### Лабораторная работа №16 «Выбор оптимального решения с помощью дерева

**решений»**

**Цель работы:** освоение основных методов и способов построения деревьев решений, приобретение практических навыков по использованию инструментария Deductor 4.

### Краткая теория

Деревья решений (decision trees) являются одним из самых мощных средств решения задачи отнесения какого-либо объекта (строчки набора данных) к одному из заранее известных классов. Дерево решений – это классификатор, полученный из обучающего множества, содержащего объекты и их характеристики, на основе обучения. Дерево состоит из узлов и листьев, указывающих на класс.

Результатом работы алгоритма является список иерархических правил образующих дерево. Каждое правило – это интуитивно-понятная конструкция вида «Если…то…» (if - then). Дерево может использоваться для классификации объектов, не вошедших в обучающее множество. Чтобы принять решение, к какому классу следует отнести некоторый объект или ситуацию, требуется ответить на вопросы, стоящие в узлах этого дерева, начиная с его корня. Вопросы имеют вид

«значение параметра А больше В?». Если ответ положительный, осуществляется переход к правому узлу следующего уровня; затем снова следует вопрос, связанный с соответствующим узлом и т. д.

Настройка назначения полей

Необходимо определить, как будут использоваться поля исходного набора данных при обучении дерева и дальнейшей практической работе с ним.

В левой части окна представлен список всех полей исходного набора данных. Для настройки поля следует выделить его в списке, при этом в правой части окна будут отображены текущие параметры поля:

1. Имя поля – идентификатор поля, определенный для него в источнике данных. Изменить его здесь нельзя.
2. Тип данных – тип данных, содержащихся в поле (вещественный, строковый, дата). Он также задается в источнике данных и здесь изменен быть не может.
3. Назначение – здесь необходимо выбрать порядок использования данного поля при обучении и работе дерева решений. Выбор производится с помощью списка, открываемого кнопкой и содержащего следующие варианты:
   * Входное – значения поля будут являться исходными данными для построения и дальнейшей практической работы дерева решений, на их основе будет производиться классификация.
   * Выходное – будет содержать результаты классификации. Выходное поле может быть только одно и оно должно быть дискретным.
   * Информационное – поле не будет использоваться при обучении дерева, но будет помещено в результирующий набор в исходном состоянии.
   * Неиспользуемое – поле не будет использоваться при построении и работе дерева решений и будет исключено из результирующей выборки. В отличие от непригодного, такое поле может быть использовано, если в этом возникнет необходимость.
   * Непригодное – поле не может быть использовано при построении и работе алгоритма, но будет помещено в результирующий набор в исходном состоянии.
4. Вид данных – указывает на характер данных, содержащихся в поле (непрерывный или дискретный). Изменить это свойство здесь нельзя.

Статус непригодного поля устанавливается только автоматически и в дальнейшем может быть изменен только на неиспользуемое или информационное. Поле будет запрещено к использованию если:

* + поле является дискретным и содержит всего одно уникальное значение;
  + непрерывное поле с нулевой дисперсией;
  + поле содержит пропущенные значения.

В случае если текущее поле содержит непрерывные (числовые) данные, отображается секция «Статистика», где показываются максимальное и минимальное значения поля, его среднее значение и стандартное отклонение. Если выделенное поле содержит дискретные (строковые) данные, то для него открывается секция «Уникальные значения», в которой отображается общее число уникальных значений поля, а также список самих уникальных значений.

Нормализация полей

Целью нормализации значений полей является преобразование данных к виду, наиболее подходящему для обработки средствами пакета Deductor. Для дерева решений данные, поступающие на вход, должны иметь числовой тип. В этом случае нормализатор может преобразовать дискретные данные к набору уникальных индексов.

Окно настройки нормализации полей вызывается с помощью кнопки «Настройка нормализации». В окне слева приведен полный список входных и выходных полей. При этом каждое поле помечено значком, обозначающим вид нормализации поля:

* + линейная – линейная нормализация исходных значений;
  + уникальные значения – преобразование уникальных значений в их индексы.

Для числовых (непрерывных) полей с линейной нормализацией дополнительные параметры недоступны. В полях «Минимум» и «Максимум» секции «Диапазон значений» можно посмотреть минимальное и максимальное значения этого поля.

Для дискретных полей справа находится список уникальных значений поля, где для каждого значения указывается количество его повторений. Поле «Количество значений» показывает общее число уникальных значений, принимаемых полем.

Настройка обучающей выборки

Обучающая выборка может быть разбита на три множества – обучающее, тестовое и валидационное.

1. Обучающее множество – включает записи (примеры), которые будут использоваться в качестве входных данных, а также соответствующие желаемые выходные значения.
2. Тестовое множество – также включает записи, содержащие входные и желаемые выходные значения, но используемое не для обучения модели, а для проверки его результатов.
3. Валидационное множество – множество примеров, используемое как для оценки результатов обучения модели, так и определения ее параметров.

Для разбиения исходного множества на обучающее, тестовое и валидационное необходимо настроить несколько параметров:

1. Из списка «Способ разделения исходного множества» выбирается порядок отбора записей во все три множества. Если выбран вариант «по порядку», то порядок следования записей при их разделении не меняется. Множества последовательно формируются в соответствии с определенным для них числом записей. Если выбран вариант «случайно», то отбор записей происходит случайным образом.
2. Затем необходимо указать, какие множества будут использоваться. Для того чтобы множество было сформировано, нужно установить флажок слева от его названия. Если флажок сброшен, то множество использовано не будет. Обучающее множество используется всегда, поэтому сбросить флажок для него нельзя.
3. Для каждого из используемых множеств необходимо задать его размер. Размер может быть задан непосредственно количеством записей или в процентах от объема исходной выборки. Для этого достаточно дважды щелкнуть мышью в соответствующей клетке и ввести нужное значение с клавиатуры. При этом размер, введенный в процентах, автоматически пересчитывается в количество строк и наоборот. В поле «Количество строк (всего)» отображается общее количество записей в исходной выборке данных, которое может быть задействовано для формирования множеств. Если суммарное число строк для всех используемых множеств меньше полного числа строк исходной выборки, то размеры множеств можно задавать произвольно. Можно, например, использовать не все записи, а только часть из них. Если же суммарное указанное число строк превышает максимальное для данной исходной выборки, то автоматически включается баланс множеств, т.е. при указании для одного из множеств размера, в результате которого будет превышено максимальное число записей в исходной выборке, размер остальных множеств будет автоматически уменьшен таким образом, чтобы суммарный размер множеств не превышал доступного числа записей. В строке

«Итого» указывается количество записей, задействованных во всех используемых множествах, а также процент от полного числа записей исходной выборки, который они составляют.

1. В столбце «Порядок сортировки» можно определить порядок следования записей внутри каждого множества. Для этого необходимо дважды щелкнуть мышью в столбце

«Порядок сортировки» для соответствующего множества и с помощью появившейся кнопки выбора открыть список, в котором выбрать один из возможных вариантов:

* + по возрастанию – записи в данном множестве будут следовать в порядке возрастания;
  + по убыванию – записи в данном множестве будут следовать в порядке убывания;
  + случайно – записи в данном множестве будут следовать в случайном порядке.

Для того чтобы обучающее множество было репрезентативным необходимо, чтобы все уникальные значения всех дискретных столбцов содержались в данном наборе данных.

Настройка параметров обучения

Необходимо установить параметры, в соответствии с которыми будет проводиться обучение дерева:

1. «Минимальное количество примеров, при котором будет создан новый узел» – задается минимальное количество примеров, которое возможно в узле. Если примеров, которые попадают в данный узел, будет меньше заданного – узел считается листом (т.е. дальнейшее ветвление прекращается).
2. «Строить дерево с более достоверными правилами в ущерб сложности» – установка данного флажка включает специальный алгоритм, который, усложняя структуру дерева, увеличивает достоверность результатов классификации. Сброс данного флажка хотя и приводит к упрощению дерева, снижает достоверность результатов классификации.
3. «Уровень доверия, используемый при отсечении узлов дерева». Значение этого параметра задается в процентах и должно лежать в пределах от 0 до 100. Эти значения выбираются из списка. Чем больше уровень доверия, тем более ветвистым получается дерево, и, соответственно, чем меньше уровень доверия, тем больше узлов будет отсечено при его построении.

### Задание

1. Разработать сценарии построения дерева решений и проведения анализа «что - если».
2. По таблице (например, продаж) создать таблицу транзакций с полями (например, Менеджер, Организация, Вид товара). Таблицу получить путем слияния соответствующих полей из разных таблиц и последующей группировки.
3. Разработать сценарии построения дерева решений с представлением правил, наиболее популярных наборов и анализа «что - если» с входными полями (например, Менеджер и Организация) и выходным полем (например, Вид товара).
4. Создать отчеты по всем разработанным сценариям.
5. Продемонстрировать проект преподавателю с использованием тестовых наборов данных и защитить работу.

### Контрольные вопросы

1. С какой целью проводится нормализация значений полей?
2. Для чего используется обучающая выборка? Из каких множеств она состоит?
3. Какие критерии используются для выбора параметров обучения?
4. Какие требования предъявляются к исходным данным при построении дерева решений?
5. Поясните смысл расчетных полей при анализе «что - если».