УДК 808.2:159.937

М. А. ЗИЛЬБЕРГЛЕЙТ, А. С. РЫЖАНКОВА, Белорусский государственный технологический университет

АНАЛИЗ УСТОЙЧИВОСТИ КЛАССИФИКАЦИОННЫХ ФУНКЦИЙ ПРИ ОБРАБОТКЕ ТЕКСТОВОЙ ИНФОРМАЦИИИ

В статье описан анализ устойчивости классификационных функций при обработке текстовой информации на основе применения различных методов преобразования данных. В силу того, что учебный текст в данном исследовании рассмотрен как формально-логическое образование, в качестве его основных характеристик определены 14 статистических параметров, составляющих факторное пространство объектов исследования. Анализ устойчивости классифицирующих правил проведен на основе анализа рассчитанных для каждого из методов преобразования коэффициентов вариации.

In article is described the analysis of stability of classification functions when processing text information on the basis of application of various methods of transformation of data. That the educational text in this research is considered as formal and logical education, as its main characteristics 14 statistical parameters making factorial space of objects of research are determined. The analysis of stability of the classifying rules is carried out on the basis of the analysis calculated for each of methods of transformation of coefficients of a variation

В настоящее время при обработке текстовой информации широкое распространение получили методы, в основе анализа которых лежит термин, определяемый в англоязычной литературе как *readability* – читабельность, либо удобочитаемость. Существует достаточное количество формул читабельности, среди которых наибольшее распространение получили формулы: Флеша; Флеша-Кинкейда; Ганнинга; Дейла-Чолл; Пауэрса-Самнера-Кера; FORCAST; Спеша; Индекс Колемана-Лиау; Индекс SMOG; Автоматический индекс читабельности; формула письма, – а также два графических метода: график Фрая и график Рэйгора [1]. Стоит отметить, что часть из них позволяет установить возрастные планки для успешного освоения текстовой информации, другая часть оценивает текст с позиций его сложности и доступности. В русскоязычной литературе число работ, посвященных рассмотрению данного понятия чрезвычайно невелико. Среди них наибольшее распространение получили исследования [2–10] и др. Однако, в указанных работахоценка информационных характеристик текста с позиций *readability*, основанная на использовании методов распознавания образов, представлена на недостаточном уровне.

Чаще всего результаты анализа по распознаванию образов представляют в виде классификационных функций, позволяющих проводить разделение новых объектов по сформулированным решающим правилам. До настоящего времени мы не встречали работ, в которых бы были рассмотрены вопросы устойчивости таких систем к различным помехам и ошибкам, связанным с использованием эмпирических данных при обработке текстовой информации. Известно, что при работе на неполных выборках существует риск получения неустойчивых результатов [11]. Сказанное выше демонстрируется на рис. 1.

В настоящей работе для поиска устойчивого решения задачи исследования был использован ряд выборок, характеризующихся набором из 14 статистических параметров: N_1 – средняя длина слов в слогах; N_2 – средняя длина слов в буквах; N_3 – средняя длина слов по Деверу; N_4 – средняя длина слов в 3 слога и более; N_5 – средняя длина слов в 4 слога и более; N_6 – средняя длина слов в 5 слогов и более; N_7 – средняя длина слов в 6 слогов и более; N_8 – средняя длина слов в 7 слогов и более; N_9 – процент односложных слов; N_{10} – средняя длина предложения в словах; N_{11} – средняя длина предложения в слогах; N_{12} – процент чисел от общего количества слов; N_{13}– от-

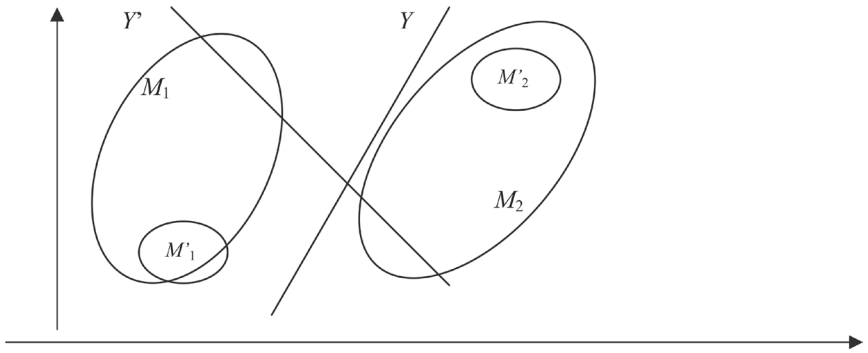

Рис. 1. Ошибки при работе на неполных выборках: M_1, M_2 – полные выборки, M'_1, M'_2 – частные выборки, Y – решающее правило на полную выборку, Y' – решающее правило на частную выборку

ношение показателя «Средняя длина слов в 3 слога и более» к показателю «Средняя длина слов в 6 слогов и более»; N_{14} – отношение показателя «Средняя длина слов в 4 слога и более» к показателю «Средняя длина слов в 6 слогов и более».

Указанные параметры описывают информационную структуру учебных изданий по специальности «Издательское дело» с позиций формально-логического подхода. В результате выполнения операций по снижению размерности факторного пространства методами корреляционных плеяд, главных компонент, кратчайшего незамкнутого пути, факторного и кластерного анализов были сформированы три класса регрессоров текстовых фрагментов, описывающих информационную структуру учебных изданий: 1 класс характеризуют факторы, описывающие среднюю длину предложения; 2 класс – факторы, описывающие дифференциальную структуру текста; 3 класс – факторы, описывающие среднюю длину слов в различных единицах измерения.

Для определения сложности каждого из текстовых компонентов был проведен эксперимент, заключающийся в реализации опроса по методу балльных оценок, методике дополнений и методу парных сравнений (735 респондентов) [12]. Полученные оценки позволили определить категории качества текста для каждого из методов опроса, а также сформировать обучающие выборки, на основе использования которых можно установить точность классификации объектов по анализу их статистической структуры.

В результате было установлено, что точность классификации объектов с использованием обучающей выборки достигает высоких результатов при анализе данных методом дискриминантного анализа (97,03% – для метода балльных оценок (МБО), 97,03% – для метода дополнений (МД), 98,02% – для метода парных сравнений (МПС)).

На данном этапе исследование можно считать завершенным, однако, возникает вопрос, насколько полученные результаты устойчивы к помехам, возникающим при замене объектов либо изменении статистических параметров текста. Известно, что для повышения устойчивости решений применяют ряд методов, изменяющих чувствительность модели к исходным данным. К ним относятся преобразования, основанные на использовании: степенной, логарифмической, квадратичной функций; трансформация по методу Бокса-Кокса и др.

Понятие устойчивости в литературе по прикладной информатике достаточно широко варьируется. Так известны различные меры устойчивости, предложенные Тьюки [14], Ляпуновым [15], Тихоновым [16] и Тагучи [13]. В настоящей работе в качестве меры устойчивости нами был принят коэффициент вариации [13], рассчитанный для каждого из методов преобразования как мера относительного разброса случайной величины, которая характеризует качество через коэффициент вариации.

При преобразовании данных методом линейного дискриминантного анализа были получены следующие результаты. Точность классификации объектов первой выборки для МБО – 84%, МД – 86%, МПС – 80%, объектов второй выборки: МБО – 86,27%, МД – 82,35%, МПС – 82,35%, объектов третьей выборки: МБО – 76,24%, МД – 83,17%, МПС – 81,19%. Статистические показатели: среднее значение – 82,17 (МБО), 83,84 (МД), 81,18 (МПС); дисперсия – 27,66 (МБО), 3,67 (МД), 1,38 (МПС).

Одним из способов преобразования был перевод значений статистических параметров

текстов в порядковую шкалу. Для этого для каждого из факторов были определены минимальные и максимальные значения и выделены границы 10 интервалов. Результаты такого преобразования: точность классификации объектов первой выборки для МБО – 80%, МД – 80,00%, МПС – 78,00%, объектов второй выборки: МБО – 72,55%, МД – 80,39%, МПС – 84,31%, объектов третьей выборки: МБО – 72,28%, МД – 89,11%, МПС – 79,21%. Статистические показатели: среднее значение – 74,94(МБО), 83,17(МД), 80,51(МПС); дисперсия – 19,20(МБО), 26,53(МД), 11,22(МПС).

Был применен метод преобразования, основанный на вычислении среднего значения для каждого из выделенных ранее интервалов. Результаты преобразования: точность классификации объектов первой выборки для МБО – 84,00%, МД – 94,00%, МПС – 80,00%, объектов второй выборки: МБО – 78,43%, МД – 90,20%, МПС – 84,31%, объектов третьей выборки: МБО – 72,28%, МД – 89,11%, МПС – 79,21%. Статистические показатели: среднее значение – 78,24 (МБО), 91,10 (МД), 81,17 (МПС); дисперсия – 34,37 (МБО), 6,59 (МД), 7,54 (МПС).

Для преобразования данных было применено преобразование, основанное на использовании логарифмической функции. Результаты: точность классификации объектов первой выборки для МБО – 86,00%, МД – 92%, МПС – 84%, объектов второй выборки: МБО – 84,31%, МД – 88,24%, МПС – 84,31%, объектов третьей выборки: МБО – 78,22%, МД – 85,15%, МПС – 80,20%. Статистические показатели: среднее значение – 82,84 (МБО), 88,46 (МД), 82,84 (МПС); дисперсия – 16,75 (МБО), 11,77 (МД), 5,24 (МПС).

Преобразование данных методом извлечения корня квадратного показало следующие результаты: точность классификации объектов первой выборки для МБО – 90%, МД – 90%, МПС – 88%, второй выборки: МБО – 86,27%, МД – 86,27%, МПС – 82,35%, третьей выборки: МБО – 76,24%, МД – 83,17%, МПС – 79,21%. Статистические показатели: среднее значение – 84,17 (МБО), 86,48 (МД), 83,19 (МПС); дисперсия – 50,64 (МБО), 11,70 (МД), 19,84 (МПС).

Одним из наиболее распространенных способов преобразования данных является трансформация методом Бокса-Кокса. Результаты такого преобразования: точность классификации объектов первой выборки для МБО – 88%, МД – 94%, МПС – 90%, объектов второй выборки: МБО – 88,24%, МД – 88,24%, МПС – 84,31%, объектов третьей выборки: МБО – 78,22%, МД – 84,16%, МПС – 83,17%. Статистические показатели: среднее значение – 84,82 (МБО), 88,80 (МД), 85,83 (МПС); дисперсия – 32,68 (МБО), 24,44 (МД), 13,39 (МПС).

В результате выполнения преобразований на основе применения степенной функции установлено: точность классификации объектов первой выборки для МБО – 92%, МД – 92%, МПС – 80%, объектов второй выборки: МБО – 90,2%, МД – 84,31%, МПС – 80%, объектов третьей выборки: МБО – 78,22%, МД – 86,14%, МПС – 78,22%. Статистические показатели: среднее значение – 86,81 (МБО), 87,48 (МД), 79,54 (МПС); дисперсия – 56,11 (МБО), 16,14 (МД), 1,34 (МПС).

В результате использования квадратичной функции для трансформации данных были получены результаты: точность классификации объектов первой выборки для МБО – 82%, МД – 88%, МПС – 88%, второй выборки: МБО – 88,24%, МД – 88,24%, МПС – 80,39%, третьей выборки: МБО – 77,23%, МД – 85,15%, МПС – 83,17%. Статистические показатели: среднее значение – 82,49 (МБО), 87,13 (МД), 83,85 (МПС); дисперсия – 30,49 (МБО), 2,95 (МД), 14,83 (МПС).

В табл. 1 приведены результаты распознавания объектов. В табл. 2 – статистические показатели проведенного анализа.

В табл. 3 представлены рассчитанные значения коэффициентов вариации.

Таблица 1. **Результаты распознавания объектов**

Наименование	Точность классификации, %								
	2-й порядок			3-й порядок			4-й порядок		
	МБО	МД	МПС	МБО	МД	МПС	МБО	МД	МПС
Объекты первой выборки	98	96	92	100	100	98	82	86	78
Объекты второй выборки	90,2	96,08	92,16	98,04	100	98,04	45,1	82,35	21,57
Объекты третьей выборки	78,22	86,14	87,13	83,17	90,1	91,09	87,13	95,05	94,06

Таблица 2. **Статистические показатели анализа**

	2-й порядок			3-й порядок			4-й порядок		
	МБО	МД	МПС	МБО	МД	МПС	МБО	МД	МПС
Среднее значение	88,81	92,74	90,43	93,74	96,7	95,71	71,41	87,80	64,54
Дисперсия	99,27	32,67	8,17	84,7	32,67	16,01	525,74	42,75	1449,51
Среднеквадратичное отклонение	9,96	5,72	2,86	9,2	5,72	4,00	22,93	6,54	38,07

Таблица 3. **Значения коэффициентов вариации**

Метод преобразования	МБО, %	МД, %	МПС, %
Линейный дискриминантный анализ	6,40	2,28	1,45
Перевод в порядковую шкалу	5,85	6,19	4,16
Среднее значение интервала	7,49	2,82	3,38
Логарифмическая функции	4,94	3,88	2,76
Извлечение корня квадратного	8,45	3,95	5,35
Метод Бокса-Кокса	6,74	5,57	4,26
Использование степенной функции	8,63	4,59	1,45
Использование квадратичной функции	6,69	1,97	4,59
Дискриминантный анализ 2-го порядка	11,22	6,16	3,16
Дискриминантный анализ 3-го порядка	9,82	5,91	4,18
Дискриминантный анализ 4-го порядка	32,11	7,45	58,99

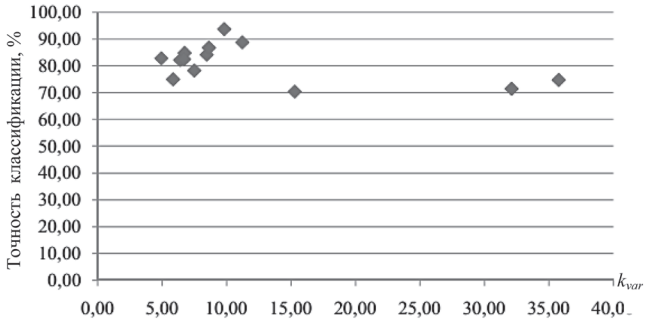

Рис. 2. Графические результаты анализа для метода балльных оценок

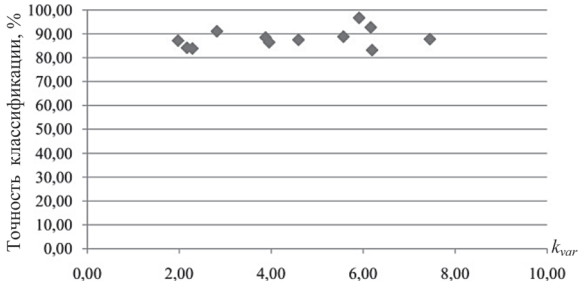

Рис. 3. Графические результаты анализа для методики дополнений

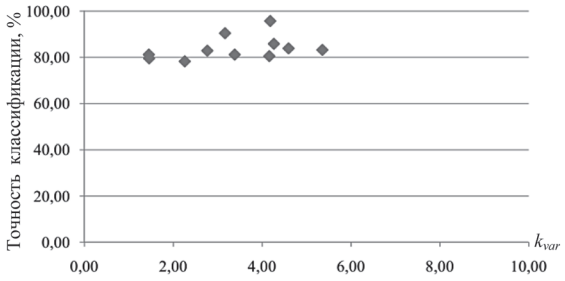

Рис. 4. Графические результаты анализа для метода парных сравнений

Коэффициенты вариации для дискриминантного уравнения пятого порядка составляют для МБО – 35,77%, для МД – 64,28%, для МПС – 59,67%. На рис. 2, 3, 4 представлены графические результаты. Ось абсцисс содержит значения коэффициентов вариации, ось ординат – средние значения процентного соотношения совпадений расчетов с результатами экспериментальной части.

Заключение

В результате поиска устойчивого к помехам решения было определено, что задачам исследования отвечают преобразования двух типов, основанных на использовании логарифмической и степенной функций. Однако наименьшие значения коэффициентов вариации наблюдаются при использовании преобразования с помощью десятичного логарифма.

Литература

1. Readability Formulas. – Режимдоступа: http://www.readabilityformulas.php.

2. **Филиппова А. В.** Управление качеством учебных материалов на основе анализа трудности понимания учебных текстов/ А. В. Филиппова. – Уфа. 2010.

3. **Мацковский М. С.** Проблемы читабельности печатного материала // Смысловое восприятие речевого сообщения в условиях массовой коммуникации / отв. ред. Т. М. Дридзе, А. А. Леонтьев. – М.: Наука, 1976. – С. 126–142.

4. **Доблаев Л. П.** Анализ и понимание текста / Сарат. гос. ун-т им. Н. Г. Чернышевского. – Саратов: Изд-во Сарат. Ун-та, 1987. – 69 с.

5. **Вул С. М.** Статистическое исследование текстов с помощью ЭВМ и дисплея в целях установления авторства // Применение ЭВМ в судебно-экспертных исследованиях и поиске правовой информации. – М.: Наука, 1975.

6. **Микк Я. А.** Оптимизация сложности учебного текста. – М.: Просвещение, 1981. – 119 с.

7. **Вильде Л. Де.** Количественный анализ текста… // Лингвистика. Межкультурная коммуникация. – Сб. науч. тр. – Вып. 1. – Курск, 1997.

8. **Криони Н. К.** Автоматизированный анализ сложности учебного текста / Криони Н. К., Никин А. Д., Филиппова А. В. // Труды международной научно-практической конференции «Новые информационные технологии в образовании».– Екатеринбург, 2009 г.

9. **Кубрякова Е. С.** О тексте и критериях его определения // Текст. Структура и семантика.– Т. 1. – М., 2001 г. – С. 72–81.

10. **Оборнева И. В.** Автоматизированная оценка сложности учебных текстов на основе статистического анализа: автореферат диссер. на соискание учен. степ. 13.00.02 Теория и методика обучения и воспитания (информатизация образования) /И. В. Оборнева.– М.: РАН Институт содержания и методов обучения, 2006.

11. **Глаз А. Б.** Параметрическая и структурная адаптация решающих правил в задачах распознавания. – Рига: Зинатне, 1988. – 167 с.

12. Сравнительный анализ методов опроса и компьютерного анализа данных для изучения восприятия текстов студентами высших учебных заведений / А. С. Малюкевич, М. А. Зильберглейт // Известия ГГУ им. Ф. Скорины. – № 6 (81) Естественные науки. – ISSN 1609–9672. – Минск, 2013. – С. 134–138.

13. Управление качеством. Робастное проектирование. Метод Тагути / Леон Р., Шумейкер А., Какар Р., Кац Л., Фадке М. и др.; Пер с англ. – М.: «СЕЙФИ» 2002. – 384 с.

14. **Орлов А. И.** Прикладная статистика. – М.: Издательство «Экзамен», 2004. – 656 с.

15. **Андреев В. С.** Теория нелинейных электрических цепей: учеб. пособ. для вузов. – М.: Радио и связь, 1982. – с. 135–136.

16. **Тихонов А. Н.** Об устойчивости обратных задач // Докл. АН СССР. Нов. сер.- 1943.– Т.39, № 5.– С. 195–198.