РЕФЕРАТ

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

ОСНОВНАЯ ЧАСТЬ

1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1.1 Графы

1.1.1 Основные определения

В данной работе графы будет использоваться в двух местах: для визуализации и анализа, а также при построении деревьев решений.

Добавлено примечание ([ДС1]): Нужно ли об этом писать в теоретической части…

Определение 1.1. Графом называется упорядоченная пара $G = (V(G), E(G))$, где $V(G)$ - непустое множество вершин (узлов) графа G, а $E(G)$ – множество ребер графа G.

Далее будут обозначены некоторые вспомогательные определения:

- количество вершин графа называется его порядком и обозначается $v = |V(G)|$. В данной работе будут рассматриваться только конечные графы, т.е. множества вершин и ребер принимают конечное число значений;
- количество рёбер графа называется его размером и обозначается $e = |E(G)|$;
- ребро $e = \{u, v\}$ соединяет вершины, называющиеся концевыми (концами) u и v;
- соседними вершинами называются такие концы, которые соединены одним и тем же ребром;
- смежными называются ребра, имеющие общую концевую вершину;
- вершина называется изолированной, если она не является конечной ни для одного из ребер;
- висячей вершиной или листом называется вершина, которая является концом ровно одного ребра.

Если не говорится об обратном, то граф считается неориентированным (пример приведен на рис 1.1), это означает, что каждое его ребро имеет два конца, порядок которых не имеет значения.

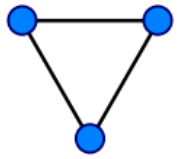

Рис. 1.1. Неориентированный граф

Пусть $(v,w)\in E$ называется дугой. Тогда вершину v называют её началом, а w — концом. Можно сказать, что дуга $v\to w$ ведёт от вершины v к вершине w.

Определение 1.2. Граф называется ориентированным (орграфом), если $G=(V(G),E(G))$ — упорядоченная пара, где $V(G)$— непустое множество вершин (узлов) графа G, а $E(G)$ — множество упорядоченных пар различных вершин, которые называются дугами. Пример такого графа приведен на рис. 1.2.

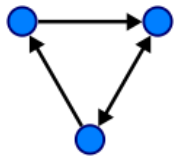

Рис. 1.2. Ориентированнный граф

Вспомогательные определения для орграфов:

- маршрутом в графе называется конечная последовательность вершин, в которой каждая вершина (исключая последнюю) соединяется со следующей в последовательности вершиной ребром;
- путем в графе называют конечную последовательность вершин и дуг, в которой каждый элемент соединен с предыдущем и последующим;
- цепью называется маршрут без повторяющихся ребер;
- циклом называют цепь, в которой начальная и конечная вершины совпадают;
- цикл называют простым, если ребра в нем не повторяются;

- цикл называют элементарным, если он простой и вершины в нем не повторяются.

Определение 1.3. Граф H называется подграфом графа G, если $V(H) \subset V(G)$ и $E(H) \subset E(G)$.

Определение 1.4. 1) Вершины a и b графа G называются связанными, если в графе существует путь между ними.

2) Граф называется связным, если любые две его вершины связаны.

Существует много несколько разновидностей графов. В данной работе будет использоваться такие структуры как дерево и лес.

Определение 1.5. 1) Деревом называется связный граф без циклов.

2) Лесом называется упорядоченное множество деревьев.

Определение 1.6. Степенью вершины называется количество инцидентных ей ребер.

Определение 1.7. Двоичным (бинарным) деревом называется:

1. неориентированное дерево, степени вершин которого не превосходят 3;
2. ориентированное дерево, в котором число исходящих из каждой вершины ребер не превосходит 2.

Вершина, из которой выходят ребра называется родительской. А связанные с ней ребрами вершины называются левой и правой дочерней.

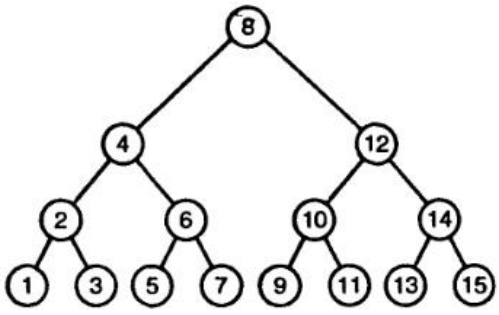

Рис. 1.3. Пример неориентированного бинарного дерева

1.1.2 Простые алгоритмы на графах

Добавлено примечание ([ДС2]): Взято из кормена

Добавлено примечание ([ДС3]): Информация взята из Кормена

В этом разделе будет приведено несколько основных алгоритмов обхода графов. Обходя граф, мы двигаемся по ребрам и проходим все вершины. При этом накапливается довольно много информации, которая полезна для дальнейшей обработки графа.

1.1.2.1 Поиск в ширину

Поиск в ширину (breadth-first search) – один из базисных алгоритмов, составляющий основу многих других.

Пусть задан граф $G = (V(G), E(G))$ и фиксирована начальная вершина *s*. Алгоритм поиска в ширину перечисляет все достижимые из *s* вершины, доступные при проходе по ребрам, в порядке возрастания расстояния от *s*. Расстоянием считается длина минимального пути из начальной вершины. Алгоритм применим как к ориентированным графам, так и к неориентированным.

Такое название объясняется тем, что в процессе поиска мы идем вширь, а не вглубь, т.е. сначала просматриваем все соседние вершины, затем соседей соседей и так далее.

Алгоритм 1.1.

0. Начинаем обход из фиксированной начальной вершины s. Пометить ее как посещенную. Добавить вершину в изначально пустую очередь.
1. Извлечь из начала очереди вершину u.
2. Если она является уже посещенной повторить шаг 1. Иначе добавить все соединенные с u не посещённые вершины в очередь и перейти к шагу 1.
3. Если очередь пустая, закончить алгоритм. В противном случае перейти к шагу 1.

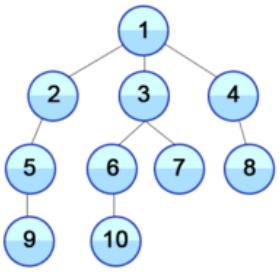

Рис. 1.4. Порядок обхода графа при поиске в ширину

На рис. 1.4. приведен пример работы алгоритма поиска в ширину. Цифры являются порядковым номером посещения вершины в процессе работы алгоритма.

1.1.2.2 Поиск в глубину

Поиск в глубину (depth-first search) наряду с вышеописанным алгоритмом также является одним из базисных методов обхода графа.

Он имеет следующую стратегию: как и в поиске в ширину, фиксируем начальную вершину *s* и начинаем от нее идти «вглубь», пока имеется такая возможность, т.е. пока существуют не пройдённые ребра, затем возвращаться и искать иной путь, в случае, когда таких ребер не осталось. Алгоритм работает, пока не обнаружит все вершины, достижимые из исходной.

Алгоритм 1.2.

0. Начинаем обход из фиксированной начальной вершины *s*.
1. Пометить текущую вершину как посещенную.
2. Если есть соседние не посещенные вершины, перейти к одной из них и выполнить для нее алгоритм начиная с шага 1. Если все соседние вершины посещены, либо отсутствуют вовсе – закончить алгоритм.

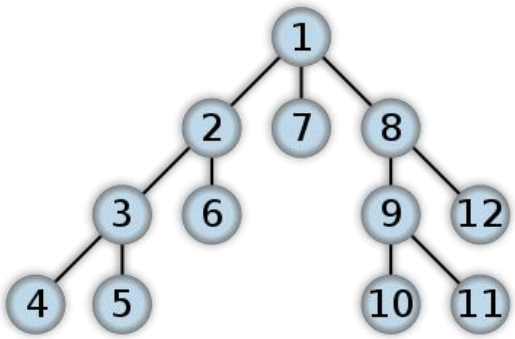

Рис. 1.5. Порядок обхода графа при поиске в глубину

На рис. 1.5. представлен пример работы алгоритма поиска в глубину. Аналогично рис. 1.4. цифры соответствуют порядковому номеру вершины при обходе графа алгоритмом 1.2.

1.2 Классификаторы

Классификацией называют один из разделов машинного обучения, посвященный решению следующей задачи. Имеется множество объектов (ситуаций), разделённых некоторым образом на классы. Задано конечное множество объектов, для которых известно, к каким классам они относятся. Классовая принадлежность остальных объектов неизвестна. Требуется построить алгоритм, способный классифицировать произвольный объект из исходного множества.

Классифицировать объект — значит, указать номер (или наименование) класса, к которому относится данный объект.

Классификация объекта — номер или наименование класса, выдаваемый алгоритмом классификации в результате его применения к данному конкретному объекту.

Существуют несколько типов классов:

- двухклассовый, число классов равно двум;
- многоклассовый, когда число классов достигает многих тысяч (при распознавании иероглифов или слитной речи).
- непересекающиеся классы.
- пересекающиеся классы. Объект может относиться одновременно к нескольким классам.
- нечёткие классы. Требуется определять степень принадлежности объекта каждому из классов, обычно это действительное число от 0 до 1.

В данной работе будет рассматриваться двухклассовый случай.

Введем некоторые базовые определения, которые будут использованы в дальнейшем.

Определение 1.8. Признаком (feature) называется результат измерения некоторой характеристики объекта. Можно сказать, что признак есть отображение $f : X \to D_f$, где D_f – множество допустимых значений признака.

В зависимости от природы этого множества, признаки делятся на нижеперечисленные типы:

- бинарный признак $D_f = \{0,1\}$;
- номинальный признак D_f — конечное множество;
- порядковый признак: D_f — конечное упорядоченное множество;
- количественный признак: $D_f = \mathbb{R}$.

Если все признаки имею одинаковый тип, то исходные данные называются однородными, в ином случае – разнородными.

Определение 1.9. Пусть имеется набор признаков $f_1,\ldots, f_n$. Признаковым описанием объекта $x \in X$ называют вектор $((f_1(x),\ldots, f_n(x))$, составленный из значений фиксированного набора признаков на данном объекте.

В задачах машинного обучения не делается различия между объектом и его признаковым описанием. Полагается, что $X = D_{f_1} \times \ldots \times D_{f_n}$

Постановка задачи классификации выглядит следующим образом. Пусть X — множество описаний (признаков) объектов. Чем является объект, определяется спецификой предметной области. Например, в задачах спортивного менеджмента объектами являются спортсмены.

Пусть Y — конечное множество номеров (имён, меток) классов. Существует неизвестная целевая зависимость — отображение $y^*: X \to Y$, значения которой известны только на объектах конечной обучающей выборки $X^m = \{(x_1, y_1),\ldots,(x_m, y_m)\}$. Требуется построить алгоритм $a(x): X \to Y$, способный классифицировать произвольный объект $x \in X$.

1.2.1 Нормализация входных данных

Зачастую, чтобы достичь адекватности работы модели, необходимо нормализовать (масштабировать) входные данные. Как будет видно далее работоспособность некоторых моделей зависит от расстояния между объектами, вследствие чего возникает необходимость проведения данной

процедуры. Проблема заключается в разных измерениях признаков. Например, если рассматривать погоду, то такие ее признаки как температура, давление, скорость ветра и т.д. измеряются в различных физических величинах, а их числовые значения могут на порядки отличаться.

Нормализовать данные можно разными способами, вот два основных.

Минимаксная нормализация:

$$x_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}. \quad (1.1)$$

Данный метод осуществляет переход от абсолютных значений признаков к относительным. Новые переменные будут принимать значения в диапазоне от 0 до 1.

Z-нормализация:

$$x_i = \frac{x_i - \bar{x}}{s}, \quad (1.2)$$

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ – выборочное среднее, $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ – выборочное среднеквадратичное отклонение.

Поскольку не все признаки имеют количественные значения, может применяться создание фиктивных переменных (dummy coding). В этом случае категориальные признаки заменяются бинарными. Например, заменяется признак «пол» на два новых: «пол мужской», «пол женский» со значения 0 и 1.

1.2.2 Метод k-ближайших соседей

Алгоритмы, основанные на анализе сходства объектов, часто называют метрическими.

Метрическим классификатором (similarity-based classifier) называют алгоритм классификации, основанный на вычислении оценок сходства между объектами. Чтобы формализовать понятие сходства вводится функция расстояния между объектами $\rho(x, x')$ в пространстве объектов X. Следует заметить, что данная функция может не всегда удовлетворят всем аксиомам метрики. Например, довольно часто не выполняется неравенство треугольника.

Метрические классификаторы опираются на гипотезу компактности. Она, в свою очередь, предполагает, что схожие объекты гораздо чаще лежат в одном классе, чем в разных. Можно сказать, что классы образуют компактно локализованные подмножества в пространстве объектов. То есть граница между классами имеет довольно простую форму.

Метод ближайшего соседа позиционируется, как один из простейших метрических классификаторов. Классифицируемый объект x относится к тому классу y_i, которому принадлежат ближайший к нему объект обучающей выборки x_i.

Метод k ближайших соседей (k-nearest neighbors algorithm, k-NN) для повышения надёжности классификации относит объект к тому классу, которому принадлежит большинство из его соседей, то есть k ближайших к нему объектов обучающей выборки x_i. В задачах с двумя классами число соседей берут нечётным, чтобы не возникало ситуаций неоднозначности, когда одинаковое число соседей принадлежат разным классам.

Пусть задана обучающая выборка $X^m = \{(x_1, y_1), ..., (x_m, y_m)\}$ и на множестве объектов задана функция расстояния $\rho(x, x')$. Данная функция должна быть достаточно адекватной моделью сходства объектов, для этого можно провести

процедуру нормализации, описанную в разделе 1.2.1. Чем больше значение этой функции, тем менее схожими являются два объекта x, x'.

Для произвольного объекта u расположим объекты обучающей выборки x_i в порядке возрастания расстояний до u: $\rho(u, x_{1;u}) \le \rho(u, x_{2;u}) \le \ldots \rho(u, x_{m;u})$, где $x_{i;u}$ обозначает объект обучающей выборки, который является i-ым соседом объекта u. Аналогичное обозначение введём и для ответа на i-ом соседе: $y_{i;u}$. Таким образом, произвольный объект u порождает свою перенумерацию выборки. В наиболее общем виде алгоритм ближайших соседей выглядит так:

$$a(u) = \arg\max_{y \in Y} \sum_{i=1}^{m} [y(x_{i;u}) = y] w(i,u), \qquad (1.3)$$

где $w(i,u)$ – заданная весовая функция, которая оценивает степень важности i-го соседа для классификации объекта u. Эта функция неотрицательна и не возрастает по i.

Различно задавая весовую функцию, получаются различные варианты методы ближайших соседей.

- $w(i,u) = [i = 1]$ – простейший метод ближайшего соседа;
- $w(i,u) = [i \le k]$ – метод k ближайших соседей;
- $w(i,u) = [i \le k] q^i$ – метод k экспоненциально взвешенных ближайших соседей, где предполагается $q < 1$ (обычно используется в случае 3-х и более классов).

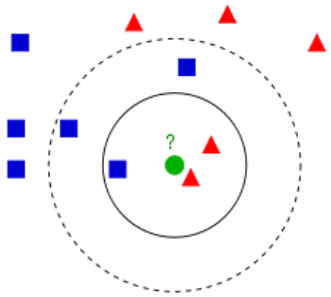

Рис. 1.6. Метод k ближайших соседей

На рис. 1.6. приведена ситуация классификации объекта, в данном случае зеленого круга. Круг должен быть классифицировать как синий квадрат, либо как красный треугольник (класс 1 и класс 2 соответственно). На иллюстрации видно два круга.

Круг, обведенный сплошной линией, показывает поведение алгоритма при $k=3$. В этом случае объект будет классифицирован как 2-ой класс, так как внутри круга находятся 2 треугольник и 1 квадрат, треугольников больше, а значит и решение принимается в сторону этого класса.

В кругу, обведенном штрихом, $k=5$. Тогда ситуация меняется, потому что количество квадратов начало превалировать над треугольниками. Соответственно и объект будет классифицирован как синий квадрат, то есть 1-й класс.

1.2.3 Дерево решений

Решающими деревьями называется семейство моделей, которые позволяют восстанавливать нелинейные зависимости произвольной сложности. Они воспроизводят логические схемы, позволяющие получить окончательное решение о классификации объекта с помощью ответов на иерархически организованную систему вопросов. Причем вопрос, задаваемый на последующем иерархическом уровне, зависит от ответа, полученного на предыдущем уровне.

Добавлено примечание ([ДС4]):

Добавлено примечание ([ДС5R4]):

Каждой из вершин дерева за исключением листьев соответствует некоторый вопрос, подразумевающий несколько вариантов ответов, соответствующих выходящим ребрам. В зависимости от выбранного варианта ответа осуществляется переход к вершине следующего уровня. Листьям поставлены в соответствие метки, указывающие на отнесение распознаваемого объекта к одному из классов.

Решающее дерево называется бинарным, если каждая внутренняя или корневая вершина инцидентна только двум выходящим рёбрам.

Определение 1.10. Рассмотрим бинарное дерево, в котором каждой внутренней вершине v приписана функция (или предикат) $\beta_v : X \to \{0,1\}$, а каждому листу v приписан прогноз $c_v \in Y$. В случае классификации листу может быть приписан вектор вероятностей. При классификации объекта $x \in X$ он проходит путь от корня дерева до некоторой концевой вершины, в соответствии с алгоритмом *a(x)*.

Теперь рассмотрим алгоритм *a(x)*, который стартует из корневой вершины v_0 и вычисляет значение функции β_{v_0}.

Алгоритм 1.3.

0. Алгоритм начинается с начальной вершины v_0.
1. Далее вычисляется значение функции β_{v_0}.
2. Если текущая вершина является листом, вернуть класс, который приписан данной вершине и закончить алгоритм.

3. Если вычисленное значение равно нулю, то алгоритм переходит в левую дочернюю, в ином случае в правую. Для новой вершины выполняем шаг 1.

Такой алгоритм называется бинарным решающим деревом.

На практике в большинстве случаев используются одномерные предикаты β_v, которые сравнивают значение одного из признаков с порогом: $\beta_v(x; j, t) = [x_j < t]$.

Объект x доходит до вершины v тогда и только тогда, когда выполняется конъюнкция $K_v(x)$, составленная из всех предикатов, приписанных внутренним вершинам дерева на пути от корня v_0 до вершины v.

Пусть T — множество всех терминальных вершин дерева. Множества объектов $\Omega_v = \{x \in X : K_v(x) = 1\}$, выделяемых терминальными конъюнкциями $v \in T$, попарно не пересекаются, а их объединение совпадает со всем пространством X. Данное утверждение легко доказывается индукцией по числу вершин дерева. Отсюда следует, что решающее дерево никогда не отказывается от классификации. А также, что алгоритм классификации $a(x): X \to Y$, реализуемый бинарным решающим деревом, можно представить в виде простого голосования конъюнкций:

$$a(x) = \arg\max_{y \in Y} \sum_{v \in T} [c_v = y] K_v(x), \qquad (1.4)$$

причем для любого $x \in X$ одно и только одно слагаемое во всех этих суммах равно единице.

Легко увидеть, что для любой выборки можно реализовать решающее дерево, которое не допустит на ней ни одной ошибки. Даже с простыми одномерными предикатами можно сформировать дерево, в каждом листе которого находится ровно по одному объекту выборки. Вероятнее всего, такое дерево будет переобученным и не сможет показать хорошее качество классификации на новых данных.

Добавлено примечание ([D6]): Не уверен, что это нужно

1.2.3.1 Построение дерева

Опишем базовый жадный алгоритм построения бинарного решающего дерева.

Алгоритм 1.4

0. В качестве начальной возьмем всю обучающую выборку X найдем наилучшее ее разбиение на две части $R_1(j,t)=\{x \mid x_j < t\}$ и $R_2(j,t)=\{x \mid x_j \geq t\}$ с точки зрения заранее заданного функционала качества $Q(x, j, t)$.
1. Найдем наилучшие значения *j* и *t*, создадим корневую вершину дерева v_0 и примем ее за текущую.
2. В текущей вершине проверяем, не выполнилось ли некоторое условие останова. Если выполнилось, то прекращаем рекурсию и объявляем эту вершину листом. Переходим к шагу 5. Если же не выполнилось, то к шагу 3.
3. Поставим в соответствие текущей вершине предикат $[x_j < t]$. Объекты разобьются на две части — одни попадут в левое поддерево, другие в правое.
4. Для каждой из подвыборок, получившихся на шаге 3, рекурсивно повторим шаг 1, построив дочерние вершины для корневой, и так далее.
5. В построенном дереве каждому листу приписывают ответ. В случае классификации – класс, к которому больше всего относится объектов в листе или вектор вероятности. Алгоритм завершен.

Выбор конкретной функции зависит от функционала качества в исходной задаче. Таким образом, конкретный метод построения решающего дерева определяется:

- видом предикатов в вершинах;
- функционалом качества $Q(x, j, t)$;

- критерием останова.

Далее подробнее рассмотрим каждый пункт из списка выше.

При построении дерева необходимо задать функционал качества, на основе которого будет осуществляться разбиение выборки на каждом шаге. Пусть R_m – множество объектов, которые попали в вершину, разбиваемую на текущем шаге, а R_l и R_r – подмножества R_m, состоящие из объектов, попадающих при заданном предикате в левое и правое поддерево соответственно. Будет использован функционал следующего вида:

$$Q(R_m, j, s) = H(R_m) - \frac{|R_l|}{|R_m|} H(R_l) - \frac{|R_r|}{|R_m|} H(R_r), \quad (1.5)$$

где $H(R)$ есть критерий информативности (impurity criterion), который оценивает качество распределения целевой переменной среди объектов множества R. Чем меньше разнообразие целевой переменной, тем меньше должно быть значение критерия информативности — и, соответственно, мы будем пытаться минимизировать его значение. Функционал качества $Q(R_m, j, s)$ мы при этом будем максимизировать.

Как уже обсуждалось выше, в каждом листе дерево будет выдавать константу — вещественное число, вероятность или класс. Исходя из этого, можно предложить оценивать качество множества объектов R тем, насколько хорошо их целевые переменные предсказываются константой (при оптимальном выборе этой константы):

$$H(R) = \min_{c \in Y} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} L(y_i, c) \quad (1.6)$$

где $L(y_i, c)$ — некоторая функция потерь. Далее мы обсудим, какие именно критерии информативности часто используют в задачах классификации.

1.2.3.2 Критерии информативности

Обозначим через p_k долю объектов класса k $(k \in \{1, ..., K\})$, попавших в вершину *R:*

$$p_k = \frac{1}{|R|} \sum_{(x_i, y_i) \in R} [y_i = k], \tag{1.7}$$

Через k_* обозначим класс, представителей которого нашлось больше остальных среди объектов, попавших в данную вершину:

$$k_* = \arg\max_k p_k . \tag{1.8}$$

Приведем в пример два критерия информативности.

Первым будет ошибка классификации. Рассмотрим индикатор ошибки, как функцию потерь. Подставим в (1.6):

$$H(R) = \min_{c \in Y} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} [y_i \neq c], \tag{1.9}$$

Заметно, что оптимальным предсказанием тут будет наиболее популярный класс k_*. Это означает, что критерий будет равен следующей доле ошибок:

$$H(R) = \min_{c \in Y} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} [y_i \neq k_*] = 1 - p_{k_*}, \tag{1.10}$$

Данный критерий считается довольно грубым, так как учитывает частоту p_{k_*} лишь одного класса.

Вторым обозначим Критерий Джини. Рассмотрим ситуацию, когда в вершине выдается не один класс, а распределение по всем классам $c = (c_1, \ldots, c_n), \sum_{k=1}^{K} c_k = 1$. Качество такого распределения можно измерять с помощью критерия Бриера (Brier score):

$$H(R) = \min_{\sum_k c_k = 1} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} \sum_{k=1}^{K} (c_k - [y_i = k])^2, \tag{1.11}$$

Можно показать, что оптимальный вектор вероятностей состоит из долей классов p_k: $c_* = (p_1, \ldots, p_K)$.

Если подставить эти вероятности в исходный критерий информативности (1.6), провести ряд преобразований, то на выходе получим критерий Джини:

$$H(R = \sum_{k=1}^{K} p_k(1 - p_k). \quad (1.12)$$

1.2.3.3 Критерии останова

Можно придумать большое количестве критериев останова. Перечислим некоторые ограничения и критерии:

- Ограничение максимальной глубины дерева.
- Ограничение минимального числа объектов в листе.
- Ограничение максимального количества листьев в дереве.
- Останов в случае, если все объекты в листе относятся к одному классу.
- Требование, что функционал качества при дроблении улучшался как минимум на s процентов.

С помощью грамотного выбора подобных критериев и их параметров можно существенно повлиять на качество дерева.

1.2.4 Случайный лес

В основе алгоритма случайный лес лежит идея использования ансамбля (комитета) решающих деревьев. Сами по себе входящие в ансамбль решающие деревья дают невысокий результат классификации, но за счет большого их количества, решение, принимаемое голосованием, получается достаточно хорошим.

Зададим обучающую выборку, количество образцов в которой равно N. Пусть размерность пространства признаков будет равна M. Параметр $m \approx \sqrt{M}$ (такое значение используется в задачах классификации) равен количеству неполных признаков, используемых для обучения.

Рассмотрим наиболее распространенный способ построения отдельных экземпляров деревьев для случайного леса.

Алгоритм 1.5.

1. Генерируется случайная подвыборку с возвращением размером n. Некоторые образцы попадут в выборку более одного раза, а некоторые не попадут вовсе.
2. Формируется решающее дерево в соответствии с алгоритмом 1.4. При создании очередного узла дерева случайным образом выбирается m признаков, на основе которых будет производится разбиение. Для выбора лучшего из этих признаков используется критерий Джини.
3. Дерево строится до срабатывания выбранного критерия останова, которые были описаны в разделе 1.2.2. Чаще всего этим критерием является полное исчерпание подвыборки.

Данный алгоритм повторяется K (количество деревьев в лесу) раз. Получившийся лес классифицирует объекты путём голосования. Процесс голосования выглядит следующим образом: каждое дерево в комитете относит объект к одному из классов. Побеждает класс, которой набрал наибольшее количество голосов.

1.3 Метрики качества

После того как модель построена и дает результаты, возникает вопрос о качестве работы данной модели и правдоподобности результатов. Прежде чем говорить о подборе той или иной метрики для оценки качества классификатора, необходимо ввести понятие матрицы ошибок.

Пусть дана обучающая выборка, над ней произведена процедура бинарной классификации, в результате которой получен вектор ответов $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n)$. Поскольку выборка обучающая, то вектор ожидаемых результатов $y = (y_1, \ldots, y_n)$ известен. Класс объектов равных 1, будем называть C_1, а объектов равных 0 – C_0.

Тогда матрица ошибок будет выглядеть следующим образом, который показан в таблице 1.1.

Таблица 1.1 Матрица ошибок

	$\hat{y}_i = 0$	$\hat{y}_i = 1$
$y_i = 0$	*TN*	*FP*
$y_i = 1$	*FN*	*TP*

Получившиеся на пересечении значения означают:

- *TN* – количество верно предсказанных классификатором объектов класса C_0.
- *TP* – количество верно предсказанных классификатором объектов класса C_1.
- *FP* – количество ошибок первого рода, которые означают, что классификатор определил класс C_0, как класс C_1.
- *FP* – количество ошибок второго рода, которые означают, что классификатор определил класс C_1, как класс C_0.

Первой очевидной метрикой, которая интуитивно приходит на ум является доля правильных ответов:

$$accuracy = \frac{TP + TN}{TP + TN + FN + FP}. \quad (1.13)$$

На практике такая метрика очень редко используется, потому что является бесполезной в задачах, где количество объектов одного класса сильно превалирует на другим.

Далее введем метрики для оценки работы модели на каждом из классов по отдельности. Первой будет точность (precision):

$$precision = \frac{TP}{TP + FP}, \quad (1.14)$$

Она показывает долю объектов, определенных классификатором как класс C_1, которые в действительности относятся к классу C_1.

Второй метрикой будет полнота (recall):

$$recall = \frac{TP}{TP + FN}, \quad (1.15)$$

она показывает, какую долю объектов класса C_1 из всех его объектов нашел классификатор.

Объединяет последние две метрики F_1-мера. Которая является средним гармоническим точности и полноты:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (1.16)$$

Очевидно, что F_1-мера будет максимальной при максимальных значениях полноты и точности, и будет близка к нулю, если один из аргументов близок к нулю.

2 ПРАКТИЧЕСКАЯ ЧАСТЬ

Практическим результатом работы является обученная модель, способная определять потенциально мошеннические транзакции. Также была спроектировано прикладное программное обеспечение, позволяющее визуализировать в виде ориентированных графов взаимодействие клиентов посредством финансовых транзакций.

2.1 Постановка задачи

Ставится задача анализа финансовых транзакций с целью предотвращения мошеннических операций. Данная задача разбивается на две подзадачи:

1. Реализовать классификатор способный отличить мошеннические транзакции от не являющихся таковыми.
2. Построить графы на основе финансовых переводов. На графах посчитать новые параметры (расстояние до ближайшего мошенника, количество мошенников в круге радиуса n), добавить их в модель и проверить значимость.

2.1.1 Описание данных

Данные, на которых будут проводится исследования, представляют собой файл с расширением *.csv*, в котором содержится информация о 6362620 финансовых транзакциях. Мошеннические транзакции в нем помечены. Следует отметить, что этот набор данных сгенерирован синтетически (по соображениям конфиденциальности) на основе реального, предоставленного африканским сервисом мобильных переводов. Временное окно данных – 1 месяц. Для отслеживания времени введена условная переменная, равная 1 часу.

В таблице 2.1 приведены признаки, которые заданы для каждой транзакции в выборке.

Таблица 2.1 Описание признаков

Название в *.csv* файле	**Смысл**	**Принимаемые значения**
step	Аналог времени (1 step = 1 час)	1–744 (30 дней)
type	тип транзакции	CASH-IN, CASH-OUT, DEBIT, PAYMENT, TRANSFER
amount	сумма перевода	Действительное число больше нуля
nameOrig	ID пользователя-отправителя	ID в формате *‘C165564’*
oldbalanceOrg	баланс отправителя до транзакции	Действительное неотрицательное число
newbalanceOrig	баланс отправителя после транзакции	Действительное неотрицательное число
nameDest	ID пользователя-получателя	ID в формате *‘C165564’* (магазин, если первая буква ID – *M)*
oldbalanceDest	баланс получателя до транзакции	Действительное неотрицательное число
newbalanceDest	баланс получателя после транзакции	Действительное неотрицательное число
isFraud	пометка о мошеннической транзакции	1 – мошенническая 0 – в противном случае

Пример строки в файле:

184,CASH_IN,90687.53,C594792269,14620285.23,14710972.76,C1485172763,166355.91,75668.37,0,0,16,1,1,0,0,0,0

2.2 Предварительный анализ

Прежде, чем перейти к построению модели, было принято решение посчитать (увидеть) некоторые очевидные вещи, которые лежат на поверхности.

В результате была получена информация, приведенная в таблице 2.2.

Таблица 2.2 Данные, полученные при предварительном анализе

Название	**Значение**
Всего транзакций	6362620
Мошеннических	8213
Средняя сумма перевода мошеннических транзакций	1467967.29
Средняя сумма перевода не мошеннических транзакций	178197.04
Уникальных клиентов	4777844
Количество магазинов	2151495
Уникальных магазинов	2150401
Клиенты, взаимодействующие друг с другом более 1 раза	0
Мошеннических транзакций с магазинами (из них мошеннических)	0
Количество типов CASH-IN	1399284 (0)
Количество типов CASH-OUT	2237500 (4116)
Количество типов DEBIT	41432 (0)
Количество типов PAYMENT	2151495 (0)
Количество типов TRANSFER	532909 (4097)

Анализируя полученные данные, можно сделать выводы:

- присутствует сильный дисбаланс между классом мошеннических и не мошеннических транзакций. Были принято решение сократить выборку до ~500.000 транзакций, причем количество мошеннических оставить равным исходному. Преимущество этого шага еще и в том, что теперь для обучения модели будет требоваться меньше памяти. Факт нехватки памяти сильно тормозил (иногда и вовсе не давал посчитать результаты) процесс исследования;
- средняя сумма мошеннических транзакций на порядок выше не мошеннических. Значит сумма является важным критерием для классификации в этом наборе данных;
- количество уникальных клиентов достаточно велико. Данный факт говорит о том, что взаимосвязь между клиентами весьма слабая;
- неправдоподобным выглядит отсутствие клиентов, взаимодействующих друг с другом более 1 раза. В жизненных условиях всегда найдутся клиенты, активно переводящие друг другу средства. Скорее всего этот недостаток, обусловлен синтетической природой данной выборки;
- уникальных магазинов также очень много, почти столько же, сколько и всего транзакций, производимых с их участием. Количество магазинов совпадает с количеством транзакций с типом PAYMENT. А также мошеннические операции не проводятся в рамках контактов с магазинами;
- мошеннические транзакции производятся в пределах двух типов: CASH-OUT и TRANSFER. Наряду с суммой перевода, этот факт будет играть одну из ключевых ролей при классификации.

2.3 Построение графов

Для построения и визуализации графов использовалась интегрированная среда разработки Visual Studio 2017. Прикладная программа была разработана на языке C# при помощи технологии WPF (Windows Presentation Foundation).

2.3.1 Структура программы

Необходимо перевести данные из построчного представления в *.csv* файле в объектный вид, чтобы использовать преимущества объектно-ориентированной разработки.

Для этого были реализованы такие классы:

- *Transaction* –данный класс является объектным представлением одной строки исходного файла. В нем абсолютно те же поля, что и набор признаков в выборке. Но вместо строчного представления имени клиента в нем хранится класс *Client* как для отправителя, так и для получателя.
- *Client* – в данном классе хранится ID клиента, список его транзакций в качестве отправителя, а также в качестве получателя, пометка о том, был ли замечен этот клиент в мошеннических переводах. Также тут может содержаться дополнительная информация, например, максимальная длина цепочки, в которой замечен клиент, и расстояние до ближайшего мошенника.

При запуске программы происходит чтение данных, преобразование их к объектному виду. В результате этого формируется список клиентов, список транзакций и формируется граф на основе словаря.

Можно упрощенно описать работу программы алгоритмом 2.1.

Алгоритм 2.1.

Перед началом инициализируем пустой список транзакций и клиентов.

1. Читается каждая строга исходного файла. Из нее создаются экземпляры классов транзакции и клиента, которые добавляются в свои списки соответственно.

2. Итеративно продвигаясь по всему списку транзакций формируется словарь, в котором ключом выступает клиент, а значением – список клиентов, с которыми он взаимодействует.
3. На сформировавшемся на шаге 2 графе, итеративно двигаясь по ключам, с помощью алгоритма 1.2 находим вспомогательные признаки (максимальная цепочка, в которой участвует клиент и расстояние до ближайшего мошенника).

2.3.2 Описание интерфейса и возможностей

После обработки данных, описанной в алгоритме 2.1 перед пользователем появляется интерфейсное окно с визуализированным графом (рис. 2.1).

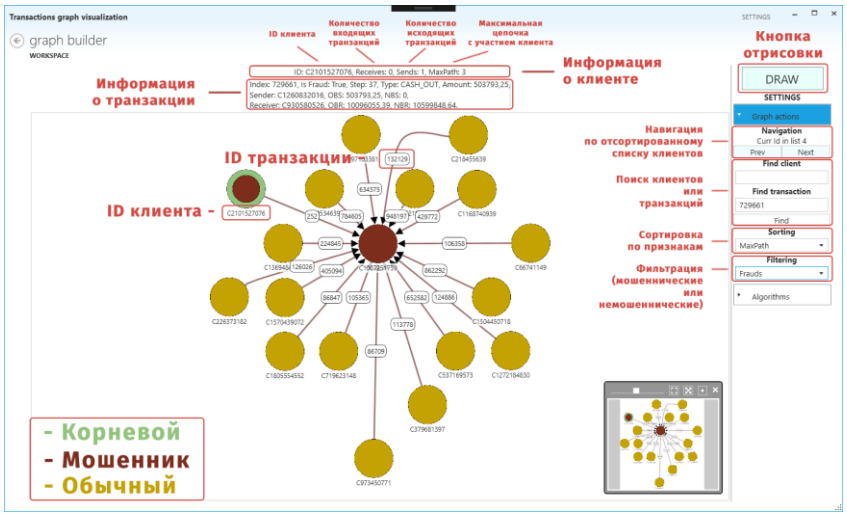

Рис. 2.1 Интерейсное окно прикладной программы

В центральной части окна располагается поле визуализации графа. Зеленым цветом помечен клиент, с которого началось построение графа. Будем называть таких клиентов корневыми. Желтым цветом помечены обычные клиенты, не замеченные в мошеннических транзакциях. Красным помечены мошенники. Если клиент является одновременно и корневым и

мошенником, то тогда он отмечается красно-зеленым цветом. Строка под кругом – ID клиента.

Направление стрелок на визуализации указывает от отправителя к получателю. Цифры посередине стрелки означают ID транзакции, который соответствует номеру строки в исходном файле.

В верхней части окна выводится информация по выбранному клиенту и транзакции. Про клиента пишется ID, количество транзакций в роли получателя, количество транзакций в роли отправителя и максимальная цепочка, в которой он был замечен. Про транзакцию также пишется ее ID и вся информация, которая доступна из строки исходного файла, то есть все признаки.

В правой части окна располагается интерфейс взаимодействия. В нем можно изменить сортировку клиентов, отфильтровать их, найти клиента или транзакцию по ID. Все изменения вступают в силу после нажатия кнопки «DRAW».

По умолчанию клиенты отсортированы по убыванию максимальной цепочки транзакций. Можно также отсортировать по числу выступлений получателем или отправителем, а также по расстоянию до ближайшего мошенника.

Фильтрация по умолчанию отключена. Но можно отфильтровать клиентов, оставив только мошенников.

2.3.3 Анализ результатов

Построив граф на отфильтрованных данных (как было предложено в разделе 2.2). Оказалось, что длина самой длиной цепи равна четырем. Предположения о разреженности графа из раздела 2.2 подтвердились. На раннем этапе исследования на полном наборе данных максимальная длина цепи была равна 9.

Соответственно предложенные для добавления в модель признаки из постановки задачи (раздел 2.1.1) не имеют смысла. Дистанция до ближайших мошенников почти у всех клиентов отсутствует, либо равна 1 или 0 (случай, когда сам клиент мошенник). Максимальной зафиксированной дистанцией была цепь длиной в 2 транзакции. Искать количество мошенников в круге радиуса *n*, также не принесет особых результатов, поскольку общая картина такова: граф разбит на кучу маленьких подграфов. Один из таких подграфов представлен на рис 2.2.

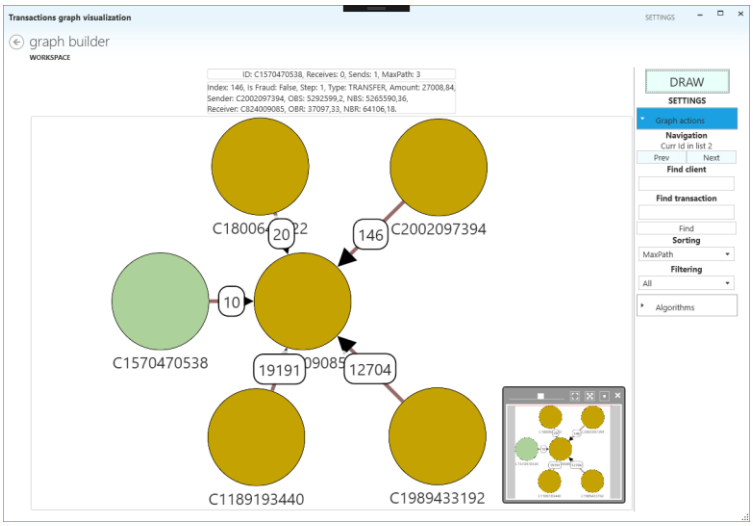

Рис. 2.2 Типичная ситуация для отдельно взятого клиента

Визуально анализируя полученные результаты, удалось установить, что типичная картина мошеннических клиентов выглядит следующим образом: клиент выступает получателем, к которому стекаются несколько транзакций, одна из которых мошенническая (рис. 2.3).

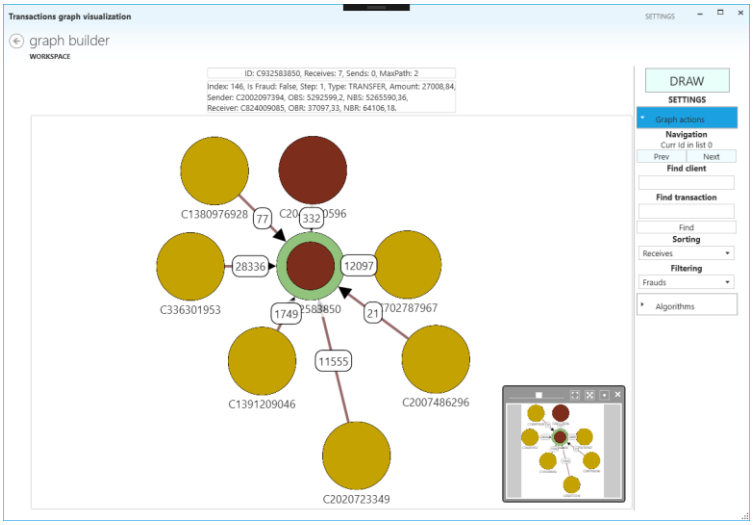

Рис. 2.3 Типичная ситуация для мошенника

Признаки, выявленные на графах, показали свою практическую неприменимость. Поэтому следует уделить большее внимание задаче классификации.

2.4 Проектирование признаков

Проектирование признаков (feature engineering) производится с целью повысить эффективность прогнозирования алгоритмов обучения путем выявления новых характеристик или исключения имеющихся из исходных данных. Данная процедура позволяет упростить процесс обучения.

Для сравнения в список используемых моделей (помимо описанных в разделах 1.2.2 и 1.2.4) добавляется наивный байесовский классификатор. Он основан на применении теоремы Байеса со строгими предположениями о том, что все объекты описываются независимыми признаками.

Проведение экспериментов происходит при помощи языка Python и библиотеки scikit-learn, внутри которой реализация моделей, а также методы расчета метрик, описанных в разделе 1.3.

Отправной точкой будет результат обучения модели на «сырых» данных, то есть никаких добавлений или исключений признаков. Результаты такого эксперимента приведены в таблице 2.3.

Таблица 2.3 Метрики качества на «сырых» данных

	precision	*recall*	F_1
Наивный байесовский	0.595	0.204	0.304
k-ближайших соседей	0.954	0.794	0.867
Случайный лес	0.994	0.831	0.905

Лучше всего себя повел случайный лес и уже показал неплохие результаты. Недавно добавленный наивный байесовский классификатор довольно плохо предсказал результат. А метод ближайших соседей повел себя уверенно, но все же уступает случайному лесу.

Далее производится отбор признаков, затем эксперимент будет еще раз проведен на новом наборе данных.

2.4.1 Отбор признаков

Для устранения сильно коррелирующих между друг другом признаков строится матрица корреляции (рис. 2.4).

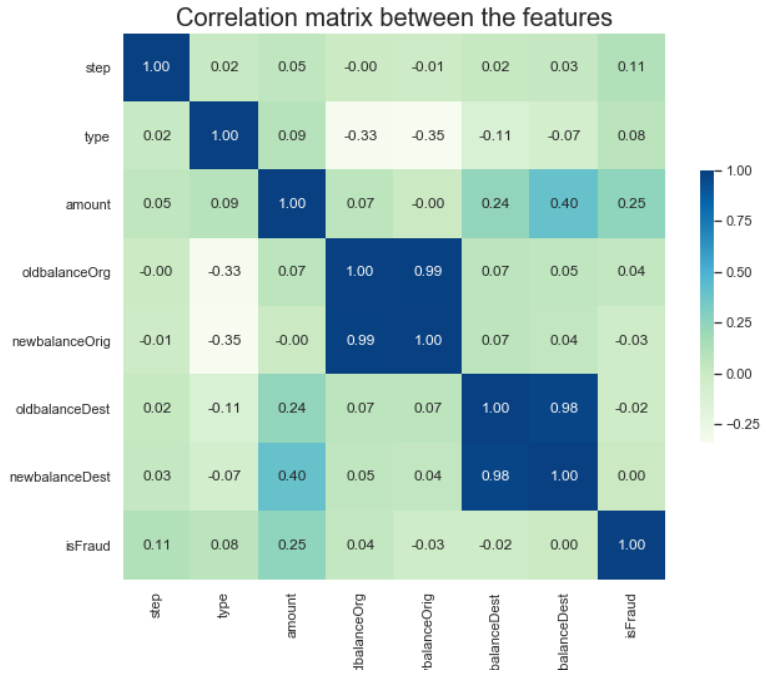

Рис. 2.4 Матрица корреляции исходных признаков

Очевидно, что балансы до и после совершения транзакции будут сильно коррелировать, поэтому можно отбросить балансы после транзакции. Информация по старому балансу получателя тоже практически ни на что не влияет, поэтому отбросим и этот признак.

Далее проектируются новые признаки на основе имеющихся. Создается новый *.csv* файл. Ниже перечисляются добавленные признаки (в скобках указано имя в новом файле):

- время (*hour*). Условный шаг из исходных данных переведен в 24-чаовой формат. Может быть мошеннические транзакции совершаются в определенное время суток;

- новый ли отправитель (*newSender).* Пометка, показывающая выступал ли ранее текущий отправитель в такой же роли;
- новый ли получатель (*newReceiver).* Пометка, показывающая выступал ли ранее текущий получатель в такой же роли;
- является ли получатель магазином (*merchant).* Пометка, показывающая является ли принимающая сторона магазина. Поскольку с магазином не совершаются злоумышленные транзакции, этот пункт может быть полезным;
- был ли хоть один из клиентов, участвующих в текущей транзакции, ранее замечен в нарушениях *(fraudsEarly).* Пометка может быть полезна, если клиент совершает кражи регулярно;
- время, прошедшее с момента последней транзакции в качестве отправителя (*LTS).* Количество часов указывается количество часов, либо ставится -1, если первое появление.
- время, прошедшее с момента последней транзакции в качестве получателя (*LTR).* Аналогично пункту, написанному выше;
- остается ли ноль на балансе отправителя (*IZoB).* Если клиент остается с 0 на балансе, то скорее всего его обчистил мошенник.

Далее с помощью встроенного в библиотеку scikit-learn метода RFE (recursive feature elimination) производится ранжирование признаков по значимости. Название метода переводится как «рекурсивное отсечение признаков». Алгоритм его работы следующий: модель обучается на исходном наборе признаков, оценивает их значимость, отсекает один или несколько наименее значимых, обучается на новых признаках и повторяет эти действия, пока не останется *n* наиболее значимых признаков.

Также у класса случайного леса из библиотеки scikit-learn есть метод, возвращающий значимость каждого признака.

Результаты работы двух этих методов приведены в таблице 2.4.

Таблица 2.4 Результаты работы алгоритма RFE и значимость признаков

	RFE	Значимость признака
step	1	0.085
type	2	0.062
amount	1	0.243
oldbalanceOrg	1	0.389
hour	1	0.110
newSender	7	7.973e-06
newReceiver	4	0.010
merchant	5	0.009
fraudsEarly	6	0.0005
LTS	8	2.28e-08
LTR	3	0.016
IZoB	1	0.071

Результаты удивляют. Самым значимым признаком оказался баланс до транзакции на счету отправителя. Не совсем понятно, с чем связана такая зависимость. Возможно, так получилось из-за синтетической природы исходной выборки. Второй по значимости признак – ожидаемо, сумма. Из новых добавленных признаков наиболее значимыми оказались время в часах, индикатор нуля на балансе, время с последней транзакции. Остальные признаки не проявили себя, их можно отбросить. Также можно отбросить шаг, у 24-часового аналога значимость больше.

Отбросив менее значимые параметры, проводится эксперимент и сравниваются результаты с предыдущей итерацией. Далее отбрасываем еще мало значимые признаки, пока не достигнем максимального результата.

2.5 Результаты

Таким образом, остались следующие признаки *type, amount, oldBalanceOrg, hour, iZoB.*

Также была проведена процедура нормализации данных, описанная в разделе 1.2.1

Результат модели на конечных данных, а также его сравнение с результатом на исходных данных приведены в таблице 2.5.

Таблица 2.5 Результат работы модели на конечных данных

		Наивный байесовский	Случайный лес	k-ближайших соседей
precision	Исходные	0.595	0.994	0.954
	Конечные	0.547	0.985	0.955
	Разница	-0.048	-0.009	0.001
recall	Исходные	0.204	0.831	0.794
	Конечные	0.189	0.915	0.959
	Разница	-0.015	0.084	0.165
F_1	Исходные	0.304	0.905	0.867
	Конечные	0.281	0.949	0.957
	Разница	-0.023	0.044	0.90

Лучше всего себя проявил метод k-ближайших соседей. По сравнению с результатами на исходных данных прогресс явно заметен. Случайный лес тоже выдал хорошее качество предсказаний. Наивный байесовский классификатор ухудшил свои результаты.

Таким образом, можно сделать вывод, что метрический метод лучше всего подходит для решения данной задачи. Наравне с ним можно применять ансамбль из решающих деревьев. А вот метод, основанный на теореме Байеса, проявил себя хуже всего, его точно не стоит использовать для решения задач такого типа.

Для более наглядного сравнения результатов построены precision-recall кривые всех трех методов (рис. 2.5).

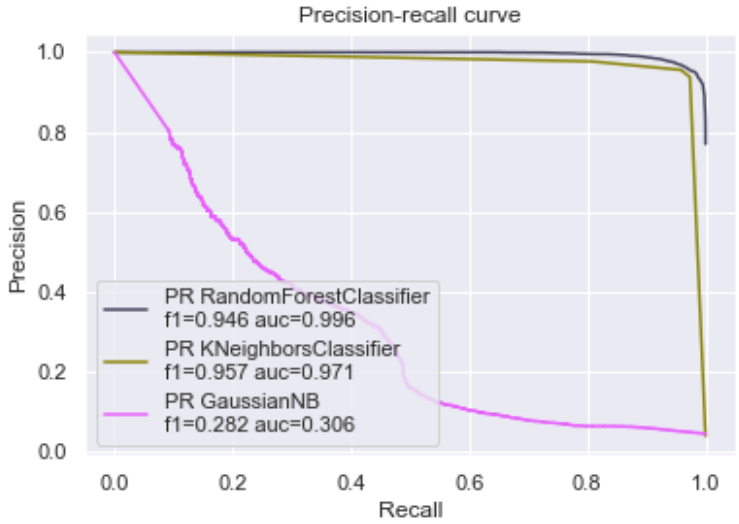

Рис. 2.5 PR-кривые использованных методов

ЗАКЛЮЧЕНИЕ

В выпускной квалификационной работе выполнены все поставленные задачи в полном объеме:

- проведен анализ финансовых транзакций с целью предотвращения мошеннических операций;
- реализована прикладная программа, визуализирующая графы на основе финансовых транзакций;
- реализована модель, способная классифицировать финансовые транзакции;
- проведён анализ эффективности реализованной модели и ее сравнение с другими методами;

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ