Title: Instance selection

URL: https://en.wikipedia.org/wiki/Instance_selection

PageID: 53279262

Categories: Category:Data mining, Category:Machine learning

Instance selection (or dataset reduction, or dataset condensation) is an important data pre-processing step that can be applied in many machine learning (or data mining) tasks. [1] Approaches for instance selection can be applied for reducing the original dataset to a manageable volume, leading to a reduction of the computational resources that are necessary for performing the learning process. Algorithms of instance selection can also be applied for removing noisy instances, before applying learning algorithms. This step can improve the accuracy in classification problems.

Algorithm for instance selection should identify a subset of the total available data to achieve the original purpose of the data mining (or machine learning) application as if the whole data had been used. Considering this, the optimal outcome of IS would be the minimum data subset that can accomplish the same task with no performance loss, in comparison with the performance achieved when the task is performed using the whole available data. Therefore, every instance selection strategy should deal with a trade-off between the reduction rate of the dataset and the classification quality.

Instance selection algorithms

The literature provides several different algorithms for instance selection. They can be distinguished from each other according to several different criteria. Considering this, instance selection algorithms can be grouped in two main classes, according to what instances they select: algorithms that preserve the instances at the boundaries of classes and algorithms that preserve the internal instances of the classes. Within the category of algorithms that select instances at the boundaries it is possible to cite DROP3, [2] ICF [3] and LSBo. [4] On the other hand, within the category of algorithms that select internal instances, it is possible to mention ENN [5] and LSSm. [4] In general, algorithm such as ENN and LSSm are used for removing harmful (noisy) instances from the dataset. They do not reduce the data as the algorithms that select border instances, but they remove instances at the boundaries that have a negative impact on the data mining task. They can be used by other instance selection algorithms, as a filtering step. For example, the ENN algorithm is used by DROP3 as the first step, and the LSSm algorithm is used by LSBo.

There is also another group of algorithms that adopt different selection criteria. For example, the algorithms LDIS, [6] CDIS [7] and XLDIS [8] select the densest instances in a given arbitrary neighborhood. The selected instances can include both, border and internal instances. The LDIS and CDIS algorithms are very simple and select subsets that are very representative of the original dataset. Besides that, since they search by the representative instances in each class separately, they are faster (in terms of time complexity and effective running time) than other algorithms, such as DROP3 and ICF.

Besides that, there is a third category of algorithms that, instead of selecting actual instances of the dataset, select prototypes (that can be synthetic instances). In this category it is possible to include PSSA, [9] PSDSP [10] and PSSP. [11] The three algorithms adopt the notion of spatial partition (a hyperrectangle) for identifying similar instances and extract prototypes for each set of similar instances. In general, these approaches can also be modified for selecting actual instances of the datasets. The algorithm ISDSP [11] adopts a similar approach for selecting actual instances (instead of prototypes).

References