УДК 004.852

Автоматизация решения задачи кластеризации при помощи эволюционных алгоритмов

Томп Д.С., Университет ИТМО, г. Санкт-Петербург
Научный руководитель – Фильченков А.А., к.ф-м.н., доц. факультета ИТиП Университета ИТМО

Введение

Задача кластеризации относится к задачам машинного обучения без учителя и часто находит применение для глубинного анализа данных. Различные методы кластеризации нашли своё применение для решения задач биоинформатики, категоризации документов, сегментации рынка, добычи данных в интернете, социальных исследований и других областей. Среди успешно применённых методов кластеризации данных присутствуют эволюционные алгоритмы – их преимущество заключается в возможности оптимизировать любой функционал качества разбиения. Для ряда других задач машинного обучения были применены техники т.н. автоматизации обучения – популярными примерами являются SMBO и активное тестирование; однако, авторами работы не было найдено ни одного свидетельства о попытках применения подобных техник к задаче кластеризации. В данной работе планируется восполнить этот пробел, разработав алгоритм автоматического подбора и настройки эволюционного алгоритма для кластеризации в зависимости от свойств задаваемой выборки.

Цель работы

Целью работы является разработка алгоритма автоматического подбора и настойки эволюционного алгоритма для кластеризации задаваемой пользователем выборки.

Базовые положения исследования

В ходе данной работы предполагается провести исследование существующих эволюционных алгоритмов кластеризации и функционалов качества, реализовать некоторые из мутаций этих алгоритмов и протестировать работу алгоритма (1+1) с этими мутациями на различных выборках данных – как реальных, так и синтезированных. После этого планируется разработать систему автоматического выбора эволюционного алгоритма и настройки этого алгоритма для кластеризации произвольной задаваемой выборки; при обучении системы автоматического выбора будут использоваться упомянутые выше результаты работы эволюционных алгоритмов на тренировочных выборках.

Предварительные результаты

На данный момент реализовано несколько мутаций и функционалов качества, и они протестированы на реальных выборках данных из популярных репозиториев машинного обучения, а также на синтезированных выборках. Разбиения, выдаваемые полученными эволюционными алгоритмами, оказались сравнимы, а иногда и опережали по качеству результат работы популярных методов кластеризации из библиотеки машинного обучения и анализа данных Scikit-Learn.

Список литературы

1. Hruschka, E.R. A Survey of Evolutionary Algorithms for Clustering / E.R. Hruschka, R.J.G.B. Campello, A.A.Freitas, A.C.P.L.F. de Carvalho // IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews, Vol. 39 – 2009 - C.133-155
2. Arbelaitz, O. An extensive comparative study of cluster validity indices. / O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J.M. Pérez, I. Perona // ELSEVIER Pattern Recognition – 2013 – 243-256