УДК [615.471:616-071]:681.513

Н. А. Тесленко, И. Г. Чурюмова

НЕЧЕТКАЯ КЛАСТЕРИЗАЦИЯ МАССИВОВ БИОМЕДИЦИНСКИХ ДАННЫХ В УСЛОВИЯХ ИЗБЫТОЧНОСТИ ИНФОРМАЦИИ

1. Введение

В последнее время пристальное внимание общественности уделяется проблеме здоровья студентов и других контингентов длительно учащейся молодежи [1, 2, 3]. Молодые люди в возрасте 16–25 лет приходят в вузы во время так называемого пика активного динамического развития. Функционально формирующиеся мышечные группы, в зависимости от адекватной физической нагрузки, оказывают прямое влияние на все внутренние органы и системы организма, управляя их работой и синхронизируя их функции [4]. Но двигательная активность студента резко ограничена как во времени, так и в пространстве. По характеру учебы он находится в состоянии постоянной гиподинамии, создаются условия для формирования, а затем развития статического мышечного перенапряжения [5]. Этот термин емко отображает функциональную недогруженность одних мышечных групп (например, мышц ног) и перегруженность других (например, мышц спины). Прежде всего это находит отражение в асинхронности мышечных сокращений и негативном влиянии на работу внутренних органов и особенно сердца.

Находясь ежедневно приблизительно в одном и том же положении, недостаточно нагруженные в физическом смысле, например, мышцы ног теряют свой тонус из-за уменьшения притока крови к ним, кровеносные сосуды сужаются, что ухудшает микроциркуляцию мышечных волокон, в расслабленных мышцах кровоток затрудняется и уменьшается. Растет общее периферическое сопротивление сосудов, что пагубно отражается на работе сердца как насоса, вынужденного с большим усилием работать, проталкивая кровь через мышечные группы. В то же время мышцы спины, находясь в тонусе, а затем переходя в состояние гипертонуса, нуждаются в усиленной работе сердца [6, 7]. Это находит отражение в центральной и периферической нервной регуляции просвета кровеносных сосудов. Сужение и расширение кровеносных сосудов нарушается. Учитывая, что жизнь учащейся молодежи связана со стресс-факторами в виде зачетов, экзаменов, контрольных и, как правило, протекает на фоне предрасполагающих болезнетворных факторов (нарушение пищевого и водного режимов, нарушение формулы сна и т. д.), а зачастую сопровождается вредными привычками (курение, алкоголь, крепкий чай и кофе, избыток соли в пище), становится понятным значение развивающегося статического перенапряжения в формировании пограничных поражений сердечно-сосудистой системы [8].

2. Цель исследования

Целью нашего исследования явилось изучение закономерностей формирования статического мышечного перенапряжения как одного из главных предрасполагающих факторов развития пограничных гипертонических состояний у учащейся молодежи и использование новых перспективных методов нечеткой кластеризации для обработки полученных данных.

Работа проводилась на базе Харьковского областного врачебно-физкультурного диспансера, кафедры внутренних болезней, спортивной медицины и лечебной физкультуры Харьковского государственного медицинского университета и кафедры биомедицинских электронных устройств и систем Харьковского национального университета радиоэлектроники. Объектами наблюдения являлись 368 студентов младших (1–2) и старших (4–5) курсов Харьковского государственного медицинского университета, Харьковской государственной академии физического воспитания, Харьковского национального университета им. Каразина и Харьковского национального университета радиоэлектроники. Обследуемые студенты, возраст которых составлял от 16 до 26 лет, были распределены на три группы.

3. Сжатие данных с помощью трехслойной нейронной сети и кластеризация данных с помощью нечеткой BSB-модели

В настоящее время задачи кластеризации данных различной природы успешно выполняются с помощью широкого класса нейронных сетей. Решение данной задачи основывается на принципах самообучения, которые заложены в основу конструкции автоассоциативной памяти как одного из видов искусственных нейронных сетей, отличающихся простотой архитектуры и наличием достаточно эффективных алгоритмов обучения.

Предварительный анализ данных показал, что исходная размерность пространства избыточна для решения задачи кластеризации, поскольку существуют неинформативные признаки. Для сокращения пространства исходных данных, т. е. уменьшения количества признаков, была реализована процедура сжатия информации с помощью автоассоциативной трехслойной нейронной сети типа «Bottle neck». Данная сеть характеризуется тем, что количество нейронов первого скрытого и третьего (выходного) слоев принято равным количеству признаков классифицируемых объектов (в нашем случае 19), а количество нейронов второго скрытого

слоя равно сокращенному количеству признаков. В таком случае выходы второго скрытого слоя представляют собой сжатые данные, а выходного слоя — восстановленные. Описанная нейронная сеть настраивается с помощью алгоритма на основе полиномиальной функции активации [9]

$$\psi_j(u_j) = 1{,}5\left(u_j(k) - \frac{u_j^3}{3}(k)\right);$$

$$-1 < \psi_j(u_j) < 1;\ -1 < u_j < 1,$$

обеспечивающей более высокую скорость настройки синаптических весов сети по сравнению с традиционными процедурами.

В результате процедуры сжатия количество признаков сокращено до 7. Для обработки полученных данных была использована искусственная нейронная сеть автоассоциативной памяти специального вида, известная под названием «Brain-State-In-A-Box Model» [10, 11]. BSB-модель представляет собой нейродинамическую нелинейную замкнутую систему с амплитудными ограничениями, охваченную положительной обратной связью. Динамика этой системы определяется парой уравнений в пространстве состояний

$$\begin{cases} y(k, \tau) = x(k, \tau) + \alpha W x(k, \tau), \\ x(k, \tau + 1) = \psi(y(k, \tau)), \end{cases}$$

где $x(k, 0) \equiv x(k)$ — $(n \times 1)$-вектор-образ, подаваемый в систему; $k = 1, 2, \ldots, l$ — номер конкретного образа в множестве фундаментальной памяти; $\tau = 0, 1, 2, \ldots, T$ — итерации машинного времени; $x(k, T)$ — вектор состояний в установившемся режиме; α — малый положительный параметр обратной связи; W — $(n \times n)$-матрица синаптических весов линейной корреляционной автоассоциативной памяти, представляющей собой по сути однослойную нейронную сеть, образованную адаптивными линейными ассоциаторами; $\psi(\bullet)$ — активационная кусочно-линейная функция с насыщением, действующая покомпонентно на элементы вектора $y(k, \tau)$ так, что

$$x_i(k, \tau + 1) = \psi(y_i(k, \tau)) =$$

$$= \begin{cases} +1, & \text{если } y_i(k, \tau) > +1; \\ y_i(k, \tau), & \text{если } -1 \le y_i(k, \tau) < +1; \\ -1, & \text{если } y_i(k, \tau) < -1; \\ i = 1, 2, \ldots, n. \end{cases}$$

Фазовое пространство BSB-нейромодели ограничено n-мерным гиперкубом с центром в начале координат и ребром длиной 2. Присутствие в модели положительной обратной связи приводит к тому, что подаваемые в модель сигналы, предварительно закодированные для работы в гиперкубе, усиливаются до тех пор, пока все состояния $x_i(k, \tau)$ не войдут в насыщение, при этом аналоговые входные сигналы преобразуются в дискретную форму, соответствующую одной из вершин гиперкуба.

BSB-модель решает задачу кластеризации заданного массива данных $x(k), k = 1, 2, \ldots, l$ благодаря тому, что вершины гиперкуба в процессе обработки действуют как точечные аттракторы с некоторыми областями притяжения, которые делят n-мерное пространство признаков. В связи с тем, что количество аттракторов, т. е. вершин гиперкуба, равно $2^n >> n$, а емкость линейной автоассоциативной памяти W не может превышать n (абсолютная емкость $l / n \le 1$), возникает ситуация при которой большинство вершин будут оставаться пустыми. В то же время данные могут принадлежать нескольким или всем кластерам одновременно, так как многие реальные задачи не могут быть решены путем однозначного разделения всех объектов на группы.

Таким образом, целесообразно ввести некоторую функцию, которая будет описывать уровень принадлежности объекта каждому из классов [12]. Для этого можно воспользоваться идеями нечеткой кластеризации, используя простую и эффективную треугольную функцию принадлежности:

$$\mu_q(x(k, T)) = 1 - \frac{d(x(k, T), x_q)}{2n},$$

где $d(x(k, T), x_q) = \sum_{i=1}^{n} |x_i(k, T) - x_{q,i}|$ — хэммингово расстояние.

В случае гиперкуба уровень принадлежности описывает близость каждой вершины объекта $x(k, T)$ к q-й вершине, в которой находится явный представитель группы (известный априори). Такая вершина определяется как центр кластера.

Качество работы BSB-модели определяется в значительной мере емкостью автоассоциативной линейной памяти, включенной в контур обратной связи, а эта емкость, в свою очередь, существенно зависит от принятой процедуры настройки n^2 синаптических весов адаптивных линейных ассоциаторов. Для настройки синаптических весов данной нейросетевой модели был использован алгоритм Уидроу—Хоффа [13], который обеспечивает достаточное быстродействие и обладает необходимыми следящими свойствами для решения задачи:

$$W(k+1) = W(k) + \eta\big(x(k+1) - W(k)x(k+1)\big)x^T(k+1),$$

где η — параметр шага обучения.

В результате обработки данных пониженной размерности с помощью нечеткой BSB-нейромодели было выявлено шесть устойчивых кластеров, соответствующих различным состояниям организма.

4. Понижение размерности входного пространства с помощью метода главных компонент и кластеризация данных fcm-методом

Поскольку, как было замечено выше, исходная размерность пространства избыточна для решения задачи кластеризации из-за наличия неинформативных признаков, то для сокращения входного пространства был использован метод главных компонент

(PCA — Principal component analysis) [14, 15]. Этот метод применяется в тех случаях, когда приходится сталкиваться с ситуациями, в которых общее число признаков $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ очень велико и, тем не менее, многомерные наблюдения $X_i = \begin{pmatrix} x_i^{(1)} \\ x_i^{(2)} \\ \vdots \\ x_i^{(p)} \end{pmatrix}$, $i = 1, 2, \ldots, n$ следует подвергнуть обработке.

Каждое из наблюдений представляется в виде вектора Z некоторых вспомогательных показателей $z^{(1)}, z^{(2)}, \ldots, z^{(p')}$ $(p' << p)$. Формально задача перехода к новому набору признаков $\tilde{z}^{(1)}, \tilde{z}^{(2)}, \ldots, \tilde{z}^{(p')}$ заключается в определении такого набора признаков $\tilde{Z}$, найденного в классе F допустимых преобразований исходных показателей $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$, чтобы

$$I_{p'}\left(\tilde{Z}(X)\right) = \max_{Z \in F}\left\{I_{p'}(Z(X))\right\},$$

где $Z(X) = (z^{(1)}(X), \ldots, z^{(p')}(X))$.

В качестве класса допустимых преобразований F определяются всевозможные линейные ортогональные нормированные комбинации исходных показателей, т. е.

$$z^{(j)}(X) = c_{j1}(x^{(1)} - \mu^{(1)}) + \ldots + c_{jp}(x^{(p)} - \mu^{(p)});$$

$$\sum_{v=1}^{p} c_{jv}^2 = 1, \; j = 1, 2, \ldots, p;$$

$$\sum_{v=1}^{p} c_{jv} c_{kv} = 0, \; j, k = 1, 2, \ldots, p, \; j \neq k;$$

(здесь $\mu^{(v)} = Ex^{(v)}$ — математическое ожидание $x^{(v)}$); $I_{p'}\left(\tilde{Z}(X)\right) = \dfrac{Dz^{(1)} + \ldots + Dz^{(p')}}{Dx^{(1)} + \ldots + Dx^{(p)}}$ — мера информативности p'-мерной системы признаков $Z(X) = (z^{(1)}(X), \ldots, z^{(p')}(X))$ (здесь D — знак операции вычисления дисперсии соответствующей случайной величины).

Для обработки данных было предложено использовать FCM-алгоритм (Fuzzy C-Means Algorithm). Это один из алгоритмов нечеткой кластеризации данных [14, 16].

Исходной информацией является выборка наблюдений, сформированная из N n-мерных векторов признаков $X = \{(x(1), x(2), \ldots, x(N)\}$, $x(k) \in X$, $k = 1, 2, \ldots, N$. Результат работы алгоритма представляет собой разбиение исходного массива данных на m классов с некоторым уровнем $w_j(k)$ принадлежности k-го вектора признаков j-му кластеру.

Поступающие на обработку данные предварительно центрируются и стандартизируются по всем признакам так, чтобы все наблюдения принадлежали гиперкубу $[-1, 1]^n$. Центрирование осуществляется либо относительно среднего, вычисляемого с помощью соотношения

$$m_i(k) = m_i(k-1) + \frac{1}{k}\left(x_i(k) - m_i(k-1)\right),$$

либо с целью придания процедуре центрирования робастных свойств (защита от аномальных наблюдений) оно осуществляется относительно медианы, которая вычисляется согласно рекуррентному соотношению

$$me_i(k) = me_i(k-1) + \eta_m \operatorname{sign}\left(x_i(k) - me_i(k-1)\right),$$
$$i = 1, 2, \ldots, n,$$

где η_m — параметр шага поиска, выбираемый в стационарном случае в соответствии с условиями Дворецкого.

Алгоритмы, основанные на целевых функциях [17], предназначены для решения задачи кластеризации путем оптимизации некоторого наперед заданного критерия качества кластеризации и являются наиболее строгими с математической точки зрения.

Целевая функция, подлежащая минимизации, имеет вид

$$E(w_j(k), c_j) = \sum_{k=1}^{N}\sum_{j=1}^{m} w_j^{\beta}(k) d^2(x(k), c_j)$$

при ограничениях

$$\sum_{j=1}^{m} w_j(k) = 1, \; k = 1, \ldots, N;$$

$$0 < \sum_{k=1}^{N} w_j(k) < N, \; j = 1, \ldots, m.$$

Здесь $w_j(k) \in [0, 1]$ — уровень принадлежности вектора $x(k)$ к j-му кластеру, c_j — прототип (центр) j-го кластера, β — неотрицательный параметр, именуемый «фаззификатором» (обычно принимается равным 2), $d^2(x(k), c_j)$ — расстояние между $x(k)$ и c_j в принятой метрике.

В результате кластеризации получаем $N \times m$ матрицу $W = \{w_j(k)\}$, называемую *матрицей нечеткого разбиения*.

Поскольку элементы матрицы W могут рассматриваться как вероятности принадлежности векторов данных определенным кластерам, то такие процедуры называются *вероятностными алгоритмами кластеризации*.

Введем функцию Лагранжа

$$L(w_j(k), c_j, \lambda(k)) = \sum_{k=1}^{N}\sum_{j=1}^{m} w_j^{\beta}(k) d^2(x(k), c_j) +$$

$$+ \sum_{k=1}^{N} \lambda(k)\left(\sum_{j=1}^{m} w_j(k) - 1\right) =$$

$$= \sum_{k=1}^{N}\left(\sum_{j=1}^{m} w_j^{\beta}(k) d^2(x(k), c_j) + \lambda(k)\left(\sum_{j=1}^{m} w_j(k) - 1\right)\right),$$

где $\lambda(k)$ — неопределенный множитель Лагранжа.

Решая систему уравнений Куна—Таккера

$$\begin{cases} \dfrac{\partial L(w_j(k), c_j, \lambda(k))}{\partial w_j(k)} = 0, \\ \nabla_{c_j} L(w_j(k), c_j, \lambda(k)) = 0, \\ \dfrac{\partial L(w_j(k), c_j, \lambda(k))}{\partial \lambda(k)} = 0, \end{cases}$$

получим решение в виде

$$w_j^{pr}(k) = \frac{\left(d^2(x(k), c_j)\right)^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m}\left(d^2(x(k), c_l)\right)^{\frac{1}{1-\beta}}},$$

$$c_j^{pr} = \frac{\sum_{k=1}^{N} w_j^{\beta}(k)x(k)}{\sum_{k=1}^{N} w_j^{\beta}(k)x(k)},$$

$$\lambda(k) = -\left(\sum_{l=1}^{m}\left(\beta d^2(x(k), c_l)\right)^{\frac{1}{1-\beta}}\right)^{\frac{1}{1-\beta}}.$$

Выбирая $\beta = 2$ и принимая эвклидово расстояние $d^2(x(k), c_j) = \left\|x(k) - c_j\right\|^2$, получаем алгоритм нечеткой кластеризации Бездека [17], называемый еще FCM-алгоритмом (Fuzzy C-Means Algorithm) [14], который и использовался в нашей работе:

$$w_j^{pr}(k) = \frac{\left\|x(k) - c_j\right\|^{-2}}{\sum_{l=1}^{m}\left\|x(k) - c_j\right\|^{-2}};$$

$$c_j^{pr} = \frac{\sum_{k=1}^{N} w_j^2(k)x(k)}{\sum_{k=1}^{N} w_j^2(k)x(k)};$$

$$\lambda(k) = -\sum_{l=1}^{m}\left(\frac{\left\|x(k) - c_j\right\|^{-2}}{2}\right)^{-1}.$$

При работе алгоритма на заранее сжатой выборке было получено 6 устойчивых классов.

5. Выводы

Результатом кластеризации сжатых данных с помощью рассмотренных выше моделей является разделение объектов на группы таким образом, что можно четко определить принадлежность каждого объекта только к одному классу.

Использование новых перспективных методов при обработке полученных данных показали высокую степень достоверности полученных результатов, что дает основание использовать такие методы в дальнейших исследованиях.

Список литературы: 1. *Бурханов А. И., Муценко Т. А.* Характеристика функции внешнего дыхания у студентов // Гигиена и санитария.– 1997.– № 2.– С. 32–34. 2. *Залимов Р. Ю.* Специфика адаптации студентов к условиям образовательного процесса и результативность их учебной деятельности в зависимости от состояния физиологических функций и личностных способностей. Физиологические основы здоровья студентов // Тр. МНС по экспериментальной и прикладной физиологии / Под редакцией К. В. Судакова.– М.: НИИНФ им. П. К. Анохина РАМН, 2001.– 10.– С. 69–83. 3. *Кашалиев М. А., Лисицин В. И., Возженкинова Р. В.* Состояние здоровья студентов медицинского института // Здравоохранение Казахстана.– 1992.– № 1.– С. 65–67. 4. *Самохвалов В. Г., Самохвалов А. В.* Динамика психологической и физиологической адаптации студентов к учебным нагрузкам. Физиологические основы здоровья студентов // Тр. МНС по экспериментальной и прикладной физиологии / Под редакцией К. В. Судакова.– М.: НИИНФ им. П. К. Анохина РАМН, 2001.– 10.– С. 84–106. 5. *Филатов О. М., Щедрина А. Г.* Роль индивидуальной изменчивости организма в формировании здоровья студентов // Гигиена и санитария.– 1996.– № 6.— С. 29–32. 6. *Аль Табак М.* Особливості впливу сполученої електростимуляції скелетних м'язів на функціональний стан організму спортсменів // Практична медицина.– 1999.– № 1–2.– С. 92–96. 7. *Босенко А. И., Белинова А. Г., Цонева Т. Н., Годына О. В.* Оценка резервных возможностей дыхания, кардио- и гемодинамики юных спортсменов // Гигиена и санитария.– 1995.– № 2.– С. 20–22. 8. *Антоненко П. А., Литвинов В. Н.* Автоматизированная система врачебного контроля функционального состояния организма спортсмена // Теория и практика физической культуры.– 1983.– № 1.– С. 42–44. 9. *Тесленко Н. А.* Алгоритм обучения автоассоциативной искусственной многослойной нейронной сети // Бионика интеллекта.– 2004.– Вып. 61(1).– С. 103–106. 10. *Anderson J. A.* Cognitive and psychological computation with neural models // IEEE Trans. on Systems, Man, and Cybernetics.– 1983.– 13.– P. 799–815. 11. *Anderson J. A., Silverstein J. W., Ritz S. A., Jones R. S.* Distinctive features, Categorical perception and probability learning: Some applications of a neural model // Neurocomputing: Foundations of Research / Ed. by J. A. Anderson, E. Rosenfeld.– Cambridge, MA: MIT Press, 1988.– P. 413–451. 12. *Бодянский Е. В., Тесленко Н. А., Чурюмова И. Г.* Нечеткая BSB-нейромодель для обработки биомедицинской информации. Интеллектуальные системы принятия решений и прикладные аспекты информационных технологий // М-лы научно-практической конференции: Т. 2.–Херсон: Изд-во Херсонского морского института, 2006.– С. 184–187. 13. *Бодянский Е. В., Колодяжный В. В., Тесленко Н. А.* Алгоритм обучения модифицированной BSB-модели в задачах кластеризации // М-лы 12-й Международной конференции по автоматическому управлению «Автоматика-2005».– Харьков: НТУ «ХПИ», 2005.– Т. 1.– С. 49–50. 14. *Nelles O.* Nonlinear System Identification: from classical approaches to neural networks and fuzzy models.– Springer, 2001. 15. *Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика. Классификация и снижение размерности.– М.: Финансы и статистика, 1989.– 607 с. 16. *Бодянский Е. В., Горшков Е. В. и др.* Об адаптивном алгоритме нечеткой кластеризации данных // АСАУ.– 2002.– 5(25). 17. *Bezdek J. C.* Pattern Recognition with Fuzzy Objective Function Algirithms. — N. Y.: Plenum Press, 1981.– 272 p.

Поступила в редколлегию 16.03.2006