МАТЕМАТИЧЕСКИЕ МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ И ВЫВОДА ПО ПРЕЦЕДЕНТАМ

Алексеева А.А, Тараник М.А.
(г. Томск, Томский политехнический университет)
e-mail: Alekseeva_92@sibmail.com

MATHEMATICAL METHODS OF INTELLIGENT DATA ANALYSIS AND OUTPUT ON THE CASE-BASED REASONING

Alekseeva A.A., Taranik M.A.
(Tomsk, Tomsk Polytechnik University)

Abstract: Case-based reasoning is a recent approach to problem solving and learning that has got a lot of attention over the last few years. This paper gives an overview of the foundational issues related to case-based reasoning, describes some of the leading methodological approaches within the field

При современном уровне развития интеллектуальных информационных систем все чаще встает вопрос о выборе эффективной системы поддержки принятия решения. Наряду с достаточно широко используемыми в ИИ методами правдоподобного вывода на основе индукции, абдукции, аргументации и аналогии в последнее время стали активно применяться методы на основе прецедентов [1-5]. Данные методы могут быть эффективны для мониторинга и поиска решения в проблемных ситуациях.

Прецедентный подход – это процесс (методология) решения новой задачи (проблемы) путем повторного использования и адаптации (при необходимости) решений, которые были ранее получены при решении подобных задач.

Суть прецедентного подхода заключается в применении накопленного опыта решения проблемы в процессе выработки решения новых задач. Данный подход базируется на принятии решения по аналогии.

Приведем основные характеристики прецедентов:

• прецедент представляет особое знание, привязанное к контексту, что позволяет использовать знания на прикладном уровне;

• прецеденты могут принимать различную форму (вид): охватывая разные по продолжительности промежутки времени; связывая решения с описаниями проблем; результаты с ситуациями и т.д.;

• прецедент фиксирует только тот опыт, который может обучить (быть полезным), фиксируемые прецеденты потенциально могут помочь специалисту (ЛПР) достичь цели, облегчить ее формулирование в будущем или предупредить его о возможной неудаче или непредвиденной проблеме

Основные типы метрик, которые могут быть использованы в задачах поиска близких прецедентов: Евклидова метрика, мера сходства Хемминга, вероятностная мера сходства, мера сходства Роджерса-Танимото, Манхэттенская метрика, расстояние Чебышева, метрика Махаланобиса, метрика Брея-Кертиса, метрика Чекановского, метрика Жаккара и обобщенное расстояние Евклида-Махаланобиса.[6]

Для проведения анализа и поиска наиболее популярных метрик был выбран электронный ресурс sciencedirect.

В ходе анализа были отобраны 3 метрики – Евклидова метрика, мера сходства Хемминга и расстояние Махаланобиса. Отбор осуществлялся в процентном соотношении, за основу были взяты статьи, в которых описывались применение и эффективность вывода при использовании данных метрик. Рассмотрим преимущества и недостатки каждой метрики:

Евклидова метрика

Евклидова метрика представляет собой геометрическое расстояние между двумя точками в многомерном пространстве. Данная величина применяется в методах таксономии, классификации и систематизации.

Расстояние Евклида показывает, как далеко друг от друга находятся два вектора, тем самым характеризуя их причастность к тому или иному классу, определенному вектором средних.

Вычисляется по классической формуле:

$$D = \sqrt{\sum_{k-1}^{n} (p_k - q_k)^2}$$

где n – мощность инварианта (количество характеристик);

p, q – сравниваемые векторы.

В задачах поиска близких прецедентов Евклидова метрика – применяется в методе "ближайшего соседа".

Преимущества метода

- Простота использования полученных результатов.
- Решения не уникальны для конкретной ситуации, возможно, их использование для других случаев.
- Целью поиска является не гарантированно верное решение, а лучшее из возможных.

Недостатки метода "ближайшего соседа"

- Данный метод не создает каких-либо моделей или правил, обобщающих предыдущий опыт, - в выборе решения они основываются на всем массиве доступных исторических данных, поэтому невозможно сказать, на каком основании строятся ответы.
- При использовании метода возникает необходимость полного перебора обучающей выборки при распознавании, следствие этого - вычислительная трудоемкость.
- Типичные задачи данного метода - это задачи небольшой размерности по количеству классов и переменных.

Мера сходства Хемминга

Расстояние Хэмминга является также расстоянием редактирования (определенной для строк одинаковой длины) с единственной допустимой операцией – заменой.

В более общем случае расстояние Хэмминга применяется для строк одинаковой длины любых k-ичных алфавитов и служит метрикой различия (функцией, определяющей расстояние в метрическом пространстве) объектов одинаковой размерности.

$$\boldsymbol{d_{ij} = \sum_{k=1}^{n} |x_{ik} - x_{jk}|}$$

Преимущества:

- Сеть работает предельно просто и быстро.
- Выходной сигнал (решение задачи) формируется в результате прохода через всего лишь один слой нейронов. Для сравнения: в многослойных сетях сигнал проходит через несколько слоев. В сетях циклического функционирования сигнал многократно проходит через нейроны сети, причем число итераций, необходимое для получения решения, бывает заранее не известно.
- Емкость сети Хемминга не зависит от размерности входного сигнала, она в точности равна количеству нейронов.

Основным недостатком расстояния редактирования Хэмминга является требование одинаковой длины строк, таким образом, расстояние Хэмминга подходит для расчета рас-

стояния редактирования с учетом таких искажений, как замена и транспозиция, но не подходит при вставках и удалениях

Сети Хемминга активно применяются в медицине при анализе геномов для разрешения проблем репродуктивной системы и развитии персонифицированной медицины.

Расстояние Махаланобиса

В математической статистике расстояние Махаланобиса – это мера расстояния между векторами случайных величин, обобщающая понятие расстояния Евклида. Оно отличается от него тем, что учитывает корреляции между переменными и инвариантно к масштабу. Данная величина широко используется в кластерном анализе и методах классификации

$$\boldsymbol{D_M(x) = \sqrt{(x-\mu)^T S^{-1} (x-\mu)}}$$

где x – вектор характеристик входного текста; μ – вектор средних для некоторого класса текстов; S – объединенная ковариационная матрица.

Данная метрика эффективна для анализа количественных данных. При её использовании учитывается зависимость признаков объекта, например, что очень важно для медицинских данных. Если данные имеют разную размерность и диапазон значений, свойства метрики Махаланобиса позволяют это учитывать. С помощью расстояния Махаланобиса можно определять сходство неизвестной и известной выборки.

Чтобы использовать расстояние Махаланобиса в задаче определения принадлежности заданной точки классу, нужно найти матрицу ковариации. Как правило, это делается на основе известных выборок. Затем необходимо подсчитать расстояние Махаланобиса от заданной точки до выделенного класса и оценить его.

Недостатком расстояния Махаланобиса является то, что данная мера расстояния плохо работает, если ковариационная матрица высчитывается на всем множестве входных данных. В то же время, будучи сосредоточенной на конкретном классе (группе данных), данная мера расстояния показывает хорошие результаты

Расстояние Махаланобиса позволяет определить качество проведения обследования пациента, у которого снимались значения ряда диагностических показателей, с помощью значения коэффициента уникальности данных пациента. Коэффициент уникальности определяется как вероятность того, что пациент с таким набором значений диагностических показателей встречается в базе данных пациентов.

Сравнительный анализ

Выбор соответствующей метрики довольно трудоемкая задача, от успешного решения которой непосредственно зависит результативность поиска и извлечения прецедентов. В каждом конкретном случае этот выбор производится по-разному, в зависимости от целей пользователя, физической и статистической природы используемой информации при управлении сложным объектом и других ограничений и факторов, влияющих на процесс поиска решения. В некоторых методах выбор соответствующей метрики достигается с помощью специальных алгоритмов преобразования исходного пространства признаков, в других – эксперт сам определяет метрику, опираясь на собственные знания о предметной области или экспериментальные данные.

Таблица 3 Сравнительный анализ выявленных метрик

Характеристики	Наименование метрики		
	Евклидова	Хемминга	Махаланобиса
Устойчивость к погрешностям	Неустойчивость к погрешностям	Неустойчивость к погрешностям	Неустойчивость к погрешностям
Простота реализации	Простая	Простая	Простая
Обработка большого количества	При неудачном выборе произ-	Ресурсоемкая задача	Плохо работает, если ковариаци-

данных	вольных начальных точек может получиться неоптимальная кластеризация		онная матрица высчитывается на всем множестве входных данных
Тип признаков	Количественные	Номинальные(качественные)	Количественные
Входные данные	Число кластеров необходимо знать заранее	Одинаковая длина строк	Конкретная группа данных, т. е. внутри кластера. Данная мера расстояния показывает хорошие результаты

Недостаток всех этих методов состоит в том, что они могут порождать разные решения в результате простого переупорядочения объектов в матрице расстояний и, кроме того, их результаты изменяются, если некоторые объекты исключаются из рассмотрения. Так же было установлено, что при больших значениях входных параметров работа алгоритма становится ресурсоемкой задачей и в случае зашумленности исходных данных или их неполноты методы становятся неэффективными

Представленные в настоящей статье достоинства и недостатки методов вывода по прецедентам характерны при их использовании в качестве единственного системного компонента. Такие модели вывода являются односложными. Использование гибридного подхода к формированию моделей позволяет устранить недостатки односложных. Среди применяемых дополнительных методов гибридных моделей наиболее распространенным является нечеткая логика. Ее использование позволяет повысить эффективность решения задач в реальных условиях неопределенности, а также сделать методы вывода более универсальными не зависимо от предметной области. Таким образом, дальнейшая работа по исследованию метрик вывода по прецедентам будет посвящена разработке гибридных моделей с использованием аппарата нечеткой логики.

В настоящей статье представлен анализ основных наиболее распространенных на практике метрик, используемых для задачи поиска прецедентов. Среди метрик были выявлены их достоинства и недостатки, а также предложен наиболее универсальный способ устранения с применением гибридного подхода.

ЛИТЕРАТУРА

1. Watson I.D., Marir F. Case-based reasoning: A review. The Knowledge Engineering Review, Vol. 9, No. 4, 1994. – PP. 355-381.

2. Карпов Л.Е., Юдин В.Н. Методы добычи данных при построении локальной метрики в системах вывода по прецедентам, М.: ИСП РАН, препринт № 18, 2006.

3. Варшавский П.Р., Еремеев А.П. Методы правдоподобных рассуждений на основе аналогий и прецедентов для интеллектуальных систем поддержки принятия решений // Новости искусственного интеллекта, № 3, 2006. – С. 39-62.

4. Варшавский П.Р., Еремеев А.П. Реализация методов поиска решения на основе аналогий и прецедентов в системах поддержки принятия решений // Вестник МЭИ, № 2, 2006. -С.77-87.

5. Берман А.Ф., Николайчук О.А., Павлов А.И., Юрин А.Ю. Использование прецедентов для обоснования мероприятий по предотвращению отказов механических систем // Труды 11-ой национальной конференции по ИИ с международным участием (КИИ-2008, г. Дубна, Россия). В 3-х т., Т. 2. –М: ЛЕНАНД, 2008. – С. 106-113.

6. Осипов Г.С. Методы искусственного интеллекта //ФИЗМАТЛИТ, 2011. - 296 С