ЭКСПЕРТНЫЕ СИСТЕМЫ И АНАЛИЗ ДАННЫХ
(Вычислительные системы)
1991 год

УДК 519.237

ОПЫТ ПРИМЕНЕНИЯ ИДЕЙ РАСПОЗНАВАНИЯ ОБРАЗОВ В ЗАДАЧАХ КОНСТРУИРОВАНИЯ ВИРУСНЫХ ПРЕПАРАТОВ

В.А.Жуков, В.Г.Устюжанинов, В.М.Чермашенцев

§1. Цель статьи и постановка задачи

Стремление разработать методы, применимые к широкому кругу задач, требует абстрагирования от конкретной практики. Поэтому разработчика методов всегда не покидает сомнение в соответствии между жизнью и принципами, положенными в основу его методов. Ему нужны примеры внедрения его методов, чтобы утвердиться в сомнении или опровергнуть его. Нам представляется,что это не только проблема авторов этой статьи. Именно поэтому материал, носящий сугубо вирусологический характер, мы предлагаем вниманию математического журнала. На примере алгоритма КОН'ЮНКЦИЯ [2] мы покажем, как используется технология порождения логических отделителей [1-3], попав в руки конструкторов вирусных препаратов (вакцин, энтомопатогенных инсектицидов и т.д.).

Это один мотив. Другой касается алгоритма КОН'ЮНКЦИЯ и связан с нашей позицией в отношении к алгоритмам вообще. Мы считаем, что алгоритмы делятся на алгоритмы с оценками и эвристические алгоритмы. Алгоритм с оценкой снабжен методикой расчета эффективности его работы. В случае его применения доказательство удовлетворительности полученного им результата про-

водится с ее использованием. Она всегда ставит ряд ограничений на условия применения алгоритма. Например, требует управляемости выборки или нормальности закона распределения помехи и т.п. Если же такие ограничения естественным образом не возникают,то применяется эвристический алгоритм. Заключение об удовлетворительности полученного эвристическим алгоритмом результата выносит эксперт (прикладник). Общее представление об эффективности работы этого алгоритма можно составить лишь на основе ряда таких экспертиз. Поэтому и важна информация о применении алгоритма в прикладных задачах. Невозможно вне условий конкретной прикладной задачи решить, какой из нескольких эвристических алгоритмов предпочтительнее. Выбор последнего осуществляет конкретный прикладник, решая конкретную прикладную задачу. При этом мнение разработчиков эвристического алгоритма принимается к сведению, но не более того. Из теории алгоритмов с оценками (например, в задачах дискретной оптимизации) известно, что незначительное изменение в конструкции последних может привести к серьезным изменениям в их эффективности. Это замечание в полной мере относится и к эвристическим алгоритмам. Поэтому необходима осторожность в заключениях типа - такой-то эвристический алгоритм мало отличается от такого-то эвристического алгоритма. Наконец, необходимо подчеркнуть, что сам эвристиче - ский алгоритм может стать объектом теоретического изучения.Если он демонстрирует хорошую результативность, то не исключено, что за экспериментами его успешного применения стоит нераскрытое явление. Алгоритм КОН'ЮНКЦИЯ есть эвристический, и отно - ситься к нему надо по законам "жанра". Да и условия настоящей работы адекватны его применению, так как обучающая выборка в нашем случае была неуправляема и относительно мала.

§2. Алгоритм КОН'ЮНКЦИЯ

В статье [2] описан алгоритм порождения логических отде - лителей, предназначенный для решения задач распознавания. Он реализует принцип простоты. Кратко напомним,о чем шла речь в [2].

Пусть $x_1, \ldots, x_l$ - семейство переменных; $Z_1, \ldots, Z_l$ - конечные множества их значений; M_1, M_2 - конечные непересекающиеся множества объектов 1-го и 2-го классов, описываемых этими переменными. Образуем список $\mathcal{P}$ всевозможных предикатов P вида $x = a$, $x \neq a$, $x > a$, $x < a$, где $a \in Z_1 \cup \ldots \cup Z_l$. Пусть K - конъюнкция некоторых предикатов из $\mathcal{P}$ и $M_1(K), M_2(K)$ - множества объектов 1-го и 2-го классов, на которых она истинна. Требуется построить такую дизъюнкцию $D = K_{\nu_1} \vee \ldots \vee K_{\nu_s}$, состоящую из конъюнкций $K_{\nu_1}, \ldots, K_{\nu_s}$ предикатов из $\mathcal{P}$, что $M_1(D) = M_1$; число s минимально,и для всех $i = 1, \ldots, s$ имеет место $M_2(K_{\nu_i}) = \emptyset$. Это задача о построении кратчайшей дизъюнктивной нормальной формы (к.д.н.ф.) частично определенной булевой функции. Алгоритм, описанный в [2], решает ее приближенно с использованием градиентных процедур. Делается это так. Сначала из всех возможных конъюнкций выбирается их подмножество - список конъюнкций (СК); затем, используя список конъюнкций,строится к.д.н.ф. Список удовлетворяет ряду требований. Сформули - руем их.

Конъюнкцию $K(I) = \underset{i \in I}{\&} P_i$ со свойством $M_2(K(I)) = \emptyset$ назовем тупиковой, если для каждого $I' \subset I$ множество $M_2(K(I')) \neq \emptyset$.

Будем говорить, что объект $\gamma \in M_1$ покрывается конъюнкцией K, если $\gamma \in M_1(K)$.

Число $kp(\gamma) = \|\{ K/K \in CK \ \& \ \gamma \in M_1(K) \}\|$ назовем кратностью покрытия объекта γ конъюнкциями из списка конъюнкций.

Требования к списку конъюнкций:

1) для всех конъюнкций $K \in CK$ множество $M_2(K) = \emptyset$;

2) мощность $\|M_1(K)\|$ должна быть по возможности больше;

3) все конъюнкции $K \in CK$ суть тупиковые;

4) количество предикатов, входящих в конъюнкцию, должно быть по возможности меньше;

5) любая пара конъюнкций $K, K' \in CK$ должна быть такой, что $M_1(K) \setminus M_1(K') \neq \emptyset$ и $M_1(K') \setminus M_1(K) \neq \emptyset$;

6) для всех объектов $\gamma \in M_1$ кратность покрытия $kp(\gamma) \geq kp \geq 1$.

Число kp является внешним параметром программы КОН'ЮНКЦИЯ и задается пользователем. Для удовлетворения первых четырех требований при построении конъюнкции K из списка конъюнкций используется градиентная процедура с возвратом. Конкретно фиксируется число $0 \leq \alpha \leq 1$. Среди предикатов из $\mathcal{P}$ разыскивается предикат P_{ξ_1}, для которого величина

$$\alpha \|M_1(P_{\xi_1})\| + (1-\alpha) \|M_2 \setminus M_2(P_{\xi_1})\| \to \max.$$

Затем к нему присоединяется предикат P_{ξ_2}, для которого

$$\alpha \|M_1(P_{\xi_1} \& P_{\xi_2})\| + (1-\alpha) \|M_2 \setminus M_2(P_{\xi_1} \& P_{\xi_2})\| \to \max,$$

и так далее. Заканчивается этот процесс построением конъюнкции $K'' = P_{\xi_1} \ \dots \ P_{\xi_s}$, когда оказывается, что $M_2(K'') = \emptyset$. Выделяя из конъюнкции K'' тупиковую, получаем требуемую конъюнкцию K. Для этой конъюнкции и всех наработанных к этому моменту конъюнкций $K' \in CK$ проверяется требование 5.

Если оно удовлетворяется, то конъюнкция K пополняет список конъюнкций. Требование 6 обеспечивается специальной процедурой обхода конъюнкции.

Для построения к.д.н.ф. с использованием списка конъюнкций также применяется градиентная процедура. А именно, среди всевозможных конъюнкций разыскивается $K_{\nu_1} \in CK$ со свойством $\|M_1(K_{\nu_2})\| \to \max$. К ней присоединяется конъюнкция $K_{\nu_2} \in CK$ со свойством $\|M_1(K_{\nu_1}) \cup M_1(K_{\nu_2})\| \to \max$ и так далее. Процесс присоединения продолжаетсся до тех пор, пока не будет достигнуто выполнение равенства $M_1 = \bigcup_{i=1}^{s} M_1(K_{\nu_i})$.

Теперь короткий комментарий к сказанному. Возникновение задачи построения к.д.н.ф. применительно к распознаванию образов восходит к [3]. Там же был описан алгоритм КОРА. Аналогичная задача, но в более узком классе д.н.ф., чем в настоящей работе и в [3], решалась в [1]. Оба источника дают алгоритмы, отличные от КОН'ЮНКЦИИ. Не известно, какой алгоритм лучше, так как все они эвристичны.

По идее [2] дизъюнкция D должна была использоваться для решения задачи распознавания. Считалось, что если D истинна на некотором объекте, то алгоритм относится к 1-му классу. Если же D на нем ложна, то - ко 2-му классу. Но биологи рассудили по-своему. Они использовали программу КОН'ЮНКЦИЯ как генератор конъюнкций. Это несложно сделать, варьируя минимальную кратность покрытия (число kp) и вынуждая программу генерировать обширные списки конъюнкций. Наработав их в значительном количестве, они воспользовались ими для определения степени влияния переменных $x_1, \ldots, x_l$ на значения классообразующего признака. Выделив наиболее влиятельные среди них: $x_{\zeta_1}, \ldots, x_{\zeta_q}$ и зафиксировав их, они затем

провели по $x_{\xi_1}, \ldots, x_{\xi_q}$ оптимизацию и решили задачу построения биопрепарата с требуемыми свойствами. Заметим, что их успех во многом определялся требованиями 1-5, предъявляемыми к списку конъюнкций.

§3. Поиск оптимальных защитных сред для микроорганизмов

Распространенным методом приготовления вирусных препаратов является лиофильное обезвоживание предварительно замороженной вирусной суспензии (высушивание в вакууме). Процесс высушивания, как правило, негативно сказывается на активности вируса. Она падает. Для уменьшения ее падения используются специальные защитные среды, компоненты которых вносятся перед замораживанием.

Литературные данные свидетельствуют о большом разнообразии эффективных комбинированных сред: биологические вещества и продукты их переработки, углеводы, естественные и синтетические полимеры и белки. Каких-либо общих закономерностей в соотношениях отдельных компонент сложных сред высушивания не установлено. В связи с этим актуальной является задача поиска эффективных защитных сред, унифицированных для применения к возможно большему кругу вирусов. Возможность существования таких сред - неспецифических - была отмечена в работе [4]: "Экспериментальные данные показывают, что очень часто среда высушивания, хорошо защищающая один вид микроорганизмов, эффективна и для других видов". Возникает задача поиска неспецифических сред для как можно большего круга вирусов. Опишем нашу методику выбора компонент защитной среды.

Данные по влиянию состава защитных сред на активность вирусов были собраны в таблице $\tilde{A}$. Она состояла из столбца Y и таблицы $A = (a_{ij})_{m \times r}$ с элементами $a_{ij} \in \{0,1\}$ и размерами $m = 503$ и $r = 40$. Строке i соответствовал эксперимент Y_i, а столбцу j - компонента x_j. Элемент a_{ij} =

= 1, если в эксперименте γ_i в состав защитной среды входила компонента x_j, и $a_{ij} = 0$ - в противном случае. На пересечении строки i и столбца Y находилась величина $y_i \in [0, \infty)$, равная падению активности вируса в результате лиофилизации в эксперименте γ_i. (Для измерения y использовались стандартные методики из [5].) Все значения величины y из таблицы данных $\tilde{A}$ умещались в интервале от 0 до 5.0. В этом интервале были выделены два подынтервала: $Y_1 = [0; 0{,}5]$ и $Y_2 = [2.1; 5.0]$. Были сформированы два класса экспериментов: $A_1 = \{\gamma_i / y_i \in Y_1\}$ и $A_2 = \{\gamma_i / y_i \in Y_2\}$. Полагая $M_1 = A_1$ и $M_2 = A_2$, к классам M_1 и M_2 применяем программу КОН'ЮНКЦИЯ и нарабатываем список конъюнкций $\mathcal{K}_1$. После этого роли A_1 и A_2 менялись: M_1 отождествлялся с классом A_2, а M_2 - с классом A_1. К новым классам M_1, M_2 вновь применялась программа КОН'ЮНКЦИЯ. В результате получался список конъюнкций $\mathcal{K}_2$. С помощью списков $\mathcal{K}_1$ и $\mathcal{K}_2$ оценивалось влияние на эффективность защитной среды присутствия в ее составе фиксированной смеси компонент. Делалось это так.

Пусть $I = \{\nu_1, \ldots, \nu_s\}$ - список номеров фиксированной смеси компонент. В нашем случае любую конъюнкцию K можно было записать в виде $K = \underset{j \in J_1}{\&} (x_j = 1) \ \underset{g \in J_0}{\&} (x_g = 0)$, где пересечение $J_1 \cap J_0 = \emptyset$. Будем говорить, что смесь компонент I участвует в конъюнкции K, если $I \subseteq J_1$. Обозначим через $S_1(I)$ число конъюнкций $K \in \mathcal{K}_1$, в которых участвует смесь компонент I. Пусть $l_i = \| \mathcal{K}_i \|$. Задавались числа $0 \le c_2 < c_1 \le 1$. Смесь компонент I считалась влияющей на эффективность защитной среды, если для нее выполнялись неравенства $S_1(I)/l_1 \ge c_1$ и $S_2(I)/l_2 \le c_2$. Такие смеси будем называть эффективными. Заметим, что на практике число c_2 почти всегда полагалось равным нулю. Приведен-

ный способ выявления эффективных смесей не дифференцирован по родам вирусов. Для того чтобы выявить смеси компонент, наиболее эффективные при работе с вирусами конкретного рода, мы поступали так. Фиксировался род вируса. Из классов экспериментов A_1 и A_2 выделялись подклассы A_1' и A_2' экспериментов с вирусами фиксированного рода. Пусть $A'(K)$ - множество экспериментов из $A_1' \cup A_2'$, на которых конъюнкция K истинна. Составлялись списки конъюнкций $\mathcal{K}_1' = \{K/K \in \mathcal{K}_1 \& A'(K) \cap A_1' \neq \emptyset\}$. Поиск смесей компонент, наиболее эффективных для вирусов интересующего нас рода, осуществлялся по описанной выше методике, но уже со списками конъюнкций $\mathcal{K}_1'$ и $\mathcal{K}_2'$.

Этот подход был применен нами к массиву фактографических литературных данных (см., например, [4,6]). Были представлены роды вирусов: тога-, фила-, руби-, парамиксо-, пикорно- (энтеро-, рино-, афто-), рабдо-, бунья-, покс-, герпес-, папова-, -арено- и реовирусы. Таблица данных содержала информацию о падении функциональной активности вируса после лиофилизации, а также о наличии в экспериментальной защитной среде компонент: сахарозы, желтка, желатина, пептона, аллантоисной жидкости, куриного белка, обезжиренного молока, декстрина, Д-глюкозы, яичного (или человеческого) альбумина, куриного желтка, бычьей (или лошадинной) сыворотки, агар-агара, альбумина, гемацела, лактозы, глюконата натрия, лактобионата, сорбита, поливинилпирролидона, аминопептида, крахмала, гидролизата, лактальбумина, глицина, гидролизата казеина, среды 199, сорбита, маннита, хлористого магния, BSM, PDA, фенола, протеина, гидролизата, фибрина, PBS, DMCO, SPGA, SPG, SDG, метилцеллюлозы. В результате обработки списков конъюнкций $\mathcal{K}_1$ и $\mathcal{K}_2$ были обнаружены следующие закономерности (состав фиксированной смеси компонент перечисляется в фигурных скобках):

1. Компоненты альбумин, пептон, желатин, сахароза, обез - жиренное молоко в конъюнкциях списка $\mathcal{K}_2$ не участвуют.

2. Смесь {сахароза, желатин} эффективна независимо от рода вируса.

3. {Пентон} и {желатин} эффективны для оболочечных РНК-содержащих вирусов (ортомиксо-, парамиксо-, тогавирусов);{пептон} эффективен для покс- и герпесовирусов и не эффективен для пикорновирусов.

4. {Альбумин} эффективен для бунья, ортомиксо- и тогавирусов.

5. Смесь {сахароза, желток} эффективна для парамиксо-, тога-, пикорновирусов.

6. {Альбумин} и смесь {пептон, желатин} эффективны для парамиксо-,тога-, пикорно-, рео-, поксвирусов.

7. {Пептон} и {желток} эффективны для орто-, парамиксо-, тогавирусов.

На основании этих закономерностей были выделены компоненты защитных сред, обладающих хорошим защитным эффектом при лиофилизации вирусов различных родов, а именно: пептон, альбумин, сахароза, желатин, желток и некоторые их комбинации: {сахаро - за, желток}, {сахароза, пептон}, {сахароза, желатин}.

Полученные результаты свидетельствуют в пользу существо - вания неспецифических защитных сред. В качестве таковой для экспериментальной проверки была выбрана контрольная среда, которая состояла из сахарозы, глюкозы, альбумина и желатина, выполнявших функцию защиты вируса от внешних воздействий, а также наполнителя - глицина. Из литературы было известно, что применение этой среды для защиты вируса гриппа A/PR- 8/34 (род - ортомиксовиридае) дало хорошие результаты. Были известны и кон - центрации компонент контрольной среды в экспериментах с грип - пом. С целью проверки ее неспецифичности она была применена для защиты вируса ВЭЛ (штамм ТС-83) (род - тоговиридае). Был

проведен эксперимент по оптимизации концентраций компонент контрольной среды. При этом достигалась минимизация падения активности вируса ВЭЛ, наступавшая вследствие фазы лиофилизации. Результатом была успешная защита вируса ВЭЛ. При этом полученные оптимальные концентрации оказались близки к концентрациям компонент, которые использовались для защиты вируса гриппа. Планирование эксперимента и статистический анализ результатов проводился по методикам [7] сотрудниками ВНИИ молекулярной биологии Фроловым В.Г. и Гусевым Ю.М.

Заключение

Существующий на сегодняшний день опыт анализа литературных данных по защитным средам при конструировании вирусных препаратов во многом зависит от искусства исследователя. Объем их велик, и они необозримы, что содержит в себе вероятность ошибочного вывода. В конце концов это приводит к материальным потерям. В этих условиях использование программы КОН'ЮНКЦИЯ позволяет перейти к новой более точной механизированной технологии выявления оптимальных комбинаций добавок, дает доказательство правильности их выбора.

Литература

1. ЛБОВ Г.С., КОТЮКОВ В.И., МАШАРОВ Ю.П. Метод обнаружения логических закономерностей на эмпирических таблицах //Эмпирическое предсказание и распознавание образов. - Новосибирск, 1976. - Вып. 67: Вычислительные системы. - С. 29-42.

2. УСТЮЖАНИНОВ В.Г. Конъюнкция - программа построения логических отделителей //Анализ разнотипных данных.- Новосибирск, 1983. - Вып. 99: Вычислительные системы. -С. 117-119.

3. БОНГАРД М.М. Проблемы узнавания: -М.: Наука, 1967.

4. ДОЛИНОВ К.Е. Основы технологии сухих препаратов. - М.: Медицина, 1969. - 343 с.

5. АШМАРИН П.П., ВОРОБЬЕВ А.А. Статистические методы в микробиологических исследованиях. - Л., 1962.

6. СЕЛЕЗНЁВА А.Ю. Влияние условий лиофилизации и длительного хранения на инфекционную активность вирусов различных групп: Автореф. дис...канд.мед.наук: 03.00.6. -М., 1979.-20 с.

7. АДЛЕР Ю.П. и др. Планирование эксперимента при поиске оптимальных условий. - М.: Наука, 1976.

Поступила в ред.-изд.отд.
12 мая 1991 года