УДК 004.93

Е.А. Гофман[1], А.А. Олейник[2], С.А. Субботин[3]

[1]Запорожский национальный технический университет,
г. Запорожье, Украина, gofman_jenek@rambler.ru;
[2]Запорожский национальный технический университет,
г. Запорожье, Украина, olejnikaa@gmail.com;
[3]Запорожский национальный технический университет,
г. Запорожье, Украина, subbotin@zntu.edu.ua

СИНТЕЗ ДЕРЕВЬЕВ РЕШЕНИЙ НА ОСНОВЕ ТЕОРИИ ПРИБЛИЖЕННЫХ МНОЖЕСТВ

Предложен новый метод идентификации деревьев решений с использованием теории приближенных множеств. Созданный метод позволяет сократить время работы и повысить эффективность синтезируемых моделей сложных объектов и систем. Проведены эксперименты по решению тестовых задач.

ДЕРЕВО РЕШЕНИЙ, ИДЕНТИФИКАЦИЯ, ПРИБЛИЖЕННОЕ МНОЖЕСТВО, ОБУЧАЮЩАЯ ВЫБОРКА

Введение

Теория приближенных множеств [1, 2] является важным математическим инструментом в области компьютерных технологий. В частности, теория приближенных множеств используется при принятии решений, интеллектуальном анализе данных, представлении знаний и др. Основным направлением, для которого может применяться теория приближенных множеств, является задача сокращения размерности входных данных, которая часто возникает при разработке различных информационных систем. Это объясняется тем, что практически любая система описывается набором признаков, определение каждого из которых зачастую требует значительных затрат материальных и временных ресурсов. Существуют различные методы, позволяющие находить сокращённый набор признаков для дальнейшей классификации на основании использования теории приближенных множеств [3, 4].

Таким образом, актуальным является применение теории приближенных множеств для синтеза моделей исследуемых объектов и систем при решении задач классификации. В данной работе предлагается выполнять идентификацию деревьев решений с предварительным применением теории приближённых множеств для обеспечения возможности построения распознающих моделей на основе сокращённого набора признаков.

Целью данной статьи является исследование и разработка метода синтеза деревьев решений с использованием теории приближённых множеств, который должен позволить повысить эффективность и скорость синтеза деревьев решений за счёт использования сокращённого набора признаков, полученного при помощи теории приближённых множеств.

Для достижения поставленной цели необходимо решить следующие задачи:

– исследование теории приближённых множеств;

– выделение операций и подходов теории приближённых множеств, которые можно использовать для решения поставленной цели;

– рассмотрение теории построения деревьев решений;

– интеграция теории приближенных множеств в теорию синтеза деревьев решений, в результате чего необходимо разработать новый метод синтеза деревьев решений;

– проведение экспериментов и сравнение разработанного метода с существующими методами идентификации деревьев решений.

1. Теория приближённых множеств

Теория приближенных множеств (rough sets) была разработана [1, 2] как математический подход для описания неопределенности, неточности и неуверенности. Эта теория основана на утверждении, что с каждым объектом универсального множества связана некоторая информация (данные, знания). Объекты, характеризуемые одинаковой информацией, являются неразличимыми (сходными). Отношение неразличимости, порождаемое таким способом, является математической основой теории приближенных (грубых) множеств.

Основой концепции теории приближенных множеств являются операции аппроксимации множеств.

В теории приближенных множеств таблица решений представляется как

$$T = (U, A, C, D),$$

где U – универсальное множество; A – множество всех признаков; C – подмножество признаков-условий; D – подмножество признаков-решений ($C, D \subset A$).

Пусть $a \in A$, $P \subseteq A$, тогда бинарное отношение IND(P) называется отношением неразличимости:

$$IND(P) = \{(x, y) \in U \times U : \forall a \in P, a(x) = a(y)\}.$$

Пусть $U/\mathrm{IND}(P)$ описывает множество всех классов эквивалентности отношения $\mathrm{IND}(P)$. Классы эквивалентности $U/\mathrm{IND}(C)$ и $U/\mathrm{IND}(C)$ называются классами условий и решений соответственно.

Пусть $R \subseteq C$ и $X \subseteq U$, тогда

$$\underline{R}X = \cup\{Y \in U/R : Y \subseteq X\},$$

$$\overline{R}X = \cup\{Y \in U/R : Y \cap X\},$$

где $\underline{R}X$ и $\overline{R}X$ – R-нижняя и R-верхняя аппроксимации X, соответственно, $(\underline{R}X, \overline{R}X)$ – R-приближённое множество. Если X является R-определяемым, то $\underline{R}X = \overline{R}X$, в противном случае – X является R-приближённым.

Граница $BN_R(X)$ определяется как $BN_R(X) = \overline{R}X - \underline{R}X$. Таким образом, если X является R-определяемым, то $BN_R(X) = \Phi$.

Пусть $c \in C$. Признак c является несущественным в T, если $\mathrm{POS}_{(C-(c))}(D) = \mathrm{POS}_C(D)$, в противном случае – признак является незаменимым в T. Признак c является независимым, если все $c \in C$ являются незаменимыми.

Множество признаков $R \subseteq C$ является сокращением C, если $T' = \{U, A, R, D\}$ является независимым и выполняется $\mathrm{POS}_r(D)$. Другими словами, сокращением является минимальное подмножество признаков, сохраняющее представленное выше условие.

$\mathrm{CORE}(C)$ определяет множество всех признаков, которые являются незаменимыми в C:

$$\mathrm{CORE}(C) = \cap\,\mathrm{REDUCT}(C),$$

где $\mathrm{REDUCT}(C)$ – множество всех сокращений C.

2. Деревья решений

Деревья решений представляют собой нисходящую систему, основанную на подходе “разделяй и властвуй”, основной целью которой является разделение дерева на взаимно непересекающиеся подмножества [5, 6]. Каждое подмножество представляет собой подзадачу классификации.

Дерево решений описывает процедуру принятия решения о принадлежности определённого экземпляра к тому или иному классу.

Дерево решений является древовидной структурой, состоящей из внутренних и внешних узлов, связанных рёбрами [7]. Внутренние узлы – модули, принимающие решение, рассчитывают значение функции решения, на основании чего определяют дочерний узел, который будет посещён далее. Внешние узлы (также называемые конечными узлами), напротив, не имеют дочерних узлов и описывают либо метку класса, либо значение, характеризующее входные данные. В общем случае, деревья решений используются следующим образом. Вначале передаются данные (обычно это вектор значений входных переменных) на корневой узел дерева решений. В зависимости от полученного значения функции решения, используемой во внутреннем узле, происходит переход к одному из дочерних узлов. Такие переходы продолжаются до тех пор, пока не будет посещён конечный узел, описывающий либо метку класса, либо значение, связанное со входным вектором значений признаков.

Для применения деревьев решений на практике в целях классификации или прогнозирования значений выходных параметров исследуемых объектов по наборам значений входных характеристик необходимо с помощью данных обучающей выборки сформировать дерево решений таким образом, чтобы оно наилучшим образом описывало исследуемый объект.

Пусть задана обучающая выборка

$$S = < X, Y >,$$

где $X = \{X_i\}$ – набор значений признаков, характеризующих рассматриваемый объект или процесс; $Y = \{y_p\}$ – массив значений выходного параметра в заданной выборке; $X_i = \{x_{ip}\}$ – i-й признак в выборке, $i = 1,2,...,L$; $x_{ip} \in [x_{\min i}; x_{\max i}]$ – значение i-го признака для p-го экземпляра выборки, $p = 1,2,...,m$; L – общее число признаков в исходном наборе; m – число экземпляров выборки.

Тогда задача построения дерева решений $T = \{t_k\}$ по заданной выборке S заключается в идентификации узлов $t_k = < c_k, l_k, r_k >$ так, чтобы значение ошибки прогнозирования или классификации E построенной модели было минимальным: $E \to \min$, где $t_k = < c_k, l_k, r_k >$ – k-й узел дерева T, представляющий собой структуру, в которой c_k – функция принятия решений на основе значений входных переменных (в случае, если узел является внутренним) или значение выходной переменной (для внешних узлов), l_k и r_k – ссылки на левого и правого потомков k-го узла соответственно, представляющих собой структуры, аналогичные t_k.

Построение деревьев решений связано с извлечением правил из обучающих выборок. Каждый путь от корня дерева к одному из его листьев может быть преобразован к логическому высказыванию – правилу типа «если А, то В», где его антецедент получается путем использования всех условий, представленных во внутренних узлах от корня к выходному листу, а правая часть правила получается из соответствующего листа дерева.

Поэтому постановку задачи синтеза дерева решений как логической модели исследуемого объекта, процесса или явления можно также представить в следующем виде. Пусть задана обучающая выборка данных, состоящая из m экземпляров, каждый из которых характеризуется L атрибутами. При этом каждый атрибут может относиться к определённому лингвистическому терму LT. Для каждого i-го экземпляра указаны вхождения к лингвистическим термам для каждого атрибута и указан лингвистический терм выходной переменной. Тогда необходимо построить такое дерево

решений, которое позволяет выполнять отнесение выходного параметра к лингвистическому терму с заданной точностью.

3. Синтез деревьев решений с использованием теории приближённых множеств

Синтез деревьев решений на основе использования теории приближённых множеств состоит из двух фаз: вычисление сокращённых наборов признаков и непосредственно идентификация дерева решений. Таким образом, достигается интеграция двух подходов, за счёт чего обеспечивается повышение эффективности разрабатываемого метода. Входная обучающая выборка может быть как дискретной, так и непрерывной. Однако, при построении дерева решений на основе непрерывных выборок, их необходимо предварительно дискретизировать [8].

Таким образом, предлагаемый метод состоит из этапов, описанных ниже.

Этап 1. Входная обработка данных. На данном этапе задаётся входная выборка данных T1. При необходимости выполняется дискретизация непрерывных признаков, в результате чего получается выборка данных T2.

Этап 2. Определение сокращённого набора признаков для T2. Данный этап предназначен для поиска сокращённого набора признаков, на основе которого будет синтезироваться дерево решений. Данный этап состоит из следующих шагов.

Шаг 1. Входная таблица данных T2 разбивается так, что последняя колонка считается колонкой решений.

Шаг 2. Строки входной таблицы данных сортируются таким образом, что экземпляр с наименьшим значением выходного признака оказывается вверху таблицы, т.е. экземпляры перераспределяются в порядке возрастания.

Шаг 3. Генерация булевой матрицы для заданной таблицы входных данных. Для этого выполняется проверка по каждому экземпляру. В случае, если значения первого и второго признаков для всех экземпляров одинаковы, то в булеву матрицу заносится значение «1», в противном случае – «0»:

$$a_{i,j} = \begin{cases} 1, x_l^j = x_l^i, \forall l = \overline{1,m}; \\ 0, \exists l : x_l^j \neq x_l^i, \forall l = \overline{1,m}, \end{cases}$$

где i, j – признаки i и j, соответственно; m – количество экземпляров входной таблицы данных; l – индекс текущего экземпляра выборки данных.

Шаг 3 повторяется для каждого признака выборки данных. Это продолжается до тех пор, пока не будет построена полная булева матрица для всех пар признаков.

Шаг 4. После выполнения шага 3 в полученной булевой матрице производится сложение значений каждого столбца:

$$\forall i = \overline{1,n} : A_i = \sum_{j=1, i \neq j}^{n} a_{i,j},$$

где n – количество столбцов входной таблицы данных.

Столбец с максимальной суммой считается незаменимым признаком:

$$x_{\inf} = x_{\max\limits_{i=\overline{1,n}} A_i},$$

где $\max\limits_{i=\overline{1,n}} A_i$ – индекс столбца с максимальной суммой.

Шаг 5. На основании полученной ранее булевой матрицы получается сокращённая булева матрица путём удаления из матрицы выбранного на шаге 4 признака:

$$A' = A / A_{\max\limits_{i=\overline{1,n}} A_i}.$$

После чего сокращённая булева матрица становится рабочей булевой матрицей:

$$A = A'.$$

Шаг 6. Шаги 4 и 5 продолжаются до тех пор, пока не выполняется одно из следующих условий.

1. Суммы столбцов сокращённой булевой матрицы равны нулю, что означает следующее: нет больше никакой информации о том, каково различие во влиянии признаков:

$$\forall i = \overline{1,n} : \sum_{j=1, i \neq j}^{n} a_{i,j} = 0.$$

2. Сокращённая булева матрица пуста:

$$A = \varnothing.$$

Шаг 7. Отобранные незаменимые признаки группируются и считаются полученным сокращённым набором признаков, который будет использоваться на следующем этапе.

Таким образом, в результате этапа 2 получается выборка данных T3.

Этап 3. Идентификация дерева решений. Синтез дерева решений выполняется на основе выборки данных T3, при этом в каждый момент времени берётся один признак и для разделения используются все узлы на одном уровне:

$$T3 \Rightarrow DT,$$

где DT – полученное дерево решений, дерево решений получается путём использования любого метода синтеза деревьев решений.

Этап 4. Генерация правил. Правила генерируются путём всевозможных переходов от корневого узла к листьям в полученном дереве решений:

$$DT \Rightarrow RB,$$

где RB – полученная база правил.

Как видно, предложенный метод включает в себя две основные фазы: вычисление сокращённых наборов признаков и идентификация дерева решений. В связи с этим сложность предложенного метода зависит от сложности каждой из этих фаз.

Таким образом, если обучающая выборка данных состоит из m экземпляров и n признаков, то задача вычисления сокращённого набора признаков

минимальной длины является NP-сложной задачей. Вычислительная сложность предобработки входной таблицы данных и её сортировки равна $O(m^2)$. При этом необходимо произвести C (m, 2) сравнений и, если каждый экземпляр характеризуется n признаками, то сложность сравнения равна $O(nm^2)$. Сложность идентификация дерева решений зависит от значений разделяющих признаков, таким образом, построенное дерево решений может быть n-арным деревом.

Предложенный метод синтеза деревьев решений на основе теории приближённых множеств был программно реализован в среде пакета Matlab 7.0.

При помощи разработанного программного обеспечения и встроенных средств пакета Matlab 7.0 проводились эксперименты. Для экспериментов использовались тестовые данные, которые были взяты из общедоступных репозиториев [9]. Экспериментальные исследования проводились на основании выборки, которая содержала информацию об эхокардиограммах пациентов с сердечными приступами. Выборка содержала информацию о 132 пациентах, каждый из которых характеризовался 12 признаками. Кроме того, для каждого пациента указывалось, жив он или умер. Предложенный метод сравнивался с методом построения деревьев решений ID3 [5, 6], а также с мультиагентным [10, 11] и эволюционным [11, 12] методами. На основании проведенных экспериментов были получены базы правил, характеризующиеся следующим качеством классификации пациентов: 95,2%, 89,4%, 92,3% и 91,1% для предложенного, ID3, мультиагентного и эволюционного методов соответственно.

Таким образом, можно отметить, что предложенный метод построения деревьев решений на основании теории приближенных множеств обеспечивает более точные результаты прогнозирования по сравнению с другими известными методами.

Выводы

В работе решена актуальная задача синтеза деревьев решений, которые могут использоваться для классификации, а также для построения базы правил с целью разработки экспертных систем.

Научная новизна работы заключается в том, что разработан новый метод построения деревьев решений с использованием теории приближённых множеств, который позволяет повысить эффективность и скорость синтеза деревьев решений за счёт использования сокращённого набора признаков, получаемого при помощи теории приближённых множеств.

Практическая ценность полученных результатов заключается в том, что на основе предложенного метода разработано программное обеспечение, позволяющее выполнять идентификацию деревьев решений и получение базы правил, на основании которых можно создавать экспертные системы с меньшей ошибкой классификации.

Список литературы: 1. *Pawlak Z.* Rough sets / Z. Pawlak // International Journal of Computer and Information Sciences. – 1982. – № 11. – P. 341-356. 2. *Pawlak Z.* Rough Sets-Theoretical Aspects and Reasoning about Data / Z. Pawlak. – Dordrecht : Kluwer Academic Publications, 1991 – 237 p. 3. *Ramadevi Y.* Knowledge Extraction Using Rough Sets / Y. Ramadevi, C.R. Rao // Classification, International conference on Bioinformatics and diabetes mellitus. – India, 2006. – P. 128-132. 4. *Олейник, Ал. А.* Модификация метода муравьиных колоний с использованием операций над чёткими множествами [Текст] / Ал. А. Олейник, С. А. Субботин // Автоматика-2008 : п'ятнадцята міжнародна науково-технічна конференція, 23–26 вересня 2008 р. : тези доповідей. – Одесса, 2008. – С. 396–398. 5. *Quinlan J. R.* Induction of decision trees / J. R. Quinlan // Machine Learning. – 1986. – № 1. – P. 81–106. 6. *Rokach L.* Data Mining with Decision Trees. Theory and Applications / L. Rokach, O. Maimon. – London : World Scientific Publishing Co, 2008. – 264 p. 7. *Classification* and regression trees / L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone. – California : Wadsworth & Brooks, 1984. – 368 p. 8. *Субботін, С. О.* Подання й обробка знань у системах штучного інтелекту та підтримки прийняття рішень : навч. посібник [Текст] / С. О. Субботін. – Запоріжжя: ЗНТУ, 2008. – 341 с. 9. *UCI* Machine Learning Repository [electronic resource] / Center for Machine Learning and Intelligent Systems. – Access mode : http://archive.ics.uci.edu/ml/datasets.html. 10. *Олейник, А. А.* Мультиагентные методы интеллектуальной оптимизации для моделирования сложных объектов и систем [Текст] / А. А. Олейник, С. А. Субботин // Нейроинформатика, её приложения и анализ данных : XVII Всероссийский семинар, 2–4 октября 2009 г. : материалы семинара. – Красноярск, 2009. – С. 79–82. 11. *Субботін, С. О.* Неітеративні, еволюційні та мультиагентні методи синтезу нечіткологічних і нейромережних моделей: монографія [Текст] / С. О. Субботін, А. О. Олійник, О. О. Олійник ; під заг. ред. С.О. Субботіна. – Запоріжжя : ЗНТУ, 2009. – 375 с. 12. *Gen M.* Genetic algorithms and engineering design / M. Gen, R. Cheng. – New Jersey : John Wiley & Sons, 1997. – 352 p.

Поступила в редколлегию 11.01.2012

УДК 004.93

Синтез дерев розв'язків на основі теорії наближених множин / Є. О. Гофман, О. О. Олійник, С. О. Субботін // Біоніка інтелекту: наук.-техн. журнал. – 2012. – № 1 (78). – С. 29-32.

Запропоновано новий метод ідентифікації дерев розв'язків з використанням теорії наближених множин. Створений метод дозволяє скоротити час роботи й підвищити ефективність синтезованих моделей складних об'єктів і систем. Проведено експерименти з вирішення тестових завдань.

Бібліогр.: 12 найм.

UDC 004.93

Synthesis of decision trees based on rough set theory / Ye. A. Gofman, O. O. Oliinyk, S. A. Subbotin // Bionics of Intelligense: Sci. Mag. – 2012. – № 1 (78). – P. 29-32.

A new method of identification of decision trees using the theory of rough sets is proposed. Created a method allows to reduce the time and improve the effectiveness of the synthesized models of complex objects and systems. Experiments on the solution of test problems are conducted.

Ref.: 12 items.