ГЛАВА 1. Профильные методы классификации

1.1.1 Статистический подход выявления информативных терминов

Для установления и оценки уровня значимости связи между различными переменными традиционно применяются статистические тесты. Наиболее известным тестом, используемым в тех случаях, когда данные представимы в виде таблиц сопряженности, в ячейках которой содержатся частоты совместной встречаемости (или невстречаемости) переменных, является χ^2-тест. Он основан на применении критерия согласия χ^2. При этом статистика χ^2- критерия использует разность между эмпирическими наблюдаемыми и теоретически ожидаемыми частотами. Расчет теоретически ожидаемых частот проводится в предположении о справедливости нулевой гипотезы о статистической независимости рассматриваемых переменных.

Для выявления информативных терминов используется таблица сопряженности размером 2x2,

Таблица 1.1 – таблица сопряженности 2x2

Q_k / X	Принадлежность классу Q_k	Непринадлежность классу Q_k	$\sum$
Наличие термина $x^{(i)}$	A	B	A+B
Отсутствие термина $x^{(i)}$	C	D	C+D
$\sum$	A+C	B+D	N

Здесь A – число документов класса Q_k, в которых встречается термин $x^{(i)}$; B – число документов другого класса (не Q_k), содержащих термин $x^{(i)}$; C – число документов класса Q_k без термина $x^{(i)}$; D – число документов других

классов (не Q_k), в которых не встречается термин $x^{(i)}$; N – общее количество наблюдений в выборке.

Величина χ^2-статистики определяется по формуле:

$$\chi^2(x^{(i)}, Q_k) = N \cdot \frac{(AD - BC)^2}{(A+B)(C+D)(A+C)(B+D)}, \quad (1.1)$$

Выбирая уровень значимости α, можно определить значение χ^2-теста с числом степеней свободы γ=1 (для случая таблицы сопряженности размером 2x2). Большие значения χ^2 в формуле (1.1) означают, что гипотеза о независимости не выполняется.

Недостатком χ^2 – критерия является его невысокая точность для редко встречающихся терминов, которые имеют малые частоты в таблице сопряженности признаков. Кроме того, выборочное распределение χ^2 хорошо аппроксимирует табличное распределение только для достаточно больших размеров выборок ($N \geq 30$), поэтому в случае малых выборок интерпретация результатов χ^2-теста может быть затруднена даже для часто встречающихся терминов.

Произвольный диапазон изменения значений χ^2-статистики тоже является недостатком и не соответствует ранее сформулированному условию о принадлежности значений весов интервалу [0;1] (или [-1;1]). Вместе с тем в литературе хорошо известны нормированные варианты χ^2- критерия. Например, коэффициент корреляции двух номинальных переменных, рассчитываемый по формуле:

PO-профиль: $$\rho(x^{(i)}, Q_k) = \sqrt{\frac{\chi^2(x^{(i)}, Q_k)}{N}} = \frac{(AD - CB)}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}. \quad (1.2)$$

Пределы измерения коэффициента корреляции заключаются в интервале от –1 до +1.

В ряде исследований вместе с мерами, основанными на χ^2 - критерии, рассматривается также идеологически близкая мера, известная как Q-статистика или коэффициент Юла. Q-статистика рассчитывается по следующей формуле:

$$\textit{Q-профиль: } Q = \frac{AD - CB}{AD + CB}, \tag{1.3}$$

Необходимо отметить высокую согласованность поведения коэффициента корреляции и Q-статистики. Однако PO-профиль становится равным единице только в случае, если A и D или C и B одновременно равны нулю. Что касается Q–статистики, то она равна единице уже в случае, если одно из чисел в клетках двумерной таблицы сопряженности равно нулю. Величина коэффициента корреляции всегда ниже значения Q-статистики.

Далее в работе будет изучаться целесообразность использования PO-профиля и Q-профиля для взвешивания терминов документов.

1.1.2 Теоретико-информационный подход выявления информативных терминов

В теории информации часто возникает задача выявления количества информации, содержащейся в одном ансамбле сообщений *{X}* относительно другого, зависящего от него ансамбля *{Q}*. Для решения задачи рассчитывается взаимная информация, показывающая количество информации о принадлежности документа $\vec{x}$ к классу Q_k, которое несет появление в нем термина $x^{(i)}$.

Используя те же обозначения, что для χ^2–критерия, формула для расчета информативности термина по критерию взаимной информации (МИ-профиль) имеет вид:

$$MI(x^{(i)}, Q_k) = \log_2 \frac{A \cdot N}{(A+B) \cdot (A+C)}, \tag{1.4}$$

В отличие от χ^2–критерия, данный подход предпочитает отбирать в качестве информативных наиболее редкие специфические слова.

Следует заметить, что веса терминов в МИ-профиле изменяются в произвольном диапазоне, поэтому в работе для расчетов весов используется нормированный МИ-профиль:

Нормированный МИ-профиль (НМИ):

$$NMI(x^{(i)}, Q_k) = \frac{A\log_2 \frac{AN}{(A+B)(A+C)}}{(A+B)\log_2 \frac{N}{A+B}} \qquad (1.5)$$

1.1.3 Эвристический подход

К данному подходу относится большой класс коэффициентов ассоциативности (КА). КА широко используются в математической статистике в качестве меры сходства между объектами, описанными бинарными признаками. Они основаны на анализе соотношения совпадающих (и/или несовпадающих) признаков у различных объектов.

Несмотря на то, что в настоящее время предложено большое число мер ассоциативности, лишь некоторые из них подверглись комплексной проверке и экспериментальным исследованиям.

Имеется две группы коэффициентов ассоциативности, которые наиболее часто используются на практике и заслуживают специального рассмотрения. К первой группе относятся *КА*, для расчета которых используются все четыре параметра из таблицы 1.1 (параметры *A,B,C,D*), вторая группа включает *КА*, вычисляемые по трем параметрам (*A,B,C*). Отметим, что использование *КА* из первой группы в ряде случаев может привести к результату, когда документы оказываются схожи главным образом за счет того, что им обоим не свойственны некоторые термины (параметр *D*), а не за счет наличия общих характеристик.

К коэффициентам ассоциативности первой группы относятся:

Простой коэффициент совстречаемости $$S(x^{(i)}, Q_k) = \frac{A+D}{A+B+C+D} \qquad (1.6)$$

Первый коэффициент несогласия $$SN1(x^{(i)}, Q_k) = \frac{C+B}{A+B+C+D} \quad (1.7)$$

Коэффициент Рассела-Рао $$RR(x^{(i)}, Q_k) = \frac{A}{A+B+C+D} \quad (1.8)$$

Коэффициент Роджерса-Танимото $$RT(x^{(i)}, Q_k) = \frac{A+D}{A+D+2(B+C)} \quad (1.9)$$

Первый коэффициент Сокала-Сниса $$SS2(x^{(i)}, Q_k) = \frac{2(A+D)}{2(A+D)+B+C} \quad (1.10)$$

Коэффициент Хаммана $$H(x^{(i)}, Q_k) = \frac{(A+D)-(B+C)}{A+B+C+D} \quad (1.11)$$

Значения коэффициентов, получаемых по формулам (1.6) - (1.11), лежат в интервале [0,1] (или [-1,1]), что облегчает интерпретацию результатов. Так, значение «1» свидетельствует о полном совпадении признаков, а значение «0» означает отсутствие связи.

Вторая группа коэффициентов ассоциативности включает:

Коэффициент Жаккара, $$J(x^{(i)}, Q_k) = \frac{A}{A+B+C} \quad (1.12)$$

Второй коэффициент несогласия, $$SN2(x^{(i)}, Q_k) = \frac{C+B}{A+B+C} \quad (1.13)$$

Второй коэффициент Сокала-Сниса , $$SS2(x^{(i)}, Q_k) = \frac{A}{A+2(B+C)} \quad (1.14)$$

Значения коэффициентов, получаемых из формул (1.12) – (1.14), также лежат в интервале [0,1].

1.2. Обучение профильных методов и построение решающего правила

На этапе обучения профильных методов проводится выявление набора информативных терминов каждого класса на основе расчета весов по одной из описанных выше формул (1.1) - (1.14). После чего составляется профиль класса – вектор, отсортированный по убыванию весов. Термины, имеющие наибольший вес, являются наиболее информативными (согласно выбранному

критерию). Единственным настраиваемым параметром всех профильных методов является значение L, которое определяет длину профиля классов L_k ($L_k < M$), т.е. сколько информативных терминов и с каким весом включается в профиль. Длина профиля L влияет на точность и быстродействие классификации. В общем случае длина профилей у различных методов может не совпадать. В данной работе используется одинаковая длина профиля для всех методов и всех классов ($L_1 = L_2 = ,..., = L_K$).

В большинстве работ в области *Text Categorization* процедуры выявления информативных терминов применяются для сокращения признакового пространства и являются частью предварительной обработки текстов. Такой подход предполагает отбор дискриминирующих терминов вне зависимости от классификатора, который планируется использовать для классификации. Результирующая подсистема терминов вряд ли будет “оптимальной” для всех методов классификации.

В данной диссертации выбор информативных терминов рассматривается как составная часть обучения метода. Важно отметить еще одно принципиальное отличие профиля от набора информативных терминов. Вычисление весов в профиле проводится только по документам одного класса, а не путем усреднения значений весов по всей выборке, как это происходит в процедурах выявления информативных терминов (наряду с усреднением в литературе предложены и другие варианты расчета веса термина). Таким образом, профиль представляет собой «терминологический портрет» класса, полученный с помощью одного из критериев (1.1) - (1.14). Это позволяет каждому профилю лаконично и корректно описывать документы своего класса. Основной выигрыш от выявления информативных терминов заключается в сокращении размерности без потерь в точности (в ряде случаев точность заметно возрастает).

В профильных методах отнесение нового документа $\vec{X}_{N+1}$, подлежащего классификации и описываемого частотой появления терминов (tf_i), проводится на основе расчета весов классов по формуле:

$$W_k = \sum_{i=1}^{Mk} tf_i \cdot \Pr of(x^{(i)}, Q_k), \qquad (1.15)$$

Здесь W_k - вес *k*-го класса, tf_i – частота встречаемости *i*–го термина в классифицируемом документе $\vec{X}_{N+1}$; M_k – количество наиболее информативных терминов, включенных в профиль *k*–го класса (в наших исследованиях все классы имеют профили одинакового размера $L=M_k$); $\Pr of(x^{(i)}, Q_k)$ - означает вес *i*-го термина в профиле, вычисленном по обучающей выборке с помощью одной из формул (1.1) - (1.14).

Классифицируемый документ $\vec{X}_{N+1}$ относится к тому классу, которому соответствует наибольшая сумма весов:

$$W_k = \max \quad (\text{для } \forall k, k = 1,..., K). \qquad (1.16)$$

Т.е. $\vec{X}_{N+1} \in Q_k$, если в $\vec{X}_{N+1}$ наиболее часто встречаются термины, которые входят в профиль *k*–го класса.