РЕФЕРАТ

Магистерская диссертация содержит 30 страниц, 2 таблицы, 3 рисунка. Список использованных источников содержит 12 позиций.

ОТБОР ПРИЗНАКОВ, ВЗАИМНАЯ ИНФОРМАЦИЯ, СОВМЕСТНАЯ ВЗАИМНАЯ ИНФОРМАЦИЯ, УСЛОВНАЯ ВЗАИМНАЯ ИНФОРМАЦИЯ, КЛАССИФИКАЦИЯ

Магистерская диссертация посвящена методам отбора признаков, основанным на теории информации. В данной работе предлагается нелинейный метод. Он направлен на преодоление ограничений современных методов отбора признаков, таких как завышение значимости, что приводит к выделению избыточных и нерелевантных признаков. Это достигается введением новой целевой функции, основанной на совместной взаимной информации и нелинейном подходе максимизации минимума.

В теоретической части рассмотрены различные методы отбора признаков и их недостатки, предложено решение для устранения недостатков метода JMI. В практической части реализована и протестирована библиотека вычисления взаимной информации и совместной взаимной информации. Реализован предложенный метод отбора признаков для больших данных. Проведено сравнение работы двух методов на наборе данных.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В настоящее время с резким ростом размера доступных наборов данных как с точки зрения количества выборок, так и количества признаков в каждой выборке, важной задачей становится уменьшение размерности данных и сохранение числа признаков как можно более низким. Основной мотивацией является сокращение времени обучения и повышение точности классификации алгоритмов.

Методы уменьшения размерности можно разделить на две основные группы: методы, основанные на извлечении признаков, и методы, основанные на отборе признаков. Методы извлечения признаков преобразуют существующие признаки в новое пространство признаков меньшей размерности. В ходе этого процесса новые признаки создаются на основе линейных или нелинейных комбинаций из исходного набора. Анализ главных компонент и линейный дискриминантный анализ являются двумя примерами таких алгоритмов. Методы отбора признаков уменьшают размерность, выбирая подмножество признаков, которое минимизирует определенную функцию потерь. В отличие от извлечения, отбор признаков не изменяет данные и, как следствие, является предпочтительным, когда требуется понимание лежащего в основе физического процесса. Извлечение признаков может быть предпочтительным, когда основную роль играет точность решения задачи (классификации, регрессии и т. д.).

Отбор признаков используется во многих прикладных областях, имеющих отношение к экспертным и интеллектуальным системам, таким как интеллектуальный анализ данных и машинное обучение, обработка изображений, обнаружение аномалий, биоинформатика и обработка естественного языка. Отбор признаков обычно используется на этапе предварительной обработки данных перед обучением классификатора.

Обычно методы отбора признаков делятся на две категории: зависимые от классификатора (оберточные и встроенные методы) и независимые от классификатора (фильтрующие методы). Методы обертки выполняют поиск в пространстве признаков и тестируют все возможные подмножества комбинаций признаков, используя точность предсказания классификатора в качестве меры качества выбранного подмножества без изменения функции обучения. Они как правило обеспечивают высокую точность классификации, потому что выбранное подмножество оптимизировано в терминах данной точности. С другой стороны, методы обертки могут страдать от чрезмерной подгонки к алгоритму обучения. Это означает, что любые изменения в модели обучения могут снизить полезность подмножества. Кроме того, эти методы очень дороги с точки зрения вычислительной сложности, особенно при обработке чрезвычайно многомерных данных.

Этап отбора признаков во встроенных методах сочетается с этапом обучения. Эти методы менее дороги с точки зрения вычислительной сложности и менее подвержены переобучению, однако они ограничены с точки зрения обобщения, поскольку они очень специфичны для используемого алгоритма обучения.

Основными преимуществами методов фильтрации являются их вычислительная эффективность, масштабируемость с точки зрения размерности набора данных и независимость от классификатора. Общим недостатком этих методов является отсутствие информации о взаимодействии признаков с классификатором и выделение избыточных и нерелевантных признаков из-за ограничений используемых целевых функций, приводящих к завышению значимости признака.

Теория информации широко применяется в методах фильтрации, где информационные меры, такие как взаимная информация (MI), используются в качестве меры релевантности и избыточности признаков. MI не делает предположения о линейности между переменными и может иметь дело с

категориальными и числовыми данными с двумя или более классами. Существует несколько альтернативных мер в теории информации, которые могут быть использованы для вычисления релевантности признаков, а именно: взаимная информация, информация взаимодействия, условная взаимная информация и совместная взаимная информация.

В данной работе предлагается нелинейный метод отбора признаков, основанный на теории информации. Он направлен на преодоление ограничений современных методов отбора признаков, таких как завышение значимости, что приводит к выделению избыточных и нерелевантных признаков. Это достигается введением новой целевой функции, основанной на совместной взаимной информации и нелинейном подходе максимизации минимума.

ОСНОВНАЯ ЧАСТЬ

1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1.1 Основные понятия теории информации

Энтропия случайной величины − это мера ее неопределенности и мера среднего количества информации, необходимой для описания случайной величины. Энтропия дискретной случайной величины $X = (x_1, x_2, \ldots, x_N)$ обозначается через $H(X)$, где x_i относится к возможным значениям, которые может принимать X.

$$H(x) = -\sum_{i=1}^{N} p(x_i)\, log(p(x_i)) \tag{1.1}$$

где $p(x_i)$ − функция вероятности. Значение $p(x_i)$, когда X дискретно, может быть оценено на основании серии экспериментов по следующей формуле:

$$p(x_i) = \frac{n_i}{N} \tag{1.2}$$

где n_i – количество наблюдений значения x_i, а N – общее количество наблюдений.

От основания логарифма в формуле (1.1) зависят единицы измерения энтропии. Одна из основных общепринятых единиц измерения количества информации и энтропии – бит, задаётся основанием логарифма равным 2. Здесь и далее основания логарифмов предполагаются равными 2. Из формулы (1.1) следует, что $0 \leq H(X) \leq 1$. Для любых двух дискретных случайных величин X и $C = (c_1, c_2, \ldots, c_M)$ совместная энтропия определяется как:

$$H(X, C) = -\sum_{j=1}^{M}\sum_{i=1}^{N} \mathrm{p}(\mathrm{x_i}, \mathrm{c_j})\, \log(\mathrm{p}(\mathrm{x_i}, \mathrm{c_j})) \tag{1.3}$$

где $p(x_i, c_j)$ − совместная функция вероятности переменных X и C. Условная энтропия переменной C при условии X, определяется как:

$$H(C|X) = -\sum_{j=1}^{M}\sum_{i=1}^{N} \mathrm{p}(\mathrm{x_i}, \mathrm{c_j}) \log(\mathrm{p}(\mathrm{c_j}|\mathrm{x_i})) \quad (1.4)$$

Условная энтропия − это величина неопределенности, оставшаяся в C при введении переменной X, поэтому она меньше или равна энтропии обеих переменных. Условная энтропия равна энтропии, если и только если две переменные независимы. Отношение между совместной энтропией и условной энтропией:

$$\mathrm{H}(\mathrm{X}, \mathrm{C}) = \mathrm{H}(\mathrm{X}) + \mathrm{H}(\mathrm{C}|\mathrm{X}) \quad (1.5)$$

$$\mathrm{H}(\mathrm{X}, \mathrm{C}) = \mathrm{H}(\mathrm{C}) + \mathrm{H}(\mathrm{X}|\mathrm{C}) \quad (1.6)$$

Взаимная информация (MI) − это статистическая функция двух случайных величин, описывающая количество информации, содержащееся в одной случайной величине относительно другой, и определяется как:

$$\mathrm{I}(\mathrm{X}; \mathrm{C}) = \mathrm{H}(\mathrm{C}) - \mathrm{H}(\mathrm{C}|\mathrm{X}) \quad (1.7)$$

Свойства взаимной информации:

- Взаимная информация является симметричной функцией случайных величин.

$$\mathrm{I}(\mathrm{X}; \mathrm{C}) = \mathrm{I}(\mathrm{C}; \mathrm{X}) \quad (1.8)$$

- Взаимная информация неотрицательна и не превосходит информационную энтропию аргументов.

$$0 \leq \mathrm{I(X,C)} \leq \min[\mathrm{H(X)}, \mathrm{H(C)}] \tag{1.9}$$

- В частности, для независимых случайных величин взаимная информация равна нулю.

$$\mathrm{I(X;C)} = \mathrm{H(X)} - \mathrm{H(X|C)} = \mathrm{H(X)} - \mathrm{H(X)} = 0 \tag{1.10}$$

- В случае, когда одна случайная величина (например, X) является детерминированной функцией другой случайной величины (C), взаимная информация равна энтропии.

$$\mathrm{I(X;C)} = \mathrm{H(X)} - \mathrm{H(X|C)} = \mathrm{H(X)} - 0 = \mathrm{H(X)} \tag{1.11}$$

MI можно выразить как объем информации, предоставляемой переменной X, что уменьшает неопределенность переменной C.

$$\mathrm{I(X;C)} = \mathrm{H(X)} - \mathrm{H(X|C)} \tag{1.12}$$

$$\mathrm{I(X;C)} = \mathrm{H(X)} + \mathrm{H(C)} - \mathrm{H(X,C)} \tag{1.13}$$

Совместная взаимная информация определяется как:

$$I(X;C|Y) = H(X|C) - H(X|C,Y) \tag{1.14}$$

$$I(X,Y;C) = I(X;C|Y) + I(Y;C) \tag{1.15}$$

где Y – дискретная переменная; $Y = (y_1, y_2, \ldots, y_N)$. Информация о взаимосвязи может быть определена как объем информации, которая является общей для всех признаков, но не находится ни в одном

подмножестве. Математически связь между информацией о взаимосвязи и MI определяется как:

$$I(X;Y;C) = I(X,Y;C) - I(X;C) - I(Y;C) \quad (1.16)$$

Высокая информационная взаимосвязь означает, что большой объем информации можно получить, рассматривая три переменные вместе. Информация о взаимосвязи может быть положительной, отрицательной или нулевой.

1.2 Обзор литературы

Теория информации используется во многих фильтрующих методах отбора признаков. Information Gain (IG) − самый простой из этих методов. Он классифицируется как одномерный метод отбора признаков, поскольку он ранжирует их, основываясь на значении взаимной информации с меткой класса. Простота и низкие вычислительные затраты − основные преимущества этого метода. Однако он не принимает во внимание зависимость между признаками, скорее, он предполагает независимость, что не всегда так. Поэтому некоторые из выбранных признаков могут нести избыточную информацию. Для решения этой проблемы были предложены новые методы отбора признаков, не являющихся избыточными по отношению друг к другу.

Для набора функций $F = \{f_1, f_2, \ldots, f_N\}$ процесс выбора признаков идентифицирует подмножество признаков S с размерностью k, где $k \leq N$, и $S \subseteq F$. Теоретически выбранное подмножество S должно максимизировать совместную взаимную информацию между меткой класса C и подмножеством S фиксированного размера k.

$$I(S;\ C) = I(f_1, f_2, \ldots, f_k;\ C) \tag{1.17}$$

Однако такой подход непрактичен из-за большого количества вычислений и ограниченного числа наблюдений, доступных для вычисления многомерной функции плотности вероятности. В результате, многие методы используют эвристические подходы для приближения идеального решения.

Как правило, критерии фильтрации основаны на понятиях релевантности признаков, избыточности и взаимодополняемости. Методы фильтрации, основанные на теории информации, можно разделить на две группы: линейные критерии, которые представляют собой линейные комбинации членов взаимной информации, и нелинейные критерии, которые используют максимальные или минимальные операции или нормализованную взаимную информацию в своих целевых функциях.

В статье [1] рассматривается алгоритм инкрементного поиска первого порядка, известный как метод взаимного выбора информационных признаков (MIFS), для выбора k наиболее релевантных признаков из начального набора n признаков. Для построения подмножества используется жадный метод отбора. Вместо того, чтобы вычислять совместную взаимную информацию между выбранными объектами и меткой класса, автор вычисляет взаимную информацию между признаком-кандидатом и классом, а также взаимосвязь между кандидатом и уже выбранными объектами.

Для повышения производительности метода MIFS за счет более точной оценки взаимной информации между входной функцией и меткой класса в статье [2] предлагается метод MIFS-U. Другой вариант метода MIFS, метод mRMR, предложен в статье [3]. Член избыточности в mRMR делится по мощности $|S|$ выбранного подмножества S, чтобы сбалансировать величину этого члена и избежать его очень большого роста по мере расширения

подмножеств. Как сообщается в литературе [3], эта модификация позволяет mRMR превзойти традиционные методы MIFS и MIFS-U.

В статье [4] предлагается метод NMIFS, который использует нормализованную взаимную информацию для отбора признаков. Нормализация взаимной информации предотвращает смещение в сторону многозначных функций и ограничивает её значение диапазоном от нуля до единицы.

В статье [5] предлагается метод под названием MIFS-ND. Метод вычисляет взаимную информацию между рассматриваемым признаком и меткой класса, а также среднее значение взаимной информации между данным признаком и всеми уже отобранными признаками. Генетический алгоритм используется для выбора признака, который максимизирует взаимную информацию с классом и минимизирует среднюю взаимную информацию с другими отобранными признаками.

В статье [6] предлагается метод отбора признаков, называемый совместной взаимной информацией (JMI). В этом методе признак-кандидат, который максимизирует совокупную сумму совместной взаимной информации с признаками выбранного подмножества, выбирается и добавляется к подмножеству. Этот метод хорошо работает с точки зрения точности и стабильности классификации.

Методы отбора признаков, основанные на теории информации, также используются для многоклассовых данных. В статье [7] предложен метод, в котором вводится новая функция оценки для измерения важности каждого признака для нескольких меток.

Двумя другими известными подходами в области фильтрующих методов отбора признаков являются применение теории неточных множеств и применение анализа охвата данных. Одной из проблем, влияющих на методы, основанные на нечетких и неточных наборах, является их неэффективность по времени. Эти методы также страдают от проблемы

больших вычислительных затрат и проблемы выбора избыточных признаков, как описано в статье [8].

1.3 Ограничения существующих методов отбора признаков

Как правило, большинство методов, перечисленных в предыдущем разделе, используют критерии, состоящие из двух элементов: релевантности и избыточности. Эти методы пытаются одновременно максимизировать значение релевантности и минимизировать значение избыточности. В литературе отмечалось, что такие методы выбора признаков имеют ряд ограничений.

Например, MIFS и MIFS-U имеют общую проблему: когда количество выбранных функций растет, значение избыточности увеличивается по величине по отношению к релевантности. В этом случае могут быть выбраны некоторые нерелевантные функции. Эта проблема была частично решена в методах mRMR, NMIFS, MIFS-ND путем разделения члена избыточности по мощности подмножества.

Другая проблема, присущая всем вышеуказанным методам (MIFS, MIFS-U, mRMR, NMIFS и MIFS-ND), заключается в том, что избыточность вычисляется на основе значения взаимной информации между признаком-кандидатом и уже отобранными признаками, никак не принимая во внимание метки класса. Признаки могут иметь взаимную информацию друг с другом, но это не означает, что они избыточны. На самом деле, они могут привносить различную информацию к целевой функции, то есть делиться с классом разной информацией.

Еще одна проблема, специфичная для методов, использующих кумулятивное суммирование и прямой поиск для аппроксимации решения уравнения (1.17) − это переоценка значимости некоторых признаков-кандидатов. Например, это может произойти, когда очередной рассматриваемый признак находится в полной корреляции с одним или

несколькими предварительно уже отобранными, но в то же время почти не зависит от большинства подмножества. В такой ситуации значение целевой функции будет высоким, несмотря на избыточность признака-кандидата для некоторых функций в подмножестве.

На практике значимость каждой из вышеперечисленных проблем зависит от данных и характеристик каждого конкретного набора данных.

1.4 Постановка задачи

В данной работе предложен метод отбора признаков, использующий совместную взаимную информацию и подход максимизации минимума. Данный метод направлен на решение проблемы завышения значимости некоторых признаков, возникающей при использовании аппроксимации кумулятивного суммирования.

Для набора признаков $F = \{f_1, f_2, \ldots, f_N\}$ набора данных D размерности N процесс выбора признаков определяет подмножество признаков S с размерностью K, где $K \leq N$ и $S \subseteq F$. Подмножество S должно обеспечивать равную или лучшую точность классификации по сравнению с набором признаков F. Другими словами, выбор признаков определяет подмножество признаков, которое максимизирует взаимную информацию с меткой класса I(S, C).

Введем понятие релевантности признаков. Признак f_i более релевантен метке класса C, чем признак f_j в контексте уже выбранных подмножеств, когда $I(f_i, S;\ C) > I(f_j, S;\ C)$.

Пусть F − полный набор признаков, а S − подмножество признаков, которые уже выбраны. Пусть $f_i \in F \setminus S$, а $f_s \in S$, а mMI − это минимальное значение совместной взаимной информации, которую признак-кандидат f_i разделяет с меткой класса C, когда он соединяется с каждым объектом в

подмножестве S индивидуально, следовательно, $mMI = min_{s=1,2,\dots,k} I(f_i, f_s;\ C)$.

Лемма 1. Для признака f_i, если совместная взаимная информация больше, чем у всех других объектов f_j, где $f_i,\ f_j \in F\ \backslash S\ (i \neq j)$, то это наиболее релевантный объект для метки класса C в контексте подмножества S.

Доказательство. Пусть $S = \{f_1, f_2, \dots, f_K\}$. Вычисляется совместная взаимная информация f_i и каждого признака в S с C. Минимальное значение этой взаимной информации (mMi) − это наименьшее количество новой информации, которую признак f_i добавляет к общей информации между S и C. Признак, который дает максимальной значение mMI − это признак, который добавляет максимальную информацию к той, которая разделяется между S и C, что означает, что это признак, который является наиболее релевантным для метки класса C в контексте подмножества S в соответствии с определением релевантности признаков.

Признак-кандидат f_i является избыточным для выбранных объектов в подмножестве S, если f_i не привносит новую информацию о классе C.

1.4.1 Максимизация совместной взаимной информации

Все методы, перечисленные в предыдущем разделе, пытаются оптимизировать соотношение между релевантностью и избыточностью при выборе признаков путем аппроксимации решения уравнения (1.17). В существующей литературе [6] сообщается, что метод JMI является методом, который выбирает наиболее релевантные признаки. Он изучает релевантность и избыточность, а также учитывает метку класса при расчете взаимной информации. Однако метод все же позволяет переоценить значимость некоторых признаков, например, когда признак-кандидат находится в полной корреляции с одним или несколькими заранее

выбранными признаками, но при этом практически не зависит от большинства подмножества. В такой ситуации значение совместной взаимной информации будет высоким, несмотря на избыточность признака-кандидата. Этот недостаток очевиден почти во всех методах, использующих аппроксимацию кумулятивной суммы.

По этой причине в данной работе предлагается метод максимизации совместной взаимной информации. Он использует совместную взаимную информацию и подход "максимизации минимума", который должен выбирать наиболее релевантные признаки в соответствии с леммой 1, следуя которой, признаки выбираются по следующему новому критерию:

$$\mathrm{f} = \mathrm{argmax}_{\mathrm{f_i} \in \mathrm{F} \setminus \mathrm{S}} (\min_{\mathrm{f_s} \in \mathrm{S}} (\mathrm{I}(\mathrm{f_i}, \mathrm{f_s};\ \mathrm{C}))) \tag{1.18}$$

$$\mathrm{I}(\mathrm{f_i}, \mathrm{f_s};\ \mathrm{C}) \ = \ \mathrm{I}(\mathrm{f_s};\ \mathrm{C}) + \mathrm{I}(\mathrm{f_i}, \mathrm{C} \ / \ \mathrm{f_s}) \tag{1.19}$$

$$I(f_i, f_s; C) = H(C) - H\left(\frac{C}{f_i}, f_s\right) \tag{1.20}$$

$$\begin{aligned} I(f_i, f_s; C) = & \left[- \sum_{c \in C} p(c) \log(p(c)) \right] \\ & - \left[\sum_{c \in C} \sum_{f_i \in F - S} \sum_{f_s \in S} log\left(\frac{p(f_i f_s, c \ / f_s)}{p(f_i \ / f_s) p(c \ / f_s)} \right) \right] \end{aligned} \tag{1.21}$$

Метод использует итеративный алгоритм прямого жадного поиска для выбора соответствующего подмножества объектов размера K в пространстве объектов, на рис. 1.1 представлена блок-схема данного алгоритма.

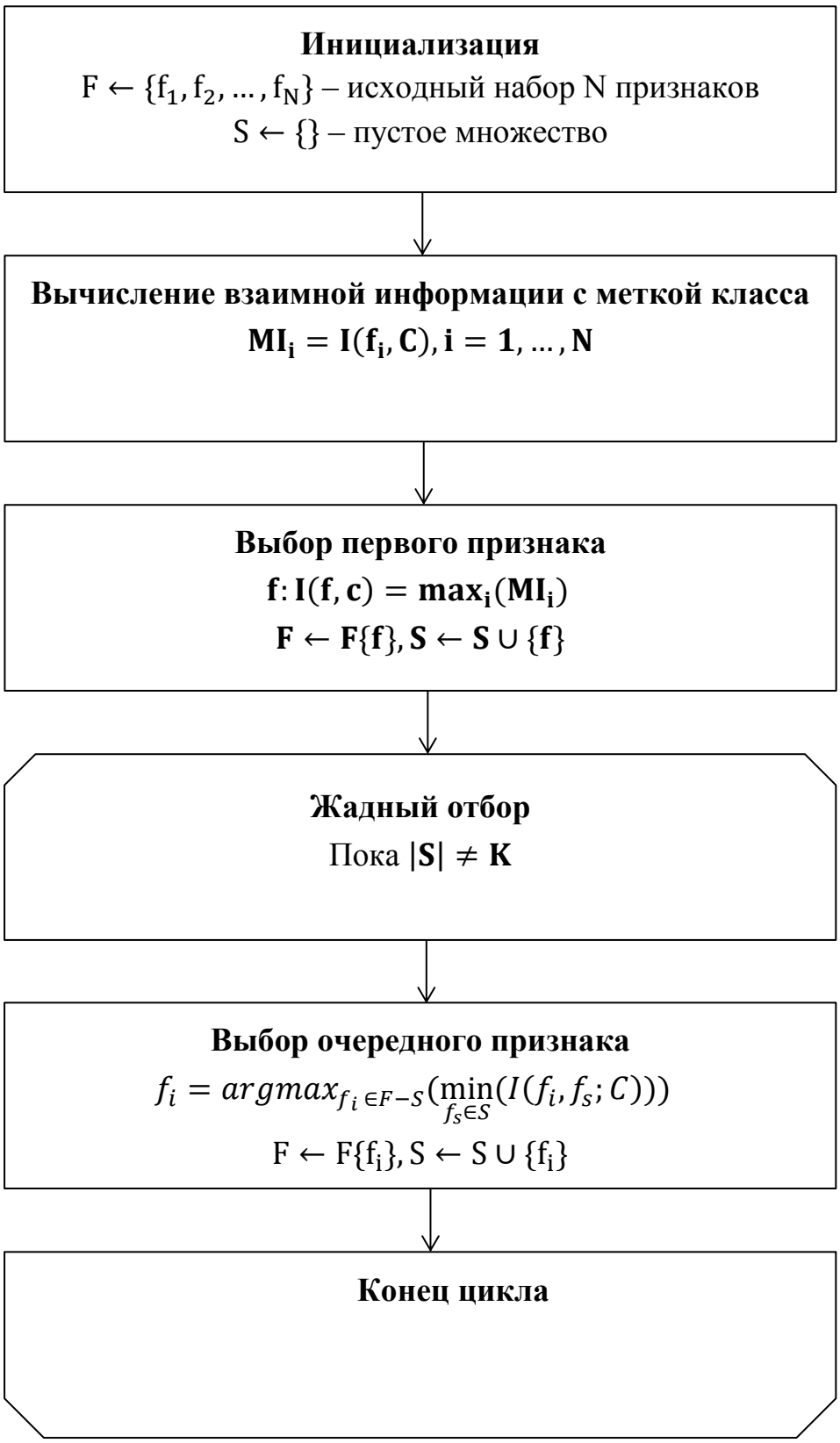

Рис. 1.1 Блок-схема алгоритма отбора признаков

1.5 Вычисление взаимной информации между дискретными и непрерывными случайными величинами

Существуют определенные сложности при вычислении взаимной информации для наборов данных с действительными значениями признаков, поскольку они по определению имеют разреженную выборку: большинство возможных значений не может быть найдено в наборе данных любого размера. Стандартное решение состоит в том, чтобы объединить непрерывные переменные в дискретные ячейки, а затем применить дискретную оценку взаимной информации. Улучшенная непрерывная оценка взаимной информации, описанная в работе [9] использует статистику расстояний между точками данных и их ближайшими соседями. Важно отметить, что метод работает только тогда, когда обе переменные являются действительными, поскольку ближайший сосед дискретной переменной не определен четко.

В работе [10] предлагается метод вычисления взаимной информации между дискретными и непрерывными величинами, основанный на методе ближайших соседей. Проводится его сравнение с наивным разбиением вещественных признаков на ячейки. В обоих методах присутствует настраиваемый пользователем параметр. В наивном методе − количество ячеек, на которые необходимо разбить данные, а в методе, использующем информацию о ближайших соседях − непосредственно число рассматриваемых соседей. Первый вывод статьи состоит в том, что существует гораздо более простой способ для выбора параметра количества соседей. Небольшое значение данного параметра стабильно дает хорошие результаты. С другой стороны оценка количества ячеек для разбиения набора данных переоценивает взаимную информацию, когда она низка, и недооценивает, когда высока, и, хотя гарантированно существует точка пересечения, где метод является точным, трудно угадать, где эта точка может

быть. Кроме того, в статье утверждается, что не существует простого способа вычислить оптимальный параметр разбиения на ячейки на основе простой статистики данных, такой как общее количество строк или частоты, с которыми встречаются разные дискретные символы.

2 ПРАКТИЧЕСКАЯ ЧАСТЬ

2.1 Постановка задачи

Рассматривается задача бинарной классификации “пользователь совершит клик на рекламный баннер”. Предоставляется выборка данных, в каждой строке которой описывается рекламный баннер и пользователь, которому данный баннер был показан. Описание содержит набор из 1200 действительных значений существующих признаков для обучения. В качестве признаков пользователя могут использоваться регион, текущее время, последние N запросов в поисковую систему, для рекламного объявления может использоваться его новизна, является ли это рекламой товара/услуги, возрастная категория объявления и другие. Также каждая строка содержит метку класса: 1, если пользователь совершил клик на рекламный баннер, и 0 – иначе. Все строки выборки отсортированы по дате показа рекламного баннера. Среди большого количества признаков необходимо отобрать подмножество, приносящее пользу при обучении классификатора.

Первой частью работы является реализация библиотеки, позволяющей вычислять взаимную информацию и условную взаимную информацию для больших данных. Необходимо протестировать работу библиотеки на искусственном наборе данных с заранее известным значением тестируемых функций.

Второй частью работы является реализация предложенного алгоритма отбора признаков, использующего максимизацию совместной взаимной информации, сравнение его работы с алгоритмом JMI.

2.2 Реализация вычисления взаимной информации

Вычисление взаимной информации реализовано на языке C++ в виде библиотеки с использованием технологии MapReduce для возможности

работы с большими данными. В работе используется MapReduce-система под названием YT, разрабатываемая Яндексом, подробно описанная в работе [11].

Библиотека состоит из двух смысловых частей: функций для предобработки данных и для вычисления взаимной информации. На первом этапе необходимо преобразовать признаки, имеющие действительные значения. В данной работе используется наивный метод равномерного разбиения действительных признаков на ячейки, так как он наиболее просто реализуем для больших данных. На вход функции подается таблица с данными, указываются имена колонок не категориальных признаков, количество ячеек, на которые необходимо разбить данные. Значения каждого признака равномерно разбиваются на заданное количество ячеек в отсортированном порядке. На вход функции, вычисляющей взаимную информацию подается таблица с данными, указываются имена колонок признаков, с предобработанными данными, указывается колонка со значением метки класса, производится вычисление совместной взаимной информации по формуле (1.21).

2.3 Анализ работы библиотеки на искусственных наборах данных

В первую очередь производится вычисление совместной взаимной информации между двумя признаками, принимающими случайное значение от 0 до 100 и случайной меткой класса, принимающей значение 0 или 1. Для таких данных очевидно, что признаки не несут никакой информации о метке класса, соответственно значение совместной взаимной информации должно быть равно 0. Действительно, для 10000 строк данных при разбиении признаков на 100 ячеек, вычисленное значение совместной взаимной информации близко к нулю.

Далее производится вычисление совместной взаимной информации для случая, когда по отдельности признаки не имеют информации о метке класса,

но в совокупности они ей обладают. Пусть метка класса принимает значение 0 или 1, а первый признак f1 будет случайным вещественным числом из интервала (0, 1), второй признак f2 будет вычисляться как разность значения метки класса и первого признака, то есть класс определяется линейной комбинацией двух признаков. Вычисление взаимной информации проводилось для 10000 строк данных, результаты представлены в таблице 2.1.

Таблица 2.1 Значение взаимной информации для наборов данных

№	f1	f2	C	I(f1;C)	I(f2;C)	I(f1, f2; C)
1.	random(0, 100)	random(0, 100)	randint(0, 1)	0	0	0
2.	random(0, 1)	C - f1	randint(0, 1)	0	0	1.3876

2.4 Анализ работы библиотеки на данных с известным распределением

Производится расчет взаимной информации между двумя признаками: x является категориальным и принимает значения 1, 2, 3, а y является вещественным с известной функцией распределения.

Пусть значения категориального признака будут равновероятны, а вероятность y для каждого x будет иметь нормальное распределение с заданным математическим ожиданием и среднеквадратическим отклонением. В таблице 2.2 представлены параметры нормального распределения вероятности признака y для каждого значения x, в ней μ — это математическое ожидание, а σ — среднеквадратическое отклонение. На рис. 2.1 представлены графики условной вероятности $p(y|x)$. Они имеют разные цвета для каждого из трех возможных значений дискретной переменной.

Таблица 2.2 Параметры функции нормального распределения

x	μ	σ
1	1.5	0.44
2	2	1
3	3	0.5

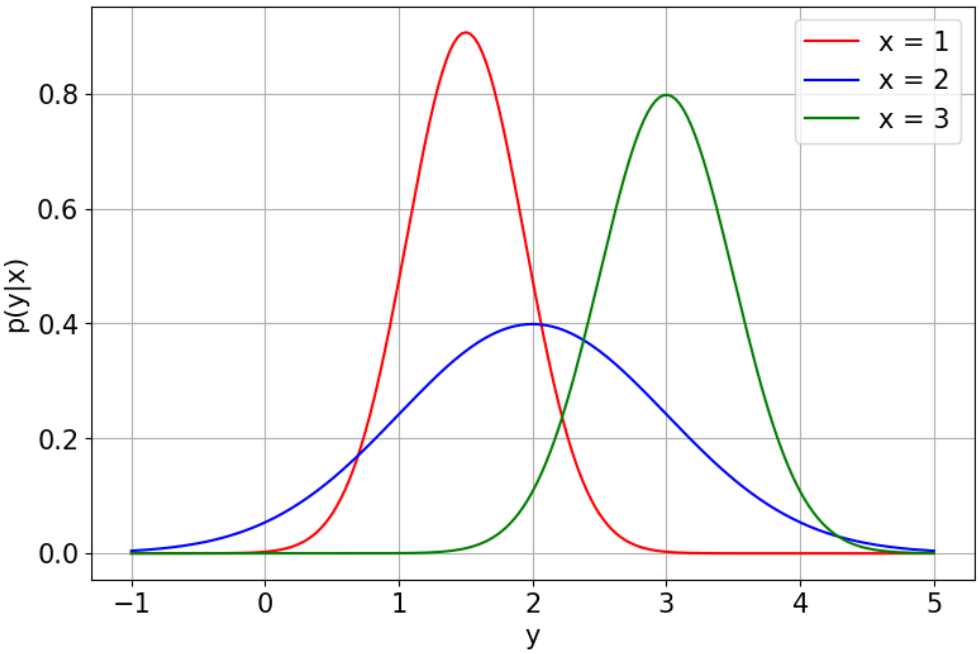

Рис. 2.1 Зависимость условной вероятности $p(y|x)$ от значения дискретной переменной x.

Вычислим взаимную информацию аналитически по формуле (1.13). Так как категориальный признак x равновероятно принимает одно из трех возможных значений, его энтропия будет вычисляться следующим образом:

$$H(x) = -\sum_{i=1}^{N} p(x_i) log(p(x_i)) = -3 * \frac{1}{3} * log\left(\frac{1}{3}\right) = 1.585 \qquad (2.1)$$

Пусть f(y) – плотность вероятности y независимо от признака x, а $f(x, y)$ – совместная плотность вероятности. Тогда,

$$f(y) = \sum_{x} f(x, y) \tag{2.2}$$

$$f(y|x) = \frac{f(x, y)}{p(x)} \tag{2.3}$$

Энтропия признака y будет вычисляться по формуле:

$$\begin{aligned} H(y) &= -\int_{-\infty}^{\infty} f(y) log\big(f(y)\big) dy \\ &= -\int_{-\infty}^{\infty} \sum_{x} \left(f(y|x)\, p(x)) \; log \left(\sum_{x} (f(y|x)\, p(x) \right) \right) dy \end{aligned} \tag{2.4}$$

Зная, что $f(y|x)$ – гауссиан с известными параметрами μ и σ, получаем $H(y) = 1.9$. Осталось вычислить совместную энтропию $H(x, y)$.

$$H(x, y) = -\sum_{x} \int_{-\infty}^{\infty} f(y|x) p(x) log\big(f(y|x) p(x)\big) \mathrm{dy} = 2.9 \tag{2.5}$$

$$I(x, y) = 1.585 + 1.9 - 2.9 = 0.585 \tag{2.6}$$

На языке Python3 с использованием библиотеки numpy была реализована генерация данных с заданным распределением. Для каждого значения x рассчитано по 1000 значений y. В результате работы библиотеки была получена оценка взаимной информации $I(x, y) = 0.52$. Она является достаточно близкой к истинному результату.

2.5 Анализ работы алгоритма отбора признаков

Эффективность работы предложенного метода сравнивается с результатами, полученными с помощью метода JMI, описанного в литературе. Эта пара алгоритмов позволяет сравнить подход "максимизации минимума" с кумулятивным суммированием для отбора признаков, используя взаимную информацию.

Анализ работы производится на описанном в постановке задачи наборе данных, имеющем 1200 признаков и 10 000 000 строк. Качество отбора признаков оценивается по точности классификации. Под точностью классификации понимается метрика accuracy – доля правильных ответов. В качестве классификатора выбран метод K ближайших соседей.

В первую очередь данные разделяются на обучающую и тестовую выборки. Задается требуемое значение количества признаков, которые нужно выбрать, каждым алгоритмом отбора получается свое подмножество. На этих подмножествах обучается и тестируется классификатор. Сравнивается точность классификации на одних и тех же обучающей и тестовой выборке, но разном наборе признаков. Затем, варьируя параметр количества признаков, можно построить график зависимости точности классификации от количества признаков для каждого алгоритма отбора. Такой график представлен на рис. 2.2.

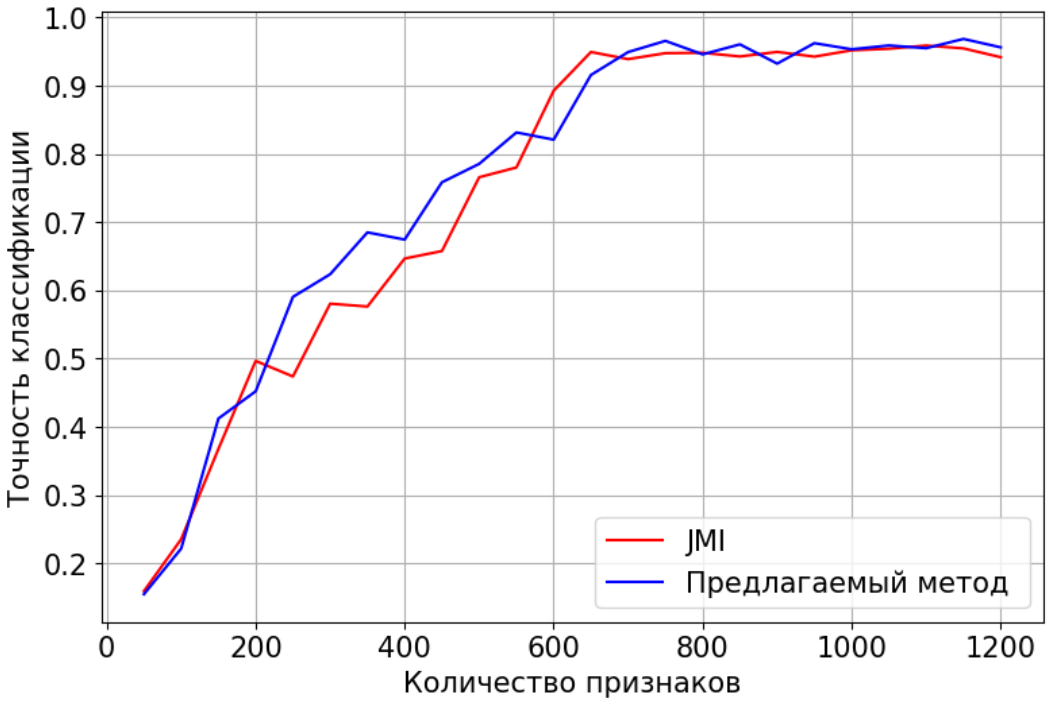

Рис. 2.2 Зависимость качества классификации от количества отобранных признаков

Можно заметить, что при малых значениях количества признаков точность классификации на обоих наборах примерно одинакова. Это означает, что отбираются одни и те же признаки. Далее, при увеличении количества признаков от 200 до 550 заметно преимущество описанного метода: классификация на признаках, отобранных с его помощью, имеет большую точность. Это обусловлено тем, что предлагаемый метод не отбирает избыточные сильно коррелирующие признаки. С дальнейшим увеличением числа признаков точность классификации вновь выравнивается, так как количество отбираемых признаков близко к мощности исходного множества признаков. Также можно заметить, что максимальная точность классификации достигается на 750 признаках, отобранных с помощью предложенного метода.

ЗАКЛЮЧЕНИЕ

В магистерской диссертации выполнены все поставленные цели. Изучена задача отбора признаков, рассмотрены различные методы ее решения с помощью теории информации. Определены недостатки существующих подходов, предложено решение для устранения недостатков метода JMI. Реализована библиотека для вычисления взаимной информации. Протестирована работа библиотеки на нескольких искусственных наборах данных с известным распределением признаков. Реализован предложенный метод отбора признаков для больших данных, используя технологию MapReduce. Для конкретной задачи классификации “Совершит ли пользователь клик на рекламный баннер” проведено сравнение работы двух методов отбора признаков. Изучена зависимость качества классификации от количества признаков. Доказана эффективность предложенного метода при наличии в данных сильно коррелирующих между собой признаков.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1] Battiti, R. Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks, 5, 537–550, 1994.

[2] Kwok, N., & Choi, C. Input feature selection for classification problems. IEEE Transactions on Neural Networks, 13, 143–159, 2002.

[3] Peng, H., Long, F., & Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27, 1226–1238, 2005.

[4] Estévez, P. A., Tesmer, M., Perez, A., & Zurada, J. M. Normalized mutual information feature selection. IEEE Transactions on Neural Networks, 20, 189–201, 2009.

[5] Hoque, N., Bhattacharyya, D. K., & Kalita, J. K. MIFS-ND: a mutual information-based feature selection method. Expert Systems with Applications, 41(14), 6371–6385, 2014.

[6] Yang, H., & Moody, J. Feature selection based on joint mutual information. In Proceedings of international ICSC symposium on advances in intelligent data analysis (pp. 22–25), 1999.

[7] Lee, J., & Kim, D. Fast multi-label feature selection based on information theoretic feature ranking. Pattern Recognition, 48, 2761–2771, 2015.

[8] Zhang, Y., Yang, C., Yang, A., Xiong, C. Y., Zhou, X., & Zhang, Z. Feature selection for classification with class-separability strategy and data envelopment analysis. Neurocomputing, 166, 172–184, 2015.

[9] Kraskov A., Stögbauer H., Grassberger P. Estimating mutual information. Physical Review E 69, 2004.

[10] Ross B.C. Mutual Information between Discrete and Continuous Data Sets. PLoS ONE 9(2): e87357, 2014.

[11] MapReduce-система Яндекса [Электронный ресурс], – https://habr.com/ru/company/yandex/blog/311104/ – статья в интернете.

[12] Fleuret F. Fast binary feature selection with conditional mutual information. Journal of Machine Learning Research, 5, 1531 – 1555, 2004.