УДК 519.237.8:81

А.И. Кубарев, О.В. Кукушкина, В.В. Поддубный, О.Г. Шевелев

ПОСТРОЕНИЕ ТАБЛИЦ СТИЛЕЙ ТЕКСТОВЫХ ПРОИЗВЕДЕНИЙ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМОВ КЛАССИФИКАЦИИ НА ОСНОВЕ ДЕРЕВЬЕВ РЕШЕНИЙ[1]

С использованием алгоритмов классификации текстов на основе деревьев решений предложен алгоритм оптимального по информационному критерию последовательного бинарного разбиения n-мерного признакового пространства стилей текстовых произведений на 2^n непересекающихся n-мерных интервалов, образующих таблицу стилей текстов, определяющую «стилевой портрет» (профили стилей) корпуса текстов. Предполагается, что признаковое пространство является частотным, т.е. образовано частотами появления в текстах наборов признаков (служебных слов, словосочетаний, биграмм и т.п.). Алгоритм реализован программно в системе «СтилеАнализатор», предназначенной для комплексного исследования корпусов текстов различных типов. На материале различных корпусов текстов проведено сравнительное исследование качества классификации текстов по авторам, жанрам, стилям и другим характеристикам текстов по алгоритмам деревьев решений и таблицам стилей текстов. Получаемые при обучении алгоритма профили стилей текстов могут быть использованы для идентификации стиля предъявляемого текста неизвестного автора, что позволяет, в частности, определять наиболее вероятное авторство текста.

Ключевые слова: *классификация текстов*, *деревья решений*, *таблицы стилей*, *стилевой профиль*, *идентификация текстов.*

В 2006 г. на факультете информатики Томского государственного университета был разработан программный комплекс «СтилеАнализатор» [1], предоставляющий пользователю необходимый инструментарий для автоматического анализа текстов статистическими методами, в том числе методами классификации текстов на основе деревьев решений. В последующие годы функциональные возможности системы «СтилеАнализатор» наращивались совместными усилиями факультета информатики ТГУ и лаборатории общей и компьютерной лексикологии и лексикографии филологического факультета Московского государственного университета им. М.В. Ломоносова. В частности, возникла необходимость построения и включения в «СтилеАнализатор» алгоритмов классификации с функциональностью экспертизы текстов, использующей «стилевые профили» текстов по авторам, жанрам и т.п. Настоящая работа посвящена построению таблиц стилей текстовых произведений с использованием алгоритмов классификации на основе деревьев решений.

1. Постановка задачи

Пусть имеется множество текстов определённого жанра (например, произведений художественной прозы ряда писателей). В предположении, что каждый пи-

[1] Работа поддержана грантами РФФИ № 11-07-00776-а, № 10-01-00462-а.

сатель обладает своим индивидуальным авторским стилем, его тексты образуют определённый авторский стилевой класс. В качестве признаков стиля можно использовать различные характеристики текста. Часто используют числовые характеристики – частоты встречаемости в текстах тех или иных слов (например, служебных или наиболее употребительных), словосочетаний, буквосочетаний (биграмм, триграмм и др.) и т.п. Набор таких числовых характеристик для каждого текста образует n-мерный вектор признаков стиля этого текста.

Задача состоит в том, чтобы по множеству «обучающих» текстов, принадлежность которых к авторским стилевым классам известна, и по набору n-мерных векторов признаков этих текстов построить алгоритм оптимального (по некоторому критерию, например, информационному) последовательного бинарного разбиения n-мерного признакового пространства стилей текстовых произведений на 2^n непересекающихся n-мерных интервалов, образующих таблицу стилей текстов. Такая таблица определит «стилевой портрет» (профили стилей) соответствующего корпуса текстов.

2. Классификация при помощи дерева решений

Для решения поставленной задачи необходимо использовать «признаковый» классификатор – алгоритм, позволяющий по n-мерному вектору признаков отнести текст, класс которого неизвестен, к тому или иному классу. Одним из наиболее эффективных классификаторов является классификатор, построенный на основе использования деревьев решений [2].

Как известно [2–4], деревом решений называется представленный в виде ациклического графа план, по которому производится классификация объектов, описанных набором признаков, по некоторому (обычно, информационному) критерию. В задаче классификации текстов объектами являются тексты, а признаки характеризуют классы текстов (в данной работе – это признаки авторского стиля, но это могут быть признаки, характеризующие жанровый или какой-либо другой стиль). Ограничимся количественными признаками стиля, такими, например, как частоты появления в тексте тех или иных служебных слов.

Дерево решений – это классификатор, полученный из обучающего множества, содержащего объекты (в данном случае тексты) и их характеристики (признаки класса), на основе обучения. Для обучения такого классификатора на его вход подаются тексты известных классов, например определённых авторов, являющиеся представителями классов авторских стилей.

Каждый узел дерева содержит условие ветвления по одному из признаков. В первый (корневой) узел дерева входят все тексты, представленные всеми имеющимися признаками. Для текущего узла дерева выбирается признак, наиболее подходящий с точки зрения выбранного критерия, и находится его оптимальное пороговое значение. На основе этого значения производится разделение обучающей выборки на две части. В одну часть попадают тексты, у которых значение признака ниже оптимального порога, в другую – у которых оно выше или равно порогу. Каждая из образовавшихся частей становится узлом дерева следующего уровня или концевым узлом, то есть листом, если она содержит тексты только одного класса. В дальнейшем ветвлении дерева этот признак больше не участвует. Для каждого неконцевого узла выбирается один из оставшихся признаков, наиболее подходящий с точки зрения того же критерия, и определяется оптимальное пороговое значение этого признака. На основе полученного порогового значения

признака все тексты, оказавшиеся в этом узле, разделяются, в свою очередь, на две части по значениям этого признака у каждого текста. Тексты со значениями признака разделения ниже порога попадают в одну часть, а тексты со значениями признака разделения выше или на уровне порога – в другую. Процесс ветвления идёт до тех пор, пока не будет исчерпано всё множество признаков или пока все узлы дерева не станут листьями. На этом обучение дерева решений завершается.

Взяв в качестве испытуемого любой произвольный текст и следуя по обученному таким образом дереву, сравнивая значение признака разделения этого текста в каждом узле дерева с соответствующим этому узлу пороговым значением признака разделения, мы окажемся в одном из листьев дерева, приписав тем самым текст классу этого листа.

3. Таблицы стилей текстов и профилей стилей текстов

Дерево решений в скрытом виде хранит в себе информацию о свойствах классов, характеризуемых выбранным набором признаков (частот), и позволяет классифицировать неизвестный текст. Но использовать эту информацию для сравнения классов практически невозможно, так как она не представлена в явном, интерпретируемом с этой точки зрения виде. Эту информацию надо извлечь из дерева решений и представить в таком виде, который позволит сравнивать классы между собой и оценивать степень схожести стиля неизвестного текста со стилем текстов какого-либо известного класса.

Для этой цели в данной работе предлагается строить специальные таблицы стилей текстов и профилей стилей текстов.

По исходной таблице значений признаков, полученных для обучающих текстов известных классов, предлагается сначала построить множество одноуровневых бинарных деревьев решений по каждому признаку. Множество обучающих текстов по каждому признаку разбивается оптимальным образом на два подмножества: подмножество текстов, для которых значения признака меньше некоторого порога, и подмножество текстов, для которого значение признака больше либо равно порогу.

Пример фрагмента таблицы входных данных для 9 служебных слов приведён на рис. 1. В обучающем множестве представлены 3 произведения Н.В. Гоголя («Вечера на хуторе близ Диканьки», ч. 1 и ч. 2, «Вий»), 3 – И.А. Гончарова («Обломов», «Обрыв», «Обыкновенная история»), 3 – Ф.И. Достоевского («Бедные люди», «Белые ночи», «Бесы»), 2 – А.И. Куприна («Колесо времени», «Яма»), 3 – М.Ю. Лермонтова («Ашик Кериб», «Вадим», «Герой нашего времени»), 3 – Н.С. Лескова («Загадочный человек», «На ножах», «Некуда»). В левом столбце таблицы приведены метки классов – авторы текстов, используемых для обучения классификатора. Каждая строка соответствует тексту художественного произведения какого-либо автора и представляет объект соответствующего авторского класса. В ячейках таблицы указаны абсолютные значения частот появления в соответствующем тексте того или иного служебного слова. В «СтилеАнализаторе» для обучения дерева решений (построения порогового правила разбиения множества текстов на два подмножества по каждому признаку) используются относительные частоты (доли абсолютных частот употребления того или иного служебного слова в общем числе употребления в тексте служебных слов). Это позволяет избежать искажений результатов классификации из-за различий в объёмах текстов, непосредственно влияющих на значения абсолютных частот. В результате

каждому тексту ставится в соответствие вектор-строка, содержащая в качестве признаков значения относительных частот.

Геометрически каждый текст может быть представлен тогда точкой в многомерном координатном пространстве относительных частот. Тексты каждого класса (например, каждого определённого автора) образуют в таком пространстве некоторое многомерное облако (диаграмму рассеяния элементов класса).

Блоки	в	на	с	о	через	из-за	хотя	когда	чтобы
ГогольНВ	611	502	272	30	12	2	3	63	53
ГогольНВ	817	675	437	66	18	11	17	103	70
ГогольНВ	292	212	135	32	3	5	3	40	15
ГончаровИА	3524	2401	2038	381	77	19	54	364	10
ГончаровИА	5088	3580	2991	558	166	41	76	576	126
ГончаровИА	1794	1208	1174	364	53	14	17	182	27
ДостоевскийФМ	761	481	333	101	13	6	15	105	62
ДостоевскийФМ	332	202	130	60	14	2	8	72	2
ДостоевскийФМ	5001	2468	2525	864	101	53	208	393	335
КупринАИ	465	246	188	116	8	1	11	74	26
КупринАИ	2032	1376	1125	247	96	22	41	204	121
ЛермонтовМЮ	62	32	26	3	1	2	1	6	1
ЛермонтовМЮ	699	543	282	115	27	4	14	95	20
ЛермонтовМЮ	859	701	421	105	36	13	39	97	9
ЛесковНС	1014	370	427	127	26	7	20	52	92
ЛесковНС	5171	3118	2729	782	169	54	104	466	358
ЛесковНС	4251	2472	2308	669	195	67	44	254	209

Рис. 1. Таблица входных данных алгоритма классификации по авторскому стилю (абсолютные частоты)

Рис. 2. Бинарное дерево решений для классификации текстов по частоте появления служебного слова «через»

На рис. 2 в качестве примера приведено бинарное дерево решений для разбиения текстов по одному из индивидуальных признаков стиля текста – частоте появления служебного слова «через».

Одноуровневая бинарная классификация разделила в этом примере множество 17 текстов 6 указанных в корне дерева русских писателей 19-го века на два подмножества, образующие листы дерева в соответствии с определённым критерием классификации, то есть критерием отнесения каждого текста к одному из двух (в бинарной классификации) стилевых классов.

Аналогичная классификация проведена и по другим признакам (другим служебным словам).

Порог для каждого признака находится однозначно по информационному критерию, лежащему в основе правила принятия решений по дереву решений [2–5]. Рассмотрим вопрос о выборе оптимальных порогов подробнее [2].

Пусть для обучения дерева решений (для отыскания оптимальных порогов бинарного разбиения текстов по каждому признаку) используется N текстов, образующих множество A текстов, принадлежащих K классам (авторам) и характеризующихся n признаками (относительными частотами V появления в каждом тексте, например, одного из n служебных слов).

Пусть N_k – число текстов k-го класса, $k = \overline{1,K}$, так что $\sum_{k=1}^{K} N_k = N$. Тогда вероятность того, что наугад взятый текст будет текстом k-го класса (эмпирическая априорная вероятность k-го класса) равна $P_k = N_k / N$. Статистическая неопределённость (энтропия) [6] множества классов выражается формулой

$$H(A) = -\sum_{k=1}^{K} P_k \log_2 P_k . \qquad (1)$$

Пусть по каждому j-му признаку, $j = \overline{1,n}$, выбирается порог h_j разделения текстов на два подмножества по значению V_{ij} j-го признака каждого i-го текста, так что если $V_{ij} < h_j$, то i-й текст относится к подмножеству A_{1j}, а если $V_{ij} \ge h_j$, то i-й текст относится к подмножеству A_{2j}. Объединение этих подмножеств по n признакам даёт множества A_1 и A_2, а объединение A_1 и A_2 – множество всех текстов A:

$$A_1 = \bigcup_{j=1}^{n} A_{1j} , \; A_2 = \bigcup_{j=1}^{n} A_{2j} , \; A = A_1 \bigcup A_2 .$$

Пусть N_{1jk} – число текстов k-го класса, попавших во множество A_{1j}. Тогда общее число текстов, попавших во множество A_{1j}, есть $\sum_{k=1}^{K} N_{1jk} = N_{1j}$. Вероятность того, что наугад взятый текст из множества A_{1j} будет текстом k-го класса (эмпирическая апостериорная вероятность k-го класса при условии, что в результате разделения текстов этот текст попал во множество A_{1j}, то есть его j-й признак оказался меньше порога h_j), равна $P_{1jk} = N_{1jk} / N_{1j}$.

Аналогично имеем N_{2jk}, A_{2j}, $\sum_{k=1}^{K} N_{2jk} = N_{2j}$. Вероятность того, что наугад взятый текст из множества A_{2j} будет текстом k-го класса (эмпирическая апостериорная вероятность k-го класса при условии, что в результате разделения текстов этот текст попал во множество A_{2j}, то есть его j-й признак оказался не меньше порога h_j), равна $P_{2jk} = N_{2jk} / N_{2j}$. Очевидно, $N_{1j} + N_{2j} = N$.

Вероятность попадания наугад взятого текста во множество A_{1j} (то есть эмпирическая вероятность того, что j-й признак текста окажется меньше порога h_j),

равна $P_{1j} = N_{1j} / N$. Аналогично вероятность попадания текста во множество A_{2j} (то есть эмпирическая вероятность того, что j-й признак наугад взятого текста окажется не меньше порога h_j), равна $P_{2j} = N_{2j} / N = 1 - P_{1j}$.

Следовательно, средняя энтропия множества классов при условии, что порог разделения текстов деревом решений по j-му признаку равен h_j, называемая «пороговой» энтропией [5] множества классов относительно j-го признака, выражается формулой

$$H\left(A \mid h_j\right) = -P_{1j} \sum_{k=1}^{K} P_{1jk} \log_2 P_{1jk} - P_{2j} \sum_{k=1}^{K} P_{2jk} \log_2 P_{2jk} . \quad (2)$$

Информативность разделения текстов по j-му признаку определяется как мера снятой по этому признаку неопределённости множества классов при фиксированном пороге h_j:

$$I_j\left(A \mid h_j\right) = H(A) - H\left(A \mid h_j\right). \quad (3)$$

Очевидно, эта величина зависит от порога h_j. Естественно выбрать порог разделения так, чтобы информативность (3) разделения текстов по j-му признаку была максимальной:

$$h_{j\,\mathrm{opt}} = \arg \max_{h_j} I_j\left(A \mid h_j\right). \quad (4)$$

Оптимальное значение $h_{j\,\mathrm{opt}}$ порога h_j разделения текстов по j-му признаку достаточно искать в конечном множестве середин интервалов между соседними значениями $V_{(i)j}$ и $V_{(i+1)j}$ в упорядоченном в порядке возрастания ряду значений V_{ij} j-го признака всех текстов:

$$h_j \in \left\{\left(V_{(i)\,j} + V_{(i+1)\,j}\right) \Big/ 2,\ i = \overline{1, N-1}\right\}, \quad (5)$$

где $V_{(i)j}$ – i-я порядковая статистика набора значений $\{V_{ij}\}$ j-го признака по всем N текстам обучающего множества.

Именно такое разделение текстов по относительной частоте употребления служебного слова «через» (соответствует $j = 5$ в таблице на рис. 1) в текстах нашего примера приведено на рис. 2 (оптимальный порог по этому признаку равен $h_{5\mathrm{опт}} = 0{,}000539$). Из рис. 2 видно, что $N_{1,5} = 10$ текстов из $N = 17$ попали во множество $A_{1,5}$ (3 текста 1-го класса «Гоголь», 3 – 2-го «Гончаров», 2 – 3-го «Достоевский», 1 – 4-го «Куприн», 1 – 5-го «Лермонтов»), а остальные $N_{2,5} = 7$ текстов – во множество $A_{2,5}$ (1 текст 3-го класса «Достоевский», 1 – 4-го «Куприн», 2 – 5-го «Лермонтов», 3 – 6-го «Лесков»). Разделение по остальным признакам аналогично.

Тем самым пространство признаков разбивается на 2^n непересекающихся n-мерных интервалов, порождающих пространственную таблицу стилей текстов. По ячейкам (n-мерным интервалам) этой таблицы распределяются все тексты обучающего множества. В каждый интервал попадает определённое число текстов каждого класса (в какие-то интервалы может не попасть ни одного текста, тогда эти интервалы остаются пустыми).

Можно найти долю текстов каждого класса, попавших в область ниже порога и выше порога по каждому признаку. Тогда эту таблицу можно представить двумерной таблицей, строками которой будут классы, а столбцами – признаки. В каждой ij-й ячейке этой таблицы можно указать долю текстов i-го класса, попавших в область ниже порога по j-му признаку и в область выше либо на уровень порога по тому же признаку.

Если доля текстов *i*-го класса, попавших по *j*-му признаку, скажем, выше оптимального порога, составляет величину, большую, например, 60% (доля текстов выше порога существенно больше доли текстов ниже порога), соответствующую ячейку таблицы можно выделить некоторым цветом (например, красным) или пометить некоторым символом (например, 1). Если эта доля составляет величину, меньшую, например, 40% (доля текстов выше порога существенно меньше доли текстов ниже порога), ячейку можно выделить другим цветом (например, зелёным) или пометить другим символом (например, 0). В противном случае (доля текстов выше порога существенно не отличается от доли текстов ниже порога) ячейку можно выделить, например, жёлтым цветом или пометить, например, символом «–» (знак неопределённости классификации). Одновременно в ячейке таблицы можно указать число и долю текстов, попавших ниже порога или выше либо на уровень порога. Такая таблица и образует таблицу профилей стилей текстов.

На рис. 3 приведён пример таблицы профилей авторских стилей для таблицы данных рис. 1. Сравнивая между собой строки таблицы рис. 3, можно оценить степень схожести классов. Предъявляя классификатору текст неизвестного класса, получим строку подобной таблицы, сравнивая которую со строками таблицы профилей стилей можно сказать, к какому классу ближе всего исследуемый текст по используемому критерию сравнения.

	в	на	с	о	через
▶ Порог	<0.0164840069111056>=	<0.0123084595158219>=	<0.0078515140143752>=	<0.00249284343540603>=	<0.00053873470662866 8>=
ГогольНВ	0.00%(0), 100.00%(3)	0.00%(0), 100.00%(3)	0.00%(0), 100.00%(3)	100.00%(3), 0.00%(0)	100.00%(3), 0.00%(0)
ГончаровИА	66.67%(2), 33.33%(1)	100.00%(3), 0.00%(0)	0.00%(0), 100.00%(3)	66.67%(2), 33.33%(1)	100.00%(3), 0.00%(0)
ДостоевскийФМ	66.67%(2), 33.33%(1)	100.00%(3), 0.00%(0)	66.67%(2), 33.33%(1)	33.33%(1), 66.67%(2)	66.67%(2), 33.33%(1)
КупринАИ	0.00%(0), 100.00%(2)	100.00%(2), 0.00%(0)	50.00%(1), 50.00%(1)	50.00%(1), 50.00%(1)	50.00%(1), 50.00%(1)
ЛермонтовМЮ	66.67%(2), 33.33%(1)	33.33%(1), 66.67%(2)	100.00%(3), 0.00%(0)	66.67%(2), 33.33%(1)	33.33%(1), 66.67%(2)
ЛесковНС	0.00%(0), 100.00%(3)	100.00%(3), 0.00%(0)	0.00%(0), 100.00%(3)	0.00%(0), 100.00%(3)	0.00%(0), 100.00%(3)

	через	из-за	хотя	когда	чтобы
▶ Порог	<0.00053873470662866 8>=	<0.00017214008377484 1>=	<0.00327638958298626>=	<0.00170842824601367>=	<0.00085736203239926>
ГогольНВ	100.00%(3), 0.00%(0)	33.33%(1), 66.67%(2)	66.67%(2), 33.33%(1)	0.00%(0), 100.00%(3)	0.00%(0), 100.00%(3)
ГончаровИА	100.00%(3), 0.00%(0)	100.00%(3), 0.00%(0)	100.00%(3), 0.00%(0)	33.33%(1), 66.67%(2)	100.00%(3), 0.00%(0)
ДостоевскийФМ	66.67%(2), 33.33%(1)	66.67%(2), 33.33%(1)	0.00%(0), 100.00%(3)	33.33%(1), 66.67%(2)	33.33%(1), 66.67%(2)
КупринАИ	50.00%(1), 50.00%(1)	50.00%(1), 50.00%(1)	0.00%(0), 100.00%(2)	0.00%(0), 100.00%(2)	0.00%(0), 100.00%(2)
ЛермонтовМЮ	33.33%(1), 66.67%(2)	33.33%(1), 66.67%(2)	0.00%(0), 100.00%(3)	0.00%(0), 100.00%(3)	100.00%(3), 0.00%(0)
ЛесковНС	0.00%(0), 100.00%(3)	0.00%(0), 100.00%(3)	33.33%(1), 66.67%(2)	100.00%(3), 0.00%(0)	0.00%(0), 100.00%(3)

Рис. 3. Таблица профилей стилей текстов

В качестве признаков, по которым можно оценивать схожесть классов, могут выступать величины, измеряемые в шкале отношений, например проценты (доли) текстов каждого класса, попавших по каждому исходному признаку (относительной частоте) в область ниже порога. Но можно использовать в качестве признаков схожести, наоборот, доли попадания относительных частот в область не ниже (выше или на уровень) порога.

А можно использовать в качестве признаков схожести также величины, измеряемые в номинальной шкале: троичную «цветовую» переменную, принимающую значения «красный», «зелёный», «жёлтый», о чём говорилось выше при описании раскраски таблицы профилей стилей. Или троичную «яркостную» переменную,

принимающую значения «тёмно-серый», «серый», «светло-серый». Им может соответствовать переменная троичной логики со значениями «1», «0», «–» («истина», «ложь», «неопределённость»).

Геометрический смысл каждой строки (класса), описывающей признаки схожести классов в шкале отношений, заключается в том, что каждому облаку точек-текстов класса в исходном частотном признаковом пространстве ставится в соответствие одна точка в пространстве признаков схожести классов, характеризующая непосредственно стиль класса. Такое представление существенно упрощает задачу сравнения стилей авторов (классов). Так, взяв за координаты точек-стилей (за признаки сравнения) проценты текстов, оказавшихся ниже порога по каждому признаку, и расставив эти точки в пространстве признаков сравнения, можно легко интерпретировать схожесть стилей: две точки достаточно близки друг к другу – стили классов схожи, две точки достаточно удалены друг от друга – стили отличаются. В этом случае в качестве меры близости точек (метрики в пространстве признаков сравнения) можно взять евклидово расстояние между точками. В случае использования в качестве признаков сравнения логических переменных «расстояние» между классами можно определять числом несовпадений значений признаков сравниваемых классов.

4. Дерево кластеризации профилей стилей текстов

С использованием шкалы отношений и выбранной меры расстояния между классами можно построить иерархическое дерево кластеризации [7] (здесь уместно использовать кластеризацию «по дальнему соседу») и визуализировать картину близости классов. Так, при евклидовой метрике сравнения классов по процентам попадания текстов в области ниже порога для рассматриваемого примера данных, представленного на рис. 1, получаем горизонтальную древовидную диаграмму кластеризации, показанную на рис. 4. Диаграмма получена из таблицы профилей авторских стилей текстов рис. 3 в программном продукте «СтилеАнализатор».

Рис. 4. Горизонтальная древовидная диаграмма кластеризации

Сначала каждый класс авторских стилей образует свой кластер (левая часть диаграммы). Горизонтальная ось представляет расстояние объединения кластеров в новые кластеры. Для каждого узла дерева кластеризации (там, где формируется новый кластер) в верхней строке рис. 4 против вертикальных точечных линий можно видеть величину расстояния, на котором соответствующие элементы (предыдущие кластеры) объединяются в новый кластер. Числа, помечающие узлы, нумеруют последовательность образования кластеров.

Обратим внимание, что в использованном здесь методе кластеризации (методе полной связи, методе «дальнего соседа») расстояние между кластерами определяется наибольшим расстоянием между двумя объектами из разных кластеров.

5. Логическое сравнение профилей стилей текстов

С использованием номинальной шкалы «расстоянием» между классами может служить число несовпадений сравниваемых строк по всем ячейкам значений введённых выше троичных переменных, характеризующих признаки схожести стилей текстов. Так, при сравнении авторских классов по «цветовой» или «яркостной» гамме из таблицы рис. 4 получается следующая таблица чисел несовпадения троичных признаков:

Таблица 1

	Гоголь	Гончаров	Достоевский	Куприн	Лермонтов	Лесков
Гоголь	0	4	6	6	5	5
Гончаров	4	0	4	7	5	7
Достоевский	6	4	0	5	5	5
Куприн	6	7	5	0	7	5
Лермонтов	5	5	5	7	0	6
Лесков	5	7	5	5	6	0

Пусть теперь имеем текст, автор которого исследователю неизвестен и который характеризуется признаками, приведёнными в первых трёх строках следующей таблицы:

Таблица 2

Признак	в	на	с	о	через	из-за	хотя	когда	чтобы
Абс. част.	1868	992	841	197	38	27	48	147	229
Отн. част.	0.0195	0.0142	0.0090	0.0021	0.0002	0.0003	0.0002	0.0027	0.0010
Профиль	1	1	1	0	0	1	0	1	1

Сравнивая относительные частоты признаков с порогами таблицы рис. 3, получим стилевой профиль этого текста. Запишем его в четвёртую строку табл. 2. Сравнивая эту строку со строками профилей стилей авторских классов таблицы рис. 3, видим, что этот профиль наиболее близок профилю авторского стиля Н.В. Гоголя (просто совпадает с ним: число несовпадений равно нулю). Действительно, предъявленный текст был произведением Н.В. Гоголя «Мёртвые души», том 2.

Отбор минимальных наборов признаков (безызбыточных, или тупиковых тестов), достаточных для распознавания классов, может быть проведён на основе методов логического распознавания (например, [8, 9]) по таблице профилей стилей текстов рис. 3.

Заключение

На основе использования алгоритмов деревьев решений предложен алгоритм построения таблицы профилей стилей текстов, принадлежащих заданным классам (авторским, жанровым и др.). Получаемые при обучении алгоритма профили стилей текстов могут быть использованы для идентификации стиля предъявляемого текста неизвестного автора, что позволяет, в частности, определять наиболее вероятное авторство текста.

ЛИТЕРАТУРА

1. *Шевелев О.Г.* Разработка и исследование алгоритмов сравнения стилей текстовых произведений: автореф. дис. … канд. техн. наук / Том. гос. ун-т. Томск, 2006. 19 с.
2. *Шевелёв О.Г.* Методы автоматической классификации текстов на естественном языке: уч. пособие. Томск: ТМЛ-Пресс, 2007. 144 с.
3. Деревья решений – общие принципы работы. [Электронный ресурс]. URL: http://www.basegroup.ru/library/analysis/tree/description/ (дата обращения 12.09.2011).
4. Деревья решений – C4.5 математический аппарат. Часть 1. [Электронный ресурс]. URL: http://www.basegroup.ru/library/analysis/tree/math_c45_part1/ (дата обращения 12.09.2011).
5. *Moore A.* Statistical Data Mining Tutorials. [Электронный ресурс]. URL: http://www.cs.cmu.edu/~awm/tutorials/ (дата обращения 12.09.2011).
6. *Волькенштейн М.В.* Энтропия и информация. М.: Наука, 1986.
7. *Мандель И.Д.* Кластерный анализ. М.: Финансы и статистика, 1988.
8. *Дюкова Е.В.* Дискретные (логические) процедуры распознавания: принципы конструирования, сложность реализации и основные модели: учебное пособие. М.: Прометей, 2003. 29 с.
9. *Lias-Rodriguez A.*, *Pons-Porrata A.* BR: A New Method for Computing All Typical Testors // E. Bayro-Corrachano and J.-O. Eklundh (Eds.): CIARP 2009, LNCS 5856. P. 433–440, 2009.

Кубарев Антон Игоревич
Поддубный Василий Васильевич
Шевелёв Олег Геннадьевич
Томский государственный университет
Кукушкина Ольга Владимировна
Московский государственный университет им. М.В. Ломоносова
E-mail: kubarev_ai@mail.ru, vvpoddubny@gmail.com,
oshevelyov@gmail.com, kukush@orc.ru

Поступила в редакцию 18 сентября 2012 г.

Kubarev Anton I., *Kukushkina Olga V.*, *Poddubny Vasiliy V.*, *Shevelyov Oleg G.* (Tomsk State University, Lomonosov Moscow State University). **Building a style sheet using text classification algorithms based on decision trees.**

Keywords: text classification, decision trees, styles sheet, styles profile, text identification.

With the use of classification algorithms based on the decision trees, we propose algorithm of building an optimal information criterion sequential binary partition of *n*-dimensional feature space of texts on 2^n disjoint *n*-dimensional intervals that make up the styles sheet that defines the «style portrait» (profile sheet) of texts set. It is assumed that space is indicative of the frequency, formed frequencies of appearance in texts set of function words, phrases, bigrams, etc. The algorithm implemented in software in the system «StyleAnalizator» intended for a comprehensive study of texts of various types. On a material of different body of texts, a comparative study of classification quality of texts by authors, genres, styles and other characteristics of the texts on algorithms for decision trees and tables of texts styles is made. The resulting algorithm for training text styles profiles can be used to identify the charge against the style of the text by an unknown author. This allows, in particular, determine the most probable authorship of the text.