

Демин А.В.
Витяев Е.Е.

Метод построения «естественной» классификации¹

В работе рассматривается метод построения «естественных» классификаций и алгоритмы, используемые при его реализации. Описание метода иллюстрируется на примере решения тестовой задачи. В заключение приводится пример построения «естественной» классификации символов, взятых из различных шрифтов.

Ключевые слова: Интеллектуальный анализ данных, Data Mining, классификация, «естественная» классификация, систематика, KDD&DM.

1. Понятие «естественной» классификации. Данная работа является продолжением работ [Витяев Е.Е., 1983; Витяев Е.Е., Костин В.С., 1992, 2009; Витяев Е.Е., Морозова Н.С., Сутягин А.С., Лапардин К.А., 2005], посвященных понятию «естественной» классификации. В этих работах было показано, что понятие «естественной» классификации отличается от стандартно понимаемых классификаций, основанных на некоторых априорных предположениях о строении классов: гипотезе компактности, различных мерах близости, по эталонам, по суперцелям и т.д.

Как отмечалось в упомянутых работах, в отличие от упомянутых классификаций, цель «естественной» классификации состоит в познании предметной области и обнаружении структурных законов строения объектов. «Естественной» является такая классификация, которая [Забродин В.Ю., 1981]:

- выражает закон природы;
- предсказывает максимум свойств объекта по его месту в классификации;
- содержит максимум общих утверждений о каждом классе и классификации в целом;
- устойчива относительно изменений классификационных признаков.

В этих работах было показано, что все эти свойства «естественной» классификации могут быть выведены из принципа [Витяев Е.Е., 1983]: «Разбиение объектов на классы должно производиться в соответствии с закономерностями, которым удовлетворяют объекты. Точнее, объекты одного класса должны подчиняться одной группе закономерностей, а объекты разных классов разным группам закономерностей. Объекты одного класса также должны обладать некоторой целостностью. Целостность – взаимная согласованность закономерностей каждой группы по взаимопредсказанию свойств объектов. У групп закономерностей могут быть, кроме того, общие закономерности, устанавливающие связь признаков объектов из разных классов».

В данной работе рассматривается практический метод построения «естественной» классификации и алгоритмы, используемые для его реализации. Описание метода иллюстрируется на примере решения тестовой задачи. В заключение приводится пример построения «естественной» классификации символов, взятых из различных шрифтов.

2. Метод построения «естественной» классификации. Метод построения «естественной» классификации можно разбить на следующие этапы:

- Определение множества отношений, описывающих признаки объектов на языке логики первого порядка;
- Обнаружение закономерностей;
- Построение «идеальных» описаний «естественных» классов объектов.

Рассмотрим все перечисленные этапы. Для простоты понимания будем иллюстрировать работу метода на примере решения следующей тестовой задачи. В таблице (рис. 1а) представлены 60 объектов (строки таблицы). Каждый объект описывается 17-ю бинарными признаками (столбцы таблицы). Все объекты разбиты на 3 класса: первый класс – строки, содержащие «1» в столбцах с номерами 1-4, второй – строки, содержащие «1» в столбцах 5-10, и третий – строки, содержащие «1» в столбцах 11-17. Далее в таблицу было добавлено 2% равномерного случайного шума (рис. 1б). Задача заключается в том, чтобы на зашумленных данных провести классификацию всех объ-

¹ Работа поддержана грантом РФФИ 08-07-00272-а; интеграционными проектами СО РАН № 47, 115, 119, а также работа выполнена при финансовой поддержке Совета по грантам Президента РФ и государственной поддержке ведущих научных школ (проект НШ-335.2008.1)

ектов и восстановить «идеальные» описания объектов всех классов.

3. Определение множества отношений, описывающих признаки объектов. Первым этапом метода «естественной» классификации является формализация различных признаков объектов и отношений между ними в языке логики первого порядка. В результате формализации получим некоторый набор предикатов \mathbb{P} , описывающий признаки объектов. В терминах этих предикатов в дальнейшем будет осуществляться поиск закономерностей.

В рассматриваемом нами примере определим набор предикатов

$$\mathbb{P} = \{P_1, P_2, \dots, P_{17}\},$$

где $P_j(i)$, $i = 1, \dots, 60$, $j = 1, \dots, 17$ имеет вид $(x_i^j = 1)$ и означает наличие 1 в j -м столбце и i строке таблицы.

Для построения «естественной» классификации будем искать закономерности следующего вида:

$$P_{j_1}(i) \& \dots \& P_{j_m}(i) \Rightarrow P_{j_0}(i),$$

где $P_{j_k} \in \mathbb{P}$, $k = 0, \dots, m$. Правила данного вида означают, что если в столбцах j_1, \dots, j_m строки i таблицы находятся 1, то с некоторой условной вероятностью p в столбце j_0 также будет находиться 1. Условная вероятность правила $P_1 \& \dots \& P_m \Rightarrow P_0$ оценивается на обучающем множестве следующим образом:

$$p(P_0 | P_1 \& \dots \& P_m) = N(P_0 \& P_1 \& \dots \& P_m) / N(P_1 \& \dots \& P_m),$$

где $N(P_0 \& P_1 \& \dots \& P_m)$ – число событий $P_0 \& P_1 \& \dots \& P_m$ на обучающем множестве, $N(P_1 \& \dots \& P_m)$ – число событий $P_1 \& \dots \& P_m$.

Отметим, что для решения более сложных задач могут быть использованы более сложные предикаты и классы гипотез [Демин А.В., Витяев Е.Е., 2008, 2009; Демин А.В., Витяев Е.Е., Полоз Т.Л., 2007]. В работах [Демин А.В., Витяев Е.Е., 2008, 2009] описан общий способ задания различных классов гипотез для задач обнаружения закономерностей в таблицах данных. Однако для понимания работы алгоритма будет достаточно рассмотреть его работу на примере обнаружения закономерностей вида (1).

4. Обнаружение закономерностей. Алгоритм поиска вероятностных закономерностей основан на методологии семантического вероятностного вывода, описанного в работах [Витяев Е.Е., 1976, 2006; Витяев Е.Е., Москвитин А.А., 1993; Kovalerchuk B., Vityaev E., 2000; Evgenii Vityaev, 2006]. Отличительной особенностью семантического вероятностного вывода является использование понятия вероятностной закономерности, которое звучит следующим образом.

Вероятностной закономерностью называется правило $P_1 \& \dots \& P_m \Rightarrow P_0$, удовлетворяющее следующим условиям:

- 1) условная вероятность $p(P_0 | P_1 \& \dots \& P_m)$ правила определена, т. е. $p(P_1 \& \dots \& P_m) > 0$;

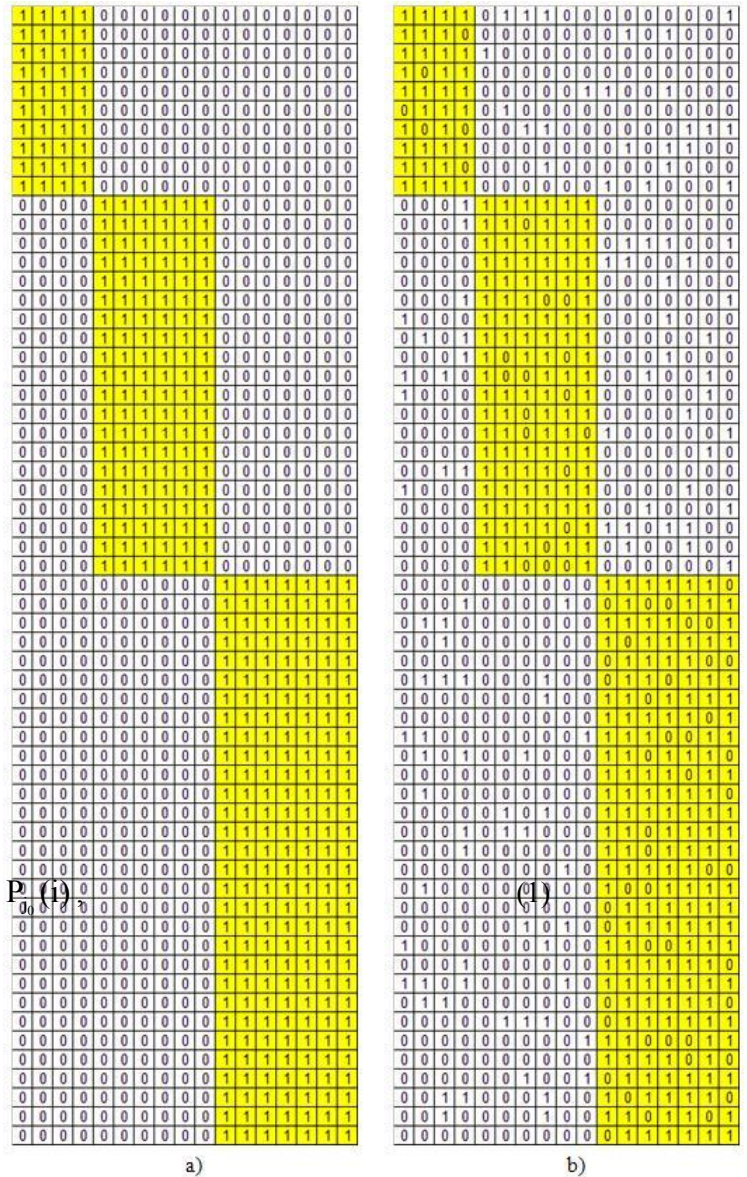


Рисунок 1. Исходная таблица данных а) и таблица с 2% шумом б).

2) условная вероятность правила строго больше условных вероятностей правил $P_{i_1} \& \dots \& P_{i_k} \Rightarrow P_0$, $\{P_{i_1}, \dots, P_{i_k}\} \subset \{P_1, \dots, P_m\}$, т.е. $p(P_0 | P_{i_1} \& \dots \& P_{i_k}) < p(P_0 | P_1 \& \dots \& P_m)$.

Семантический вероятностный вывод позволяет находить все статистически значимые вероятностные закономерности вида $P_1 \& \dots \& P_m \Rightarrow P_0$.

Для дальнейшего описания введем несколько определений. Длиной правила $R = P_1 \& \dots \& P_m \Rightarrow P_0$ будем называть величину $\text{len}(R)$, равную количеству предикатов, входящих в посылку правила. Правило $P_1 \& \dots \& P_m \& P_{m+1} \Rightarrow P_0$ является *уточнением* правила $P_1 \& \dots \& P_m \Rightarrow P_0$, если оно получено добавлением в посылку правила $P_1 \& \dots \& P_m \Rightarrow P_0$ произвольного предиката P_{m+1} . Будем обозначать $\text{Spec}(\text{RUL})$ множество уточнений всех правил из RUL , где RUL – произвольное множество правил вида (1). Опишем алгоритм поиска закономерностей, реализующий семантический вероятностный вывод:

На первом шаге генерируем множество RUL_1 всех правил единичной длины, имеющих вид $R = P_i \Rightarrow P_0$, $P_i \in \mathbb{P}$, $\text{len}(R) = 1$. Все правила из RUL_1 проходят проверку на выполнение условий 1) и 2) вероятностных закономерностей. Правила, прошедшие проверку, будут являться вероятностными закономерностями. Обозначим через REG_1 множество всех вероятностных закономерностей, обнаруженных на первом шаге, т. е. $\text{REG}_1 = \{R_i\}$, где $i \in I_1$, $R_i = P_j \Rightarrow P_0$, $P_j \in \mathbb{P}$, $\text{len}(R_i) = 1$, R_i – вероятностная закономерность.

На шаге $k \leq d$ генерируется множество RUL_k всех уточнений правил, сгенерированных на предыдущем шаге, $\text{RUL}_k = \text{Spec}(\text{RUL}_{k-1})$. Все правила из RUL_k проходят проверку на выполнение условий 1) и 2) вероятностных закономерностей. Обозначим через REG_k множество всех вероятностных закономерностей, обнаруженных на данном шаге, т. е. $\text{REG}_k = \{R_i\}$, где $i \in I_k$, $R_i = P_1 \& \dots \& P_k \Rightarrow P_0$, $P_j \in \mathbb{P}$, $\text{len}(R_i) = k$, R_i – вероятностная закономерность.

На шаге $k > d$ генерируется множество RUL_k всех уточнений всех вероятностных закономерностей, обнаруженных на предыдущем шаге, $\text{RUL}_k = \text{Spec}(\text{RUL}_{k-1})$. Все правила из RUL_k проходят проверку на выполнение условий для вероятностных закономерностей. Обозначим через REG_k множество всех вероятностных закономерностей, обнаруженных на данном шаге, т. е. $\text{REG}_k = \{R_i\}$, где $i \in I_1$, $R_i = P_1 \& \dots \& P_k \Rightarrow P_0$, $P_j \in \mathbb{P}$, $\text{len}(R_i) = k$, R_i – вероятностная закономерность.

Алгоритм останавливается, когда невозможно далее уточнить ни одно правило, т. е. когда $\text{RUL}_k = \text{Spec}(\text{RUL}_{k-1}) = \text{REG}_{k-1} = \emptyset$. Результирующее множество всех закономерностей REG будет равно объединению всех REG_i : $\text{REG} = \bigcup_i \text{REG}_i$.

Шаги алгоритма $k \leq d$ называются базовым перебором, а шаги $k > d$ – дополнительным перебором. Величина d называется глубиной базового перебора и является параметром алгоритма.

Чтобы проверить при помощи обучающего множества, является ли некоторое правило $P_1 \& \dots \& P_m \Rightarrow P_0$ вероятностной закономерностью, необходимо проверить выполнимость вероятностных неравенств 1) и 2), и оценить его статистическую значимость.

Для проверки статистической значимости правила используется статистический критерий Фишера (точный критерий независимости Фишера для таблиц сопряженности) [Кендал М., Стюарт А., 1973]. Если правило удовлетворяет этому критерию с некоторым доверительным уровнем α , а также удовлетворяет условиям 1) и 2), то оно будет являться вероятностной закономерностью.

Для нахождения всех закономерностей вида (1) мы последовательно берем каждый предикат из \mathbb{P} в качестве целевого и при помощи описанного алгоритма находим множество предсказывающих его закономерностей.

5. Построение «идеальных» описаний объектов «естественной» классификации. Следующим этапом построения «естественной» классификации является воссоздание «идеальных» описаний объектов классов. Если объекты классов обладают некоторой целостностью, то она про-

явится в структуре закономерных связей, связывающих части – признаки объектов в единое целое. Структура закономерных связей и будет определять связь частей – признаков объекта в единое целое.

Процедура идеализации сводится к дополнению описания реального объекта признаками, которые с высокой вероятностью предсказываются закономерностями, применимыми к имеющимся признакам, и, наоборот, удаляем те признаки, которые опровергаются закономерностями (предсказывается отсутствие признака), применимыми к имеющимся признакам. Эта процедура продолжается до тех пор, пока не будут включены все предсказанные признаки и не будут отсеяны все опровергающиеся. Эта процедура регулируется критерием взаимосогласованности закономерностей, который при каждом шаге должен строго увеличивать своё значение.

Определим наборы предикатов, являющиеся «идеальным» описанием классов. Для этого введем критерий взаимной согласованности закономерностей по предсказанию на этих объектах.

Будем говорить, что закономерность $P_1 \& \dots \& P_m \Rightarrow P_0$ применима к набору $\{P_{i_1}, \dots, P_{i_k}\}$, если $\{P_1, \dots, P_m\} \subseteq \{P_{i_1}, \dots, P_{i_k}\}$. Если закономерность $P_1 \& \dots \& P_m \Rightarrow P_0$ применима к набору $\{P_{i_1}, \dots, P_{i_k}\}$ и ее заключение P_0 выполнимо на этом наборе: $P_0 \in \{P_{i_1}, \dots, P_{i_k}\}$, то будем говорить, что эта закономерность *подтверждается* на этом наборе. Если закономерность применима к набору, но ее заключение не выполняется на этом наборе: $P_0 \notin \{P_{i_1}, \dots, P_{i_k}\}$, то будем говорить, что она *опровергается* на этом наборе.

Определим *критерий взаимной согласованности закономерностей* по предсказанию на наборе предикатов $\{P_1, \dots, P_m\}$ следующим образом:

$$\Gamma\{P_1, \dots, P_m\} = \sum_{R \in \Pi} p(R) - \sum_{R \in O} p(R),$$

где Π – множество закономерностей, подтверждающихся на наборе $\{P_1, \dots, P_m\}$, O – множество закономерностей, опровергающихся на этом наборе.

«Идеальным» описанием объектов «естественного» класса будем называть такой набор предикатов $\{P_1, \dots, P_m\}$, для которого критерий Γ имеет локальный максимум: при добавлении в набор или удалении из набора любого предиката значение критерия строго уменьшается.

Рассмотрим алгоритм построения «идеальных» описаний объектов класса. Пусть мы получили множество закономерностей REG при помощи описанного выше алгоритма поиска закономерностей. Задача заключается в том, чтобы, используя множество закономерностей REG, найти все наборы предикатов $\{P_1, \dots, P_m\}$, удовлетворяющие определению «идеального» описания объектов «естественного» класса.

Алгоритм построения «идеального» описания объектов класса итеративный. Началом итерации может служить произвольный набор предикатов $I = \{P_1, \dots, P_m\}$. Рассмотрим алгоритм построения «идеального» описания из некоторого начального набора $I = \{P_1, \dots, P_m\}$.

1. Рассчитываем критерий взаимной согласованности закономерностей для набора I : $\Gamma(I) = \Gamma\{P_1, \dots, P_m\}$.
 - а. Находим максимальное значение критерия при добавлении одного нового предиката к набору. Для этого мы перебираем все предикаты P из \mathbb{P} , не входящие в набор I : $P \in \mathbb{P}/\{P_1, \dots, P_m\}$. Для каждого предиката $P \in \mathbb{P}/\{P_1, \dots, P_m\}$ мы строим новый набор $I^{\text{new}} = \{P_1, \dots, P_m, P\}$ и рассчитываем для него значение критерия $\Gamma(I^{\text{new}}) = \Gamma\{P_1, \dots, P_m, P\}$. Обозначим тот набор I^{new} , на котором было получено максимальное значение критерия Γ , через I^+ .
 - б. Находим максимальное значение критерия при удалении одного предиката из набора. Для этого мы перебираем все предикаты $P \in \{P_1, \dots, P_m\}$, входящие в набор I . Для каждого предиката $P \in \{P_1, \dots, P_m\}$ мы строим новый набор $I^{\text{new}} = \{P_1, \dots, P_{m-1}\}$, $P \notin \{P_1, \dots, P_{m-1}\}$, не содержащий P , и рассчитываем для него значение критерия Γ :

$\Gamma(I^{\text{new}}) = \Gamma\{P_1, \dots, P_{m-1}\}$. Обозначим тот набор I^{new} , на котором было получено максимальное значение критерия Γ , через I^- .

2. Если $\Gamma(I) > \Gamma(I^+)$ и $\Gamma(I) > \Gamma(I^-)$, то локальный максимум критерия Γ найден, набор I будет являться «идеальным» описанием объектов класса и алгоритм останавливается. В противном случае мы переопределяем начальный набор следующим образом: если $\Gamma(I^+) \geq \Gamma(I^-)$, то $I = I^+$, если $\Gamma(I^+) < \Gamma(I^-)$, то $I = I^-$, и переходим к шагу 1.

Для того, чтобы найти все «идеальные» описания объектов классов, необходима некоторая стратегия генерации начальных наборов предикатов, из которых при помощи описанного алгоритма будут конструироваться «идеальные» описания. Нами предлагается следующий алгоритм генерации начальных наборов.

Пусть REG – множество закономерностей, полученное при помощи описанного выше алгоритма поиска закономерностей. IDL – множество найденных «идеальных» описаний, изначально $IDL = \emptyset$. Для каждой закономерности $R = P_1 \& \dots \& P_m \Rightarrow P_0$, $R \in REG$ инициализируется набор предикатов $I = \{P_1, \dots, P_m\}$, состоящий из всех предикатов, входящих в условную часть правила R . При помощи описанного выше алгоритма из набора I конструируется «идеальное» описание I_{ideal} . Если $I_{\text{ideal}} \notin IDL$, то добавляем его в IDL . Алгоритм останавливается после того как будет осуществлен перебор всех закономерностей $R \in REG$.

В рассматриваемом нами примере при использовании данного алгоритма было получено 3 «идеальных» описания объектов классов:

1. $I_1 = \{P_1, P_2, P_3, P_4\}$;
2. $I_2 = \{P_5, P_6, P_7, P_8, P_9, P_{10}\}$;
3. $I_3 = \{P_{11}, P_{12}, P_{13}, P_{14}, P_{15}, P_{16}, P_{17}\}$.

Легко видеть, что полученные «идеальные» описания полностью соответствуют трем классам объектов, заданных в таблице.

6. Классификация объектов. Каждому найденному «идеальному» описанию соответствует некоторое множество объектов, которое является «естественным» классом. Множество всех классов образует «естественную» классификацию. Рассмотрим, каким образом используя полученные классы и их «идеальные» описания осуществить классификацию некоторого объекта.

Пусть REG – множество обнаруженных закономерностей, IDL – множество найденных «идеальных» описаний, а i – некоторый объект (в рассматриваемом нами примере – некоторая строка таблицы).

Пусть $IDL = \{I_1, \dots, I_n\}$, т.е. было найдено n классов и соответствующих им «идеальных» описаний. Чтобы определить, к какому классу отнести объект i , мы для каждого класса рассчитываем критерий взаимосогласованности закономерностей данного класса на объекте i и выбираем тот класс, на котором этот критерий будет максимален. Рассмотрим эту процедуру более подробно.

Пусть $\{P_1, \dots, P_m\}$ – набор предикатов, описывающих объект i . Описанием объекта i является набор всех предикатов $\{P_1, \dots, P_m\}$, $P_j \in \mathbb{P}$, $j = 1, \dots, m$, принимающих значение «истина» на объекте i , т.е. $P_j(i) = 1$, $j = 1, \dots, m$. Пусть $REG_{I_k} \subseteq REG$ – множество закономерностей из REG , применимых к «идеальному» описанию I_k .

Для каждого «идеального» описания I_k , $k = 1, \dots, n$ мы рассчитываем критерий взаимосогласованности следующим образом:

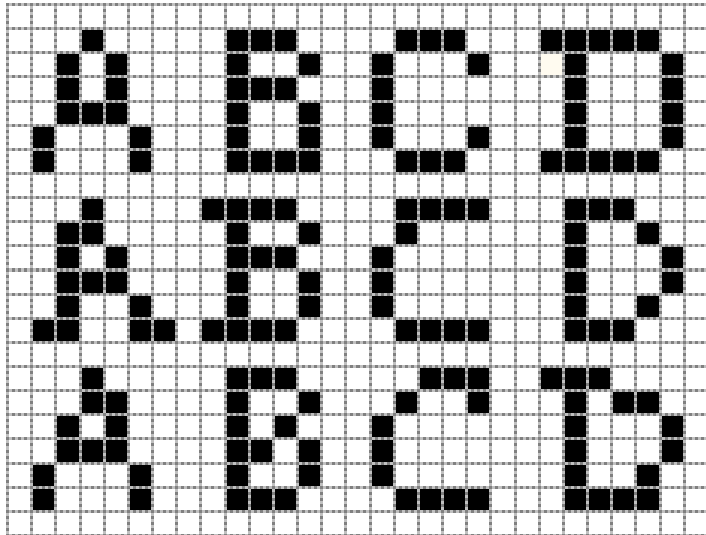


Рисунок 2. Набор символов.

$$\Gamma_{I_k}(i) = \sum_{R \in \Pi} p(R) - \sum_{R \in O} p(R),$$

где $\Pi \cup O$ – множество закономерностей из REG_{I_k} , применимых к объекту i , Π – множество закономерностей, подтверждающихся на объекте i , O – множество закономерностей, опровергающихся на объекте i . Искомым классом для объекта i будет являться тот класс, на котором критерий $\Gamma_{I_k}(i)$ будет максимальным.

В рассматриваемом нами примере данным методом классификации все строки таблицы были расклассифицированы правильно: строки 1-10 были отнесены к первому классу, строки 11-30 – ко второму, строки 31-60 – к третьему.

На данном примере также можно увидеть влияние параметра α (доверительный уровень при проверке статистической значимости правил), используемого при обнаружении закономерностей, на результат построения «естественной» классификации. Поскольку в данном примере был добавлен 2% шум, то, при использовании слишком высокого доверительного уровня, алгоритмом будет обнаружено много статистически неподтвержденных закономерностей, и, как следствие, могут быть построены ложные классы. С другой стороны, первый класс объектов (строки 1-10) представлены в таблице данных только десятью примерами, к тому же зашумленными, поэтому при использовании слишком низкого доверительного уровня, алгоритм может не обнаружить на этих примерах ни одной статистически значимой закономерности, что приведет к тому, что данный класс не будет идентифицирован. В наших экспериментах все три класса были правильно построены при использовании $\alpha = 0.005$.

7. Пример построения «естественной» классификации символов. В качестве демонстрации того, как метод «естественной» классификации группирует объекты, рассмотрим задачу классификации символов «А», «В», «С» и «D», взятых из трех различных шрифтов (рисунок 2). Все символы изображены на сетке размером 6×6 пикселей. Чтобы представить символы в виде таблицы объект-признак, мы закодировали каждый символ в виде строки, содержащей 36 позиций – по количеству пикселей в изображении. Если пиксель имел черный цвет, то в соответствующей позиции строки записывалась цифра «1», если белый – «0». Таким образом, была сформирована таблица данных, содержащая 12 строк – по числу символов, и 36 столбцов – по числу пикселей в каждом символе.

Для описания объектов нами было определено 36 предикатов, которые, по аналогии с предыдущей задачей, принимали значение «истина», если в соответствующей позиции строки находилась цифра «1».

Далее, в соответствии с описанным методом, мы извлекли из таблицы данных множество закономерностей и осуществили процедуру построения «идеальных» описаний объектов классов. В результате было получено 4 «идеальных» описания объектов и, соответственно, 4 класса. По полученным «идеальным» описаниям мы воссоздали их графическое представление, которое приве-

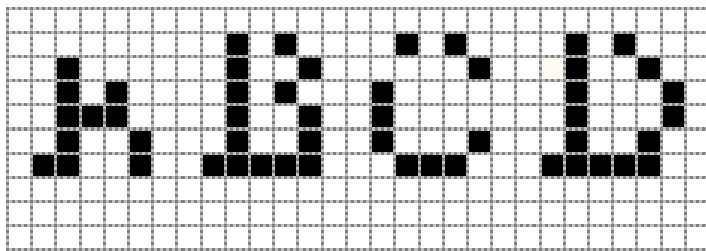


Рисунок 3. «Идеальные» описания объектов классов.

дено на рисунке 3. Как видно из рисунка, графические представления «идеальных» объектов соответствуют четырем классам символов: «А», «В», «С» и «D».

В заключении мы провели классификацию всех 12 символов. В результате все три символа «А» были отнесены к первому классу, символы «В» ко второму, «С» – к третьему, «D» – к четвертому.

Литература

- Демин А.В., Витяев Е.Е. Разработка универсальной системы извлечения знаний «Discovery» и ее применение // Вестник НГУ, серия: Информационные технологии. – 2009. – Т. 7. – Вып. 1. – С. 73-83.
- Демин А.В., Витяев Е.Е. Реализация универсальной системы извлечения знаний «Discovery» и ее применение в задачах финансового прогнозирования // Информационные технологии работы со знаниями: обнаружение, поиск, управление. – Новосибирск, 2008. – Вып. 175: Вычислительные системы. – С. 3-47.
- Демин А.В., Витяев Е.Е., Полоз Т.Л. Реализация универсальной системы извлечения знаний «Discovery» и ее применение в задачах медицинской диагностики // Труды Всероссийской конференции с международным участием «Знания – Онтологии – Теории», Том 1, Новосибирск, 2007. – С. 63–70.
- Витяев Е.Е. Классификация как выделение групп объектов, удовлетворяющих разным множествам согласованных закономерностей // Анализ разнотипных данных. – Новосибирск, 1983. – Вып. 99: Вычислительные системы. – С. 44-50.
- Витяев Е.Е., Костин В.С. Естественная классификация как закон природы // Интеллектуальные системы и методология. Материалы научно-практического симпозиума "Интеллектуальная поддержка деятельности в сложных предметных областях", Вып.4, Новосибирск, 1992. – С. 107-115.
- Витяев Е.Е., Костин В.С. Естественная классификация, систематика, онтология. Информационные технологии в гуманитарных исследованиях, Вып. 13, ИАЭТ СО РАН, Новосибирск, 2009, стр. 65-75
- Витяев Е.Е., Морозова Н.С., Сулягин А.С., Лапардин К.А. Естественная классификация и систематика как законы природы // Анализ структурных закономерностей. – Новосибирск, 2005. – Вып. 174: Выч. системы. – С. 80-92.
- Витяев Е.Е. Извлечение знаний из данных. Компьютерное познание. Модели когнитивных процессов. – Новосибирск: НГУ, 2006. – 293 с.
- Витяев Е.Е., Москвитин А.А. Введение в теорию открытий. Программная система Discovery // Логические методы в информатике. – Новосибирск, 1993. – Вып. 148: Вычислительные системы. – С.117-163.
- Витяев Е.Е. Метод обнаружения закономерностей и метод предсказания // Эмпирическое предсказание и распознавание образов. – Новосибирск, 1976. – Вып. 67: Вычислительные системы. – С.54-68.
- Забродин В.Ю. О критериях естественной классификации // НТИ, 1981. Сер.2, №8.
- Кендал М., Стюарт А. Статистические выводы и связи. – М.: Наука, 1973. – 899 с.
- Kovalerchuk B., Vityaev E. Data Mining in Finance: Advances in Relational and Hybrid methods. – Kluwer Academic Publishers, 2000. – p.308.
- Evgenii Vityaev. The logic of prediction. In: Mathematical Logic in Asia. Proceedings of the 9th Asian Logic Conference (August 16-19, 2005, Novosibirsk, Russia), edited by S.S. Goncharov, R. Downey, H. Ono, World Scientific, Singapore, 2006, pp.263-276