УДК 681.5

Введение в область нечеткой кластеризации качественных и количественных данных

А.В. Егоров[1], Н.И. Куприянова[1]

Abstract – **In many applications of clustering objects described by both quantitative and qualitative characteristics. A variety of algorithms have been proposed for "uncontrolled classification", which is the best thing for fuzzy partitions, and described the cluster prototypes. However, most of these methods are designed for data sets that can be represented in a limited type (only classes or only the metric). We propose a new approach to fuzzy clustering based on the probability measure distances.**

***Keywords* – fuzzy clustering, clustering, clustering methods, the mixed model of clustering.**

Кластеризация смешанных типов, наборов данных часто встречается в анализе данных. Это может произойти, например, в области моделирования - при добыче описательных данных для группировки пользователей с учетом их конкретных интересов и поведения.

Мало результатов существует в области выявления и сравнения алгоритмов, формирующих «характеристические описания» (прототипы) нечетких кластеров на основе данных, описываемых смесью количественных и качественных данных. Исмаил и Эль-Сонбати были первыми, кто применил понятие «нечеткость» для разделения символьных объектов. Их символьный нечеткий c- means подход может быть применим для обработки широкого спектра различных типов атрибутов, включая порядковые переменные и интервалы, а также для формирования прототипов кластеров. Например, этими прототипами могут являться взвешенные частоты условий в кластере. Они применяют дидайевскую меру несходства для символьных данных. Хотя эта мера позволяет рассчитать несходство также для непрерывных атрибутов, расчет центров кластеров не подходит для данных количественного типа. Символьный нечеткий c – means алгоритм неэффективен при анализе смеси разных наборов данных. Ян, Хуан и Чень преодолели это ограничение, изменив и расширив его мерой несходства, а также построив прототипы кластеров. Этот подход кластеризует символьные и непрерывные переменные. Значения непрерывных атрибутов могут быть воплощены в соответствующей параметризацией с использованием трапециевидных нечетких чисел. Таким образом, все атрибуты объектов смешанного типа используются для нахождения кластеров.

Другие связанные методы имеют больше ограничений на множестве ограниченных типов атрибутов. Нечеткий k-modes алгоритм ограничивается в разрезе символьных переменных и находит нечеткие кластеры с использованием простого совпадения меры несходства для различных категорий. Имеется расширение алгоритма под названием К-прототипный алгоритм для работы как с метрикой и особенными категориями. Этот алгоритм позже получил название сложного k-modes алгоритма (работа с категорийными атрибутами). В соответствии с ним несходство рассчитывается отдельно для каждого типа переменных.

Для агрегированной меры несходства вес параметра определяет степень влияния номинальных (первоначальных) характеристик объекта на результирующее значение. [1] Другой подход, ограничивающий себя в разрезе категорийных переменных, это нечетко-статистический алгоритм. «Чистые» прототипы предполагаются только для теоретического обоснования, но на практике их построить нельзя. Кроме того, все реляционные алгоритмы кластеризации подходят для выполнения задачи классификации для смешанного типа объектов. Они требуют матрицы расстояний в качестве входных данных, не ссылаясь на фактически вводимые данные, и ,как результат, они рассчитывают индексы с учетом наиболее типичных объектов в кластерах. Таким образом, прототипы кластеров, которые были бы полезны для их описания и дальнейших расчетов, не построены. В нашем подходе мы хотим преодолеть следующие ограничения рассмотренных методов:

Проблема 1: Формирование «представительного» прототипа кластера

«Мотивация» для прототипов, которые способны представлять кластеров, лежит в требовании, что они должны быть информативными для пользователей, так как они являются прототипами результата анализа данных.

Проблема № 2: Расчет несходство. Для расчета меры несходства между центрами кластеров и объектов доступны различные меры несходства. [2], [3] Тем не менее, особое внимание стоит уделить тому моменту, что расстояние, в разрезе рассмотрения качественных и количественных характеристик, должно рассчитываться отдельно на первом этапе, а затем объединяться в общую несхожесть. Тогда должно быть обеспечено, что качественные и количественные компоненты соизмеримы, и отсутствует преобладание одного типа атрибутов.

Это может быть достигнуто за счет стандартизации числовых данных и/или введения весов, как дополнительную меру расстояний.

Однако, определение веса параметра всегда проблематично, так как цель выбора значения часто

[1] Таганрогский Технологический институт ЮФУ пер.Некрасовский 44, Таганрог, 347928, РОССИЯ, E-mail: egor@tsure.ru

бывает трудна. [1] Кроме того, результаты классификации легко могут быть искажены или можно манипулировать ими, когда веса выбираются пользователем. [2]

Теперь рассмотрим непосредственно алгоритм кластеризации качественных и количественных данных (контролируемая классификация).

Для нее были предложены различные алгоритмы, если необходимо нечеткое распределение и прототипы кластеров. Однако, большинство этих методов разработаны для множества данных с переменными, которые измеряются таким же типом шкалы (безусловным или метрическим). Мы предлагаем новый подход к нечеткой кластеризации, который основан на вероятностном способе измерения.

Большинство алгоритмов нечеткой кластеризации основаны на целевой функции. Они определяют оптимальное разбиение данного набора данных $X = \left\{ \vec{x}_j \mid j = 1,..., n \right\}$ на кластеры путем сокращения целевой функции.

$$J(X, U, C) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} d_{ij}, \qquad (1)$$

определяемое условиями

$$\sum_{j=1}^{n} u_{ij} > 0, \text{ для всех } i \in \{1,..., c\} \qquad (2)$$

$$\sum_{i=1}^{n} u_{ij} = 1, \text{ для всех } j \in \{1,..., n\} \qquad (3)$$

где $i \in \{1,..., c\}$ - характеристика принадлежности элемента данных $\vec{x}_j$ к кластеру i, а d_{ij} это расстояние между данной $\vec{x}_j$ и кластером i. C×n – размерности матрица $U = (u_{ij})$ называется матрицей нечеткого разделения , а C - описывает множество кластеров, определяя положение параметров (т.е. центр кластера) и, возможно, размер и форму параметра для каждого кластера. Параметр m, m > 1 называется элементом первичной обработки параметров (фазификатором) или образцом весового показателя. Он определяет нечеткость классификации: чем выше значение для m, тем более нечеткими становятся границы между кластерами. Чем ниже параметры, тем границы становятся четче. Обычно выбирается условие m=2.

Выражение (2) обеспечивает условие, когда ни один кластер не остается пустым, а выражение (3) обеспечивает условие, когда характеристика принадлежности данной к кластерам суммируется до 1 и, тем самым, каждый элемент данных имеет влияние. Из-за второго условия этот подход обычно называется вероятностной нечеткой кластеризацией, поскольку с ее помощью характеристика принадлежности кластеру совпадает с возможностью быть членом соответствующего кластера. Особенность алгоритма вероятностной кластеризации, которая «определяет» вес элемента для отнесения его в разные кластеры, обусловлена также выражением 2.

К сожалению, целевая функция J не может быть напрямую сокращена. Поэтому, алгоритм используется повторно, оптимизируя степень принадлежности и параметры кластеров. Таким образом, во-первых, степень принадлежности оптимизируется для постоянных параметров кластера, а затем параметры кластеров оптимизируются для постоянной степени принадлежности. Главное преимущество этой схемы заключается в том, что в каждом из этих случаев оптимальные условия можно вычислить сразу. Повторяя оба шага можно получить единые оптимальные условия (но нельзя гарантировать, что будут определены всеобщие оптимальные условия – алгоритм может остановиться на локальном минимуме целевой функции J).

Обновленные формулы получаются путем установления производной целевой функции J по отношению к параметрам для того, чтобы оптимизировать равенство до ноля (необходимое условие для минимума).

Будучи независимым от измерения выбранного расстояния мы получаем следующую формулу для степени принадлежности [10]

$$u_{ij} = \frac{d_{ij}^{-\frac{1}{m-1}}}{\sum_{t=1}^{c} d_{tj}^{-\frac{1}{m-1}}} \qquad (4)$$

Обновленная формула для параметров кластера, конечно, зависит от того, какие параметры используются для описания кластера (расположение, форма, размер) и от измерения выбранного расстояния. Поэтому, нельзя предоставить общую обновленную формулу.

Подход, который возможно создать в дальнейших исследованиях, основан на смешанной модели для процесса, который формирует данные. из этого предположения, реализованного выше, мы получаем модель с измерением расстояния, которое в обратной зависимости пропорционально вероятности, созданной кластером, что похоже на идею расчета вероятности нечеткого максимума [3].

Список ссылок

[1] Z.Huang, “Clustering large data sets with mixed numeric and categories values”, in Proceedings of the First Pacific –Asia Conference on Knowledge Discovery and Data Mining, Ser. Lecture Notes in Artificial Intelligence,1997, pp.21-34

[2] F. Hooper, F.Klawonn, R.Kruse, T.Runkler, Fuzzy Clustering. Chichester, United Kingdom:Wiley,1999

[3] Chidananda Gowda and E.Diday, “Symbolic clustering using a new dissimilarity measure”, Pattern Recognition, ug. 1995, pp. 77-80.