Exercise 1

Data Mining Algorithms 1 - WS 2015/16

Davide Pedranz (362504)

October 30, 2015

Exercise 1.3

We denote the set of data already present in the Data Warehouse with D_w and the new entries with D_{new}. We have $D = D_w \cup D_{new}$, where D is the set of all data after the entries are added.

(a) Mean

The *mean* is defined as follows:

$$mean(D) = \frac{sum(D)}{count(D)}$$

The functions *sum* and *count* are distributive, so we have:

$$mean(D) = mean(D_w \cup D_{new}) = \frac{sum(D_w \cup D_{new})}{count(D_w \cup D_{new})} = \frac{sum(D_w) + sum(D_{new})}{count(D_w) + count(D_{new})}$$

It is sufficient to store the values $sum(D_w)$ and $count(D_w)$ from the last aggregation in order to accelerate the computation of the new one.

Since the *mean* can be computed by applying the distributive functions *sum* and *count* a limited number of times (twice in this case), we conclude that *mean* is an algebraic measure.

(b) 25% quantile of the values

The 25% quantile is defined as a value that partition the dataset into 2 subsets, which contains respectively the 25% and 75% of the values. To efficiently determine this value, we have to order the values. The computation of the next aggregation can be accelerated storing the data in order, that decreses the complexity of the ordering phase.

The storage size needed for the computation is linear to the data's dimension, so the 25% quantile is a holistic measure.

(c) Variance

The *variance* is defined as follows:

$$\delta^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Since adding new data to the Data Warehouse affects the mean, we have to recompute $(x_i - \overline{x})$ for each value in the dataset. Thus, the variance is a holistic measure.

If an approximation of the *variance* is sufficient, we can use this formula:

$$\delta^2 = \frac{1}{n-1}\Big[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2\Big]$$

Since *sum* and *count* are distributive measures, we have the following:

$$\begin{aligned} variance(D) &= \frac{1}{count(D)-1}\Big(sum(\{d^2|d \in D\}) - \frac{sum(D)^2}{count(D)}\Big) = \\ &= \frac{1}{count(D_w)+count(D_{new})-1}\Big(sum(\{d^2|d \in D_w\}) + sum(\{d^2|d \in D_{new}\}) - \\ &\qquad - \frac{sum(D_w)^2 + sum(D_{new})^2}{count(D_w)+count(D_{new})}\Big) \end{aligned}$$

Thus, it is sufficient to store the values $count(D_w)$, $sum(D_w)$ and $sum(\{d^2|d \in D_w\})$ to accelerate the next aggregation. Since these values are indipendent from $|D_w|$, this approximation of the variance is an algebraic measure.