КЛАССИФИКАЦИЯ ОБЪЕКТОВ В ПРОСТРАНСТВЕ ДВОИЧНЫХ ПРИЗНАКОВ

В. В. Никифоров

Санкт-Петербургский институт информатики и автоматизации РАН
199178, Санкт-Петербург, 14-я линия В.О., д.39
<nik@iias.spb.su>

УДК 681.3
Никифоров В. В. **Классификация объектов в пространстве двоичных признаков** // Труды СПИИРАН, Вып. 3, т. 2. — СПб: СПИИРАН, 2006.

Аннотация. *Рассматриваются сравнительные возможности различных групп алгоритмов классификации объектов. В частности, рассматривается группа спектральных алгоритмов, оперирующих данными о спектрах расстояний между объектами. Рассматриваются также более широкие в общем случае группы разностных алгоритмов, оперирующих данными о шкалах различий между объектами, и еще более широкой группы алгоритмов, инвариантных к смене кодировки признаков. Показано, что для множеств в пространстве двоичных признаков возможности группы спектральных алгоритмов совпадают с возможностями разностных алгоритмов, а для множеств с нечетным числом объектов — и с возможностями алгоритмов, инвариантных к смене кодировки признаков.* — Библ. 5 назв.

UDC 681.3
Nikiforov V. **Object Classifying of objects in the space of binary features** // SPIIRAS Proceedings. Issue 3, vol. 2. — SPb: SPIIRAS, 2006.

Abstract. *The comparative potential possibilities are regarded of various groups of algorithms for object classification. Particularly, the spectrum group of algorithms, that operate spectrum of distances between objects, is regarded. The group of difference algorithms, which operate with scales of differences between objects, and the group of coding invariant algorithms that are invariant to feature coding, are also regarded. It is shown, that for the sets of objects in the space of binary features the possibilities of spectrum group of algorithms are equal to the possibilities of the group of difference algorithms and more over, for the set of odd number of objects are also equal to the possibilities of the group of coding invariant algorithms.* — Bibl. 5 items.

1. Введение

Методы и алгоритмы классификации, распознавания, предсказания на основе обработки описаний множества объектов, представляемых наборами дискретных значений признаков используются при разработке распространенных инструментов информационных технологий в различных сферах науки и производства [1]. К ряду задач, возникающих в этой области, относится поиск алгоритмов, обеспечивающих эффективное (в отношении того или иного заданного критерия) разбиение анализируемого множества на составляющие подмножества объектов, а также отыскание таких поднаборов признаков, которые обеспечивают наилучшие результаты для решения задач диагностики, классификации, предсказания [2].

В ходе поиска нужных алгоритмов исследователи и разработчики прибегают к использованию тех или иных эвристических принципов [3]. С этим непосредственно связан вопрос о потенциальных возможностях той группы $\mathfrak{R}$ алгоритмов, в рамках которой в соответствии с принятыми эвристическими принципами ищется решение. Аналогичный вопрос возникает в случае применения для этих целей методов генетического программирования [4]. Так, если из-

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

вестно, что выбор способа кодировки некоторого признака не существенен (например, содержательное значение имеет лишь факт совпадения или различия значений признака), то нет смысла искать эффективные решения в излишне широкой группе $\Re_{\text{codfelt}}$ алгоритмов, учитывающих изменения кодировки признаков в описаниях объектов анализируемого множества. То есть, в этом случае требуемое решение следует искать в более узкой группе $\Re_{\text{codinv}}$ алгоритмов, инвариантных к смене кодировки значений признаков.

При поиске алгоритмов, которые предназначены для определения особенностей конфигурации анализируемых множеств объектов, возникает вопрос о перспективности поиска эвристических решений в рамках группы $\Re_{\text{dif}}$ таких алгоритмов, которые оперируют лишь данными о попарных различиях в значениях признаков по всевозможным парам объектов анализируемого множества (разностные алгоритмы [5]). Аналогичный вопрос возникает в отношении группы $\Re_{\text{dispec}}$ алгоритмов, оперирующих лишь данными о спектрах расстояний между объектами по поднаборам признаков (спектральные алгоритмы [5]).

Очевидно, что в отношении выделения особенностей конфигурации анализируемых множеств возможности группы $\Re_{\text{dispec}}$ спектральных алгоритмов, по крайней мере, не шире, чем возможности группы $\Re_{\text{dif}}$ разностных алгоритмов. Здесь сравнительная широта возможностей понимается в том смысле, что все особенности конфигурации, которые могут быть выявлены в рамках алгоритмов группы $\Re_{\text{dispec}}$, могут быть выявлены и алгоритмами группы $\Re_{\text{dif}}$.

Действительно, средствами алгоритмов группы $\Re_{\text{dif}}$ могут быть вычислены все данные о спектрах расстояний в анализируемом множестве объектов — то есть, все те данные, которые являются исходными для алгоритмов группы $\Re_{\text{dispec}}$.

Будем считать, что обозначение $\Re_1 \supseteq \Re_2$ отражает сравнительные возможности алгоритмов групп $\Re_1$ и $\Re_2$ в отмеченном выше смысле: возможности группы $\Re_2$ не шире, чем возможности группы $\Re_1$. Если же известно, что возможности алгоритмов группы $\Re_1$ строго шире, будем использовать обозначение $\Re_1 \supset \Re_2$. В случае равных возможностей алгоритмов групп $\Re_1$ и $\Re_2$ в отношении выделения особенностей конфигурации анализируемых множеств будем использовать обозначение $\Re_1 \supset \Re_2$. В принятых обозначениях $\Re_{\text{dif}} \supseteq \Re_{\text{dispec}}$. Поскольку входные данные для разностных алгоритмов не изменяются при смене кодировок значений признаков, имеет место и отношение.

В работе [5] отмечены, в частности, следующие два факта.

1) В пространстве двоичных признаков любой из разностных алгоритмов $\Re_{\text{dif}}$ может быть смоделирован в рамках группы спектральных алгоритмов $\Re_{\text{dispec}}$ в том смысле, что для любого алгоритма из $\Re_{\text{dif}}$ найдется эквивалентный (в отношении получаемых результатов о структурных свойствах анализируемого множества) алгоритм из $\Re_{\text{dispec}}$. Такое утверждение означает, что $\Re_{\text{dif}} \subseteq \Re_{\text{dispec}}$. Но поскольку по определению $\Re_{\text{dif}} \supseteq \Re_{\text{dispec}}$, получаем

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

$\Re_{dif} = \Re_{dispec}$. Следовательно, для множеств объектов, представленных в пространстве двоичных признаков, потенциальные возможности поиска эффективных алгоритмов классификации в группе $\Re_{dispec}$ столь же широки, как и в группе $\Re_{dif}$.

2) Для множеств с нечетным числом объектов, представленных в пространстве двоичных признаков, для любого алгоритма группы $\Re_{codinv}$ (любого алгоритма, инвариантного к смене кодировки значений признаков) найдется эквивалентный (в отношении получаемых результатов) алгоритм из группы $\Re_{dif}$ разностных алгоритмов. Следовательно, для таких множеств потенциальные возможности поиска эффективных алгоритмов классификации в группе $\Re_{dif}$ столь же широки, как и в группе $\Re_{codinv}$.

В настоящей работе приведены формальные доказательства этих фактов. Доказательства проведены на конструктивной основе — показано, как строить алгоритмы группы $\Re_{dispec}$, эквивалентные алгоритмам группы $\Re_{dif}$ и как для множеств с нечетным числом элементов строить алгоритмы группы $\Re_{dif}$, эквивалентные алгоритмам группы $\Re_{codinv}$.

2. Множества объектов в пространстве двоичных признаков

Рациональный выбор способов представления данных об объектах (способов формальных описаний объектов) является важным шагом в разработке подходов к анализу структурных свойств множеств объектов. Одним из способов составления формальных описаний является представление свойств объектов упорядоченными наборами значений двоичных признаков. В ряду алгоритмов обработки информации о множествах объектов, представленных наборами значений признаков, выше были выделены следующие группы алгоритмов:

- группа $\Re_{codfelt}$ — алгоритмы, учитывающие способ кодировки признаков;
- группа $\Re_{codinv}$ — алгоритмы, инвариантные к способу кодировки признаков;
- группа $\Re_{dif}$ — разностные алгоритмы (инвариантные к способу кодировки признаков и порядку следования описаний объектов);
- группа $\Re_{dispec}$ — спектральные алгоритмы.

Ниже рассматривается также группа $\Re_{iden}$ алгоритмов, основанных на учете числа пар объектов с идентичными частичными описаниями (т.е. описаниями по поднаборам исходного набора признаков). Группа алгоритмов $\Re_{iden}$ является подмножеством группы $\Re_{dispec}$ спектральных алгоритмов. Если алгоритмы группы $\Re_{dispec}$ оперируют данными о спектрах расстояний между объектами по поднаборам признаков, то алгоритмы группы $\Re_{iden}$ используют по каждому поднабору признаков лишь один из параметров спектра расстояний — число пар объектов, неразличимых (имеющих

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

идентичные описания) по данному поднабору признаков. Согласно такому определению имеет место соотношение $\Re_{\text{dispec}} \supseteq \Re_{\text{iden}}$. Ниже будет показано, что для множеств в пространстве двоичных признаков, для любого спектрального алгоритма, найдется эквивалентный (в отношении получаемых результатов) алгоритм из группы $\Re_{\text{iden}}$. То есть, для множеств в пространстве двоичных признаков $\Re_{\text{dispec}} = \Re_{\text{iden}}$.

Таблица шкал значений. Пусть для представления свойств объектов используются векторы (шкалы значений) признаков: каждому, объекту соответствует вектор (шкала значений) некоторая конкретная вершина n-мерного единичного куба. Значения компонент вектора признаков конкретного объекта зависят не только от свойств самого объекта, но и от выбранного способа представления. Например, вес вектора двоичных признаков конкретного объекта (число ненулевых компонент) определяется двумя факторами: особенностями представляемого объекта и определением базиса формального описания объектов, то есть, определением способа кодировки значений признаков. Описание множества объектов задается таблицей шкал значений (ТШЗ): строки ТШЗ соответствуют объектам, столбцы — признакам.

Для объектов, представляемых значениями n двоичных признаков существует 2^n базисов описаний, различающихся способами кодировки компонент описания (способами кодировки значений признаков). Описания, представленные в одном базисе, преобразуются в описания, представленные в другом базисе инверсией значений одного или нескольких признаков. Для конкретного множества объектов в ряду базисов описаний выделяются минимизирующие базисы, обеспечивающие минимизацию суммарного веса векторов представляемого множества объектов. Очевидно, что для множества с нечетным числом объектов существует единственный минимизирующий базис.

Пусть в качестве описаний актуальных свойств объектов x из некоторого множества объектов X используются наборы значений двоичных признаков $p_i(x)$, $1 < i \le n$; каждый из признаков p_i принимает для конкретного объекта $x \in X$ одно из двух возможных значений: 0 или 1. Возможные комбинации значений признаков $p_i(x)$ образуют пространство описаний объектов: каждый объект x соответственно значениям $p_i(x)$ проецируется в определенную точку пространства описаний (в определенную вершину n-мерного единичного куба). Признаки $(p_1, p_2, ..., p_n)$ образуют базис пространства описаний. Вектор $[p_1(x), p_2(x), ..., p_n(x)]$ будем называть шкалой значений признаков объекта, x и обозначать символом $[x]$.

В табл. 1 приведен пример конфигурации множества описаний шести объектов a, b, c, d, e, f в пространстве описаний (базисе), образуемом парой двоичных признаков (p_1, p_2). В первом из четырех столбцов табл. 1 описания объектов множества $A = \{a, b, c, d, e, f\}$ представлены в виде двоичных шкал $[a]$, … , $[f]$ значений признаков p_1 и p_2. Три из шести объектов (объекты b, c, d) неразличимы в базисе (p_1, p_2): они проецируются в одну и ту же точку пространства описаний, соответствующие им шкалы значений признаков совпадают по всем компонентам, $[b] = [c] = [d]$.

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Для множества A минимизирующим является базис (p_1^*, p_2^*), связанный с исходным базисом (p_1, p_2) соотношениями $p_1^*(x) = \neg p_1(x)$ и $p_2^*(x) = \neg p_2(x)$ — оба признака инвертированы относительно исходного базиса. В последнем столбце табл. 1 приведены шкалы значений признаков для объектов множества A в минимизирующем базисе (p_1^*, p_2^*). Очевидно, что базисы (p_1, p_2) и (p_1^*, p_2^*) информационно эквивалентны, соответствующие комплекты шкал значений признаков несут идентичную информацию, но различаются по форме, поскольку относятся к различным системам отсчета. Суммарный вес шкал $[x]$ множества A в базисе (p_1, p_2) равен восьми, а в базисе (p_1^*, p_2^*) — четырем.

Таблица 1

Значения признаков объектов множества A в различных базисах

Объекты	**Коды признаков**			
	Базис (p_1, p_2)	Базис (p_1^*, p_2)	Базис (p_1, p_2^*)	Базис (p_1^*, p_2^*)
a	[0,1]	[1,1]	[0,0]	[1,0]
b	[1,1]	[0,1]	[1,0]	[0,0]
c	[1,1]	[0,1]	[1,0]	[0,0]
d	[1,1]	[0,1]	[1,0]	[0,0]
e	[0,0]	[1,0]	[0,1]	[1,1]
f	[1,0]	[0,0]	[1,1]	[0,1]

Таким образом, вид представления множества объектов векторами двоичных признаков зависит от способа кодировки признаков (от выбора базиса представления). В тех случаях, когда выбор способа кодировки отдельных признаков играет содержательную роль, использование алгоритмов, инвариантных к выбору базиса представления, может привести к потере важной информации об особенностях анализируемого множества. Следовательно, во избежание подобных потерь информации алгоритмы обработки описаний анализируемого множества следует искать в ряду алгоритмов группы $\Re_{\text{codfelt}}$. Алгоритмы этой группы могут использовать всю информацию, имеющуюся в ТШЗ — как информацию о различиях между объектами, так и информацию о выбранной кодировке значений отдельных признаков.

Таблица шкал различий. В тех случаях, когда выбор способа кодировки является произвольным (не несет содержательной информации) представляется целесообразным использование такого метода представления особенностей анализируемого множества, который бы не зависел от выбора базиса представления. Таким способом отображения структурных особенностей анализируемого множества является составление таблицы шкал различий (ТШР): каждая строка ТШР представляет собой шкалу (вектор) различий значений признаков для конкретной пары объектов анализируемого множества. Разряд конкретной шкалы различий равен единице, если значения соответствующего признака пары объектов совпадают (в противном случае этот разряд равен нулю). В случае использования базиса описаний из n двоичных признаков

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

$(p_1, p_2, ..., p_n)$ каждой паре объектов x и y соответствует шкала различий $(d_1(x,y), ..., d_n(x,y))$ где $d_i(x,y) = 0$ если $p_i(x) = p_i(y)$, в противном случае $d_i(x,y) = 1$.

В соответствии с определением содержимое ТШР не изменяется при изменении базиса (при инверсии значений признаков исходного описания объектов).

Условимся упорядочивать строки ТШР лексикографически: тогда вид ТШР становится инвариантен и по отношению к исходному порядку следования описаний объектов анализируемого множества. Таким образом, ТШР отражает особенности структуры анализируемого множества в виде, отвлеченном и от произвольного выбора базиса представления и от произвольного выбора порядка следования описаний объектов в ТШЗ.

Отметим, что произвольная лексикографически упорядоченная таблица двоичных шкал с числом строк C_m^2 не обязательно является таблицей шкал различий какого-то множества из m объектов.

В табл. 2 представлено множество B из шести объектов с соответствующей ему таблицей шкал различий. Пятнадцать строк ТШР представляют пятнадцать шкал попарных различий объектов множества B.

Как было отмечено выше, для некоторых прикладных задач выбор системы отсчета или порядок следования описаний объектов (порядок следования в исходной таблице шкал значений признаков) могут нести смысловую нагрузку. Тогда часть существенной информации о структуре множества X заведомо теряется при переходе от исходного представления к представлению в виде таблицы шкал различий.

Если же порядок объектов в исходном представлении и использование конкретной системы отсчета несущественны (выбираются случайным образом), то способ представления структуры множества X в виде ТШР имеет то преимущество, что является инвариантным к таким несущественным случайным факторам.

Таблица 2

Пример таблицы шкал различий для множества из шести объектов

Таблица шкал значений		**Таблица шкал различий**					
Объекты	Шкалы значений	Пары объектов	Шкалы различий	Пары объектов	Шкалы различий	Пары объектов	Шкалы различий
a	[0,1]	$a-b$	[0,0]	$a-b$	[0,1]	$a-b$	[1,0]
b	[0,1]	$c-d$	[0,0]	$c-d$	[0,1]	$c-d$	[1,1]
c	[0,0]	$e-f$	[0,0]	$e-f$	[1,0]	$e-f$	[1,1]
d	[0,0]	$c-e$	[0,1]	$c-e$	[1,0]	$c-e$	[1,1]
e	[1,0]	$c-f$	[0,1]	$c-f$	[1,1]	$c-f$	[1,1]
f	[1,0]						

Таблица шкал различий включает лишь данные, касающиеся различий между кодами любой пары объектов анализируемого множества. Это обстоятельство дает основание использовать термин «разностные алгоритмы» в случае, когда в роли входных данных алгоритма выступают только данные ТШР

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

(такие алгоритмы относятся к группе $\Re_{\text{dif}}$). Посредством обработки данных из ТШР разностные алгоритмы выделяют информацию о структурных свойствах исходного множества объектов. Для реализации разностных алгоритмов нет необходимости строить ТШР в явном виде — при реализации алгоритма могут использоваться и исходные описания (векторы значений двоичных признаков в каком-либо из базисов представлений) с оперативным вычислением значений, соответствующих компонентам ТШР. Определяющим свойством разностных алгоритмов является их независимость от инверсии значений признаков в исходном представлении множества и от порядка следования описаний объектов в исходном представлении.

Спектры расстояний. Простейшими примерами разностных алгоритмов (алгоритмами, определяющими наиболее общие структурные характеристики исходного множества) являются вычисления максимального, минимального и среднего расстояния между элементами исходного множества. Спектр расстояний и другие характеристики, извлекаемые из ТШР, могут дать информацию о расслоении исходного множества на ряд подмножеств тесно сгруппированных объектов.

Заметим, что при представлении объектов наборами двоичных признаков вес вектора признаков (число единиц в векторе признаков), представляющего конкретный объект, зависит от того, какой из базисов выбран для представления значений свойств объектов. Вместе с тем, вес w шкалы различий, соответствующий конкретной паре объектов, не зависит от выбора базиса, так же, как и количество $D(w)$ элементов ТШР с весом, равным w. Функция $D(w)$ в концентрированном виде представляет информацию об анализируемом множестве объектов — она дает спектр расстояний между объектами. Функция $D(w)$ представляет важную информацию о структурных особенностях анализируемого множества в виде, не связанном с выбором базиса представления описаний объектов.

Примеры. Как видно из трех нижеследующих примеров, вид функции $D(w)$, представляющей спектр расстояний, существенно различается для множеств с радикально различными структурами.

Пример 1: множество, состоящее из 2^n объектов, равномерно распределенных по вершинам n-мерного куба. График $D(w)$ спектра расстояний для такого множества, представленный логарифмическом масштабе [5], имеет характерный вид купола $D(w) = C_n^w$, где C_n^w — число сочетаний из n по w.

Пример 2: множество элементов кода с проверкой на четность. Для такого множества общий ход графика $D(w)$ спектра расстояний имеет то же куполообразный вид, но $D(w) = 0$ для любого нечетного w.

Пример 3: множество из четного числа m элементов разделено на два равные подмножества X и Y. Коды всех элементов подмножества X соответствуют некоторой вершине φ единичного n-мерного единичного куба. Коды всех элементов подмножества Y соответствуют диаметрально противоположной вершине $\overline{\varphi}$. Для такого множества крайним значениям аргумента w графика $D(w)$ спектра расстояний соответствуют величины

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

$D(0) \approx D(n) \approx C_m^2/2$, всем остальным значениям аргумента w соответствуют нулевые значения функции $D(w)$.

Поднаборы свойств и частные спектры расстояний. Каждому поднабору ψ признаков (поднабору свойств в описаниях объектов) соответствует неполный набор столбцов ТШР и, соответственно, частный спектр $D(\psi,w)$ расстояний между кодами таких неполных описаний объектов. Все частные спектры расстояний $D(\psi,w)$ однозначно определяются содержимым ТШР. Алгоритмы, для которых в роли входных данных выступают только данные о спектре расстояний $D(w)$ и о частных спектрах расстояний $D(\psi,w)$, относим к группе $\mathfrak{R}_{\text{dispec}}$ (спектральные алгоритмы).

Идентичные объекты. Значение $D(0)$ спектра расстояний указывает количество пар объектов с полностью совпадающими описаниями. Выполнение неравенства $D(0) \neq 0$ означает, что в ТШЗ имеются полностью идентичные описания объектов. Для множества m объектов в пространстве n двоичных признаков равенство $D(0) = 0$ может выполняться только в том случае, если $m \leq 2^n$. Случаю вырожденного множества, для которого (в выбранном базисе) описания всех m объектов идентичны, соответствует равенство $D(0) = C_m^2$.

Значение $D(\psi,0)$ частного спектра расстояний указывает количество пар объектов с полностью совпадающими описаниями по поднабору ψ признаков. Выполнение неравенства $D(\psi,0) \neq 0$ означает, что в ТШЗ имеются описания объектов, идентичные по поднабору ψ признаков объектов. Для множества из m объектов в пространстве n двоичных признаков равенство $D(\psi,0) = 0$ может выполняться только в том случае, если $m \leq 2^k$, где k — число признаков в поднаборе ψ. Равенство $D(\psi,0) = C_m^2$ означает совпадение значений признаков всех m объектов по всем признакам поднабора ψ (это свидетельствует о том, что признаки поднабора ψ неинформативны).

Алгоритмы, использующие только значения $D(\psi,0)$ для различных поднаборов признаков и значение $D(0)$, относятся к группе $\mathfrak{R}_{\text{iden}}$ (алгоритмы на основе учета числа совпадающих описаний объектов по поднаборам признаков).

Потеря данных о структуре множества при переходе от ТШЗ к ТШР. В конкретных предметных областях возникает большое разнообразие эвристических алгоритмов, критериев оценки свойств множества X по информации, содержащейся в таблице шкал различий. При этом встает вопрос: каковы потенциальные возможности такого рода критериев, алгоритмов, сколь полно ТШР отражает сведения о конфигурации множества, представляемого исходными шкалами значений признаков.

Рассмотрение примеров, приведенных в табл. 1 и табл. 2, показывает, что в общем случае двум множествам объектов с принципиально различной структурой может соответствовать одна и та же ТШР. Действительно, структура множества A в табл. 1 отличается от структуры множества B в табл. 2: множе-

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

ство A содержит подмножество из трех идентичных объектов (объекты b, c, d), множество B содержит три подмножества, каждое из которых пару объектов с идентичными описаниями.

Таким образом, структуры множеств A и B радикально различны. Между тем, нетрудно проверить, что ТШР для множества A в табл. 1 полностью совпадает с ТШР для множества B в табл. 2. Следовательно, в общем случае при переходе к представлению посредством ТШР может утрачиваться существенная часть информации о структуре представляемого множества объектов, что приводит к выводу об ограниченности возможностей разностных алгоритмов. Оказывается, однако, что существует широкий класс множеств, для которых таблицы шкал различий обладают всей полнотой информации о конфигурации представляемых ими множеств и, следовательно, любые оценки структурных особенностей представляемого множества могут быть реализованы посредством применения разностных алгоритмов. Цель дальнейшего изложения — показать, что такой полнотой обладают таблицы шкал различий для множеств со сколь угодно большим нечетным числом объектов. Ниже приводится доказательство утверждения, из которого следует, что если множество X содержит нечетное число m объектов, то соответствующая ТШР содержит всю информацию о структурных особенностях множества X. Приводимое ниже доказательство этого факта носит конструктивный характер: строится процедура вычисления по содержимому ТШР соответствующего набора строк ТШЗ в минимизурующем базисе. Формулируются условия однозначности возможных решений на каждом шаге процедуры, однозначности общего результата работы процедуры.

3. Процедура восстановления шкал значений признаков

Для конструирования процедуры необходимо ввести некоторые обозначения. Областью значений переменных, обозначаемых греческими буквами, будем считать множество векторов (шкал) из n двоичных символов, при необходимости отмечая вес шкалы (число единиц в ней): символ $\varphi^{(k)}$ обозначает некоторую шкалу, содержащую k единиц. Единственная шкала веса ноль $\varphi^{(0)}$ состоит из n нулей; единственная шакала $\varphi^{(n)}$ веса n состоит из n единиц. Для обозначения шкалы, получаемой из φ инвертированием всех ее компонент, используем символ $\overline{\varphi}$. Запись $\varphi \le \psi$ означает, что значения компонент шкалы φ не превосходят значений соответствующих компонент шкалы ψ. Двоичные шкалы будут использоваться, в частности, для указания поднаборов признаков базиса $(p_1, p_2, ..., p_n)$, для выделения соответствующих подтаблиц составляемых из части столбцов ТШР, для выделения частичных описаний объектов (т.е., описаний, содержащих значения лишь части признаков базисного набора).

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Число нулевых строк в подтаблице φ таблицы шкал различий обозначим символом $E(\varphi)$. Например, для ТШР табл. 2 $E(\varphi^{(2)}) = 3$.

Будем обозначать символом $q(\varphi)$ число объектов множества X, которым в выбранном базисе соответствует шкала φ. Другими словами $q(\varphi)$ указывает количество строк ТШЗ, совпадающих с φ. Символ $z(\varphi)$ указывает число нулевых строк в подтаблице ТШЗ, составленной из тех столбцов, которые соответствуют единицам шкалы ψ. Условимся считать $z(\psi^{(0)})$ равным числу объектов множества X. Так, для представления множества A объектов в первом столбце табл.1 имеем:

$$q(00) = 1, \quad q(01) = 1, \quad q(10) = 1, \quad q(11) = 1,$$
$$z(00) = 6, \quad z(01) = 2, \quad z(10) = 2, \quad z(11) = 1.$$

Покажем с использованием введенных обозначений, как может быть построена процедура восстановления ТШЗ. Эта процедура обеспечивает построение исходного набора шкал признаков по информации из соответствующей ТШР. В ходе выполнения процедуры последовательно восстанавливаются частичные описания представляемого множества объектов по неполным наборам признаков (т.е., все $2^n - 2$ подтаблиц ТШЗ, составляемые из подмножеств ее столбцов). Реализация процедуры восстановления ТШЗ, по-видимому, не имеет смысла, но ее существование показывает потенциальные возможности группы разностных алгоритмов. В основу процедуры положено следующее утверждение.

Утверждение 1. Пусть для некоторой ТШЗ восстановлены все $2^n - 2$ подтаблицы, составленные из ее столбцов. Тогда количество $q(\varphi)$ строк, совпадающих с φ в полной ТШЗ, может быть выражено через количество $q(\varphi^{(0)})$ нулевых строк в полной ТШЗ и количество нулевых строк в каждой из $2^n - 2$ подтаблиц ТШЗ.

Можно показать (индукцией по числу столбцов в ТШЗ), что величины $q(\varphi)$ и $z(\psi)$ связаны соотношением:

$$q(\varphi) = \sum_{\psi \le \varphi} z(\overline{\psi})(-1)^{(\|\varphi\| - \|\psi\|)}. \qquad (1)$$

Для шкалы $\varphi^{(k)}$, содержащей k единиц, правая часть представления $q(\varphi^{(k)})$ по формуле (1) содержит 2^k слагаемых. Отметим, что в принятых обозначениях количество нулевых строк в полной ТШЗ может быть представлено двумя способами:

$$q(\varphi^{(0)}) = z(\varphi^{(n)}). \qquad (2)$$

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Равенство (2) справедливо в силу того, что как в левой, так и в правой его части стоит обозначенное различными способами число объектов x, которым соответствуют нулевые значения всех признаков базиса. Следовательно, одно из слагаемых суммы в правой части формулы (1) (а именно, слагаемое, соответствующее нулевому значению индекса суммирования), с учетом (2) может быть представлено как $(-1)^k q(\varphi^{(0)})$:

$$q(\varphi) = (-1)^{\|\varphi\|} q(\varphi^{(0)}) + M(\varphi), \tag{3}$$

а величина оставшейся части $M(\varphi)$ суммы $\sum_{\psi \le \varphi}$ зависит только от частичных описаний объектов по неполным наборам признаков:

$$M(\varphi) = \sum_{\psi \neq \eta^{(0)}, \psi \le \varphi} z(\overline{\psi})(-1)^{(\|\varphi\| - \|\psi\|)}, \tag{4}$$

где суммирование ведется по $\psi \le \varphi$ с весом от 1 до $\|\varphi\|$. Утверждение 1 доказано, поскольку выражения $M(\varphi)$ во всех формулах типа (3) зависят только от данных из $2^n - 2$ подтаблиц ТШЗ.

Эквивалентность разностных и спектральных алгоритмов. Заметим, что равенство (1) выполняется для любой прямоугольной таблицы, содержащей строки двоичных кодов. В качестве такой таблицы может выступать не только ТЩЗ, но и ТШР. Таким образом, из равенства (1) следует, что ТШР может быть однозначно восстановлена по информации о числе нулевых строк в каждой из подтаблиц ТШР. Число нулевых строк в подтаблице ТШР равно числу пар объектов, неразличимых по соответствующему поднабору признаков (входная информация для алгоритмов группы $\Re_{\text{iden}}$).

Следовательно, ТШР, (содержащая всю входную информацию для алгоритмов группы $\Re_{\text{dif}}$), может быть построена по входным данным для алгоритмов группы $\Re_{\text{iden}}$. А это значит, что в силу справедливости равенства (1) для множеств объектов в пространстве двоичных кодов выполняется соотношение $\Re_{\text{iden}} \supseteq \Re_{\text{dif}}$. Но, вместе с тем, в соответствии с определением групп алгоритмов $\Re_{\text{dif}}$, $\Re_{\text{dispec}}$ и $\Re_{\text{dispec}}$, выполняются и соотношения $\Re_{\text{dif}} \supseteq \Re_{\text{dispec}} \supseteq \Re_{\text{iden}}$. Из этого непосредственно следует вывод об эквивалентности всех трех групп алгоритмов для множеств объектов в пространстве двоичных признаков ($\Re_{\text{dif}} = \Re_{\text{dispec}} = \Re_{\text{iden}}$).

Таким образом, если какие-то сведения о структурных свойствах анализируемого множества могут быть получены в рамках выполнения разностных алгоритмов, то эти сведения могут быть получены и путем обработки данных об удельном весе пар объектов с совпадающими описаниями (по каждому поднабору двоичных признаков).

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Утверждение 2. Количество $q(\varphi^{(0)})$ нулевых строк в полной ТШЗ может быть восстановлено путем решения квадратного уравнения по данным из полной ТШР и из предварительно восстановленных $2^n - 2$ подтаблиц ТШЗ.

Очевидно, что число $E(\eta^{(n)})$ нулевых строк в полной ТШР, определяется выражением

$$E(\eta^{(n)}) = \sum_{\varphi} C^2_{q(\varphi)}, \tag{5}$$

где суммирование ведется по всем шкалам значений φ, существующим в ТШЗ ($C^2_1 = 0$). Очевидно также, что формула (5) может быть переписана в виде

$$2E(\eta^{(n)}) + m = \sum_{\varphi} q^2(\varphi), \tag{6}$$

где m – число строк в ТШЗ и суммирование ведется уже по всем 2^n возможным значениям индекса φ. Подставляя (3) и (4) в (6) имеем

$$2E(\eta^{(n)}) + m = \sum_{\varphi} ((-1)^{\|\varphi\|} q(\varphi^{(0)}) + \sum_{\psi \neq \eta^{(0)}, \psi \leq \varphi} z(\overline{\psi})(-1)^{(\|\varphi\| - \|\psi\|)})^2, \tag{7}$$

где индекс φ внешнего суммирования пробегает все возможные 2^n значений, индекс ψ внутренней суммы принимает ненулевые значения, ограниченные неравенством $\psi \leq \varphi$.

Поскольку левая часть выражения (7) зависит только от содержания ТШР, а величины $z(\overline{\psi})$ зависят только от частичных описаний (от подтаблиц ТШЗ), утверждение 2 доказано.

Значение $q(\varphi^{(0)})$, получаемое решением уравнения (7), позволяет по формулам (3) восстановить шкалы значений признаков для всех объектов множества X, то есть, восстановить полную ТШЗ.

Общая схема процедуры восстановления ТШЗ по информации из ТШР состоит из n шагов.

Шаг 1. По каждому отдельному признаку p_i минимизурующего базиса вычисляется количество объектов $x \in X$, для которых $p_i(x) = 0$. В качестве входных данных используются соответствующие значения $E(\varphi^{(1)})$.

Шаг 2. Для всех поднаборов $\varphi^{(2)}$, т.е. для всех пар признаков (p_j, p_k), $j \neq k_i$ определяется количество объектов из X, для которых значения признаков p_j и p_k одновременно равны нулю. В качестве входных данных используются: а) результаты, полученные на первом шаге процедуры; б) соответствующие значения $E(\varphi^{(2)})$.

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

. . .

Шаг n. С учетом результатов предварительных вычислений, полученных на предшествовавших $n-1$ шагах, определяется набор полных шкал значений, представляющих объекты множества в минимизирующем базисе.

Таким образом, по данным, содержащимся в ТШР, восстанавливается исходное описание множества в виде набора строк ТШЗ в минимизурующем базисе.

4. Условия однозначности

Итак, ТШР содержит информацию, по которой, следуя рассмотренной процедуре, можно восстановить соответствующий ей комплект описаний объектов множества X. Если при переходе к представлению множества X в виде ТШР не происходит потери информации о структуре множества, то каждый шаг процедуры должен давать однозначное решение. В противном случае процедура восстановления множества шкал значений объектов из X дает более одного решения (как в случае ТШР, представленной в табл. 2), и это означает, что при представлении множества X в виде ТШР часть информации о структуре множества X теряется. Рассмотрим с этой точки зрения шаг 1. Число нулей в каждом из столбцов ТШР определяется выражениями

$$E(\varphi^{(1)}) = C^2_{z(\varphi^{(1)})} + C^2_{m-z(\varphi^{(1)})},$$

представляющими собой квадратные уравнения относительно $z(\varphi^{(1)})$. Если корни уравнения совпадают, решение однозначно. Если они различны, то только один из них соответствует минимизирующему базису, то есть, и в этом случае решения, получаемые на шаге 1 однозначны.

Решения, получаемые на последующих шагах процедуры, в общем случае могут быть неоднозначными. Они получаются решением квадратных уравнений типа (7). Раскрытие скобок под знаком суммы приводит уравнение (7) к виду:

$$2^n q^2(\varphi^{(0)}) + 2Sq(\varphi^{(0)}) + G = 0, \qquad (8)$$

где G — некоторое целое число, S определяется выражением

$$S = \sum_{\varphi} (-1)^{\|\varphi\|} \sum_{\psi \neq \eta^{(0)}, \psi \leq \varphi} z(\overline{\psi})(-1)^{(\|\varphi\| - \|\psi\|)} = \sum_{\varphi} \sum_{\psi \neq \eta^{(0)}, \psi \leq \varphi} z(\overline{\psi})(-1)^{\|\psi\|}, \qquad (9)$$

индекс суммирования φ пробегает все 2^n возможных значений. Анализ показывает, что в двойной сумме в (9) слагаемое $z(\overline{\psi})(-1)^{\|\psi\|}$ появляется $2^{n-\|\psi\|}$ раз. В частности, слагаемое, $z(\overline{\psi}^{(n)})(-1)^{\|\psi\|}$ появляется ровно один раз, а все остальные слагаемые появляются четное число раз. Поскольку

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

$z(\overline{\psi}^{(n)}) = z(\eta^{(0)})$ равно числу m строк ТШЗ, равенство (9) можно переписать в виде

$$S = m(-1)^n + 2v\,,$$

где m — число строк в ТШЗ, v — некоторое целое число. Тогда корни уравнения (8) представляются дробью:

$$q(\varphi^{(0)}) = \frac{-(m(-1)^n + 2v) \pm \sqrt{r}}{2^n}\,. \qquad (10)$$

Выражение (10) является рабочей формулой для получения результатов на всех шагах рассмотренной процедуры, начиная со второго. Условия однозначности результата, получаемого при выполнении процедуры, сводятся к однозначности выбора корня выражения (7) на каждом шаге. Если хотя бы на одном шаге оба значения, представляемые выражением (7), имеют смысл, комплект шкал значений восстанавливается по ТШР неоднозначно. Пример такой ситуации — обработка ТШР, представленной в табл. 2; второй шаг процедуры дает для этой ТШР два возможных решения: $q(\varphi^{(0)}) = 1$ и $q(\varphi^{(0)}) = 2$. Первое решение соответствует структуре множества A, представленного в табл. 1, второе — структуре множества B, представленного в табл. 2.

При больших значениях m и n можно найти такие ТШР, каждой из которых соответствует очень большое число различных структур представляемого множества объектов. Это делает особенно примечательным следующий факт.

Утверждение 3. При сколь угодно больших нечетных значениях m и любых $n \geq 2$ два корня, представляемые выражением (10) не могут быть одновременно целочисленными.

Справедливость утверждения следует из того, что при $n \geq 2$, нечетном целом x и целом y дробная часть выражения $R = \frac{x + 2y}{2^n}$ не может быть равна нулю или степени двойки. Но в таком случае при любом a выражения $R - a$ и $R + a$ не могут быть одновременно целочисленными, иначе целочисленной была бы их сумма, равная $2R$, и тогда дробная часть R составляла бы 0 или 2^{-1}.

Следствием приведенного утверждения является взаимнооднозначное соответствие между таблицами шкал различий и структурными свойствами анализируемых множеств объектов в пространстве двоичных признаков, если рассматриваемые множества содержат (со сколь угодно большое) нечетное число объектов. Таким образом, для множеств с нечетным числом объектов в пространстве двоичных признаков группы алгоритмов инвариантных к перекодировке признаков $\mathfrak{R}_{\text{codinv}}$ и разностных алгоритмов $\mathfrak{R}_{\text{dif}}$ эквивалентны ($\mathfrak{R}_{\text{codinv}} = \mathfrak{R}_{\text{dif}}$).

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

5. Заключение

Для решения задач классификации, распознавания, предсказания используются алгоритмы, определенные над множествами описаний объектов, представленных фиксированными наборами значений признаков. При этом возникает необходимость поиска эффективных (в отношении того или иного критерия) алгоритмов анализа свойств представляемых множеств объектов (например, алгоритмов разбиения анализируемого множества на составляющие подмножества, алгоритмов отыскания поднаборов признаков, которые обеспечивают наилучшие результаты в области диагностики, классификации, предсказания).

Выбор рамок поиска эффективных алгоритмов подобного рода во многих случаях определяется тем или иными эвристическими принципами. При этом важно знать, в какой мере выбранные рамки сужают пространство поиска. Слишком широкие рамки означают снижение продуктивности поиска, слишком узкие рамки могут означать, что наиболее эффективные алгоритмы останутся вне сферы поиска.

Рассмотрены пять групп алгоритмов. Группа $\mathfrak{R}_{\text{codfelt}}$ — результаты зависят от способа кодировки признаков. Группа $\mathfrak{R}_{\text{codinv}}$ — результаты инвариантны к способу кодировки. Группа $\mathfrak{R}_{\text{dif}}$ — разностные алгоритмы, результаты зависят только от попарных различий в описаниях объектов. Группа $\mathfrak{R}_{\text{dispec}}$ — спектральные алгоритмы, результаты зависят только от спектров расстояний между объектами по подмножествам признаков. Группа $\mathfrak{R}_{\text{iden}}$ — результаты зависят только от частоты совпадающих пар описаний объектов по подмножествам признаков. Непосредственно из определений рассматриваемых групп алгоритмов следует, что

$$\mathfrak{R}_{\text{codfelt}} \supseteq \mathfrak{R}_{\text{codinv}} \supseteq \mathfrak{R}_{\text{dif}} \supseteq \mathfrak{R}_{\text{dispec}} \supseteq \mathfrak{R}_{\text{iden}}$$

на каждом шаге продвижения по цепочке групп алгоритмов сравнительные возможности рассматриваемых групп алгоритмов не расширяются — все структурные особенности анализируемых множеств, которые могут быть выявляемые алгоритмами из групп в правой части цепочки, могут быть выявлены и алгоритмами предыдущих групп.

В общем случае (с учетом возможности использования многозначных признаков) рассмотренные группы алгоритмов строго упорядочены в отношении возможностей обнаружения структурных свойств анализируемого множества:

$$\mathfrak{R}_{\text{codfelt}} \supset \mathfrak{R}_{\text{codinv}} \supset \mathfrak{R}_{\text{dif}} \supset \mathfrak{R}_{\text{dispec}} \supset \mathfrak{R}_{\text{iden}}.$$

Простые примеры показывают, что в общем случае на каждом шаге продвижения по приведенной цепочке групп алгоритмов теряется часть тех структурных свойств, которые могут быть обнаружены в анализируемых множествах средствами очередной группы алгоритмов.

В то же время, в случае двоичных признаков соотношение возможностей рассматриваемых групп алгоритмов принципиально иное. Как показано в настоящей работе, для множеств объектов в пространстве двоичных признаков сравнительные возможности групп алгоритмов $\mathfrak{R}_{\text{dif}}$, $\mathfrak{R}_{\text{dispec}}$ и $\mathfrak{R}_{\text{iden}}$ полностью эквивалентны:

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

$$\mathfrak{R}_{\text{dif}} = \mathfrak{R}_{\text{dispec}} = \mathfrak{R}_{\text{iden}}.$$

Кроме того, для множеств с нечетным числом объектов эта эквивалентность распространяется и на группу $\mathfrak{R}_{\text{codinv}}$ алгоритмов, инвариантных к способу кодировки двоичных признаков:

$$\mathfrak{R}_{\text{codinv}} = \mathfrak{R}_{\text{dif}} = \mathfrak{R}_{\text{dispec}} = \mathfrak{R}_{\text{iden}}.$$

Отсюда, в частности, следует, что при определении эвристических принципов поиска эффективных алгоритмов классификации, распознавания, прогнозирования следует принять к сведению, что для множеств в пространстве двоичных признаков все структурные свойства анализируемого множества, обнаруживаемые средствами алгоритмов группы $\mathfrak{R}_{\text{dif}}$ (или даже, в случае множеств с нечетным числом элементов, средствами алгоритмов группы $\mathfrak{R}_{\text{codinv}}$), могут быть обнаружены и средствами алгоритмов групп $\mathfrak{R}_{\text{dispec}}$ и $\mathfrak{R}_{\text{iden}}$. Другими словами, все структурные свойства анализируемого множества, обнаруживаемые разносными алгоритмами (а в случае множеств с нечетным числом элементов и любыми алгоритмами, инвариантными к перекодировке значений двоичных признаков), могут быть обнаружены и путем обработки данных (по поднаборам двоичных признаков) об удельном весе пар объектов с совпадающими описаниями.

Литература

1. Математические методы в распознавании образов и дискретной оптимизации / Под ред. Журавлева Ю. И. М.: ВЦ АН СССР, 1987. 113 с.
2. *Журавлев Ю. И., Никифоров В. В.* Алгоритмы распознавания, основанные на вычислении оценок // Кибернетика. 1971. № 3. С. 3–10.
3. *Загоруйко Н. Г., Елкина В. Н., Лобов Г. С.* Алгоритмы обнаружения эмпирических закономерностей. Новосибирск: Наука, 1985. 110 с.
4. *Goldberg D.* Genetic Algorithms n Search, Optimization and Learning. Boston: Addison–Wesley, 2002. 412 p.
5. *Nikiforov V.* Difference Algorithms Abilities in Decision Support Systems // Industrial Applications of Artificial Intelligence. Amsterdam: Elsevier Science Publishers, 1991. P. 308–313.