УДК 519.8

О СИНТЕЗЕ УПРУГИХ КАРТ ЭВОЛЮЦИОННЫМИ АЛГОРИТМАМИ

Гасанова Т.О.
Научный руководитель — профессор Семенкин Е.С.
Сибирский федеральный университет

Несмотря на то, что модели, использующие обучение с учителем, способны решать многие прикладные задачи, в реальности мы редко знаем требуемый выход, поэтому в биологических системах более обоснованной является модель обучения без учителя. Такая модель использует только предъявляемые ей входные векторы, выделяет статистические свойства объектов и группирует их так, чтобы похожие объекты оказались в одном классе, а непохожие – в разных.

Широко известна нейросетевая архитектура, предложенная Тойво Кохоненом и используемая для автоматической кластеризации (классификации без учителя). Такая архитектура учитывает информацию о взаимном расположении нейронов, образующих решетку. Сигнал поступает сразу на все узлы, а соответствующие синапсы интерпретируют как координаты положения данного узла, выход сети формируется таким образом, что отличный от нуля сигнал имеет ближайший узел (к подаваемому на вход объекту). Сеть обучается таким образом, чтобы узлы решетки оказались в местах локального сгущения данных, то есть моделировали кластерную структуру множества данных, а связи между узлами описывают отношения соседства между кластерами в пространстве признаков.

В настоящее время, в биоинформатике для анализа многомерных данных широко используется метод, получивший название «упругие карты». Этот метод обладает большей регулярностью и предсказуемостью, чем карты Кохонена, и способен решать те же задачи. Основа метода «упругих карт» – минимизация «энергии упругой деформации» карты, погружённой в пространство данных. Задача построения карты ставится как оптимизационная, то есть построенная карта будет решением задачи на оптимизацию функционала от положения узлов относительно входных векторов. После построения сетки остаются неопределенными два параметра, которые можно интерпретировать как упругость карты по отношению к растяжению и упругость по отношению к изгибу. С одной стороны, более упругая сетка является более гладкой моделью данных, обладает большей обобщающей способностью, но, как следствие, хуже описывает отклонения от предполагаемого закона. С другой стороны, менее упругая карта точнее описывает данные, но и воспроизводит при этом случайные шумы, которые обычно присутствуют в реальных данных, то есть менее упругая сетка обладает меньшей обобщающей способностью. Известно, что применяемый для решения оптимизационной задачи метод расщепления может сходиться в локальный минимум. Поэтому целесообразно рассмотреть возможность применения в методе упругих карт алгоритмов глобальной оптимизации.

Известно, что эволюционные, в частности – генетические, алгоритмы способны эффективно решать многоэкстремальные задачи оптимизации с целевыми функциями, заданными неявно (таблично, алгоритмически и т.п.) на сложных структурах данных (дискретных, комбинаторных, смешанных). Генетические алгоритмы являются стохастическими процедурами прямого поиска на множестве бинарных переменных, и упомянутые трудности оптимизации не создают для них дополнительных проблем.

В данной работе рассматривается применение гибридного генетического алгоритма для синтеза упругих карт, решающих задачу кластеризации многомерных данных. Гиб-

ридизация состоит в использовании покоординатного спуска для наилучшего найденного решения.

Тестирование подхода проводилось на задаче из репозитория UCI «Ирисы Фишера». Ирисы Фишера состоят из данных о 150 экземплярах ириса, по 50 экземпляров из трёх сортов — iris setosa, iris virginica и iris versicolor. Для каждого экземпляра измерялись четыре характеристики (в сантиметрах):

Длина чашелистика (англ. sepal length);

Ширина чашелистика (англ. sepal width);

Длина лепестка (англ. petal length);

Ширина лепестка (англ. petal width).

На основании этого набора данных требуется построить правило классификации, определяющее сорт растения по данным измерений. Это задача многоклассовой классификации, так как имеется три класса — три сорта ириса. Один из классов (iris setosa) линейно разделим от двух остальных (рис. 1).

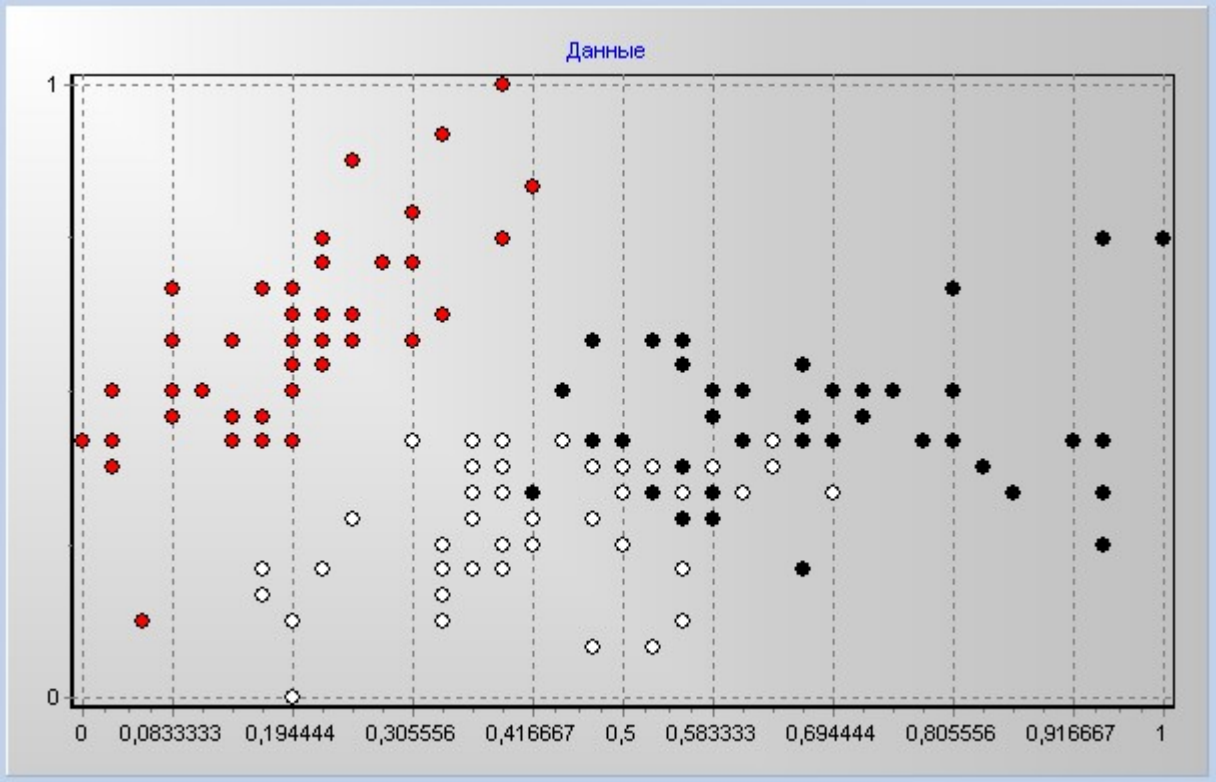

Рис. 1. Ирисы Фишера (по горизонтальной оси – длина чашелистника, по вертикальной оси – ширина чашелистника)

Процедура оптимизации энергии упругой деформации гибридным генетическим алгоритмом и результаты тестирования работоспособности подхода обсуждаются в докладе.