МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

ПРИЛОЖЕНИЕ МЕТОДОВ КЛАССИФИКАЦИИ В ЗАДАЧАХ МЕДИЦИНСКОЙ ДИАГНОСТИКИ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студентки 2 курса 248 группы
направления 09.04.03 — Прикладная информатика

механико-математического факультета
Максимкиной Анастасии Эдуардовны

Научный руководитель
доцент, к. ф.-м. н., доцент _____ Е. В. Гудошникова

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2024

ВВЕДЕНИЕ

Актуальность темы исследования. Машинное обучение представляет собой обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов для анализа данных и получения выводов и выноса решения или предсказания в отношении чего-либо. Подход, при котором прошлые данные или примеры используются для первоначального формирования и совершенствования схемы предсказания, называется методом машинного обучения. Общая задача машинного обучения заключается в восстановлении зависимости между входными и выходными элементами с целью предсказания будущего выхода по заданному входу. Целью машинного обучения является построение максимально точной модели на основе данных и затем использования этой модели для предсказаний в будущем.

В зависимости от наличия или отсутствия прецедентной информации различают ряд категорий машинного обучения: контролируемое обучение или “обучение с учителем”, неконтролируемое обучение, обучение с подкреплением. Большая вариативность позволяет применять методы машинного обучения для данных различных типов в самых разных областях: биоинформатике, медицинской диагностике, технике. Широкий спектр приложений методы машинного обучения получили в экономике. Так они используются для обнаружения мошенничества, кредитного скоринга, биржевого технического анализа. В современных условиях функционирования социально-экономических систем проблема получения приемлемого прогноза может быть решена за счет комбинирования традиционных классических методов совместно с методами интеллектуального прогнозирования.

Для получения релевантных результатов необходимы подходящие инструменты и корректные алгоритмы, в связи с чем методы машинного обучения и data mining получили широкое применение при анализе и прогнозировании. Существует множество методов машинного обучения: искусственные нейронные сети, деревья принятия решений, логистическая регрессия, генетический алгоритм.

Актуальность определила выбор темы данной работы: "Приложение методов классификации в задачах медицинской диагностики".

Цель работ: получение теоретических и практических навыков построения случайного леса и k-ближайших соседей.

Для достижения поставленных целей в работе необходимо решить следующие задачи:

— Применение методов классификации на практике.
— Проведение классификации методом:
  – случайного леса;
  – k-ближайших соседей.
— Автоматизация решения задачи классификации на языке R.

Практическая значимость работы заключается в разработке программных продуктов для построения случайного леса и k-ближайших соседей.

Основное содержание работы

Магистерская работа состоит из: введения, трех теоретических и трех практических глав, заключения, списка использованных источников, приложения.

Введение содержит основные положения: актуальность темы исследования (цель, объект, предмет, задачи исследования);практическую значимость исследования.

Первый раздел "Задача классификации" описывает теоретические основы классификации.

Проведение классификации в R. В задаче классификации зависимая переменная является категориальной, то есть может принимать конечное число значений. Классификация относится к кассическим задачам машинного обучения. Состоит в прогнозировании класса входного вектора на основе одной зависимой переменной.

Также будет рассмотрен вопрос, связанный с оценкой качества модели классификации. Алгоритм классификации вычисляет вероятность принадлежности к одному классу, затем входному вектору присваивается тот класс, вероятность принадлежности к которому вектора больше.

Практическое применение методов классификации.

Проблемы, при решении которых возникает задача классификации:

— классификация как необходимый предварительный этап статистической обработки данных;
— классификация в задачах прогнозирования экономико-социологических ситуаций для отдельных показателей.

Постановка задачи классификации. Задача заключается в том, чтобы построить такую программу, которая, используя обучающую последовательность, вырабатывала бы правило, позволяющее классифицировать вновь предъявляемые «незнакомые» ситуации (вообще говоря, отличные от входивших в обучающую последовательность).

Способность к обучению характеризуется двумя понятиями:
— качеством полученного решающего правила (вероятностью неправильных ответов — чем меньше эта вероятность, тем выше качество);
— надежностью получения решающего правила с заданным качеством (вероятностью получения заданного качества — чем выше эта вероятность, тем выше надежность успешного обучения).

Математическая постановка задачи обучения. В среде, которая характеризуется распределением вероятностей $P(x)$, случайно и независимо появляются ситуации x. Существует «учитель», который классифицирует их, то есть относит к одному из k классов (для простоты $k = 2$). Пусть он делает это согласно условной вероятности $P(t|x)$, где $t = 1$ означает, что вектор x отнесен к первому классу, а $t = 0$ — ко второму. Ни характеристика среды $P(x)$, ни правило классификации $P(t|x)$ нам не известны. Однако известно, что обе функции существуют, то есть существует совместное распределение вероятностей

$$P(x, t) = P(x) \cdot P(t|x).$$

Пусть теперь определено множество Ω решающих правил $F(x, \alpha)$. В этом множестве каждое правило определяется заданием параметра α (обычно это вектор). Все правила F(x,α) — характеристические функции, то есть могут принимать только одно из двух значений — нуль или единицу:

$$F(x, \alpha) = \begin{cases} 1, x - \text{принадлежит первому классу}, \\ 0, x - \text{принадлежит второму классу}, \end{cases}$$

Для каждой функции $F(x,\alpha) \in \Omega$ может быть определено качество $Q(\alpha)$ как вероятность различных классификаций ситуаций x с учителем.

1. В случае, когда пространство X дискретно и состоит из точек $x^1, ..., x^N$

$$Q(\alpha) = \sum_{t=0}^{1} \sum_{t=1}^{N} (t - F(x^i, \alpha))^2 P(x^i) P(t|x^i),$$

где $P(x^i)$ — вероятность возникновения ситуации x^i.

2. В случае, когда в пространстве X существует плотность распределения $p(x)$,

$$Q(\alpha) = \sum_{t=0}^{1} \int (t - F(x, \alpha))^2 p(x) P(t|x) dx.$$

3. В общем случае можно сказать, что в пространстве X задана вероятностная мера $P(x,t)$, тогда

$$Q(\alpha) = \int_{x,t} (t - F(x, \alpha))^2 dP(x,t).$$

Среди всех функций $F(x,\alpha)$ есть такая $F(x,\alpha^0)$, которая минимизирует вероятность ошибок. Эту функцию (или близкую к ней) и следует найти. Так как совместное распределение вероятностей $P(x,t)$ неизвестно, поиск ведется с использованием обучающей последовательности

$$(x^1, t_1), (x^2, t_2), ..., (x^N, t_N),$$

то есть случайной и независимой выборки примеров фиксированной длины N. Нельзя найти алгоритм, который по конечной выборке безусловно гарантировал успех поиска. Успех можно гарантировать лишь с некоторой вероятностью $1 - \eta$.

Таким образом, задача заключается в том, чтобы для любой функции $P(x,t)$ среди характеристических функций $F(x,\alpha)$ найти по обучающей последовательности фиксированной длины N такую функцию $F(x,\alpha^*)$, о которой с надежностью, не меньшей $1 - \eta$, можно было бы утверждать, что ее качество отличается от качества лучшей функции $F(x,\alpha^0)$ на величину, не превышающую ϵ.

Второй раздел "Применение алгоритма случайного леса к задаче классификации". Алгоритм индукции случайного леса может быть представлен в следующем виде:

1. Для $i = 1, 2, ..., B$ (здесь B — количество деревьев в ансамбле) выполнить:
   - Сформировать бутстреп выборку S размера l по исходной обучающей выборке $D = \{x_i, y_i\}_{i=1}^{l}$;
   - По бутстреп выборке S индуцировать неусеченное дерево решений T_i с минимальным количеством наблюдений в терминальных вершинах равным n_min, рекурсивно следуя следующему подалгоритму:
     - из исходного набора n признаков случайно выбрать p признаков;
     - из p признаков выбрать признак, который обеспечивает наилучшее расщепление;
     - расщепить выборку, соответствующую обрабатываемой вершине, на две подвыборки;
2. В результате выполнения шага 1 получаем ансамбль деревьев решений $\{T_i\}_{i=1}^{B}$;
3. Предсказание новых наблюдений осуществлять следующим образом:
   - для регрессии:
   $$\widehat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{i=1}^{B} T_i(x);$$
   - для классификации: пусть $\widehat{\omega}(x) \in \{\omega_1, \omega_2, ..., \omega_c\}$ - класс, предсказанный деревом решений T_i, т.е. $T_i(x) = \widehat{\omega}_i(x)$; тогда $\widehat{\omega}_{rf}^{B}(x)$ - класс, наиболее часто встречающийся в множестве $\{\widehat{\omega}_b(x)\}_{i=1}^{B}$.

Одно из достоинств случайных лесов состоит в том, что для оценки вероятности ошибочной классификации нет необходимости использовать кросс–проверку или тестовую выборку. Оценка вероятности ошибочной классификации случайного леса осуществляется методом "Out–Of–Bag"(OOB), состоящем в следующем. Известно, что каждая бутстреп выборка не содержит примерно 37% наблюдений исходной обучающей выборки (поскольку выборка с возвращением, то некоторые наблюдения в нее не попадают, а некоторые попадают

несколько раз). Классифицируем некоторый вектор $x \in D$. Для классификации используются только те деревья случайного леса, которые строились по бутстреп выборкам, не содержащим x, и как обычно используется метод голосования. Частота ошибочно классифицированных векторов обучающей выборки при таком способе классификации и представляет собой оценку вероятности ошибочной классификации случайного леса методом OOB. Практика применения оценки OOB показала, что в случае, если количество деревьев достаточно велико, эта оценка обладает высокой точностью. Если число деревьев мало, то оценка имеет положительное смещение.

Третий раздел "Применение алгоритма k-ближайших соседей к задаче классификации".

Алгоритм KNN предсказывает метки тестового набора данных, просматривая метки его ближайших соседей в пространстве объектов обучающего набора данных. “K” - это наиболее важный гиперпараметр, который можно настроить для оптимизации производительности модели.

KNN - это простой и интуитивно понятный алгоритм, который обеспечивает хорошие результаты для широкого круга задач классификации. Его легко реализовать и понять, и он применим как к небольшим, так и к большим наборам данных. Однако у нее также есть некоторые недостатки, и главный недостаток заключается в том, что она может быть дорогостоящей с точки зрения вычислений для больших наборов данных или многомерных пространств объектов.

Алгоритм KNN используется в системах рекомендаций электронной коммерции, распознавания изображений, обнаружения мошенничества, классификации текстов, обнаружения аномалий и многих других. В данной работе этот алгоритм будет использоваться для прогнозирования возникновения болезни.

Алгоритм классификации KNN работает путем поиска K соседей (ближайших точек данных) в обучающем наборе данных для новой точки данных. Затем новым точкам данных присваивается метка мажоритарного класса среди соседей.

Четвертый раздел "Описание набора данных".

Этот набор состоит из 371 субъекта в возрасте от 60 до 96 лет. Каждый

субъект был просканирован во время двух или более посещений с интервалом не менее одного года. Для каждого субъекта включены 3 или 4 отдельных МРТ. Все испытуемые были правшами и включали как мужчин, так и женщин. Часть испытуемых характеризовались как здоровые на протяжении всего исследования. Часть включенных субъектов были охарактеризованы как страдающие деменцией во время их первых посещений и оставались таковыми при последующих наблюдениях, включая 51 человека с болезнью Альцгеймера легкой и умеренной степени тяжести. Еще несколько субъектов были охарактеризованы как здоровые во время их первоначального посещения и впоследствии были охарактеризованы как нездоровые при более позднем посещении.

Пятый раздел "Прогнозирование деменции с помощью алгоритма случайного леса".

Для начала разделим набор данных на две выборки: обучающую и тестовую выборки.

Далее используются модель дерева решений для прогнозирования. Модель может быть чрезмерно сложной. Как правило, после построения модели дерева решений возможно сократить модель, свернув определенные ребра, узлы и листья вместе без существенной потери производительности. Сокращение модели будет достигнуто путём k-кратной перекрестной проверки в рамках обучающего набора.

После того, как модель построена, она обычно может хорошо классифицировать примеры из обучающей выборки. Этап оценки точности предсказания модели и ошибки прогнозирования на новом наборе тестовых данных является крайне важным. Фактический результат каждого примера из набора тестовых данных известен, поэтому оценку эффективности предсказанной силы модели можно проводить на основе сравнения предсказанных моделью значений с известными значениями.

Далее сравниваются фактические результаты с прогнозируемыми. Для этого используется матрица ошибок (Confusion Matrix). Confusion Matrix представляет собой таблицу, которая описывает эффективность классификации для каждой модели на основе тестовых данных.

Чтобы построить модель случайного леса, необходимо вызвать функ-

цию randomForest. Функция randomForest имеет общий вид:

```
randomForest(formula, data, ntree, mtry, classwt, cutoff, nodesize,
maxnodes, do.trace=FALSE, importance=FALSE),
```

где:

- formula= задает формулу в формате: Зависимая переменная Предиктор1 + Предиктор2 + Предиктор3 + др.;
- data= задает таблицу данных для анализа. Если таблица данных названа data, можно просто указать data;
- ntree= задает количество деревьев в ансамбле. Значение по умолчанию — 500;
- mtry= задает m — количество переменных (признаков), случайно отбираемых при разбиении;
- classwt= задает априорные вероятности;
- cutoff= задает пороговое значение вероятности для прогнозирования класса;
- nodesize= задает минимальный размер терминальных узлов;
- maxnodes= задает максимальное количество терминальных узлов, которое могут иметь деревья в случайном лесе.

Потребуется график зависимости ошибок классификации по методу OOB от количества деревьев в ансамбле. График показывает зависимость ошибок классификации по методу OOB (речь идет об общей доле ошибочно классифицированных наблюдений и доле ошибочно классифицированных наблюдений по каждой категории зависимой переменной) от количества деревьев в ансамбле.

Кроме того, для визуализации результатов генерируется график зависимости правильности модели от количества переменных для разбиения.

Затем строится ROC-кривая. У высокоэффективного классификатора будет ROC-кривая, которая круто поднимается в верхний левый угол, то есть он будет правильно идентифицировать множество примеров одного класса без ошибочной классификации на множестве примеров другого класса как примеров из первого класса.

Шестой раздел "Прогнозирование деменции с помощью алгоритма k-ближайших соседей".

Для алгоритма k-ближайших соседей категориальной переменной используется CDR. Для алгоритма случайного леса использовались 4 значения данной переменной: 0; 0.5; 1; 2. Для алгоритма KNN значения переменной CDR преобразуем на 2 группы - люди больные деменцией и здоровые.

Для начала избавимся от пустых значений. Далее проводится преобразование соответствующих переменных в факторные переменные и разабиение переменной CDR на больных деменцией и здоровых.

Далее следует преобразование числовых переменных. Важно проверить, не коррелируют ли какие-либо предикторы друг с другом. Если они коррелируют, то эффективность прогнозирования может ухудшиться и возникнет числовая нестабильность.

Затем числовые переменные подвергаются центрированию и масштабированию.

Данные разбиваются на два множества: обучающий набор и тестовый набор. Для начала проводится поиск значения K, то есть выбрать количество соседей. Выбор K-значения очень важен для точности модели. Для этого можно использовать метод локтя и искать значение K из диапазона значений k и проверять точность модели при каждом поиске. Найдено оптимальное k=6, количество ближайших значений, которые необходимо собрать, чтобы сделать прогноз.

Используя полученную модель предсказывается класс для тестового набора и визуализирована матрица ошибок. Далее строится ROC-кривая.

Основные результаты

1. Рассмотрены теоретические основы задачи классификации, проведение классификации на языке программирования R, практическое применение данного класса задач и постановка задачи.
2. Изучены основы построения метода случайного леса, разработан программный код на языке программирования R, позволяющий провести классификацию на медицинских данных.
3. Описаны результаты построенных моделей случайного леса.
4. Рассмотрены основы построения метода k - ближайших соседей, построен программный код на языке программирования R, позволяющий

легко воспроизвести расчеты.

5. Описаны результаты построенных моделей k - ближайших соседей.

Программный код приводится в **приложении А.**