УДК 004

АВТОМАТИЗАЦИЯ МЕТОДА ИЗОЛИРУЮЩЕГО ЛЕСА

Магистрант Каменко Д. А.
Кандидат физ.-мат. наук, доцент Гундина М. А.
Белорусский национальный технический университет, Минск, Республика Беларусь

Идея изолирующего леса основана на принципе Монте-Карло: проводится случайное разбиение пространства признаков такое, что в среднем изолированные точки отсекаются от нормальных, кластеризованных данных.

Алгоритм изолирующего дерева заключается в построении случайного бинарного решающего дерева. Корнем дерева является все пространство признаков; в очередном узле выбирается случайный признак и случайный порог разбиения. Критерием остановки является тождественное совпадение всех объектов в узле, то есть решающее дерево строится полностью. Ответом в листе, который также соответствует f алгоритма, объявляется глубина листа в построенном дереве.

Утверждается, что аномальным точкам свойственно оказываться в листьях с низкой глубиной, то есть в листьях, близким к корню, когда же для разбиения гиперплоскостями кластера нормальных данных дереву потребуется построить еще несколько уровней.

При «случайном» способе построения деревьев выбросы будут попадать в листья на ранних этапах (на небольшой глубине дерева), т. е. выбросы проще «изолировать». Выделение аномальных значений происходит на первых итерациях работы алгоритма.

Точка данных определяется как выброс, если ее число изоляции ниже порогового значения.

Порог определяется на основе расчетного процента выбросов в данных, что является отправной точкой этого алгоритма обнаружения выбросов.

Воспользуемся библиотекой встроенных данных в систему Wolfram Mathematica.

```
data=ResourceData["Sample Data: Fisher's Irises"][[All,{"PetalLength","SepalWidth"}]]
```

Найдем аномалии:

```
outliers=FindAnomalies[data,Method->"DecisionTree"]
```

Представим графически эти данные:

```
ListPlot[{data,outliers},Sequence[PlotStyle->{Directive[Opacity[0.8],PointSize[0.02]],Directive[Red,PointSize[0.025]]},Frame->True,FrameLabel->{"PetalLength","SepalWidth"},PlotLegends->{"Data","Anomalies"},ImageSize->Medium,AspectRatio->Automatic]]
```

Результат представлен на рис. 1.

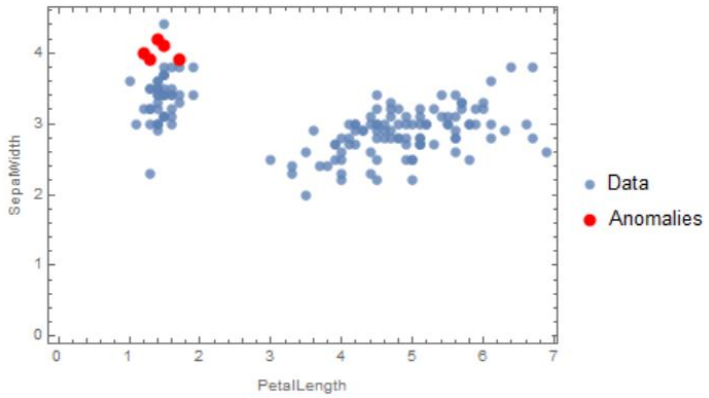

Рис. 1. Графическое представление данных

Алгоритм обладает рядом существенных преимуществ. Алгоритм распознает аномалии различных видов: как изолированные точки с низкой локальной плотностью, так и кластеры аномалий малых размеров. Он не требует существенных затрат по памяти, в отличие от, например, метрических методов, зачастую требующих построения матрицы попарных расстояний. Метод инвариантен к масштабированию признаков; не требует задания метрики или другой априорной информации об устройстве данных.