Анализ структур отношений между описаниями объектов классов и оценки их компактности

Е.Н. Згуральская[1]

[1]Ульяновский технический Университет. Институт авиационных технологий и управления, проспект Созидателей 13А, Ульяновск, Россия, 432072

Аннотация. В работе проводится исследование по оценке компактности описаний объектов классов на числовой оси и в многомерном признаковом пространстве. Вычисление компактности возможно лишь в определяемых границах областей признакового пространства. В одномерном случае границы вычисляются по частоте встречаемости значений признаков объектов классов в интервале. В многомерном случае используется подмножество граничных объектов классов по заданной метрике. Приведён сравнительный анализ значений меры компактности по латентным признакам на числовой оси и по наборам исходных признаков, из которых они синтезированы.

1. Введение

В теории распознавания образов разбиение объектов на классы базируется на гипотезе о компактности. Согласно этой гипотезе, «близкие» объекты должны лежать в одном классе. Требуется специальное уточнение (разъяснение) понятий «близость» и «компактность» объектов.

Не существует единого общепринятого определения понятия «компактность». В [1] предложена мера компактности непересекающихся классов, множество допустимых значений которой определено в (0, 1] и зависит от структуры отношений между объектами. Отмечены следующие факторы, влияющие на значение компактности:

- выбор метрики для вычисления расстояний между объектами;
- значение размерности признакового пространства;
- выбор способа масштабирования и нормирования данных;
- использование методов отбора информативных наборов признаков;
- условия отбора и удаления шумовых объектов из выборки данных;
- количество объектов–эталонов минимального покрытия обучающей выборки;
- линейные и нелинейные преобразования признакового пространства для описания объектов.

Компактность предполагает наличие границы между областями признакового пространства с описанием объектов из разных классов.

Различаются между собой и численные методы для количественного оценивания компактности. В одномерном случае для этого используются интервальные методы, в многомерном – вычисление меры компактности объектов классов и выборки в целом по заданной метрике. Общим для одномерного и многомерного случаев является наличие областей признакового пространства, в границах которых вычисляется мера компактности.

В одномерном случае на числовой оси можно производить сравнение объектов по значениям их исходных и латентных признаков, используя отношения «больше», «меньше» или «равно».
При вычислении меры компактности в многомерном случае в [1] применялось отношение связанности объектов по подмножеству (оболочке) граничных объектов непересекающихся классов. На основе этого отношения производилось разбиение объектов на непересекающиеся группы. Связанность объектов S_i, S_j рассматривалась, как свойство логических закономерностей в форме гипершаров, центрами которых они являлись. Объекты S_i и S_j считались связанными, если в пересечении их гипершаров содержались объекты оболочки. Любую пару объектов (S_i, S_j) из одной группы всегда можно соединить цепочкой из связанных объектов. В идеале все объекты класса представляют одну группу из связанных объектов.
В данной работе исследуются структуры отношений между описаниями объектов классов на числовой оси. В качестве инструмента для исследования предлагаются меры компактности, вычисляемые по результатам разбиения значений признаков (исходных и латентных) на непересекающиеся интервалы. Значения мер компактности используются для поиска скрытых закономерностей в данных. Скрытые закономерности могут рассматриваться как новое знание, полученное в рамках информационных моделей для слабо структурированных предметных областей.

2. Критерии для разбиения признаков на интервалы
Рассмотрим два вычислительных алгоритма для оптимизации критериев разбиения значений признаков на непересекающиеся интервалы, предложенные в [2, 3]. Для удобства изложения обозначим эти критерии CR1 и CR2.
При вычислении по CR1 число интервалов на упорядоченной последовательности значений признака равно числу непересекающихся классов. Значения границ интервалов определяются по максимуму произведения внутриклассового сходства и межклассового различия. В идеале каждый интервал должен быть представлен всеми значениями признака объектов одного класса.
Для критерия CR2 число классов равно 2, число интервалов больше либо равно 2. При вычислении границ непересекающихся интервалов, число которых изначально неизвестно, используется абсолютная разность частот встречаемости значений признаков (как исходных, так и латентных) в описании объектов двух классов. Значения признаков на числовой оси образуют последовательность кластеров (интервалов). Не должно существовать двух соседних кластеров, в которых доминировали бы (по частоте встречаемости) представители одного класса. Идеальными в смысле устойчивости считаются разбиения, при которых в границах каждого интервала содержатся значения (не обязательно всех) объектов только одного класса.
Множество допустимых значений критерия CR1 и устойчивое разбиение признаков на интервалы по CR2 содержится в отрезке [0;1] и далее рассматривается как мера их компактности. Значение 1 соответствует идеальному разбиению по CR1 и CR2. О степени отклонения от идеала можно судить по значениям меньше 1.
Комбинированное использование критериев CR1 и CR2 необходимо для обнаружения скрытых закономерностей в данных. Поиск закономерностей производится по результатам вычислительного эксперимента. Для интерпретации результатов эксперимента используются известные формы логических закономерностей (гипершар, полуплоскость, параллелепипед).
Пусть задано множество объектов $E_0=\{S_1,\ldots,S_m\}$, содержащее представителей d непересекающихся классов $K_1,\ldots,K_d$. Описание объектов производится с помощью набора из n разнотипных признаков $X(n)$, δ ($\delta<n$) из которых измеряются в номинальной, $n–\delta$ в интервальных шкалах. Допускается наличие пропусков и повторяющихся значений в данных.
Большой интерес представляет отыскание латентных, то есть срытых признаков, которые могут оказаться очень информативными при классификации, что и составляет одну из задач настоящего исследования. Считается, что для разбиения значений количественного признака (как исходного, так и латентного) на непересекающиеся интервалы используются критерии CR1 и CR2. Латентные признаки могут представлять комбинации из номинальных и количественных признаков. Требуется определить:

- способ вычисления латентных признаков;
- границы интервалов и значения критерия CR1 на исходных и латентных признаках;
- число интервалов, значения их границ и устойчивость разбиения исходных признаков по критерию CR2.

Многообразие способов формирования латентных признаков и критериев для разбиения их значений на непересекающиеся интервалы необходимо для поиска скрытых закономерностей по базам данных предметных областей. Латентные признаки из набора $X(n)$ будем формировать в виде комбинаций x_i*x_j и x_i/x_j. Если число градаций номинального признака равно числу непересекающихся классов объектов, то им всегда можно поставить в соответствие набор целых чисел $a_1,\ldots,a_d$, где $a_i \neq 0$, $i=1,\ldots,d$ и $a_{j+1}-a_j=const$, где $j=1,\ldots,d-1$. Каждый непересекающийся интервал по CR1 будет представлен одним значением. Например, при числе градаций равной 2, удобной для вычисления формой представления является выбор значений из [-1,1].

3. Вычислительный эксперимент

Рассмотрим результаты разбиения количественных признаков на непересекающиеся интервалы по критериям CR1, CR2 на выборке данных из [4]. Выборка состоит из двух классов K_1, K_2 и содержит данные о сердечно-сосудистых заболеваниях. Описание объектов производится набором признаков $X(13)=(x_1,\ldots,x_{13})$. Число объектов класса K_1 равно 150, K_2 равно 120. Признаки x_1, x_4, x_5, x_8, x_{10}, x_{11}, x_{12} – количественные, x_2, x_3, x_6, x_7, x_9, x_{13} – номинальные. Номинальные признаки x_2, x_6, x_9 имеют две градации (т.е. число градаций признака равно числу классов) Компактность количественных признаков из $X(13)$ и границы интервалов по CR1 приводятся в таблице 1.

Таблица 1. Границы интервалов и значения компактности по CR1.

	Название признака	Границы интервалов	Компактность
x_1	Возраст	[29..54] (54..77]	0.2871
x_4	Покоящееся кровяное давление	[94..135] (135..200]	0.2548
x_5	Холесторальная сыворотка в мг./дл.	[126..252] (252..564]	0.2684
x_8	Достигнутый максимальный сердечный ритм	[71..147] (147..202]	0.3413
x_{10}	Oldpeak = депрессия ST, вызванная упражнениями относительно покоя	[0..1.6] (1.6..6.2]	0.3177
x_{11}	Наклон пикового упражнения ST сегмент	(1..2] (2..3]	0.3246
x_{12}	Количество основных сосудов (0-3), окрашенных флоусопой	[0..1] (1..3]	0.3772

Для вычисления весов номинальных признаков по CR1 (как и для вычисления компактности количественных признаков) используется произведение внутриклассового сходства и межклассового различия. Если значения градаций в описании объектов каждого класса не пересекаются друг с другом, то вес номинального признака равен 1. В таблице 2 представлены значения всех шести номинальных признаков.

Таблица 2. Веса номинальных признаков.

Признак	Название признака	Вес
x_2	Пол	0.2727
x_3	Тип боли в груди	0.3203
x_6	Уровень сахара в крови натощак> 120 мг./дл.	0.1873
x_7	Результаты электрокардиографии покоя	0.2762
x_9	Упражнение индуцированной стенокардии	0.3453
x_{13}	thal: 3 = нормальный; 6 =фиксированный дефект; 7 = обратимый дефект.	0.4193

Как видно из таблицы 1 и таблицы 2, компактность количественных по CR1 и значения весов номинальных признаков сильно отличаются от идеала. Значение количественного признака в границах непересекающегося интервала по CR1 можно рассматривать как градацию (номер

интервала) в номинальной шкале измерений. Вес признака при таком описании объектов в номинальной шкале будет совпадать со значением компактности по критерию CR1.
Число непересекающихся интервалов и устойчивость разбиения по критерию CR2 приводится в таблице 3.

Таблица 3. Устойчивость признаков и границы интервалов по критерию CR2.

Признак	Границы интервалов	Устойчивость
x_1	[29..54], [55..70], [71..76], [77..77]	0.6571
x_4	[94..122], [123..200]	0.5585
x_5	[126..160], [164..174], [175..245], [246..353], [354..394], [407..409], [417..564]	0.6309
x_8	[71..147], [148..194], [195..195], [202..202]	0.7030
x_{10}	[0..0.8], [0.9..6.2]	0.6957
x_{12}	[0.. 0], [1..3]	0.7316

Как видно из таблицы 1 и таблицы 3, относительно высокие значения компактности получены по признаку x_{12}.
Разбиение на два интервала латентных признаков, полученных по операциям умножения и деления значений исходных признаков, приводится в таблице 4.

Таблица 4. Границы интервалов для латентных признаков и значения компактности по CR1.

Латентный признак	Границы интервалов	Компактность
x_4*x_9	[-192..105] (105..200]	0.3552
x_8*x_9	[-202..-115] (-115..186]	0.3718
$x_{10}*x_{11}$	[1..3.3] (3.3..21.6]	0.3684
x_2/x_8	[-0.0104..0.0067] (0.0067..0.0140]	0.3597
x_8/x_{10}	[16.8182..66.6667] (66.6667..202]	0.3726
x_8/x_{11}	[32..75] (75..202]	0.3555
x_9/x_4	[-0.0106..-0.0062] (-0.0062..0.0106]	0.3504
x_9/x_8	[-0.0140..0.0061] (0.0061..0.0113]	0.3523
x_{10}/x_8	[0.0049..0.0149] (0.0149..0.0594]	0.3726
x_{11}/x_8	[0.0049..0.0132] (0.0132..0.0312]	0.3555

Анализ результатов из таблицы 4 и таблицы 1 показывает целесообразность поиска скрытых закономерностей по латентным признакам, компактность которых выше, чем каждого из исходных признаков, входящих в их состав.

4. Заключение

Для анализа структуры отношений между объектами непересекающихся классов по количественным исходным и латентным признакам предложено использовать два интервальных метода. Численные оценки структуры отношений по этим методам различаются между собой тем, что число непересекающихся интервалов известно изначально или определяется алгоритмом.
Рассмотренные методы рекомендуются использовать для поиска скрытых закономерностей в данных при разработке информационных моделей, основанных на знаниях.

5. Литература

[1] Ignatyev, N.A. Structure Choice for Relations between Objects in Metric Classification Algorithms // Pattern Recognition and Image Analysis. – 2018. – Vol. 28, № 4. – P. 590-597.
[2] Згуральская, Е.Н. Устойчивость разбиения данных на интервалы в задачах распознавания и поиск скрытых закономерностей // Известия Самарского научного центра Российской академии наук. – 2018. – Т. 14, № 4(3). – С. 826-829.

[3] Згуральская, Е.Н. Выбор информативных признаков для решения задач классификации с помощью искусственных нейронных сетей / Е.Н. Згуральская // Нейрокомпьютеры: разработка, применение. – 2012. – № 2. – С. 20-27.
[4] UCI repository of machine learning databases [Electronic resource]. – Access mode: http://archive.ics.uci.edu (03.12.2018).

Analysis of the structure of the relationship between the descriptions of objects of classes and evaluation of their compactness

E.N. Zguralskaya[1]

[1]Ulyanovsk Technical University. Institute of Aviation Technologies and Management, 13A Avenue, Ulyanovsk, Russia, 432072

Abstract. The study is conducted to assess the compactness of descriptions of objects of classes on the numerical axis and in the multidimensional attribute space. The computation of compactness is possible only in the defined boundaries of areas of the attribute space. In the one-dimensional case, the boundaries are calculated by the frequency of occurrence of the values of features of objects of classes in the interval. In the multidimensional case, a subset of the boundary objects of the classes is used for a given metric. A comparative analysis is given of the values of the compactness measure by latent attributes on the numerical axis and by the sets of initial features from which they are synthesized.