Distributed Systems seminar, 2014
Clustering 1-dimensional continuous data

Elis Kõivumägi

Abstract

How to derive meaningful information from continuous data is the central topic of this work. When age data about 10 000 or 100 000 people is given, how can we represent the data in a way that is understandable to the analyst and also reveals significant meta-information about the people included in the datasets? In this work I am reviewing different clustering algorithms and proposing the most suitable for inducing previously unknown information.

Introduction

Grouping continuous data such as age, data usage and ARPU has been usually done by some person having knowledge about the industry or domain the data is being used in. For example, when describing the age of the population of a country, age is being grouped by five-year age groups [1] – 0-4, 5-9, 10-14, 15-19 etc. The same data in marketing is usually grouped into another set of intervals, for example, the penetration of smartphones in relation to age [2] – 12-17, 18-24 etc. In Demograft [3] applications, age is also grouped arbitrarily.

If the purpose of an application is more than just descriptive statistics, then this approach is not the best as it does not help the end-user to really understand the data. To tackle this shortcoming, data mining can help. The purpose of data mining in general is to detect the unknown in the data. As there are many different clustering algorithms to choose from, the question that needs to be answered is which one of them has the greatest promise – which one can automatically detect clusters in the one dimensional continuous data.

There are also two limitations that have to be considered when choosing the correct algorithm. The first is related to speed – algorithm has to be fast enough to provide clusters nearly real-time. The delay may not exceed 20 seconds for 50 000 data points and the range of values may vary – it may be 100 for age and up to 100 000 for income.

During my work I reviewed 3 different clustering algorithms and I will present the results in the following sections.

K-Means clustering

K-means clustering algorithm is arguably the most popular clustering algorithm out there. It has been used in lots of different application domains. The purpose of the algorithm is to cluster different observations (in this case, the age values) into k different clusters [4].

The idea behind k-means is to find k centroids as far away from each other as possible. The next step is to associate each data point in the dataset with its nearest centroid. After the first loop has been completed, new barycenters have to be chosen from each cluster and repeat the nearest centroid detection for each data point. Repeat this process until no changes will occur during the assignment process.

The most troubling aspect of k-means clustering is that one needs to know how many clusters there will be, which basically means you have to know your data very well. The other problem about k-means is that it is usually used for clustering 2 dimensional data.

This means that the algorithm is not suitable for solving our problem in this stage, but it shows promise during the future work – when several data attributes (for example age and income) have been clustered using an algorithm which can do the clustering on one dimension only, k-means clustering can be applied to further cluster the data, but now using 2 dimensions. More about this in the future work section.

Hierarchical clustering

Another possibility was to use hierarchical clustering [5] which has the objective to cluster observations into hierarchy where different hierarchy levels will have different number of clusters and it would be possible to choose the most suitable number of clusters that suits the purpose of the analysis.

How hierarchical clustering works by dividing dataset into N clusters and then uses the distance function to find similarities between clusters. The two most similar pairs of clusters will be merged together. The next step includes calculating distances for new and old clusters. These two steps are repeated until there is one cluster containing all the data points. One known issue with hierarchical clustering is it speed. It is very slow when large datasets need to be clustered.

Another limitation is that a pre-determined distance function needs to be in place in order to use this clustering technique.

In addition, with continuous data, it becomes clear that the cluster will be obvious and equal in ranges which basically yields the same results as making the groups beforehand. Hierarchical clustering could be useful when clustering cities by their distance differences [6].

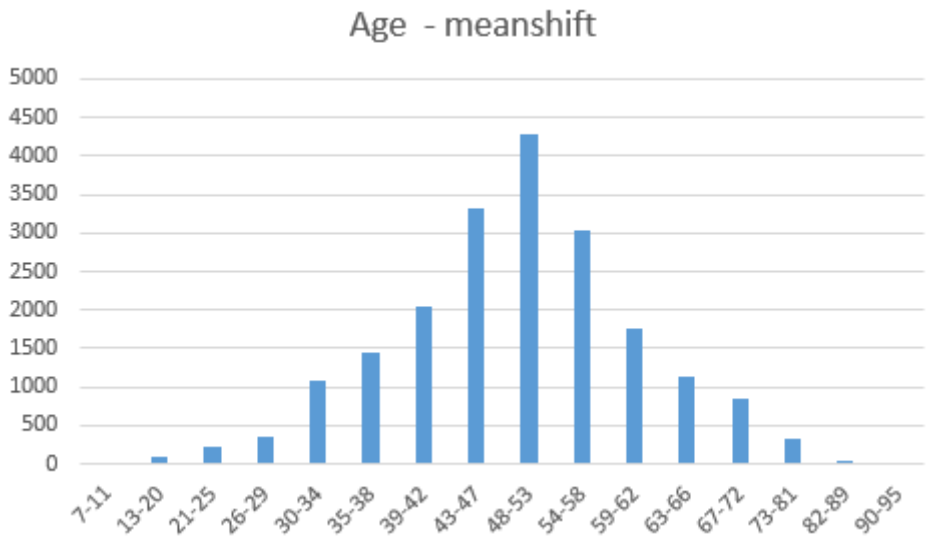

Figure 1: Applying mean-shift clustering on age data yielded clusters that offer more insight into the data than arbitrary grouping.

Mean-shift clustering

The mean-shift clustering technique is about clustering dense data areas together without prior knowledge of the number of the clusters nor is the shape of the clusters restricted [7][8]. Dense regions in the data correspond to the modes of the underlying distribution.

Like hierarchical clustering the mean-shift clustering also suffers from performance issues. Which means that there's a limit for audience size that can be clustered in real-time. The Python SkLearn implementation worked fast enough for audiences with up to 20 000 data points, but beyond that, the algorithm became very slow.

Other than speed, there are no major limitations that would render mean-shift clustering not usable for the problem at hand. Therefore, this algorithm was chosen to cluster continuous data and estimate the results – whether or not this approach is suitable.

Test setup and results

To conduct a test a randomly generated data of 20 000 points was generated.

Two different attributes were created: Age – the age of the subscribers, values range from 7 – 95 Data usage – the data usage in MB per subscriber, values range from 0 – 40 000

The chosen values are the most common among mobile operators for classifying their subscribers.

I will run the mean-shift clustering algorithm on both datasets and compare the results against regular classification used by mobile operators and estimate the difference. The results of the clustering are depicted in the figures 1-4.

Conclusion

For age attribute, the mean-shift algorithm yields great results as it offers more insight into the data when compared to arbitrary clusters. In addition to clusters which are range-wise more or less equal but not exactly the same as by-five division referenced earlier.

While analyzing data usage mean-shift clustering, it can be seen instantly that results aren't as good as age attribute's. The arbitrary groups give much better understanding of the data and mean-shift groups actually show only that lot of people use up to 3GB of data.

What I expected was a much more uneven creation of clusters (range-wise), but it seems that the implementation that was used always tries to keep the range of clusters similar. Therefore it is impossible to learn anything valuable from the data usage as the most important part of data is included into one cluster and reveals nothing.

This means that mean-shift algorithm work well for normal and uniform distribution, which is not a satisfying result as this cannot be guaranteed while running different queries in Demograft application.

Future work

There are two main areas of research for the upcoming seminar in the next semester.

First, I have to find a better clustering solution which would handle different kinds of distributions. It may appear that for different types of distributions, a different type of algorithm must be used. Or use a hybrid solution – run mean-shift clustering several times while grouping non-satisfactory clusters together.

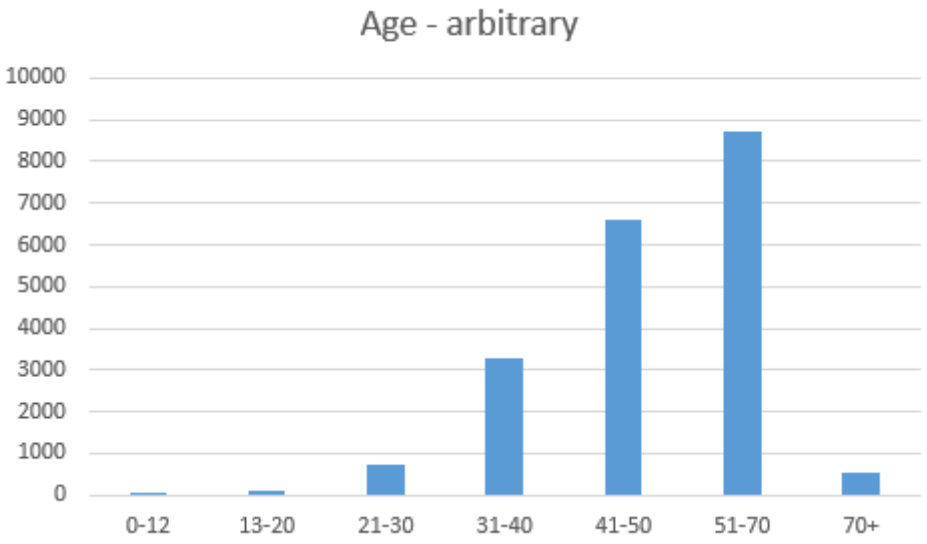

Figure 2: Arbitrary grouping of the same age data.

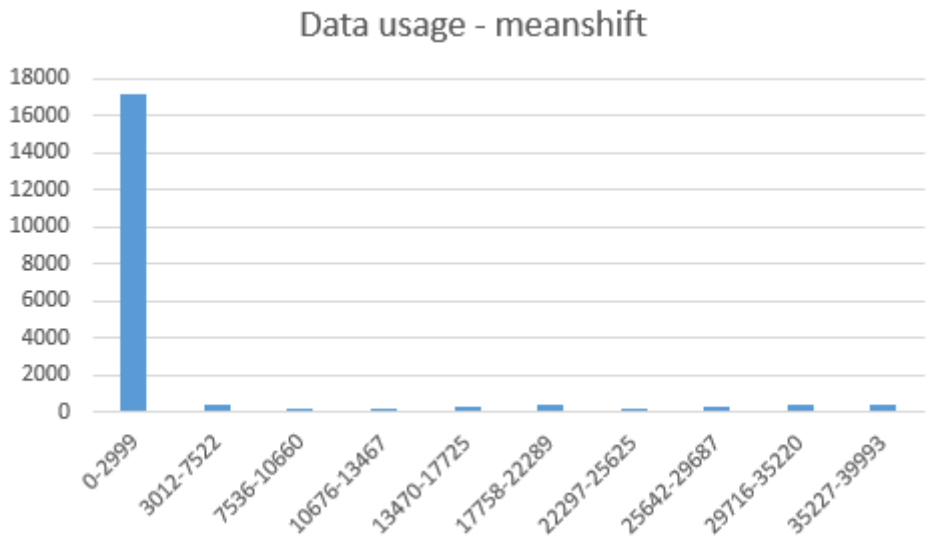

Figure 3: Data usage clustering with mean-shift algorithm does not reveal anything interesting about the data. In fact, it makes matters worse.

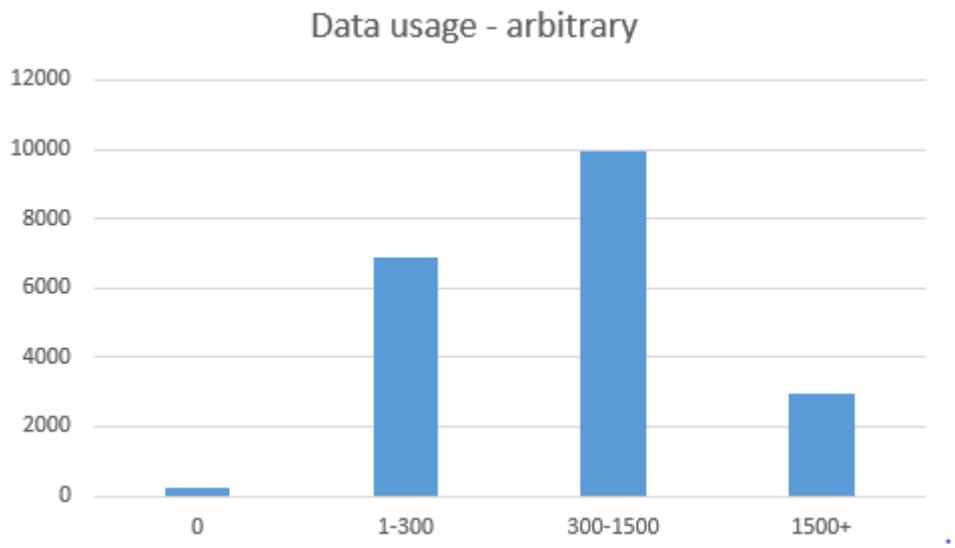

Figure 4: Arbitrary grouping gives much better understanding how data usage by volume is divided between subscribers.

Another possibility is to try coming up with a good distance function and use hierarchical clustering, but this needs a lot of testing.

Secondly, I intended to run regression analysis on the clustered data, but as the results are partly unsatisfactory, this will be delayed until the clustering issues are solved. Using regression also had a prerequisite of having interest data for each subscriber that I was unable to get my hands on. I will be able to get the data soon, but still, the clustering has to be solved beforehand.

References

[1]http://www.statcan.gc.ca/concepts/definitions/class-age1-eng.htm

[2]http://heidicohen.com/wp-content/uploads/Edison_Research_Arbitron-Smartphone-Ownership-by-Age.png

[3]https://demo.demograft.com/emirates/targeter

[4]http://en.wikipedia.org/wiki/K-means_clustering

[5]http://en.wikipedia.org/wiki/Hierarchical_clustering

[6]http://www.analytictech.com/networks/hiclus.htm

[7]http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/TUZEL1/MeanShift.pdf

[8]http://www.cse.yorku.ca/~kosta/CompVis_Notes/mean_shift.pdf