Заключение. Таким образом, предлагаемые подходы, методы и алгоритмы и автоматизированная система онлайн мониторинга, построенная на их основе, обладают следующими преимуществами, такими как универсальность применения, высокая скорость обработки данных (за счет естественного параллелизма обработки информации в нейронных сетях), высокое качество распознавание данных [1], - возможность повышения степени интеллектуализации принимаемого решения за счет наличия процедур коррекции результатов и переобучения нейросетевого "ядра" системы.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Кононов С.В., Коровин С.Я.* Развитие информационных систем основного производства ОАО "Сургутнефтегаз" для задач поддержки принятия решений // Нефтяное хозяйство. – 2006. – № 10.
2. *Галуев Г.А., Коровин Я.С., Коровин С.Я., Матвеев С.Н.* Комплексный подход к поддержке принятия решений для управления производственными процессами в нефтяной промышленности на основе нейрокомпьютерных и мультиагентных технологий // Нейрокомпьютеры: разработка, применение". – 2006. – № 3. – С. 42-49.
3. *Коровин Я.С.* Система поддержки принятия решений по контролю состояния УЭЦН на основе нейронной сети: архитектура, реализация, перспективы // Нефтяное хозяйство. – 2007. – № 1. – С. 80-85.

Статью рекомендовал к опубликованию д.т.н. В.А. Гандурин.

Коровин Яков Сергеевич – Научно-исследовательский институт многопроцессорных вычислительных систем имени академика А.В. Каляева федерального государственного автономного образовательного учреждения высшего профессионального образования «Южный федеральный университет»; e-mail: korovin_yakov@mail.ru; 347928, г. Таганрог, ул. Чехова, 2, ГСП-284; тел.: 88634365883; к.т.н.; зав. лабораторией нейросетевых систем.

Korovin Yakov Sergeevich – SFedU Acad. Kalyaev Scientific Research Institute of Multiprocessor Computer Systems; e-mail: korovin_yakov@mail.ru; GSP-284, 2, Chekhov street, Taganrog, 347928, Russia; phone: +78634365883; cand. of eng. sc.; head of neuronetwork systems laboratory.

УДК 004.4

А.В. Егоров, Н.И. Куприянова

ПРОБЛЕМА ВЫБОРА КРИТЕРИЕВ В МЕТОДАХ НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ

Целью исследования является описание проблемы, вызванной многообразием существующих критериев в нечетком кластерном анализе. Описываются подгруппы критериев, служащие для определения кластеров, а также для оценки процесса кластеризации и его результатов. Основные критерии иллюстрируются на базовых видах данных: матрицах связи и матрицах объект–признак. Далее рассматривается механизм расчета данных критериев. Особое место занимают оптимизационные критерии, в состав входят эвристические, аппроксимационные и критерии статистического оценивания. В завершении данные критерии соотносятся с критерием k-средних, как одним из базовых для процессов нечеткого кластерного анализа. В завершении описываются функционалы качества, являющиеся типовыми примерами использования понятия «критерий» в процессах обработки данных.

Оптимизационные критерии; критерии выбора методов кластеризации; функционалы качества кластеризации.

A.V. Egorov, N.I. Kupriyanova

PROBLEM OF THE CHOICE OF CRITERIA IN METHODS OF THE FUZZY CLUSTERING

Research objective is the description of the problem caused by variety of existing criteria in the indistinct cluster analysis. The subgroups of criteria serving for definition of clusters, and also for an assessment of process of a clustering and its results are described. The main criteria are illustrated on basic types of data: matrixes of communication and matrixes object – a sign. Further the mechanism of calculation of these criteria is considered. In end these criteria correspond to criterion of k-averages, as one of basic criteria of the indistinct cluster analysis.

Optimizing criteria; criteria of a choice of methods of a clustering; functionalities of quality of a clustering.

На сегодняшний день кластерный анализ является одним из наиболее перспективных направлений в обработке данных. Но в нем выделяют особую область, называемую выбор критериев. В разрезе кластерного анализа понятие «критерий» рассматривается с нескольких точек зрения, выделяя оптимизационные критерии, служащие для определения кластеров, многокритериальную задачу кластеризации нечетких объектов, описывающие работу с качественными данными, а также использование критериев в оценке качества решений задач нечеткой кластеризации.

Оптимизационные критерии. Оптимизационные критерии кластер – анализа могут быть разделены на три типа:

- *эвристические*

В таких критериях формализуется интуитивная идея, что объекты внутри кластеров должны быть близки друг к другу, а в разных кластерах – далеки друг от друга [1];

- *аппроксимационные*

Такие критерии основаны на представлении искомой кластерной структуры математическими объектами того же типа, что и данные, обычно в виде матриц, так что в качестве критерия выступает степень близости между матрицей исходных данных и матрицей формируемой кластер–структуры.

- *статистического оценивания*

Обычно это критерий максимального правдоподобия какой-либо статистической модели, такой как смесь распределений.

Рассмотрим примеры эвристических критериев для двух основных видов данных: матрицы связи и матрицы объект–признак.

1. Матрица связей. При заданной матрице связи $A = (a_{ij})$ на множестве I будем делить I на две части S_1 и S_2, так что связи между заданными двумя частями были минимальны, а внутри множеств S_1 и S_2 – максимальной [2]. Естественная формализация такого критерия может быть выражена в терминах суммарной связи. Обозначая суммарную связь между S_f и S_k через $a(S_f, S_k)$, получаем формулу суммарной связи (1) и ее основное свойство (2), при условии, что матрица A – симметрична:

$$a(S_f, S_k) = \sum_{i \in S_f} \sum_{j \in S_k} a_{ij}. \quad (1)$$

$$a(S_1, S_2) = a(S_1, S_2). \quad (2)$$

Поскольку сумма $a(S_1, S_2) + a(S_2, S_1) + a(S_1, S_1) + a(S_2, S_2)$ равна $a(I)$ всех связей и постоянна, то минимальный разрез (естественный критерий минимизации) одновременно максимизирует сумму всех связей внутри кластера (3):

$$aw(S_1, S_2) = a(S_1, S_1) + a(S_1, S_2), \tag{3}$$

что является удачным свойством критерия. Тем не менее этот критерий не работает для наиболее часто встречаемых, неотрицательных матриц связи, так как он приводит к очевидному – и тривиальному – оптимальному решению: собрать все объекты в один кластер, выделив в другой кластер только один объект – тот, у которого самые слабые связи с остальными [1]. Такое несбалансированное решение, конечно, не может рассматриваться по-настоящему кластерной структурой, что приводит к необходимости модификации критерия. Такая модификация была предложена Дорофеюком и Браверманом, которые предложили делить сумму внутренних связей кластера на его численность, так что максимизируемый критерий превращается в выражение (4):

$$ab(S_1, S_2) = a(S_1, S_1) / |S_1| + a(S_2, S_2) / |S_2|. \tag{4}$$

Оптимизационный критерий (4) приводит к более сбалансированным структурам. Также следует отметить, что данный критерий зачастую используется в апроксимационных задачах и задачах нечеткой кластеризации, так как напрямую связан с критерием *k*-средних.

2. Матрица объект–признак. Критерием для таблицы объект–признак является сумма расстояний от объектов до центров соответствующих кластеров.

Рассмотрим предобработанную матрицу объект – признак $Y = (y_{iv})$, где столбцы $v = 1,.....V$ соответствует признакам, а строки $i \in I$ – объектам. Кластерная структура задается разбиением S множества объектов на K непересекающихся кластеров $S = \{S_1, S_2, ..., S_n\}$, представляемые таким образом через кластеры S_k и центроиды $c_k = (c_{k1}, c_{k2}, ..., c_{kV})$, $k \in (1, K)$. Тогда минимизируемым критерием становится сумма расстояний $d(y_i, c_k)$ от объектов y_i до соответствующих центроидов c_k

$$W(S, c) = \sum_{k=1}^{K} \sum_{i \in S_k} d(y_i, c_k). \tag{5}$$

Данный критерий лежит в основе метода *k*-средних и его производных методик. В теории нечеткой кластеризации метод *k*-средних является одним из основных методов, хотя он поддерживает ряд допущений, а именно: в результате использования данного метода увеличивается равномерность распределения объектов по кластерам, что не всегда соответствует реальному расположению рассматриваемых объектов.

Помимо критериев, характеризующих сами методики кластеризации, в кластерном анализе существуют критерии выбора метода кластеризации. К ним относятся:

- объем информации;
- размерность информации;
- тип атрибутов;
- априорное представление о количестве кластеров, перекрываемости и форме получаемых кластеров;
- качество кластеризации.

В процессе оценки качества кластеризации также используются критерии, позволяющие оценить результативность разбиения данных на кластеры:

1. Визуальные средства оценки степени разнесенности и компактности выделенных групп объектов – отображение объектов в одно-, дву-, трехмерном пространстве с указанием их групповой принадлежности в виде гистограмм или диаграмм рассеивания [2]. Также это могут быть графики средних.
2. Оценка качества кластеризации с помощью функционалов качества – проверяются неформальные требования к образованному разбиению: внутри группы объекты должны быть тесно связаны между собой; объекты различных групп должны быть далеки друг от друга; распределение объектов по кластерам должно быть равномерным (сумма квадратов расстояний до центров классов, сумма внутриклассовых расстояний между объектами, метод проверки на основе кофенетической корреляции и др).

В результате следует отметить, что критерии используются в кластерном анализе на всех этапах вне зависимости от методики кластерного анализа. Они участвуют как в построении кластерных структур и распределении объектов, так и в формировании результатов процесса кластеризации и его оценок. Объективное расширение количества используемых критериев и их взаимозаменяемость является шагом к получению новых адаптивных алгоритмов и метод кластеризации, в том числе и нечетких.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Егоров А.В., Куприянова Н.И.* Особенности методов кластеризации данных // Известия ЮФУ. Технические науки. – 2011. – № 11 (124). – С. 174-178.
2. *Миркин Б.Г.* Методы кластер – анализа для поддержки принятия решений: обзор. – М.: Высшая школа экономики, 2011.

Статью рекомендовал к опубликованию д.т.н., профессор В.П. Карелин.

Егоров Александр Вадимович – Федеральное государственное автономное образовательное учреждение высшего профессионального образования «Южный федеральный университет»; e-mail: egor@tti.sfedu.ru; 347928, г. Таганрог, пер. Некрасовский, 44; тел.: 88634383652; кафедра прикладной информатики; к.т.н.; доцент.

Куприянова Наталия Игоревна – e-mail: ultra-n@list.ru; кафедра прикладной информатики; аспирантка.

Yegorov Alexander Vadimovich – Federal State-Owned Autonomy Educational Establishment of Higher Vocational Education “Southern Federal University”; e-mail: egor@tsure.ru; 44, Nekrasovskiy, Taganrog, 347928, Russia; phone: +78634383652; the department of applied information science; cand. of eng. sc.; associate professor.

Kupriyanova Natalia Igorevna – e-mail: ultra-n@list.ru; the department of applied information science; postgraduate student.