УДК 004.048

Бабичев С. А.

Канд. техн. наук, доцент, доцент кафедры информатики и компьютерных наук, Херсонский национальный технический университет, Украина

ОПТИМИЗАЦИЯ ПРОЦЕССА ПРЕДОБРАБОТКИ ИНФОРМАЦИИ В СИСТЕМАХ КЛАСТЕРИЗАЦИИ ВЫСОКОРАЗМЕРНЫХ ДАННЫХ

Представлена методика выбора оптимального метода нормализации при построении кластерной структуры объектов, отличительной особенностью которых является высокая размерность признакового пространства. В качестве основного критерия оценки качества предобработки данных использовался критерий энтропия Шеннона и относительное изменение энтропии в процессе трансформации данных. Понижение размерности признакового пространства исследуемых объектов производилось при помощи компонентного анализа. Построена модель системы кластеризации с использованием алгоритма нечеткой кластеризации fuzzy C-means, при помощи которой произведена оценка качества кластеризации при использовании различных методов предобработки данных. Показано, что для исследуемых данных наилучшим методом нормализации является метод десятичного масштабирования, при котором энтропия обработанного сигнала принимает наименьшее значение, при этом в процессе трансформации данных компонентным анализом относительное изменение энтропии не превышает допустимых норм.

Ключевые слова: Кластеризация, размерность признакового пространства, нормализация, энтропия.

НОМЕНКЛАТУРА

ДНК – дезоксирибонуклеиновая кислота;

ALL – острая лимфобластная лейкемия;

AML – острая миелоидная лейкемия;

A_i – аппроксимирующие коэффициенты на i-м уровне вейвлет-декомпозиции сигнала;

СКМ система компьютерной математики;

D_i – диагональные детализирующие коэффициенты на i-м уровне вейвлет-декомпозиции сигнала;

E_{0i} – энтропия нормализованных признаков i-го объекта;

E_i – энтропия выделенных главных компонент i-го объекта;

$\sigma(i)$ – стандартное отклонение значений в i-й строке;

H_i – горизонтальные детализирующие коэффициенты на i-м уровне вейвлет-декомпозиции сигнала;

i – количество строк в массиве данных;

j – количество признаков в i-й строке;

K – положительная константа для согласования размерности;

m – количество признаков i-го объекта;

n – количество объектов;

p_i – вероятность осуществления i-го события;

sym4 – вейвлет симплет-4;

V_i – вертикальные детализирующие коэффициенты на i-м уровне вейвлет-декомпозиции сигнала;

$XN(i, j)$ – нормализованное значение j-го элемента в i-й строке;

$X(i, j)$ – исходное значение j-го элемента в i-й строке;

$X_{\max}(i)$ – максимальное значение в i-й строке;

$X_{\min}(i)$ – минимальное значение в i-й строке;

$\overline{X}_{us}(i)$ – усеченное среднее в i-й строке (отброшены 2 % максимальных и минимальных значений);

$\overline{X}_{op}$ – среднее значение элементов в априори выбранном опорном векторе исходного массива данных;

$X(op, j)$ – значение j-го элемента в опорном векторе массива;

$\hat{M}(i, j)$ – значение функции регрессии, соответствующее j-му признаку i-й строки массива.

ВВЕДЕНИЕ

В настоящее время во многих областях научных исследований возникает необходимость в разработке моделей и систем кластеризации объектов сложной природы, отличительной особенностью которых является высокая размерность признакового пространства. Одним из основных методов, используемых в данное время для сокращения размерности пространства признаков, является метод главных компонент, позволяющий существенно уменьшить количество признаков, характеризующих объект при максимальном сохранении полезной информации об объекте. Основной проблемой, возникающей при использовании данного метода, является его высокая чувствительность к методам предобработки данных, вследствие чего характер разбиения объектов на кластеры будет существенно различаться в зависимости от комбинации используемых методов предобработки. Одним из ключевых методов предобработки данных является их нормализация, в результате которой происходит приведение исходных данных к требуемому диа-

DOI 10.15588/1607-3274-2014-2-19

пазону и распределению. Вследствие этого возникает актуальная проблема создания методики обоснованного выбора метода нормализации высокоразмерных данных сложной природы, позволяющей при дальнейшем сокращении размерности признакового пространства максимизировать объективность процесса кластеризации исследуемых объектов.

1 ПОСТАНОВКА ЗАДАЧИ

При построении модели исходные данные представляются в виде матрицы, строками в которой являются исследуемые объекты, а столбцами – признаки, характеризующие соответствующий объект (1):

$$X = \{x_{ij}\},\ i = 1,...,n\ ,\quad j = 1,...,m\ . \tag{1}$$

Выбор набора признаков осуществляется в результате сужающего отображения:

$$\{X^m\} \xrightarrow{F} \{X^k\},\ k < m\ , \tag{2}$$

при котором достигается экстремум некоторого критерия качества $J_X(F)$. F в (2) представляет собой функционал преобразования множества $\{X^m\}$ в множество $\{X^k\}$, k – размерность нового признакового пространства. Каждому отображению (2) ставится в соответствие некоторое значение критерия $J_X(F)$. Результатом полученных отображений является функция $g(J_X)$. Задача сокращения размерности признакового пространства заключается в нахождении такого отображения, при котором достигается экстремум функции $g(J_X)$. В данной работе для решения поставленной задачи в качестве функции $g(J_X)$ предлагается использовать зависимость критерия энтропии Шеннона от используемого метода нормализации [1], определяемую по формуле (3):

$$H = -K \cdot \sum_{i=1}^{n} p_i \ln(p_i). \tag{3}$$

Значения критерия энтропии вычисляется для оригинального ненормализованного сигнала, после нормализации, и для главных компонент исследуемых объектов. В соответствии с принципом максимума энтропии [2]: «Для данных знаний в форме ограничений существует только одно распределение, удовлетворяющее этим ограничениям, которое можно выбрать с помощью процедуры, удовлетворяющей «аксиомам согласованности». Это уникальное распределение определяется максимизацией энтропии». В соответствии с вышесказанным предлагается следующая методика проведения эксперимента:

– фильтрация исходных данных с целью минимизации шумовой компоненты исследуемых данных;

– вычисление энтропии Шеннона ненормализованного сигнала;

– нормализация данных различными методами нормализации;

– вычисление энтропии нормализованных данных;

– вычисление главных компонент каждой матрицы нормализованных данных;

– вычисление энтропии векторов главных компонент исследуемых объектов;

– кластеризация тестового множества исследуемых объектов методом fuzzy C-means;

– построение графиков полученных результатов, их интерпретация и анализ.

2 ОБЗОР ЛИТЕРАТУРЫ

Анализ публикаций по обозначенной проблеме [3–7] показывает, что большинство методов и алгоритмов кластеризации, используемых в настоящее время в различных областях человеческой деятельности, ориентированы на небольшую размерность вектора признаков исследуемого объекта (не более 1000). В [8, 9] процесс кластеризации представляется в виде модели, что позволило перенести в теорию кластерного анализа все основные методы теории самоорганизации моделей на основе метода группового учета аргументов, а именно: многоэтапность поиска лучшей кластеризации; критериальный подход для оценки качества кластеризации; использование методов формирования признакового пространства и формирования кластеров; выбор мер сходства между объектами, кластерами и объектом и кластером. В последнее время с развитием методов биоинформатики созданы базы данных клеток биологических объектов, особенностью которых является высокая размерность (≈ 80000) и высокая степень зашумленности, определяемая биологическими и технологическими факторами, возникающими в процессе подготовки и проведения эксперимента по их формированию [10, 11]. Вследствие этого возникает необходимость в разработке эффективных для данного типа данных методов их предобработки и сокращения размерности признакового пространства без существенной потери информации об исследуемых объектах. Методам нормализации высокоразмерных данных биологической природы посвящены работы [12, 13]. В общем случае все методы нормализации можно разделить на две подгруппы: методы, использующие опорное множество эталонного объекта и методы, использующие всю совокупность исследуемых данных. В первом случае методы нормализации подразделяются на линейные и нелинейные, во втором случае используют метод циклической локальной регрессии, метод контрастов и квантильную нормализацию. В работе [13] представлены результаты сравнительного анализа существующих методов нормализации на примере выборок данных, взятых с двух микромассивов ДНК. Результаты их исследований показали, что методы, не использующие эталон, дают лучшее качество нормализации данных. При попарной нормализации более эффективным оказался метод квантильной нормализации. Кроме того, метод квантильной нормализации ока-

зался самым быстрым из используемых методов. Однако, несмотря на достигнутые успехи в данной предметной области существует ряд нерешенных или частично решенных проблем.

К нерешенным частям общей проблемы относится отсутствие эффективной методики выбора комплекса методов предобработки информации и сокращения размера признакового пространства в системах кластеризации высокоразмерных данных сложной биологической природы.

Целью статьи является разработка методики выбора оптимального метода нормализации на этапе предварительной обработки высокоразмерных данных сложной биологической природы с последующим сокращением размерности признакового пространства и кластеризацией исследуемых объектов, основным критерием оценки качества обработки информации в которой является энтропия Шеннона.

3 МАТЕРИАЛЫ И МЕТОДЫ

В соответствии с предложенной методикой проведения эксперимента первым этапом является фильтрация данных, целью которой является минимизация шумовой компоненты исследуемых данных. Фильтрация данных осуществлялась при помощи вейвлетов, представляющих собой локализованные в пространстве функции, способные отслеживать и нужным образом обрабатывать локальные особенности исследуемых объектов. Эффективность использования вейвлетов для предварительной обработки высокоразмерных данных обусловлена характером распределения признакового пространства исследуемых объектов. Например, в случае анализа микромассивов ДНК признаками являются уровни экспрессии соответствующих генов, при этом гены на микромассиве распределяются так, что различающиеся гены с различным уровнем экспрессии будут чередоваться с частотой, существенно ниже частоты фоновой компоненты освещенности и частоты фоновых помех. Уровень экспрессии генов можно оценить по интенсивности освещенности в той или иной точке. В процессе вейвлет-декомпозиции из исходного сигнала выделяется его низкочастотная компонента, при этом сохраняются локальные особенности исходных данных. Основная идея вейвлет-декомпозиции сигнала заключается в следующем: исходное пространство интенсивностей освещенностей, соответствующее различным уровням экспрессии генов, разлагается на систему подпространств так, что каждое последующее подпространство является вложенным в предыдущее (4):

$$I \subset I_1 \subset I_2 \subset I_3 \subset \ldots \subset I_n. \quad (4)$$

Полученные подпространства при этом должны удовлетворять свойству (5):

$$\bigcap_{i=0}^{n} I_i = \{0\}. \quad (5)$$

Каждое из полученных подпространств разлагается, в свою очередь, с использованием вейвлет-функции и масштабирующей функции на пространства аппроксимирующих и детализирующих коэффициентов (6):

$$I_i = \{A_i, H_i, V_i, D_i\}. \quad (6)$$

В случае, если объект характеризуется одним вектором признаков, результатом вейвлет-декомпозиции будет вектор аппроксимирующих и один вектор детализирующих коэффициентов. Аппроксимирующие коэффициенты несут информацию о низкочастотной составляющей сигнала, детализирующие коэффициенты содержат информацию о высокочастотной составляющей сигнала. В большинстве случаев полезный сигнал содержится в низкочастотной составляющей, а детализирующие коэффициенты несут информацию о локальных особенностях сигнала. Шумовая составляющая представляет собой высокочастотную компоненту, которая так же содержится в детализирующих коэффициентах. Вследствие вышесказанного можно сделать вывод, что обработав нужным образом детализирующие коэффициенты, и восстановив сигнал из аппроксимирующих коэффициентов и обработанных детализирующих коэффициентов, получаем матрицу признаков исследуемых объектов при минимальном уровне шумовой компоненты, что способствует повышению точности решения поставленной задачи.

В [14] представлены результаты исследований по определению типа вейвлета, уровня вейвлет-декомпозиции и значений соответствующих коэффициентов в системе фильтрации хроматограмм. В качестве основного критерия оценки качества фильтрации использовался критерий энтропии. В данном случае информацией являлась шумовая компонента сигнала, вследствие чего оптимальный набор параметров фильтра определялся из условия минимума энтропии. С учетом результатов, полученных в [14], при построении модели кластеризации объектов на этапе фильтрации использовался вейвлет симплет-4 при четвертом уровне вейвлет декомпозиции с мягкой обработкой детализирующих коэффициентов и значением порогового коэффициента, равного 4.

Целью этапа нормализации является приведение эмпирических данных микромассивов и их характеристик к одинаковому диапазону и распределению, чаще всего к нормальному. При проведении исследований использовались следующие методы нормализации данных:

– минимаксная нормализация (7):

$$XN(i,j) = \frac{X(i,j) - X_{\min}(i)}{X_{\max}(i) - X_{\min}(i)}; \quad (7)$$

– десятичное масштабирование (8):

$$XN(i,j) = \frac{X(i,j)}{10^k}, \quad (8)$$

где параметр k подбирается так, чтобы максимальное значение элемента в массиве было меньше 1;

– нормализация при помощи стандартного отклонения (9):

$$XN(i,j)=\frac{X(i,j)-\overline{X}_{us}(i)}{\sigma(i)}; \quad (9)$$

– метод линейной нормализации, предложенный разработчиками Affymetrix применительно к анализу микромассивов ДНК (10):

$$XN(i,j)=\frac{\overline{X}_{op}}{\overline{X}_{us}(i)}X(i,j); \quad (10)$$

– нелинейная нормализация с использованием логистической передаточной функции (11):

$$XN(i,j)=\frac{1}{1+\exp\left(-\frac{X(i,j)-\overline{X}_{us}(i)}{\sigma(i)}\right)}; \quad (11)$$

– метод контрастов, в котором предполагается что между векторами M и A существует линейная регрессионная зависимость. Векторы M и A для каждого элемента массива данных вычисляются по формулам (12) и (13):

$$M(i,j)=\log_2\left(\frac{X(i,j)}{X(op,j)}\right); \quad (12)$$

$$A(i,j)=\log_2\left(X(i,j)\cdot X(op,j)\right). \quad (13)$$

Нормирующая поправка вычисляется по формуле (14):

$$\delta M(i,j)=M(i,j)-\hat{M}(i,j). \quad (14)$$

Нормализованное значение вычисляется по формуле (15):

$$XN(i,j)=2^{\left(A(i,j)+\frac{\delta M(i,j)}{2}\right)}; \quad (15)$$

– квантильная нормализация основана на предположении, что признаки, характеризующие n-объектов, имеют одинаковые распределения данных. Тогда график квантилей в n-мерном пространстве будет представлять собой линию, лежащую вдоль диагонали, координаты которой представлены вектором: $\left(\frac{1}{\sqrt{n}},\ldots,\frac{1}{\sqrt{n}}\right)$. Нормализованные значения признаков исследуемых векторов представляют собой проекции точки n-мерного графика квантилей на данную диагональ.

Оценка качества предобработки признаков исследуемых объектов оценивалось посредством анализа характера изменения критерия «Энтропия Шеннона» (3). Если под текущим событием понимать значение признака j в векторе i $X(i,j)$, то вероятность осуществления данного события $p(i,j)=X(i,j)^2$. Тогда учитывая, что как мера количества информации в исследуемом векторе энтропия является величиной безразмерной, формулу (3) для i-й строки можно представить как (16):

$$H_i=-\sum_{j=1}^{c}X(i,j)^2\cdot\ln\left(X(i,j)^2\right). \quad (16)$$

В соответствии с концепцией связи энтропии и информации о состоянии системы максимальное количество информации соответствует минимуму энтропии, при этом полная информация соответствует нулевой энтропии. Тогда естественно предположить, что наиболее качественная предобработка признаков объекта может быть выбрана из условия минимума энтропии Шеннона, что соответствует максимальной информации об исследуемых объектах, при этом в процессе проведения компонентного анализа изменение энтропии должно быть минимальным, что свидетельствует о минимальной потере информации в процессе выделения главных компонент.

4 ЭКСПЕРИМЕНТЫ

Моделирование процесса кластеризации исследуемых объектов производилось с использованием программной платформы KNIME-2.10.1, которая является свободно доступным программным продуктом и предназначена для обработки данных различной природы и извлечения с них информации, а также для моделирования всевозможных процессов и систем. Структурная блок-схема используемой модели кластеризации представлена на рисунке 1. В качестве экспериментальной базы для проведения исследований использовалась база данных больных лейкемией [15], представляющая собой массив размером 72×7131. Каждая строка содержит информацию об уровне экспрессии генов больных клеток отдельного человека. Объекты делятся на два класса. Один класс представляет собой пробы клеток, взятых у больных острой миелоидной лейкемией, а дугой представляет клетки больных острой лимфобластной лейкемией.

Фильтрация исходных данных, их нормализация каждым из вышеописанных методов и выделение главных компонент производилось с использованием CKM MATLAB. Этот выбор определялся тем обстоятельством, что платформа KNIME имеет ограничения на размерность признакового пространства исследуемых объек-

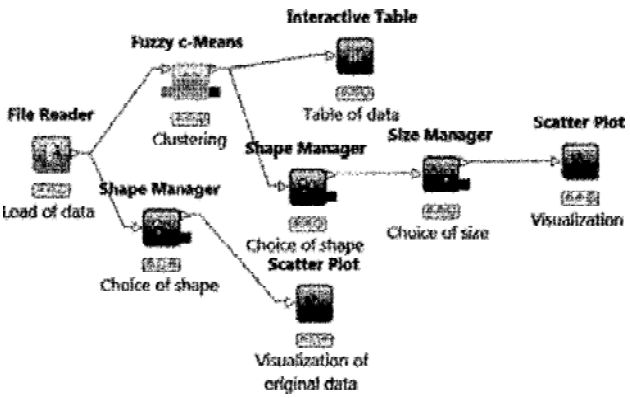


Рисунок 1 – Структурная блок-схема модели кластеризации объектов

тов. Далее матрица главных компонент загружалась в программную среду KNIME и подвергалась обработке с целью кластеризации исследуемых объектов. В качестве алгоритма кластеризации был выбран алгоритм нечеткой кластеризации C-means. Этот выбор определяется характером распределения объектов в пространстве. Вариации признаков исследуемых объектов в силу их природных особенностей не позволяют однозначно отнести объект к тому или иному кластеру. Имеет смысл только утверждение о степени принадлежности объекта тому или иному кластеру, вследствие чего для анализа такого типа данных наиболее целесообразны нечеткие методы кластеризации, одним из которых является метод C-means.

5 РЕЗУЛЬТАТЫ

Диаграмма распределения процентного содержания корректно распределенных объектов в зависимости от используемого метода нормализации представлена на рис. 2. На рис. 3 представлены графики распределения энтропии исследуемых объектов для оригинального и нормализованного различными методами сигналов. На рис. 4 показаны аналогичные графики для главных компонент исследуемых объектов.

Для оценки степени изменения энтропии в процессе вычисления главных компонент исследуемого объекта рассчитывалось относительное изменение энтропии в процессе трансформации исследуемых векторов признаков объектов (17):

$$\frac{dE_i}{E_{0i}} = \frac{\left|E_i - E_{0i}\right|}{\left|E_{0i}\right|}. \qquad (17)$$

Графики относительного изменения энтропий нормализованного и трансформированного векторов при использовании различных методов нормализации представлены на рис. 5.

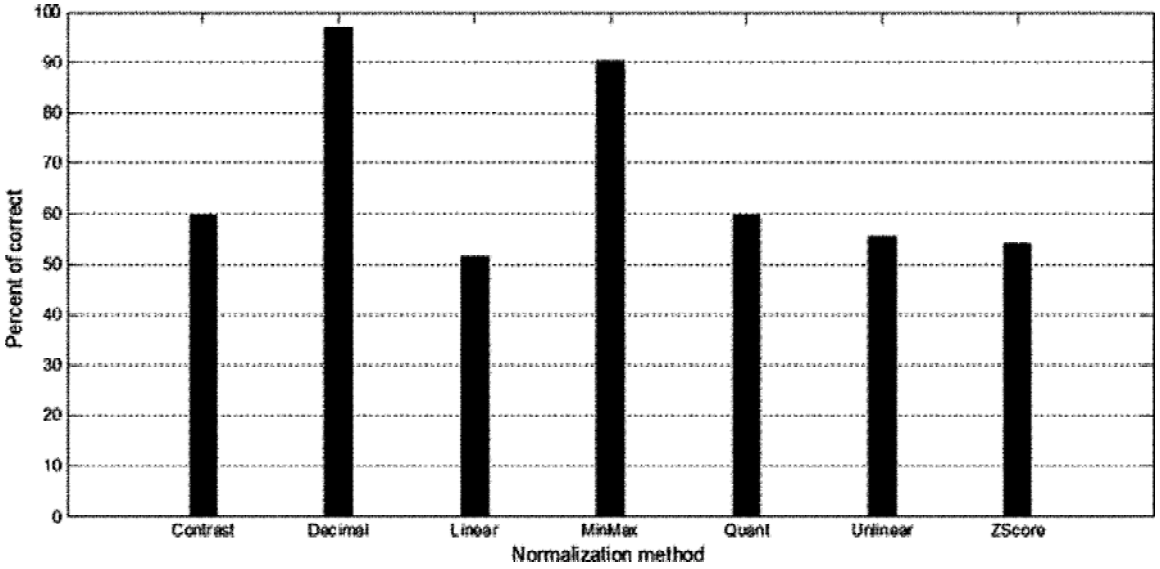


Рисунок 2 – Диаграмма распределения процентов корректно распределенных по кластерам объектов в зависимости от метода нормализации

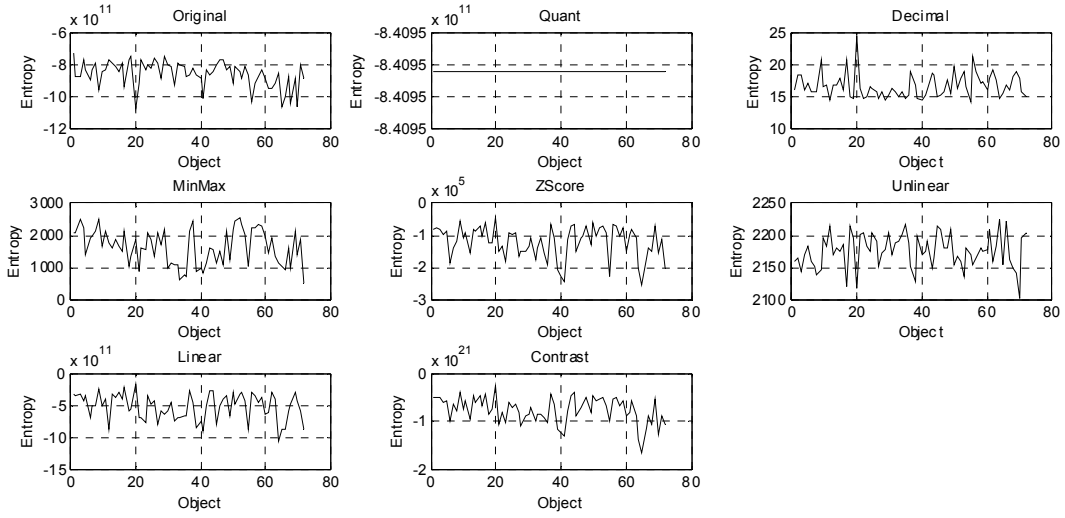


Рисунок 3 – Графики энтропий оригинального и нормализованного различными методами сигналов: Entropy – значение энтропии Шеннона, Object – исследуемый объект

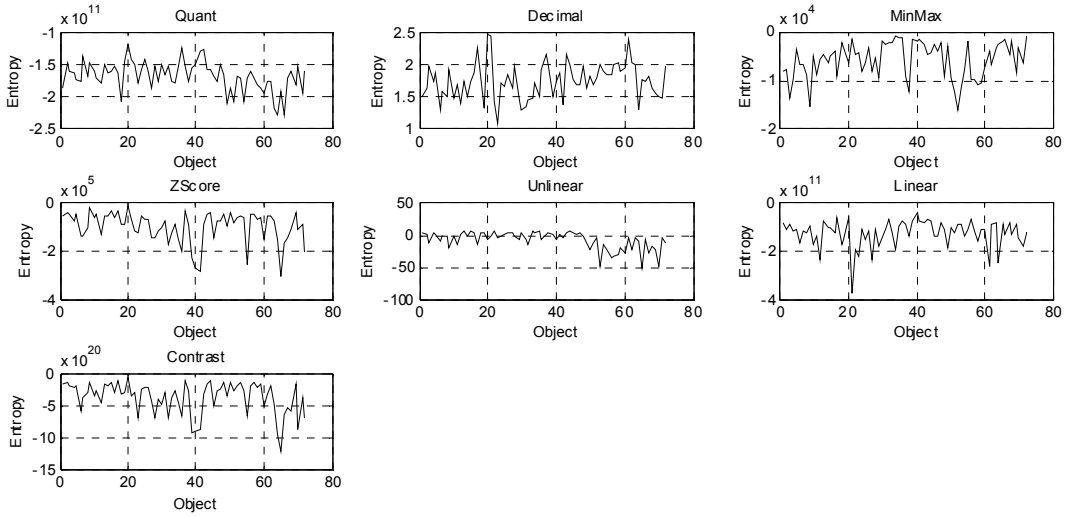


Рисунок 4 – Графики энтропий главных компонент исследуемых объектов

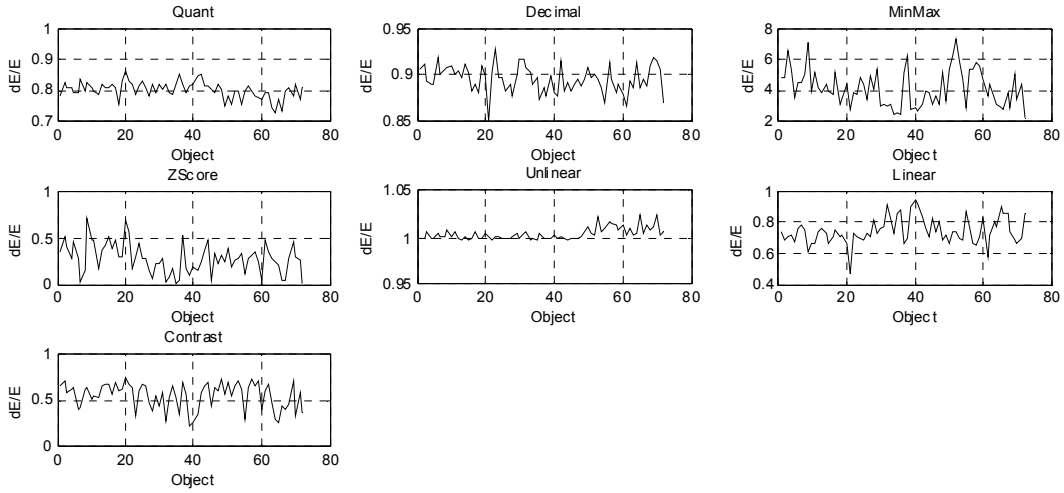
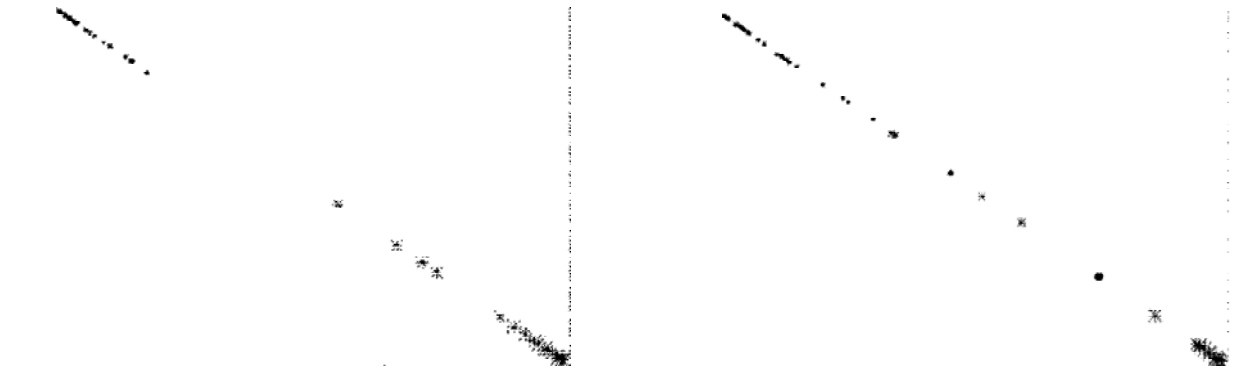


Рисунок 5 – Графики относительного изменения энтропии при использовании различных методов нормализации: dE/E – относительное изменение энтропии; Object – исследуемый объект

6 ОБСУЖДЕНИЕ

Анализ рис. 3 позволяет сделать вывод, что наименьшую энтропию имеют вектора, нормализованные методами decimal-scaling, min-max и unlinear. При этом метод десятичного масштабирования оказывается по критерию энтропии более предпочтительным. Однако абсолютное значение критерия энтропии не является объективным, поскольку диапазон варьирования значений признаков исследуемых объектов изменяется в зависимости от используемого метода нормализации, что оказывает непосредственное влияние на абсолютное значение критерия энтропии.

О величине потери информации вследствие трансформации вектора (вычислении главных компонент) можно судить по относительному критерию изменения энтропии. Из анализа графиков, изображенных на рис. 3–5 следует, что при использовании методов квантильной нормализации, десятичного масштабирования, Z-масштабирования, линейной нормализации и нормализации методом контрастов происходит приблизительно одинаковая потеря информации. Однако по сумме двух критериев метод десятичного масштабирования является наиболее предпочтительным для нормализации исследуемого типа сигналов. При использовании минимаксного и нелинейного методов нормализации наблюдается более существенное возрастание энтропии, а значит и большая потеря информации об исследуемом объекте, что подтверждается результатами кластеризации, представленными на рис. 2. На рис. 6 показаны диаграммы распределения объектов по кластерам при использовании decimal-scaling и min-max методов нормализации. Анализ распределения объектов по кластерам, представленного на рис. 6, подтверждает полученные результаты о более высокой эффективности нормализации методом decimal-scaling по сравнению с другими используемыми методами при кластеризации объектов, отличительной особенностью которых является высокая размерность признакового пространства.



а б

Рисунок 6 – Результаты разбиения объектов на кластеры при использовании: а – decimal-scaling нормализации; б – min-max нормализации

ВЫВОДЫ

В работе получила дальнейшее развитие технология использования критерия «Энтропия Шеннона» в качестве информационного критерия для оценки количества полезной информации в исследуемых данных, отличительной особенностью которых является высокая размерность пространства признаков. Предложен комплексный подход по оценке качества обработки информации на основе критерия энтропии и относительного изменения энтропии в процессе трансформации данных с использованием компонентного анализа. Как показали исследования, метод главных компонент имеет высокую чувствительность к методам нормализации данных на этапе их предобработки, вследствие чего высокую актуальность приобретают исследования по оптимизации использования методов предобработки высокоразмерных данных с целью повышения качества их кластеризации.

Практический интерес представляет применение предложенной технологии для обработки микромассивов ДНК, объектами в которых являются клетки больного органа, а признаками – уровень экспрессии генов, определяющих состояние соответствующей клетки. Особенностью исследуемых данных является высокий уровень шума и высокая размерность пространства признаков, что ограничивает использование традиционных методов обработки информации. Результаты исследований, показали, что для данных биологической природы, представляющих собой уровни экспрессии генов больных клеток, наименьшая потеря информации наблюдается при использовании decimal-scaling метода нормализации данных. Количество правильно распределенных по кластерам объектов в этом случае составляет 97%. Исследования также показали, что в случае отсутствия априорной информации о принадлежности объекта тому или иному кластеру оптимальную комбинацию методов предобработки можно выбрать на основании критерия Энтропия Шеннона. Значения данного критерия для наиболее качественно предобработанного сигнала является минимальным, что соответствует максимальному количеству сохраненной информации. Кроме того, относительное изменение энтропии в процессе трансформации сигнала компонентным анализом в случае применения метода decimal-scaling не превышает допустимого значения, что свидетельствует о незначительной потере информации в процессе трансформации данных.

Перспективными направлениями дальнейших исследований является нахождение аналитического выражения комплексного критерия оценки качества предобработки информации, одну из составляющих которого будет составлять энтропия Шеннона.

БЛАГОДАРНОСТИ

Работа выполнена в рамках госбюджетной научно-исследовательской темы Херсонского национального технического университета «Исследование искусственных иммунных систем и методов мультифрактального анализа в задачах идентификации и классификации биологических сигналов» (номер государственной регистрации 0109U009003).

СПИСОК ЛИТЕРАТУРЫ

1. Shannon C. E. A mathematical theory of communication / C. E. Shannon // Bell System Technical Journal. – 1948. – Vol. 27. – P. 379–423, 623–656.
2. Shore J. E. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy / J. E. Shore, R. W. Johnson // IEEE Transactions on Information theory. – 1980. – Vol. IT–26. – P. 26–37.
3. Data Analysis of Bio-Medical Data Mining using Enhanced Hierarchical Agglomerative Clustering / [Krishnaiah J. V., Chandra Sekar D. V., Ramchand K., Rao H.] // International Journal of Engineering and Innovative Technology. – 2012. – Vol. 2, Issue 3. – P. 43–49.
4. Liang J. Computational analysis of microarray gene expression profiles: clustering, classification, and beyond / J. Liang, S. Kachalo // Chemometrics and Intelligent Laboratory Systems. – 2002. – No. 62. – P. 199–216.
5. Rezankova H. Cluster analysis of economic data / H. Rezankova // Statistica. – 2014. – No. 94(1). – P. 73–86.
6. Li Y. Text document clustering based on frequent word meaning sequences / Y. Li, S. M. Chung, J. D. Holt // Data & Knowledge Engineering. – 2008. – No. 64(1). – P. 381–404.
7. Jain A. K. Data clustering: A review / A. K. Jain, M. N. Murty, P. J. Flynn // ACM Computing Surveys. – 1999. – Vol. 31, No. 3. – P. 264–323.
8. Ивахненко А. Г. Объективная кластеризация на основе теории самоорганизации моделей / А. Г. Ивахненко // Автоматика. – 1987. – № 5. – С. 6–15.

9. Ивахненко А. Г. Алгоритмы метода группового учета аргументов (МГУА) при непрерывных и бинарных признаках / А. Г. Ивахненко. – К. : Институт кибернетики АН Украины, 1992. – 49 с.
10. Ивахно С. С. Методы кластеризации в программе Microarraytool для анализа данных ДНК-микроарреев / С. С. Ивахно, А. И. Корнелюк, О. П. Минцер // Медична інформатика та інженерія. – 2008. – № 3. – С. 33–40.
11. Івахно С. С. Огляд технологій та аналіз даних / С. С. Івахно, О. І. Корнелюк // Український біохімічний журнал. – 2004. – № 2 (76). – С. 5–19.
12. Bioinformatics and Computational Biology. Solutions Using R and Bioconductor /[R. Gentleman, V. Carey, W. Huber et al.]. – New York : Springer, 2005. – 473 p.
13. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias / [Bolstad B. M., Irizarry R. A., Astrand M., Speed T. P.] // Bioinformatics. – 2003. – Vol. 19. – P. 185–193.
14. Фильтрация хроматограмм с помощью вейвлет-анализа с использованием критерия энтропии / [С. А. Бабичев, Н. И. Бабенко, А. А. Дидык и др.] // Системные технологии. – 2010. – № 6 (71). – С. 17–22.
15. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring / [T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, et al.] // Science. – 1999. – Vol. 286. – P. 531–537.

Статья поступила в редакцию 24.10.2014.
После доработки 04.11.2014.

Бабічев С. А.
Канд. техн. наук, доцент, доцент кафедри інформатики та комп'ютерних наук, Херсонський національний технічний університет, Україна

ОПТИМІЗАЦІЯ ПРОЦЕСУ ПЕРЕДОБРОБКИ ІНФОРМАЦІЇ У СИСТЕМАХ КЛАСТЕРИЗАЦІЇ ВИСОКОРОЗМІРНИХ ДАНИХ

Представлено методику вибору оптимального методу нормалізації при побудові кластерної структури об'єктів, відмінною особливістю яких є висока розмірність простору ознак. Як основний критерій оцінки якості передобробки даних використовувався критерій ентропія Шеннона і відносна зміна ентропії у процесі трансформації даних. Зниження розмірності простору ознак досліджуваних об'єктів здійснювалося за допомогою компонентного аналізу. Побудовано модель системи кластеризації з використанням алгоритму нечіткої кластеризації fuzzy C-means, за допомогою якої зроблено оцінку якості кластеризації при використанні різних методів передобробки даних. Показано, що для досліджуваних даних найкращим методом нормалізації є метод десяткового масштабування, при якому ентропія обробленого сигналу приймає найменше значення, при цьому в процесі трансформації даних компонентним аналізом відносна зміна ентропії не перевищує допустимих норм.

Ключові слова: кластеризація, розмірність простору ознак, нормалізація, ентропія.

Babichev S. A.
PhD., Associate Professor, Department of Informatics and Computer Science, Kherson National Technical University, Kherson, Ukraine

OPTIMIZATION OF INFORMATION PREPROCESSING IN CLUSTERING SYSTEMS OF HIGH DIMENSION DATA

The methodic of choice of optimal normalization method for object cluster structure of creation, with high dimension of feature space, is shown. The Shannon entropy criterion and entropy relative change were used as main criterions of estimating the data preprocessing quality during the data transformation. Decreasing of feature space dimension of tested objects was realized by component analysis. Model of system clustering with the use of fuzzy C-means algorithm was constructed, which the help of whith the estimate of clustering quality was established by the use of different data preprocessing methods. It's shown that the best normalization method for tested data is decimal-scaling method, by which the entropy of processed signal gets minimal significance, and relative change of entropy doesn't exceed permissible norms during the process of data transformation by component analysis.

Keywords: clustering, the feature space dimension, normalization, entropy.

REFERENCES

1. Shannon C. E. A mathematical theory of communication, *Bell System Technical Journal*, 1948, Vol. 27, pp. 379–423, 623–656.
2. Shore J. E., Johnson R. W. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy, *IEEE Transactions on Information theory,* 1980, Vol. IT–26, pp. 26–37.
3. Krishnaiah J. V., Chandra Sekar D. V., Ramchand K., Rao H. Data Analysis of Bio-Medical Data Mining using Enhanced Hierarchical Agglomerative Clustering, *International Journal of Engineering and Innovative Technology,* 2012, Vol. 2, Issue 3, pp. 43–49.
4. Liang J., Kachalo S. Computational analysis of microarray gene expression profiles: clustering, classification, and beyond, *Chemometrics and Intelligent Laboratory Systems,* 2002, No. 62, pp. 199–216.
5. Rezankova H. Cluster analysis of economic data, *Statistica,* 2014, No. 94(1), pp. 73–86.
6. Li Y. Text document clustering based on frequent word meaning sequences, *Data & Knowledge Engineering,* 2008, No. 64(1), pp. 381–404.
7. Jain A. K., Murty M. N., Flynn P. J. Data clustering: A review, *ACM Computing Surveys,* 1999, Vol.31, No. 3, pp. 264–323.
8. Ivahnenko A. G. Objektivnaja klasterizacija na osnove teorii samoorganizacii modelej, *Avtomatika*, 1987, No.5, pp. 6–15.
9. Ivahnenko A. G. Algoritmy metoda gruppovogo ucheta argumentov (MGUA) pri nepreryvnyh i binarnyh priznakah. Kiev, Institut kibernetiki AN Ukrainy, 1992, 49 p.
10. Ivahno S. S., Korneljuk A. I., Mincer O. P. Metody klasterizacii v programme Microarraytool dlja analiza dannyh DNK-mikroarreev, *Medichna informatika ta inzhenerija,* 2008, No. 3, pp. 33–40.
11. Ivahno S. S., Korneljuk O. I. Ogljad tehnologij ta analiz danih, *Ukraïnskij biohimichnij zhurnal*, 2004, No. 2(76), pp. 5–19.
12. Gentleman R. Carey V., Huber W., Irizarry R., Dudoit S. Bioinformatics and Computational Biology. Solutions Using R and Bioconductor. New York, Springer, 2005, 473 p.
13. Bolstad B. M., Irizarry R. A., Astrand M., Speed T. P. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias, *Bioinformatics*, 2003, Vol. 19, pp. 185–193.
14. Babichev S. A., Babenko N. I., Didyk A. A., Litvinenko V. I., Fefelov A. A., Shkurdoda S. V. Fil'tracija hromatogramm s pomoshh'ju vejvlet-analiza s ispol'zovaniem kriterija jentropii, *Sistemnye tehnologii,* 2010, No. 6(71), pp. 17–22.
15. Golub T. R., Slonim D. K., Tamayo P., Huard C., at al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *Science,* 1999, Vol. 286, pp. 531–537.