ИССЛЕДОВАНИЕ ПРИЗНАКОВ-ОТНОШЕНИЙ В ЗАДАЧАХ КЛАССИФИКАЦИИ И РАСПОЗНАВАНИЯ

К. П. Коршунова
Кафедра вычислительной техники, Филиал ФГБОУ ВО "НИУ "МЭИ"в г. Смоленске
Смоленск, Российская Федерация
E-mail: ksenya-kor@mail.ru

В задачах распознавания (классификации) образов выбор и выделение признаков играют центральную роль. В большинстве практических задач необходим компромисс между доступностью набора признаков и качеством классификации. В статье вводится понятие признаков-отношений (бинарных и тернарных), а также рассмотрен подход к решению классификационных задач на их основе.

Введение

Рассматриваемый подход был описан и применен к задаче медицинской диагностики (диагностика рака молочной железы по результатам лабораторного исследования [1-2]), а также к задаче распознавания поэтических текстов (классификация поэтических текстов А.С. Пушкина и М.Ю. Лермонтова по периодам творчества: ранний и поздний период – на основе числовых признаков, характеризующих текст).

I. Признаки-отношения

Пусть объекты характеризуются набором признаков классификации: $\{S_1, S_2, S_3, S_4, S_5\}$. Рассмотрим для наглядности задачу бинарной классификации – с двумя классами $\{C_1, C_2\}$. Для распознавания класса полезно проанализировать n-местные отношения между признаками. В данной статье ограничимся бинарными (между элементами двух множеств) и тернарными (трех множеств) отношениями, характеризующими рассматриваемые классы, т.е. отношениями эквивалентности на множествах $\{S_1, S_2, S_3, S_4, S_5\}$. Отношения между элементами множеств представляют собой подмножества декартова произведения данных множеств [3]. Рассмотрим координатные области – декартово произведение множеств двух признаков (для анализа бинарных отношений) и трех признаков (для анализа тернарных отношений). Нанесем точки, соответствующие каждому объекту обучающей выборки, и присвоим им разные метки: o – класс C_1, x – класс C_2 (см.рисунки 1-2).

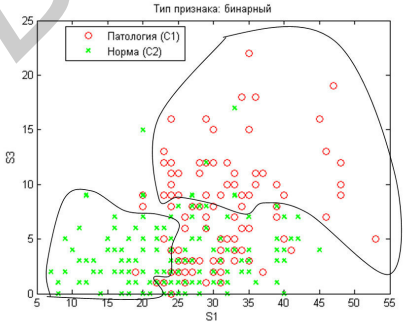

Рис. 1 – Пример бинарного признака-отношения

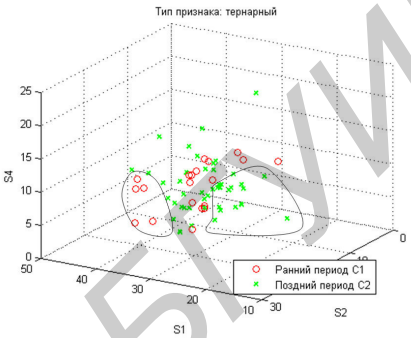

Рис. 2 – Пример тернарного признака-отношения

В ряде случаев в рассматриваемых областях можно выделить области сгущения однородных точек (объектов одного класса) и решить задачу распознавания, используя лишь подмножество признаков, не прибегая к автоматическим методам анализа данных. Подобные "срезы"признакового пространства назовем *признаками-отношениями*: бинарными в случае двумерных "срезов"(рис.1) и тернарными в случае трехмерных "срезов"(рис.2). Как видно из рисунков, изучение признаков-отношений позволяет получить наглядное представление о внутренней структуре задачи распознавания, а также сделать предположение о принципиальной возможности/невозможности ее качественного решения: хаотично разбросанные и сильно пересекающиеся множества точек (см.рис.3) свидетельствуют о неразделимости классов и/или о малой информативности исходных признаков.

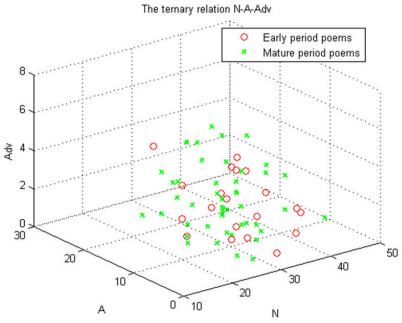

Рис. 3 – Пример тернарного признака-отношения

Визуальный анализ признаков-отношений для рассматриваемых задач показывает, что классы C_1 и C_2 являются пересекающимися и плохо разделимыми по крайней мере с учетом имеющихся признаков. Однако для небольшого подмножества образов («типичных» представителей класса) возможно решить задачу распознавания с достаточной достоверностью (см. выделенные области на рисунках), причем используя не полный набор, а лишь подмножество исходных признаков (например, отнести произведение к позднему периоду творчества по значениям признаков V и N, рис.4).

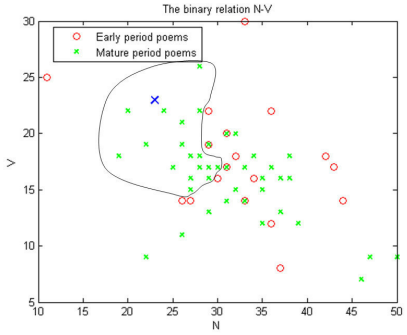

Рис. 4 – Пример решения задачи распознавания

II. Результаты решения задач распознавания

При построении автоматической классификации будем учитывать признаки-отношения – метки областей на рассматриваемых срезах в двумерном и трехмерном пространствах, в которую попадает изучаемый объект. Мы решаем задачу классификации исходя из принципа голосования: объект относится к тому классу, для которого количества попаданий значений признаков данного объекта в области соответствующего класса (метки класса для признаков-отношений и статистические градации для исходных признаков) больше количества попаданий в области другого класса.

В качестве показателя качества решения задачи воспользуемся понятием точности (*Precision* [4]). Приведем результаты решения рассматриваемых задач: классификации поэтических текстов по периодам творчества автора (класс C_1 – ранний период, класс C_2 – поздний период творчества) и медицинской диагностики (класс C_1 – патология, класс C_2 – норма).

Таблица 1 – Результаты оценки качества решения задачи распознавания

Задача	Точность, % в классе C_1	Точность, % в классе C_2
Классификация поэтических текстов	70%	57%
Медицинская диагностика	91,3%	86,7%

Из таблицы видно, что решение задачи классификации поэтических текстов сильно смещено в сторону одного из классов. Задача медицинской диагностики, напротив, решена более качественно. Подобные результаты вполне ожидаемы, исходя из предварительного анализа признаков-отношений (см.рис. 1-2).

III. Выводы

Полученные результаты показали, что применение рассматриваемого подхода на основе анализа отношений между множествами признаков классифицируемых объектов к задаче распознавания позволяет получить новый материал для извлечения знаний об изучаемых объектах и является интересным предметом для дальнейших исследований.

Исследование признаков-отношений в задачах распознавания образов позволяет:

1. Сделать выводы о качестве и информативности имеющихся признаков и предположение о принципиальной возможности/невозможности качественного решения задачи: при рассмотрении "срезов"признакового пространства в некоторых случаях можно легко визуально выделить области, принадлежащие одному из классов, – имеющаяся информация обеспечивает успешное распознавание классов; если же оба класса практически полностью пересекаются, это говорит о невозможности качественного распознавания по крайней мере на основе значений данных признаков;
2. Получить наглядное представление о внутренней структуре задачи: являются ли классы разделимыми, пересекающимися, эквивалетными, входит ли один класс в состав другого и т.д. – и тем самым осуществить точную постановку решаемой задачи;
3. В ряде случаев решить задачу распознавания, используя лишь подмножество доступных признаков, не прибегая к автоматическим методам анализа данных (см. рис.4).

1. Абросимов С.Ю. Проверка гипотезы о возможности идентификации стромы биологических тканей в норме, при предопухолевых и опухолевых процессах, научный отчет по проведенному научному исследованию, Смоленск, 2006.
2. Коршунова К.П. Нейросетевой способ классификации сложных объектов на основе признаков-отношений // Нейрокомпьютеры: разработка, применение, издательство Радиотехника, 2016. №6.
3. Шрейдер Ю.А. Равенство. Сходство. Порядок. Москва: Наука, 1971. -– 256 с.
4. Manning C., Raghavan P., Schutze H. An Introduction to Information Retrieval. Cambridge: Cambridge UniversityPress, 2009.