УДК 004.8+004.93’14+519.25
doi:10.18720/SPBPU/2/id24-204

Больщиков Виталий Андреевич,
аспирант

ИСПОЛЬЗОВАНИЕ САМООРГАНИЗУЮЩИХСЯ КАРТ КОХОНЕНА ДЛЯ МЕТРОЛОГИЧЕСКИ ОБОСНОВАННОЙ КЛАСТЕРИЗАЦИИ РЕЗУЛЬТАТОВ ИЗМЕРЕНИЙ

Россия, Санкт-Петербург,
Санкт-Петербургский политехнический университет Петра Великого,
boltshikov.va@edu.spbstu.ru

Аннотация. В данной работе предложен алгоритм метрологически обоснованной кластеризации неточных данных на основе самоорганизующихся карт Кохонена. Выполнены численные исследования, показавшие устойчивость результатов представленного метода при комбинировании его с классическими методами кластеризации.

Ключевые слова: кластеризация, обработка неточных данных, метрологически обоснованная кластеризация, самоорганизующиеся карты Кохонена.

Vitaly A. Bolschikov,
Postgraduate (PhD) Student

USING SELF-ORGANIZING MAPS FOR METROLOGICALLY REASONABLE CLUSTERING OF MEASUREMENT RESULTS

Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia,
boltshikov.va@edu.spbstu.ru

Abstract. This paper proposes an algorithm for metrologically reasonable clustering of inaccurate data based on self-organizing maps. Numerical studies have been carried out to show the stability of the results of the presented method when combined with classical clustering methods.

Keywords: clustering, processing of inaccurate data, metrologically reasonable clustering, Kohonen self-organizing maps.

Введение

Ускоренные темпы проводимой глубокой цифровизации технологических процессов на промышленных производствах влекут за собой соответствующий рост объема собираемой информации. Поскольку получаемые результаты измерений должны выступать основой в дальнейшем для выработки решений, возникает необходимость выполнения их специальной математической обработки, сводящейся к задаче анализа больших объемов неразмеченных данных для выявления скрытых взаимосвязей и внутренней структуры. Для решения данной задачи принято использовать алгоритмы кластеризации, суть которых заключается в выполнении отображения множества $\mathbf{X} = \{\mathbf{x}_i\}$, образованного полученными векторами $\mathbf{x}_i$ результатов измерений состояния технологического процесса, на множество C, образованного перечнем допустимых классов, для каждого из которых определены алгоритмы принятия решений, т. е. в выполнении операции

$$a : \mathbf{X} \to \mathbf{X}^* \to \mathrm{C},$$

где $C = \{c_1, c_2, \ldots, c_k, \ldots, c_q\}$ — множество кластеров, удовлетворяющих условию $c_k = \left\{ \mathbf{x}_i, \mathbf{x}_j \middle| \mathbf{x}_i \in \mathbf{X}, \mathbf{x}_j \in \mathbf{X}, d(\mathbf{x}_i, \mathbf{x}_j) < \varepsilon \right\}$; $d(\mathbf{x}_i, \mathbf{x}_j)$ — расстояние между векторами $\mathbf{x}_i$ и $\mathbf{x}_j$; $\varepsilon > 0$ — величина допустимой степени близости элементов внутри кластера; $\mathbf{X}^*$ — множество результатов предварительной обработки элементов множества $\mathbf{X}$, заключающейся в снижении размерности векторов $\mathbf{x}_i$ и их преобразовании к виду, обеспечивающему наибольшую контрастность последующей кластеризации.

Таким образом, математическая обработка результатов измерений, получаемых при изучении поведения сложных объектов с целью формирования их математического описания, обязательно включает в себя задачу кластеризации как естественный способ выделить группы близких

состояний, характеризующихся схожим образом. Получение соответствующей классификации позволяет в дальнейшем различить состояния или режимы работы объекта и соответствующим образом выполнить управление им или принять необходимые решения.

1. Метрологически обоснованная кластеризация

Поскольку обрабатываемые данные представляют собой результаты измерений, то, как впервые было наглядно проиллюстрировано в [1–2], кластеризация измерительной информации обязана учитывать неопределенность исследуемых данных, вызванную их погрешностями. Учет метрологических характеристик анализируемых количественных сведений при выполнении кластерного анализа измерительных данных позволяет согласовать получаемые результаты с точностью обрабатываемых данных, снижая чувствительность результатов к возможным выбросам и риск возникновения необоснованных интерпретаций по результатам выполнения исследования. Выработка подходов к кластеризации неточной информации является достаточно обсуждаемой в литературе задачей [3–10], однако систематически данная задача так и не решена. В практике метрологии эту процедуру в основном производят без принятия во внимание погрешности исходных данных, что часто приводит к неверным выводам. Предложенные в работах [11–13] общие подходы к проведению кластерного анализа с учетом метрологических характеристик изучаемых сведений были ранее реализованы [14–17] и применены автором для широко известных и наиболее популярных подходов и алгоритмов кластеризации информации — комбинации метода главных компонент PCA, анализа сингулярного спектра SSA или преобразования Фурье FFT в качестве алгоритмов получения элементов $\mathbf{X}^*$, содержащих результаты предобработки и снижения размерности элементов множества $\mathbf{X}$, и методов кластеризации k-means, DBSCAN, иерархических методов кластеризации Single-linkage или методов кластеризации MST на основе построения минимального остовного дерева.

2. Использование самоорганизующихся карт Кохонена

С ростом вычислительных возможностей стало возможным применение нейросетевых алгоритмов анализа исследуемых данных для получения оценки внутренних скрытых в них зависимостей. Для выполнения кластеризации оказывается удобным задействовать не широко используемые нейросетевые модели с обучением с учителем, а так называемые самоорганизующиеся карты Кохонена (SOM), которые представляют собой класс алгоритмов, для которых не требуется обучающая выборка, что и делает их популярными при проведении кластерного анализа. В настоящей работе рассматривается возможность использования само-

организующихся карт Кохонена для метрологически обоснованной кластеризации. Данная работа продолжает исследования, начатые в [11].

При проведении метрологически обоснованной кластеризации с использованием самоорганизующихся карт Кохонена алгоритмы настройки весов нейронов соответствует описанию, представленному в [11]. Пример результатов процесса обучения синаптических весов нейронов, настроенных на двумерные вектора $\mathbf{x}_i^* \in \mathbf{X}^* \subseteq R^2$, подлежащие кластеризации, проиллюстрированы на рисунке 1.

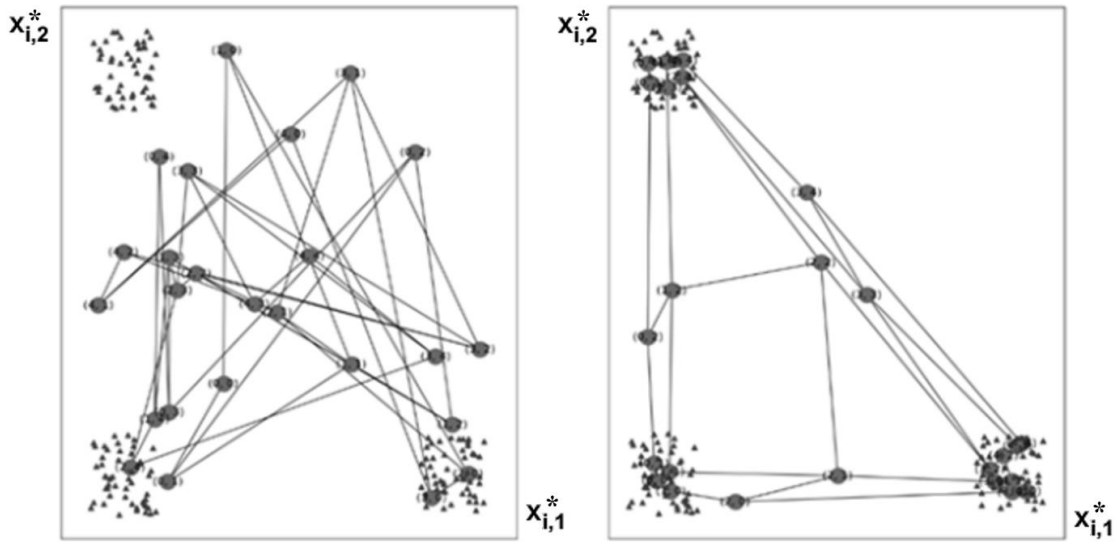

Рис. 1. Синаптические связи нейронов сети Кохонена, наложенные на плоскость обрабатываемых данных: а) – в начале процесса обучения (случайные значения $\mathbf{X}^*$); б) – в конце процесса обучения

На рисунке 1 маленькими точками отмечены значения векторов $\mathbf{x}_i^*$, подвергающихся кластеризации, а большими — положение фокуса нейронов сети Кохонена в системе координат, задаваемой первой и второй компонентой векторов из множества.

Продолжая рассматривать пример двумерных векторов $\mathbf{x}_i^*$, отметим, что чаще всего на практике для проекции нейронов и их весов на двумерную плоскость R^2 для дальнейшего визуального анализа и группирования исходных данных $\mathbf{x}_i \in \mathbf{X}$, строят унифицированную матрицу расстояний U, чья структура проиллюстрирована на рисунке 2.

Несмотря на то, что унифицированная матрица расстояний позволяет визуально оценить, на какое именно количество кластеров можно разделить множество исходных данных $\mathbf{X}$, в самом алгоритме кластеризации на основе самоорганизующихся карт Кохонена никак не учитывается то обстоятельство, что исходные измерительные данные $\mathbf{x}_i$ сопровождены погрешностями и не могут быть известны точно. Для выполнения метрологически обоснованной кластеризации с применением карт Кохо-

нена отсутствие учета неопределенности исходных обрабатываемых данных недопустимо. К тому же принятие решения о прекращении итерационной процедуры кластеризации согласно [11–13] принимается на основе количественной оценки неточности результатов кластеризации, унаследованной от погрешности исходных данных, которая в непосредственном виде не содержится ни в весах синоптических связей нейронов, ни в матрице *U*, ни в результатах ее визуального представления.

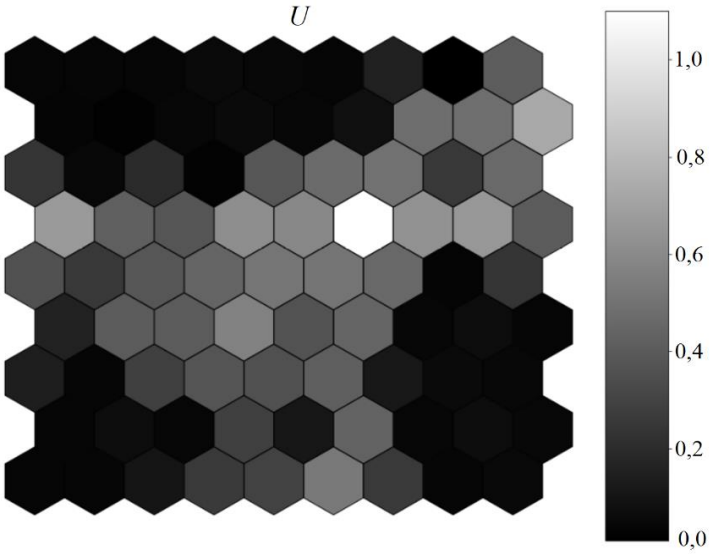

Рис. 2. Визуализация унифицированной матрицы расстояний U для исходных данных, представленных на рисунке 1

Для выполнения метрологически обоснованной кластеризации с применением карт Кохонена автором предложено формализовать элементы матрицы расстояний *U* (как конечный результат процесса проецирования исходных данных на структуру карты Кохонена) следующими числовыми характеристиками:

1) значениями нейронов-ячеек матрицы *U* (соответствуют среднему расстоянию между синаптическими весами ω_t соседних нейронов), определяемых как

$$\overline{d_t} = \frac{1}{n(t)} \cdot \sum_{j=1}^{n(t)} T(\omega_t, \omega_j),$$

где $T(\omega_t, \omega_j)$ — расстояние между векторами синаптических весов ω_t и ω_j; $j = 1, 2, \ldots, n(t)$ — индекс номера нейронов, являющихся соседями к заданному нейрону ω_t;

2) количеством активаций нейрона в процессе обучения сети;

3) координаты нейрона в матрице (значения абсциссы и ординаты для двумерного случая).

Данный набор характеристик нейронов позволяет обоснованно сформировать множество векторов признаков $\mathbf{y}_t \in \mathbf{Y} \subseteq R^4$, которое может быть использовано в процессе метрологически обоснованной кластеризации по алгоритму [13].

3. Результаты численных расчетов

Для апробации предложенного подхода к метрологически обоснованной кластеризации и соответствующей предварительной обработке данных методом самоорганизующихся карт Кохонена был написан пакет программ [18] на языке Python с привлечением библиотек Numpy, Sklearn, Matplotlib.

Были выполнены численные расчеты с применением созданных программных средств. Полученные результаты для синтетического набора данных, описанного в [11], сведены в таблице 1.

Таблица 1

Результаты численных расчетов по применению метрологически обоснованной кластеризации

Предел относительной погрешности исходных данных, %	Алгоритм предварительной обработки исходных данных	Алгоритм кластеризации	Максимальное количество обнаруженных кластеров
3	SOM	k-means	10
		MST	9
		Single-linkage	9
4	SOM	k-means	7
		MST	7
		Single-linkage	7
5	SOM	k-means	4
		MST	4
		Single-linkage	4
8	SOM	k-means	2
		MST	2
		Single-linkage	2

Из таблицы 1 видно, что при одинаковой величине пределов относительной погрешности исходных данных и одном и том же методе предварительной обработки данных (самоорганизующихся карт Кохонена) результаты дальнейшей кластеризации сходятся к одинаковому количеству кластеров вне зависимости от того, какой именно алгоритм кластеризации использован. Данное обстоятельство указывает на то, что использование предложенного подхода на основе карт Кохонена позволяет получать обоснованные выводы о внутренней структуре данных с учетом их точности и определять, на какое максимально возможное количество групп можно выделить в исходных данных в соответствии с их метрологическими характеристиками. Наблюдаемые при небольших

значениях пределов относительной погрешности исходных данных возможные отличия в определении максимального количества распознаваемых кластеров связаны с тем обстоятельством, что в таких условиях большее влияние начинают оказывать особенности алгоритмов, лежащих в основе используемых методов кластеризации, нежели чем сама неточность обрабатываемых данных. Также стоит отметить, что при определенных значениях предела погрешности исходных данных могут возникать пограничные состояния, когда выделяемое количество кластеров разными алгоритмами колеблется от реализации случайной погрешности к реализации (но не более чем на единицу).

Заключение

В данной работе предложен алгоритм осуществления метрологически обоснованной кластеризации неточных данных, возмущенных погрешностями, с применением самоорганизующихся карт Кохонена. Представленный подход отличается высокой контрастностью результатов предварительной обработки кластеризуемых данных, приводящей к устойчивым результатам при дальнейшем применении алгоритмов разбиения на группы в широком диапазоне допустимых пределов погрешности анализируемых данных. Полученные результаты указывают на перспективность использования представленного комбинированного алгоритма кластеризации на основе карт Кохонена.

Список литературы

1. Больщиков В.А., Семенов К.К. Кластеризация сигналов измерительной информации с учетом их метрологических характеристик // В сборнике: Измерения в современном мире – 2017. Сборник научных трудов 6-ой Всероссийской научно-практической конференции. – СПб.: СПбПУ, 2017. – С. 130–134.

2. Больщиков В.А., Семенов К.К. Обработка неточных данных при их кластеризации, согласованная с их точностью // В сборнике: Неделя науки СПбПУ. Материалы научной конференции с международным участием. Институт компьютерных наук и технологий. – СПб.: СПбПУ, 2018. – С. 184–187.

3. de Souza R.M., de Carvalho F.D.A. Clustering of interval data based on city–block distances // Pattern Recognition Letters. – 2004. – Vol. 25(3). – Pp. 353–365. – DOI: 10.1016/j.patrec.2003.10.016.

4. Asharaf S., Murty M.N., Shevade S.K. Rough set based incremental clustering of interval data // Pattern Recognition Letters. – 2006. – Vol. 27(6). – Pp. 515–519. – DOI: 10.1016/j.patrec.2005.09.018.

5. Peng W., Li T. Interval data clustering with applications // In: Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06). – IEEE, 2006. – Pp. 355–362. – DOI: 10.1109/ICTAI.2006.71.

6. Irpino A., Verde R. Dynamic clustering of interval data using a Wasserstein-based distance // Pattern Recognition Letters. – 2008. – Vol. 29(11). – Pp. 1648–1658. – DOI: 10.1016/j.patrec.2008.04.008.

7. De Carvalho F.D.A., Tenório C.P. Fuzzy K-means clustering algorithms for interval-valued data based on adaptive quadratic distances // Fuzzy Sets and Systems. – 2010. – Vol. 161(23). – Pp. 2978–2999. – DOI: 10.1016/j.fss.2010.08.003.

8. Leski J.M. Fuzzy c-ordered medoids clustering for interval-valued data // Pattern Recognition. – 2016. – Vol. 58. – Pp. 49–67. – DOI: 10.1016/j.patcog.2016.04.005.

9. Ji Z., Xia Y., Sun Q., Cao G. Interval-valued possibilistic fuzzy C-means clustering algorithm // Fuzzy Sets and Systems. – 2014. – Vol. 253. – Pp. 138–156. – DOI: 10.1016/j.fss.2013.12.011.

10. Le Hegarat-Mascle S., Bloch I., Vidal-Madjar D. Application of Dempster-Shafer evidence theory to unsupervised classification in multisource remote sensing // IEEE transactions on geoscience and remote sensing. – 1997. – Vol. 35(4). – Pp. 1018–1031. – DOI: 10.1109/36.602544.

11. Semenov K.K., Bolschikov V.A. The metrologically reasonable clustering of measurement results // Joint IMEKO TC1-TC7-TC13-TC18 Symposium, 2–5 July 2019. Journal of Physics: Conference Series. – 2019. – Vol. 1379. – Paper 012054. – DOI: 10.1088/1742-6596/1379/1/012054.

12. Bolschikov V.A., Semenov K.K. Clustering of inaccurate data using information on its precision // International Scientific Conference “Digital Transformation on Manufacturing, Infrastructure and Service”, 21–22 November 2018. IOP Conference Series: Materials Science and Engineering. – 2019. – Vol. 497. – Paper 012023. – DOI: 10.1088/1757-899X/497/1/012023.

13. Больщиков В.А., Семенов К.К. Обобщенный алгоритм определения максимального количества распознаваемых кластеров в неточных данных // Системный анализ в проектировании и управлении. сборник научных трудов XXV Международной научной и учебно-практической конференции. В 3 ч. Ч. 3. Санкт-Петербург, 2021. – СПб.: ПОЛИТЕХ-ПРЕСС, 2021. – С. 459–471.

14. Больщиков В.А., Семенов К.К. Программа для определения максимального количества кластеров, в обрабатываемых неточных данных, при использовании для кластеризации алгоритма Прима построения минимального остовного дерева / Свидетельство о регистрации программы для ЭВМ RU 2023680225, 27.09.2023. Заявка от 25.09.2023.

15. Больщиков В.А., Семенов К.К. Программа для определения максимального количества кластеров, в принципе различимых в обрабатываемых неточных данных, при использовании алгоритма кластеризации k-means / Свидетельство о регистрации программы для ЭВМ RU 2023680534, 02.10.2023. Заявка от 25.09.2023.

16. Больщиков В.А., Семенов К.К. Программа для определения максимального количества кластеров, в принципе различимых в обрабатываемых неточных данных, при использовании алгоритма кластеризации DBSCAN / Свидетельство о регистрации программы для ЭВМ RU 2023680718, 04.10.2023. Заявка от 25.09.2023.

17. Больщиков В.А., Семенов К.К. Программа для определения максимального количества кластеров, в принципе различимых в обрабатываемых неточных данных, при использовании для кластеризации иерархического алгоритма одиночной связи / Свидетельство о регистрации программы для ЭВМ RU 2023681054, 10.10.2023. Заявка от 25.09.2023.

18. Больщиков В.А., Семенов К.К. Программа для определения максимального количества кластеров, в принципе различимых в обрабатываемых неточных данных, при использовании для кластеризации нейронных сетей по типу самоорганизующихся карт Кохонена / Свидетельство о регистрации программы для ЭВМ RU 2023680348, 28.09.2023. Заявка от 25.09.2023.