ДСМ-ПОДОБНЫЕ СИСТЕМЫ, ИСПОЛЬЗУЮЩИЕ АППАРАТ НЕЧЕТКОГО ВЫВОДА

О.М.Аншаков, В.А.Ковтун

РГГУ, кафедра МЛиИС

ДСМ-метод

Кто и когда предложил ДСМ-метод?

ДСМ-метод автоматического порождения гипотез был предложен В.К.Финном в конце 1970-х.

Для чего используется ДСМ-метод?

В настоящее время ДСМ-метод рассматривается как оригинальная совокупность технологий интеллектуального анализа данных, использующая *формализованные правила правдоподобных рассуждений*.

Что такое "ДСМ"

Аббревиатура "ДСМ" — это инициалы Джона Стюарта Милля, формализованные правила индуктивной логики которого являются идеологическим фундаментом ДСМ-метода.

Фазы ДСМ-метода

ДСМ-метод включает две фазы:

1. фазу обнаружения закономерностей в данных (фазу обучения),
2. фазу предсказания.

Соглашение

ДСМ-метод может предсказывать наличие или отсутствие *множества* целевых свойств. Но в этой работе мы будем предполагать, что целевое свойство (признак) у нас *единственное*.

При исследовании результатов фазы предсказаний ДСМ-метода уместно ставить вопрос о *полноте* и *точности* предсказаний.

Чтобы иметь возможность определять *полноту* и *точность* применительно к ДСМ-методу нам необходимо ввести некоторые термины.

Объект

Объектом в ДСМ-методе называется сущность, относительно которой задаются вопросы об обладании или не обладании набором целевых свойств (или одним целевым признаком). ДСМ-метод работает со *структурой* объекта, которая представляется, как правило, в виде *множества атомов* из некоторого *универсума*.

Введем следующие обозначения:

1. через $\mathbf{O}_{\text{Test}}$ будем обозначать множество *тестовых примеров* — множество объектов, для которых мы собираемся предсказывать наличие или отсутствие целевого признака,
2. через $\mathbf{O}^{+}_{\text{Test}}$ будем обозначать множество *положительных тестовых примеров* — множество объектов, для которых было предсказано *наличие* целевого признака,
3. через $\mathbf{O}^{-}_{\text{Test}}$ будем обозначать множество *отрицательных тестовых примеров* — множество объектов, для которых было предсказано *отсутствие* целевого признака,

4. через $\mathbf{O}^{+,\mathbf{T}}_{\text{Test}}$ будем обозначать множество *правильно предсказанных положительных тестовых примеров* — множество объектов, для которых было предсказано *наличие* целевого признака, и они действительно этим признаком *обладают*,
5. через $\mathbf{O}^{-,\mathbf{T}}_{\text{Test}}$ будем обозначать множество *правильно предсказанных отрицательных тестовых примеров* — множество объектов, для которых было предсказано *отсутствие* целевого признака, и они действительно этим признаком *не обладают*.

Через $|M|$, как обычно, будем обозначать *мощность множества* M.

Тогда *полноту* R (recall) и *точность* P (precision) фазы предсказаний ДСМ-системы можно определить с помощью следующих равенств:

$$R = \frac{\left|\mathbf{O}^{+}_{\text{Test}} \cup \mathbf{O}^{-}_{\text{Test}}\right|}{\left|\mathbf{O}_{\text{Test}}\right|}, \qquad P = \frac{\left|\mathbf{O}^{+,\mathbf{T}}_{\text{Test}} \cup \mathbf{O}^{-,\mathbf{T}}_{\text{Test}}\right|}{\left|\mathbf{O}^{+}_{\text{Test}} \cup \mathbf{O}^{-}_{\text{Test}}\right|}.$$

Мы предполагаем, что исходные данные представлены в виде *числовых векторов*. Точнее, следует рассматривать исходное представление каждого объекта как *строку таблицы реляционной базы данных*.

Введем следующие обозначения:

1. через U^{Attr} обозначим универсум атрибутов — множество заголовков столбцов таблицы, представляющей совокупность объектов (другими словами, U^{Attr} — это схема отношения, представленного таблицей);
2. исходное представление каждого объекта будем рассматривать как отображение $f : U^{\text{Attr}} \to \mathbf{R}$, где $\mathbf{R}$ — множество вещественных чисел;

3. в множестве U^{Attr} выделим два особых атрибута: ID — идентификатор объекта — и Tar — целевой признак;
4. через U^{MA} обозначим множество подлежащих обработке (анализируемых) атрибутов (minable attributes), т.е. атрибутов, отличных от ID и Tar

$$U^{\mathrm{MA}} = U^{\mathrm{Attr}} \setminus \{\mathrm{ID}, \mathrm{Tar}\};$$

5. нас будет интересовать проекция вещественного вектора, представляющего объект, на множество U^{MA}, т.е., сужение функции f на множество U^{MA}, это сужение будем обозначать через f^{MA}.

Исходную таблицу, чтобы ее можно было обрабатывать с помощью ДСМ-системы, необходимо представить в виде *семейства множеств*.

Очевидный способ такого представления состоит в том, что для каждого атрибута из U^{MA} задается *совокупность числовых промежутков*.

Например для атрибута "Температура тела человека" могут быть заданы промежутки: до 36°, от 36° до 37°, больше 37°.

Числовые промежутки задают разбиение домена атрибута на непересекающиеся подмножества. *Множество классов такого разбиения* для атрибута A будем обозначать через $\mathrm{Patrition}\,(A)$.

Универсум атомов

Универсум атомов будем обозначать через U^{Atom}. Положим по определению:

$$U^{\mathrm{Atom}} = \left\{ \langle A, C \rangle \,|A \in U^{\mathrm{MA}}, C \in \mathrm{Patrition}\,(A) \right\}.$$

Множество, соответствующее строке таблицы

Каждой строке f исходного табличного представления поставим в соответствие множество

$$\mathrm{Set}\,(f) = \left\{ \langle A, C \rangle \in U^{\mathrm{Atom}} | f\,(A) \in C \right\}.$$

Каждый числовой промежуток, на которые разбивается домен атрибута, можно интерпретировать как *лингвистический терм*.

В примере с температурой тела человека промежуткам “до 36°”, “от 36° до 37°” и “больше 37°” соответствуют лингвистические термы *“Пониженная”*, *“Нормальная”* и *“Повышенная”*, соответственно.

Каждому атому в этом случае соответствует *нечеткое множество* со своей *функцией принадлежности*.

Возможные причины снижения полноты

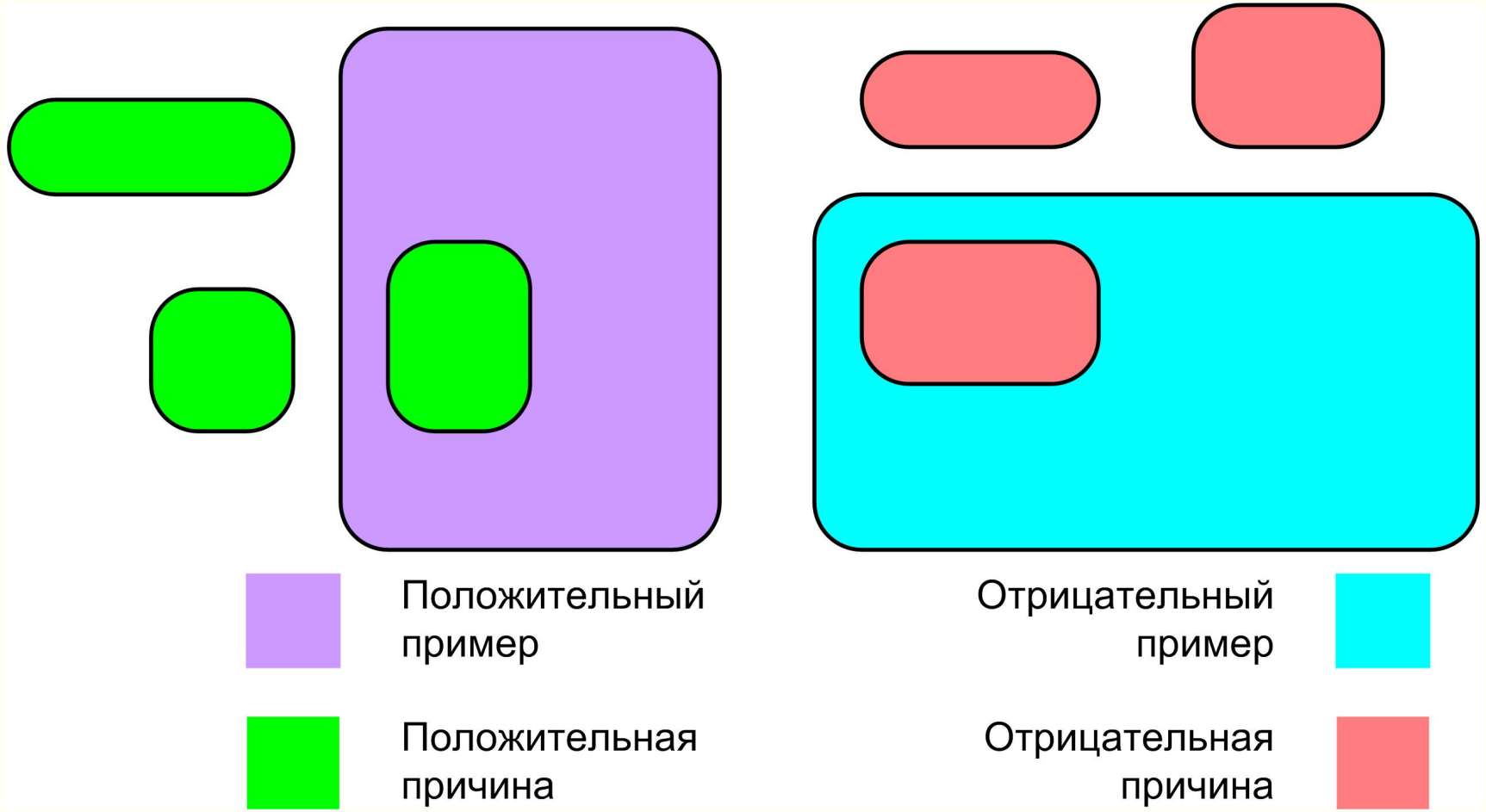
Неформальное объяснение

Говоря упрощенно, ДСМ-система предсказывает *наличие целевого признака* у неопределенного примера (т.е., объекта из тестовой выборки), если в объект (представленный в виде множества) включается хотя бы одна *причина наличия целевого признака* и не включается ни одной *причины отсутствия этого признака*.

ДСМ-система предсказывает *отсутствие целевого признака* двойственным образом, т.е., в том случае, если в объект включается хотя бы одна *причина отсутствия целевого признака* и не включается ни одной *причины наличия* этого признака.

Возможные причины снижения полноты

Неформальное объяснение



Возможные причины снижения полноты

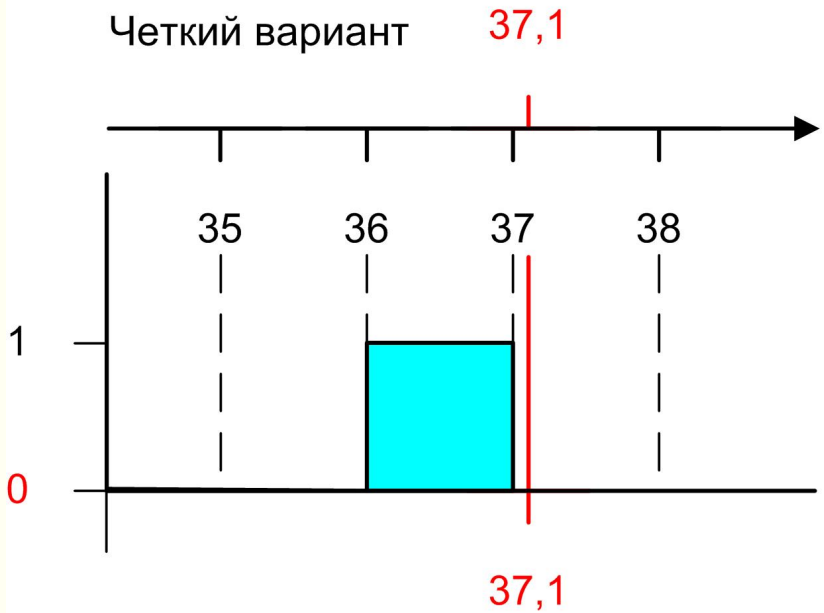
Неформальное объяснение

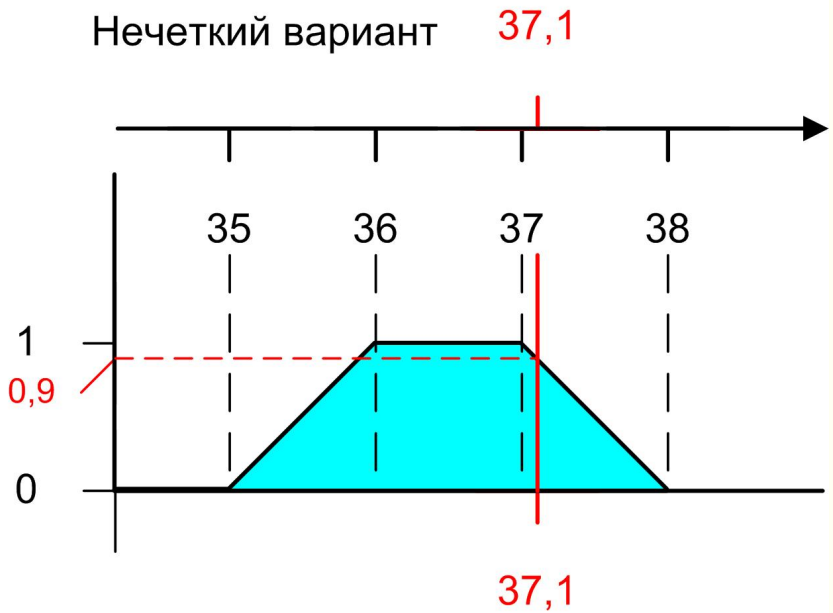
Включение в объект возможной причины означает принадлежность объекту каждого атома из этой возможной причины. Но атом $\langle A, C \rangle \in \mathrm{Set}\,(f)$ тогда и только тогда, когда $f\,(A) \in C$, т.е., когда значение атрибута A в исходном представлении объекта *попадает в числовой промежуток* C.

Но может случиться так, что значение атрибута A не попадает в промежуток C, но *находится рядом* с ним. Например, температура $35{,}9^{\circ}$ или $37{,}1^{\circ}$ не слишком отличаются от нормальной. Однако обычный ДСМ-метод в этом случае ничего не заметит и *не сформирует предсказание*.

Возможные причины снижения полноты

Неформальное объяснение





Как повысить полноту?

Чтобы повысить полноту, логично было бы говорить не о **попадании** или *непопадании* в промежуток, а о *степени принадлежности* нечеткому интервалу.

Что предлагается?

Разработанное приложение представляет собой *прототип системы анализа данных*, которая порождает гипотезы о возможных причинах следуя *алгоритмам ДСМ-метода*, а при предсказании наличия или отсутствия целевого признака использует *отдельные процедуры систем нечеткого вывода*.

Работу системы на фазе предсказаний можно интерпретировать следующим образом:

1. каждой *возможной причине* s *наличия целевого признака* Tar ставится в соответствие нечеткая продукция в стиле Мамдани, а именно

$$\bigwedge_{\langle A,C\rangle\in s} A \text{ is } C \to \text{Tar is Present},$$

   где C интерпретируется как лингвистический терм;

2. в случае, если s является *возможной причиной отсутствия целевого признака*, заключение продукции "Tar is Present" заменяется на "Tar is Absent";

3. значение каждой формулы A is C вычисляется как *степень принадлежности* нечеткому интервалу, соответствующему паре $\langle A, C \rangle$, где A рассматривается как *лингвистическая переменная*, C — как *лингвистический терм*, $f(A)$ в этом случае рассматривается как числовое значение переменной A;
4. для каждой причины s наличия целевого свойства значение посылки и заключения соответствующей продукции находится *так же*, как в *системах нечеткого вывода*;

1. итоговое значение формулы “Tar is Present” находится как *максимум* (или *алгебраическая сумма*) значений этой формулы для всех продукций, соответствующих причинам наличия свойства;
2. аналогично находится итоговое значение формулы “Tar is Absent”;
3. вывод о *наличии или отсутствии целевого признака* делается по разности значений формул “Tar is Present” и “Tar is Absent”, пороговое значение этой разности может меняться.

Источник данных

Для оценки эффективности работы системы использовались результаты исследований *онкологических заболеваний*, взятые из общедоступного репозитория UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/, архив 177548), содержащего данные для тестирования различных методов машинного обучения.

Объем

Общий объем выборки составляет 569 объектов, каждый включает в себя 30 признаков и 1 целевой атрибут — положительный или отрицательный результат анализа.

Оценка полноты и точности

Часть из поступивших на вход объектов отводилась на *тестовую выборку*. По этим объектам оценивалась *точность* и *полнота* предсказаний.

В экспериментах использовался прототип системы интеллектуального анализа данных, позволяющий порождать гипотезы как с помощью *обычных процедур* ДСМ-метода, так и с использованием *техники нечеткого вывода*.

Эксперименты показали, что нечеткий вариант всегда позволяет добиться *больших значений полноты*, по сравнению с обычным ДСМ-методом, при практически *не меняющейся точности*. Прирост полноты при разных условиях экспериментов: от 5% до 30%.

Был проведен ряд экспериментов, чтобы выявить необходимое количество примеров, требуемое для обучения системы.

Процент тест. прим.	Полнота		Точность	
	Нечеткий	ДСМ	Нечеткий	ДСМ
70	66,9	35,3	84	81
60	68,2	36,9	86	84
50	68,4	39,7	87	85
40	70,2	40,2	87	87
30	73,9	44,9	91	92
20	70,1	45,4	89	91

Из полученных результатов можно сделать вывод, что полнота предложенного метода *мало зависит* от объема обучающей выборки. Кроме этого, не исключена возможность так называемого *“переобучения”* системы, когда при слишком большой обучающей выборке точность на тестовых примерах падает.

Оптимальный процент обучающих примеров — 70%, что примерно составляет 398 из 569 объектов используемых в качестве входных данных.

Перспективы дальнейших исследований

Компьютерные эксперименты показали, что использование техники нечеткого вывода действительно *может повысить полноту* предсказаний ДСМ-метода *при сохранении или незначительной потере точности*.

Продолжение компьютерных экспериментов является одним из основных направлений дальнейшей работы. В этом направлении необходимо выполнить следующие действия:

1. провести *сравнение* результатов предсказаний различных ДСМ-систем с результатами предсказаний системы, использующей аппарат нечеткого вывода,
2. провести эксперименты с *различными наборами данных*, в том числе и *новыми данными*, для которых компьютерных экспериментов еще не проводилось.

Перспективы дальнейших исследований

Еще одним направлением дальнейшей работы является *доработка системы*. В этом направлении необходимо:

1. разработать модули, позволяющие пользователю *выбирать различные способы* разбиения домена атрибута на числовые промежутки,
2. добавить в систему возможность *работы с нечисловыми данными*,
3. добавить в систему возможность использования *различных алгоритмов* порождения гипотез о возможных причинах,
4. улучшить *пользовательский интерфейс*.

Спасибо за внимание!