МЕТОДИКА АВТОМАТИЗИРОВАННОГО ДИАГНОСТИРОВАНИЯ ЗАБОЛЕВАНИЯ

Мотовилова Ю.Е.

В наше время уже существует и продолжает появляется огромный поток научной информации на бумажных носителях в виде статей, монографий и сборников, с которой врачи просто не в состоянии справиться (ежегодно публикуется около 2 млн. статей в 40 тыс. биомедицинских журналах, словарь международных клинических классификаций насчитывает около 30 000 основных терминов). Использование для лечения больных значительного количества лекарственных препаратов, имеющих противопоказания, побочные действия, сложные взаимодействия с другими препаратами, определенные дозировки и схемы применения (Сейчас известно уже более 12 тысяч лекарственных препаратов и симптомов заболеваний, свыше 4000 болезней, помнить эту информацию не представляется возможным). Существование немалого процента врачебных ошибок, сопровождаемого дальнейшей потерей здоровья больного или летальным исходом из-за неточности в оценке его состояния.

Очень часто для принятия медицинских решений характерны недостаточность знаний, ограниченность временных ресурсов, отсутствие возможности привлечения компетентных экспертов, неполнота информации о состоянии больного. Указанные факторы являются причинами врачебных ошибок, которые могут привести к дальнейшей потере здоровья пациента.

Исходными данными для оценки являются совокупность характеристик (симптомов) пациента, по которым определяют состояние организма пациента и наличие болезни или ее отсутствие, т.е диагноз.

Подход к кластеризации, предусматривающий использование показателя качества, связан с разработкой процедур, которые обеспечивают минимизацию или максимизацию выбранного показателя качества.

Одним из наиболее популярных показателей является сумма квадратов ошибки

$$J = \sum_{j=1}^{N_c} \sum_{x \in S_j} \left\| x - m_j \right\|^2 \qquad 1.1$$

где N_c — число кластеров, т.е заболеваний, S_j — множество образов, т.е симптомов, относящихся к j-му заболеванию,

$$m_j = \frac{1}{N_j} \sum_{x \in S_j} x \qquad 1.2$$

вектор выборочных средних значений для множества S_j; N_j - число симптомов, входящих во множество S_j.

Показатель качества (1.1) определяет общую сумму квадратов отклонений характеристик всех образов, входящих в некоторый кластер, от соответствующих средних значений по кластеру. Алгоритм, основанный на этом показателе качества, рассматривается ниже. (следовательно здесь мы находим показатель ,показывающий отклонение от значения кластера).

Алгоритм k внутригрупповых средних, представленный ниже, минимизирует показатель качества, определенный как сумма квадратов расстояний всех точек, входящих в кластерную область, до центра кластера. Эта процедура, которую часто называют алгоритмом, основанным на вычислении k внутригрупповых средних, состоит из следующих шагов.

Шаг 1. Выбираются k исходных центров кластеров $z_1(l)$, $z_2(l)$, ..., $z_k(l)$. Этот выбор производится произвольно, и обычно в качестве исходных центров используются первые k результатов выборки из заданного множества образов.

Шаг 2. На k-м шаге итерации заданное множество образов {x} распределяется по k кластерам по следующему правилу:

$$x \in S_j(k), \ \textit{если} \quad \|x - z_j(k)\| < \|x - z_i(k)\| \qquad (2.6)$$

для всех i - 1, 2 ..... k, $i \neq j$, где S_j (k) — множество образов, входящих в кластер с центром z_j(k). В случае равенства в (2.6) решение принимается произвольным образом.

Шаг 3. На основе результатов шага 2 определяются новые центры кластеров z_j(k+1), j= 1, 2 ..... k, исходя из условия, что сумма квадратов расстояний между всеми образами, принадлежащими множеству S_j (k), и новым центром кластера должна быть минимальной.

Другими словами, новые центры кластеров z_j(k+1) выбираются таким образом, чтобы минимизировать показатель качества

$$J_j = \sum_{x \in S_j(k)} \|x - z_j(k+1)\|^2, \quad j = 1,2....K \qquad (2.7)$$

Центр $z_j(k+1)$, обеспечивающий минимизацию показателя качества, является, в сущности, выборочным средним, определенным по множеству $S_j(k)$. Следовательно, новые центры кластеров определяются как

$$z_j(k+1) = \frac{1}{N_j} \sum_{x \in S_j(k)} x, \quad j = 1,2.....K \tag{2.8}$$

где N_j - число выборочных симптомов, входящих в множество $S_j(k)$.

Очевидно, что название алгоритма «k внутригрупповых средних» определяется способом, принятым для последовательной коррекции назначения центров кластеров.

Шаг 4. Равенство $z_j(k+1) = z_j(k)$, при j=1, 2 ..... k является условием сходимости алгоритма, и при его достижении выполнение алгоритма заканчивается. В противном случае алгоритм повторяется от шага 2.

Качество работы алгоритмов, основанных на вычислении k внутригрупповых средних, зависит от числа выбираемых центров кластеров, от выбора исходных центров кластеров, от последовательности осмотра симптомов и, естественно, от геометрических особенностей данных. Хотя для этого алгоритма общее доказательство сходимости не известно, приемлемые результаты можно ожидать в тех случаях, когда данные образуют характерные гроздья, отстоящие друг от друга достаточно далеко. В большинстве случаев практическое применение этого алгоритма потребует проведения экспериментов, связанных с выбором различных значений параметра k и исходного расположения центров кластеров.

При таком подходе решающую роль играют интуиция и опыт. Он предусматривает задание набора правил, которые обеспечивают использование выбранной меры сходства для отнесения образов(симптомов) к одному из кластеров(заболеваний). Поскольку, близость двух образов является относительной мерой их подобия, обычно приходится вводить порог, чтобы установить приемлемые степени сходства для процесса отыскания кластеров.

После определения показателя качества его необходимо сравнить с эталонным показателем для заболевания.

При использовании данного метода станет возможным упрощение для врача процедуру проведения диагностики человека и определения заболевания.

1 Список литературы

1. Проект [Электронный ресурс]:Википедия – свободная энциклопедия – режим доступа : http://www.machinelearning.ru/wiki/index.php ;
2. Milenova, B., Campos, M. Clustering large databases with numeric and nominal values using orthogonal projections, Oracle Data Mining Technologies, 2002.
3. David Arthur, Sergei Vassilvitskii, "How Slow is the k-means Method?". Proceedings of the 2006 Symposium on Computational Geometry (SoCG)/2006
4. Журавлев Ю. И., Рязанов В. В., Сенько О. В. «Распознавание». Математические методы. Программная система. Практические применения. — М.: Фазис, 2006.
5. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. — Springer, 2001.