УДК 004.832.3

РАЗРАБОТКА МЕТОДОВ СНИЖЕНИЯ ВЛИЯНИЯ ШУМА В АЛГОРИТМАХ ОБОБЩЕНИЯ

***Вадим Николаевич Вагин**, д-р техн. наук, проф. каф. ПМ,*
e-mail: vagin@appmat.ru,
Национальный исследовательский университет «МЭИ»,
http://mpei.ru,
***Александр Викторович Суворов**, д-р техн. наук,*
проф. каф. информационной безопасности,
e-mail:avsuvorov@list.ru,
Финансовый университет при правительстве Российской Федерации,
http://www.fa.ru,
***Марина Владимировна Фомина**, канд. техн. наук, доц. каф. ВТ,*
e-mail: m_fomina2000@mail.ru,
***Олег Леонидович Моросин**, канд. техн. наук, каф. ПМ,*
e-mail: oleg@morosin.ru,
Национальный исследовательский университет «МЭИ»,
http://mpei.ru

Целью представленной работы является исследование влияния шума в данных на работу алгоритмов обобщения, основанных на построении деревьев решений. Рассматриваются различные модели шума и различные способы внесения шума в обучающие и экзаменационные выборки. Для улучшения работы алгоритма обобщения предлагается использовать подход на основе аргументации. Приводятся результаты машинного моделирования, подтверждающие эффективность предложенных методов.

Ключевые слова: индуктивное формирование понятий; пересматриваемые рассуждения; аргументация; степени обоснования; немонотонный вывод, обобщение.

Работа выполнена при поддержке грантов РФФИ № 14-07-00862, 15-01-05567, 16-37-00309.

DOI: 10.21777/2312-5500-2016-3-59-68

В.Н. Вагин

Введение

Современные системы интеллектуального анализа данных имеют возможность перерабатывать и анализировать «сырые» данные, предоставляя извлеченную информацию скорее и успешнее, чем аналитик мог бы найти ее самостоятельно [1]. Одним из важных классов таких интеллектуальных систем являются системы индуктивного формирования понятий, которые имеют возможность обобщать опыт экспертов по управлению сложными техническими объектами и строить классы ситуаций, в которых принимались сходные решения. Системы индуктивного формирования понятий, таким образом, могут быть использованы в интеллектуальных системах поддержки принятия решений (ИСППР).

В настоящее время ИСППР работают с динамическими, сложно организованными техническими объектами и системами, которые плохо поддаются формализации. Человек-эксперт может успешно решать задачу управления сложной технической системой на основе накопленного опыта, используя информацию, поступающую от такой системы. Однако разнородная информация, поступающая с реальных объектов, иногда может быть зашумленной: неточной, недостоверной и даже противоречивой. Алгоритмы обобщения должны иметь возможность получать формальное описание опыта эксперта по управлению системой на основе анализа этой разнородной информации.

Таким образом, исследование проблемы влияния шума в исходных данных на точность индуктивных моделей, сформированных системой индуктивного формирования понятий, является важной проблемой, которая будет рассмотрена в данной статье.


М.В. Фомина

Индуктивный вывод на основе деревьев решений

Важной частью любой интеллектуальной системы является подсистема логического вывода. Традиционно основой процесса формирования рассуждений является дедуктивный вывод, основанный на получении заключения из посылок. Классические логические модели играют важную роль в экспертных системах, поскольку в таких системах необходимы средства логического вывода, позволяющие проводить рассуждения от фактов к заключениям. Однако задачи, решаемые при управлении сложными техническими объектами, часто являются некорректными в том смысле, что они требуют применения эвристик и не предполагают полноты знаний.


О.Л. Моросин

Реализация индуктивных рассуждений позволяет получить правдоподобные выводы. Одной из наиболее успешных моделей представления знаний для индуктивного вывода является модель деревьев решений.

Деревья решений используются при решении классификационных задач и реализуют процедуру отнесения предъявленного примера к одному из возможных классов на основании анализа свойств (атрибутов), приписанных этому примеру. Классами могут быть, например, множества ситуаций, в которых требуется выполнять однотипные управляющие действия. Дерево решений можно рассматривать, таким образом, как особую форму теста, задающего последовательность проверок значений атрибутов конкретного примера, для которого выполняется классификация.


А.В. Суворов

Классификация примера начинается с корня дерева решений, где выполняется проверка атрибута, приписанного данному узлу (тест для данного атрибута), затем выбирается путь для движения вниз по одной из ветвей дерева в соответствии со значением атрибута. Процесс повторяется в узле, которым заканчивается выбранная ветвь, и так далее до тех пор, пока не будет достигнут конечный узел (лист). Конечному узлу приписан один из возможных ответов (решение).

Представление знаний с помощью решающих деревьев с успехом было использовано в ряде систем обучения с учителем, например в алгоритмах ID3 и C4.5 Куинлана [2, 3].

Построение дерева решений выполняется на основе множества примеров, для которых заранее известен результат классификации. Такое множество примеров K называется обучающей выборкой. Дерево решений строится с корневого узла (вершина дерева) вниз к конечным узлам (листьям). Различные алгоритмы построения деревьев решений используют разные критерии выбора очередного атрибута и условия проверки. Например, в алгоритме ID3 [2] на каждом этапе построения для выбора атрибута, на основании которого происходит ветвление в данной точке, используется информационная связь между классификационным и исследуемым атрибутами. Эта связь между классификационным атрибутом и исследуемым атрибутом называется также приростом информации (*information gain*) и определяется на основе частоты появления значений признаков атрибута в тестовом множестве примеров.

Далее предлагается использовать модели деревьев решений вместе с продукцион-

ными моделями. Основными чертами таких моделей являются универсальность, простота реализации и удобство преобразования дерева решений в продукционные правила.

Шум в исходных данных

Как было показано в [4], исходными данными для решения задачи обобщения является обучающая выборка, которая содержит примеры формируемых понятий. При использовании признакового описания понятий обучающая выборка имеет вид таблицы, которая может храниться в базе данных (БД).

Допустим, что примеры в обучающих выборках K содержат шум, т. е. значения атрибутов могут быть искажены. Различные типы искажений будем называть моделями шума, которые будут рассмотрены далее. Причины возникновения шума изложены в [5].

Одним из основных параметров исследования является уровень шума. Пусть обучающая выборка K (размер обучающей выборки обозначим $|K| = m$) содержит описания примеров, причем для описания каждого примера используются r атрибутов A_1, A_2, …, A_r. Далее называем такие атрибуты информационными. Область допустимых значений каждого атрибута A_k обозначим $Dom(A_k)$. Выборка K может быть представлена таблицей с m строками и r столбцами, такая таблица имеет $N = m \cdot r$ ячеек. Каждая строка таблицы соответствует одному примеру, а каждый столбец – одному из атрибутов A_k, где $1 \le k \le r$.

Примеры в K, на основе которых формируется дерево решений, принадлежат нескольким различным классам. Для отнесения примеров к конкретным классам вводим специальный атрибут, обозначенный далее d . Такой атрибут назовем решением или решающим атрибутом, его область допустимых значений $Dom(d)$ содержит два или более возможных значений. Таким образом, каждый объект из обучающего множества задан значениями информационных атрибутов и значением решающего атрибута.

Уровень шума – это величина p_0, которая представляет вероятность того, что значение атрибута в обучающем или тестовом множестве будет отличаться от истинного. Таким образом, среди всех N ячеек $N \cdot p_0$ ячеек в среднем будут неверными. Моделирование шума включает в себя модели шума, а также методы внесения шума в таблицу.

Для исследования были выбраны две модели шума: «отсутствующие значения», «искаженные значения». В первом случае для заданного уровня шума с вероятностью p_0 известное значение атрибута в таблице удаляется. Второй вариант внесения шума связан с заменой известного значения атрибута на другое, допустимое, но неверное для данного примера. Значения для замены выбираются из областей $Dom(A_k)$, $1 \le k \le r$, величина p_0 определяет вероятность такой замены.

Если шум связан с отсутствием в таблице некоторых значений атрибутов, необходимо выбрать способ обработки «отсутствующих значений». Предлагается два пути: пропуск такого примера и восстановление отсутствующих значений, используя метод «ближайшего соседа» [6].

Существуют различные способы внесения шума в обучающие и экзаменационные множества [7]. Рассмотрим три варианта внесения шума в таблицу.

1. Шум вводится равномерно во всю таблицу с одинаковым уровнем шума для всех атрибутов.

2. Шум вводится равномерно в один или несколько явно указанных атрибутов.

3. Был предложен новый способ неравномерного внесения шума в таблицу. Здесь уровень шума для каждого столбца (информативный атрибут) отличается в зависимости от вероятности прохождения случайно выбранного примера через вершину дерева, помеченную этим атрибутом. При этом:

– суммарный шум, внесенный в обучающую выборку, соответствует заданному уровню шума;

– искажениям подвергаются все информативные атрибуты, значения которых проверяются в узлах дерева решений;

– чем более «важен» атрибут, тем выше уровень искажений для его значений.

Предложены принципы расчета уровня шума для третьей нерегулярной модели. Пусть дерево решений *T* было построено на основе обучающей выборки *K*. Очевидно, случайно выбранный пример пройдет далеко не через все узлы дерева. Следовательно, задача состоит в том, чтобы эффективно распределить этот шум между атрибутами в соответствии со статистическим анализом обучающей выборки, сохранив заданный средний уровень шума p_0.

Поставим в соответствие каждому атрибуту A_k коэффициент распределения шума S_k в соответствии с вероятностью прохождения некоторого примера через узел, помеченный A_k. Очевидно, *каждый* выбранный пример из *K* пройдет через корень дерева решений. Поэтому присвоим значение 1 соответствующему коэффициенту распределения шума S_k для наиболее информативного атрибута (корневого атрибута).

Все другие узлы дерева, которые не являются листьями, имеют одного предка и несколько потомков. Пусть один такой узел отмечен атрибутом A_i и имеет предка A_q. Ребро между этими узлами отмечено значением атрибута x_j , где $x_j \in Dom(A_q)$. Пусть m – количество примеров в *K* и m_j – число примеров в *K*, удовлетворяющих условию «значение атрибута A_q *равно* x_j».

Тогда норма распределения шума для атрибута A_i: $S_{A_i} = S_q \frac{m_j}{m}$.

Всем коэффициентам для атрибутов, не использующихся в дереве решения, присвоим значение 0.

Введем норму

$$S = \sum_{i=1}^{r} S_{A_i}.$$

Таким образом, каждый атрибут A_i подвергается воздействию шума уровня $d_{A_i} = \frac{S_{A_i}}{S} \cdot p_0 \cdot r$,

где p_0 – заданный средний уровень шума, r – количество атрибутов.

Нетрудно заметить, что $\left(\sum d_{A_i}\right)/r = p_0$. Таким образом, среднее значение шума остается равным заданному.

В дальнейшем будем рассматривать работу алгоритма обобщения при наличии шума в исходных данных. Наша цель – оценить точность классификации примеров в обучающих выборках при повышении уровня шума в них. В данной работе будет исследовано влияния шума двух первых типов.

Методы борьбы с шумом

Наличие шума в обучающих и тестовых выборках способно ухудшить результаты работы алгоритмов обобщения как на этапе обучения, так и на этапе «экзамена». Рассмотрим основные средства, позволяющие снизить влияние шума в данных на результаты обобщения.

Если мы имеем дело с моделью «неизвестные значения», наиболее разумным представляется заполнить потерянное значение атрибута данными. Из возможных методов восстановления неизвестных значений предлагается использовать метод «ближайших соседей» и метод «выбор среднего» [8].

В табл. 1 представлены данные о влиянии шума «отсутствующие значения» на точность классификации тестовых примеров для алгоритма C4.5. Представленные в таблице наборы данных взяты из коллекции данных UCI Repository of Machine Learning Datasets Калифорнийского университета [9]. Шум вносился равномерно во все информативные атрибуты тестового множества. Из табл. 1 видно, что метод «ближайших соседей» дает лучшие результаты, чем замена неизвестного значения на среднее. Табл. 2

демонстрирует влияние шума «искаженные значения» на точность классификации тестовых примеров для алгоритма C4.5. Из табл. 2 видно, что шум «искаженные значения» снижает точность классификации тестовых примеров существенно больше, чем шум «отсутствующие значения». При использовании модели шума «искажение значений» основной проблемой является невозможность определить, какие конкретные значения в тестовом множестве являются недостоверными и нуждаются в корректировке. Таким образом, для этой модели шума невозможно применить такие методы обработки зашумленных данных, как метод «ближайших соседей» и метод «выбор среднего».

Таблица 1

Влияние шума «отсутствующие значения» на точность классификации тестовых примеров для алгоритма C4.5

Наборы данных	**Метод обработки «зашумленных» примеров**	**Точность классификации примеров с шумом, %**				
		Нет шума	**Шум 5%**	**Шум 10%**	**Шум 20%**	**Шум 30%**
MONKS1	Выбор среднего	81,71	81,94	81,94	82,64	83,1
	k ближайших соседей		81,99	82,04	81,88	82,11
MONKS2	Выбор среднего	67,36	66,44	66,9	66,67	64,58
	k ближайших соседей		67,64	67,55	66,62	65,34
MONKS3	Выбор среднего	94,68	94,21	93,75	93,15	92,29
	k ближайших соседей		94,03	93,94	93,75	93,56

Таблица 2

Влияние шума «искаженные значения» на точность классификации тестовых примеров для алгоритма C4.5

Наборы данных	**Нет шума**	**Шум 5%**	**Шум 10%**	**Шум 15%**	**Шум 20%**
MONKS1	81,71	80,56	78,47	78,24	79,17
MONKS2	67,36	65,74	66,2	63,19	62,27
MONKS3	94,68	91,67	90,28	88,89	87,96

В следующем разделе мы рассмотрим подход, основанный на методе аргументации.

Аргументация как средство снижения влияния зашумленных данных

Под аргументацией обычно понимают процесс построения предположений, относительно некоторой анализируемой проблемы. Как правило, этот процесс включает в себя обнаружение конфликтов и поиск путей их решения. Наиболее перспективным для применения в задаче обобщения выглядит использование теории аргументации, основанной на пересматриваемых рассуждениях, предложенной Джоном Поллоком [10].

Более подробно аргументация, основанная на пересматриваемых рассуждениях, описана авторами в [11, 12]. Известно, что основным критерием качества для построенного обобщенного понятия является успешность классификации с помощью полученного набора правил $\mathbb{R}$ примеров тестовых выборок, то есть примеров, не входящих первоначально в обучающее множество *U*.

Предлагается использовать методы аргументации применительно к построенным наборам продукционных правил с целью получения улучшенного набора $\mathbb{R}^*$, способного классифицировать тестовые примеры с большей точностью, чем исходный набор $\mathbb{R}$.

Для получения обобщенных понятий в виде наборов продукционных правил используются обучающие выборки. Качество получаемых правил зависит в первую очередь от представительности обучающей выборки.

Базовая идея заключается в разбиении обучающей выборки примеров U на два подмножества $U1$ и $U2$ таких, что $U1 \cup U2 = U$, $U1 \cap U2 = \varnothing$, и раздельном обучении на каждом из подмножеств. В данной работе будем считать, что способ разбиения не детерминирован, однако $|U1| = |U2| = \frac{|U|}{2}$, если $|U|$ четно, и $|U1| = |U2| - 1 = \left\lfloor \frac{|U|}{2} \right\rfloor$ в противном случае. После разбиения обучение проводится независимо на каждом из подмножеств, при этом можно полностью абстрагироваться от конкретных механизмов обобщения (единственное требование – в результате работы алгоритма формируются правила классификации вида «**ЕСЛИ** <*условия*>, **ТО** <*искомое понятие*>»). Пусть на обучающих выборках $U1$ и $U2$ построены наборы правил $\mathbb{R}1 = \{R1_1, R1_2, \ldots, R1_p\}$ и $\mathbb{R}2 = \{R2_1, R2_2, \ldots, R2_q\}$, где $R1i$ и $R2_j$ – классификационные правила, полученные на $U1$ и $U2$, p и q – количество таких правил. Нашей целью является построение множества $\mathbb{R}^* = \mathbb{R}1 \cup \mathbb{R}2$ такого, что оно не порождает конфликтов при классификации примеров из обучающей выборки U. Критерием успешности полученных объединенных наборов правил $\mathbb{R}^*$ будет повышение точности распознания тестовых наборов данных, а именно отсутствие конфликтов при классификации всех тестовых примеров. Для построения $\mathbb{R}^*$ будут использованы методы теории аргументации.

Формализация проблемы обобщения в терминах аргументации

Как уже было сказано выше, пусть на обучающих выборках $U1$ и $U2$ построены наборы правил $\mathbb{R}1$ и $\mathbb{R}2$. Правила из $\mathbb{R}1$ и $\mathbb{R}2$ имеют вид «**ЕСЛИ** <*условия*>, **ТО** <*искомое понятие*>».

Такие правила можно считать пересматриваемыми правилами вывода для аргументационной системы. Далее мы будем записывать такие правила в форме аргументационных пересматриваемых правил вывода X|=>Y, где X – условия, а Y –значение решающего атрибута. Так, например, правило «*Если* **(A_2 = 1) & (A_1 = 1)** *то* **CLASS = 1**» можно записать в виде пересматриваемого правила аргументационной системы:

{(A_2 = 1) & (A_1 = 1)}|=> CLASS = 1.

Кроме того, предполагается, что возможно только два возможных значения решающего атрибута: CLASS = 1 и CLASS = 0. Следовательно, во всех пересматриваемых правилах заключение CLASS = 0 может быть заменено на ¬ (CLASS = 1).

Требуется определить, имеются ли конфликты между правилами $\mathbb{R}1$ и $\mathbb{R}2$, построенными на обучающих выборках $U1$ и $U2$. Для этого необходимо для каждого объекта $X_i = <z_1^i, z_2^i, \ldots, z_K^i>$ (z_k^i – значение атрибута A_k для объекта X_i) с решающим атрибутом d_i, принадлежащего обучающей выборке $U = U1 \cup U2$, проверить, порождает ли он конфликты на некотором наборе решающих правил $\mathbb{R}1 \cup \mathbb{R}2$.

Для формирования непротиворечивого множества, объединяющего $\mathbb{R}1$ и $\mathbb{R}2$, предлагается использовать механизм *степеней обоснования (justification degrees)* [13] для пересматриваемых правил вывода.

Для задания количественной оценки достоверности аргумента в системах аргументации применяется механизм степеней обоснования. В данной статье для задания степеней обоснования используется числовая шкала [0, 1], где 0 соответствует пораженному аргументу, 1 – наиболее обоснованному аргументу. Степени обоснования могут быть двух типов [13]:

1) степени обоснования исходных аргументов;
2) степени обоснования пересматриваемых правил.

Первый тип степеней обоснования присваивается каждому исходному аргументу, и представляет собой некую оценку достоверности источника, из которого получен данный аргумент. Второй тип степеней обоснования связан с неопределенностью пересматриваемых правил, которые предлагается использовать при построении бесконфликтных множеств классификационных правил в задаче обобщения. Степени обоснования пересматриваемых правил будем обозначать как $Jus(R_i)$, $R_i \in \mathbb{R}$.

Ставится задача: определить степени обоснования всех правил вывода таким образом, чтобы все конфликты, возникающие на обучающей выборке, стали разрешимыми. Приведем предлагаемую процедуру обучения для поиска таких степеней обоснования, что конфликты становятся разрешимыми.

Процедура обучения

1. Задать всем правилам $R1_i \in \mathbb{R}1$, $1 \le i \le |\mathbb{R}1|$, и $R2_j \in \mathbb{R}2$, $1 \le j \le |\mathbb{R}2|$, степень обоснования, равную 1. Задать все правила вывода в качестве пересматриваемых правил аргументационной системы.

2. Для каждого примера $X_i = <z_1^i, z_2^i,\ldots, z_K^i>$ из обучающей выборки выполнить следующие шаги:

2.1. Подать Arg_1: $a_1 = z_1^i$, Arg_2: $a_2 = z_2^i,\ldots$, Arg_K: $a_K = z_K^i$ на вход аргументационной системы в качестве начальных аргументов со степенью обоснования, равной 1. Выполнить поиск конфликтов в полученной системе аргументации.

2.2. Если система обнаруживает конфликты, то есть в графе вывода имеются два конфликтующих аргумента *Arg** и *Arg***, перейти к шагу 2.3, в противном случае – к шагу 2.1.

2.3. Выбрать аргумент Arg^+ из {*Arg**, *Arg***} такой, что его заключение совпадает со значением решающего атрибута *d*, и Arg^-, не совпадающий с *d*.

2.4. Получить два множества правил $\mathbb{R}c^+$ и $\mathbb{R}c^-$, таких что правила из $\mathbb{R}c^+$ поддерживают аргумент Arg^+, а $\mathbb{R}c^-$ поддерживают Arg^-.

2.5. Степень обоснования правил, относящих рассматриваемый объект к правильному классу следует увеличить. Для этого для всех $R_j \in \mathbb{R}c^+$, $1 \le j \le |\mathbb{R}c^+|$, пересчитаем значение функции $Jus(R_j)$, по формуле

$$Jus(R_j) = \begin{cases} Jus(R_j)(1+\Delta), & \text{если } Jus(R_j)(1+\Delta) < 1, \\ 1 & \text{в противном случае.} \end{cases}$$

Значение Δ выбирается в интервале (0, 1) эмпирически в зависимости от количества правил вывода в $\mathbb{R}1$ и $\mathbb{R}2$. В приведенном эксперименте $\Delta = 0{,}05$.

2.6. Степень же правил, производящих неверную классификацию следует понизить. Пересчитаем степень обоснования всех $R_i \in \mathbb{R}c^-$, $1 \le i \le |\mathbb{R}c^-|$ по формуле

$$Jus(R_j) = (1 - \Delta)Jus(R_j).$$

3. Провести классификацию примеров из обучающей выборки *U* с учетом полученных степеней обоснования. В случае если на тестовой выборке остались конфликты, выполнить пункты 2.1–2.6. В противном случае завершить обучение.

С помощью приведенной процедуры обучения можно построить улучшенный набор классификационных правил, объединяющий классификационные правила из $\mathbb{R}1$ и $\mathbb{R}2$. Далее приведем результаты экспериментов по применению предложенного метода.

Результаты экспериментов

Приведем основные результаты, полученные в ходе выполнения компьютерного эксперимента.

В качестве базового алгоритма индуктивного формирования понятий использовался классический алгоритм C4.5 [3]. В качестве тестовых данных использовался набор данных MONKS3 из репозитория UCI [9].

Для оценки результатов сравнивались результаты работы алгоритмов:

1. Классический алгоритм C4.5. Обучение проводилось на полной обучающей выборке *U* с помощью алгоритма C4.5. Классификация примеров из тестовой выборки происходит на полном множестве $\mathbb{R}$.

2. Алгоритм C4.5 с применением аргументации. Обучающее множество *U* делится на два подмножества *U*1 и *U*2. Проводится независимое обучение на каждом подмножестве, и получаются два множества классификационных правил $\mathbb{R}1$ и $\mathbb{R}2$.

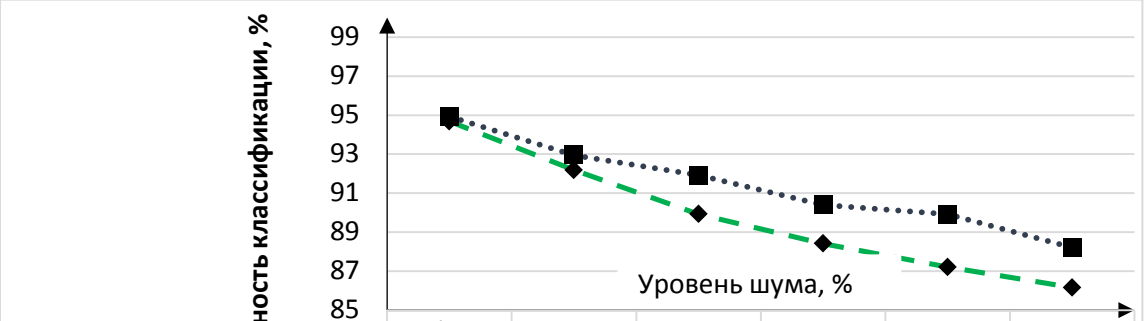
Множество $\mathbb{R}^*$ получается применением аргументационного подхода. Классификация проводится на объединенном множестве $\mathbb{R}^*$.

Были рассмотрены три типа шума: внесение искажений равномерно во все информационные атрибуты, внесение искажений в решающий атрибут и внесение шума типа «отсутствующие значения» в информационные атрибуты.

При внесении шума равномерно во все информационные атрибуты наблюдалось плавное снижение качества получаемых классификационных моделей, при этом применение аргументации позволило несколько снизить влияние шума (рис. 1).

Были проведены три новых эксперимента по применению аргументации при следующих типах шума: внесение искажений равномерно во все информационные атрибуты, внесение искажений в решающий атрибут и внесение шума типа «отсутствующие значения» в информационные атрибуты.

При внесении шума равномерно во все информационные атрибуты наблюдалось плавное снижение качества получаемых классификационных моделей. Применение аргументации позволило несколько снизить влияние шума (см. C4.5 + аргументация на рис. 1). Применение аргументации при таком типе шума позволило улучшить результаты классификации за счет отбраковки некоторых правил, полученных на основе примеров, содержавших искажения.



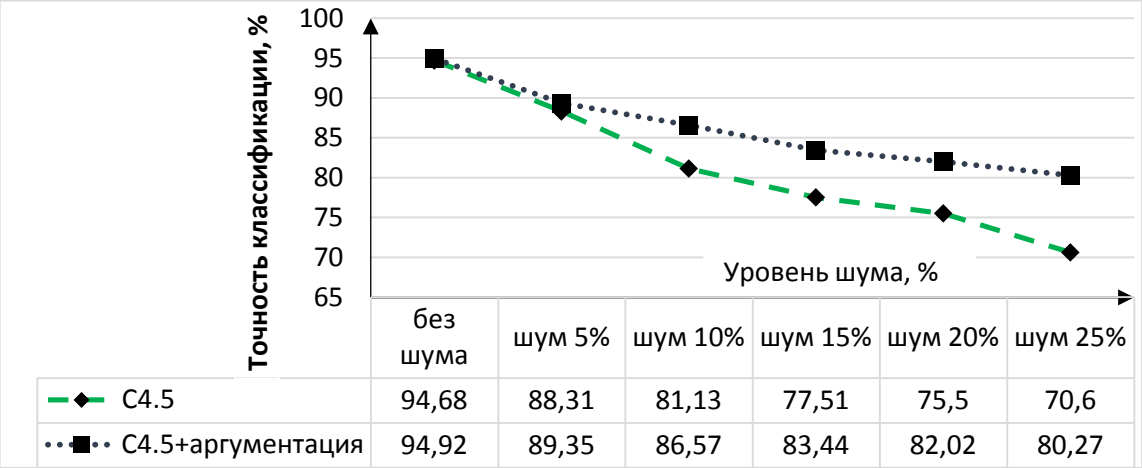
	без шума	шум 5%	шум 10%	шум 15%	шум 20%	шум 25%
C4.5	94,68	92,18	89,93	88,42	87,21	86,16
C4.5+аргументация	94,92	92,95	91,92	90,4	89,92	88,21

Рис. 1. Равномерное распределение шума типа «искажение» по всем информационным атрибутам

Внесение шума в решающий атрибут имеет наиболее сильное негативное влияние на результаты обобщения, так как при возрастании уровня шума в обучающей выборке увеличивается количество примеров, для которых неверно указан класс. Такие примеры сильно влияют на качество модели, так как на их основе создаются неверные классификационные правила, применение которых существенно ухудшает качество классификации. Предложенный метод применения аргументации позволяет находить и удалять такие правила из результирующего множества классификационных правил, что приводит к существенному улучшению (см. C4.5 + аргументация на рис. 2) результатов классификации.

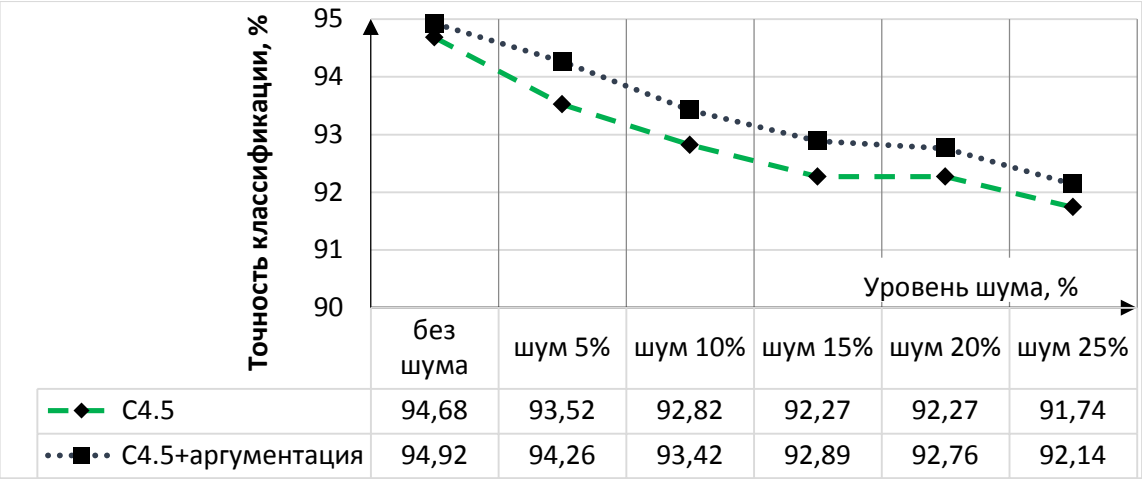
Кроме того, были проведены эксперименты по выявлению влияния шума типа «отсутствующие значения» на результаты обобщения. Для борьбы с данным типом шума применялся алгоритм восстановления отсутствующих значений методом ближайших соседей [8]. Восстановление отсутствующих значений среди информационных атрибутов позволяет успешно справляться с шумом при условии, что уровень шума невысок (5–10%). Шум «отсутствующие значения» оказывает наименьшее влияние на результаты работы алгоритма обобщения по сравнению с другими рассмотренными типами шума. Тем не менее, несмотря на то что алгоритм восстановления отсутствующих значений показывает довольно хорошие результаты, восстановленные значения не все-

гда оказываются верными и применение аргументации позволяет частично скорректировать результаты его работы, что приводит к увеличению качества классификационных моделей на 1–3%. (рис. 3).



	без шума	шум 5%	шум 10%	шум 15%	шум 20%	шум 25%
C4.5	94,68	88,31	81,13	77,51	75,5	70,6
C4.5+аргументация	94,92	89,35	86,57	83,44	82,02	80,27

Рис. 2. Внесение шума типа «искажение» в решающий атрибут обучающей выборки



	без шума	шум 5%	шум 10%	шум 15%	шум 20%	шум 25%
C4.5	94,68	93,52	92,82	92,27	92,27	91,74
C4.5+аргументация	94,92	94,26	93,42	92,89	92,76	92,14

Рис 3. Равномерное распределение шума типа «отсутствующее значение» по всем информационным атрибутам

Заключение

Как показало проведенное исследование, успешность работы алгоритмов обобщения тесно связана с использованием качественных обучающих выборок. Однако при работе с реальными наборами данных задача получения «чистых» обучающих выборок, не содержащих искажений и неточностей, является весьма сложной задачей, и на практике часто приходится использовать данные, которые заведомо могут содержать искаженную и даже противоречивую информацию. В таких условиях разработка методов и алгоритмов, позволяющих снизить влияние шума в обучающих выборках, является крайне важной. В данной работе были рассмотрены основные типы шумов, которые встречаются в обучающих множествах и предложены различные методы снижения их влияния. В частности, для борьбы с шумами предложен метод, основанный на применении аппарата аргументации. Из результатов машинного эксперимента можно сделать вывод, что метод аргументации наиболее эффективен при наличии шума «искаженные значения» в решающем атрибуте обучающего множества. Такие шумы приводят к наиболее ощутимым потерям качества классификационных моделей, поскольку наличие неверно классифицированных объектов в обучающих выборках часто приводит к формированию неверных классификационных правил. Применение методов аргумен-

тации позволило уменьшить влияние таких некорректных правил вывода за счет их согласования, что является значимым результатом.

Литература

1. *Finn V.K.* The synthesis of cognitive procedures and the problem of induction // NIT [Moscow, Russia: VINITI]. 1999. Series 2 (1–2). P. 8–44.
2. *Quinlan J.R.* Induction of Decision Trees // Machine Learning. 1986. Vol. 1. P. 81–106.
3. *Quinlan J.R.* Improved Use of Continuous Attributes in C4.5 // Journal of Artificial Intelligence Research. 1996. Vol. 4. P. 77–90.
4. *Вагин В.Н., Головина Е.Ю., Загорянская А.А., Фомина М.В.* Достоверный и правдоподобный вывод в интеллектуальных системах / под ред. В.Н. Вагина, Д.А. Поспелова. 2-е изд., доп. и испр. – М.: Физматлит, 2008. 712 с.
5. *Mookerjee V.S., Mannino M.V., Gilson R.* Improving the Performance Stability of Inductive Expert Systems under Input Noise // Information Systems Research. 1995. Vol. 6. No. 4. P. 328–356.
6. *Vagin V., Fomina M.* Methods and Algorithms of Information Generalization in Noisy Databases // Advances in Soft Computing: 9th Mexican Intern. Conference on AI, MICAI, Pachuca, 2010. P. 44–55.
7. *Fomina M., Eremeev A., Vagin V.* Noise models in Inductive Concept Formation // Proceedings of ICEIS 2013: 15th International Conference on Enterprise Information Systems. –Angers, France, 2013. Vol. 1. P. 413–419.
8. *Vagin V., Fomina M.* Problem of Knowledge Discovery in Noisy Databases // International Journal of Machine Learning and Cybernetics. 2011. Vol. 2. No. 3. P. 135–145.
9. *Merz C., Murphy P.* UCI Repository of Machine Learning Datasets. – Information and Computer Science University of California, 1998. http://archive.ics.uci.edu/ml.
10. *Pollock J.L.* How to Reason Defensibly // Artificial Intelligence. 1992. Vol. 57. P. 1–42.
11. *Вагин В.Н., Моросин О.Л.* Обзор методов нахождения степеней обоснования в системах аргументации // 14-я национальная конференция по искусственному интеллекту с международным участием КИИ-2014. Труды конференции. Т. 1. – Казань: Школа, 2014. С. 5 – 13.
12. *Моросин О.Л.* Аргументация с применением степеней обоснования в интеллектуальных системах. // Известия ЮФУ. Технические науки. 2014. № 7. С. 142–152.
13. *Pollock J.L.* Defeasible reasoning with variable degrees of justification // Artificial intelligence. 2001. Vol. 133. No. 1. P. 233–282.

Development of methods for decreasing noise influence on generalization algorithms

***Vadim Nikolaevich Vagin**, prof., department of Applied Mathematics, National Research University "MPEI"*

***Alexander Viktorovich Suvorov**, prof., department of information security Financial University under the Government of the Russian Federation*

***Marina Vladimirovna Fomina**, PhD, department of Applied Mathematics, National Research University "MPEI",*
***Oleg Leonidovich Morosin**, PhD, department of Applied Mathematics, National Research University "MPEI"*

This paper is devoted to study of the influence of noise in data on the work of generalization algorithms based on building decision trees. Different types of noise and various ways of introducing noise in the learning and test sets are viewed. To improve the efficiency of generalization algorithms, it is proposed to use an argumentation based approach. The results of computer simulation, confirming the effectiveness of the proposed methods and algorithms are presented.

Keywords: inductive notion formation; defeasible reasoning; argumentation; justification degrees; non-monotonic reasoning; generalization.