

УДК 330.42

*С.Ю. Балахонов, Д.С. Дмитриев**

ОСНОВЫ ТЕОРИИ РАСПОЗНАВАНИЯ ОБРАЗОВ ПРИ РАБОТЕ С ПОДСИСТЕМОЙ ЭФФЕКТИВНОСТИ ДЕЯТЕЛЬНОСТИ ПРЕДПРИЯТИЯ

В статье определены основные аспекты математической теории распознавания образов при проектировании, разработке и интеграции в информационную структуру предприятия подсистемы обработки показателей эффективности деятельности предприятия.

Ключевые слова: распознавание образов, информационная метрика, показатели деятельности предприятия, эффективность деятельности предприятия, информационная система.

В настоящее время коммерческие и государственные предприятия различной сферы деятельности характеризуются высоким уровнем интеграции информационных технологий и высоким уровнем развития информационных систем: бизнес-процессы различных направлений сегодня являются в значительной мере автоматизированными, в соответствии с чем и инфраструктура предприятий выходит на более высокий технологический этап развития, активно внедряются принципы процессного подхода [1], выделяются шаги построения модели оптимизации прибыли [2], уделяется значительное внимание образовательному аспекту в деятельности предприятия [3], в том числе интеграции образовательных программных продуктов [4; 5]. Все более сложный математический аппарат находит свое применение в прикладных бизнес-системах, задачей которых становится повышение эффективности функционирования организаций. Одним из наиболее активно развивающихся математических направлений, реализуемых в автоматизированных информационных системах, является теория распознавания образов – специализированный раздел кибернетики, в котором изучается развитие основ и методов теоретической классификации, идентификации различных явлений, предметов и т. д., характеризующихся конечным набором признаков и свойств.

Распознавание образов определяется еще и как отнесение исходного набора данных к определенному классу на основе методики выделения существенных признаков, характеризующих выбранные данные из общей структуры. Распознаванием образов на современном этапе начинают овладевать системы с элементами искусственного интеллекта. Методы распознавания образов играют ведущую роль в системах распознавания речи. Поэтому в прикладных системах, созданных для понимания текста, все больше применяются подобные методы распознавания образов вместо традиционных способов распознавания речи.

Создание информационных систем с элементами распознавания образов остается нетривиальной задачей, необходимость решения которой возникает в различных областях. Однако математические аспекты теории распознавания образов имеют

* © Балахонов С.Ю., Дмитриев Д.С., 2015

Балахонов Станислав Юрьевич (stblhmv@gmail.com), магистрант кафедры математики и бизнес-информатики, *Дмитриев Денис Сергеевич* (denisdmitriev000@gmail.com), кафедра математики и бизнес-информатики, Самарский государственный университет, 443011, Российская Федерация, г. Самара, ул. Акад. Павлова, 1.

недостаточное обоснование и практическое применение при построении прикладных информационных систем. В данной статье рассмотрены основы алгебраической теории информации в задаче распознавания образов, предложенные В.Д. Гоппой [6], при реализации доступа в подсистему анализа эффективности деятельности предприятия.

1. Теоретические основы алгебраической теории информации в задаче распознавания образов.

Первое упоминание необходимости применения распознавания образов – это задача моделирования операций, совершаемых живыми организмами, которая была исследована Ф. Розенблатом в процессе взаимодействия живого организма с внешней средой. Ф. Розенблат сконструировал «Перцептрон» – техническое устройство, способное производить автоматическую классификацию объектов, что позволило решить одноименную задачу классификации.

Каждый выбранный для изучения объект может быть охарактеризован определенным набором признаков, обозначенных X_1, X_2, \dots, X_n . Все признаки можно классифицировать по двум типам: во-первых, когда $X = \{1, 2, \dots, m\}$, и, во-вторых, количественным, когда X является любым действительным числом. В случае произвольного действительного числа данный признак с помощью квантования может быть преобразован в количественный признак. Для этого его можно представить в виде слова – конечной последовательности целых чисел.

Основной проблемой задачи классификации является выявление признаков, на основе которых будет происходить распознавание. Задачи, для которых удастся удачно выделить этот набор признаков, решаются с помощью описываемого ниже алгоритма. Вместе с тем все объекты, выбранные в качестве контрольных, обязательно должны принадлежать к определенному классу из множества $Y = \{1, 2, \dots, k\}$.

Рассматриваемый признак X – конечная последовательность (слово) конечного алфавита, а также столбец информационной матрицы. Целью является определение по этому признаку значения основного признака Y , т. е. определение принадлежности объекта классу. Таким образом, информативность признака может определяться количеством информации, которое присутствует в данном признаке относительно значения основного признака:

$$I_0(X:Y) = I_0(Y) - I_0(Y|X) = I_0(X) - I_0(X|Y), \quad (1)$$

где I_0 – информационная метрика.

Например, если $Y = (1,1,1,2,2,2,3,3,3)$, то $I_0(X:Y) = 1,58$ бита.

Затем для того, чтобы определить информативность, конструируется матрица перехода, которая связывает буквы слова X в буквы слова Y . Структура представлена в таблице.

Выберем первый параметр настройки, обозначенный через α , с помощью которого возможно определить минимально допустимый уровень информативности признака, т. е. не учитывать признаки, информативность которых меньше α .

Сходные признаки, повторяющие друг друга, объединяются в кластеры. Зависимость признаков изучается с помощью метрики (2).

$$\rho(X_i, X_j) = \frac{1}{2} (I_0(X_j, X_i) + I_0(X_i, X_j)). \quad (2)$$

Таблица

Пример матрицы перехода букв слова X в буквы слова Y

	1	2	3	
1	0	1	0	1
2	3	2	3	8
	3	3	3	9

Если расстояние $\rho(X_i, X_j)$ мало, то X_i и X_j зависимы (повторяют друг друга). Для независимых признаков $\rho(X_i, X_j)$ приближается к максимуму. Метрика $\rho(X_i, X_j)$ показывает «расстояние» признаков X_i и X_j от взаимной эквивалентности (насколько различны признаки X_i и X_j). В предельном случае, когда $\rho(X_i, X_j)$, существует функциональная зависимость в обе стороны:

$$X_i \rightarrow X_j, X_j \rightarrow X_i. \quad (3)$$

На основе вычисления метрики $\rho(X_i, X_j)$ и ведется формирование кластеров из признаков, т. е. способность проводить классификацию по кластеру на основе распознавания значения признака Y из признаков, сходных с признаком X . Если выбрать один признак X_i , то неопределенность будет рассмотрена лишь частично, следовательно, при раскрытии полного спектра необходимо выбирать сложные признаки (наборы).

Выберем следующий параметр настройки β , который будет характеризовать минимально допустимую информативность набора признаков. Величина β обычно колеблется от 60 % $I_0(Y)$ до $I_0(Y)$. Сначала находятся все двойные признаки, удовлетворяющие этому условию. При выбранном значении β производится поиск всех сложных признаков, информативность которых превышает β . Затем составляется матрица перехода букв слова в буквы признака Y для каждого из отобранных сложных признаков.

Введем заключительный параметр настройки, обозначаемый γ . Данный параметр настройки принимает только целочисленные значения $1, 2, \dots$. Параметр работает следующим образом: пусть у некоторого объекта в «старом» пространстве признаки X_2, X_4, X_8 принимают значения 0, 1, 1 соответственно. Находим в соответствующей переходной матрице сочетание (011) и соответствующую строку (C_0, C_1, C_2) (в данном случае $(C_0 = 0, C_1 = 1, C_2 = 0)$). Находим максимальное значение, предположим, C_1 . Если окажется, например, что $C_1 - C_0 \geq \gamma, C_1 - C_2 \geq \gamma$, то соответствующий признак в «новом» пространстве принимает значение, равное 1. Если же в

строке (C_0, C_1, C_2) нет такого «лидера», то записываем пробел (отказ от распознавания по данному признаку для данного объекта). Таким образом, параметр γ определяет «степень доверия» к данному сложному признаку. Число отобранных сложных признаков определяет размер нового пространства.

Если в пространстве признаков основной является информационная метрика $\rho(X_i, X_j)$, то кластеризация объектов нового пространства производится по метрике Хэмминга $d(k, m)$. Степень кластеризации определяется параметром настройки β . Параметр настройки γ определяет размер нового пространства признаков.

Настройку программы осуществляют изменением параметра γ , затем β и потом α . При работе необходимо выбирать как можно большие значения γ , но при этом число признаков в новом пространстве может сократиться. Это можно компенсировать уменьшением параметра β , но тогда образы начинают «расплываться», т. е. ухудшается кластеризация по Хэммингу. Первый параметр настройки α не следует выбирать слишком малым. При малых значениях параметра α может произойти искажение результатов распознавания. Если и в этом случае настройка программы не получается, необходимо произвести изменения в обучающей выборке: объекты, которые использовались ранее для контроля, применить для обучения и наоборот.

2. Реализация доступа в подсистему анализа эффективности деятельности предприятия.

Принцип доступа в подсистему обработки показателей эффективности работы организации ЗАО «Самарская лука» был разработан на основе логики распознавания образов по описанной ранее теории с помощью программного приложения на языке программирования Python. Была поставлена задача распознавания изображений цифр от 0 до 9, которые вводятся в виджет браузера при входе через веб-интерфейс в информационную систему предприятия. В ходе решения была использована информационная метрика, предложенная В.Д. Гоппой [6]. Каждая картинка была приведена к размеру 100 x 150 пикселей и поделена на 150 равных прямоугольников. В нашем случае все признаки являются качественными и принимают значения из множества $\{0, 1\}$ в зависимости от цвета прямоугольника (белый или черный). В поставленной задаче количество классов равно 10 (по количеству цифр 0–9), каждому из которых принадлежит по 8 изображений цифр. При разработке приложения 80 объектов помещаются в обучающую выборку и 40 – для контроля.

На предварительном этапе работы были созданы функции, реализующие действие перестановок на множестве слов фиксированной длины, функции, которые вычисляют орбиты и стабилизаторы для слов произвольной длины. Также были реализованы функции, определяющие работу с изображениями. Функция разбиения изображения на произвольное количество прямоугольников позволяет достичь результата очень высокой точности. Нами сочтен оптимальным вариант из 150 признаков. Функция перевода полученных прямоугольников в двоичное слово заданной длины: в ходе написания этой функции был введен первый дополнительный параметр настройки, решающий, какое значение примет признак. Он принимает значение в процентах и показывает долю заполнения прямоугольника черным цветом, чтобы целиком считаться черным. Были реализованы функции для подсчета информативности признаков и введения метрики. Функции, которые вычисляют ноль-информацию, условную и взаимную информации, также работают со словами произвольной длины. И наконец, была построена информационная матрица для базы картинок. Матрица записывается в файл.