УДК 005

ДИАГНОСТИКА ЗАБОЛЕВАНИЙ ПЕЧЕНИ ПО СТАТИСТИЧЕСКИМ ПЕРЕМЕННЫМ В СРЕДЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ RAPIDMINER

Пивень О.И.

Сибирский государственный университет науки и технологий

E-mail: 9piveneg@gmail.com

В статье рассматриваются методы кластеризации набора данных по диагностированию заболеваний печени в среде интеллектуального анализа RapidMiner, дана характеристика каждому методу. В определенной мере полученный результат может служить дополнительным ориентиром при принятии решения о качестве набора данных с целью постановки точного диагноза.

Ключевые слова: печень, кластеризация, кластерный анализ, методы кластерного анализа, RapidMiner.

DIAGNOSTICS OF DISEASES OF THE LIVER ON STATISTICAL VARIABLES IN THE ENVIRONMENT OF DATA ANALYSIS RAPIDMINER

Piven O.I.

The article discusses the methods of clustering a set of data on the diagnosis of liver diseases in the RapidMiner intellectual analysis environment, characterizing each method. To a certain extent, the result obtained can serve as an additional guideline when deciding on the quality of a data set in order to make an accurate diagnosis.

Keywords: liver, clustering, cluster analysis, cluster analysis methods, RapidMiner

На данный момент существует множество компаний, нуждающихся в системах аналитики, так как большой объем поступающей и накопленной информации на предприятиях, а также различных заведениях и учреждениях, при должной обработке, способен принести значительную пользу. Немаловажным направлением применения аналитических ресурсов является медицинская сфера. На сегодняшний день госпитали оснащены по современным стандартам оказания медицинских услуг, в том числе и современным цифровым оборудованием, которое позволяет собрать определенные статистические данные. В данной статье рассматривается набор данных по диагностированию заболеваний печени для выявления подходящего метода кластеризации. Обработка данных будет осуществляться в среде интеллектуального анализа RapidMiner

В данной работе используются следующие методы кластеризации:

1. Наивный Байесовский классификатор. Данный метод привлекателен своей простотой и фундаментальностью. Данный метод зарекомендовал себя как эффективный метод в решении

простых задач. Минусы: плохо работает на данных с большим количеством атрибутов; на коротких выборках склонен к переобучению.

2. Random Forest. Данный метод выбран из-за того, что является случайным и соответственно обладает способностью эффективно решать сложные, многомерные задачи. Метод обрабатывает как дискретные, так и непрерывные признаки. В качестве недостатка метода можно выделить большой размер получающихся моделей, длительность работы.

3. Метод k-ближайших соседей. Традиционный, наиболее классический и обоснованный метод кластеризации. Метод привлекателен своей наглядностью. Минусы: подверженность проклятию размерности; метод считает все признаки равными по значению; слабая обоснованность выбора значения k.

4. Дерево принятия решений. Выбран в силу того, что зарекомендовал себя как крайне эффективный метод в решении задач малой и средней сложности. Метод способен работать как с категориальными, так и с интервальными величинами. Минусы метода: склонность к переобучению; алгоритм склонен ошибочно давать больший вес тем или иным атрибутам, исходя только из их количественных характеристик.

5. Логистическая регрессия. Данный метод выбран потому, что традиционно подобные методы редко используются в задачу кластеризации. Результат работы данного метода не очевиден, и тем самым интересен. К минусам метода относится склонность к переобучению и то, что метод хорошо работает только при сильных теоретико - вероятностных предположениях.

Каждый из методов впоследствии был оптимизирован и произведен сравнительный анализ работы каждого метода, а также обработка данных проходила различными ансамблями этих методов. Приведена оценочная характеристика качества классификации для каждого из них.

Сведения о каждом из методов

Байесовский классификатор - широкий класс алгоритмов классификации, основанный на принципе максимума апостериорной вероятности (применение теоремы Байеса со строгими (наивными) предположениями о независимости). Для классифицируемого объекта вычисляются функции правдоподобия каждого из классов, по ним вычисляются апостериорные вероятности классов. Объект относится к тому классу, для которого апостериорная вероятность максимальна. Достоинством наивного байесовского классификатора является малое количество данных для обучения, необходимых для оценки параметров, требуемых для классификации [3].

Random forest (с англ. — «случайный лес») – метод основан на построении большого числа (ансамбля) деревьев решений (это число является параметром метода), каждое из которых строится по выборке, получаемой из исходной обучающей выборки. Классификация

осуществляется с помощью голосования классификаторов, определяемых отдельными деревьями, и побеждает класс, за который проголосовало наибольшее число деревьев [6]. Оптимальное число деревьев подбирается таким образом, чтобы минимизировать ошибку классификатора на тестовой выборке. В случае её отсутствия, минимизируется оценка ошибки out-of-bag: доля примеров обучающей выборки, неправильно классифицируемых комитетом, если не учитывать голоса деревьев на примерах, входящих в их собственную обучающую подвыборку.

Метод K-ближайшего соседа (англ.: k-nearest neighbors method, k-NN) – один из методов решения задачи классификации. В основе k-NN лежит следующее правило: объект считается принадлежащим тому классу, к которому относится большинство его ближайших соседей. Под «соседями» здесь понимаются объекты, близкие к исследуемому в том или ином смысле [4]. Применяя метод k-NN в пространстве признаков объектов необходимо определить некоторую метрику (т.е. функцию расстояния). Классический вариант определения дистанции — дистанция в евклидовом пространстве. Предполагается, что объекты с близкими значениями одних признаков будут близки и по другим признакам (т.е. относиться к одному и тому же классу).

Дерево принятия решений (также может называться деревом классификации или регрессионным деревом) — способ представления правил в иерархической, последовательной структуре, где каждому объекту соответствует единственный узел, дающий решение. Является средством поддержки принятия решений, использующееся в статистике и анализе данных для прогнозных моделей [1]. Структура дерева состоит из узлов, соединенных друг с другом ребрами, что представляют собой «листья» и «ветки». На ребрах («ветках») дерева решения записаны атрибуты, от которых зависит целевая функция, в узлах («листьях») записаны значения целевой функции, а в остальных узлах — атрибуты, по которым различаются случаи. Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение.

Применяется для предсказания вероятности возникновения некоторого события по значениям множества признаков [5]. Для этого вводится так называемая зависимая переменная y, принимающая лишь одно из двух значений — как правило, это числа 0 (событие не произошло) и 1 (событие произошло), и множество независимых переменных (также называемых признаками, предикторами или регрессорами), на основе значений которых требуется вычислить вероятность принятия того или иного значения зависимой переменной.

Каждый из данных методов будет оптимизирован. В качестве параметров для оптимизации выбраны:

Таблица 1. Оптимизируемые параметры

Метод	Оптимизируемые параметры
Логистическая регрессия	Параметр C ядра
Метод k-ближайших соседей	Величина k
Дерево принятия решений	Максимальная глубина дерева
Наивный Байесовский классификатор	Laplace correction
Random Forest	Количество деревьев

3 Вычислительные эксперименты

Исследования будут осуществляться в среде **RapidMiner** - среде интеллектуального анализа данных.

Данные представляют из себя статистические переменные. В каждом наборе данных присутствуют 7 атрибутов, 5 из них – анализы крови, которые считаются чувствительными к нарушениям работы печени, которые могут возникнуть в результате чрезмерного употребления алкоголя. Шестой атрибут – количество напитков на эквивалент 0,25 мл. алкогольных напитков, выпитых в день. Седьмой атрибут – поле селектора, чтобы разделить данные на испытательные установки (введен исследователями BUPA). Каждая строка в наборе данных представляет собой запись одного человека мужского пола.

Сравнение алгоритмов классификации будет осуществляться с помощью **скользящего контроля** (*cross-validation*) [2]. Скользящий контроль - процедура эмпирического оценивания обобщающей способности алгоритмов, обучаемых по прецедентам.

Метод скользящего контроля работает следующим образом: фиксируется некоторое множество разбиений исходной выборки на две подвыборки: обучающую и контрольную. Для каждого разбиения выполняется настройка алгоритма по обучающей подвыборке, затем оценивается его средняя ошибка на объектах контрольной подвыборки. Оценкой скользящего контроля называется средняя по всем разбиениям величина ошибки на контрольных подвыборках.

Если выборка независима, то средняя ошибка скользящего контроля даёт несмещённую оценку вероятности ошибки. Это выгодно отличает её от средней ошибки на обучающей выборке, которая может оказаться смещённой (оптимистически заниженной) оценкой вероятности ошибки, что связано с явлением переобучения.

Скользящий контроль является стандартной методикой тестирования и сравнения алгоритмов классификации, регрессии и прогнозирования.

Ансамбли из трех лучших и худших методов строились из методов, включая оптимизированные и неоптимизированные. Ансамбли из пяти методов строились только из неоптимизированных методов, так можно определить, что эффективнее, ансамбль методов или отдельный метод.

Первый набор данных:

Этот набор данных включает 7 атрибутов и 110 измерений.

Таблица 2. Результаты работы алгоритмов на первом наборе данных

Метод	Качество классификации
Наивный Байесовский классификатор	85.67%
Random Forest	78.44%
K-nn	81.33%
Дерево решений	85.33%
Логистическая регрессия	67.41%
3 худших метода	70.41%
3 лучших метода	85.41%
5 методов	73.95%
Наивный Байесовский классификатор (оптимизированный)	85.89%
Random Forest (оптимизированный)	80.44%
K-nn (оптимизированный)	81.67%
Дерево решений (оптимизированный)	85.33%
Логистическая регрессия (оптимизированный)	79.16%

Наилучший результат на первом наборе данных показал оптимизированный метод «Наивный байесовский классификатор» – 85.89% правильно классифицированных значений. Хуже справился неоптимизированный «Наивный Байесовский классификатор» – 85.67%. На третьем месте по качеству ансамбль из трех лучших методов – 85.41%.

Второй набор данных:

Этот набор данных включает 7 атрибутов и 112 измерений.

Таблица 3. Результаты работы алгоритмов на втором наборе данных

Метод	Качество классификации
Наивный Байесовский классификатор	82.56%
Random Forest	84.56%
K-nn	84.56%
Дерево решений	79.33%
Логистическая регрессия	77.88%
3 худших метода	77.95%
3 лучших метода	84.75%
5 методов	81.45%
Наивный Байесовский классификатор (оптимизированный)	84.67%
Random Forest (оптимизированный)	85.56%
K-nn (оптимизированный)	84.56%
Дерево решений (оптимизированный)	87.56%
Логистическая регрессия (оптимизированный)	81.35%

Наилучший результат на первом наборе данных показал оптимизированный метод «Дерево решений» – 87.56% правильно классифицированных значений. Хуже справился

оптимизированный метод «Random Forest» - 85.56%. На третьем месте по качеству ансамбль из трех лучших методов – 84.75%.

Третий набор данных:

Этот набор данных включает 7 атрибутов и 123 измерения.

Таблица 4. Результаты работы алгоритмов на третьем наборе данных

Метод	Качество классификации
Наивный Байесовский классификатор	77.89%
Random Forest	76.11%
K-nn	72.78%
Дерево решений	90.22%
Логистическая регрессия	69.57%
3 худших метода	73.77%
3 лучших метода	85.49%
5 методов	82.08%
Наивный Байесовский классификатор (оптимизированный)	80.78%
Random Forest (оптимизированный)	79.22%
K-nn (оптимизированный)	75.00%
Дерево решений (оптимизированный)	91.33%
Логистическая регрессия (оптимизированный)	78.23%

Наилучший результат на первом наборе данных показал оптимизированный метод «Дерево решений» – 91.33% правильно классифицированных значений. Далее идёт неоптимизированный метод «Дерево решений» – 90.22%. На третьем месте по качеству ансамбль из трех лучших методов – 85.49%.

На данных, рассматриваемых в этой работе, нет возможности однозначно выделить лучший метод, но, стоит заметить, что наилучше всего себя показал оптимизированный метод «Дерево решений», оказавшись на первом месте при работе со вторым и третьим набором данных. Ансамбль из трех лучших методов показывал стабильный результат на всех трех наборах данных, работая лучше, чем один из лучших методов, но, как и ансамбль из пяти методов, уступал лучшему методу. Ансамбль из трёх худших методов всегда работал лучше, чем один худший метод, но не лучше отдельного лучшего.

В работе с текущими данными оптимизация метода всегда приводила к существенному или хотя бы несущественному улучшению качества работы. Во всех трёх случаях самый лучший результат показывал именно какой-то оптимизированный метод.

В определенной мере полученный результат может служить дополнительным ориентиром при принятии решения о качестве набора данных с целью постановки точного диагноза.

Список литературы

1. Алекс Дж. Шампандар. Глава 26. Деревья классификации и регрессии // Развитие игрового интеллекта: синтетические существа с обучением и реактивным поведением — М.: Вильямс, 2007. — С. 385–401. — 385 с.

2. Воронцов К. В. Комбинаторный подход к оценке качества обучаемых алгоритмов. — Математические вопросы кибернетики / Под ред. О. Б. Лупанов. — М.: Физматлит, 2004. — Т. 13. — С. 5–36.

3. Машинное обучение, наивный байесовский классификатор [Электронный ресурс]. URL: http://ru.cybernetics.wikia.com/wiki/наивный_байесовский_классификатор (дата обращения 20.11.17).

4. Метод K-ближайших соседей для решения задачи классификации [Электронный ресурс]. URL: https://edu.kpfu.ru/pluginfile.php/78207/mod_resource/content/1/kNN.pdf (дата обращения 20.11.17).

5. Справочник по прикладной статистике. В 2-х т. Т. 1: Пер. с англ. / Под ред. Э. Ллойда, У. Ледермана, Ю. Н. Тюрина. — М.: Финансы и статистика, 1989. — 510 с.

6. Чистяков С. П. СЛУЧАЙНЫЕ ЛЕСА: ОБЗОР: - М.: Труды Карельского научного центра РАН № 1. 2013. С. 117–136. – 118 с.