ПОСТОРОЕНИЕ ВЗВЕШЕННЫХ ОБУЧАЮЩИХ ВЫБОРОК НА ОСНОВЕ АЛГОРИТМА КРАСКАЛА

Волченко Е.В.

ГВУЗ «Донецкий национальный технический университет», г. Донецк

В работе рассматривается задача предобработки больших объемов данных при построении обучающихся систем распознавания. Одним из эффективных подходов к формированию обучающих выборок является построение взвешенных обучающих выборок *w*-объектов. Каждый объект такой выборки описывается не только набором признаков, но и дополнительным параметром (множеством параметров), содержащих информацию о расположении *w*-объекта в пространстве признаков относительно других объектов, особенностях его формирования и т.п. В случае, когда исходные объекты характеризуются только набором признаков, построение каждого *w*-объекта происходит путем выделения некоторого подмножества близких по значениям признаков объектов (образующего множества) и их замены на один или несколько *w*-объектов. При этом большинство дополнительных параметров содержит информацию об объектах образующего множества (их количестве, удаленности друг от друга и др.).

Разделение исходного множества объектов на образующие множества формально может быть представлено как задача кластеризации данных, а каждый кластер в дальнейшем может представлять один или несколько *w*-объектов. За основу первичного разделения объектов на кластеры предложено использовать алгоритм Краскала построения минимального остовного дерева, как один из наиболее эффективных алгоритмов кластеризации большого числа однотипных данных, когда существует возможность представить эти данные в виде взвешенного графа. В данной работе традиционно объекты представляются вершинами графа, а веса ребер соответствуют расстоянию между обучающими объектами в пространстве признаков.

Для рассматриваемой задачи априорное задание требуемого количества кластеров является неверным, поскольку размер обучающей выборки заранее неизвестен. Поэтому выделение кластеров из сформированного остовного дерева происходит путем удаления ребер, вес которых превышает заданное пороговое значение, зависящее от среднего значения веса ребер остовного дерева и всех ребер полного графа. После удаления таких ребер по каждому связному подграфу формируется *w*-объект, значения признаков которого рассчитываются как средние по всем объектам исходной выборки, составившим этот подграф. В качестве дополнительных параметров *w*-объекта используется количество объектов подграфа и среднее расстояние между ними.

Для оптимизации полученной выборки *w*-объектов после её формирования может быть проведена процедура разделения *w*-объектов, содержащих существенно большее по сравнению с другими *w*-объектами количество объектов исходной выборки, и объединения *w*-объектов, построенных по малому количеству объектов.