2. Puzhaylo A. F., Spiridovich E. A., Lisin V. N. et al. Method of detection of the gas pipeline sections of the pipelines predisposed to stress corrosion cracking (stress corrosion); Pat. 2147098 Ros Federatsiya. № 99111247/6; zayavl. 03.06.1999; opuvbl. 27.03.2000, Byul. № 9.

3. VRD 39-1.11-020-99. Inspection technique for pipeline sections that are predisposed to stress corrosion cracking / OAO «Giprogaztsentr». Nizhniy Novgorod, 1999. 21 s.

4. Diagnostics and monitoring of the gas pipelines technical state while ensuring reliability, environmental security and gas transport controllability. Scientific publication / Pod red. V. E. Kostyukova. Nizhniy Novgorod: Nizhegorodsky un-ta, 2007. 204 s.

5. Puzhaylo A. F., Spiridovich E. A., Kostyukov V. E. et al. Software package for automated expert-analytical evaluation and analysis and forecasting of the gas pipelines technical condition (AES GP): programma dlya EVM. Sv. GR. № 2001610952. Opubl. 01.08.2001, Byul. №3.

УДК 519.876.2:378

Применение методов автоматического анализа при формировании контрольных цифр приема в учреждения профессионального образования

Ю.А. Шичкина[a], Ю.В. Планкова[b]

Братский государственный университет, ул. Макаренко 40, Братск, Россия
[a]cpk@brstu.ru, [b]strange.y@mail.ru
Статья поступила 02.02.2013, принята 15.05.2013

На формирование контрольных цифр приема граждан в образовательные учреждения профессионального образования оказывает влияние ряд факторов, таких, как демографический фактор, показатели приемных кампаний прошлых лет, количество выпускников, записавшихся на сдачу ЕГЭ по профильному предмету и др. Оптимизация контрольных цифр приема в учреждения профессионального образования обучающихся за счет средств федерального бюджета позволяет обеспечить качественный прием абитуриентов. В статье рассмотрено применение одного из методов автоматического исследования (анализа), а именно построение дерева решений при формировании контрольных цифр приема вуза. Дерево принятия решений (их можно также назвать деревьями классификации или регрессионными деревьями) используется в области статистики и анализа данных для прогнозных моделей. Структура дерева представляет собой «листья» и «ветки». На ребрах («ветках») дерева решения записаны атрибуты, от которых зависит целевая функция, в «листьях» – значения целевой функции, а в остальных узлах – атрибуты, по которым различаются случаи. Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение. Подобные деревья решений широко используются в интеллектуальном анализе данных. Цель состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной на основе нескольких переменных на входе. Приемная кампания – это процесс, который сопровождается достаточно широким множеством параметров, на него влияющих. Деревья решений – это способ представления правил в иерархической, последовательной структуре, где каждому объекту соответствует единственный узел, дающий решение.

Ключевые слова: дерево решений, алгоритм, приемная кампания, бюджетные места.

Applying automatic analysis techniques while forming admission quotas to enter institutions of higher education

Yu.A. Shichkina[a], Yu.V. Plankova [b]

Bratsk State University, 40 Makarenko st., Bratsk, Russia
[a]cpk@brstu.ru, [b]strange.y@mail.ru
Received 02.02.2013, accepted 15.05.2013

A number of factors exert influence on forming admission quotas for citizens to enter the institutions of higher professional education such as a demographic factor, admission figures of the previous years, the number of graduates wishing to pass the USE on the profession-oriented subject and others. The admission quotas optimization to enter the institutions of professional education studying from federal budget resources allows ensuring high quality of graduates' admission. This article considers applying one of the automated research (analysis) techniques, specifically, creating (constructing) the decision tree while forming admission quotas for citizens to enter the institutions of higher education. The decision tree (it can also be called the classification tree or regression tree) is used in the field of statistics and data analysis for forecast models. The tree structure is represented by "leaves" and "branches". The attributes determining the objective function are on the tree edges ("branches"), the values of the objective function are on the "leaves", and all the other nodes contain the attributes to distinguish among the cases. To classify a new case, it is necessary to go down the tree to the leaf and give a corresponding decision. These decision trees are wildly used in the intellectual data analysis. The aim is to create a

model which forecasts the goal variable on the basis of some input variables. The admission campaign is a process accompanied by a rather wide range of parameters influencing it. Decision trees demonstrate a method of representing rules in the hierarchical consecutive order (structure) where every object has its proper point giving its own decision.

Keywords: decision tree, algorithm, admission campaign, state-funded places.

Приемная кампания вуза – это процесс, который обусловлен достаточно широким множеством параметров, на него влияющих. На сегодняшний день не существует четкого формализованного подхода к отбору этих параметров и расчету контрольных цифр приема (КЦП), оптимизированных как минимум по двум показателям: своевременного закрытия КЦП с заданным минимальным средним балом ЕГЭ и максимального привлечения студентов на коммерческой основе. В связи с этим становится актуальной задача построения математической модели формирования КЦП на основе ряда параметров, наиболее на них влияющих, с учетом связей между этими параметрами.

Данную задачу по организации оптимальной с экономической точки зрения приемной кампании можно разбить на несколько подзадач:

– определение совокупности параметров, оказывающих наибольшее влияние на результат формирования КЦП;

– вывод совокупности правил, позволяющих принимать решения в рамках приемной кампании, в зависимости от сложившихся обстоятельств и конкретных значений параметров, влияющих на результат формирования КЦП;

– организация компактной формы хранения данных, собираемых в ходе мониторинга приемной кампании.

Все эти задачи можно решить с помощью деревьев решений.

На сегодняшний день существует значительное число алгоритмов, реализующих деревья решений – CART, C4.5, NewId, ITrule, CHAID, CN2 и т. д. Наибольшее распространение и популярность получили следующие два:

– CART – это алгоритм построения бинарного дерева решений – дихотомической классификационной модели. Каждый узел дерева при разбиении имеет только двух потомков. Из названия алгоритма видно, что он решает задачи классификации и регрессии.

– C4.5 – алгоритм построения дерева решений. Количество потомков у узла в данном алгоритме не ограничено. Алгоритм не может работать с непрерывным целевым полем, поэтому решает только задачи классификации.

Построим дерево решений по формированию контрольных цифр приема в Братском государственном университете на основе CART-алгоритма. CART (сокращение от Classification And Regression Tree, переводится как «дерево классификации и регрессии) – алгоритм бинарного дерева решений, впервые опубликованный Бриманом и др. в 1984 году [1]. Алгоритм предназначен для решения задач классификации и регрессии.

Для построения дерева решений будем применять данные приемных кампаний Братского государственного университета, полученные за 5-10-летний период, условные обозначения которых представлены в таблице 1 [2]:

Таблица 1

Условные обозначения показателей приемной кампании

Показатели	**Условное обозначение**
Среднее количество бюджетных мест на одно направление	n_c
Количество направлений	К
Количество выпускников Братска	Вп
Количество выпускников, записавшихся на ЕГЭ по профильному предмету	ЕГЭ
Количество договорников	Дв
Количество заявлений	Зв
Конкурс по предыдущему году	Кнс
Проходной балл	Бл

С учетом обозначений, приведенных в таблице 1, общее число бюджетных мест можно рассчитать по формуле:

$$N_c = K \cdot n_c . \quad (1)$$

Так как эмпирически установлено, что ежегодно от 15 до 40 % из общего числа выпускников средних образовательных учреждений уезжает из Братска и от 10 до 30 % – приезжает из Братского района, то количество потенциальных абитуриентов Братского государственного университета можно рассчитать по формуле:

$$\text{Аб} = Bn \cdot 0{,}6 + Bn \cdot 0{,}3 = 0{,}9Bn . \quad (2)$$

Предполагается, что конкурс на направление не должен быть ниже единицы, поэтому:

$$Кнт = \frac{\text{Аб}}{Nc} \geq 1 . \quad (3)$$

Еще одним немаловажным показателем является количество выпускников средних образовательных учреждений, записавшихся на ЕГЭ по профильному предмету. С учетом, того, что часть абитуриентов уез-

жает из Братска, конкурс среди потенциальных абитуриентов с профильным экзаменом ЕГЭ можно рассчитать по формуле:

$$Ke = \frac{\text{ЕГЭ}}{n_{\text{фс}}} \geq 1, \qquad (4),$$

где $n_{\text{фс}}$ – количество бюджетных мест с определенным профильным экзаменом.

С учетом планируемого среднего числа бюджетных мест на направление введем обозначение рассчитываемых с помощью дерева решений интервалов бюджетных мест на направление (таблица 2).

Таблица 2

Условные обозначения интервалов

Наименование интервала, *I*	**Диапазон значений**
A	$0{,}8n_c$
B	n_c
C	$0{,}8n_c$+Пл
D	n_c+Пл

Процесс создания дерева происходит сверху вниз, т. е. является нисходящим. В ходе процесса алгоритм должен найти такой критерий расщепления, иногда также называемый критерием разбиения, чтобы разбить множество на подмножества, которые бы ассоциировались с данным узлом проверки. Каждый узел проверки должен быть помечен определенным атрибутом. Существует правило выбора атрибута: он должен разбивать исходное множество данных таким образом, чтобы объекты подмножеств, получаемых в результате этого разбиения, являлись представителями одного класса или же были максимально приближены к такому разбиению. Последняя фраза означает, что количество объектов из других классов, так называемых «примесей», в каждом классе должно стремиться к минимуму.

Существуют различные критерии расщепления. Наиболее известные – мера энтропии и индекс *Gini*.

В алгоритме CART применяется индекс *Gini*, предложенный Брейманом (Breiman) и др. При помощи этого индекса атрибут выбирается на основании расстояний между распределениями классов.

Если дано множество *T*, включающее примеры из n классов, то индекс *Gini* определяется по формуле:

$$Gini\ (T) = 1 - \sum_{j=1}^{n} Pj^{\,2}, \qquad (5)$$

где *T* – текущий узел; *pj* – вероятность класса *j* в узле *T;* *n* – количество классов.

Если набор T разбивается на две части, T1 и T2, с числом примеров в каждом N1 и N2 соответственно, тогда показатель качества разбиения будет равен:

$$Gini_{split}(T) = \frac{N_1}{N} \cdot Gini(T_1) + \frac{N_2}{N} \cdot Gini(T_2). \qquad (6)$$

Наилучшим считается то разбиение, для которого $Gini_{split}(T)$ минимально.

Пример: Пусть имеются данные за десятилетний период для специальности СДМ (таблица 3).

Таблица 3

Исходные данные для специальности СДМ

Год	$\mathbf{n_c}$	**К**	**Вп**	**ЕГЭ**	**Зв**	**Кнс**	**Бл**	**Аб**	**Nc**	**Кнт**	$\mathbf{n_{фс}}$	**Ке**	**Т**
2002	30	29	2734	1715	121	2,5	50	2461	870	2,83	720	2,38	Y
2003	30	31	2680	1854	138	2,3	53	2412	930	2,59	673	2,75	Y
2004	30	32	2344	1541	143	2	47	2110	960	2,2	695	2,22	Y
2205	30	34	2615	1573	69	1,3	36	2354	1020	2,3	655	2,4	N
2006	30	33	2313	1278	94	1,4	40	2082	990	2,1	680	1,88	Y
2007	40	32	2220	1232	83	1	30	1998	1280	1,6	610	2,02	N
2008	36	31	1936	960	99	1,3	33	1742	1116	1,6	485	1,98	N
2009	25	31	1594	483	114	1	34	1435	775	1,9	330	1,46	N
2010	25	30	1214	568	154	1	41	1093	750	1,5	310	1,83	N
2011	20	31	908	419	128	1	46	817	620	1,3	265	1,58	N

ПРИМЕЧАНИЕ. Чем больше объем данных, тем точнее будут выведены правила для принятия решений.

Индекс *Gini* всей выборки по целевой функции Т (выполнение/невыполнение КЦП) составляет:

$$Gini(T) = 1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2 = 0{,}48.$$

В следующей таблице приведены значения, по которым исходная таблица, в зависимости от выбранного показателя, будет разбиваться на две части.

Таблица 4

Пороги делимости каждого критерия

n_c	К	Вп	ЕГЭ	Зв	Кнс	Бл	Аб	Nc	Кнт	$n_{фс}$	Ке
29,6	31,4	2055,8	1162	114	1,48	41	1850	931	2,28	542,3	2,05

Индексы *Gini* для каждого параметра в соответствии с формулой 6 приведены в таблице 5.

Таблица 5

Индекс Gini

n_c	К	Вп	ЕГЭ	Зв	Кнс	Бл	Аб	Nc	Кнт	$n_{фс}$	Ке
0,34	0,47	0,27	0,27	0,4	0,18	0,32	0,27	0,47	0,42	0,27	0,32

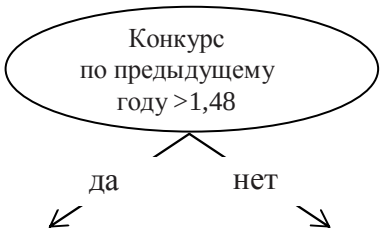

Рис. 1. Построение дерева решений

Выбираем параметр «Конкурс по предыдущему году». Следовательно, данный параметр имеет минимальное значение индекса *Gini* и является наиболее весомым при принятии решения о закрытии контрольных цифр приема. Этот параметр помещаем в вершину дерева решений для специальности СДМ (рис. 1).

Само множество решений будет при этом разбито на два новых подмножества *T*1 (таблица 6) и *T*2 (таблица 7).

Таблица 6

Подмножество T_1

Год	n_c	К	Вп	ЕГЭ	Зв	Кнс	Бл	Аб	Nc	Кнт	$n_{фс}$	Ке	Т
2002	30	29	2734	1715	121	2,5	50	2461	870	2,83	720	2,38	Y
2003	30	31	2680	1854	138	2,3	53	2412	930	2,59	673	2,75	Y
2004	30	32	2344	1541	143	2	47	2110	960	2,2	695	2,22	Y

Таблица 7

Подмножество T_2

Год	n_c	К	Вп	ЕГЭ	Зв	Кнс	Бл	Аб	Nc	Кнт	$n_{фс}$	Ке	Т
2005	30	34	2615	1573	69	1,3	36	2354	1020	2,3	655	2,4	N
2006	30	33	2313	1278	94	1,4	40	2082	990	2,1	680	1,88	Y
2007	40	32	2220	1232	83	1	30	1998	1280	1,6	610	2,02	N
2008	36	31	1936	960	99	1,3	33	1742	1116	1,6	485	1,98	N
2009	25	31	1594	483	114	1	34	1435	775	1,9	330	1,46	N
2010	25	30	1214	568	154	1	41	1093	750	1,5	310	1,83	N
2011	20	31	908	419	128	1	46	817	620	1,3	265	1,58	N

На следующем шаге с минимальным значением индекса Gini будет целый ряд параметров. Из них параметры К (количество направлений подготовки) и $n_{фс}$ (количество бюджетных мест с определенным профилем) – это те параметры, которые также, как и количество бюджетных мест, указываются при формировании заявки и не являются в данном случае определяющими. Параметр Аб (количество потенциальных абитуриентов) рассчитывается по формуле (2) и зависит от параметра Вп (количество выпускников). Поэтому параметр Аб не может быть выбран в качестве основного критерия для продолжения дерева. Для дальнейшего построения можно взять один из атрибутов Вп (количество выпускников) или ЕГЭ (количество выпускников, записавшихся на ЕГЭ по профильному предмету). Очевидно, что параметр ЕГЭ является более значимым, и в связи с этим для дальнейшего построения дерева выберем его. Продолжая алгоритм, отберем еще один параметр, влияющий на формирование КЦП по специальности СДМ, – это Ке.

Окончательный вариант дерева решений представлен на рис.2.

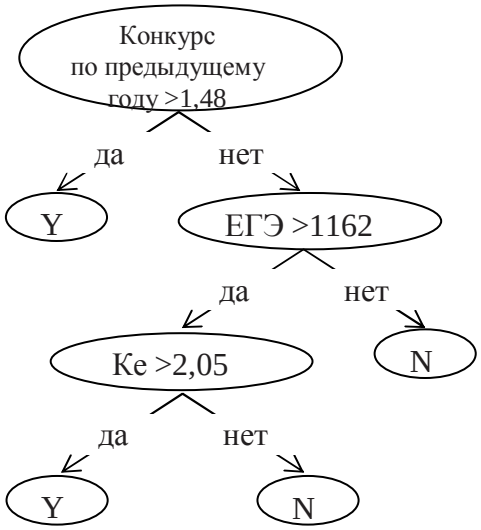

Рис. 2. Дерево решений

На основании построенного дерева для специальности СДМ можно сформулировать следующие правила.

Правило 1. Если конкурс на специальность СДМ по предыдущему году составлял более 1,48 чел., то число бюджетных мест на данную специальность можно оставить прежним (по прошлому году) или увеличить при наличии в прошлом году абитуриентов, поступавших на платной основе.

Правило 2. Если конкурс на специальность СДМ по предыдущему году составлял менее 1,48 чел. и число выпускников, записавшихся на ЕГЭ по профильному предмету, меньше чем 1162 чел., то число бюджетных мест на данную специальность необходимо сократить.

Правило 3. Если конкурс на специальность СДМ по предыдущему году составлял менее 1,48 чел., число выпускников, записавшихся на ЕГЭ по профильному предмету, превысило 1162 чел. и конкурс среди потенциальных абитуриентов с профильным экзаменом ЕГЭ выше 2,05 чел., то число бюджетных мест на данную специальность можно оставить прежним (по прошлому году) или увеличить при наличии в прошлом году абитуриентов, поступавших на платной основе.

Правило 4. Если конкурс на специальность СДМ по предыдущему году составлял менее 1,48 чел., число выпускников, записавшихся на ЕГЭ по профильному предмету, превысило 1162 чел, но конкурс среди потенциальных абитуриентов с профильным экзаменом ЕГЭ ниже 2,05 чел., то число бюджетных мест на данную специальность необходимо сократить.

Полученные правила можно представить в виде таблицы 8.

Таблица 8

Правила формирования КЦП на специальность СДМ

Конкурс по предыдущему году > 1,48	**ЕГЭ > 1162**	**Ке > 2,05**	**Наличие договорников**	**Интервал**	**Диапазон значений**
+	+/–	+/–	+	D	n+Пл
+	+/–	+/–	–	B	n
–	+	–	–	A	0,8n
–	+	–	+	C	0,8n+Пл
–	+	+	–	B	n
–	+	+	+	D	n+Пл
–	–	+/–	–	A	0,8n
–	–	+/–	+	C	0,8n+Пл

Пример. В 2010 году на специальность СДМ было выделено 25 бюджетных мест, n = 25. Конкурс в 2010 году на данную специальность составлял 1 чел./место. Следовательно, из таблицы 8 необходимо убрать первые две строки:

Конкурс по предыдущему году > 1,48	**ЕГЭ > 1162**	**Ке > 2,05**	**Наличие договорников**	**Интервал**	**Диапазон значений**
–	+	–	–	A	0,8n
–	+	–	+	C	0,8n+Пл
–	+	+	–	B	n
–	+	+	+	D	n+Пл
–	–	+/–	–	A	0,8n
–	–	+/–	+	C	0,8n+Пл

Следующий параметр – ЕГЭ. В 2011 году на ЕГЭ по профильному предмету записались 419 человек, что ниже порогового значения 1162. Следовательно, из таблицы 8 убираем еще четыре верхних строки:

Конкурс по предыдущему году > 1,48	**ЕГЭ > 1162**	**Ке > 2,05**	**Наличие договорников**	**Интервал**	**Диапазон значений**
	–	+/–	–	A	0,8n
–	–	+/–	+	C	0,8n+Пл

В 2010 году на специальность СДМ не поступило ни одного абитуриента на договорной основе. Поэтому в таблице 8 остается единственная строка:

Конкурс по предыдущему году > 1,48	**ЕГЭ > 1162**	**Ке > 2,05**	**Наличие договорников**	**Интервал**	**Диапазон значений**
–	–	+/–	–	A	0,8n

Таким образом, планируя прием абитуриентов в следующем году на специальность СДМ с целью выполнения контрольных цифр приема, необходимо сократить число бюджетных мест до значения, попадающего в интервал [0, 0,8n], т. е. [0, 20].

Литература

1. Чикалов И.В. Алгоритм построения деревьев решений с минимальным суммарным весом вершин // Вестн. ННГУ. Математическое моделирование и оптимальное управление. 2000. Вып. 1(22). С. 200-204.

2. Шичкина Ю.А. Матричный метод построения деревьев решения // Математическое моделирование, численные методы и комплексы программ: межвуз. темат. сб.тр. СПб., 2012. Вып. 11. С.101-109.

Preferences

1. Chikalov I. V. The algorithm for constructing decision trees with minimum aggregate weight of tree's point // Vestnik NNGU. Matematicheskoye modelirovaniye i optimal’noye upravleniye. Vyp. 1(22). Nizhniy Novgorod: Izd-vo Nizhegorodskogo un-ta, 2000. S. 200-204.

2. Shichkina Yu.A. Matrix technique for decision tree constructing // Matematicheskoye modelirovaniye, chislennye metody i kompleksy programm: Mezhvuz. temat. sb. tr. SPb., 2012. Vyp.11. S. 101-109.