КЛАССИФИКАЦИОННЫЙ ПОДХОД К ВЫРАБОТКЕ МЕДИЦИНСКОГО ДИАГНОЗА ПО РЕЗУЛЬТАТАМ ПОДГОТОВЛЕННЫХ АНАЛИЗОВ

Корноушенко Е. К.

Институт проблем управления им. В.А. Трапезникова РАН,
Россия, г. Москва, ул. Профсоюзная д.65
ekorno@ipu.ru

Аннотация: Предлагается алгоритмическая процедура выработки медицинского диагноза по результатам анализов, удовлетворяющих заданным требованиям. Изложение ведется на примере конкретной выборки Dermatology из известного репозитория UCI Machine Learning. Полученные результаты свидетельствуют о том, что предлагаемая процедура может быть полезным помощником врачу при выработке окончательного диагноза.

Ключевые слова: вид заболевания, анализ, вес анализа, «профиль» пациента, «портрет» пациента.

Введение

Применение компьютерных методов в медицинском диагностировании составляет отдельную ветвь в медико-биологических исследованиях, характеризующуюся использованием новых технологий при исследовании чрезвычайно сложных объектов в условиях неполноты и неоднозначности получаемой информациии, затрудняющих построение формальных моделей и компьютерных программ медицинского диагностирования. Подобные программы известны лишь для отдельных случаев [1], характеризующихся тем, что в исследуемых случаях можно выделить классы с известными значениями входящих в них переменных, а при дальнйшем исследовании подобного случаев такие классы сохраняются при изменении значений входящих в них переменных. В последнее время появились публикации по компьютерному диагностированию в случаях принципиального остутствия каких-либо сохраняющихся априорных классов. В таких случаях выработка диагноза для пациента сводится к распознаванию структуры того или иного класса с привлечением результатов соответствующих обследований пациента. В литературе такие диагнозы получили название классификации без обучения (*unsupervised classification*) [2]

В данной работе рассматривается процедура медицинского диагностирования по полученным результатам так называемых подготовленных анализов (т.е. анализов, удовлетворяющих определенным требованиям – см. ниже). Каждый этап этой процедуры предусматривает выполнение определенных действий над получаемой информацией, при этом в силу её неполноты и неоднозначности невозможны какие-либо прогнозирования окончательного результата. Предлагаемая процедура является, таким образом, процедурой классификации без обучения. Для простоты изложения действенность этой процедуры показывается на примере конкретной выборки Dermatology из известного репозитория UCI Machine Learning [3]. Полученные на этой выборке результаты свидетельствуют о том, что предлагаемая процедура могла бы стать полезным помощником врачу при определении им окончательного диагноза.

1 Исходная информация, содержащаяся в выборке Dermatology

Выборка Dermatology содержит результаты 33 клинических и гистопатологических анализов[110], проведенных над 343 пациентами с целью определения (классификации) у них тех или иных кожных заболеваний из 6 видов возможных заболеваний. Более подробная информация об используемых анализах, представляющая интерес для читателя, содержится в [4, 5]. В частности, в [4] указано, какие анализы являюся клиническими, а какие – гистопатологическими. Результаты каждого из анализов оцениваются качественно в единой для всех анализов шкале 0, 1, 2, 3. Выборка содержит также результаты классификации (диагноз) для каждого из 343 пациентов. Подчеркнем, что эти диагнозы никоим образом не используются в предлагаемой процедуре диагностирования, а служат лишь для оценки точности классификации самой процедуры. Базисным понятием в процедуре является понятие подготовленных анализов, положенное в основу анализа взаимно многозначных отображений типа «вид заболевания – применяемые анализы». Это понятие предусматривает наложение на совокуность используемых анализов следующих. требований: а) для каждого вида заболевания должно быть определено множество анализов, допустимых для этого вида; б) анализы в каждом таком множестве ранжируются по «важности» для соответствующего вида заболевания. В работе [4] эти требования выполняются в виде приведенной таблицы (см. далее табл. 1), в которой

[110] *По причинам, не относящимся к данной работе, в [4] для диагностирования используются лишь 26 анализов (с сохранением их прежней нумерации).*

каждому виду заболевания соответствуют 16 возможных анализов и эти анализы упорядочены по «важности» для каждого заболевания.

2 Формальная процедура выработки диагноза (классификация)

Введем следующие обозначения: a_i – i-й анализ, i = 1,…, 33, d_j – j-й вид заболевания, j = 1,…, 6, $C(d_j)$ – множество анализов, относящихся к d_j (см. табл. 1). Совокупность множеств вида $C(d_j)$, содержащих анализ a_i, обозначим как $S(a_i)$, число таких множеств в $S(a_i)$ характеризует степень неоднозначности N_i анализа a_i Каждому упорядоченному множеству $C(d_j)$ сопоставим вектор $R(d_j)$, в котором s–й координате приписывается ранг $r(s, d_j)$, равный порядковому номеру этой координаты в $R(d_j)$, так что на каждом множестве $C(d_j)$ вводится ординальное ранжирование.

Пусть $a_i \in C(d_j)$ и a_i является s–й координатой в $R(d_j)$. Считаем, что *важность* $q(a_i, d_j)$ анализа a_i для заболевания d_j обратно пропорциональна рангу $r(s, d_j)$, умноженному на N_i: $q(a_i, d_j)= 1/ r(s , d_j)N_i$.

Предположим, что анализ ai имеет некоторое ненулевое значение $\hat{a}_i$ в шкале 1, 2, 3. Вес w(ai, dj) анализа ai при заболевании dj определим как

(1) $\quad w(ai, dj) = \hat{a}_i q(a_i, d_j).$ x_M^*

Пусть pk – k-й пациент из выборки Dermatology, которому соответствует k-я строка в этой выборке и nk ненулевых анализов в этой строке. Множество Pk ненулевых анализов в k-й строке выборки назовем профилем k–го пациента, а множество значений ненулевых анализов из Pk обозначим через $\hat{P}_k$. Каждый анализ из Pk назовем компонентой профиля Pk. Вес заболевания dj у k–го пациента определим как

(2) $$w(p_k, d_j) = \sum_{i=1}^{n_k} \hat{a}_i q(a_i, d_j) \,|\hat{a}_i \in \hat{P}_k) \,.$$

Замечание к формулам (1) и (2). В теории принятия решений вопросы ранжирования и взвешивания критериев (в данном случае – анализов) занимают важное место (см., например, [6], где рассматриваются различные варианты таких процедур). Согласно [6], веса, приписываемые ранжированным переменным (ранжированные веса), имеют наиболее важные свойства. Именно такими весами являются веса, приписываемые согласно (1) ранжированным анализам. Отличие формул (1) и (2) от ивестных определений весов в том, что данные формулы носят условный характер, т.е. они определены по отношению к соответствующей альтернативе (в данном случае – к конкретному виду того или иного заболевания).

Суть предлагаемой процедуры диагностирования для k–го пациента состоит в нахождении для него весов $w(a_i, d_j)$ (согласно (1)) компонент профиля Pk для каждого заболевания из 6 возможных, суммирования этих весов для каждого заболевания dj и выбора из найденных сумм наибольшей суммы. Заболевание, соответствующее наибольшей сумме, интерпретируем как диагноз для k–го пациента.

Для оценки точности диагностирования предлагаемой процедуры её диагноз для каждого пациента сравнивался с диагнозом, содержащимся в выборке Dermatology, и в случае их совпадения диагноз процедуры считался правильным (*wrapper approach*). Точность диагностирования процедуры без коррекций составила 76% (261 правильный диагноз). Ниже показано, как с помощью некоторых эвристических коррекций профиля пациента можно повысить точность диагностирования.

2 Эвристические приемы повышения точности диагностирования

Продолжим аналогию в терминах распознавания образов и перейдем от профиля пациентиа к «портрету» пациента. Под портретом пациента будем понимать совокупное распределение D(ak, dj) весов анализов по компонентам профиля Pk, k = k1, …,k_{n_k} , и по видам заболеваний dj, j = 1, …, 6. Эвристические приемы повышения точности диагностирования предусматривают введение следующих коррекций в портрет пациента.

А). Поскольку точность классификации зависит от выбора шкалы, используемой для оценки результатов анализов, исходные оценки анализов в выборке Dermatology корректировались с использованием нелинейной монотонной функции $f(\hat{a}_i) = log_{10}(1 + \hat{a}_i{}^2)$.

Б). При вычислениях веса для того или иного заболевания согласно (2) возможны длинные «хвосты» последовательных сложений «важностей» анализов с возрастающими рангами (≤ 16), эти «хвосты» могут ухудшать точность классификации. Для «отсечения» таких «хвостов» был проведен одномерный поиск по нахождению «точки отсечения» (*cut off*) длинных «хвостов». В данной процедуре из формулы (2) исключались анализы с рангами, большими 11.

В). Допустим, что для k-го пациента распределение $D(a_k, d_j)$ таково, что в нем существует (j, i)-й элемент, являющийся единственным ненулевым элементом для j-й строки и i-го столбца, Это

означает, что анализ a_i взаимно однозначно связан с видом заболевания d_j. В таком случае заболевание d_j считаем диагнозом для k-го пациента, а анализ a_i назовем *маркером* для этого пациента. Возможны случаи, когда для некоторых пациентов существует несколько маркеров, тогда диагноз связывается с маркером наибольшего веса.

Учет маркеров в распределении $D(a_k, d_j)$ позволил повысить точность диагностирования до 88.4% (302 правильных диагноза). Заметим, что в рамках классификации без обучения каждый рассматриваемый случай является уникальным, поэтому невозможно проводить сравнение результатов классификации для несравнимых случаев. В то же время результат предложенной здесь процедуры диагностирования (*unsupervised*) представляется вполне «приличным».

Г). Представляется полезным следующий этап процедуры: для каждого из 41 оставшегося «без диагноза» пациента процедура указывает пару заболеваний (диагнозов), при этом для 14 пациентов в такой паре один из диагнозов являлся правильным («серое» диагностирование). Цель этого этапа состоит в сокращении вариантов возможных диагнозов для каждого пациента, что может ускорить процесс выработки врачом результирующего диагноза.

Заключение

Выработка медицинского диагноза – это сложный многоэтапный процесс: предварительное обследование, подбор необходимых анализов, интерпретация результатов анализов, принятие окончательного решения. Предлагаемая в докладе формальная процедура относится к последнему этапу – принятию решения по результатам анализов с учетом качественных измерений результатов анализов, принципиальной неоднозначности в интерпретации самих анализов и т.п. Подобные трудности возникают при интерпретации совершаемых действий и выработке решений и в других системах (финансовых, технических, социальных и т.п.), где подобная формальная процедура могла бы оказаться весьма полезной.

Литература

1. Fatima M., Pasha M. Survey of machine learning algorithms for disease diagnostic //J. Intel. Learning Syst.&Appl. 2017. Vol. 9. P. 1–16.http://www.scirp.org/journal/jilsa

2. Deo R. C. Machine learning in medicine // Circulation. 2015. Vol. 132. № 20. P. 1920–1930. lhttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5831252/

*3.*UCI Machine learning repository // archive.ics.uci.edu/ml/datasets.html

4. Govenir H. A., Demiroz G., Ilter N. Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals // Artif. Intelligence in Medicine, 1998. Vol.13. pp. 147 – 165.

5. El-Baz A. H. Filter based feature selection for automatic detection of erythemato-squamous diseases // British J. of Math. and Comp. Sci. 2015. Vol. 9. № 5. pp. 394 – 406. www.journalrepository.org/.../El-Baz952015BJMCS17618.p

*6. E. Roszkowska, "*Rank ordering criteria weighting methods – a comparative overview," Optimum. Studia Ekonomiczne, 2013, vol. 65, No 5. https://www.researchgate.net/publication/280081585 ...