evaluating each feature independently [141]. The feature vector $\tilde{V}$ can then be simplified according to

$$V = [\mathrm{IC}, f_1, f_2, f_3].$$

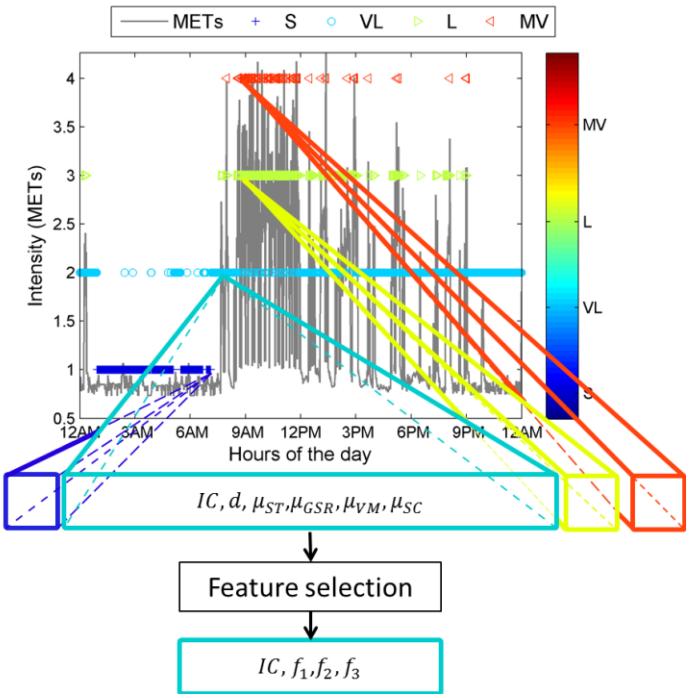

Figure 38 Continuous data stream representing METs value (grey line) is first quantized in IC (blue is sleeping, ciano is very light, green is light, and red is moderate to vigorous intensity). IC of the same intensity form bouts. Features are computed within the bouts, and the most relevant for each intensity, according to the feature selection procedure, will be combined to create PA descriptors.

The features not selected by the feature selection algorithm can be interpreted as wildcard characters representing multiple terms in the vector space ([IC, f_1, f_2, f_3, f_4] and [IC, f_1, f_2, f_3, f_4, f_5] are two such terms of [IC, f_1, f_2, f_3]). As intuitively explained later in this section, wildcarding and the removal of frequent words help in generating sparse topics [139]. The selected features $f_{j\epsilon\{1,2,3\}}$ obviously might be different for each intensity. To generate the vocabulary of words, each of the selected features was first standardized, and then, mapped into a set of discrete levels (which could be interpreted as the letters of our words) using a k-means clustering algorithm. The algorithm automatically selects for each feature the number of levels K^{fj} in a way that the resulting clustering levels $L_{p\epsilon\{1:K\}}$ are the most stable under small perturbations of the input dataset [58]. Levels are sorted according to their mean value L_p in ascending order such that the first level (L_1) represents clusters with the smallest feature values, and the last level (L_K) represents clusters with the highest feature values. Features selected and levels are shown in Table XII. Mean value and variance of the levels were stored and used to create the documents as described in section 6.4.4. The vocabulary of terms was built by allowing all the possible