As locations (city or country) have many properties in common like: area and population, we can apply the subpopulation method based on these properties and that will give us more consistent grouping of our samples which are the observations that were recorded in certain locations.

4.3.1 RDF data model:

4.3.1.1 The Data Cube Vocabulary overview:

[3] https://www.usna.edu/Users/mecheng/ratcliff/EM375/handouts/Statistics/06-OutlierElimination.pdf

Data Cube Vocabulary represents datasets as data cubes, i.e. collections of data that comprises of observed values (observations), associated dimensions, and metadata. The DCV provides a set of classes and properties for representing the data cubes in RDF and publishing them according to the linked data principles [19].

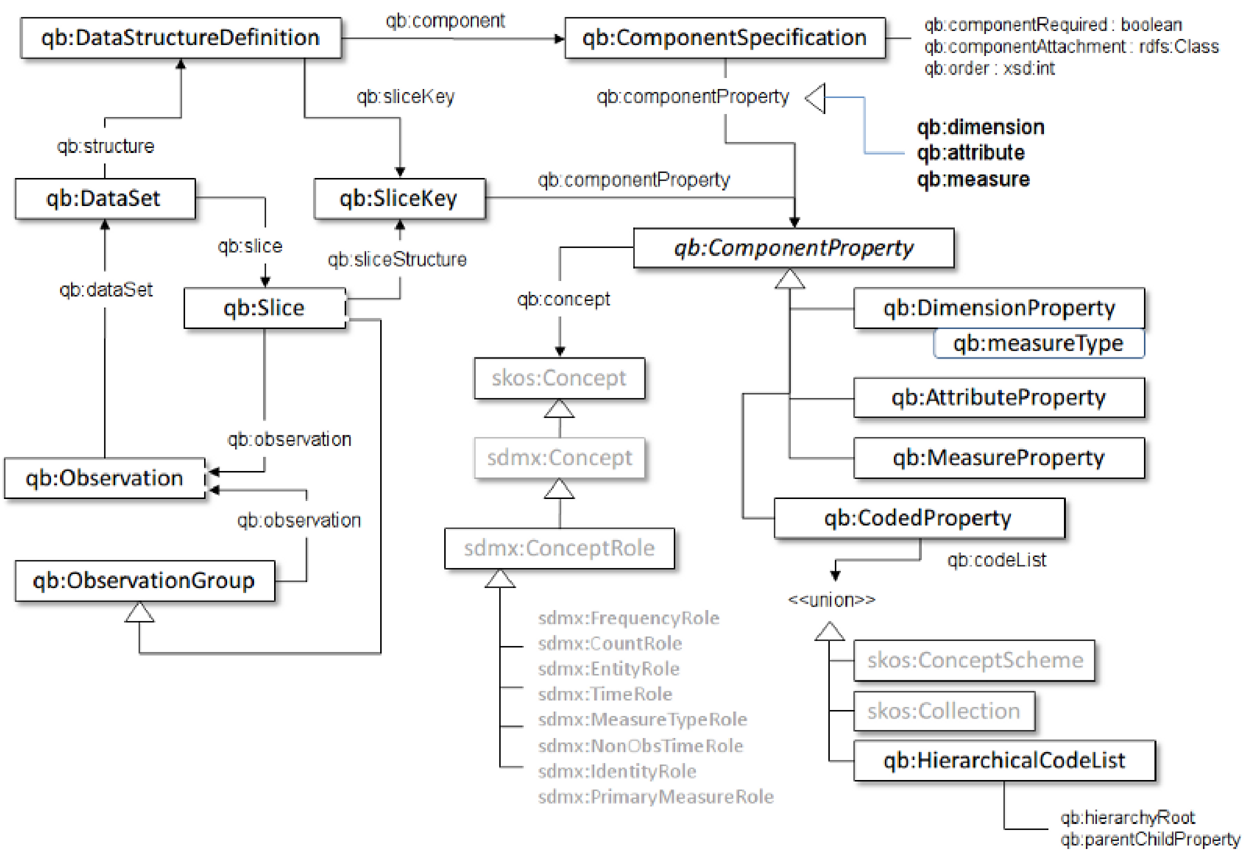


Figure 13 : Key terms and relationships in The RDF Data Cube Vocabulary

source: [21]

For every dataset (qb:DataSet) a definition of its structure (qb:DataStructureDefinition) needs to be developed. This structure is made of specifications of its components properties (qb:ComponentProperty). There are 3 types of components properties:

- **Measures** (qb:MeasureProperty) – measure properties specify the types of the observed values in the dataset.

- **Dimensions** (qb:DimensionProperty) – dimension properties specify dimensions used in the dataset to organize the observed values in a multidimensional space.
- **Attributes** (qb:AttributeProperty) – attribute properties specify additional attributes of the observed values, such as currency or accuracy.

Datasets using the DCV are made of observations (qb:Observation). An observation might be seen as a record of measures (one or more observed values) and the respective values of the specified dimensions and attributes. By selecting specific values of one or more dimensions, a view on the data called slice (qb:Slice) can be defined.

4.3.1.2 Observations, DataSet, and Data Structure Definition :

The RDF Data Cube Vocabulary builds upon an abstract cube model, i.e. a multidimensional space where measured values are indexed by multiple dimensions.

Using an excerpt of the total general government expenditure expressed as percentage of GDP published by Eurostat [22] as an example.

Reference area\year	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
EU (28 countries)	:	:	:	45,6	44,9	46,5	50,3	50	48,6	49	48,6	48,2
EU (27 countries)	:	:	:	45,6	44,9	46,5	50,3	50	48,6	49	48,6	48,2
Euro area (19 countries)	:	:	:	46	45,3	46,6	50,7	50,5	49,1	49,7	49,6	49,4
Euro area (18 countries)	:	:	:	46,1	45,3	46,6	50,7	50,5	49,1	49,8	49,6	49,4
Euro area (17 countries)	:	:	:	46,1	45,4	46,6	50,7	50,5	49,1	49,8	49,7	49,4

Total general government expenditure (% of GDP, excerpt)

source: [22]

The total general government expenditure expressed as percentage of GDP is the measured phenomenon, as illustrated in Table 1, which is indexed by two dimensions: **reference area** and **year**. The total government expenditure in EU28 in 2010 represents a single observation. The collection of observations forms a dataset, i.e. a data cube.

```
@prefix rdfs:            <http://www.w3.org/2000/01/rdf-schema#> .
@prefix qb:              <http://purl.org/linked-data/cube#> .
@prefix sdmx-attribute: <http://purl.org/linked-data/sdmx/2009/attribute#> .
@prefix xsd:             <http://www.w3.org/2001/XMLSchema#> .

@prefix ex-dimension:  <http://data.example.org/ontology/dsd/dimension/> .
@prefix ex-dsd:        <http://data.example.org/resource/dsd/> .
@prefix ex-measure:    <http://data.example.org/ontology/dsd/measure/> .

ex-dsd:total-general-government-expenditure a qb:DataStructureDefinition ;
  rdfs:label "Total general government expenditure"@en ;
  # Dimensions
  qb:component [
       qb:dimension ex-dimension:refPeriod ;
       qb:order 1 ;
       rdfs:label "Dimension representing a year for which the total general
government expenditure is reported"@en
  ] ;
  qb:component [
       qb:dimension ex-dimension:refArea ;
       qb:order 2 ;
       rdfs:label "Dimension representing a state or group of states for which the
total general government expenditure is reported"@en
  ] ;
  # Measure
  qb:component [
       qb:measure ex-measure:total-general-government-expenditure ;
```

Figure 14:data structure definition for the dataset described in the table above

4.3.2 Data Linking Module:

We use DBPedia as an external data repository to enrich our local dataset uploaded by the user.

By using special SPARQL queries we fetch following information:

- Area and population values for both location (usually a city) and the country where this city is located.
- Latitude and longitude values for the location
- GDP per capita value for the country

Each and every observation in the provided RDF dataset must be linked to a location (usually a city) so that that location can be used to retrieve the formerly mentioned information from DBPedia endpoint.

<HTTP://DATA.OPENBUDGETS.EU/RESOURCE/DATASET/EXAMPLE-BUDGET-CITIES/OBSERVATION/1A> A QB:OBSERVATION ;

<HTTP://DATA.OPENBUDGETS.EU/ONTOLOGY/DSD/DIMENSION/ORGANIZATION>

<HTTP://DBPEDIA.ORG/RESOURCE/PARIS>;

<HTTP://DATA.OPENBUDGETS.EU/ONTOLOGY/DSD/MEASURE/AMOUNT> 14881;

4.3.2.1 Enrichment stage steps:

1. We go through all locations we have to fetch the relevant country, geo-spatial, area and population information

```
+ "CONSTRUCT \n"
+ "{\n"
+ " ?location dbo:populationTotal ?pop.\n"
+ " ?location geo:lat ?lat.\n"
+ " ?location geo:long ?long.\n"
+ " ?location dbo:areaTotal ?area. \n"
+ " ?location dbo:country ?country ."
+ " ?country dbp:gdpPppPerCapita ?countryGDP ."
+ " ?country dbo:populationTotal ?countryPopulation ."
+ " ?country dbo:areaTotal ?countryArea ."
+ "\n"
+ "}  \n"
+ " where {\n"
+ " ?subject datacube:isLocation ?location .\n"
+ " \n"
+ " SERVICE <http://dbpedia.org/sparql> { \n"
+ " ?location dbo:populationTotal ?pop.\n"
+ " ?location geo:lat ?lat.\n"
+ " ?location geo:long ?long.\n"
+ " OPTIONAL {"
+ " ?location dbo:areaTotal ?area ."
+ " ?location dbo:country ?country ."
+ " ?country dbp:gdpPppPerCapita ?countryGDP ."
+ " ?country dbo:populationTotal ?countryPopulation ."
+ " ?country dbo:areaTotal ?countryArea ."
+ "}\n"
```

2. We go through all observations we have and according to its location we attach the country and geo-spatial information again to the observation

```
+ "CONSTRUCT \n"
+ "{\n"
+ " ?observation datacube:info_lat ?lat.\n"
+ " ?observation datacube:info_long ?long.\n"
+ " ?observation datacube:info_country ?country.\n"
+ "\n"
+ "}  \n"
+ " where {\n"
+ " ?observation a qb:Observation .\n"
+ " ?observation " + locationPredName + " ?location.\n"
+ " ?location geo:lat ?lat.\n"
+ " ?location geo:long ?long.\n"
+ " OPTIONAL {"
+ " ?location dbo:country ?country."
+ "} \n"
```

In our example the "locationPredName" is HTTP://DATA.OPENBUDGETS.EU/ONTOLOGY/DSD/DIMENSION/ORGANIZATION

3. Finally we again go through all observations and attach the GDP, area and population information for both country and location.

```
+ "CONSTRUCT \n"
+ "{\n"
+ "   ?observation datacube:info_countrypopulation \"low\".\n"
+ "}  \n"
+ "where {\n"
+ " ?observation a qb:Observation .\n"
+ " ?observation datacube:info_country ?country .\n"
+ " ?country dbo:populationTotal ?pop .\n"
+ "  filter(?pop > " + Ranges.countryPopLow + " &&  ?pop < " + Ranges.countryPopMid + ").\n"
```

Figure 15: Example SPARQL query on how to generate county population information

```
+ "CONSTRUCT \n"
+ "{\n"
+ "   ?observation datacube:info_area \"mid\".\n"
+ "}  \n"
+ "where {\n"
+ " ?observation a qb:Observation .\n"
+ " ?observation " + locationPredName + " ?location .\n"
+ " ?location dbo:areaTotal ?area .\n"
+ "  filter(?area > " + Ranges.locationAreadMid + " &&  ?area < " + Ranges.locationAreadHigh + ").\n"
```

Figure 16: Example SPARQL query on how to generate location area information

The following ranges are used for classification of area, population and GDP information

```
public static double locationAreadLow = 1000000d;
public static double locationAreadMid = 1962500000d;
public static double locationAreadHigh = 3925000000d;

public static double locationPopLow = 1000d;
public static double locationPopMid = 1000000d;
public static double locationPopHigh = 2000000d;

public static double countryAreadLow = 1000000d;
public static double countryAreadMid = 178584000000d;
public static double countryAreadHigh = 357168000000d;

public static double countryPopLow = 10000d;
public static double countryPopMid = 20000000d;
public static double countryPopHigh = 60000000d;

public static double countryGDPLow = 1000d;
public static double countryGDPMid = 32000d;
public static double countryGDPHigh = 45000d;
```

Figure 17: classification of area, population and GDP information

The information generated during this step will used as constraints for the subpopulation method.

By the end of these steps the dataset is ready to go through the subpopulation stage.

4.4 User Input and Interface:

The user is required to input different information to start a new session. During creating the session the system goes through enrichment and preprocessing (subpopulation) and then the user can select which bucket (bin) to find outliers for.

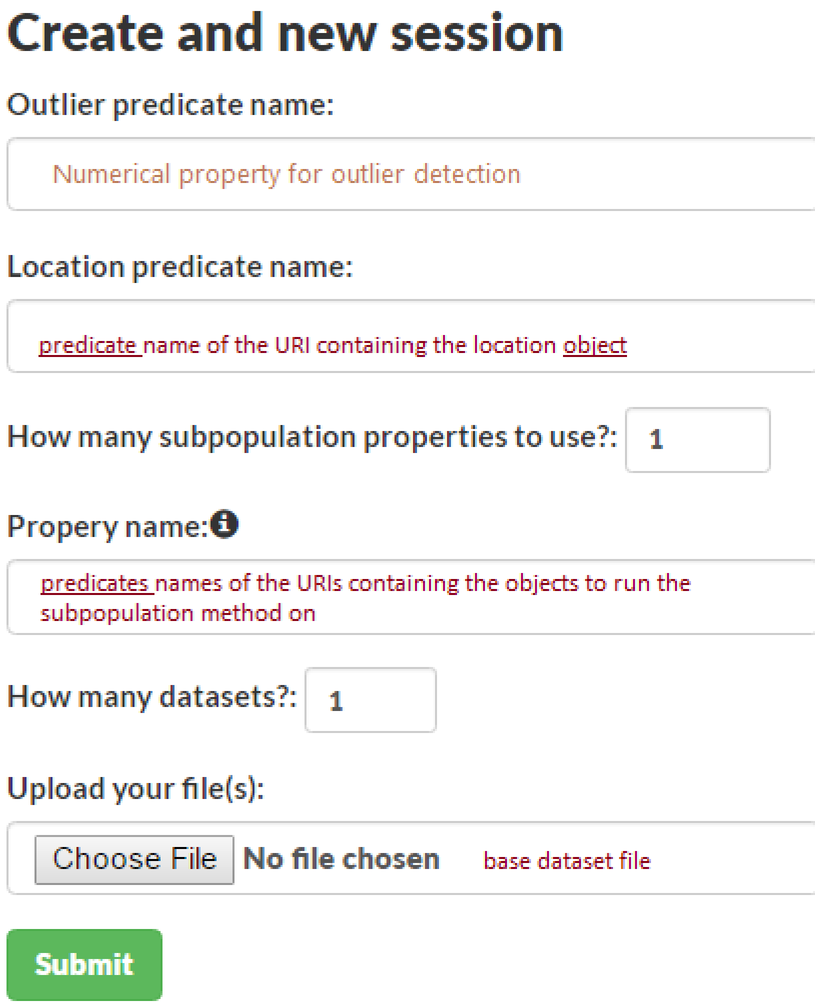


Figure 18: User Interface

After the session is created the user should keep the shown session ID to view the session again in the future without going through the above-mentioned stages.

4.5 Results in each level:

Figure 19: Subpopulation Results

By choosing certain level the user can view the different buckets that we generated for each level and choose one bucket to apply one of the two implemented outlier detection methods.

4.6 Property Name and constraint:

For the property that are used in subpopulation:

- info_population.
- info_countryarea
- budgetPhase

Depending on this property, the system analysis the data set automatically to find the constraint depending on these properties and its values.

For example:

- info_population=high , info_population= mid
- info_population=low if it is available in data set
- info_countryarea=mid…. Etc.
- budgetPhase=approved…. Etc.

4.7 Choose a bin(bucket) to apply outlier detection:

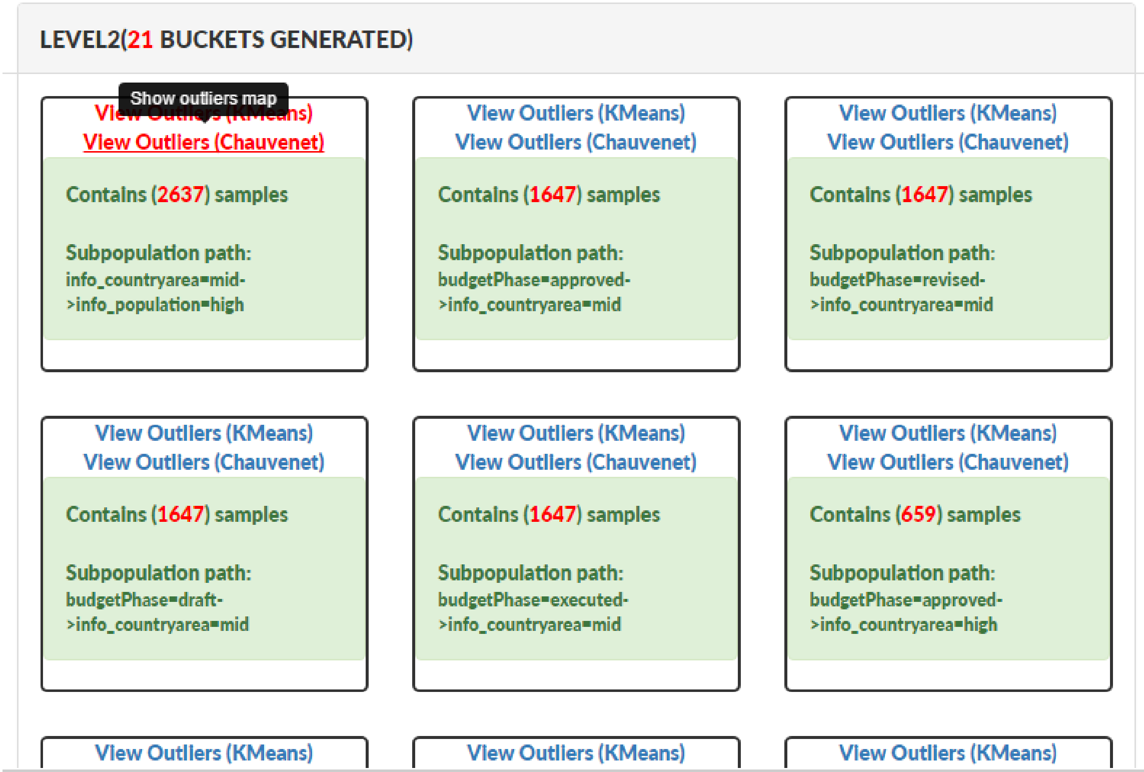


Figure 20 : Details of results sub population

Visualization of the results:

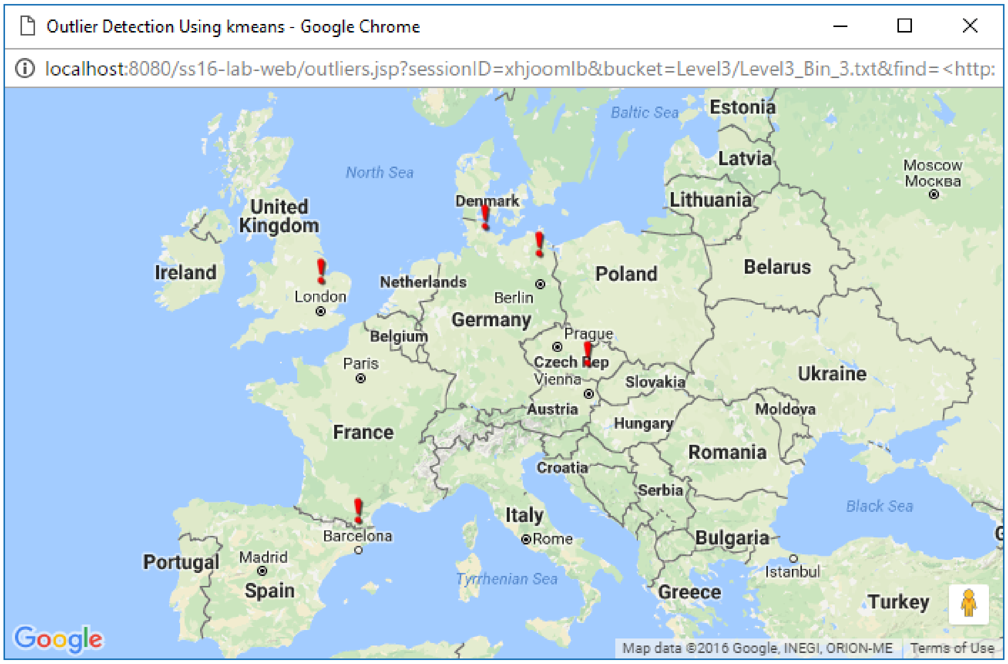


Figure 21: A Google Map shows outliers found in a number of cities for chosen bin

After the outlier results are found, a cache entry is created for that specific outlier results based on the applied method, session ID and bucket name so that these results can be lately retrieved directly from the cache in both web interface and restful interface.

4.7.1.1 Restful Service Interface:

The system also provides a web service that can be used to retrieve the results of a previously cached outlier method in a certain session for a certain bucket, therefore the user is required to provide three pieces of information using the PUT method, and then can use the GET method to retrieve the results in a JSON format.

```
HttpClient client = new DefaultHttpClient();
HttpPut put = new HttpPut("http://localhost:8080/ss16-lab-web/resources/outliers/session");
put.setEntity(new StringEntity("upenkwbq"));// session ID
client.execute(put);
put.releaseConnection();

put = new HttpPut("http://localhost:8080/ss16-lab-web/resources/outliers/bucket");
put.setEntity(new StringEntity("Level1/Level1_Bin_1.txt")); // bucket name
client.execute(put);
put.releaseConnection();

put = new HttpPut("http://localhost:8080/ss16-lab-web/resources/outliers/method");
put.setEntity(new StringEntity("chauvenet")); // method name
client.execute(put);
put.releaseConnection();

HttpGet get = new HttpGet("http://localhost:8080/ss16-lab-web/resources/outliers");
HttpResponse response = client.execute(get);
HttpEntity en = response.getEntity();
InputStreamReader i = new InputStreamReader(en.getContent());
```

Figure 22: Example illustrates how to build restful client

5 Comparison of Outlier Detection Methods:

Since different outlier detection algorithms are based on disjoints sets of assumption, a direct comparison between them is not always possible. In many cases, the data structure and the outlier generating mechanism on which the study is based dictate which method will outperform the others. There are few works that compare different classes of outlier detection methods. In particular, the methods depend on: whether or not the data set is multivariate normal, the dimension of the data set, the type of the outliers, the proportion of outliers in the dataset.

5.1 Comparing Example by our system:

The project can process the dataset dynamically and this point is one of the strongest points that support the project. Our data set that is used in testing the project and comparing the result is “Budget-EU.ttl”. The statistics in the tables below are generated automatically inside the project folder in the *statistics* folder and include all information that we used to analysis and compare the results.

There are three criteria for comparing the result:

1. Comparing number of outliers in each bin in each level for each method.
2. The degree of agreement.
3. Compare the states where outlier results from K-means are a subset of outlier's results from Chauvenet .additionally, states where outlier results from K- Chauvenet are a subset of outlier's results from K-means.

- **Comparing number of outliers in each bin in each level for each method.**

In these criteria, we will calculate the number of all observations that are specified by K-means method as an outlier and compare it with a number of all observations that are specified by Chauvenet method.

Number of the bins	Number of instances (samples) in each bin	Outliers found by K-means	Outliers found by Chauvenet	Degree of agreement
Level 1				
Bin 1.1	5273	15	26	0
Bin 1.2	5265	9	9	1
Bin1.3	6585	14	29	0
Bin1.4	2633	5	5	1
Bin1.5	1317	1	1	1
Bin1.6	2964	2	6	0
Bin1.7	2964	11	12	0
Bin1.8	2964	8	9	0
Bin1.9	2964	7	10	0
Level 2				

Bin2.1	2637	13	20	0
Bin2.2	1317	2	2	1
Bin2.3	1319	2	5	0
Bin2.4	1319	8	9	0
Bin2.5	1319	3	5	0.6
Bin2.6	3949	6	6	1
Bin2.7	1317	1	1	1
Bin2.8	1317	423	0	0
Bin2.9	1317	5	5	1
Bin2.10	1647	2	6	0
Bin2.11	1647	6	7	0
Bin2.12	1647	8	9	0
Bin2.13	1647	12	8	0
Bin2.14	659	190	0	0
Bin2.15	659	4	4	1
Bin2.16	659	199	0	0
Bin2.17	659	1	1	1
Bin2.18	330	88	0	0
Bin2.19	330	1	1	1
Bin2.20	330	105	0	0
Bin2.21	330	84	0	0
Level 3				
B3.1	660	2	4	0
B3.2	660	8	8	1
B3.3	660	3	4	0.75
B3.4	330	102	0	0
B3.5	330	106	0	0
B3.6	330	71	0	0
Sum	62254	1527	211	0
Results				

Outliers found by K-means	1488	2.3%	
Outliers found by Chauvenet	169	0.27%	
Shared outliers K-means and Chauvenet	43	0.067%	

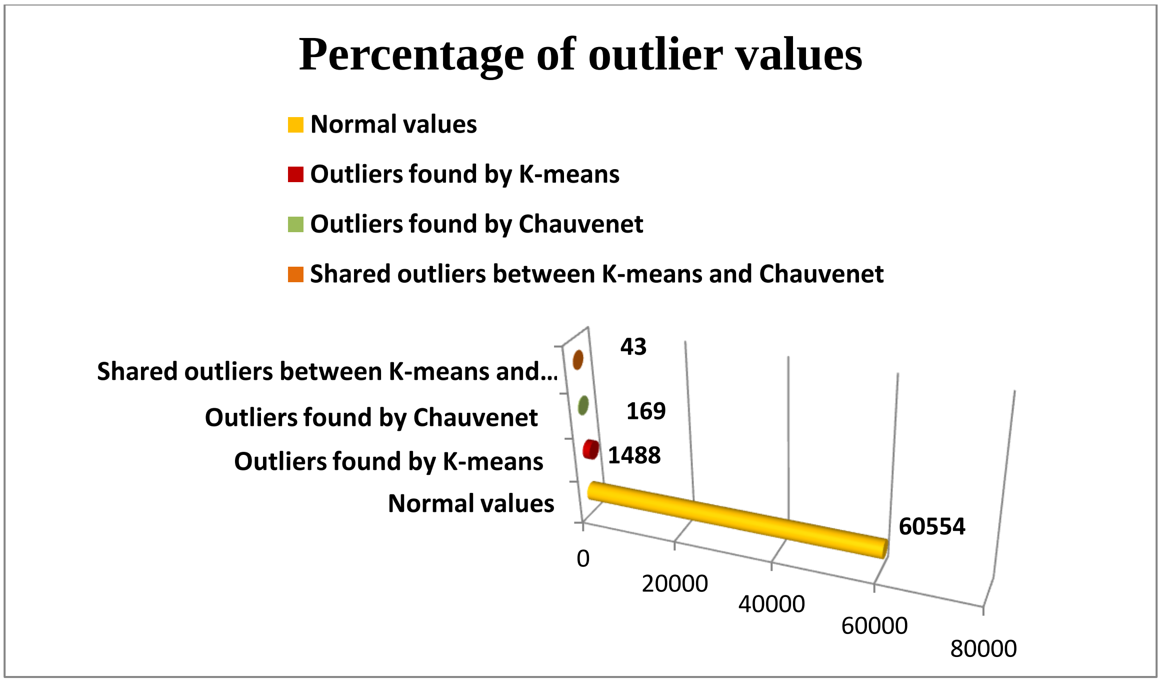


Figure 23: Percentage of outlier values

Using “Budget-Eu” data set as input to the project will give about 3% outliers for k-means method comparing with Chauvenet method that gave only 0.3 % outliers. Depending on this comparison we can conclude that k-means algorithm is more sensitive to outlier compared to Chauvenet method but that does not mean k-means is better than Chauvenet because each algorithm follows specific processes and special rules . As a result, we can conclude following points related to K-means:

- Sensitivity to outliers and noise
- Number of cluster and initial seed value need to be specified beforehand.

For instance:

1. We applied K-means for different values of k and the best results started to appear when k value > = 4.
2. For centroid value can be chosen depending on zero value , smallest value , mean value and highest value.

- **The degree of agreement.**

We followed a special scale in the measure of the degree of agreement between the results of outliers in two methods. The scale ranges between 0 and 1 display the different degree of agreement.

The degree of agreement is the number of the shared outliers between both methods divided by the maximum number of found outliers by each method. That being said, if the value of agreement is 1 then that’s considered as a <u>full-agreement</u> and it means that both methods yielded the same outliers (in number and values). However, if the value of agreement is 0, for example, then that’s considered as a full-disagreement indicating that both methods yielded no common outliers. Therefore, the degree of agreement can take any value in [0,1] reflecting how much the two methods “agreed” on the outputted outliers.

<u>For instance:</u>

Method A gave 4 outliers values: Value _A= 4.

Method B gave 4 outliers values: Value_B =4.

The number of shared outliers: Value_C

Maximum value when comparing value_A and value_B : Value_D

e.g. status 1:

If number of shared outliers between both methods is 4, that means there is a full matching (agreement) in the results

$$\frac{Value\_C}{Value\_D} = \frac{4}{4} = 1\ full\ agreement$$

e.g. status 2:

If number of shared outliers between both methods is 0, that means there is no matching in the results

$$\frac{Value\_C}{Value\_D} = \frac{0}{4} = 0\ full\ disagreement$$

e.g. status 3:

If number of shared outliers between both methods is 2, that means there is no matching in the results .

$$\frac{Value\_C}{Value\_D} = \frac{2}{4} = 0.5\ degree\ agreement$$

Number of the bins	Number of instances (samples) in each bin	Outliers found by K-means	Outliers found by Chauvenet	Degree of agreement	Formula
Level 1					
Bin 1.1	5273	15	26	0	A=15 B=26 C=0 D=26

Bin 1.2	5265	9	9	1	A=9 B=9 C=9 D=9
Bin1.3	6585	14	29	0	A=14 B=29 C=0 D=29
Bin1.4	2633	5	5	1	A=5 B=5 C=5 D=5
Bin1.5	1317	1	1	1	A=1 B=1 C=1 D=1
Bin1.6	2964	2	6	0	A=2 B=6 C=0 D=6
Bin1.7	2964	11	12	0	A=11 B=12 C=0 D=12
Bin1.8	2964	8	9	0	A=8 B=9 C=0 D=9
Bin1.9	2964	7	10	0	A=7 B=10

					C=0 D=10
Level 2					
Bin2.1	2637	13	20	0	A=13 B=20 C=0 D=20
Bin2.2	1317	2	2	1	A=2 B=1 C=1 D=1
Bin2.3	1319	2	5	0	A=2 B=5 C=0 D=5
Bin2.4	1319	8	9	0	A=8 B=9 C=0 D=9
Bin2.5	1319	3	5	0.6	A=3 B=5 C=3 D=5
Bin2.6	3949	6	6	1	A=6 B=6 C=6 D=6
Bin2.7	1317	1	1	1	A=1 B=1 C=1

					D=1
Bin2.8	1317	423	0	0	A=423 B=0 C=0 D=423
Bin2.9	1317	5	5	1	A=5 B=5 C=5 D=5
Bin2.10	1647	2	6	0	A=2 B=6 C=0 D=6
Bin2.11	1647	6	7	0	A=6 B=7 C=0 D=7
Bin2.12	1647	8	9	0	A=8 B=9 C=0 D=9
Bin2.13	1647	12	8	0	A=12 B=8 C=0 D=12
Bin2.14	659	190	0	0	A=190 B=0 C=0 D=190
Bin2.15	659	4	4	1	A=4 B=4

					C=4 D=4
Bin2.16	659	199	0	0	A=199 B=0 C=0 D=199
Bin2.17	659	1	1	1	A=1 B=1 C=1 D=1
Bin2.18	330	88	0	0	A=88 B=0 C=0 D=88
Bin2.19	330	1	1	1	A=1 B=1 C=1 D=1
Bin2.20	330	105	0	0	A=105 B=0 C=0 D=105
Bin2.21	330	84	0	0	A=84 B=0 C=0 D=84
Level 3					
B3.1	660	2	4	0	A=2 B=4 C=0

					D=4
B3.2	660	8	8	1	A=8 B=8 C=8 D=8
B3.3	660	3	4	0.75	A=3 B=4 C=3 D=4
B3.4	330	102	0	0	A=102 B=0 C=0 D=102
B3.5	330	106	0	0	A=106 B=0 C=0 D=106
B3.6	330	71	0	0	A=71 B=0 C=0 D=71
Sum	62254	1527	211	0	
Results		1527	211		

Agreement value = 1	11
Disagreement value = 0	23
Partial agreement	2

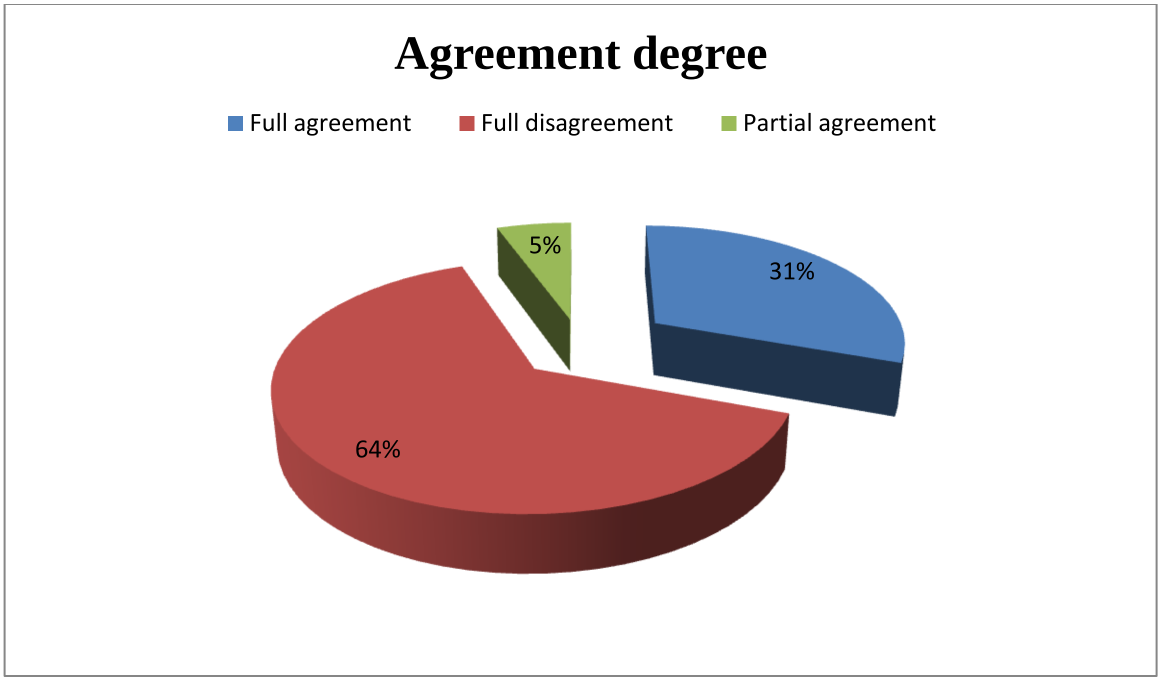


Figure 24:Agreement degree

We can conclude from the graph that the agreement between these two methods (k-means and Chauvenet) is low in the most cases and that reflects the differences in strategy of the work between these methods . However, there are still some cases where we had full-agreement and that is due to the high deviations of found outliers from the other “normal” values.

- Compare the states where outlier results from K-means are a subset of outlier’s results from Chauvenet .additionally, states where outlier results from K- Chauvenet are a subset of outlier’s results from K-means.

	Number of instances in each bin (samples)	Chauvenet subset of Kmeans	Kmeans subset of Chauvenet	
Level 1				
Bin 1.1	5273	false	false	0
Bin 1.2	5265	true	true	1
Bin1.3	6585	false	false	0
Bin1.4	2633	true	true	1
Bin1.5	1317	true	true	1
Bin1.6	2964	false	false	0
Bin1.7	2964	false	false	0
Bin1.8	2964	false	false	0
Bin1.9	2964	false	false	0
Level 2				
Bin2.1	2637	false	false	0
Bin2.2	1317	true	true	1
Bin2.3	1319	false	false	0
Bin2.4	1319	false	false	0
Bin2.5	1319	false	true	0.6
Bin2.6	3949	true	true	1
Bin2.7	1317	true	true	1
Bin2.8	1317	false	false	0
Bin2.9	1317	true	true	1
Bin2.10	1647	false	false	0
Bin2.11	1647	false	false	0
Bin2.12	1647	false	false	0
Bin2.13	1647	false	false	0
Bin2.14	659	false	false	0

Bin2.15	659	true	true	1
Bin2.16	659	false	false	0
Bin2.17	659	true	true	1
Bin2.18	330	false	false	0
Bin2.19	330	true	true	1
Bin2.20	330	false	false	0
Bin2.21	330	false	false	0
Level 3				
B3.1	660	false	false	0
B3.2	660	true	true	1
B3.3	660	false	true	0.75
B3.4	330	false	false	0
B3.5	330	false	false	0
B3.6	330	false	false	0
No intersection		23 cases		
Chauvenet subset of Kmeans		0 cases	Chauvenet subset of Kmeans= true Kmeans subset of Chauvenet=false	
Kmeans subset of Chauvenet		2 cases	Kmeans subset of Chauvenet = true Chauvenet subset of Kmeans =false	
Kmeans subset of Chauvenet **AND** **Chauvenet subset of Kmeans**		11 cases	Kmeans subset of Chauvenet =true Chauvenet subset of Kmeans = true	

- We had **11** cases (31% of the cases) where "**Chauvenet subset of Kmeans**"=true and "**Kmeans subset of Chauvenet**" = true. In other words, we had **11** cases of full-agreement between the two methods.
- We had **2 (0.05%)**cased where "**Kmeans subset of Chauvenet**" = true. In other words, all outliers found by Kmeans were also found by Chauvenet.

- We had **0** cased where "**Chauvenet subset of Kmeans**" = true. In other words, all outliers found by Chauvenet were also found by Kmeans.
- We had **23** cases (64% of the cases) of **no intersection** of full-disagreement.

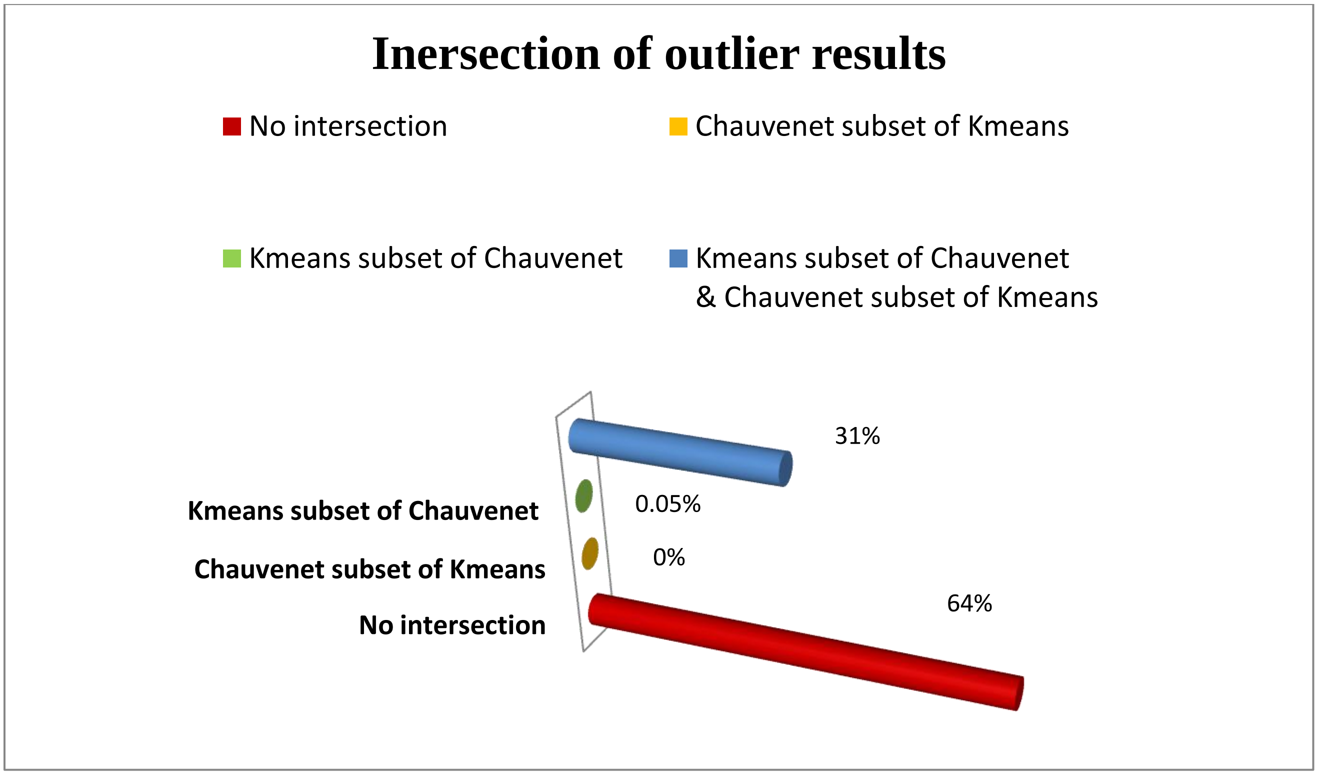


Figure 25:Inersection of outlier results

5.2 Clarification example depending on Rapid Minder:

We use in rapid miner the same samples of example that are used in the above evaluation.

5.2.1 Nearest Neighbor Based: Local Outlier Factor (LOF)[4]

The LOF anomaly detection calculates the anomaly score according to the local outlier factor algorithm proposed by Breunig[10]. There are several steps in the calculation of the LOF. The initial step involves getting the nearest neighbors set. The definition states that the k-distance(p) has at least k neighbors with distinct spatial coordinates that have a distance less

[4] http://www.dbs.ifi.lmu.de/Publikationen/Papers/LOF.pdf

than or equal it and at most k-1 of such neighbors with distance strictly less than it. The reachability distance (reach-dist(p,o)) is the maximum of the distance between point p and o and the k-distance(o). The local reachability is the inverse of the average reachability distance over the nearest neighborhood set. Finally the LOF is calculated as the average of the ratio of the local reachability density over the neighborhood set. The values of the LOF oscillates with the change in the size of the neighborhood. Thus a range is defined for the size of the neighborhood. The maximum LOF over that range is taken as the final LOF score.

Statistical Based: Histogram Based Outlier Score (HBOS)[5]:

Calculates an outlier score by creating an histogram with a fixed or a dynamic binwidth.
This method calculates a separate univariate histogram for every column in the Example Set. There are two modes, one with a static and one with a dynamic bandwidth[11]. In the static mode every bin has the same binwidth equally distributed over the value range. In the dynamic mode the binwidth can vary, but you can specify a minimum number of examples contained in a bin. The parameter number of bins sets the total number of bins used for either mode. The binwidth / minimum number values per bin is then calculated automatically. In the dynamic mode it is possible that there are less bins then specified if some bins contain more than the minimum number of values. To compute the outlier score, the histograms are normalized to one in height first. Then, the score is inverted, so that **<u>anomalies have a high score</u>** and normal examples a low score.

[5] http://www.dfki.de/KI2012/PosterDemoTrack/ki2012pd13.pdf

We choose some bins from all levels of the results and applied different outlier methods in Rapidminer as follow :

Example 1:

834	2452d	11504.0	
835	2585d	12808.0	
836	2452c	11400.0	
837	3212d	12954.0	
838	3345c	14415.0	
839	3212b	13418.0	
840	3345d	13464.0	
841	3212c	12966.0	
842	3345a	14412.0	
843	3345b	13651.0	
844	3212a	14058.0	
845	2147a	13233.0	
846	2013d	15473.0	
847	2146d	10255.0	
848	2013c	15375.0	
849	1915d	18044.0	
850	2013b	13202.0	
851	2146c	10585.0	Level1_Bin2
852	2146b	10903.0	
853	2013a	14606.0	
854	2146a	10974.0	
855	1915a	16949.0	
856	2279b	12095.0	
857	2279c	13573.0	
858	1915c	18414.0	
859	2279d	11908.0	
860	1915b	15690.0	

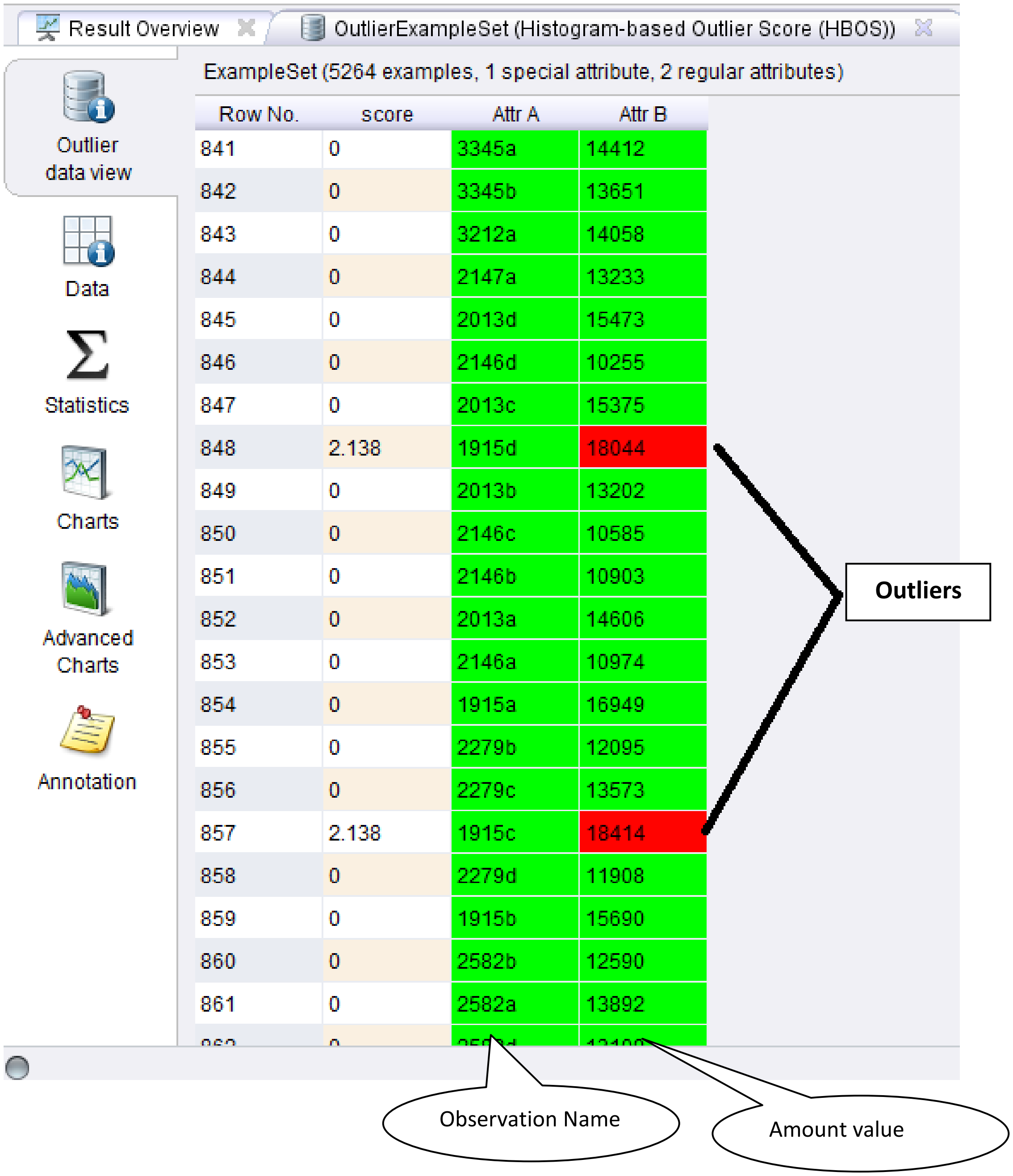
Observation Name

Amount value

Figure 26:Data sample

Outliers by Statistical Based: Histogram Based Outlier Score (HBOS):

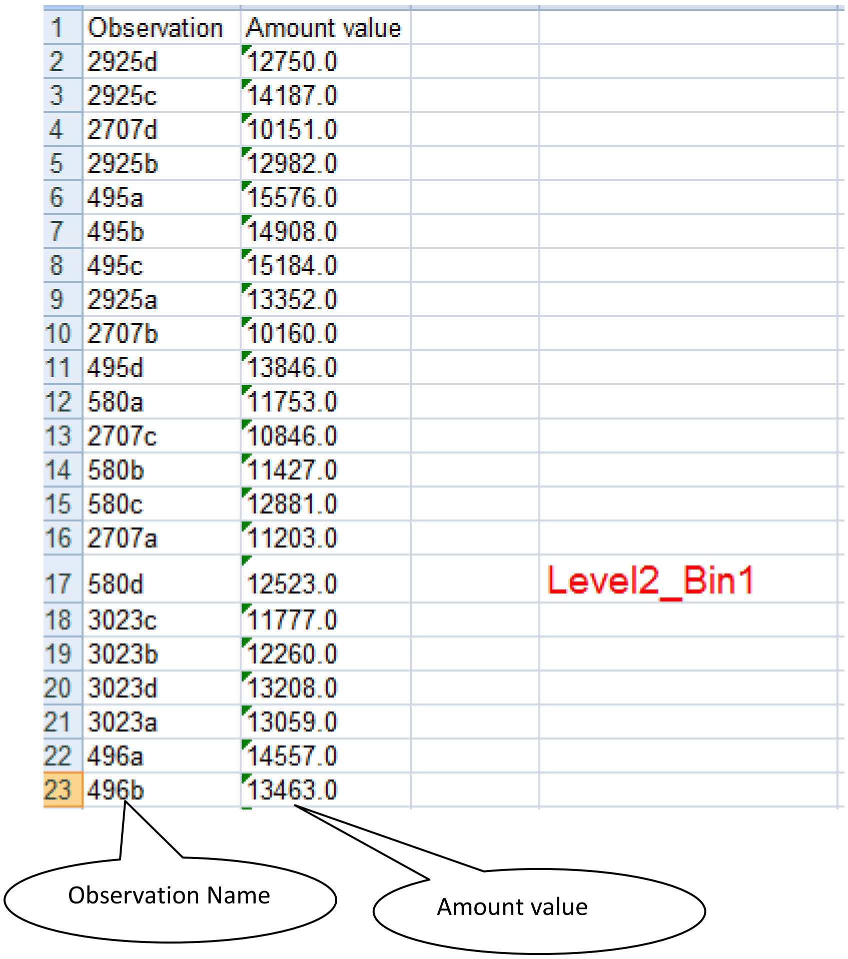
<u>**Criteria**: anomalies have a high score and normal examples a low score</u>



Row No.	score	Attr A	Attr B
841	0	3345a	14412
842	0	3345b	13651
843	0	3212a	14058
844	0	2147a	13233
845	0	2013d	15473
846	0	2146d	10255
847	0	2013c	15375
848	2.138	1915d	18044
849	0	2013b	13202
850	0	2146c	10585
851	0	2146b	10903
852	0	2013a	14606
853	0	2146a	10974
854	0	1915a	16949
855	0	2279b	12095
856	0	2279c	13573
857	2.138	1915c	18414
858	0	2279d	11908
859	0	1915b	15690
860	0	2582b	12590
861	0	2582a	13892

Figure 27: outlier in Rapidminer

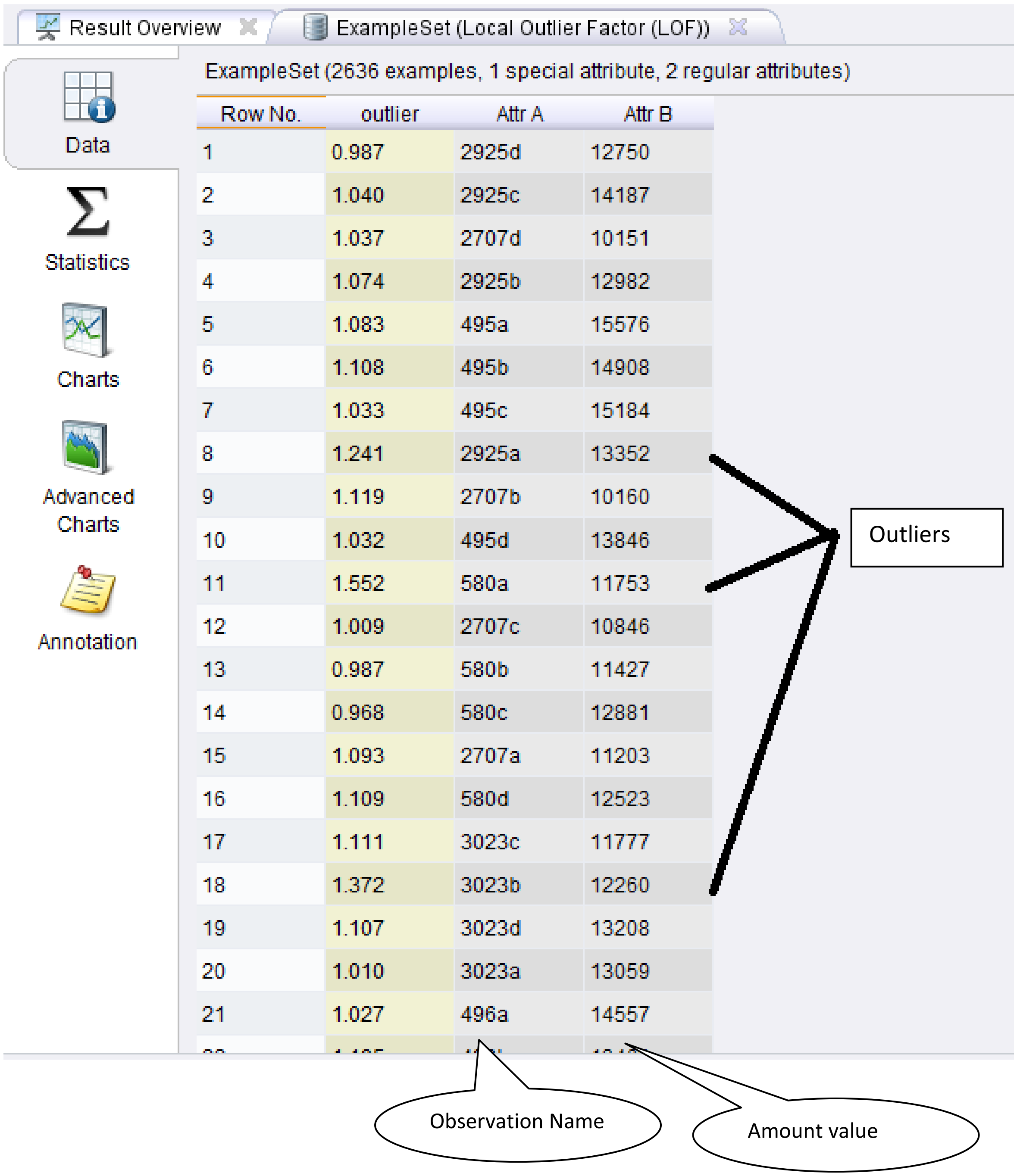
Example 2 :



1	Observation	Amount value		
2	2925d	12750.0		
3	2925c	14187.0		
4	2707d	10151.0		
5	2925b	12982.0		
6	495a	15576.0		
7	495b	14908.0		
8	495c	15184.0		
9	2925a	13352.0		
10	2707b	10160.0		
11	495d	13846.0		
12	580a	11753.0		
13	2707c	10846.0		
14	580b	11427.0		
15	580c	12881.0		
16	2707a	11203.0		
17	580d	12523.0		Level2_Bin1
18	3023c	11777.0		
19	3023b	12260.0		
20	3023d	13208.0		
21	3023a	13059.0		
22	496a	14557.0		
23	496b	13463.0		

Figure 28: Data Sample

Outliers by Nearest Neighbor Based: Local Outlier Factor (LOF):

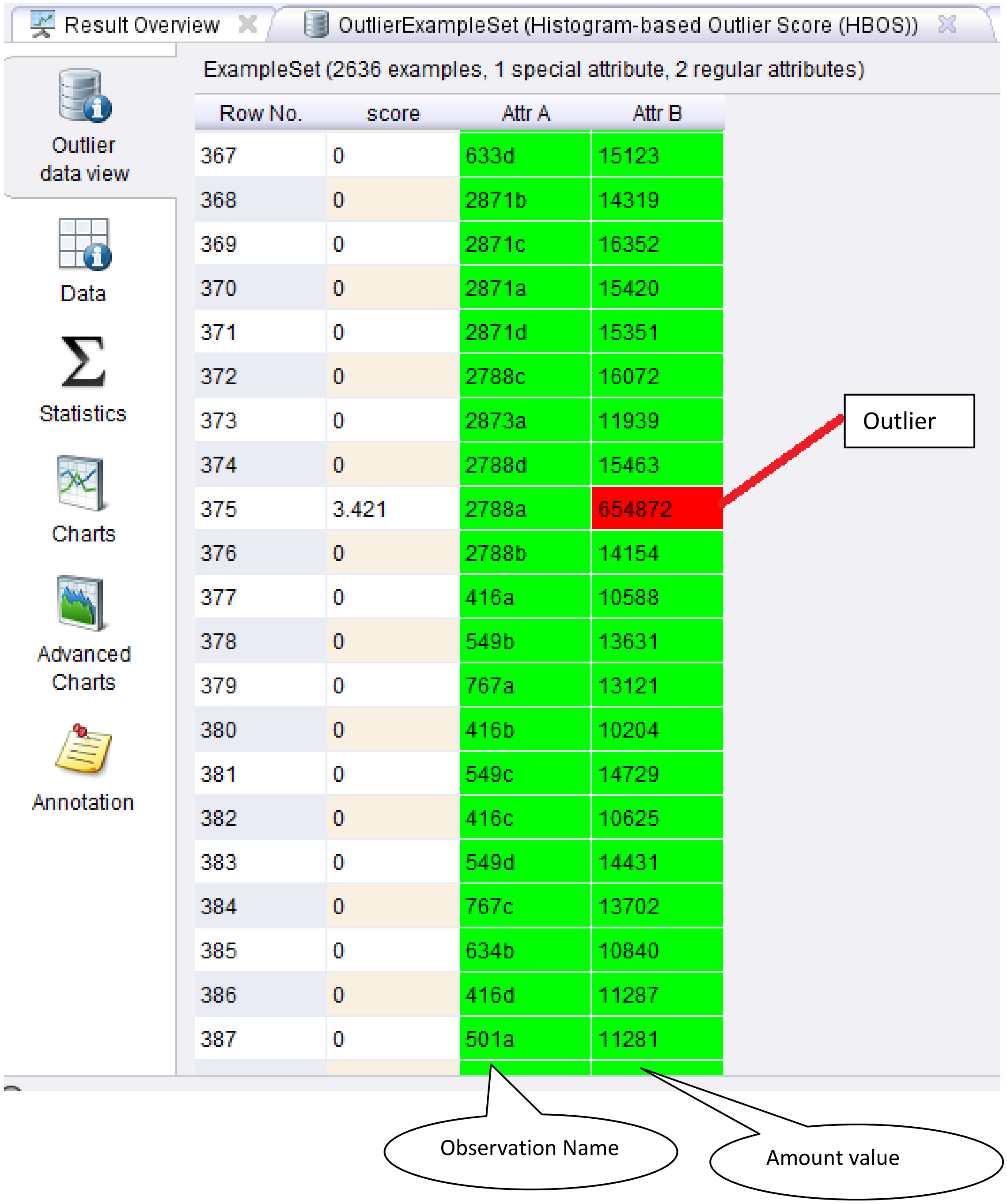


Row No.	outlier	Attr A	Attr B
1	0.987	2925d	12750
2	1.040	2925c	14187
3	1.037	2707d	10151
4	1.074	2925b	12982
5	1.083	495a	15576
6	1.108	495b	14908
7	1.033	495c	15184
8	1.241	2925a	13352
9	1.119	2707b	10160
10	1.032	495d	13846
11	1.552	580a	11753
12	1.009	2707c	10846
13	0.987	580b	11427
14	0.968	580c	12881
15	1.093	2707a	11203
16	1.109	580d	12523
17	1.111	3023c	11777
18	1.372	3023b	12260
19	1.107	3023d	13208
20	1.010	3023a	13059
21	1.027	496a	14557

Figure 29: result outlier in Rapidminer

Outliers by Statistical Based: Histogram Based Outlier Score (HBOS):

Criteria: anomalies have a high score and normal examples a low score



Row No.	score	Attr A	Attr B
367	0	633d	15123
368	0	2871b	14319
369	0	2871c	16352
370	0	2871a	15420
371	0	2871d	15351
372	0	2788c	16072
373	0	2873a	11939
374	0	2788d	15463
375	3.421	2788a	654872
376	0	2788b	14154
377	0	416a	10588
378	0	549b	13631
379	0	767a	13121
380	0	416b	10204
381	0	549c	14729
382	0	416c	10625
383	0	549d	14431
384	0	767c	13702
385	0	634b	10840
386	0	416d	11287
387	0	501a	11281

Figure 30:results outlier in Rapidminer

6 Conclusion:

Most real-world data sets contain outliers that have unusually large or small values when compared to others in the data set. Outliers may provide useful information about data when we look into an unusual response to a given study. Before starting the outlier detection, the dataset went through two important stages: enrichment and subpopulation.

During the enrichment stage , we linked our dataset to an external RDF data source called **DBPedia** to enrich it with more information that can be later used in subpopulation as constraints.

The subpopulation stage is a cruical part of our system and it must be done before starting the outlier detection stage, and that is to divide our dataset into smaller and more consistent datasets , called **buckets** or bins after the subpopulation is finised.

After we have our **buckets** from the subpoplation stage, we applied on each one of them two outlier detection methods, **Chauvenet and K-means**.

One of the main aims of our project is to compare the results (outliers) we got from the above-mentioned outlier detection methods. In our comparison we used statistical approach to calculate multiple indicators collected from outliers found in the whole buckets space we had. One of these indicators is the **degree of agreement** between Chauvenet and K-means, by which we found out that for the most of outlier detection cases(64%) applied on the different buckets, there was a full-disagreement compared to 31% of the cased where we have full-agreement.

Before applying the two outlier detection methods, we assumed that we would get very similar outlier results from the two methods but that was not the case.

As for future works, it would be an important addition to the system to make it support multiple vocabularies other than Data Cube vocabulary which is the only vocabulary considered by our system. The system can also be extened to support the linking to different external RDF data sources other than DBPedia.

7 References:

[1]. Heath, T. and Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space.Synthesis Lectures on the Semantic Web: Theory and Technology, 1(1), pp.1-136.

[2]. Techniques for Anomaly Detection in Network Flows. (2016). [online] Available at: http://cahsi.cs.utep.edu/cahsifiles/Files/PostersFinal/TechniquesAnomalyDetectionNetwork.pdf [Accessed 14 Oct. 2016].

[3]. Chandola, Varun and Banerjee, Arindam and Kumar, Vipin: "Anomaly detection: A survey", ACM computing surveys (CSUR), 2009.

[4]. Detecting Errors in Numerical Linked Data using Cross-Checked Outlier Detection. (2016). [online] Mannheim, Germany: Daniel Fleischhacker, Heiko Paulheim, Volha Bryl, Johanna Volker?, and Christian Bizer. Available at: http://www.heikopaulheim.com/docs/iswc_2014.pdf [Accessed 14 Oct. 2016].

[5]. Yan, M. (2005). Methods of Determining the Number of Clusters in a Data Set and a New Clustering Criterion. Keying Ye, Chair Samantha Bates Prins Eric P. Smith Dan Spitzner. Faculty of the Virginia Polytechnic Institute and State University.

[6].Selection of K in K-means clustering. (2016). [online] Cardiff University, Cardiff, UK. Available at: https://www.ee.columbia.edu/~dpwe/papers/PhamDN05-kmeans.pdf [Accessed 14 Oct. 2016].

[7].Clustering and information visualization. (2016). [online] University of Helsinki. Available at: https://www.cs.helsinki.fi/bioinformatiikka/mbi/courses/06-07/itb/slides/clustering.pdf [Accessed 14 Oct. 2016].

[8]. Heath, T. and Bizer, C. (2011) 'Linked data: Evolving the web into a global data space', Linked Data Evolving the Web into a Global Data Space, 1(1), pp. 1–136, page 7.

[9].Melo, A., Theobald, M., Volker, J.: Correlation-based refinement of rules with numerical attributes. In: Proc. of the 27th International Florida Artificial Intelligence Research Society Conference (2014).

[10]. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Comput.Surv. (2009).

[11]. Euzenat, J., Shvaiko, P.: Ontology Matching, Second Edition. Springer (2013).

[12]. Hautamaki,V.Cherednichenko, S.,Karkkainen, I.,Kinnunen, T.,and Franti, P.2005. Improving K-Means by Outlier Removal. In: SCIA 2005, pp.978-987.

[13]. Murugavel, P., and Punithavalli, M. 2011. Improved Hybrid Clustering and Distance-based Technique for Outlier Removal, International Journal on Computer Science and Engineering (IJCSE).

[14]. J. R. Taylor, "An Introduction to Error Analysis," 1st Ed., University Science Books, CA, 1982.

[15]. Breunig, M., Kriegel, H., Ng, R. and Sander, J. (2000). LOF. ACM SIGMOD Record, 29(2), pp.93-104.

[16]. Mennatallah Amer and Markus Goldstein. Nearest-neighbor and clustering based anomaly detection algorithms for rapidminer. In Proc. of the 3rd RCOMM 2012.

[17]. Yan, M. (2005). Methods of Determining the Number of Clusters in a Data Set and a New Clustering Criterion. Keying Ye, Chair Samantha Bates Prins Eric P. Smith Dan Spitzner. Faculty of the Virginia Polytechnic Institute and State University.

[18].Allen R., Tommasi D. (eds.) (2001): Managing public expenditure: a reference book for transition countries. http://www1.worldbank.org/publicsector/pe/oecdpemhandbook.pdf .

[19].Berners-Lee, T. (2006): Linked Data - Design Issues, http://www.w3.org/DesignIssues/LinkedData.html .

[20].Brickley D., Guha R. (2014): RDF Schema 1.1, http://www.w3.org/TR/rdf-schema/ .

[21].Cyganiak R, Reynolds D. (2014): The RDF Data Cube Vocabulary, http://www.w3.org/TR/vocab-data-cube/

[22].Eurostat (2015a): Gross domestic product at market prices, http://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=tec00001&plugin=1 .

[23].Eurostat (2015b): Total general government expenditure, http://ec.europa.eu/eurostat/tgm/table.do?tab=table&plugin=1&language=en&pcode=tec00023.

[24].Eurostat (2015c): Real GDP growth rate – volume, http://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=tec00115&plugin=1

[25].Selection of K in K-means clustering. (2016). [online] Cardiff University, Cardiff, UK. Available at: https://www.ee.columbia.edu/~dpwe/papers/PhamDN05-kmeans.pdf [Accessed 14 Oct. 2016].

[26]. Clustering and information visualization. (2016). [online] University of Helsinki. Available at: https://www.cs.helsinki.fi/bioinformatiikka/mbi/courses/06-07/itb/slides/clustering.pdf [Accessed 14 Oct. 2016].