УДК 519.7:004.8

Е.В. БОДЯНСКИЙ

Харьковский национальный университет радиоэлектроники, Украина

МЕТОДЫ ВЫЧИСЛИТЕЛЬНОГО ИНТЕЛЛЕКТА ДЛЯ АНАЛИЗА ДАННЫХ

Рассмотрены методы нечеткой (фаззи) кластеризации данных в условиях перекрывающихся классов на основе вероятностного и возможностного подходов. Предложены адаптивные алгоритмы, реализующие эти подходы и позволяющие обрабатывать данные по мере их поступления в реальном времени. Введены робастные процедуры кластеризации, основанные на целевых функциях, устойчивых к аномальным выбросам.

фаззи-кластеризация, вероятностный и возможностный подходы, адаптивный алгоритм, робастность, целевая функция

Введение

Интеллектуальный анализ данных (Data Mining) является процессом извлечения предварительно неизвестных, нетривиальных, практически полезных и интерпретируемых знаний из "сырых" неструктурированных данных в больших массивах или базах данных. В общем случае задача, которую решают с помощью интеллектуального анализа, состоит в обнаружении закономерностей в данных различной природы. Более подробно он сочетает в себе построение математических моделей, прогнозирование, классификацию, кластеризацию, генерацию правил, обобщение, понижение размерности, визуализацию и т.п.

Так как данные являются часто неточными, их распределения и скрытые зависимости не известны и достаточно сложны, возникает необходимость в методах, которые смогут справиться с дефицитом информации, сложностью, неточностью, загрязненностью данных аномальными выбросами т.п. Среди таких методов технологии, основанные на мягких вычислениях [1] получают все более широкое распространение.

Фаззи-системы, являющиеся одной из основ таких технологий, базируются на теории нечетких множеств, предложенных Л. Заде [2]. В отличие от классической теории множеств в теории фаззи-множеств любой объект может принадлежать нескольким множествам одновременно с определенными значениями уровней принадлежности, которые выражаются вещественными числами в интервале $[0,1]$.

Важным применением фаззи-систем в анализе данных является использование нечетких правил для интерпретации лингвистических данных [3]. Фаззи-системы могут также применяться для классификации и моделирования нелинейных зависимостей, когда акцент больше ставится на интерпретируемость, нежели на точность. Нечеткие правила могут определяться с помощью фаззи-кластеризации. В традиционных подходах в кластеризации предполагается, что каждое наблюдение принадлежит только одному классу. Примерами таких подходов являются, например, алгоритм *k*-средних [4] и правило ближайшего соседа [5]. Более естественным предположением является то, что каждое наблюдение может принадлежать сразу нескольким кластерам с определенными значениями функций принадлежности. Это предположение является основой фаззи-кластерного анализа [6, 7]. В настоящее время известно достаточно много алгоритмов фаззи-кластеризации, например, алгоритм Бездека [6], алгоритм Густафсона-Кесселя [8], горная кластеризация Ягера-Филева [9] и т.д. Все отмеченные методы предполагают, что массив

данных, подлежащих обработке, задан заранее, данные не содержат аномальных выбросов, а их характер не меняется с течением времени. Естественно, что эти ограничения существенно сужают возможности фаззи-кластеризации и требуют разработки новых подходов к решению проблемы.

1. Формулирование проблемы

Исходной информацией для решения задач кластеризации является множество из N данных в виде n-мерных векторов признаков $X = \{x_1, x_2, \ldots, x_N\}$, $x_k \in X$, $k = 1, 2, \ldots, N$. Результат представляет собой разделение исходных данных на m кластеров с некоторым значением уровня принадлежности $w_{k,j} \in [0, 1]$ k-го вектора признаков к j-му кластеру. При этом предполагается расчет $N \times m$ матрицы $W = \{w_{k,j}\}$, которая называется матрицей нечеткого разбиения.

Когда элементы матрицы W рассматриваются как вероятности гипотез о принадлежности вектора x_k к определенным кластерам, кластеризация называется вероятностной [6, 8, 10]. Наиболее существенным недостатком вероятностного подхода является требование, чтобы сумма функций принадлежности каждого вектора данных была равна единице. Преодолеть это ограничение позволяют так называемые возможностные методы фаззи- кластеризации, которые впервые были предложены в работе [11].

Если массив данных, требующий обработки, является очень большим, обработка их в пакетном режиме может быть очень медленной или вообще невозможной. Обработка таких наборов данных может быть ускорена при использовании рекуррентных алгоритмов фаззи-кластеризации [12, 13].

В настоящей статье изложен подход к синтезу рекуррентных адаптивных вероятностных и возможностных алгоритмов фаззи-кластеризации, предназначенных для анализа данных, «загрязненных» искажениями различного рода, в реальном времени по мере поступления.

2. Анализ данных с помощью алгоритмов адаптивной фаззи-кластеризации

Алгоритмы фаззи-кластеризации, основанные на целевых функциях [6], предназначены для решения задачи путем оптимизации некоторого наперед заданного критерия качества кластеризации и являются наиболее строгими с математической точки зрения.

Целевая функция, подлежащая минимизации, имеет вид

$$E(w_{k,j}, c_j) = \sum_{k=1}^{N} \sum_{j=1}^{m} w_{k,j}^{\beta} d^2(x_k, c_j) \qquad (1)$$

при ограничениях

$$\sum_{j=1}^{m} w_{k,j} = 1, \quad k = 1, \ldots, N, \qquad (2)$$

$$0 < \sum_{k=1}^{N} w_{k,j} < N, \quad j = 1, \ldots, m. \qquad (3)$$

Здесь c_j – прототип (центр) j-го кластера, β – неотрицательный параметр, именуемый «фаззификатором» (обычно $\beta = 2$), $d^2(x_k, c_j)$ – расстояние между x_k и c_j в принятой метрике.

Заметим, что поскольку элементы матрицы нечеткого разбиения W могут рассматриваться как вероятности гипотез принадлежности векторов данных определенным кластерам, процедуры, порождаемые (1) при ограничениях (2), (3), называются вероятностными алгоритмами кластеризации. Это наиболее исследованный класс процедур фаззи-кластеризации.

2.1. Пакетные алгоритмы фаззи-кластеризации, основанные на целевых функциях

Вводя функцию Лагранжа

$$L(w_{k,j},c_j,\lambda_k)=$$

$$=\sum_{k=1}^{N}\sum_{j=1}^{m}w_{k,j}^{\beta}d^2(x_k,c_j)+\sum_{k=1}^{N}\lambda_k(\sum_{j=1}^{m}w_{k,j}-1)=$$

$$=\sum_{k=1}^{N}(\sum_{j=1}^{m}w_{k,j}^{\beta}d^2(x_k,c_j)+\lambda_k(\sum_{j=1}^{m}w_{k,j}-1)) \quad (4)$$

(здесь λ_k - неопределенный множитель Лагранжа) и решая систему уравнений Куна-Таккера

$$\begin{cases}\partial L(w_{k,j},c_j,\lambda_k)/\partial w_{k,j}=0,\\ \nabla_{c_j}L(w_{k,j},c_j,\lambda_k)=0,\\ \partial L(w_{k,j},c_j,\lambda_k)/\partial\lambda_k=0,\end{cases} \quad (5)$$

несложно получить искомое решение в виде

$$w_{k,j}=\frac{(d^2(x_k,c_j))^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m}(d^2(x_k,c_l))^{\frac{1}{1-\beta}}}, \quad (6)$$

$$c_j=\frac{\sum_{k=1}^{N}w_{k,j}^{\beta}x_k}{\sum_{k=1}^{N}w_{k,j}^{\beta}}, \quad (7)$$

$$\lambda_k=-\left(\sum_{l=1}^{m}(\beta\, d^2(x_k,c_l))^{\frac{1}{1-\beta}}\right)^{1-\beta}. \quad (8)$$

Уравнения (6)-(8) порождают широкий класс процедур кластеризации. Выбирая $\beta=2$ и принимая евклидову метрику $d^2(x_k,c_j)=\|x_k-c_j\|^2$, получаем простой и эффективный алгоритм нечеткой кластеризации Бездека (fuzzy c-means) [6]:

$$w_{k,j}=\frac{\|x_k-c_j\|^{-2}}{\sum_{l=1}^{m}\|x_k-c_l\|^{-2}}, \quad (9)$$

$$c_j=\frac{\sum_{k=1}^{N}w_{k,j}^{2}x_k}{\sum_{k=1}^{N}w_{k,j}^{2}}, \quad (10)$$

$$\lambda_k=-\sum_{l=1}^{m}\left(\frac{\|x_k-c_l\|^{-2}}{2}\right)^{-1}. \quad (11)$$

К вероятностным алгоритмам кластеризации относятся также алгоритмы Густафсона-Кесселя [8], Гата-Гевы [10] и ряд других. Основные недостатки вероятностного подхода связаны с ограничениями (2). В простейшем случае двух кластеров ($m=2$) несложно видеть, что наблюдение x_k, равноправно принадлежащее обоим кластерам, и наблюдение x_p, не принадлежащее ни одному из них, могут иметь одинаковые уровни принадлежности $w_{k,1}=w_{k,2}=w_{p,1}=w_{p,2}=0.5$. Естественно, что данное обстоятельство, ухудшающее точность классификации, привело к появлению возможностных подходов к нечеткой классификации [11, 14].

В возможностных алгоритмах кластеризации целевая функция имеет вид

$$E(w_{k,j},c_j)=\sum_{k=1}^{N}\sum_{j=1}^{m}w_{k,j}^{\beta}d^2(x_k,c_j)+$$
$$+\sum_{j=1}^{m}\mu_j\sum_{k=1}^{N}(1-w_{k,j})^{\beta}, \quad (12)$$

где скалярный параметр $\mu_j>0$ определяет расстояние, на котором уровень принадлежности принимает значение 0.5, т.е. если $d^2(x_k,c_j)=\mu_j$, то $w_{k,j}=0.5$.

Минимизация (12) по $w_{k,j}$, c_j и μ_j дает очевидное решение

$$w_{k,j}=\left(1+\left(\frac{d^2(x_k,c_j)}{\mu_j}\right)^{\frac{1}{\beta-1}}\right)^{-1}, \quad (13)$$

$$c_j=\frac{\sum_{k=1}^{N}w_{k,j}^{\beta}x_k}{\sum_{k=1}^{N}w_{k,j}^{\beta}}, \quad (14)$$

$$\mu_j=\frac{\sum_{k=1}^{N}w_{k,j}^{\beta}d^2(x_k,c_j)}{\sum_{k=1}^{N}w_{k,j}^{\beta}}. \quad (15)$$

Видно, что возможностные и вероятностные алгоритмы очень похожи и переходят друг в друга заменой выражения (13) на формулу (6), и наоборот. Общим недостатком рассмотренных алгоритмов является их вычислительная сложность и невозможность работы в реальном времени.

Работа алгоритма (6)-(8) начинается с задания

начальной (обычно случайной) матрицы разбиения W^0. На основе ее значений рассчитывается начальный набор прототипов c_j^0, которые затем используются для вычисления новой матрицы W^1. Затем в пакетном режиме пересчитываются $c_j^1, W^2, \ldots, W^t, c_j^t, W^{t+1}$ и т.д., пока разность $\left\| W^{t+1} - W^t \right\|$ не станет меньше некоторого наперед заданного порога ε. Таким образом, вся имеющаяся выборка данных обрабатывается многократно.

Решение, полученное с помощью вероятностного алгоритма, рекомендуется использовать в качестве начальных условий для возможностного алгоритма (13)-(15) [14, 15]. Параметры расстояния μ_j инициализируются в соответствии с (15) по результатам работы вероятностного алгоритма.

2.2. Рекуррентные алгоритмы фаззи-кластеризации

Анализ уравнения (6) показывает, что для расчета уровней принадлежности $w_{k,j}$ вместо лагранжиана (4) можно использовать его локальную модификацию

$$L_k(w_{k,j}, c_j, \lambda_k) = \sum_{j=1}^{m} w_{k,j}^{\beta} d^2(x_k, c_j) + \\ + \lambda_k \left(\sum_{j=1}^{m} w_{k,j} - 1 \right). \tag{16}$$

Оптимизация выражения (16) с помощью процедуры Эрроу-Гурвица-Удзавы приводит к алгоритму

$$w_{k,j} = \frac{(d^2(x_k, c_{k,j}))^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m} (d^2(x_k, c_{k,l}))^{\frac{1}{1-\beta}}}, \tag{17}$$

$$c_{k+1,j} = c_{k,j} - \eta_k \nabla_{c_j} L_k(w_{k,j}, c_{k,j}, \lambda_k) = \\ = c_{k,j} - \eta_k w_{k,j}^{\beta} d(x_{k+1}, c_{k,j}) \nabla_{c_j} d(x_{k+1}, c_{k,j}), \tag{18}$$

где η_k – параметр скорости обучения;

$c_{k,j}$ – прототипы *j*-го кластера, вычисленные на выборке из *k* наблюдений.

Процедура (17), (18) близка к алгоритму обучения Чанга-Ли [16] и для $\beta = 2$ совпадает с градиентной процедурой кластеризации Парка-Дэггера [17]:

$$w_{k,j} = \frac{\left\| x_k - c_{k,j} \right\|^{-2}}{\sum_{l=1}^{m} \left\| x_k - c_{k,l} \right\|^{-2}}, \tag{19}$$

$$c_{k+1,j} = c_{k,j} + \eta_k w_{k,j}^2 (x_{k+1} - c_{k,j}). \tag{20}$$

В рамках возможностного подхода локальный критерий принимает форму

$$E_k(w_{k,j}, c_j) = \sum_{j=1}^{m} w_{k,j}^{\beta} d^2(x_k, c_j) + \\ + \sum_{j=1}^{m} \mu_j (1 - w_{k,j})^{\beta}, \tag{21}$$

а результат его оптимизации имеет вид

$$w_{k,j} = \left(1 + \left(\frac{d^2(x_k, c_{k,j})}{\mu_j} \right)^{\frac{1}{\beta - 1}} \right)^{-1}, \tag{22}$$

$$c_{k+1,j} = c_{k,j} - \\ - \eta_k w_{k,j}^{\beta} d(x_{k+1}, c_{k,j}) \nabla_{c_j} d(x_{k+1}, c_{k,j}), \tag{23}$$

где параметр расстояния μ_j инициализируется согласно (15).

В этом случае N в уравнении (15) будет объемом множества данных, используемых для инициализации.

В квадратичном случае алгоритм (22), (23) преобразуется в достаточно простую конструкцию

$$w_{k,j} = \frac{\mu_j}{\mu_j + \left\| x_k - c_{k,j} \right\|^2}, \tag{24}$$

$$c_{k+1,j} = c_{k,j} + \eta_k w_{k,j}^2 (x_{k+1} - c_{k,j}), \tag{25}$$

где параметр расстояния μ_j инициализируется по результатам вероятностной кластеризации (например, с помощью алгоритма Бездека (9), (10)) согласно уравнению

$$\mu_j = \frac{\sum_{k=1}^{N} w_{k,j}^2 \left\| x_k - c_j \right\|^2}{\sum_{k=1}^{N} w_{k,j}^2}. \tag{26}$$

Предложенный адаптивный алгоритм может

использоваться как в пакетном режиме для итеративной обработки заданной выборки, так и в режиме реального времени, где количество наблюдений k определяется текущим дискретным временем $k = 1,2,\ldots,N,N+1,\ldots$. В этом случае алгоритм последовательно обрабатывает наблюдения, поступающие на вход, настраивая уровни принадлежности и прототипы кластеров под новые данные.

Тестирование предложенных рекуррентных алгоритмов [12, 18] на наборах данных из UCI-репозитория [19] показало их высокую эффективность по сравнению с известными процедурами.

А в частности, предложенный алгоритм (24), (25) исследовался в задачах классификации данных и сравнивался с результатами алгоритма Бездека, пакетным алгоритмом возможностной кластеризации и рекуррентным алгоритмом кластеризации Парка-Дэггера. Для тестирования использовались три широко известных набора данных: данные «Вина», данные «Ирисы» и данные «Щитовидная железа». Данные «Ирисы» содержат описания 150 экземпляров цветов ириса, равномерно распределенных на три вида. Цветы описываются четырьмя атрибутами. Данные «Вина» содержат 178 результатов химического анализа вин, полученных из винограда трех разных сортов, выращенных в одном регионе Италии. Анализ определял количество 13 компонент, присутствующих в каждом из трех типов вин. Данные «Щитовидная железа» содержат 215 результатов (разделенных на три класса) медицинских анализов по пяти параметрам. Задача классификации состоит в отнесении каждой представленной комбинации признаков к определенному классу.

Наборы данных были поделены на обучающую и тестовую выборки, содержащие 70 и 30% данных соответственно. Для лучшей работы рекуррентных алгоритмов наборы данных были случайным образом перемешаны. Обучающие выборки использовались для инициализации классификатора с помощью нечеткой кластеризации, а тестовые выборки – для сравнения точности классификации.

Введенные рекуррентные алгоритмы во всех экспериментах показали более высокую точность кластеризации по сравнению с известными процедурами [18].

Еще одним достоинством адаптивного подхода является возможность обработки в реальном времени рядов различной природы [20].

3. Робастные адаптивные алгоритмы фаззи-кластеризации

Рассмотренные подходы позволяют эффективно решать задачу классификации в условиях существенного пересечения кластеров, при этом, однако, предполагается, что данные внутри каждого кластера располагаются достаточно компактно без резких (аномальных) выбросов. Вместе с тем следует отметить, что реальные данные, как правило, загрязнены выбросами, доля которых по некоторым оценкам [21-23] составляет до 20%, так что говорить о компактном расположении данных не всегда корректно.

В связи с этим последнее время большое внимание уделяется задачам нечеткого кластер-анализа данных, плотность распределения которых отличается от нормального наличием «тяжелых хвостов». В работах [24-28] предложены различные модификации упомянутых выше процедур кластеризации, предназначенные для обработки наблюдений, содержащих выбросы.

3.1. Робастный рекуррентный вероятностный алгоритм нечеткой кластеризации

Для предварительно стандартизованных векторов признаков (стандартизация выполняется покомпонентно так, чтобы все исходные векторы принадлежали единичному гиперкубу $[0,1]^n$)

введем целевую функцию

$$E\left(w_{k,j},c_j\right)=\sum_{k=1}^{N}\sum_{j=1}^{m}w_{k,j}^{\beta}D\left(x_k,c_j\right) \qquad (27)$$

при ограничениях

$$\sum_{j=1}^{m}w_{k,j}=1,\quad k=1,\ldots,N, \qquad (28)$$

$$0<\sum_{k=1}^{N}w_{k,j}\le N,\quad j=1,\ldots,m. \qquad (29)$$

Здесь $D\left(x_k,c_j\right)$ – расстояние между x_k и c_j в принятой метрике. Результатом кластеризации предполагается $N\times m$ матрица $W=\{w_{k,j}\}$, называемая «матрицей нечеткого разбиения».

Как правило, в качестве функции расстояния $D\left(x_k,c_j\right)$ применяется L^p метрика Минковского [29]

$$D\left(x_k,c_j\right)=\left(\sum_{i=1}^{n}\left|x_{k,i}-c_{j,i}\right|^p\right)^{\frac{1}{p}},\ p\ge 1, \qquad (30)$$

где $x_{k,i}$, $c_{j,i}$ - i-е компоненты $(n\times 1)$- векторов x_k, c_j соответственно.

Оценки, связанные с квадратичными целевыми функциями, являются оптимальными в случаях, когда данные принадлежат классу распределений с ограниченной дисперсией, наиболее известным представителем которых является гауссово. Варьирование параметра p позволяет улучшить свойства робастности процедур кластеризации, однако качество оценивания определяется видом распределения данных. Так, оценки при $p=1$ оптимальны для лапласовского распределения данных, однако их построение связано с большими вычислительными затратами.

Достаточно реалистичным является класс приближенно нормальных распределений [30]. Приближенно нормальные распределения представляют собой смесь гауссовой плотности и распределения некоторой произвольной плотности, загрязняющего нормальное выбросами. Оптимальной целевой функцией в этом случае является квадратично-линейная, причем линейный характер проявляется по мере удаления от минимума.

Наиболее ярким представителем приближенно нормальной плотности распределения является функция

$$p(x_i,c_i)=Se(c_i,s_i)=\frac{1}{2s_i}\operatorname{sech}^2\frac{x_i-c_i}{s_i}, \qquad (31)$$

где c_i и s_i - параметры, задающие центр и ширину распределения соответственно.

Эта функция напоминает гауссову в окрестности центра, однако отличается более тяжелыми хвостами. С распределением (31) связана целевая функция [31, 32]

$$f_i(x_i,c_i)=\beta_i\ln\cosh\frac{x_i-c_i}{\beta_i}, \qquad (32)$$

где параметр β_i задает крутизну этой функции, при этом в окрестности минимума функция весьма близка к квадратичной, стремясь по мере роста к линейной.

Интересен также тот факт, что производная этой функции

$$f_i'(x_i)=\varphi(x_i)=\tanh\frac{x_i}{\beta_i} \qquad (33)$$

является стандартной активационной функцией искусственных нейронных сетей [33].

Используя в качестве метрики конструкцию

$$\begin{aligned}D^R\left(x_k,c_j\right)&=\sum_{i=1}^{n}f_i\left(x_{k,i},c_{j,i}\right)=\\&=\sum_{i=1}^{n}\beta_i\ln\cosh\frac{x_{k,i}-c_{j,i}}{\beta_i},\end{aligned} \qquad (34)$$

можно ввести целевую функцию робастной классификации

$$\begin{aligned}E^R\left(w_{k,j},c_j\right)&=\sum_{k=1}^{N}\sum_{j=1}^{m}w_{k,j}^{\beta}D^R\left(x_k,c_j\right)=\\&=\sum_{k=1}^{N}\sum_{j=1}^{m}w_{k,j}^{\beta}\sum_{i=1}^{n}\beta_i\ln\cosh\frac{x_{k,i}-c_{j,i}}{\beta_i},\end{aligned} \qquad (35)$$

и соответствующую функцию Лагранжа

$$L\left(w_{k,j},c_j,\lambda_k\right)= \\ =\sum_{k=1}^{N}\sum_{j=1}^{m}w_{k,j}^{\beta}\sum_{i=1}^{n}\beta_i \ln\cosh\frac{x_{k,i}-c_{j,i}}{\beta_i}+ \\ +\sum_{k=1}^{N}\lambda_k\left(\sum_{j=1}^{m}w_{k,j}-1\right), \tag{36}$$

где λ_k - неопределенный множитель Лагранжа, обеспечивающий выполнение ограничений (28), (29).

Седловая точка функции Лагранжа (36) может быть найдена путем решения системы уравнений Куна-Таккера

$$\begin{cases}\dfrac{\partial L(w_{k,j},c_j,\lambda_k)}{\partial w_{k,j}}=0,\\ \dfrac{\partial L(w_{k,j},c_j,\lambda_k)}{\partial \lambda_k}=0,\\ \nabla_{c_j}L(w_{k,j},c_j,\lambda_k)=0.\end{cases} \tag{37}$$

Решения первого и второго уравнений системы (37) приводят к известному результату

$$\begin{cases}w_{k,j}=\dfrac{\left(D^R(x_k,c_j)\right)^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m}\left(D^R\left(x_k,c_l\right)\right)^{\frac{1}{1-\beta}}},\\ \lambda_k=-\left(\sum_{l=1}^{m}\left(\beta D^R\left(x_k,c_l\right)\right)^{\frac{1}{1-\beta}}\right)^{1-\beta},\end{cases} \tag{38}$$

однако третье уравнение

$$\nabla_{c_j}L(w_{k,j},c_j,\lambda_k)= \\ =\sum_{k=1}^{N}w_{k,j}^{\beta}\nabla_{c_j}D^R\left(x_k,c_j\right)=0, \tag{39}$$

очевидно, не имеет аналитического решения. Решение уравнения (39) может быть получено с помощью локальной модификации функции Лагранжа [24] и рекуррентных алгоритмов нечеткой кластеризации [12]. Поиск седловой точки локальной функции Лагранжа

$$L_k(w_{k,j},c_j,\lambda_k)= \\ =\sum_{j=1}^{m}w_{k,j}^{\beta}D^R\left(x_k,c_j\right)+\lambda_k\left(\sum_{j=1}^{m}w_{k,j}-1\right) \tag{40}$$

с помощью процедуры Эрроу-Гурвица-Удзавы приводит к алгоритму

$$\begin{cases}w_{k,j}^{pr}=\dfrac{\left(D^R(x_k,c_j)\right)^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m}\left(D^R\left(x_k,c_l\right)\right)^{\frac{1}{1-\beta}}},\\ c_{k+1,j,i}=c_{k,j,i}-\\ -\eta_k\dfrac{\partial L_k\left(w_{k,j},c_j,\lambda_k\right)}{\partial c_{j,i}}=\\ =c_{k,j,i}+\eta_k w_{k,j}^{\beta}\tanh\dfrac{x_{k,i}-c_{k,j,i}}{\beta_i},\end{cases} \tag{41}$$

где η_k - параметр скорости обучения, $c_{k,j,i}$ - i-й компонент j-го прототипа, вычисленный на k-м шаге.

Несмотря на низкую вычислительную сложность, алгоритм (41) имеет недостаток, присущий всем вероятностным алгоритмам кластеризации.

3.2. Робастный рекуррентный возможностный алгоритм нечеткой кластеризации

Для возможностных алгоритмов нечеткой кластеризации критерием является выражение

$$E^R(w_{k,j},c_j,\mu_j)=\sum_{k=1}^{N}\sum_{j=1}^{m}w_{k,j}^{\beta}D^R\left(x_k,c_j\right)+ \\ +\sum_{j=1}^{m}\mu_j\sum_{k=1}^{N}\left(1-w_{k,j}\right)^{\beta}. \tag{42}$$

При минимизации (42) по параметрам $w_{k,j}$, c_j и μ_j получают систему уравнений

$$\begin{cases}\partial E^R(w_{k,j},c_j,\mu_j)/\partial w_{k,j}=0,\\ \partial E^R(w_{k,j},c_j,\mu_j)/\partial \mu_j=0,\\ \nabla_{c_j}E^R(w_{k,j},c_j,\mu_j)=0,\end{cases} \tag{43}$$

при этом решение первых двух уравнений системы

(43) приводит к известному результату

$$\begin{cases} w_{k,j}^{pos} = \left(1 + \left(\dfrac{D^R\left(x_k, c_j\right)}{\mu_j}\right)^{\frac{1}{\beta-1}}\right)^{-1}, \\ \mu_j = \dfrac{\sum_{k=1}^{N} w_{k,j}^{\beta} D^R\left(x_k, c_i\right)}{\sum_{k=1}^{N} w_{k,j}^{\beta}}, \end{cases} \quad (44)$$

а третье

$$\nabla_{c_j} E^R(w_{k,j}, c_j, \mu_j) = \\ = \sum_{k=1}^{N} w_{k,j}^{\beta} \nabla_{c_j} D^R\left(x_k, c_j\right) = 0 \quad (45)$$

полностью соответствует (39).

Вводя локальную модификацию (42)

$$E_k^R\left(w_{k,j}, c_j, \mu_j\right) = \\ = \sum_{j=1}^{m} w_{k,j}^{\beta} D^R\left(x_k, c_j\right) + \sum_{j=1}^{m} \mu_j \left(1 - w_{k,j}\right)^{\beta} = \\ = \sum_{j=1}^{m} w_{k,j}^{\beta} \sum_{i=1}^{n} \beta_i \ln \cosh \frac{x_{k,i} - c_{j,i}}{\beta_i} + \\ + \sum_{j=1}^{m} \mu_i \left(1 - w_{k,j}\right)^{\beta} \quad (46)$$

и оптимизируя ее, имеем

$$\begin{cases} w_{k,j}^{pos} = \left(1 + \left(\dfrac{D^R\left(x_k, c_j\right)}{\mu_j}\right)^{\frac{1}{\beta-1}}\right)^{-1}, \\ c_{k+1,j,i} = c_{k,j,i} - \\ -\eta_k \dfrac{\partial E_k^R\left(w_{k,j}, c_j, \mu_j\right)}{\partial c_{j,i}} = \\ = c_{k,j,i} + \eta_k w_{k,j}^{\beta} \tanh \dfrac{x_k - c_{k,j,i}}{\beta_i}, \end{cases} \quad (47)$$

где параметр расстояния $\mu_{k,j}$ может рассматриваться в соответствии со вторым уравнением системы (44), в случае k наблюдений вместо объема всей выборки N.

Следует отметить, что последние уравнения систем (41) и (47) полностью идентичны и определяются лишь выбором метрики. Это обстоятельство позволяет использовать любую подходящую метрику для конкретного случая, которая будет определять только процедуру настройки прототипов, при этом уравнение расчета весов останется прежним.

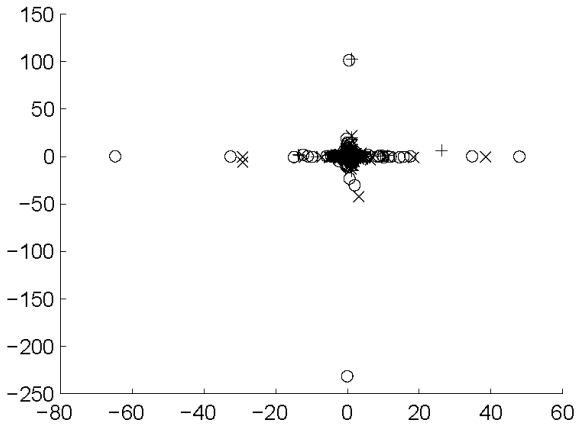
Рассмотренные робастные рекуррентные методы могут применяться как в многопроходном пакетном режиме, так и в режиме поступления наблюдений. В последнем случае – номер наблюдения k будет представлять собой дискретное время.

Эксперименты с репозиторными данными, искаженными аномальными выбросами, показали высокую эффективность предложенных алгоритмов при обработке информации, заданной как в виде таблиц и «объект-свойство» [34, 35], так и в форме временных рядов [36].

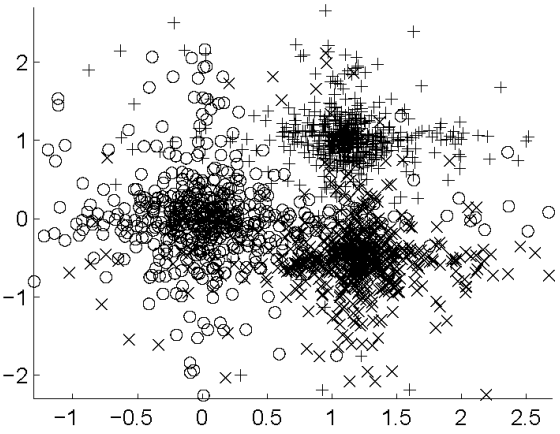
В частности, рассматривалась задача классификации данных на специально искусственно сгенерированной выборке, содержащей три двумерных кластера данных, наблюдения которых помечены символами «о», «х» и «+» (см. рис. 1). Точки в каждом кластере выборки распределены согласно плотности распределения Лапласа, имеющего «тяжелые хвосты»

$$p(x_i) = \sigma(1 + (x_i - c)^2)^{-1}, \quad (48)$$

где σ и c - ширина и ожидание соответственно.



а



б

Рис. 1. Полная выборка (а) и ее центральная часть (б)

Выборка содержит 9000 наблюдений (3000 в каждом кластере) и разделена на обучающую (7200 наблюдений) и проверочную (1800 наблюдений) подвыборки. Следует отметить, что некоторые наблюдения находятся очень далеко от центров кластеров (рис. 1,а). Прототипы кластеров находятся в центральной области данных, как показано на рис. 1,б. Для корректного нахождения прототипов алгоритм кластеризации должен быть не чувствительным к выбросам.

Для всех алгоритмов, участвующих в сравнении, процедура эксперимента была следующей. Вначале обучающая выборка была кластеризована соответствующим алгоритмом и прототипы кластеров были найдены. Затем обучающая и проверочная выборки были классифицированы по результатам кластеризации. Принадлежность наблюдения к каждому кластеру в процессе классификации вычислялась в соответствии с уравнениями (9), (41) или (47) в зависимости от типа алгоритма кластеризации. Кластер, к которому принадлежало наблюдение с максимальной степенью принадлежности, определял класс этого наблюдения. Классификация и обучения проводились в режиме поступления наблюдений при $\beta = 2$, $\beta_1 = \beta_2 = \beta_3 = 1$, $\eta(k) = 0.01$.

Результаты приведены в таблице

Результаты классификации

Алгоритм	**Ошибка классификации**	
	на обучающей выборке	**на проверочной выборке**
«fuzzy C-means» Бездека	17.1 % (1229 набл.)	16.6 % (299 набл.)
Робастный вероятностный (17)	15.6 % (1127 набл.)	15.6 % (281 набл.)
Робастный возможностный (23)	15.2 % (1099 набл.)	14.6 % (263 набл.)

На рис. 2 легко можно заметить, что центры кластеров (прототипы), полученные с помощью алгоритма «fuzzy C-means» Бездека, смещены от визуальных центров кластеров, благодаря наличию «тяжелых хвостов» плотности распределения данных, в отличие от методов с робастной целевой функцией (41) и (47), при которых прототипы найдены более точно, что подтверждается меньшей ошибкой классификации.

Заключение

Вычислительный интеллект является мощной методологией для решения широкого круга задач анализа данных.

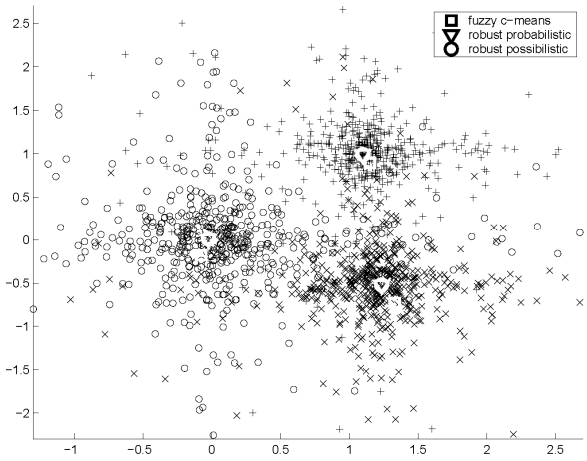


Рис. 2. Прототипы кластеров

Постоянный рост успешного применения

технологий вычислительного интеллекта в области анализа данных подтверждает многосторонность этого подхода. В то же время реальные задачи, которые возникают при обработке очень больших баз данных, осложняют использование существующих алгоритмов и требуют усовершенствования инструментария для решения задач интеллектуального анализа данных в реальном времени с использованием парадигм самообучения и мягких вычислений.

Литература

1. Jang J.-S. R., Sun C.-T., Mizutani E. Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence. – Upper Saddle, NJ: Prentice Hall, 1997. – 614 p.

2. Zadeh L. A. Fuzzy sets // Information and Control. – 1965. – 8. – P. 338–353.

3. Kruse R., Nauck D., Borgelt C. Data mining with fuzzy methods: status and perspectives // Proc. 7th European Congress on Intelligent Techniques and Soft Computing (EUFIT'99, Aachen, Germany). – Aachen: Verlag Mainz. – 1999. – CDROM.

4. MacQueen J. On convergence of k-means and partitions with minimum average variance // Ann. Math. Statist. – 1965. – 36. – P. 1084-1093.

5. Cover T. M. Estimates by the nearest-neighbor rule // IEEE Trans. on Information Theory. – 1968. – 14. – P. 50–55.

6. Bezdek J. C. Pattern Recognition with Fuzzy Objective Function Algorithms. – New York: Plenum Press, 1981 – 272 p.

7. Hoeppner F., Klawonn F., Kruse R. Fuzzy-Klusteranalyse. Verfahren fuer die Bilderkennung, Klassifikation und Datenanalyse. – Braunschweig: Vieweg, 1996. – 280 S.

8. Gustafson E. E., Kessel W. C. Fuzzy clustering with a fuzzy covariance matrix // Proc. IEEE CDC. – San Diego, California. – 1979. – P. 761–766.

9. Yager R. R., Filev D. P. Approximate clustering via the mountain method // IEEE Trans. on Syst., Man and Cybern. – 1994. – 24. – P. 1279–1284.

10. Gath I., Geva A. B. Unsupervised optimal fuzzy clustering // IEEE Trans. on Pattern Analysis and Machine Intelligence. – 1989. – 11. – P. 773–781.

11. Krishnapuram R., Keller J. A possibilistic approach to clustering // IEEE Trans. on Fuzzy Systems. – 1993. – 1. – P. 98–110.

12. Bodyanskiy Ye., Kolodyazhniy V., Stephan A. Recursive fuzzy clustering algorithms // Proc. East-West Fuzzy Colloquim 2002. – Zittau/ Goerlitz:HS, 2002. – P. 164-172.

13. Bodyanskiy Ye., Chaplanov O., Kolodyazhniy V. Soft computing techniques for data mining // Proc. Pre-Conf. Workshop 29th Int. Conf. on Very Large Data Bases VLDB 2003. Emerging Database Research in East Europe. – Cottbus: Brandenburg University of Technology, 2003. – P. 1-4.

14. Krishnapuram R., Keller J. Fuzzy and possibilistic clustering methods for computer vision // IEEE Trans. on Fuzzy Systems. – 1993. – 1. – P. 98–110.

15. Klawonn F., Kruse R. Constructing a fuzzy controller from data // Fuzzy Sets and Systems. – 1997. – 85. – P. 177–193.

16. Chung F. L., Lee T. Fuzzy competitive learning // Neural Networks. – 1994. – 7. – P. 539–552.

17. Park D. C., Dagher I. Gradient based fuzzy c-means (GBFCM) algorithm // Proc. IEEE Int. Conf. on Neural Networks. – 1984. – P. 1626–1631.

18. Bodyanskiy Ye. Computational intelligence techniques for data analysis // Lecture Notes in Informatics. – V. P-72. – Bonn: GI, 2005. – P. 15-36.

19. The UCI Repository of Machine Learning Databases and Domain Theories. http://www.ics.uci.edu/mlearn/MLRepository.html.

20. Resursive fuzzy clustering algorithm for segmentation of biomedical time series / Bodyanskiy Ye.,Gorshkov Ye., Kokshenev I.,

Kolodyazhniy V., Shilo O. // Proc. East West Fuzzy Colloquium 2006. – Zittau / Goerlitz: HS, 2006. – P.130-139.

21. Barnett V., Lewis T. Outliers in Statistical Data. – Chichester-New York-Brisbane-Toronto: John Wiley and Sons, 1978. – 584 p.

22. Rey W. J. J. Robust Statistical Methods // Lecture Notes in Mathematics. – Berlin-Heidelberg-New York: Springer-Verlag. – 1978. – 314 p.

23. Huber P. J. Robust Statistics. – New York: John Wiley and Sons, 1981. – 320 p.

24. Looney C. G. A fuzzy clustering and fuzzy merging algorithm. – Technical Report, CSUNR- 101-1999, 1999. http://sherry.ifi.unizh.ch/looney 99fuzzy.html.

25. Looney C. G. A fuzzy classifier with ellipsoidal Epanechnikovs. – Technical Report, Computer Science Department, University of Nevada, Reno, NV, 2001. http://sherry.ifi.unizh.ch/looney01fuzzy.html.

26. Sequential fuzzy cluster extraction and its robustness against noise / Tsuda K., Senda S., Minoh M., Ikeda K. // Systems and Computers in Japan. – 1997. – 28. – P. 10–17.

27. Hoeppner F., Klawonn F. Fuzzy clustering of sampled functions // Proc. 19th Int. Conf. of the North American Fuzzy Information Processing Society (NAFIPS), Atlanta, USA. – 2000. – P. 251–255.

28. Georgieva O., Klawonn F. A clustering algorithm for identification of single clusters in large data sets // Proc. East-West Fuzzy Colloquium 2004. – Zittau –Goerlitz: HS. – 2004. – P. 118–125.

29. Pau L. F. Failure Diagnosis and Performance Monitoring. - NY: Marcel Dekker Inc., 1981. – 427 p.

30. Цыпкин Я.З. Основы информационной теории идентификации. – Москва: Наука. ГРФМЛ, 1984. – 320 с.

31. Holland P. W., Welsh R. E. Robust regression using iteratively re-weighted least squares // Comm. Statist. Theory and Methods. – 1977. – P. 813–827.

32. Welsh R. E. Nonlinear statistical data analysis // Proc. Comp. Sci. and Statist. Tenth-Ann. Symp. Interface. Nat'l Bur. Stds. Gaithersburg, MD. – 1977. – P. 77–86.

33. Chichocki A., Unbehauen R. Neural Networks for Optimization and Signal Processing. – Stuttgart: Teubner, 1993. – 526 p.

34. Robust recursive fuzzy clustering algorithms / Bodyanskiy Ye., Gorshkov Ye., Kokshenev I., Kolodyazhniy V. // Proc. East West Fuzzy Colloquim 2005. - Zittau/ Goerlitz: HS, 2005. – P. 301-308.

35. Outlier resistant recursive fuzzy clustering algorithm / Bodyanskiy Ye., Gorshkov Ye., Kokshenev I., Kolodyazhniy V. // In "Computational Intelligence: Theory and Applications". – Ed. By B. Reusch – Advances in Soft Computing. – Berlin-Heidelberg: Springer-Verlag. – Vol. 38. – 2006. – P. 647-652.

36. Robust recursive fuzzy clustering-based segmentation of biomedical time series / Bodyanskiy Ye.,Gorshkov Ye., Kokshenev I., Kolodyazhniy V., Shilo O. // Proc. 2006 Int. Symp. on Evolving Fuzzy Systems. – Lancaster, UK. – 2006. – P.101-105.

Поступила в редакцию 08.02.2007

Рецензент: лауреат Государственной премии Украины, д-р. техн. наук, проф. А. С. Кулик, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков.