СПОСОБ РЕШЕНИЯ ЗАДАЧИ КЛАССИФИКАЦИИ СЛОЖНЫХ ОБЪЕКТОВ НА ОСНОВЕ УЧЕТА ОТНОШЕНИЙ МЕЖДУ ВЕРОЯТНОСТНЫМИ ПРИЗНАКАМИ

К.П. Коршунова
Кафедра вычислительной техники,
филиал ФГБОУВПО «Национальный исследовательский университет «МЭИ»
Смоленск, Российская Федерация
E-mail: ksenya-kor@mail.ru

Предложен способ решения задачи классификации сложных объектов, позволяющий увеличить объем используемой информации за счет учета отношений между вероятностными признаками. Рассмотрен пример, демонстрирующий применение способа. Произведена оценка информативности.

Известные подходы к решению задач классификации предполагают рассмотрение отдельных признаков объекта и установление класса по их конкретным значениям. Предлагаемый подход основан на рассмотрении объекта как сложной системы, состоящей в свою очередь из подсистем (признаков) и всевозможных связей между ними. Поэтому для решения задач анализа данных недостаточно учитывать отдельные признаки, а требуется также учет взаимосвязей (отношений) между ними [1]. Отметим, что речь идет не только о числовых зависимостях, но об отношениях между множествами (признаками объекта) общего вида. Это обосновано в том числе тем, что признаки могут быть измерены в разных шкалах и применение известных методов установления зависимостей в этом случае затруднено. Обнаружение отношений общего вида (в отличие от тех же числовых зависимостей) и оценка их значимости (с точки зрения рассматриваемой классификации) представляет основную сложность. После этого отношения могут быть рассмотрены как дополнительные признаки, тогда для построения классификации могут быть применены известные приемы и технологии решения задач анализа данных.

Постановка задачи приведена в [2]. Дано множество всех возможных объектов распознавания A. Каждый элемент этого множества (каждый объект) характеризуется набором признаков распознавания. Множество признаков распознавания P: $P = \{S_1, S_2, \ldots, S_n\}$. Во множестве объектов A выделено некоторое подмножество, для элементов которого распределение по классам известно, – обучающая выборка.

Для простоты рассмотрим бинарные отношения. Бинарное отношение (подмножество декартова произведения двух множеств) может быть задано простым перечислением пар, в него входящих. Будем рассматривать декартово произведение $\langle S_i, S_j \rangle, i \neq j$ в координатной плоскости, по оси абсцисс которой – значения признака S_i, а по оси ординат – S_j. Нанесем точки, соответствующие каждому объекту обучающей выборки. Точки, относящиеся к разным классам, обозначим разными метками (например, цветовыми). Проанализируем полученную картину (см. рис. 1). В ряде случаев возможно выделить в полученной координатной плоскости кластеры – области сгущения точек одного типа (цвета). Области пересечения кластеров разных типов будем считать областями неопределенности. Выделенные кластеры и представляют собой подмножества $\langle S_i, S_j \rangle$ – бинарные отношения. Таким образом, подзадачей классификации становится специфическая задача кластеризации, метод решения которой может быть различным.

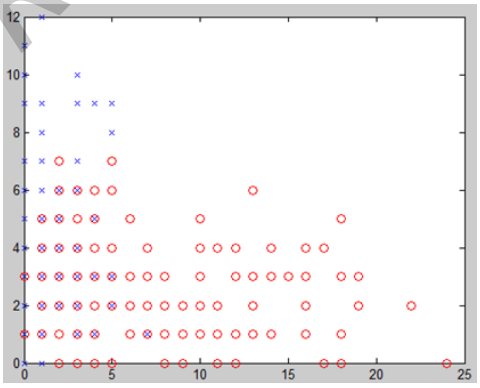

Рис. 1 - Пример координатной плоскости

Номер кластера в дальнейшем будем рассматривать как еще один признак классификации S_{ij}. Рассматривая весь набор признаков попарно, получим дополнительно n' признаков (в дальнейшем будем называть их интегрированными признаками), где n' равно числу сочетаний из n по 2.

Рассмотрим пример реальной задачи медицинской диагностики: идентификация рака молочной железы [3]. Множество распознаваемых классов состоит из 2 элементов: $C = \{C_1, C_2\}$, где C_1 – наличие заболевания, C_2 – отсутствие заболевания. Для диагностики заболевания исследовались образцы биологических тканей пациентов и производился подсчет клеток пяти типов (фибропласты Фб, фиброциты Фц, лимфоциты Лф, макрофаги Мф, плазмоциты Пл). Таким

образом, каждый объект характеризуется пятью признаками, каждый из которых выражен натуральным числом. В нашем распоряжении имеется обучающая выборка, состоящая из 340 элементов.

Рассмотрим координатную плоскость $\langle S_1, S_3 \rangle$ (декартово произведение множеств Фб и Лф). Каждой точке присвоена цифровая метка: 1 – класс C_1, 3 – класс C_2, меткой 2 выделены точки, в которые попали объекты обучающей выборки как из класса C_1, так из класса C_2 (см. рис. 2).

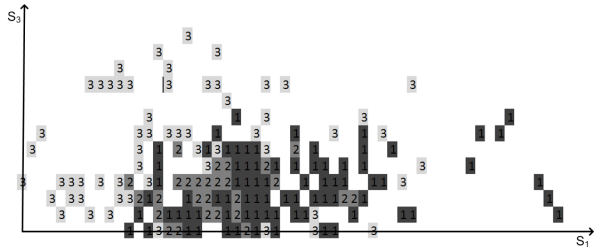

Рис. 2 - Координатная плоскость $\langle S_1, S_3 \rangle$

После применения простого алгоритма обработки изображения получили разбивку всей координатной плоскости (диапазона изменения признаков) на следующие области (границы кластеров расширены таким образом, чтобы учитывать точки, не попавшие в обучающую выборку) (см. рис. 3).

Рис. 3 - Разбиение координатной плоскости $\langle S_1, S_3 \rangle$ на области

Таким образом, мы получаем $n' = 10$ новых интегрированных признаков, каждый из которых имеет три градации: 1, 2, 3.

Оценим информативность исходных и новых признаков, используя меру информативности Кульбака [4] (см.табл. 1, 2, рис. 4).

Таблица 1 - Информативность исходных признаков

Признак	Информативность
Фб	1,256252
Фц	0,446849
Лф	1,951345
Мф	6,298487
Пл	0

Таблица 2 - Информативность интегрированных признаков

	Фб	Фц	Лф	Мф	Пл
Фб	-	1,05	3,93	6,44	0,67
Фц	-	-	1,34	2,66	0,56
Лф	-	-	-	2,95	2,09
Мф	-	-	-	-	3,55
Пл	-	-	-	-	-

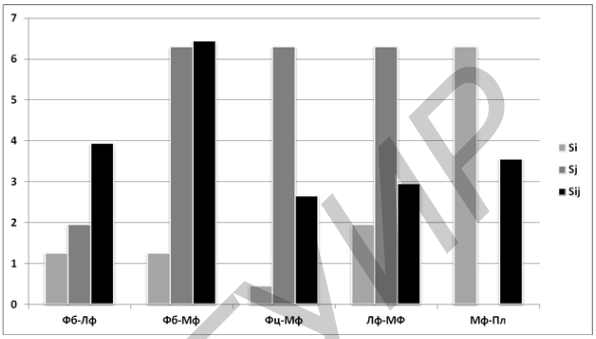

Рис. 4 - Гистограмма информативности исходных и интегрированных признаков для нескольких случаев

Как видно из таблиц и гистограммы, информативность интегрированных признаков сопоставима с информативностью исходных, а в ряде случаев превышает ее. Кроме того, активно «участвует» в отношениях и признак Пл с нулевой информативностью (см.рис.4., декартово произведение Мф-Пл). Можно сделать вывод о том, что предложенный способ позволяет получить дополнительно несколько информативных признаков, которые при решении задачи могут быть рассмотрены наравне с исходными. Таким образом, введя в последовательность решения дополнительный этап (подзадача кластеризации), появляется возможность существенно увеличить объем используемой информации, что не может не отразиться на качестве решения.

1. Лямец Л. Л. Подход к формальному описанию объектов в задачах распознавания на основе принципа системности / Математическая морфология. Электронный математический и медико-биологический журнал. - Т. 13. - Вып. 2. - 2014.
2. Коршунова К.П., Лямец Л.Л. Постановка задачи классификации объектов по множествам вероятностных признаков и отношениям между ними / Сборник трудов XI международной научно-технической конференции студентов и аспирантов «Информационные технологии, энергетика и экономика». В 3 т. Т 1. – Смоленск: "Универсум 2014. - С.171-173.
3. Абросимов С.Ю. Проверка гипотезы о возможности идентификации стромы биологических тканей в норме, при предопухолевых и опухолевых процессах. [Текст]: научный отчет по проведенному научному исследованию / д.м.н. Абросимов Сергей Юрьевич. – Смоленск, 2006. – 45 с.
4. Кульбак С. Теория информации и статистика / С. Кульбак. - М.: Наука, 1967. - 408 с.