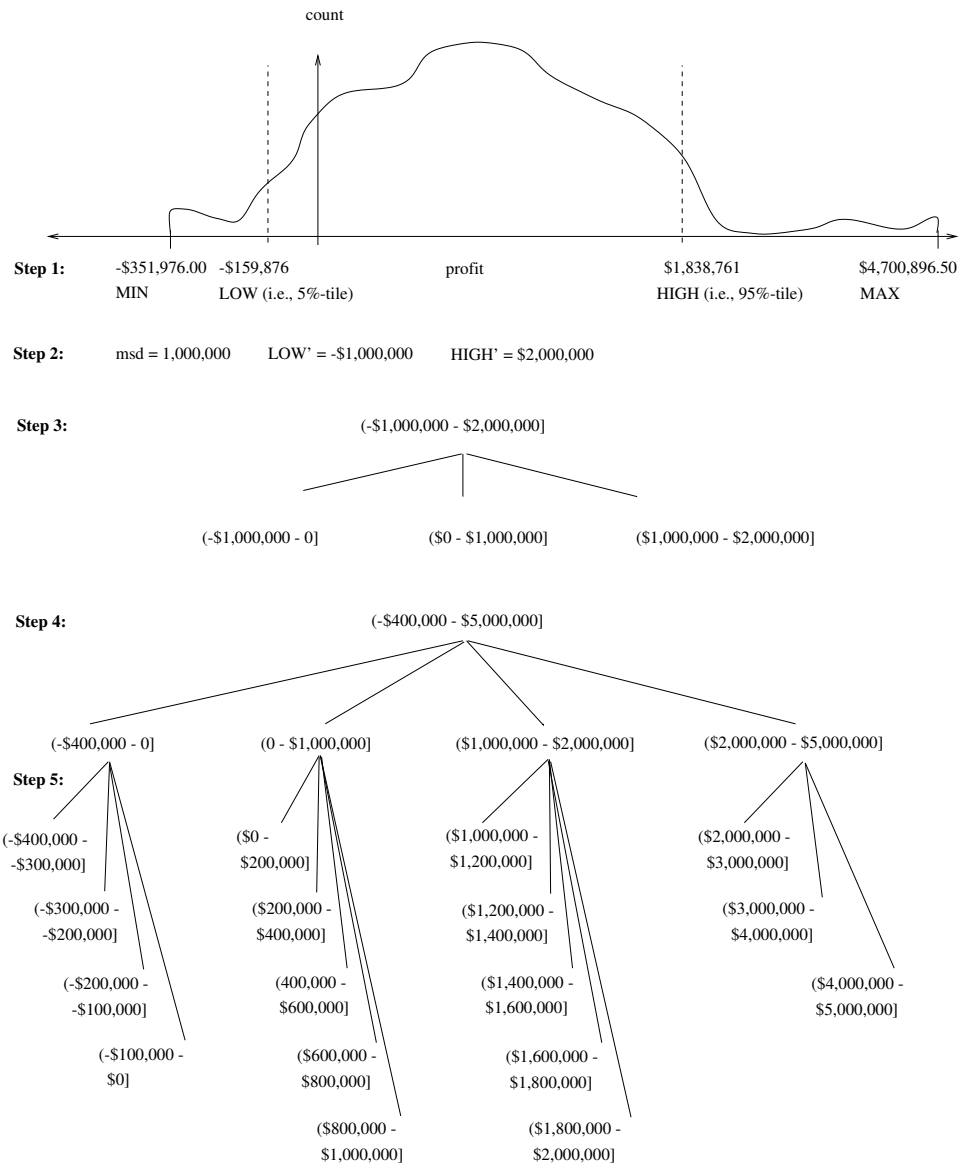

Figure 3.14: Automatic generation of a concept hierarchy for *profit* based on the 3-4-5 rule.

(c) if it covers 1, 5, or 10 distinct values at the most significant digit, then partition the range into 5 equi-width intervals.

The rule can be recursively applied to each interval, creating a concept hierarchy for the given numeric attribute. Since there could be some dramatically large positive or negative values in a data set, the top level segmentation, based merely on the minimum and maximum values, may derive distorted results. For example, the assets of a few people could be several orders of magnitude higher than those of others in a data set. Segmentation based on the maximal asset values may lead to a highly biased hierarchy. Thus the top level segmentation can be performed based on the range of data values representing the majority (e.g., 5%-tile to 95%-tile) of the given data. The extremely high or low values beyond the top level segmentation will form distinct interval(s) which can be handled separately, but in a similar manner.

The following example illustrates the use of the 3-4-5 rule for the automatic construction of a numeric hierarchy.

Example 3.5 Suppose that profits at different branches of *AllElectronics* for the year 1997 cover a wide range, from −$351,976.00 to $4,700,896.50. A user wishes to have a concept hierarchy for *profit* automatically