АНАЛИЗ РАЗНОТИПНЫХ ДАННЫХ
(Вычислительные системы)
1983 год Выпуск 99

УДК 519.237

КЛАССИФИКАЦИЯ КАК ВЫДЕЛЕНИЕ ГРУПП ОБЪЕКТОВ, УДОВЛЕТВОРЯЮЩИХ РАЗНЫМ МНОЖЕСТВАМ СОГЛАСОВАННЫХ ЗАКОНОМЕРНОСТЕЙ

Е.Е. Витяев

I. В настоящее время известно много принципов построения классификаций [I-5]: на основе гипотезы компактности и различных мер близости в некотором пространстве; по эталонам - сходством с эталонами, преобразованиями эталонов, выбором эталонов или типичных представителей; по суперцелям (например, для последующего распознавания); по различным критериям качества классификации и функционалам качества; путем разделения смесей распределений и другие.

В данной работе предлагается следующий принцип построения классификаций: разбиение на кластеры производится так, чтобы объекты одного кластера подчинялись одним и тем же закономерностям, объекты разных кластеров подчинялись разным группам закономерностей. Объекты одного кластера, кроме того, должны обладать некоторой целостностью. Целостность определяется как взаимная согласованность закономерностей каждой группы. Различным определениям понятий закономерности и взаимной согласованности соответствуют различные алгоритмы классификации. В данной работе эти понятия определяются для качественных признаков (измеренных в шкале наименований).

2. Пусть A - генеральная совокупность объектов и $x_1,\ldots,x_n$ - признаки, определенные на A, $I_i = \{x_i(a) \mid a \in A\} = \{x_{i1},\ldots,x_{ik_i}\}$, $i = 1,\ldots, n$. Определим атомарные одноместные отношения $P_{ij}(a) \Leftrightarrow x_i(a) = x_{ij}$, $x_{ij} \in I_i$, $i = 1,\ldots,n$; $j = 1,\ldots,k_i$. Множество атомарных отношений обозначим через X. Детерминированной закономерностью на A назовем истинную на

А формулу вида $\forall a\, \Phi(a)$, составленную из атомарных отношений или их отрицаний.

ОПРЕДЕЛЕНИЕ I. И м п л и к а т и в н о й д е т е р м и н и р о в а н н о й з а к о н о м е р н о с т ь ю назовем истинную на А формулу вида

$$\Phi = \forall a\, (P^{\varepsilon_1}_{i_1 j_1} \& \ldots \& P^{\varepsilon_1}_{i_1 j_1} \to P^{\varepsilon_0}_{i_0 j_0}). \qquad (1)$$

Для упрощения записи объект а при отношениях опущен; $\varepsilon_1 = I(0)$, если отношение берется без отрицания (с отрицанием). Эта формула удовлетворяет следующим условиям:

а) среди атомарных отношений $P^{\varepsilon_0}_{i_0 j_0}, P^{\varepsilon_1}_{i_1 j_1}, \ldots, P^{\varepsilon_1}_{i_1 j_1}$ нет повторений и нет одновременно отношения и его отрицания;

б) если убрать из конъюнкции $P^{\varepsilon_1}_{i_1 j_1} \& \ldots \& P^{\varepsilon_1}_{i_1 j_1}$ одно из отношений, либо заменить отношение $P^{\varepsilon_0}_{i_0 j_0}$ на Л(ложь), то полученная в результате формула уже не будет истинной на А.

В [6] доказано, что любая детерминированная закономерность логически эквивалентна совокупности импликативных детерминированных закономерностей. Расширим понятие детерминированной закономерности, введя закономерности, истинные в определенном смысле в большинстве случаев. Предположим, что задана некоторая процедура случайного выбора объектов из генеральной совокупности А и определены вероятности $\mathcal{P}(P_{ij})$ отношений из X .

ОПРЕДЕЛЕНИЕ 2. Формулу вида (I) назовем в е р о я т н о с т н о й з а к о н о м е р н о с т ь ю (на А), если

I) $\mathcal{P}(P^{\varepsilon_1}_{i_1 j_1} \& \ldots \& P^{\varepsilon_1}_{i_1 j_1}) > 0$;

2) $\mathcal{P}(P^{\varepsilon_0}_{i_0 j_0} / P^{\varepsilon_1}_{i_1 j_1} \& \ldots \& P^{\varepsilon_1}_{i_1 j_1}) > \mathcal{P}(P^{\varepsilon_0}_{i_0 j_0} / P^{\varepsilon_1}_{i_1 j_1} \& \hat{\ldots} \hat{\&} P^{\varepsilon_1}_{i_1 j_1})$,

где $\hat{\ldots}\hat{\&}$ обозначает отсутствие одного или нескольких отношений в конъюнкции;

3) При добавлении к конъюнкции $P^{\varepsilon_1}_{i_1 j_1} \& \ldots \& P^{\varepsilon_1}_{i_1 j_1}$ любого отношения из X (или его отрицания) нарушается одно из первых двух

условий. Обозначим условную вероятность $\mathcal{P}(P^{\varepsilon_0}_{i_0 j_0} / P^{\varepsilon_1}_{i_1 j_1} \& \ldots \ldots \& P^{\varepsilon_1}_{i_1 j_1})$ формулы Φ через $\mathcal{P}(\Phi)$.

ЛЕММА. Импликативные детерминированные закономерности являются вероятностными закономерностями.

ДОКАЗАТЕЛЬСТВО. Так как импликативная детерминированная закономерность перестает быть истинной на A при замене отношения $P^{\varepsilon_0}_{i_0 j_0}$ на Л, то конъюнкция $P^{\varepsilon_1}_{i_1 j_1} \& \ldots \& P^{\varepsilon_1}_{i_1 j_1}$ не всегда ложна на A. Следовательно, $\mathcal{P}(P^{\varepsilon_1}_{i_1 j_1} \& \ldots \& P^{\varepsilon_1}_{i_1 j_1}) > 0$.

Для импликативной детерминированной закономерности Φ имеем $\mathcal{P}(\Phi) = 1$. Так как Φ перестает быть истинной на A при удалении каких-либо отношений из $\{P^{\varepsilon_1}_{i_1 j_1}, \ldots, P^{\varepsilon_1}_{i_1 j_1}\}$, то $\mathcal{P}(P^{\varepsilon_0}_{i_0 j_0} / P^{\varepsilon_1}_{i_1 j_1} \& \hat{\ldots} \ldots \hat{} \& P^{\varepsilon_1}_{i_1 j_1}) < 1$, откуда вытекает второе условие. При добавлении к конъюнкции $P^{\varepsilon_1}_{i_1 j_1} \& \ldots \& P^{\varepsilon_1}_{i_1 j_1}$ любого отношения $P \in X$ (или его отрицания) она становится либо тождественно ложной на A, и тогда нарушается первое условие, либо, поскольку из истинности $P^{\varepsilon} \& P^{\varepsilon_1}_{i_1 j_1} \& \ldots \& P^{\varepsilon_1}_{i_1 j_1}$ следует истинность $P^{\varepsilon_1}_{i_1 j_1} \& \ldots \& P^{\varepsilon_1}_{i_1 j_1}$, а значит и истинность $P^{\varepsilon_0}_{i_0 j_0}$, получаем, что условная вероятность $\mathcal{P}(P^{\varepsilon_0}_{i_0 j_0} / P^{\varepsilon} \& P^{\varepsilon_1}_{i_1 j_1} \& \ldots \& P^{\varepsilon_1}_{i_1 j_1})$ также равна I, и, следовательно, нарушается второе условие. Лемма доказана.

Из леммы следует, что понятие вероятностной закономерности является расширением понятия детерминированной закономерности. Вероятностные закономерности можно с некоторым уровнем доверия об - наруживать на выборке из A. В [6] приведен метод обнаружения вероятностных закономерностей, использующий для проверки условия 2 точный критерий независимости Фишера для таблиц сопряженности признаков. Применяя этот метод с некоторым уровнем доверия α для критерия, получим множество формул F_α вида (I). Формулы из F_α будем называть закономерностями. Для каждой закономер-

ности $\Phi \in F_\alpha$ и некоторого доверительного уровня β в методе [6] определяется нижняя доверительная граница $\mathcal{P}^\beta(\Phi)$ условной вероятности $\mathcal{P}(\Phi)$.

3. Кластеры будем определять через наборы значений признаков. Набором значений признаков $x_{s_1},\ldots,x_{s_m}$ (признаки не повторяются) будем называть множество $\{Y_{s_1},\ldots,Y_{s_m}\}$, $Y_{s_t} \subset I_{s_t}$, $Y_{s_t} \neq \emptyset$, $t = 1,\ldots,m$. Каждый такой набор $\{Y_{s_1},\ldots,Y_{s_m}\}$ выделяет в произвольном множестве объектов $B \subset A$ подмножество $M_B(\{Y_{s_1},\ldots,Y_{s_m}\}) \rightleftarrows \{a \in B \mid a = \langle x_{1j_1},\ldots,x_{nj_n}\rangle,\ x_{s_tj_{s_t}} \in Y_{s_t},\ t=1,\ldots,m\}$. Будем говорить, что закономерность $P^{\varepsilon_1}_{i_1j_1} \& \ldots \& P^{\varepsilon_l}_{i_lj_l} \Rightarrow P^{\varepsilon_0}_{i_0j_0}$ применима к набору $\{Y_{s_1},\ldots,Y_{s_m}\}$, если $\{i_0, i_1,\ldots,i_l\} \subset \{s_1,\ldots,s_m\}$ и $x_{i_tj_t} \in Y_{i_t}$ ($x_{i_tj_t} \notin Y_{i_t}$) при $\varepsilon_t=1(0)$, $t=1,\ldots,l$ (заметим, что $t=0$ отсутствует). Если закономерность применима к набору $\{Y_{s_1},\ldots,Y_{s_m}\}$ и ее заключение (отношение $P^{\varepsilon_0}_{i_0j_0}$) выполнимо на этом наборе $x_{i_0j_0} \in Y_{i_0}$ ($x_{i_0j_0} \notin Y_{i_0}$) при $\varepsilon_0=1(0)$, то будем говорить, что эта закономерность подтверждается на этом наборе. Если закономерность применима к набору, но ее заключение невыполнимо на этом наборе $x_{i_0j_0} \notin Y_{i_0}$ ($x_{i_0j_0} \in Y_{i_0}$) при $\varepsilon_0 = 1(0)$, то будем говорить, что она опровергается на этом наборе.

Кластеры, в идеале, должны описываться такими наборами, на которых подтверждаются все применимые к ним закономерности. Однако по целому ряду причин это условие не выполняется. Поэтому для определения наборов, описывающих кластеры, необходимо ввести дополнительный критерий взаимной согласованности закономерностей, подтверждающихся и опровергающихся на этих наборах. По аналогии с антиципацией и ее подтверждением в психологии восприятия определим следующий критерий:

$$\mathcal{T}(\{Y_1,\ldots,Y_m\}) \rightleftarrows -\Big(\sum_{\Phi\in\Pi} \ln(1-\mathcal{P}^\beta(\Phi)) - \sum_{\Phi\in O} \ln(1-\mathcal{P}^\beta(\Phi))\Big),$$

где П - множество закономерностей из F_α, подтверждающихся на наборе $\{Y_1,\dots, Y_m\}$, О - множество закономерностей из F_α, опровергающихся на этом наборе.

ОПРЕДЕЛЕНИЕ 3. К л а с т е р о м (образом, таксоном) является такой набор значений признаков $\{Y_1, \dots, Y_m\}$, для которого критерий $\mathcal{T}$ имеет локальный максимум: при изменении любого из множеств $Y_1,\dots,Y_m$ на один элемент значение критерия строго уменьшается.

Определение 3 имеет статус гипотезы, состоящей в том, что реальным кластерам при представительной обучающей выборке соответствует набор значений признаков, удовлетворяющий этому определению.

Кластеры образуют иерархию: кластер $\{Y_1,\dots,Y_m\}$ может уточняться кластером $\{Y_1',\dots,Y_m',Y_{m+1}',\dots,Y_n'\}$, $Y_i' \subset Y_i$, $i = 1,\dots,m$.

4. Классификация. Возьмем выборку B из A и получим на ней множество закономерностей F_α^B. Каждому кластеру $\{Y_1,\dots,Y_m\}$ в выборке B соответствует подмножество $M_B(\{Y_1, \dots ,Y_m\})$, которое также назовем кластером. Множество всех кластеров образует классификацию. Критерий Фишера достаточно чувствителен и позволяет даже при небольших выборках обнаруживать закономерности, выделяющие кластер, состоящий из одного элемента, если этот элемент обладает своеобразным сочетанием значений признаков. В B могут быть также объекты, во всем похожие на объекты из некоторого кластера, за исключением одного - двух, существенных для этого кластера, значений признаков. Такие объекты могут не входить ни в один кластер.

Объединение всех множеств $Y_1,\dots,Y_m$ всех кластеров $\{Y_1,\dots,Y_m\}$ дает информативную систему значений признаков.

5. Распознавание. Пусть B - выборка из A и b - новый, случайно выбранный из A объект. Обнаружим на множестве $B \cup b$ закономерности $F_\alpha^{B\cup b}$. Определим все кластеры на $B \cup b$. Возможны три случая: I) объект b входит в некоторые кластеры, содержащие также объекты из выборки B; 2) объект b составляет одноэлементный кластер; 3) объект b не входит ни в один кластер. В собственном смысле распознавание - соотнесение объекта b с другими объектами выборки - происходит только в первом случае.

6. Предсказание. Пусть для объекта b требуется предсказать одно или несколько неизвестных значений признаков. Проведем распознавание объекта b, как описано в предыдущем пункте. Это воз-

можно, так как метод [6] обнаруживает закономерности при наличии пробелов. Объект $b = \langle x_{1j_1}, \ldots, x_{nj_n} \rangle$ с пропущенными значениями признаков будем относить к кластеру $M_{B \cup b}(\{Y_{s_1}, \ldots, Y_{s_m}\})$ при выполнении условий $x_{s_t j_{s_t}} \in Y_{s_t}$, $t = 1, \ldots, m$ для всех $x_{s_t j_{s_t}}$, определенных в b. Если объект b не входит ни в один кластер, предсказание не делается. Если объект b составляет самостоятельный кластер $b = M_{B \cup b}(\{Y^b_{s_1}, \ldots, Y^b_{s_m}\})$, то для неопределенных признаков объекта b множества $Y^b_{s_t}$ должны быть пусты. Поэтому предсказание в этом случае также не делается. Пусть объект b входит в кластеры $M^1_{B \cup b}(\{Y^1_{s_1}, \ldots, Y^1_{s_{m_1}}\}), \ldots, M^k_{B \cup b}(\{Y^k_{s_1}, \ldots, Y^k_{s_{m_k}}\})$, и нам надо предсказать неизвестное значение признака x_s. Так как объекты каждого кластера подчиняются своей группе закономерностей, то эти закономерности будут л о к а л ь н о - к о м п е т е н т н ы м и [7] для предсказания неизвестных значений признаков объектов данного кластера. Если в кластере $M^i_{B \cup b}$ есть множество значений Y^i_s признака x_s, то для объекта b предсказываются значения Y^i_s. Если ни один из кластеров $M^1_{B \cup b}, \ldots, M^k_{B \cup b}$ не содержит значений признака x_s, то предсказание не делается. Пусть $i_1, \ldots, i_v$ – номера кластеров, имеющих значения признака x_s. Тогда предсказанием по всем кластерам будет множество значений $Y_s = Y^{i_1}_s \cap \ldots \cap Y^{i_v}_s$. Если множество Y_s непусто, то оно является искомым предсказанием, если пусто – предсказание по разным кластерам противоречиво, и поэтому не осуществляется.

7. Замечания. Если, как и в [6], использовать закономерности, включающие произвольные многоместные отношения, то определение кластера можно распространить и на данные, измеренные в других шкалах.

Если принять, что определение кластера является формализацией клеточного ансамбля [8], а понятие закономерности формальной моделью нейрона, то процедуры классификации, распознавания и предсказания могут быть положены в основу формальной модели восприятия. При этом необходимо отвергнуть гипотезу об электрической сум-

мации импульсов на поверхности нейрона и принять концепцию интегративной деятельности нейрона, изложенную в [9].

Литература

I. ЗАГОРУЙКО Н.Г. Таксономия в анизотроном пространстве. - В кн. Эмпирическое предсказание и распознавание образов (Вычислительные системы, вып. 76). Новосибирск, 1978, с. 26-33.

2. Классификация и кластер /Под ред. Дж.Вэн Райзин, перев.под ред. Ю.И.Журавлёва. -М.: Мир, 1980. - 389 с.

3. ЗАГОРУЙКО Н.Г. Методы распознавания и их применение.-М.: Сов.радио, 1972. -206 с.

4. ДУДА Р., ХАРТ П. Распознавание образов и анализ сцен.-М.: Мир, 1976. - 510 с.

5. ТУ Дж., ГОНСАЛЕС Р. Принципы распознавания образов. - М.: Мир, 1978. -410 с.

6. ВИТЯЕВ Е.Е. Метод обнаружения закономерностей и метод предсказания. -В кн.: Эмпирическое предсказание и распознавание образов (Вычислительные системы, вып. 67). Новосибирск, 1976, с.54-68.

7. ЗАГОРУЙКО Н.Г., ЁЛКИНА В.Н., ТИМЕРКАЕВ В.С. Алгоритм заполнения пропусков в эмпирических таблицах (Алгоритм ZET).-В кн.: Вычислительные системы, вып. 61. Новосибирск, 1975, с. 3-27.

8. HEBB D.O.The organization of behavior.- New York: Willey, 1949.- 500 p.

9. АНОХИН П.К. Системный анализ интегративной деятельности нейрона. -В кн.:Анохин П.К. Очерки по физиологии функциональных систем. М., Медицина, 1975, с. 347-440.

Поступила в ред.-изд.отд.
11 ноября 1983 года