УДК 519.81, 004.8

АНАЛИЗ ВОЗМОЖНОСТЕЙ МЕТОДА ПРОГНОЗИРОВАНИЯ ПО ПРЕЦЕДЕНТАМ

Свитина А.С. , Иващенко А.Б.
Донецкий национальный технический университет.
Кафедра компьютерных систем мониторинга.
E–mail: svitina@inbox.ru

Аннотация

Свитина А.С., Иващенко А.Б. Анализ возможностей метода прогнозирования по прецедентам. *Рассмотрено прогнозирование с помощью метода прецедентов: основные сферы применения, достоинства и недостатки метода. Определен основной алгоритм прогнозирование на основе прецедентов. Представлен метод "ближайшего соседа" или системы рассуждений на основе аналогичных случаев, как один из способов прогнозирования ситуации на основе прецедентов.*

Общая постановка задачи

Прецедент – это описание проблемы или ситуации в совокупности с подробным указанием действий, предпринимаемых в данной ситуации или для решения данной проблемы [1].

Прогнозирование с помощью метода прецедентов является противоречивым. С одной стороны он не достаточно точен в прогнозах, при его использовании можно получить не гарантированное верное решение, но с другой стороны он обладает, большим преимуществом по сравнению с другими методами прогнозирования:

– является логически обоснованным;

–область его применения достаточно широкая, что дает возможность использовать полученные результаты в других областях;

– основным источником информации о событии является опыт, а не теория.

Актуальность метода обусловлена многочисленностью задач, суть метода заключается в нахождении подходящего решения там, где нет четко сформулированного правила или метода.

Область применения метода прогнозирования с помощью прецедентов является очень широкой, примеры немногих из них представлены ниже:

Задачи медицинской диагностики

Можно решать различные задачи: классифицировать вид заболевания (дифференциальная диагностика); определять наиболее целесообразный способ лечения; предсказывать длительность и исход заболевания; оценивать риск осложнений; находить синдромы наиболее характерные для данного заболевания совокупности симптомов.

Ценность такого рода систем в том, что они способны мгновенно анализировать и обобщать огромное количество прецедентов – возможность, недоступная специалисту – врачу.

Предсказание месторождений полезных ископаемых

Признаками являются данные геологической разведки. Задача решается путём поиска закономерностей в имеющемся массиве данных. В процессе решения выделяются короткие наборы признаков, обладающие наибольшей информативностью – способностью наилучшим образом разделять классы.

Оценивание кредитоспособности заёмщиков

Эта задача решается банками при выдаче кредитов. Объектами в данном случае являются физические или юридические лица, претендующие на получение кредита. В случае

физических лиц признаковое описание состоит из анкеты, которую заполняет сам заёмщик, и, возможно, дополнительной информации, которую банк собирает о нём из собственных источников. Примеры бинарных признаков: пол, наличие телефона. Номинальные признаки – место проживания, профессия, работодатель. Порядковые признаки – образование, занимаемая должность. Количественные признаки – сумма кредита, возраст, стаж работы, доход семьи, размер задолженностей в других банках. Обучающая выборка составляется из заёмщиков с известной кредитной историей. В простейшем случае принятие решений сводится к классификации заёмщиков на два класса: «хороших» и «плохих».

Прогнозирование потребительского спроса

Для эффективного управления торговой сетью необходимо прогнозировать объёмы продаж для каждого товара на заданное число дней вперёд. На основе этих прогнозов осуществляется планирование закупок, управление ассортиментом, формирование ценовой политики, планирование промоакций. Для увеличения точности прогнозов необходимо также учитывать различные внешние факторы, влияющие на потребительский спрос: уровень инфляции, погодные условия, рекламные кампании, социально–демографические условия, активность конкурентов.

Принятие инвестиционных решений на финансовом рынке

В этой задаче умение хорошо прогнозировать самым непосредственным образом превращается в прибыль. Задача инвестора–спекулянта в том, чтобы правильно предугадать направление будущего изменения цены – роста или падения. Большой популярностью пользуются автоматические торговые стратегии – алгоритмы, принимающие торговые решения без участия человека [2].

Описание сути метода прогнозирования на основе прецедентов

Вывод на основе прецедентов – это метод принятия решений, в котором используются знания о предыдущих ситуациях или случаях (прецедентах). При рассмотрении новой проблемы (текущего случая) отыскивается похожий прецедент в качестве аналога. Вместо того, чтобы искать решение каждый раз сначала, можно пытаться использовать решение, принятое в сходной ситуации, возможно, адаптировав его к изменившейся ситуации текущего случая. После того, как текущий случай будет обработан, он вносится в базу прецедентов вместе со своим решением для его возможного последующего использования в будущем.

Прецедент включает:

– описание проблемы,

– решение этой проблемы,

– результат (обоснованность) применения решения.

Имеется множество способов представления прецедента: от записей в базах данных, древовидных структур – до предикатов и фреймов. Конкретное выбранное представление прецедентов должно соответствовать общим целям системы. Проблема представления прецедента – прежде всего проблема выбора информации, которую надо включать в описание прецедентов, нахождение соответствующей структуры для описания содержания прецедента, а также определения, каким образом должна быть организована и индексирована база знаний прецедентов для эффективного поиска и многократного использования [1].

Декомпозиция метода (основные фазы)

Подход, основанный на прецедентах, в целом состоит из следующих компонентов:

- извлечение релевантных прецедентов для текущего случая из библиотеки прецедентов;
- адаптация выбранного решения для текущего случая, если это необходимо;
- анализ решения, оценка применения (проверка корректности);
- сохранение;
- добавление текущего случая в базу прецедентов.

Более подробная схема метода изображена на рисунке 1.

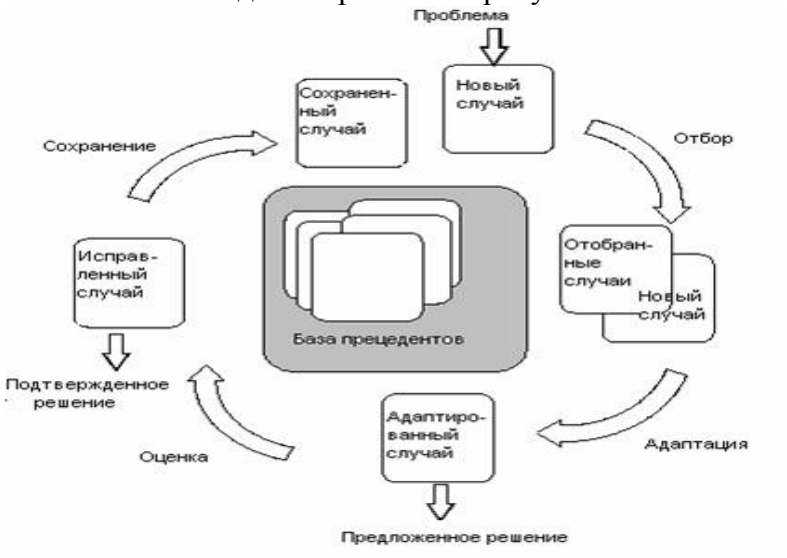

Рисунок 1. Декомпозиция метода

Проблема выбора подходящего прецедента является одной из самых важных в таких системах. Естественно искать подходящий прецедент в той области пространства поиска, где находятся решения сходных проблем, иначе говоря, поиск должен быть организован сообразно цели [1].

Метод "ближайшего соседа" или системы рассуждений на основе аналогичных случаев

Следует сразу отметить, что метод "ближайшего соседа" относится к классу методов, работа которых основывается на хранении данных в памяти для сравнения с новыми элементами. При появлении новой записи для прогнозирования находятся отклонения между этой записью и подобными наборами данных, и наиболее подобная (или ближний сосед) идентифицируется.

При таком подходе используется термин "k–ближайший сосед". Термин означает, что выбирается k "верхних" (ближайших) соседей для их рассмотрения в качестве множества "ближайших соседей". Поскольку не всегда удобно хранить все данные, иногда хранится только множество "типичных" случаев.

Данный метод по своей сути относится к категории "обучение без учителя", т.е. является "самообучающейся" технологией, благодаря чему рабочие характеристики каждой базы прецедентов с течением времени и накоплением примеров улучшаются. Разработка баз прецедентов по конкретной предметной области происходит на естественном для человека языке, следовательно, может быть выполнена наиболее опытными сотрудниками компании – экспертами или аналитиками, работающими в данной предметной области.

Преимущества метода "ближайшего соседа":

– 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;

– существует сложность выбора меры "близости" (метрики), высокая зависимость результатов классификации от выбранной метрики.

– при использовании метода возникает необходимость полного перебора обучающей выборки при распознавании, следствие этого – вычислительная трудоемкость [4].

Решение задачи прогнозирования методом "ближайшего соседа"

Рассмотрим подробно принципы работы метода k–ближайших соседей для решения задач классификации и регрессии (прогнозирования). Принцип работы метода k–ближайших соседей для решения задачи регрессии. Регрессионные задачи связаны с прогнозированием значения зависимой переменной по значениям независимых переменных набора данных.

Рассмотрим график, показанный на рисунке 2. Изображенный на ней набор точек (прямоугольники) получен по связи между независимой переменной x и зависимой переменной y (кривая на графике). Задан набор некоторых объектов (т.е. набор примеров); мы используем метод k–ближайших соседей для предсказания выхода точки запроса X по данному набору.

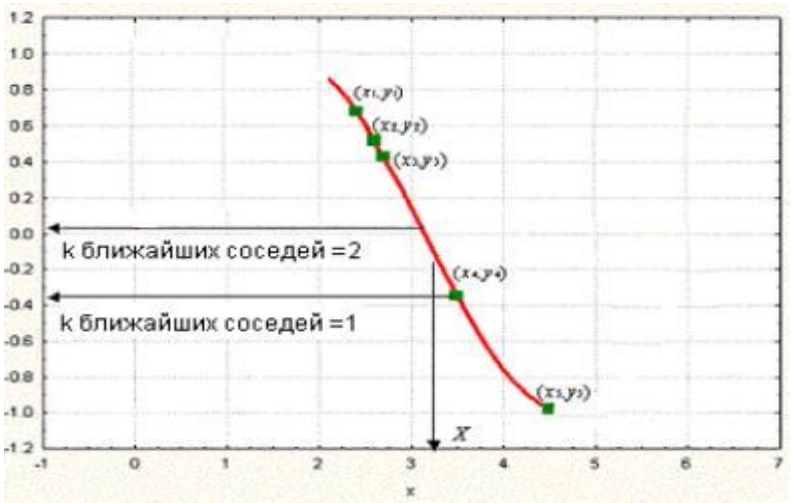

Рисунок 2. Графическое представление поиска соседей

Сначала рассмотрим в качестве примера метод k–ближайших соседей с использованием одного ближайшего соседа, т.е. при k, равном единице. Мы ищем набор примеров (прямоугольники) и выделяем из их числа ближайших к точке запроса X. Для нашего случая ближайший пример – точка $(x_4; y_4)$. Выход x_4 (т.е. y_4), таким образом, принимается в качестве результата предсказания выхода X (т.е. Y). Следовательно, для одного ближайшего соседа можем записать: выход Y равен y_4 (Y = y_4). Далее рассмотрим ситуацию, когда k равно двум, т.е. рассмотрим двух ближайших соседей. В этом случае мы выделяем уже две ближайшие к X точки. На нашем графике это точки y_3 и y_4 соответственно. Вычислив среднее их выходов, записываем решение для Y в виде Y = (y_3 + y_4)/2. Кроме арифметического среднего, применяют и другие методы адаптации решения. Решение задачи прогнозирования осуществляется путем переноса описанных выше действий на использование произвольного числа ближайших соседей таким образом, что выход Y точки запроса X вычисляется как среднеарифметическое значение выходов k–ближайших соседей точки запроса.

Независимые и зависимые переменные набора данных могут быть как непрерывными, так и категориальными. Для непрерывных зависимых переменных задача рассматривается как задача прогнозирования, для дискретных переменных – как задача классификации.

Предсказание в задаче прогнозирования получается усреднением выходов k–ближайших соседей, а решение задачи классификации основано на принципе "по большинству голосов". Критическим моментом в использовании метода k–ближайших

соседей является выбор параметра k. Он один из наиболее важных факторов, определяющих качество прогнозной либо классификационной модели.

Если выбрано слишком маленькое значение параметра k, возникает вероятность большого разброса значений прогноза. Если выбранное значение слишком велико, это может привести к сильной смещенности модели. Таким образом, мы видим, что должно быть выбрано оптимальное значение параметра k. То есть это значение должно быть настолько большим, чтобы свести к минимуму вероятность неверной классификации, и одновременно, достаточно малым, чтобы k соседей были расположены достаточно близко к точке запроса [3].

Выводы

Метод прогнозирования с помощью прецедентов является гибким методом. Он применим в различных сферах деятельности, прост в использовании.

Прогнозирование заключается в том, что при анализе нового события отыскивается похожий прецедент, в качестве аналога. Данный метод значительно упрощает поиск и решения и увеличивает его скорость.

Эксперт исследуемой области не способен хранить и использовать в полной мере огромной количество прецедентов, для этого специально созданы данные системы. Достоинства систем такого вида в том, что можно хранить большой объем информации, обобщать ее и анализировать.

Основной алгоритм прогнозирование на основе прецедентов включает:

1. Нахождение самых адекватных прецедентов для конкретных задач из базы данных прецедентов.

2.Нужно адаптировать найденное решение для конкретного случая, рассматриваемого в данный момент.

3. Проанализировав решение, применяем его к нашему случаю.

4. Проверяем корректность решения и сохраняем его.

5. Добавляем решение в базу данных прецедентов, и имеем решение задачи для конкретного случая на основе предыдущих прецедентов.

Основной тип представления данных при решении задачи прогнозирования методом "ближайшего соседа"– база данных прецедентов.

Технологией прогнозирования методом «ближайшего соседа» является накопление характеристики события, что гарантирует в дальнейшем улучшение с течением времени базы прецедентов. Основной целью поиска решений в методе «ближайшего соседа» является не гарантированно верное решение, а лучшее из возможных.

Литература

1. Методы добычи данных при построении локальной метрики в системах вывода по прецедентам. Режим доступа: http://citforum.ru/consulting/BI/data_mining/2.shtml#1 (15.03.2011).
2. Режим доступа: http://ru. wikipedia.org/wiki/Задачи_прогнозирования (15.03.2011).
3. Методы классификации и прогнозирования. Метод опорных векторов. Метод "ближайшегососеда".Режимдоступа:http://www.intuit.ru/department/database/datamining/10/datamining_10.html (12.03.2011).
4. Методы классификации и прогнозирования. Режим доступа: http://www.neural.forekc.ru/dm1/index_metod_quot_blijaishego_soseda_quot_ili_sistemy_rassujdenii_na_osnove_analogichnyh_sluchaev.htm (13.03.2011).