ПАРАЛЛЕЛЬНЫЙ АЛГОРИТМ ОБУЧЕНИЯ ДЛЯ ИНТЕРАКТИВНОГО ПОЛИТОМИЧЕСКОГО ОПРЕДЕЛИТЕЛЯ БИОЛОГИЧЕСКИХ ВИДОВ

А.Т. Вахитов, О.Н. Граничин, А.Г. Кирейчук, А.Л. Лобанов

Идентификация таксономической принадлежности биологического объекта по его характеристикам, доступным наблюдателю, является важной задачей, имеющей много практических приложений. В последнее время было разработано значительное число программных продуктов, призванных так или иначе облегчить решение этой задачи [1-3]. Все они работают по общему алгоритму:

- Предлагается набор вопросов;
- Наблюдателем выбирается один вопрос и дается ответ на него;
- Переходим к шагу 1 до тех пор, пока не останется 0 или 1 таксон.

Если под ответы наблюдателя подходит не один таксон или таксон, включающий несколько низших таксонов, то следует перейти к следующему набору вопросов. А в случае, когда ответ на вопрос позволяет установить определенный таксон самого нижнего уровня, определение завершается.

Важной характеристикой определителя или ключа является число входов в ключ, т.е. число признаков, с которых можно начать новый диагноз или очередной его шаг. Обычно выделяют одновходовые ключи, в которых у пользователя нет выбора - как на первом шаге, так и на последующих, и он должен пользоваться единственным предъявленным ему признаком; и многовходовые ключи, в которых на каждом шаге пользователю предоставляются несколько признаков и он выбирает из них наиболее удобный и надежный. Другой важной характеристикой является число состояний, выделяемых в ключе для каждого признака. Обычно по этому основанию классификации выделяют дихотомические ключи, в которых во всех признаках имеются строго два состояния; и политомические, в которых хотя бы в части признаков могут быть три и более состояний.

Биологические классификации и базы данных на их основе насчитывают сотни, тысячи и более различных таксонов низшего уровня, характеризующиеся иногда сотнями признаков. Для компьютерных определителей на основе таких баз данных очень больших объемов очень важной функцией является ранжирование вопросов, подсказывающее наблюдателю те вопросы, ответы на которые были бы наиболее полезны. Полезность может определяться минимизацией общего числа задаваемых системой вопросов об объекте, или степенью надежности определителя и минимизацией ошибок в определении. В этой статье будет подробнее рассмотрена проблема минимизации числа задаваемых вопросов при заданном уровне надежности.

Для известного множества таксонов поиск минимальной последовательности вопросов, разделяющей каждую пару из этого множества, является NP-полной задачей [4]. Поэтому при описанных объемах биологических баз данных оптимальную последовательность вопросов найти не удается и актуальной является разработка субоптимальных алгоритмов.

Данная работа посвящена разработке алгоритма с элементами обучения, реализованного в виде Грид-приложения, для поиска близких к минимальным последовательностей вопросов. Алгоритм опирается на нейронную сеть, данные для обучения которой получаются в результате параллельного анализа базы данных таксонов. Обучение нейронной сети производится на небольших подмножествах таксонов приемлемого размера для получения оптимального решения. Поиск минимальных последовательностей вопросов для этих подмножеств является хорошо параллелизуемой задачей, что и используется в описанной в этой статье системе. В следующих разделах будет подробнее рассмотрен алгоритм оптимального решения задачи, проанализированы известные автору подходы к построения субоптимального решения задачи, описана система и сделано сравнение результатов, доставляемых ею, и результатов работы известных алгоритмов.

Оптимальное решение. Для определения принадлежности объекта какому-то таксону необходимо удостовериться в том, что никакой другой таксон не характеризуется признаками этого объекта. Каждая пара таксонов i и j в определителе отделима с помощью кого-либо признака. Обозначим набор признаков, отделяющих i от j, как $c^{ij}_1,\ldots,c^{ij}_{kij}$. Пусть T общее число таксонов. Рассмотрим следующую формулу булевой логики, приняв c^{ij}_k за переменные со значениями *истина* и *ложь*, отражающими, выбран ли признак в ходе конкретного процесса определения:

$$(c^{i1}_1 \wedge c^{i1}_2 \wedge \ldots \wedge c^{i1}_{ki1}) \vee (c^{i2}_1 \wedge c^{i2}_2 \wedge \ldots \wedge c^{i2}_{ki2}) \vee \ldots \vee (c^{iT}_1 \wedge c^{iT}_2 \wedge \ldots \wedge c^{iT}_{kiT}). \qquad (1)$$

Для того, чтобы таксон был отделен ото всех остальных, необходимо и достаточно, чтобы формула принимала значение *истина* при подстановке значения *истина* в переменные, соответствующие выбранным признакам, и *ложь* в переменные, соответствующие не выбранным признакам.

Для построения аналогичной формулы для группы таксонов, нужно перебрать все пары таксонов и соединить знаком конъюнкции формулы для каждой пары таксонов.

Задача построения кратчайшей определяющей последовательности для таксона эквивалентна нахождению минимального по числу элементов набора переменных (c^{ij}_k), обращающего формулу в истинную

при подстановке значения *истина* в переменные набора. Найти такую последовательность можно только преобразовав формулу (1), записанную изначально в конъюнктивной нормальной форме, в дизъюнктивную нормальную форму. Эта задача NP-полна, и точное решение ее, как выше было сказано, невозможно.

Существуют эвристики, позволяющие на практике сильно сократить число перебираемых вариантов [5]:

1. Если cijk1 для какого-то k1 является подмножеством cijk2, то из формулы можно исключить дизъюнкцию (ci21 ^ ci22 ^ … ^ ci2ki2)
2. Если cij1 входит в формулу только в тех же дизъюнкциях, что и cij2, то можно исключить из формулы все вхождения cij1.

Несмотря на использование эвристик, задача построения оптимального решения для достаточного объема таксонов является слишком вычислительно сложной. Для построения субоптимальных решений используются алгоритмы, описанные в следующем разделе.

Известные подходы. Традиционно для построения определителя применяется подход на основании расчета энтропии распределения возможных ответов для каждого из вопросов (*признаков*) [6]. Иными словами, для каждого вопроса C_i множество таксонов T делится на возможно пересекающиеся подмножества $\{t_{i1},...,t_{ik}\}$, соответствующие возможным ответам (используется термин *состояния*) на данный вопрос. Далее производится ранжирование вопросов по критерию максимизации энтропии $E_i=-(t_{i1}/T \ln t_{i1}/T + t_{i2}/T \ln t_{i2}/T +... + t_{ik}/T \ln t_{ik}/T)$. Этот подход и родственные ему, имеющие в основе другие формулы для определения ранга вопроса, например $L_i = -(t_{i1}\text{^}2 + ... + t_{ik}\text{^}2)$, не учитывают возможные особенности задачи, связанные с наличием взаимосвязей между признаками, отражающихся на распределении ответов. Более подробно, поскольку нужно с помощью признаков отделить каждый таксон от каждого, признак необходимо использовать, если только он может отделить один из таксонов.

Для подтверждения этого ограничения рассмотрим следующий пример.

Пример. Пусть таксоны появляются на входе определителя равновероятно и матрица определения выглядит следующим образом:

Признак\Таксон	α	B	Г	δ	M
A	1	2	3	4	1,2,3,4
B	1	1	1	1	2

Для признака A $L_A=-16$, $L_B=-17$ ($E_A=-1{,}6 \ln 0{,}4 \approx 1{,}46$, $E_B=0{,}8 \ln 0{,}8 \approx 0{,}17$). При этом первым будет всегда задаваться вопрос A, затем вопрос B. Поскольку вопрос A не отделяет ни один таксон от таксона μ, получается, что всегда необходимо задавать 2 вопроса. В то же время, если первым задавать вопрос B, то в 25% случаев можно ограничиться одним вопросом. Оптимальный алгоритм сделает вывод о том, что B должен задаваться первым. В реальной системе оптимальный алгоритм не может быть использован напрямую, что мотивирует исследователей предлагать новые способы решения проблемы, один из которых представляет настоящее исследование.

Другим известным эвристическим методом решения задачи построения минимальной определяющей последовательности является алгоритм построения градиентного покрытия [4]. Применительно к рассматриваемой задаче, алгоритм заключается в ранжировании признаков по числу вхождений в формулу (1). Он также отдал бы предпочтение в рассматривавшемся примере признаку A.

Предлагаемое решение. Авторы предлагают испольовать обучаемую нейронную сеть, обучающая последовательность для которых будет генерироваться параллельно на большом числе независимых вычислителей.

Для построения обучающей последовательности выбирается большое число небольших случайных выборок таксонов, и для них находится оптимальные решения задачи. Затем из них формируется обучающая последовательность, подаваемая на вход нейронной сети, ранжирующей признаки.

При работе с определителем на вход сети подается вектор из вероятностей p_i того, что i-й таксон является искомым. На выходе сеть выдает ранги r_j, определяющие позицию признака с номером j в списке.

Нейронная сеть. Предлагается использовать нейронную сеть для определения ранжирования признаков. На входе сети находится набор вероятностей того, что наблюдаемый объект принадлежит тому или иному таксону. На выходе сеть выдает ранги признаков, на основании которых строится список признаков, выдаваемый наблюдателю.

Нейронная сеть является многослойным персептроном [7]. Она состоит из 3х слоев, входного, число нейронов в котором равно числу таксонов, выходного, с числом нейронов, равным числу признаков, и промежуточного.

Обучение сети производится на основании выделения малых подмножеств таксонов, для которых еще можно найти наименьшую разделяющую последовательность признаков оптимальным алгоритмом. Подмножества таксонов формируются как случайные выборки из множества всех таксонов с равномерным распределением вероятностей. После этого производится обучение сети на серии примеров таких подмножеств.

Алгоритм обучения — это обратное распространение ошибки [7], метод стохастической оптимизации с пробным одновременным возмущением [8], метод имитации отжига [9] либо сочетание последнего метода и одного из первых двух.

Реализация. В СПБГУ на математико-механическом факультете на несколько десятков компьютеров установлены клиентские системы Грид на основе набора инструментов GPE [10]. Ведутся работы по установке и развертыванию Грид Condor, ставшего де-факто стандартом настольных Грид-вычислений [11]. В рамках этих инфраструктур реализуется и описываемое приложение.

Применение системы. Система разработана как элемент существующего программного обеспечения для исследований по биоразнообразию, включающего в себя средство для заполнения баз данных с возможностями проверки корректности и надежности данных, и средства интерактивного определения, доступного в Интернет. В настоящее время она развивается на портале ЗИН РАН для зоологических объектов. Однако она представляет собой достаточно универсальное средство, применимое для любых биологических объектов.

Сейчас посредством специально разработанной оригинальной программы (Swingfiller) создаются и наполняются базы данных по таксонам различных групп животных, объединяющие информацию о систематической иерархии таксонов, различных состояний диагностических признаков у каждого из них и иллюстрациями, включающими рисунки и фотографии как в этих базах данных, так и в Интернете. В качестве основы для проверки определяемых объектов будут использоваться атласы видов на сайтах портала ЗИН РАН. К настоящему времени уже создан крупнейший в мире Интернет-атлас по жесткокрылым насекомым <http://www.zin.ru/Animalia/Coleoptera/rus/atlas.htm;

http://www.zin.ru/Animalia/Coleoptera/eng/atlas.htm>, который преимущественно включает виды, встречающиеся на территории России и сопредельных стран.

В дальнейшем планируется интеграция баз данных определителя с уже имеющимися Интернет-страницами атласов на портале ЗИН РАН, а также создание новых страниц атласов по другим группам животных и интеграция их с информационными ресурсами портала ЗИН РАН и другими источниками, чтобы в результате определения наблюдатель получал максимальный объем полезной информации относительно таксона, к которому принадлежит определяемый объект. Предполагается также, что создаваемая система станет наиболее полным и разносторонним справочным источником по группам животных, для которых будут созданы соответствующие определители, интегрирующие всю информацию по каждой из групп.Работа поддержана грантом РФФИ 09-04-00789-а.

ЛИТЕРАТУРА:

1. М.Б. Дианов, А.Л. Лобанов PICKEY - Программа для определения организмов
2. с интерактивным использованием изображений // Базы данных и компьютерная графика в зоологических исследованиях (Труды Зоологического института, т. 269). 1997. С. 35-39.
3. А.Л. Лобанов, А.Г. Кирейчук, И.С. Смирнов, А.Т. Вахитов, М.Б. Дианов К реализации идеального интерактивного определителя биологических объектов в Интернете. Материалы Всероссийской научной конференции "Научный сервис в сети ИНТЕРНЕТ: технологии параллельных вычислений" (18-23 сентября 2006 г., г. Новороссийск). 2006. С. 202-204.
4. M.J. Dallwitz, T.A. Paine and E.J. Zurcher. 2000 onwards. Principles of interactive keys. http://delta-intkey.com
5. С.А. Ложкин. Лекции по основам кибернетики. М.: Издательский отдел ф-та ВМиК МГУ, 2004. 251 с.
6. R. Valdes-Perez, V. Pericliev, F. Pereira. Concise, Intelligible, and Approximate Profiling of Multiple Classes // Int. J. of Human-Computer Studies. Vol. 53, No. 3, September 2000. Pp. 411-436.
7. А.В. Свиридов. Ключи в биологической систематике: теория и практика. М.: Изд-во Московского Университета, 1994. 224 с.
8. С. Осовский. Нейронные сети для обработки информации. М.: Финансы и статистика. 2004. 344 с.
9. О.Н. Граничин, О.А. Измакова. *Рандомизированный алгоритм стохастической аппроксимации в задаче самообучения* // Автоматика и телемеханика, 2005, № 8, с. 52-63.
10. S. Kirkpatrick, C.D Gelatt Jr., and M.P. Vecchi. Optimization by Simulated Annealing // Science, 220. pp. 671-680. 1983.
11. Grid Programming Environment Documentation (http://gpe.sourceforge.net)
12. D. Sepulveda, S. Goasguen. The Deployment and Maintenance of a Condor-Based Campus Grid. In Proceedings of Grid and Pervasive Computing (GPC) 2009 Conference, Geneva, 3-7 May 2009, pp. 165-176.