

## İNFORMATİKA

УДК 01.23

НЕЧЕТКАЯ КЛАСТЕРИЗАЦИЯ ОБЪЕКТОВ В РАМКАХ  
НЕЧЕТКОЙ ЛОГИКИ В ШИРОКОМ СМЫСЛЕ

У.Ш.РЗАЕВА

*Институт Кибернетики Академии Наук Азербайджана**u.rzayeva@aseu.az*

*Рассматривается задача кластеризации объектов с нечеткими признаками, принимающие лингвистические значения, для решения которой используется нечеткая логика в широком смысле (FLb). Представлен метод распознавания на основе естественного языка, по которому определяется специальный класс лингвистических синтагм в FLb. Для построения алгоритма классификации задачи использованы лингвистические описания, дана схема, по которой устанавливаются зависимости между объектами и между объектами и классами.*

**Ключевые слова:** синтагма, предикат, интенсивность, лингвистическое описание, нечеткая логика в широком смысле, импликация.

Распознавание образов или классификация объектов является одной из наиболее активно развивающихся областей прикладной математики и математической кибернетики, что обусловлено, в первую очередь, потребностями практики. На практике часто встречаются задачи, в которых рассматриваются достаточно сложные процессы или явления. По сей день еще не разработан строгий формальный аппарат, предназначенный для построения математических моделей сложных процессов и явлений, адекватных моделируемым процессам или явлениям. Математическое моделирование пока еще является искусством, а качество моделей в значительной мере зависит от интуиции и изобретательности их разработчиков. Одним из способов обойти проблему математического моделирования является построение систем распознавания на основе накопленной информации (признаковое пространство, отношения между признаками и т.д.) [2,3].

Основной подход состоит в описании классов объектов через определенные значения признаков, и каждому объекту ставится в соответствие вектор признаков, по которому происходит его распознавание. Модели распознавания используют различные процедуры и функции, разде-

ляющие объекты на классы. Эксперимент обучения без учителя при решении задачи распознавания образов можно сформулировать как задачу кластерного анализа. Таким образом, целью алгоритма кластеризации является автоматическая классификация множества объектов, которые задаются векторами признаков в признаковом пространстве. Большую актуальность имеют методы кластеризации общего типа, не зависящие от типа используемых признаков и обрабатывающие разнотипные данные, характерные для описания нелинейных систем. Среди таких методов наиболее приспособленными являются методы вычисления оценок [3]. Не уступая методам, ориентированным на анализ данных одного типа в вычислительном отношении, методы вычисления оценок превосходят последние по легкости интерпретации результатов, возможностям учесть разную априорную информацию, резерву адаптивности к новым требованиям и задачам. При их практической реализации используется аналогия между построением образа при помощи соединения различными способами подобразов и построением фраз и предложений языка при помощи соединения слов, состоящих, в свою очередь, из букв. Таким образом, этот метод распознавания связан с использованием аналогии между структурой образов и синтаксисом естественного языка [1].

По сей день не существует критерия для установления полноты пространства признаков, поэтому для установления истины берется как можно больше признаков, затем синтезом признаков выбирается их оптимальное количество. Известно, что информация содержится не только в отдельных признаках, в основном она содержится в их сочетаниях (информативных сочетаниях). В ряде задач для описания данных, включающих показатели нечеткости [12], требуется уточнение смысла признаков, которые приносят информативность. Для управления результатом при выполнении импликации в таких задачах используется представление данных триплетом “объект-атрибут-значение”, поскольку большее количество информативных признаков ведет к усложнению вычислений.

Цель нашей работы заключается в моделировании информационных сочетаний, посредством которых происходит распознавание образа, с помощью синтагм в рамках нечеткой логики в широком смысле (FLb) [7].

Нечеткая логика в широком смысле (FLb) имеет возможность расширить возможности классической логики в тех областях, где классическая логика не может дать удовлетворительных решений. Существуют некоторые проблемы, связанные с естественным языком, для которого посредством FLb можно построить лучшую модель, нежели это возможно в классической логике. В данной работе FLb служит помощью при ответе на вопросы, поставленные в теории кластерного анализа в аспекте делимости объектов на кластеры в рамках установленных правил. В

работе представленная формализация относится лишь к небольшому подмножеству естественного языка, а именно, к тому, что используется при оценке поведения динамических систем и различных систем принятия решений. В рамках данной работы описываются две основные схемы рассуждений: элементарная дедукция на основе *modus ponens* [11], а также более сложная схема, целью которой является установление жесткой или функциональной зависимости между объектами и между объектами и кластерами.

Программа нечеткой логики связана со смыслом некоторых выражений естественного языка, которые называются синтагмами. *Синтагма* – это сочетание двух членов предложения, связанных тем или иным отношением с неравноправной направленностью членов, где один член является определяемым, а другой – определяющим. Между членами синтагм могут быть разные отношения, из которых главным является *предикативное*, которое выражает зависимость двух членов с обязательной связью времени и наклонения.

Перевод выражений естественного языка – синтагм - в термины нечеткой логики позволяет сформировать систему FLb по аналогии с любой другой формальной логической системой [7]. Элементами множества синтагм  $S$  являются, так называемые, оценочные синтагмы, то есть лингвистические выражения, которые определяют некоторую градацию («малый, средний, большой, очень маленький и т.д.»). Элементарные синтагмы называются также атомарными.

Пусть высказывания (синтагмы)  $Z_1, \dots, Z_n$  означают какие-либо свойства распознаваемых объектов, например, различные болезни, скрытые дефекты и неисправности в сложном механизме и т.д. Пусть высказывания  $X_1, \dots, X_m$  - регистрируемые в эксперименте признаки, по определенным совокупностям которых (информативным сочетаниям) можно установить наличие одних и отсутствие других свойств распознаваемых объектов.

Распознать наблюдаемый объект  $\omega'$  – значит установить, опираясь на априорные зависимости

$$E(X_1, \dots, X_m; Z_1, \dots, Z_n) = 1,$$

где  $E$  - известная булева функция и экспериментальные данные  $X_1, \dots, X_m$ , какими свойствами из числа  $Z_1, \dots, Z_n$  этот объект обладает, а какими не обладает. Если каждый тип объекта рассматривать как класс и обозначать  $K_j, j=1, \dots, n$ , то придем к стандартной постановке задачи распознавания: на основе вышеприведенных зависимостей и информативных сочетаниях  $X_1, \dots, X_m$ , определить, к какому классу  $K_1, \dots, K_n$  относится данный объект. [1,5]

**Определение 1.** Пусть  $\mathcal{A}$  является оценочной синтагмой. Тогда синтагма

$$\langle \text{Существительное} \rangle \text{ есть } \mathcal{A}$$

называется *оценочным предикатом*.

**Определение 2.** *Интенсивностью* лингвистической синтагмы называется соотношение ее к различным значениям истинности в различных возможных контекстах.

Фиксируем некоторый многосортный язык  $J$ , который имеет конечное число сортов  $l$  и ставим в соответствие синтагмы из  $S$  элементам  $J$ .  $F_j$  - множество корректно построенных формул соответствующих синтагм.

**Определение 3.** Пусть  $\mathcal{A} \in S$  является синтагмой, а  $A(x_1, \dots, x_n) \in F_j$  является соответствующей ей формулой. Тогда множество

$$\mathbf{A}_{(x_1, \dots, x_n)} = \{a_{t_1, \dots, t_n} / A_{x_1, \dots, x_n}[t_1, \dots, t_n] \mid t_1 \in M_{l_1}, \dots, t_n \in M_{l_n}\},$$

где  $M_{l_1}, M_{l_2}$  - множества термов вычислимых формул, называемых *мультиформулой*, является интенсивностью  $\mathcal{A}$ .

Исходя из вышеизложенных понятий и определений, постановку основной задачи аналогично задаче кластер-анализа сформулируем следующим образом [4,6,8]:

Пусть  $X$  - матрица размерности  $m \times n$ :

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

С таблицей  $X$  однозначно связаны упорядоченная последовательность ее строк

$$Y = \{y_1, y_2, \dots, y_m\} \quad (2)$$

такая, что для

$$\forall i \left( (y_i \in Y) \rightarrow (y_i = (x_{i1}, x_{i2}, \dots, x_{in})) \right), \quad (3)$$

и упорядоченная последовательность ее столбцов

$$Y^t = \{y_1^t, y_2^t, \dots, y_n^t\} \quad (2a)$$

такая, что

$$\forall j \left( (y_j^t \in Y^t) \rightarrow (y_j^t = (x_{1j}, x_{2j}, \dots, x_{mj})) \right). \quad (3a)$$

Строки (2) таблицы  $X$  будем также называть объектами (или векторами), а элементы строк (3) – признаками объектов (или компонентами векторов). Таким образом, задание числовой таблицы  $X$  (1) равносильно заданию упорядоченного множества  $Y$  из  $m$  объектов (2), каждый из которых характеризуется набором из  $n$  признаков (3), или, что то же, таб-

лица  $X$  задает координаты  $m$  векторов (2) в признаковом пространстве  $\mathbb{R}^N$ . Признаки предварительно шкалированы и в  $\mathbb{R}^N$  таким признакам сопоставляется набор гиперповерхностей, проходящих через узлы соответствующих шкал. Требуется разбить строки матрицы  $X$  на классы по некоторому критерию сходства.

Так как синтагмы  $A \in S$  являются названиями некоторых свойств объектов, присваивая им предикатные формулы, получим не только описания объектов, а также связи между объектами.

Для построения алгоритма классификации задачи использованы разные условные клаузы, являющиеся импликациями, описанными на естественном языке [6,7,8]. Множество таких утверждений называется лингвистическим описанием. Далее использован двухэтапный алгоритм классификации [5,8], в ходе первого этапа формируется предварительный набор кластеров, на втором этапе, то есть этапе уточнения классификации, используется нечеткая аппроксимация отношений между объектами  $X_i, i = 1, 2, \dots, m$  и классами  $K_j = 1, 2, \dots, n$ .

**Определение 4.** Пусть  $\mathcal{A}_i \in S, i = 1, \dots, m$  являются синтагмами с интенсивностями  $\mathbf{A}_i$  [6,7].

*Формальная теория FLb* есть

$$\mathcal{T} = \{ \mathcal{A}_0[\mathbf{A}_0], \dots, \mathcal{A}_m[\mathbf{A}_m] \} \quad (4)$$

Так как интенсивности являются мультиформулами, теория  $\mathcal{T}$  в (4) примыкает к нечеткой теории в узком смысле (FLn)  $T$  [3]:

$$T = BT \cup \mathbf{A}_0 \cup \dots \cup \mathbf{A}_m,$$

где  $BT$  является вспомогательной нечеткой теорией. Таким образом, все основные операции FLb могут трансформироваться в FLn.

Введем специальную синтагму

$$\mathcal{R} = \text{“} \langle \text{существительное} \rangle_1 \text{ в отношении с } \langle \text{существительное} \rangle_2 \text{”}$$

с интенсивностью

$$\mathbf{R}_{\langle x, y \rangle} = \{ r_{ts} / R_{x,y}[t, s] \mid t \in M_{l_1}, s \in M_{l_2} \},$$

где  $R$  является некоторым бинарным предикатным символом.

**Определение 5.** Пусть  $\mathcal{T}$  является теорией в FLb. Лингвистическое утверждение  $\mathcal{A}$ , которое может являться условной клаузой, верно в  $\mathcal{T}$ , если оно имеет интенсивность

$$\mathbf{A}_{(x_1, \dots, x_n)} = \{ 1 / A_{x_1, \dots, x_n}[t_1, \dots, t_n] \mid t_1 \in M_{l_1}, \dots, t_n \in M_{l_n} \}.$$

Данное отношение можно рассматривать как некоторое группирование пар элементов. Разложим его на пары и характеризуем каждую из пар элементов, используя оценочные утверждения

$$\text{“} \langle \text{существительное} \rangle_1 \text{ есть } \mathcal{A} \text{ и } \langle \text{существительное} \rangle_2 \text{ есть } \mathcal{B} \text{”},$$

где  $\langle \text{существительное} \rangle_1$  есть имя первого элемента каждой пары, а  $\langle \text{су-}$

*существительное*<sub>2</sub> - имя второго. Таким образом, каждая из рассматриваемых частей может быть описана на естественном языке с использованием условной клаузы вида

$$\mathcal{F} := \text{ЕСЛИ } \langle \text{существительное} \rangle_1 \text{ есть } \mathcal{A}u \langle \text{существительное} \rangle_2 \\ \text{есть } \mathcal{B}, \text{ ТО } \mathcal{R} \text{ (5)}$$

Условная клауза (5) является лингвистическим утверждением и рассматривается как истинное. Это означает, что интенсивность всей клаузы  $\mathcal{F}$  есть

$$\mathbf{P}_{\langle x,y \rangle} := \{1 / ((A_x[t] \wedge B_y[s]) \Rightarrow R_{x,y}[t,s]) \mid t \in M_{l_1}, s \in M_{l_2}\}$$

при том, что *существительное*<sub>1</sub> ставится в соответствие переменной  $x$ , а *существительное*<sub>2</sub> - переменной  $y$ .

**Определение 6.** Пусть  $\mathcal{A}_j, \mathcal{B}_j$  являются оценочными предикатами с соответствующими интенсивностями  $\mathbf{A}_j, \mathbf{B}_j$ . Тогда лингвистическое описание в FLb есть либо конечное множество  $\mathcal{LD}^I$ , либо конечное множество  $\mathcal{LD}^A$  следующих высказываний

$$\mathcal{LD}^I = \{ \mathcal{R}_1^I, \dots, \mathcal{R}_m^I \},$$

где  $\mathcal{R}_j^I = \text{ЕСЛИ } \mathcal{A}_j, \text{ ТО } \mathcal{B}_j, j=1, \dots, m$  являются условными клаузами;

$$\mathcal{LD}^A = \{ \mathcal{R}_1^A, \dots, \mathcal{R}_m^A \},$$

где  $\mathcal{R}_j^A = \mathcal{A}И \mathcal{B}_j, j=1, \dots, m$  являются составными оценочными предикатами.

Из этих рассуждений следует, что в FLb существуют два метода работы с лингвистической переменной. Первый из них работает с лингвистическим описанием  $\mathcal{LD}^I$ , состоящим из сформулированных на естественном языке логических импликаций. Второй метод основывается на дополнительном предположении и работает с лингвистическим описанием  $\mathcal{LD}^A$ , состоящем из конъюнкций лингвистических предикатов. Возникает вопрос о различии этих двух методов.

**Теорема 1.** Пусть  $\mathcal{LD}^A$  является простым лингвистическим описанием, состоящим из  $m$  правил и  $\mathcal{A}_j, j=1, \dots, m$  являются истинными условными клаузами вида (5). Предположим, что *существительное*<sub>1</sub> ставится в соответствие переменной  $x$ , а *существительное*<sub>2</sub> - переменной  $y$ , и пусть  $\mathcal{A}$  является простым оценочным предикатом с интенсивностью  $\mathbf{A}'_{\langle x \rangle}$ . Тогда существует теория в FLb

$$\mathcal{T} = \{ \mathcal{LD}^A, \mathcal{A}_j, \mathcal{A} \mid j=1, \dots, m \}$$

такая, что можно вывести следующее:

а. можно получить вывод  $\mathcal{B}$  в  $\mathcal{T}$  с интенсивностью

$$\mathbf{B}'_{\langle y \rangle} = \{ b'_s = \bigvee_{t \in M_{l_1}} (a'_t \wedge \bigvee_{j=1}^m (a_{j,t} \wedge b_{j,s})) \mid B_y[s] \mid s \in M_{l_2} \}$$

где  $B'(y) := (\exists x)(A'(x) \wedge R(x, y))$  и все  $b'_s = \bigvee_{t \in M_{l_1}} (a'_t \wedge \bigvee_{j=1}^m (a_{j,t} \wedge b_{j,s}))$  являются максимальными;

б. лингвистическое высказывание

ЕСЛИ  $\mathcal{LD}^A$  ТО  $\mathcal{R}$

верно в  $\mathcal{T}$  и порождает кластеризацию между объектами  $X_i, i = 1, 2, \dots, m$  и классами  $K_j = 1, 2, \dots, n$ .

**Доказательство.**

а. Так как  $\mathcal{LD}^A$  является простым лингвистическим описанием, а формула  $R(x, y)$  с интенсивностью  $\mathbf{R}_{\langle x, y \rangle} = \{r_{ts}/R_{x,y}[t, s] \mid t \in M_{l_1}, s \in M_{l_2}\}$  является атомарной, мы можем построить нечеткую теорию  $T$ , присоединенную к  $\mathcal{T}$ , в которой соотношение получено дизъюнкцией конъюнкций независимых формул.

б. Используя тавтологию  $T \vdash (A \Rightarrow C) \Rightarrow ((B \Rightarrow C) \Rightarrow (A \vee B) \Rightarrow C)$  и правило modus ponens, получаем  $T \vdash \bigvee_{i=1}^m (A_x[t] \wedge B_y[s]) \Rightarrow R_{x,y}[t, s]$  для всех  $t \in M_{l_1}, s \in M_{l_2}$ . Это означает, что лингвистическое описание имеет интенсивность

$$\{1 / \bigvee_{i=1}^m (A_x[t] \wedge B_y[s]) \Rightarrow R_{x,y}[t, s] \mid t \in M_{l_1}, s \in M_{l_2}\}$$

Определим  $\mathcal{S}$  иначе, чем это было показано выше, а именно, используя условные клаузы вида

$$\mathcal{S} = \text{ЕСЛИ } \mathcal{R} \text{ и } \langle \text{существительное} \rangle_1 \text{ есть } \mathcal{A}, \\ \text{ТО } \langle \text{существительное} \rangle_2 \text{ есть } \mathcal{B} \quad (6)$$

**Теорема 2.** Пусть  $\mathcal{S}_j, j = 1, \dots, m$  являются истинными условными клаузами вида (6), определяющими теорию FLb

$$\mathcal{T} = \{ \mathcal{S}_j \mid j = 1, \dots, m \}$$

Тогда существует лингвистическое описание  $\mathcal{LD}^T$  такое, что условная клауза

$$\text{ЕСЛИ } \mathcal{R} \text{ ТО } \mathcal{LD}^T \quad (7)$$

верна в  $\mathcal{T}$ .

**Доказательство.** Нечеткая теория  $T$ , присоединенная к  $\mathcal{T}$ , имеет вид

$$T = \{1 / ((R_{x,y}[t, s] \& A_{j,x}[t]) \Rightarrow B_{j,y}[s]) \mid j = 1, \dots, m, t \in M_{l_1}, s \in M_{l_2}\}.$$

Тогда, используя тавтологии

$$T \vdash (A \Rightarrow C) \Rightarrow ((B \Rightarrow C) \Rightarrow (A \vee B) \Rightarrow C)$$

и

$$T \vdash A \Rightarrow (B \Rightarrow (A \& B)),$$

закключаем, что

$$T \vdash R_{x,y}[t, s] \Rightarrow \bigwedge_{j=1}^m (A_{j,x}[t] \Rightarrow B_{j,y}[s]),$$

для всех  $t \in M_{l_1}, s \in M_{l_2}$ . Следовательно, мы получаем мультиформулу

$$\mathbf{P}_{\langle x, y \rangle} := \{1 / R_{x,y}[t, s] \Rightarrow \bigwedge_{j=1}^m (A_{j,x}[t] \Rightarrow B_{j,y}[s]) \mid t \in M_{l_1}, s \in M_{l_2}\},$$

которая является интенсивностью истинной условной клаузы (7) в  $\mathcal{T}$ .

Из этой теоремы следует, что лингвистическое описание  $\mathcal{LD}'$  делает возможным определение отношения  $\mathcal{R}$ : всякий раз, когда некоторые объекты состоят в рассматриваемом отношении, они могут быть определены, используя выводы оценочных предикатов. Описание  $\mathcal{LD}'$  используется, когда необходимо вывести утверждения из некоторых фактов, тогда как описание  $\mathcal{LD}^A$  применяется, когда существует надобность описать отношение или функцию [9,10].

### **Заключение**

Рассматривается задача кластеризации объектов с нечеткими признаками, принимающие лингвистические значения в рамках FLb. На основе введенных понятий с помощью синтагм описаны объекты, определены понятия близости двух объектов по одному признаку, и исходя из этого близостей объектов друг к другу и близости объекта к классу. Представлена попытка устранения существующего разрыва между приложением естественного языка и средствами его интерпретации и обеспечения точного выражения смысла высказываний, чтобы понимание естественного языка было более осмысленным как при описании объектов, подлежащих кластеризации, так и при построении алгоритма для решения данной проблемы.

### **ЛИТЕРАТУРА**

1. Горелик А.Л., Гуревич И.Б., Скрипкин В.А. Современное состояние проблемы распознавания. М.: Радио и связь, 1985, 160 с.
2. Дородницын А.А. Проблемы математического моделирования в описательных науках // Кибернетика, 1983, №4, с.6-10.
3. Журавлев Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики, 1978, №33, с.5-68.
4. Зенкин А.А., Зенкин А.М. Задача построения оптимальных классификаций // Сборник работ по математической кибернетике, 1981, с. 20-33.
5. Керимов А.К. Новое решающее правило в задаче автоматической классификации // Изв. НАН Азербайджана, Серия физико-математических и технических наук, 2005, т. XXV, № 3, с.95-98.
6. Керимов А.К., Давудова Р.И. Эволюционный алгоритм для решения задачи автоматической классификации // Искусственный интеллект и принятие решений, 2009, №4 с.74-79.
7. Новак В., Перфильева И., Мочкорж И. Математические принципы нечеткой логики. М.: Физматлит, 2006, 352 с.
8. Kerimov A., Moharrami S. Automatic Classification with Neural Networks Using New Decision Rule. // International Journal of Applied Mathematics & Statistics, 2010, №4, p.90-96.
9. Kerimov A.K., Rzayeva U.Sh. About one problem of objects' fuzzy clusterization on the basis of fuzzy logic in broad sense. / The Third Congress of the World Mathematical Society of Turkic Countries, Almaty, 2009, p.147-152.
10. Kerimov A.K., Rzayeva U.Sh. Fuzzy interpolation of partial functions of membership,

- characterizing affinity of objects to each other and objects to class / Ninth International Conference on Applications of Fuzzy Systems and Soft Computing, Prague, 2010, p.229-235.
11. Mamdani E.H., Assilian S. An experiment in linguistic synthesis with fuzzy logic controller // International Journal of Man-Machine, 1975, v. 7, №1, p.1-13.
  12. Zadeh L.A. The concept of a linguistic variable and its application to approximate reasoning I, II, III // Information Sciences, 1975, v. 8, p. 199-249, 301-357; v. 9, p. 43–80.

## **GENİŞ MƏNADA QEYRİ-SƏLİS MƏNTİQ ƏHATƏSİNDƏ OBYEKT LƏRİN QEYRİ-SƏLİS KLASTERİZASİYASI**

**Ü.Ş. RZAYEVA**

### **XÜLASƏ**

Qeyri-səlis əlamətli, linqvistik qiymətləri qəbul edən və həllində geniş mənada qeyri-səlis məntiqdən (FLb) istifadə edən obyektlərin klasterizasiyası məsələsinə baxılır. Təbii dil əsasında tanıma metodu təklif olunub ki, ondan istifadə etməklə FLb-də linqvistik sintaqmaların xüsusi sinfi müəyyən olunur. Məsələnin təsnifat alqoritminin qurulması üçün linqvistik təsvirlərdən istifadə olunub, müəyyən sxem əsasında obyektlər arası və obyektlərlə sinif arasında asılılıqlar qurulub.

**Açar sözlər:** sintaqma, predikat, intensivlik, linqvistik təsvir, geniş mənada qeyri-səlis məntiq, implikasiya.

### **FUZZY CLUSTERIZATION OF OBJECTS WITHIN FUZZY LOGIC IN BROAD SENSE**

**U.Sh.RZAYEVA**

### **SUMMARY**

The article considers the problem of clusterization of objects possessing fuzzy characteristics and taking linguistic values, the solution of which uses fuzzy logic in a broad sense (FLb). The submitted recognition method bases on the natural language that determines a special class of linguistic syntagmas in FLb. The algorithm of the classification uses linguistic descriptions, and the scheme which establishes the relationships between objects and between objects and classes is given.

**Key words:** syntagma, predicate, intensity, linguistic description, fuzzy logic in a broad sense, implication.

*Поступила в редакцию: 05.03.2011 г.*

*Принято к печати: 17.06.2011 г.*