ВЕСТНИК НАЦИОНАЛЬНОГО ИССЛЕДОВАТЕЛЬСКОГО ЯДЕРНОГО УНИВЕРСИТЕТА "МИФИ", 2020, том 9, № 3, с. 270–278

ПРИКЛАДНАЯ МАТЕМАТИКА И ИНФОРМАТИКА

УДК 004.422.635.5

МЕТОД КОМПЛЕКСИРОВАНИЯ ДАННЫХ В СИСТЕМЕ РАСПРЕДЕЛЕННОГО МОНИТОРИНГА

© 2020 г. А. А. Моисеев*

Технос – РМ, Мытищи, 141002, Россия

**e-mail: slow.coach@yandex.ru*

Поступила в редакцию 28.01.2020 г.
После доработки 26.03.2020 г.
Принята к публикации 08.09.2020 г.

Современные системы распределенного мониторинга включают устройства наблюдения и регистрации различных типов. Разнородность формируемых ими данных порождает проблемы, связанные:
– с объединением разнородных данных от различных источников, зачастую слабо связанных;
– противоречивостью, неполнотой и неточностью данных в отсутствие априорной идентификации наблюдаемых объектов;
– требованием оперативной обработки большого объема разнородной информации.
В данной работе рассматриваются алгоритмы обработки разнородных данных, обеспечивающие агрегирование последних с целью приведения к обозримому виду, удобному для получения аналитических выводов, повышения их надежности и принятия решений. Предлагаемый подход к комплексированию данных от разнородных источников состоит в использовании объединенного вектора признаков объектов. Дополнительной проблемой при этом является необходимость реализации комплексирования на единой формальной основе. Однако задача настолько назрела, что попытки создания такой основы на базе метода функционального шкалирования представляются вполне оправданными. Внутри этого направления развивается подход, основанный на переходе от исходных показателей к обобщенным, обрабатываемым численными методами. Последние ориентированы на решение следующих задач:
– снижение размерности векторов признаков за счет предварительного отбора наиболее информативных показателей;
– рациональная оцифровка анализируемых признаков;
– разбиение совокупности объектов на некоторое число однородных классов в рамках автоматической классификации без учителя;
– статистический анализ эффективности проведенного разбиения.
Для отображения признаков объекта используются номинальные (бинарные), порядковые (целочисленные) и относительные (действительные) показатели (переменные), нормированные к диапазону (0.1). В работе продемонстрировано преимущество евклидовой и манхэттенской метрик в пространстве нормированных переменных, состоящее в возможности естественным образом сформировать порог различения на базе критерия Неймана–Пирсона. Приведены также примеры формирования переменных различного типа и их использования на практике.

Ключевые слова: распределенный мониторинг, разнородные данные, агрегирование (комплексирование) данных, функциональное шкалирование, евклидова метрика, манхэттенская метрика, критерий Неймана–Пирсона

DOI: 10.1134/S2304487X20030062

Наиболее значимой составляющей автоматизированной обработки информации является автоматический анализ данных. Наряду с классическими процедурами статистического анализа – факторного, дисперсионного, дискриминантного и др. – он также включает ряд дополнительных процедур, не связанных напрямую со статистическим анализом [1]. К ним, в частности, относится процедура распознавания в отсутствие априорной идентификации объектов. В рамках данной процедуры особо выделяется направление, связанное с агрегированием (комплексированием) данных от разнородных источников. Целью этого комплексирования является, в конечном счете, предварительная идентификация объекта. Исходной при решении этой задачи традиционно является совокупность стандартизированных параметров, определяющая образ объекта в рамках его системного окружения. Обычно этот образ характеризуется положением в пространстве пара-

метров, а задача идентификации эквивалентна привязке этого положения к соответствующей области, сформированной в ходе предварительной классификации. Последняя в этих условиях сводится к разбиению этого пространства на области решений, соответствующие характеристикам отдельных объектов. Возможным методом решения этой задачи является предварительная кластеризация указанного множества точек на этапе обучения без учителя с последующей фиксацией границ областей. Предварительная оценка параметров объектов может проводиться методом редукции на основе модели измерения, базирующейся, например, на основном соотношении пассивной локации [2]. Настройка параметров модели, интерпретируемых как параметры наблюдаемого сигнала, осуществляется с помощью статистического или прямого перебора. Таким образом, метод редукции представляет собой модельно-ориентированный метод оценки параметров отдельного источника по результатам наблюдения принятого от него сигнала.

ЗАДАЧА КОМПЛЕКСИРОВАНИЯ ПАРАМЕТРОВ

В рамках проводимого рассмотрения источниками данных являются системы распределенного мониторинга, которые имеют сложную структуру и включают устройства наблюдения и регистрации различных типов. Данные устройства предназначены для определения разных параметров, связанных с выполнением тех или иных функций: обнаружения, принятия решений и др. При этом возникают проблемы, связанные:

– с объединением разнородных данных от различных источников, зачастую слабо связанных;

– противоречивостью, неполнотой и неточностью данных в отсутствие априорной идентификации наблюдаемых объектов;

– требованием оперативной обработки большого объема разнородной информации.

В данных условиях требуется разработка новых методов и моделей, обеспечивающих агрегирование разнородных данных с целью приведения последних к обозримому виду, удобному для получения аналитических выводов, повышения их надежности и принятия решений [3].

Для анализа информации требуется не только ее сбор, но и представление в приемлемом для дальнейшего анализа виде в ходе стандартизации. Дополнительной проблемой при этом является то, что реализация комплексирования на единой формальной основе осложняется мозаичным характером используемых на практике моделей и алгоритмов. Однако задача настолько назрела, что первые попытки создания такой основы уже предпринимаются, например, на базе метода функционального шкалирования [4]. В свою очередь, внутри этого направления развивается подход, основанный на переходе от исходных показателей к обобщенным, позволяющим, в частности, снизить размерность номенклатуры используемых показателей. К этому направлению относятся, в частности, методы метрического и неметрического шкалирования, предназначенные для формирования векторов обобщенных показателей (переменных) для признаков.

Одним из наиболее перспективных методов агрегирования, основанном на этой базе, является метод динамических сгущений [5], ориентированный на решение следующих основных задач:

– снижение размерности векторов признаков за счет предварительного отбора наиболее информативных показателей;

– рациональная оцифровка анализируемых признаков;

– разбиение совокупности объектов на некоторое число однородных классов в рамках автоматической классификации без учителя;

– статистический анализ эффективности проведенного разбиения.

Возможный подход к комплексированию данных от разнородных источников состоит в использовании объединенного вектора признаков объектов, общего для всех них. Для различных типов данных используется, разумеется, лишь часть компонент этого вектора, и лишь объединение данных позволяет сформировать его полностью. Безусловно, номенклатуры признаков для разных объектов могут отличаться. В этих условиях в составе единого вектора признаков должны отображаться факты наличия или отсутствия тех или иных признаков.

Естественным методом этого отображения является введение в состав вектора признаков соответствующих бинарных компонент, которые фактически представляют собой номинальные переменные. Помимо них в состав признаков могут входить номинальные компоненты с несколькими состояниями. Ради сохранения единообразия представления компонент было предложено характеризовать различные состояния номинальных признаков отдельными бинарными компонентами [6]. Относительные переменные в ходе стандартизации нормируются в их диапазоне в соответствии с соотношением $x = \dfrac{N - N_{\min}}{N_{\max} - N_{\min}} \in (0,1)$, где N – значение переменной, а $(N_{\min}, N_{\max})$ – ее диапазон. Аналогичным образом нормируются порядковые переменные: $x = \dfrac{O - O_{\min}}{O_{\max} - O_{\min}} \in (0,1)$, где O – текущее значение порядковой переменной, а $(O_{\min}, O_{\max})$ – диапазон номеров ее состояний в порядке возрастания. Подобная нормировка вы-

глядит довольно грубой, однако, с точки зрения принципа недостаточного основания, представляется вполне логичной. Таким образом, в результате проведенной декомпозиции номинальных переменных и нормировки относительных и порядковых все компоненты вектора признаков представляются в диапазоне (0.1).

МЕРЫ БЛИЗОСТИ В ПРОСТРАНСТВЕ ПРИЗНАКОВ

Возможной мерой близости в пространстве векторов номинальных признаков является коэффициент корреляции по Пирсону, определяемый соотношением:

$$k_P = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+c)(b+d)}},$$

где a – число признаков, которыми обладают оба объекта; d – число признаков, которыми не обладают оба объекта; b – число признаков, которыми обладает только объект 1; c – число признаков, которыми обладает только объект 2.

Этот коэффициент может служить мерой близости между объектами только в случае, когда используются исключительно номинальные (бинарные) признаки, что, разумеется, снижает его ценность.

Противоположная ситуация возникает при использовании информационной (шенноновской) метрики. В случае использования нормировки по диапазону нормированные переменные x_{i1}, x_{i2}, относящиеся к разным объектам, могут быть использованы для формирования шенноновских норм вида [7]:

$$I_1 = \ln n - \sum_{i=1}^{n} x_{1i} \ln \frac{1}{x_{1i}},$$

$$I_2 = \ln n - \sum_{i=1}^{n} x_{2i} \ln \frac{1}{x_{2i}},$$

где n – размерность вектора признаков. Смысл этих норм – избыток информации в векторах признаков, формально рассчитанной по переменным x_{1i}, x_{2i}. Последние при этом интерпретируются как частные вероятности признаков, а избыток рассчитывается относительно ситуации равновероятности последних. Эти нормы могут быть использованы для формального определения шенноновской метрики вида $M_S = |I_1 - I_2|$. Очевидно, однако, что эффективными здесь являются только относительные переменные, поскольку бинарные в обеих нормах автоматически обнуляются.

В этих условиях более естественными представляются евклидова метрика, базирующаяся на соотношении $M_E = \sum_{i=1}^{n} \frac{(x_{1i} - x_{2i})^2}{\sigma^2}$, а также манхэттенская метрика вида $M_M = \sum_{i=1}^{n} \frac{|x_{1i} - x_{2i}|}{\sigma}$. Здесь $i = 1, ..., n$, номера компонент вектора признаков, n – его размерность, а 1, 2 – номера сравниваемых объектов, $\sigma^2 = \sigma_1^2 + \sigma_2^2$, где $\sigma_1^2 = \sigma_2^2 = \frac{1}{12}$ – дисперсии нормированных переменных, предполагаемых равномерно распределенными в диапазоне (0, 1). С учетом нормировки величину евклидовой метрики можно интерпретировать как случайную величину с распределением Пирсона $\chi^2_{n-1}(M_E)$. Это обстоятельство является очень удобным, поскольку позволяет задать порог различения объектов по критерию Неймана–Пирсона, исходя из естественного требования по вероятности ошибки первого рода [8].

Близкий подход использовался в задаче формирования эталонов в ходе параметрической дискриминации [9], базирующейся на метрике Кларка вида:

$$\rho(x_1, x_2) = \sum_{i=1}^{n} \left(\frac{x_{1i} - x_{2i}}{\frac{x_{1i} + x_{2i}}{2}} \right)^2. \quad (1)$$

Предположим, что при наличии эталона x_0, в параметрическом пространстве наблюдается совокупность параметров x. В этом случае строится промежуточный эталон $x_1 = \frac{x_0 + x}{2}$ и в соответствии с (1) рассчитываются метрики $\rho(x_1, x_0)$, $\rho(x_1, x)$. Эти метрики, в свою очередь, используются для формирования решающей статистики вида [10]:

$$f_{0x} = \frac{\max(\rho(x_1, x_0), \rho(x_1, x))}{\min(\rho(x_1, x_0), \rho(x_1, x))}. \quad (2)$$

Поскольку определенная выше статистика близка к оценке дисперсии для вариации, статистика f_{0x} имеет распределение Фишера $F_{0x} = F(f_{0x}, n-1, n-1)$. При этом решение $x \in O(x_0)$ принимается по критерию Фишера, если $F_{0x} < P$, где $P \in (0.9, 0.999)$ – доверительная вероятность различения совокупностей параметров, дополнительная к вероятности ошибки первого рода [8]. При невыполнении этого неравенства принимается решение $x \notin O(x_0)$.

Алгоритм параметрической дискриминации включает поэтапное обучение. Пусть на первом этапе обучения наблюдаются совокупности x_1, x_2. Как и ранее формируется промежуточная совокупность $x_3 = \frac{x_1 + x_2}{2}$. В соответствии с (1) для нее

формируются метрики $\rho(x_1, x_3)$, $\rho(x_2, x_3)$, а в соответствии с (2) – решающая статистика вида:

$$f_{12} = \frac{\max(\rho(x_1, x_3), \rho(x_2, x_3))}{\min(\rho(x_1, x_3), \rho(x_2, x_3))}. \quad (3)$$

В соответствии с критерием Фишера, если выполняется условие $F_{12} = F(f_{12}, n-1, n-1) < P$, считается, что $x_1, x_2 \in O(x_0)$. В противном случае x_1, x_2 интерпретируются как разные эталоны.

Предположим, что совокупность x наблюдается при наличии нескольких эталонов. Для каждой новой совокупности x находится эталон x_0, обеспечивающий минимум метрики $\rho(x_0, x)$ среди всех определенных эталонов. Для этого эталона с использованием описанного выше метода проверяется условие $x \in O(x_0)$. При выполнении этого условия x классифицируется как относящийся к эталону x_0. В противном случае x интерпретируется как новый эталон x_{00} и для него инициируется этап обучения установкой показателя $k = 1$. При отнесении x к x_0, т.е. при выполнении условия $x \in O(x_0)$, обучение состоит в коррекции соответствующего эталона для очередной совокупности x в соответствии с соотношениями:

$$k \to k+1,$$
$$x_{00} \to \left(1 - \frac{1}{k+1}\right) x_{00} + \frac{x}{k+1}. \quad (4)$$

Показатель для прочих эталонов корректируется в соответствии с соотношением $k \to k - 0.05$. Решение о дальнейшей судьбе эталона принимается по величине показателя k. Если в ходе обучения оказывается, что $k < 0.1$, соответствующий эталон удаляется из набора эталонов. Если $k > 10$, обучение завершается и эталон фиксируется. Введенные параметры коррекции показателя k указаны по умолчанию и могут быть изменены в ходе настройки алгоритма.

При отсутствии в составе используемого программного пакета встроенных статистических функций для распределения Фишера можно использовать аппроксимацию, базирующуюся на использовании бета-распределения [11]:

$$F(f, \nu_1, \nu_2) = I_x\left(\frac{\nu_2}{2}, \frac{\nu_1}{2}\right),$$
$$x = \frac{\nu_2}{\nu_2 + \nu_1 f},$$

где F – распределение Фишера; f – статистика Фишера вида (3); ν_1, ν_1 – числа степеней свободы; I_x – бета-распределение.

Как и выше, указанное распределение используется для формирования порога различения по допустимой вероятности ошибки первого рода в соответствии с критерием Неймана–Пирсона. Предложенный подход имеет очевидные преимущества в сравнении с традиционным [12]. Они связаны, во-первых, с тем, что необходимость в предварительном центрировании вектора признаков отпадает. А во-вторых, с тем, что снимается также вопрос о произвольности выбора порога различения.

Что касается манхэттенской метрики, то в общем случае распределение ее отдельного слагаемого q определяется соотношением [13]:

$$\rho(q) = \frac{\exp\left(-\frac{(q-q_0)^2}{2\sigma^2}\right)}{\sigma\sqrt{2\pi}} + \frac{\exp\left(-\frac{(q+q_0)^2}{2\sigma^2}\right)}{\sigma\sqrt{2\pi}}, \quad (5)$$
$$q_0 = \frac{m_1 - m_2}{\sigma},$$

где m_1, m_2 – математические ожидания нормированных компонент для сравниваемых объектов, σ^2 – дисперсия этих компонент. У объектов одного класса эти математические ожидания естественно считать одинаковыми. В этом случае $q_0 = 0$ и нормированное слагаемое имеет распределение $\rho(q) = 2\frac{\exp\left(-\frac{q^2}{2\sigma^2}\right)}{\sigma\sqrt{2\pi}}$, т.е. близко к нормальному. Его дисперсия на интервале переменной (0,1) составляет [14]:

$$d = 2\int_0^1 \frac{\exp\left(-\frac{q^2}{2\sigma^2}\right)}{\sigma\sqrt{2\pi}} q^2 dq - \left(2\int_0^1 \frac{\exp\left(-\frac{q^2}{2\sigma^2}\right)}{\sigma\sqrt{2\pi}} q dq\right)^2,$$

или, после замены переменной $u = \frac{q^2}{2\sigma^2}$:

$$d = \frac{2\sigma^2}{\sqrt{\pi}} \int_0^{\frac{1}{2\sigma^2}} \sqrt{u}\, e^{-u}\, du - \left(\sigma\sqrt{\frac{2}{\pi}} \int_0^{\frac{1}{2\sigma^2}} e^{-u}\, du\right)^2.$$

После вычисления интегралов получаем [14]:

$$d = 2\sigma^2 \gamma\left(\frac{1}{2\sigma^2}, \frac{3}{2}\right) - \frac{2\sigma^2}{\pi}\left(1 - \exp\left(-\frac{1}{2\sigma^2}\right)\right)^2, \quad (6)$$

где γ – функция гамма-распределения.

В предположении о равномерном распределении нормированных компонент векторов признаков получаем для дисперсии $\sigma^2 = \frac{1}{6}$. Подставляя это значение в (2), находим $d = 0.200$.

С учетом композиции нормальных сигналов [15] находим, что в сделанных предположениях распределение манхэттенской метрики на положи-

тельной полуоси может быть аппроксимировано следующим образом:

$$F(M_M) = 2\Phi\left(\frac{M_M}{\sqrt{D}}\right) - 1, \qquad D = nd, \tag{7}$$

где $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{x^2}{2}}dx$ – распределение Лапласа.

Таким образом, и в этом случае порог различения может быть сформирован по заданной вероятности ошибки первого рода. Использование евклидовой и манхэттенской метрик значительно упрощает проведение кластеризации в рое объектов. В принципе, ее проведение можно начинать с любого из них. К тому же классу при этом относят объекты, метрика которых относительно исходного меньше порога различения. Относительно отобранных объектов операция повторяется. В конечном итоге к классу относят все объекты из роя, метрика которых отличается от соседей менее, чем на величину порога различения. Процесс продолжается для объектов вне кластера до тех пор, пока они не будут отнесены к тому или иному классу. Настроечным параметром процесса является упомянутая выше вероятность ошибки первого рода, которая соответствует также вероятности некорректности кластеризации в целом.

ПРИМЕРЫ ФОРМИРОВАНИЯ ПЕРЕМЕННЫХ

Остановимся теперь на некоторых примерах формирования переменных различного типа. Как уже указывалось, в простейшем случае бинарная переменная отображает факт наличия или отсутствия того или иного признака или состояния номинальной переменной. В качестве примера формирования относительных переменных рассмотрим практически важный вопрос отображения в векторе признаков формы объекта в картинной плоскости. К такому отображению обычно предъявляется требование инвариантности по отношению к таким геометрическим преобразованиям как перенос, поворот и масштабирование [16].

Будем считать, что для формирования соответствующих инвариантов используется метод центральных моментов, рассчитанных относительно центра тяжести плоской фигуры, наблюдаемой в картинной плоскости. Для расчета моментов используется представление контура объекта в виде функции $r = r(\varphi)$ в полярных координатах с полюсом в центре тяжести. Считается при этом, что полярный угол изменяется в интервале $(-\pi, \pi)$, т.е. отсчитывается от отрицательной полуоси используемой системы координат в направлении против часовой стрелки. Первый нецентральный момент полярной функции рассчитывается в соответствии с соотношением вида $R = \int_{-\pi}^{\pi} r(\varphi)d\varphi$. Он представляет собой радиус круга, который приближает контур наилучшим в среднем образом. В соответствии со стандартным подходом [8], второй центральный момент рассчитывается в соответствии с соотношением $\sigma^2 = \mu_2 = \int_{-\pi}^{\pi}(r(\varphi) - R)^2 d\varphi$. Он интерпретируется как дисперсия отклонения контура от круга наилучшего приближения. Асимметрия и эксцесс контура относительно этого круга также определяются стандартным образом:

$$A = \frac{\mu_3}{\sigma^3},$$

$$E = \frac{\mu_4}{\sigma^4} - 3,$$

где μ_3, μ_4 – центральные моменты третьего и четвертого порядка.

Смысл нулевых значений асимметрии и эксцесса в данном случае следующий: контур в среднем симметричен относительно круга наилучшего приближения, а его отклонение от этого круга подчиняется нормальному закону. Инвариантность этих величин относительно упомянутых геометрических преобразований интуитивно очевидна, а методика расчета хорошо известна. Исходя из этих соображений, естественно именно их использовать в качестве относительных переменных, определяющих форму объекта в картинной плоскости. Заметим, что опыт подобного подхода в форме применения динамогенетических диаграмм “асимметрия–эксцесс” хорошо известен в гранулометрии [17].

Примером формирования порядковой переменной является оценка объема группы. Последняя представляет собой совокупность элементов – источников импульсного сигнала [18]. Очевидно, что сопровождению элементов предшествует этап их селекции, включающий оценку объема группы. Предположим, что селекция и сопровождение элементов являются апостериорными, т.е. осуществляются после накопления массива первичных данных. В этом случае для оценки объема группы может быть использован гистограммный метод [19], базирующийся на предварительном построении гистограммы для длительности импульсов или для периодов их следования.

Гистограмма для периодов следования строится в ситуации идентичности длительностей импульсов для разных объектов. Построение осуществляется в предположении, что периоды следования импульсов от разных объектов приблизительно одинаковы и лишь смещены по времени. Отсюда следует, что порядок следования импульсов от разных объектов не меняется. Это

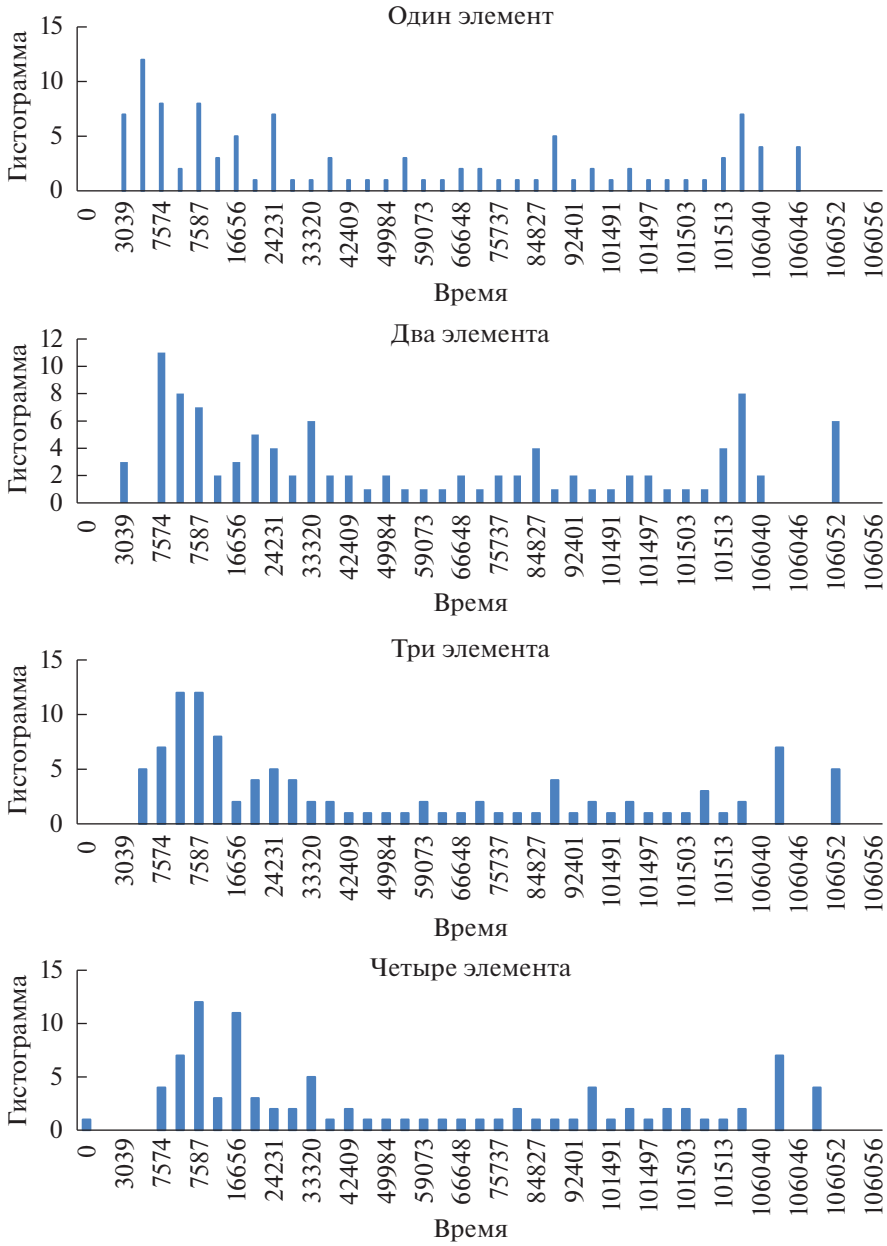

Рис. 1. Объединенные гистограммы.

предположение безусловно оправдано в случае активной локации, а также в некоторых ситуациях пассивной [20].

Особенность метода состоит в том, что гистограммы строятся для импульсов, разнесенных на заданное число номеров. При этом гистограммы, сформированные для пачек импульсов, объединялись и дальнейшая обработка осуществлялась над ними. В каждой объединенной гистограмме строились следующие вариации:

– относительно среднего момента прихода (фронта) импульса в пачке $v_{a\Delta}$;

– относительно максимальной моды в пачке $v_{t\Delta}$;

– относительно максимальной моды в стробе сопровождения $v_{\delta\Delta}$.

В качестве решающей функции для величины разнесения Δ использовалось произведение $s_\Delta = v_{a\Delta}v_{t\Delta}v_{\delta\Delta}$, а решающее правило состояло в выборе точки минимума s_Δ. Найденное значение Δ интерпретировалось как объем группы. Связано это было с тем, что минимальность s_Δ соответствует близости периодов следования исследуемых импульсов, т.е. их принадлежности одному объекту. Дополнительным преимуществом использования этой величины является ее инвари-

Таблица 1.

объекты / параметры	1	2	3	4
решающая функция	0.02517	0.017803	0.019035	0.031864
смещение строба	3045	7574	101497	101503
полуширина строба	6055.5	8328	6827.5	6827.5

антность во времени, т.е. независимость от временного масштаба разностей.

На практике для реализации описанного выше подхода использовался массив накопленных данных, включающий:

– последовательность номеров импульсов в накопленном массиве;

– моменты регистрации передних и задних фронтов наблюдаемых импульсов;

– наблюдаемые амплитуды импульсов;

– наблюдаемые частоты.

Примеры объединенных гистограмм приведены на рис. 1, а значения решающих функций и параметров сопровождения – в табл. 1. Минимальное значение решающей функции имеет место для объема группы два элемента.

Проведенное рассмотрение показывает, что, в отсутствие пеленгационной информации, принятое предположение о близости периодов следования импульсов от разных элементов является необходимым и обеспечивает решение задачи селекции в этих условиях. Отказ от него делает селекцию невозможной и требует дополнения номенклатуры данных, например, сведениями о пеленгах наблюдаемых объектов.

ВЫВОДЫ

1. В рамках распознавания образов выделяется процедура комплексирования данных, целью которой является идентификация объекта по совокупности признаков – стандартизированных показателей (параметров), определяющих образ объекта в условиях его системного окружения.

2. Образ объекта характеризуется положением в пространстве признаков, а задача идентификации эквивалентна привязке этого положения к соответствующей области, сформированной в ходе предварительной классификации. Последнюю предлагается осуществлять в ходе кластеризации имеющейся статистической выборки в ходе обучения без учителя.

3. Комплексирование данных в системах распределенного мониторинга порождает проблемы, связанные:

– с объединением разнородных данных от различных источников, зачастую слабо связанных;

– противоречивостью, неполнотой и неточностью данных в отсутствие априорной идентификации наблюдаемых объектов;

– требованием оперативной обработки большого объема разнородной информации.

4. В качестве основы для комплексирования используется функциональное шкалирование. В рамках последнего развиваются методы, предназначенные для формирования обобщенных показателей в составе объединенного вектора признаков. Номинальным признакам при этом сопоставляются бинарные показатели, порядковым – целочисленные, относительным – вещественные. Предварительная нормировка позволяет свести эти показатели к одному диапазону.

5. Возможными мерами близости в пространстве признаков являются коэффициент корреляции по Пирсону, информационная метрика, а также евклидова и манхэттенская метрики. Ценность первых двух мер ограничена, поскольку они не охватывают все типы показателей в составе объединенного вектора признаков. В то же время евклидова и манхэттенская метрики имеют установленные статистические распределения. Это позволяет формировать порог различения естественным образом – по заданной вероятности ошибки первого рода на основе критерия Неймана–Пирсона. Это обстоятельство обуславливает их предпочтительность.

СПИСОК ЛИТЕРАТУРЫ

1. *Моисеев А.А.* Модификация некоторых процедур автоматического анализа данных // Наукоемкие технологии в космических исследованиях Земли. 2017. Т. 9. № 2. С. 48–53.
2. *Моисеев А.А.* Оценка параметров наблюдения методом редукции // Радиопромышленность. 2017. № 3. С. 19–25.
3. *Бекенева Я.А.* Преобразование данных от разнородных систем мониторинга. Программные продукты и системы. 2019. Т. 32. № 2. С. 197–206.
4. *Torgerson W.* Theory and methods of scaling. NY.: Wiley, 1958. 460 p.
5. *Diday E.* Optimisationen classification automatique et reconnaissance des forms // Operations Research, 1972. V. 6. № V3. P. 61–95.
6. Statistical methods for digital computers. ed. by Enslein K. ea. N.Y.: Wiley, 1977. 464 p.

7. *Gray M.* Entropy and information theory. NY.: Springer, 2007. 289 p.
8. *Wilks S.* Mathematical statistics. NY.: Wiley, 1962. 644 p.
9. *Моисеев А.А.* Параметрическая дискриминация// Наукоемкие технологии. 2018. Т. 19. № 4. С. 20–22.
10. *Lindley D., Scott W.* New Cambridge statistical tables. Cambridge: Cambridge university press. 1995. 96 p.
11. Handbook of mathematical functions.ed. by Abramovitz M.,National bureau of standards. 1964. 1046 p.
12. Empirical process techniques for dependent data.ed. by Dehling H. ea., Boston.: Birkhauser, 2002. 383 p.
13. *Sidnyaev N., Andreytseva K.* Independence of the Residual Quadratic Sums in the Dispersion Equation with Noncentral χ2-Distribution // Applied Mathematics. 2011. V. 2. № 10. P. 1303–1308.
14. *Siegel A.* Practical business statistics. NY.: McGraw – Hill, 2000. 640 p.
15. *Kallenberg O.* Foundations of modern probability. NY.: Springer, 2002. 650 p.
16. *Mardia K.* Statistic analysis of directional data. NY.: Academic press.: 1972. 389 p.
17. *Passega R., Byramjee R.* Grain-size image of clastic deposits// Sedimentology, 1969. V. 13. № 3–4. P. 233–252.
18. *Моисеев А.А.* Апостериорное сопровождение элементов групповой цели // Радиопромышленность. 2020. Т. 30. № 2. С. 25–31.
19. *Lancaster H.* An introduction to medical statistics. N.Y.: Wiley, 1974. 305 p.
20. *Scolnik M.* Radar handbook. N.Y.: McGraw–Hill, 2008. 1352 p.

Vestnik Natsional'nogo issledovatel'skogo yadernogo universiteta "MIFI", 2020, vol. 9, no. 3, pp. 270–278

Data Fusion Method in a Distributed Monitoring System

A. A. Moiseev[#]

Research and Production Enterprise Radio Monitoring Technologies and Systems, Mytishchi, 141002 Russia

[#]*e-mail: slow.coach@yandex.ru*

Received January 28, 2020; revised March 26, 2020; accepted September 8, 2020

Abstract—Modern distributed monitoring systems include observation and detection devices of different types. Heterogeneity of corresponding data creates problems associated with such data unification, with inconsistency, deficiency, and inaccuracy of data, as well as with necessity of big information volume processing. Algorithms providing heterogeneous data fusion for convenient representation of last ones, for getting reliable conclusions, and for decision making are considered in this work. The proposed approach to heterogeneous data fusion based on associate index vector application and data fusion realization on unified formal base. Start attempts of such base creation are quite justified. In this field, the approach based on transfer from original indices to generalized ones is developed. The aims of the generalized index processing methods are as follows: (i) reduction of the vector dimension, (ii) rational index nominalization, (iii) object classification without teaching, (iv) statistical analysis of classification efficiency.

For the representation of indices, nominal (binary), ordinal (integer), and relative (real) variables normalized in range (0, 1) are used. It is demonstrated that the advantage of Euclidian and Manhattan metrics is the possibility of forming the diversity threshold based on the Neyman–Pearson criterion. Examples of indices of different types and their application are also presented.

Keywords: distributed monitoring, heterogeneous data, data fusion, functional scaling, Euclidian metric, Manhattan metric, Neyman–Pearson criterion

DOI: 10.1134/S2304487X20030062

REFERENCES

1. Moiseev A. Modifikatsiya nekotorych protsedur avtomaticheskogo analiza dannych [Some procedures modification of automatic data analysis] // H&ES (Russia), 2017. vol. 9. no. 2. pp. 48–53 (in Russian).
2. Moiseev A. Otsenka parametrov nablyudeniya metodom reduktsii Evaluation of observation parameters by reduction method] // Radio industry (Russia), 2017. no. 3. pp. 19–25 (in Russian).
3. Bekeneva Ya. Preobrazovaniye dannykh ot raznorodnykh system monitoringa [Data conversion from heterogeneous systems of monitoring] // Program products and systems. 2019. vol. 32. no. 2. pp. 197–206 (in Russian).
4. Torgerson W. Theory and methods of scaling. NY.: Wiley, 1958. 460 pp.

5. Diday E. Optimisationen classification automatiqueet reconnaissance des forms // Operations Research. 1972. vol. 6. no. V3, pp. 61–95.
6. Statistical methods for digital computers. ed. by Enslein K. ea. NY.: Wiley, 1977. 464 p.
7. Gray M. Entropy and information theory. NY.: Springer, 2007. 289 p.
8. Wilks S. Mathematical statistics. NY.: Wiley, 1962. 644 p.
9. Moiseev A. Parametricheskaya diskriminatsiya [Parametrical discrimination] // Science intensive technologies (Russia), vol. 19. N 4. 2018. pp. 20–22 (in Russian).
10. Lindley D., Scott W. New Cambridge statistical tables. Cambridge: Cambridge university press, 1995. 96 pp.
11. Handbook of mathematical functions. ed. by Abramovitz M., National bureau of standards. 1964. 1046 p
12. Empirical process techniques for dependent data. ed. by Dehling H. ea., Boston.: Birkhauser, 2002. 383 pp.
13. Sidnyaev N., Andreytseva K. Independence of the Residual Quadratic Sums in the Dispersion Equation with Noncentral χ2-Distribution // Applied Mathematics. 2011, vol. 2. no. 10. pp. 1303–1308.
14. Siegel A. Practical business statistics. NY.: McGraw – Hill, 2000. 640 pp.
15. Kallenberg O. Foundations of modern probability. NY.: Springer, 2002. 650 pp.
16. Mardia K. Statistic analysis of directional data. NY.: Academic press, 1972. 389 pp.
17. Passega R., Byramjee R. Grain-size image of clastic deposits // Sedimentology. 1969, v. 13, N 3–4, p. 233–252.
18. Moiseev A.A. Posterior tracking of multiple target elements. Radio industry (Russia), 2020, vol. 30, no. 2, pp. 25–31 (in Russian).
19. Lancaster H. An introduction to medical statistics. NY.: Wiley, 1974. 305 pp.
20. Scolnik M. Radar handbook. NY.: McGraw–Hill, 2008. 1352 pp.