СИСТЕМНЫЙ АНАЛИЗ И УПРАВЛЕНИЕ

УДК 004.422.635.3

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ МЕТОДА ДЕРЕВЬЕВ РЕШЕНИЙ ДЛЯ РЕАЛИЗАЦИИ ЗАДАЧ КЛАССИФИКАЦИИ И ПРОГНОЗИРОВАНИЯ

Т. В. ЗАЙЦЕВА[1]
Н. В. ВАСИНА[2]
О. П. ПУСНАЯ[1]
Н. Н. СМОРОДИНА[1]

[1] Белгородский государственный национальный исследовательский университет

[2] Тульский государственный университет

e-mail:
zaitseva@bsu.edu.ru
natavasina71@yandex.ru
pusnaya@bsu.edu.ru
smorodina@bsu.edu.ru

Применение гибридных методов технологии Data Mining позволяет эффективно использовать их при решении задач, которые направлены на автоматический анализ и выявление закономерностей в большом объеме данных.

В статье рассматривается метод деревьев решений с учетом вероятностной неопределенности классификации. Дерево решений строится автоматически в зависимости от статистических данных.

Приведен пример принятия решения о выдаче кредита потребителю.

Ключевые слова: деревья решений, теорема Байеса, принятие решения, классификация, прогнозирование, правила-продукции.

Развитие компьютерных технологий послужило значительному увеличению объема хранимых данных. Что в свою очередь привело к тому, что человеку стало все труднее проанализировать их. Хотя необходимость проведения такого анализа вполне очевидна, ведь в этих «сырых данных» заключены знания, которые могут быть использованы при принятии решений. Поэтому стали развиваться методы, позволяющие проводить автоматический анализ данных.

Data Mining –процесс обнаружения в «сырых» данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Большинство аналитических методов, используемые в технологии Data Mining – это известные математические алгоритмы и методы. Новым в их применении является возможность их использования при решении тех или иных конкретных проблем, обусловленная появившимися возможностями технических и программных средств.

Задачи интеллектуального анализа данных можно разделить по типу извлекаемой информации: классификация; кластеризация; выявление ассоциаций; выявление последовательностей; прогнозирование. Наиболее часто в экономической практике встречаются задачи классификации и прогнозирования. Одним из старейших и наиболее популярных методов решения данных задач является метод деревьев решений (decision trees).

Преимущества деревьев решений.

1. Интуитивность деревьев решений.

2. Деревья решений дают возможность извлекать правила из базы данных на естественном языке.

3. Алгоритм конструирования дерева решений не требует от пользователя выбора входных атрибутов.

4. Высокая точность создаваемых моделей.

5. Быстрый процесс обучения.

6. Большинство алгоритмов конструирования деревьев решений имеют возможность специальной обработки пропущенных значений.

Рассмотрим задачу определения кредитонадежности заемщика. База данных, на основе которой должно осуществляться прогнозирование, содержит следующие ретроспективные данные о клиентах банка, являющиеся ее атрибутами: возраст, наличие недвижимости, образование, социальное положение, среднемесячный доход, вернул ли клиент вовремя кредит и т.д. В принципе условия выдачи кредита в разных банках являются различными, однако, все вышеперечисленные атрибуты присутствуют в явном или неявном виде. Задача состоит в том, чтобы на основании перечисленных выше данных (кроме последнего атрибута) определить, стоит ли выдавать кредит новому клиенту. Такая задача решается в два этапа: построение классификационной модели и ее использование. На этапе построения модели строится дерево классификации или создается набор неких правил. На этапе использования модели построенное дерево, или путь от его корня к одной из вершин, являющийся набором правил для конкретного клиента, используется для ответа на поставленный вопрос «Выдавать ли кредит?». Правилом является логическая конструкция, представленная в виде «если : то :».

Качество построенного дерева решения весьма зависит от правильного выбора критерия расщепления. Традиционно дерево решений строится, начиная с первого атрибута (то есть в данном примере с возраста), не учитывая характер и силу влияния каждого атрибута. Более эффективным является подход, основанный на учете вероятностной неопределенности классификации. Другими словами, событие, состоящее в установлении соответствия между значениями цепочки атрибутов и определенным классом, является случайным и характеризуется некоторой вероятностью. При использовании нескольких атрибутов в качестве первого атрибута для анализа выбирается тот, который обеспечивает максимальное снижение неопределенности классификации по отношению к исходному множеству (т.е. минимальное значение энтропии).

Согласно предложенной методике построения дерево решений начинает строиться с атрибута, который больше всего уменьшает неопределенность (в рассмотренном примере это факт возврата кредита). Далее по формуле Байеса находятся апостериорные условные вероятности, которые будут использованы для построения правил-продукций.

Далее отдельно рассматриваются те данные, записи которых соответствуют положительному значению рассмотренного атрибута, и данные, записи которых соответствуют отрицательному значению. Аналогично выбирается следующий из критериев расщепления дерева решений и т.д. По полученному дереву решений строится система продукционных правил:

ЕСЛИ (С10 = Да) И (С5 ≤ 5) И (С8 > 200) И (С3 = КД), ТО клиент К1 с вероятностью 100%.

ЕСЛИ (С10 = Да) И (С5 ≤ 5) И (С8 > 200) И (С3 = НКД), ТО клиент К1 с вероятностью 86%; клиент К2 с вероятностью 14%. И т.д.

Совокупность полученных правил-продукций после небольшой доработки преобразуется в законченную базу знаний и может быть использована в продукционных или гибридных экспертных системах.

Программная поддержка на примере определения кредитонадежности заемщика реализована в идее программных модулей: получения дерева решений по статистическим данным; создания продукционных правил для экспертной системы.

Рассмотрим пользовательский интерфейс прототипа системы «Дерево решений». При запуске программной системы «Дерево решений» на экране монитора появляется диалоговая форма, представленная на рисунке 1.

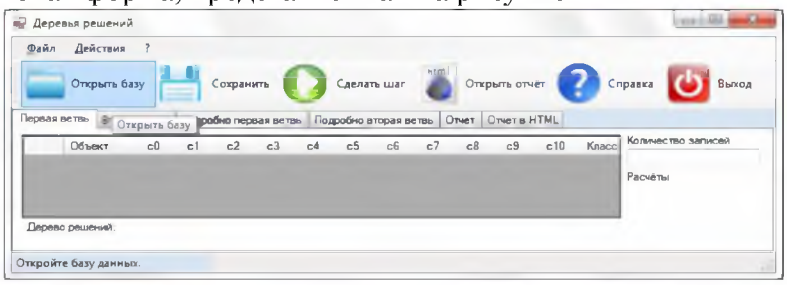

Рис. 1. Интерфейс программы

После загрузки базы данных, программная система принимает вид, отображенный на рисунке 2.

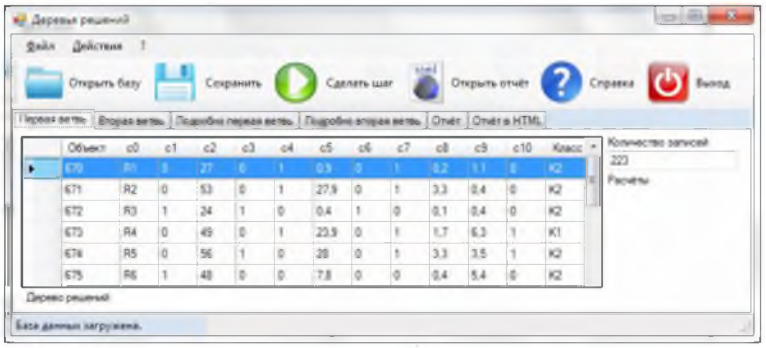

Рис. 2. Открытая база данных

В поле «Количество записей» выдается общее число записей в базе данных. Есть возможность упорядочить данные по одному из полей (атрибутов). Далее выбирая пункт меню или используя кнопку на панели инструментов «Сделать шаг», вычисляем энтропии для каждого атрибута, находим минимальное значение энтропии и ему соответствующий номер атрибута (рисунок 3). Построение дерева решений начнем именного с этого атрибута.

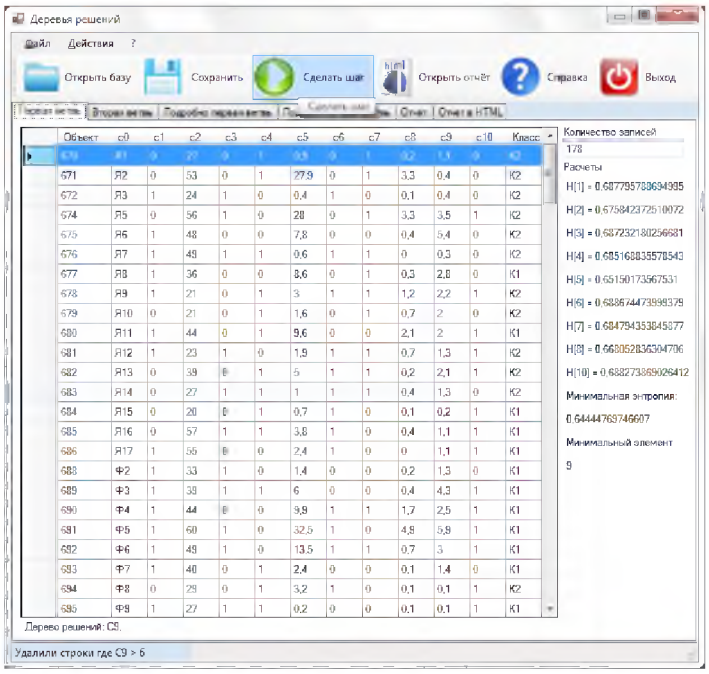

Рис. 3. Первый шаг построения дерева решений

Далее по формуле Байеса находим апостериорные условные вероятности при условии, что атрибут C9 принял одно из двух значений в данной ветке C9<6. Выбираем минимальную энтропию и находим номер атрибута, которому соответствует минимальная энтропия. Вторая ветвь (C9>=6) приводит к готовому продукционному правилу Построение дерева решений происходит с расчётом всех возможных комбинаций. Во вкладке «Вторая ветвь» выводятся удалённые строки из базы данных, и по ним производится параллельный расчёт. Аналогично первой ветке во второй происходит расчёт энтропии и нахождение минимального элемента. Количественные характеристики, а также энтропии по атрибутам можно просмотреть во вкладке «Первая ветвь подробно» (рис. 4). Подробные расчёты второй ветки дерева можно посмотреть, открыв вкладку «Вторая ветвь подробно».

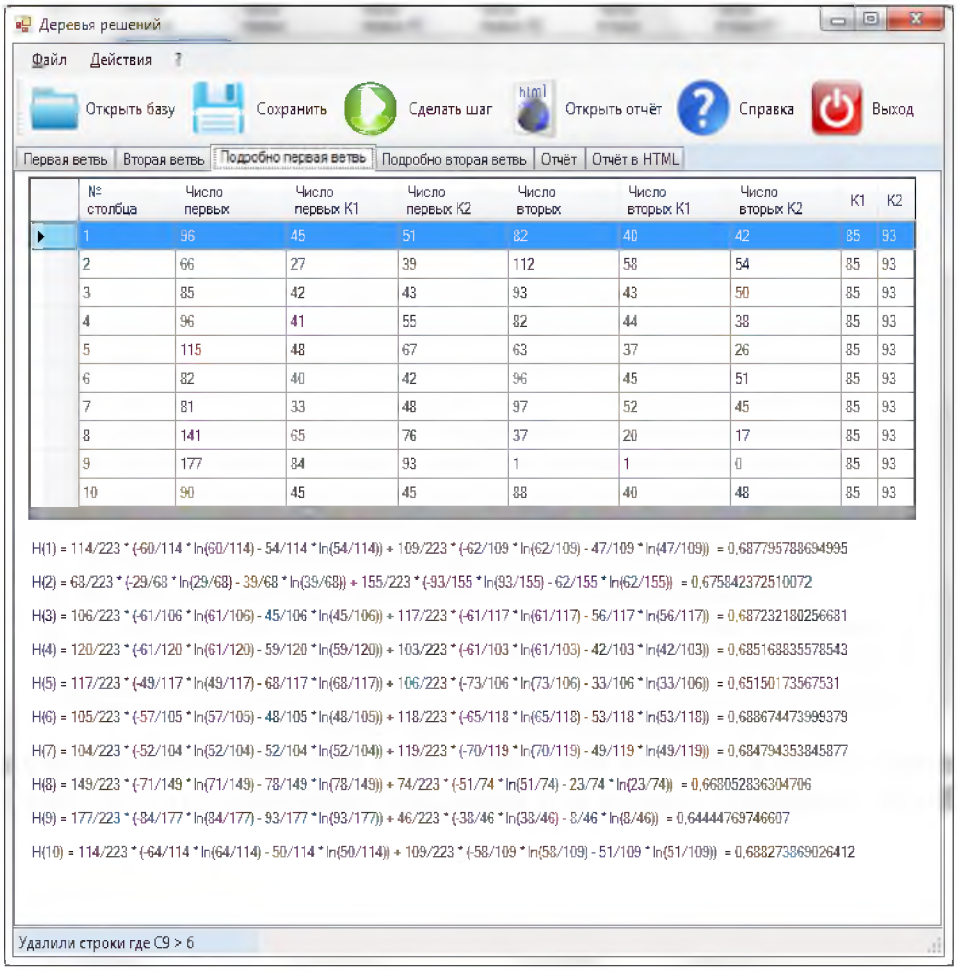

Рис. 4. Подробные расчёты энтропии первой ветки дерева

Ветвь с C9<6 разбиваем на подмножества по следующему выбранному атрибуту. Критерием выбора атрибута, по которому должно пойти разбиение соответствующего подмножества, является минимальная энтропия. Далее шаги повторяются до тех пор, пока не получим вершину, для которой апостериорная вероятность принадлежности объекта к определенному классу равна единице. На последнем шаге можно увидеть атрибуты, влияющие на построение дерева решений, их порядок и значения энтропии. После выполнения последнего шага можно посмотреть готовое дерево решений либо в приложении на вкладке «Отчёт», либо во внешнем браузере, выбрав пункт меню «Открыть отчёт». Результаты построения дерева показаны на рис. 5. Для проверки работоспособности и эффективности разработанного алгоритма было проведено сравнение результатов классификации с использованием разных алгоритмов.

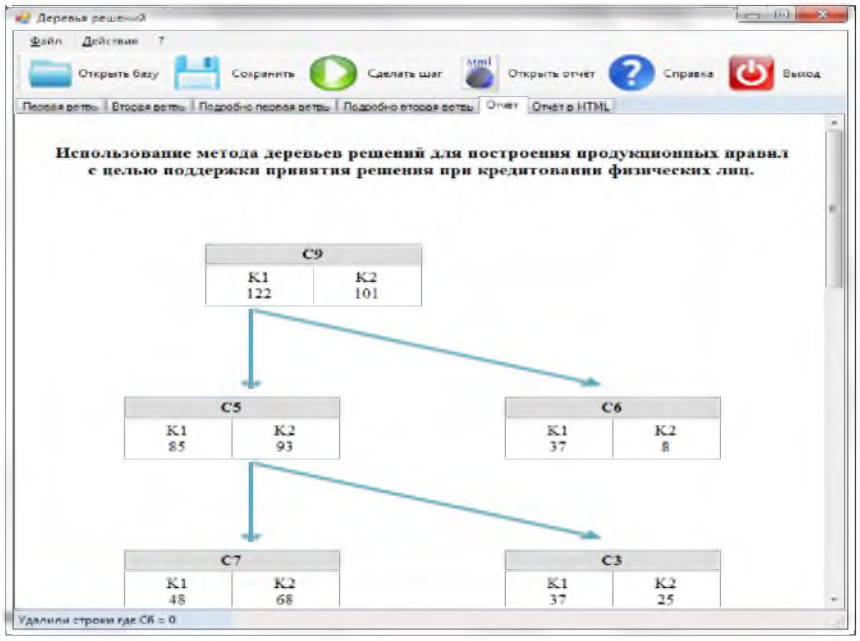

Рис. 5. Дерево решений

После построения дерева решений можно формировать продукционные правила. Выбирая пункт меню «Сформировать правила» (рис. 6), формируем продукционные правила вида IF () AND () AND () ... AND () THEN (). Сформированные правила автоматически сохраняются в текстовом формате (рис. 7).

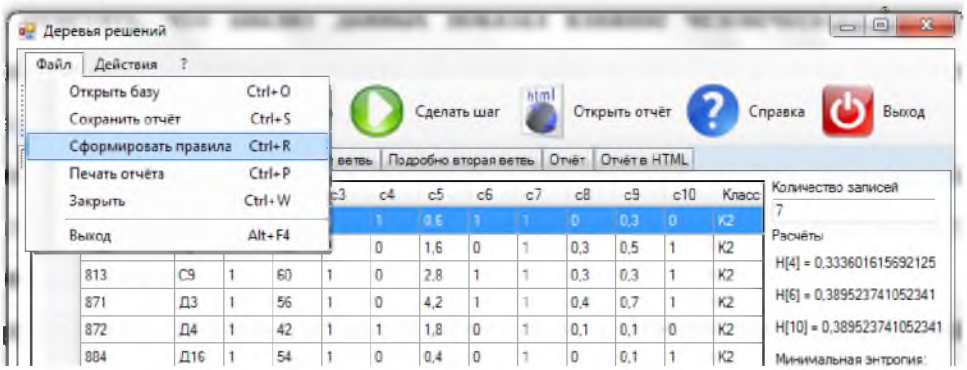

Рис. 6. Формирование продукционных правил

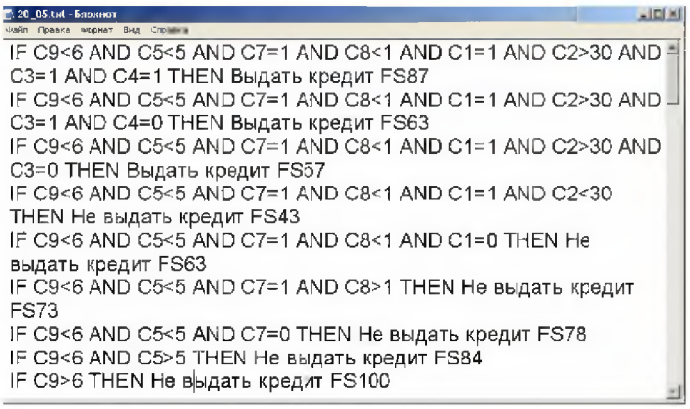

Рис. 7. Текстовый файл со сформированными правилами

Сравнение результатов, полученных различными методами интеллектуального анализа данных, и с применением предлагаемого алгоритма проводилось на примере базы данных одного из региональных банков за 2006-2010 года (бралась выборка в размере 140 записей за каждый год). Все записи в базе данных (отдельно по годам и в целом за 5 лет) были разбиты на две группы:

– обучающая;

– тестовая.

Следует отметить, что анализ данных показал влияние человеческого фактора при принятии решения о выдаче кредита. В базе данных из 700 записей было отмечено 96, когда было принято неправильное решение:

– 83 записи о выданных кредитах, которые в последствии не были возвращены;

– 13 записей о не выданных кредитах заемщикам, которые удовлетворяют всем критериям для выдачи.

То есть работники в среднем ошибаются в 14 случаях из 100 при принятии решений.

Сравнивались результаты, полученные следующими методами:

1) предлагаемая модель;

2) нейронная сеть на примере программы Neural Network Wizard;

3) линейная регрессия (Microsoft Linear Regression) пакета Microsoft SQL Server 2005;

4) решающие деревья (Sipina с алгоритмом C4.5).

Полученные результаты по процентам ошибок по годам, в целом за 5 лет и в среднем за 5 лет приведены в таблице.

Предлагаемая модель в среднем выдает 7-8 неправильных рекомендаций из 140 за год по вопросу выдачи кредитов. При этом, как показал анализ, 4-6 рекомендаций относятся к упущенной выгоде банка (программа рекомендует не выдавать кредит благонадежному заемщику) и только 2-3 рекомендации можно отнести к потери банком собственных денежных средств (программа рекомендует выдавать кредит, при этом он не был возвращен). При рассмотрении дерева решений с количеством ветвей больше двух процент ошибок возрастает незначительно, максимальная ошибка составила 7,2%.

Таблица

Процент ошибок для различных методов

Методы	По годам, %					За 5 лет	Среднее за 5 лет
	2006	2007	2008	2009	2010		
Предлагаемая модель	5	4,6	5,5	4,9	5,8	5,2	5,2
Нейронные сети	14	13,1	13,3	12,9	14,9	13,7	13,64
Нейронные сети (с предобработкой)	4,7	4,3	4,8	5,1	5,7	5	4,92
Линейная регрессия	18,5	17	23	19,8	22,5	35,2	20,16
Решающие деревья	13,4	16	13,8	15,6	14,2	15,6	14,6
Решения, принятые работниками банка	14,3	12,9	13,6	12,9	15	13,7	13,74

При этом никакой дополнительной предобработки данных не производилось, все результаты были получены полностью в автоматическом режиме. Результаты, полученные с помощью нейронной сети, оказались очень близкими к решениям, которые принимали работники банка (проценты ошибок сопоставимы).

После проведения предобработки данных (были удалены все записи с неправильно принятыми решениями сотрудниками банка) нейронные сети показали очень хороший результат. Процент ошибок в этом случае был сопоставим с процентом ошибок по предлагаемой модели. Однако, для получения такого результата необходимо проводить предварительную обработку базы данных с удалением части записей, что потребует дополнительных временных затрат или создание дополнительного программного модуля отбора данных.

Результаты, полученные с помощью пакета Microsoft SQL Server 2005 (линейная регрессия), являются неоднозначными. Если рассматривать отдельно каждый год, то про-

цент ошибки не превышает 23%, а при рассмотрении данных за 5 лет процент ошибки вырастает до 35%. Это связано с особенностями данного метода – автоматическим выбором наиболее значимых результатов. Если при рассмотрении по годам значимыми критериями являлись 7-8 (при этом в различные годы значимыми оказывались различные критерии), то при рассмотрении данных за 5 лет значимыми критериями остались только 5 из 10 рассматриваемых. То есть данный метод можно использовать в течение небольшого временного периода (1-2 года) для предварительной оценки. Кроме того, полученные результаты, представленные в графическом виде, являются сложными для восприятия и понимания без хороших знаний статистических пакетов. Результаты, полученные с помощью алгоритма C4.5, показали среднюю величину ошибки в 15% .

Это достаточно хороший результат в случае предварительного анализа данных. Однако, при построении деревьев с количеством ветвей больше двух, процент ошибок возрастает значительно.

Преимуществами разработанного алгоритма являются:

1) быстрый процесс обучения;

2) генерация правил в областях, где эксперту трудно формализовать свои знания;

3) извлечение правил на естественном языке;

4) понятная на интуитивном уровне классификационная модель;

5) высокая точность прогноза, сопоставимая с другими методами (статистика, нейронные сети)

В заключение следует отметить, что использование вариационных алгоритмов в задачах классификации является весьма актуальным в связи с постоянным ростом вычислительной мощности компьютеров. Такого рода алгоритмы позволяют добиться хороших (адекватных) результатов. Но в связи с большой долей эвристики исследование их свойств сильно затрудняется. Таким образом, имеет смысл продолжать исследования в данном направлении и создавать новые алгоритмы, использующие вариационный принцип, которые будут более универсальными и адекватными.

Литература

1. Зайцева Т.В., Игрунова С.В., Путивцева Н.П., Пусная О.П., Манзуланич М.Ю. Компьютерная технология генерации правил для гибридных продукционно–фреймовых экспертных систем // Вопросы радиоэлектроники. Серия Электронная вычислительная техника. 2011. Вып. 1. С. 105–115.

2. Зайцева Т.В., Нестерова Е.В, Игрунова С.В., Пусная О.П., Путивцева Н.П., Смородина Н.Н. Байесовская стратегия оценки достоверности выводов // Научные ведомости БелГУ Серия История. Политология. Экономика. Информатика. Белгород: Изд-во БелГУ. 2012. №13(132). Выпуск 23/1. – С. 180-183.

3. Зайцева Т.В., Устинов Р.М., Пусная О.П. Компьютерная реализация алгоритма обработки статистических данных с учетом вероятностной неопределенности классификации // Вопросы радиоэлектроники. Серия Электронная вычислительная техника. 2012. Вып. 12. – С. 119-130.

SOFTWARE IMPLEMENTATION METHOD OF DECISION TREE FOR A PARTICULAR PURPOSE OF CLASSIFICATION AND PREDICTION

T. V. ZAITSEVA[1]
N. V. VASINA[2]
O. P. PUSNAYA[1]
N. N. SMORODINA[1]

[1])Belgorod National Research University
[2])Tula State University

e-mail:
zaitseva@bsu.edu.ru
natavasina71@yandex.ru
pusnaya@bsu.edu.ru
smorodina@bsu.edu.ru

The use of hybrid methods of Data Mining technology can effectively use them to solve problems that are aimed at carrying out the automatic analysis and identification of patterns in large data.

The article discusses a method of decision trees based probabilistic uncertainty classification. A decision tree is built automatically based on statistical data.

The article is an example of the decision to grant credit to consumers.

Keywords: decision trees, Bayes' theorem, decision making, classification, forecasting, rules-products.