Cartification: Turning Similarities into Itemset Frequencies*

Bart Goethals

University of Antwerp, Belgium
bart.goethals@ua.ac.be

Extended Abstract

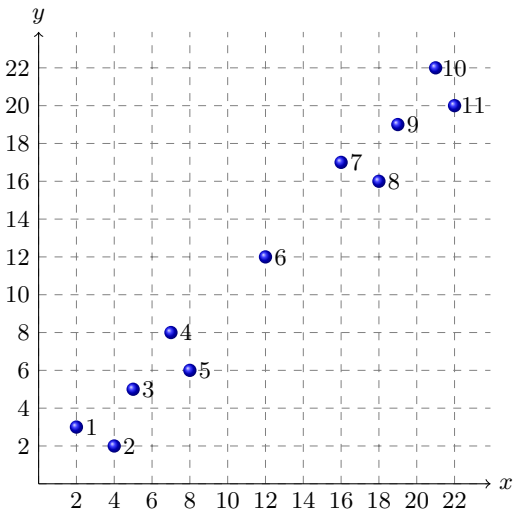
Suppose we are given a multi-dimensional dataset. For every point in the dataset, we create a transaction, or cart, in which we store the k-nearest neighbors of that point for one of the given dimensions. This is repeated for every dimension. The resulting collection of carts can then be used to mine frequent itemsets; that is, sets of points, or clusters, that are frequently seen together in one or more of the dimensions. Essentially, this transformation, that we call *cartification*, combines multiple distance measures without suffering from the curse of dimensionality.

An important observation to make in order to see the potential of cartified data is, that when the frequency of a single item is high, we know it is often found in the neighborhoods of many points in one or more of the dimensions; in fact, we can say that this item lies central in a cluster of data points, and, if it is most central, its frequency will be among the highest of all items in that cluster. Moreover, if an item is indeed part of a cluster, it is easy to see that it will mainly receive its support from those transactions in the database that correspond to the *relevant* dimensions of the cluster, as for the other dimensions it will have wildly varying neighborhoods. This is very important, as it allows us to identify which dimensions are relevant for the cluster, as well as to circumvent the dreaded curse of dimensionality. This observation also goes for sets of items: those itemsets that have relatively high frequency lie centrally in a sub-structure of the data, and will mainly receive their support from the dimensions relevant to that sub-structure. In fact, the latter effect will be even more pronounced for (large) sets than for single items, as it is increasingly unlikely that all points in the itemset lie closely together in an irrelevant dimension.

For example, assume we are given the two-dimensional dataset as shown in Figure 1. For cartification, we first need to choose a threshold k, the number of the nearest neighbors that can be added to each cart. Table 1b shows the resulting database for $k = 3$. The first two columns show the cartification for the x-dimension, and the second two columns for the y-dimension. Every row corresponds to the cart generated from one of the data points. For example, for point 3, the three nearest neighbors in the x-dimension are points $2, 3$, and 4, while for the y-dimension these are $1, 3$, and 5.

In Figure 2a, we plot the frequencies of all items (points) in this cartified database. This plot shows there are two points, 3 and 9, with a high frequency.

* work in collaboration with Jilles Vreeken

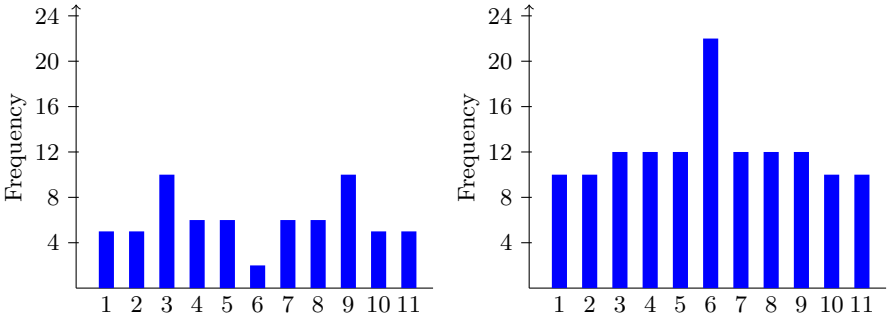


(a) original

x	**cart**	y	**cart**
1	$\{1,2,3\}$	1	$\{1,2,3\}$
2	$\{1,2,3\}$	2	$\{1,2,3\}$
3	$\{2,3,4\}$	3	$\{1,3,5\}$
4	$\{3,4,5\}$	4	$\{3,4,5\}$
5	$\{3,4,5\}$	5	$\{3,4,5\}$
6	$\{5,6,7\}$	6	$\{4,6,8\}$
7	$\{7,8,9\}$	7	$\{7,8,9\}$
8	$\{7,8,9\}$	8	$\{7,8,9\}$
9	$\{8,9,10\}$	9	$\{7,9,11\}$
10	$\{9,10,11\}$	10	$\{9,10,11\}$
11	$\{9,10,11\}$	11	$\{9,10,11\}$

(b) cartified

Fig. 1. Example dataset



(a) k=3

(b) k=6

Fig. 2. Item Frequencies

As Figure 1 shows, these points are in the centers of the clusters $\{1, 2, 3, 4, 5\}$ and $\{7, 8, 9, 10, 11\}$ respectively. Additionally, point 6 has a very low frequency which corresponds to the point being an outlier w.r.t. all other points in the figure.

Obviously, the choice of k is important with regard to which clusters, centers or outliers can be identified; if only, as for k only itemsets of length k or smaller can receive support. Essentially, k affects the granularity at which we are considering centers and outliers. For example, if we look back at the data in Figure 1, as cartified in Table 1b using $k = 3$, we already identified items 3 and 9 as cluster centers, and item 6 as an outlier. If we choose $k = 6$ instead, the resulting frequencies per item are shown in Figure 2b.

Now item 6 has the maximum frequency of 22, far more than all other items, and represents the center of the cluster consisting of all items. At this granularity, no outspoken outliers can be identified, yet the most remote elements (i.e., $1, 2, 10$, and 11) are still identifiable as they have the least support. Hence, increasing k is like zooming out on the data, and so taking a more global point of view.

Let us illustrate this in another example, where we show that not only the centroid items, but also the clusters themselves can be clearly identified. When keeping a relatively close zoom, with $k = 3$, for instance, the largest itemsets with a frequency greater than 2, are $\{1, 2, 3\}, \{3, 4, 5\}, \{7, 8, 9\}$, and $\{9, 10, 11\}$. These sets correspond to the clusters in the data when we aim at finding small clusters. Next, we zoom out to $k = 5$, and now find $\{1, 2, 3, 4, 5\}$ and $\{7, 8, 9, 10, 11\}$ as the largest itemsets with support greater than 2. These sets represent the clusters in the data well. Note however, these are good clusters *at the current level of zoom*. Indeed, if we zoom in 'too much', every point is regarded as a cluster, and if we zoom out 'too much' all of the data automatically becomes one big cluster. Clearly, what is 'enough' for the task at hand is subjective, and as clustering is explorative in nature, it is ultimately up to the data analyst to decide what types of clusters are potentially of interest, and hence, how to set k.

To conclude, preliminary cartification experiments on real data show that it allows us to efficiently discover centroid and outlying *sets of points*, subspace clusters, and clusterings, while not suffering from the curse of dimensionality. Multiple dimensions, numerical or categorical, as well as multiple distance measures, can be combined, and become represented in the frequency of clusters. Moreover, several efficient and scalable itemset mining techniques can be effectively applied on the cartified database, resulting in meaningful and interesting discoveries.