А.К. Керимов, У.Ш. Рзаева

Кластеризация объектов с помощью лингвистических описаний в рамках теории FLb

Аннотация. Рассматривается задача кластеризации объектов с нечеткими признаками, принимающие лингвистические значения, для решения которой используется нечеткая логика в широком смысле (FLb). Определяется специальный класс лингвистических синтагм в FLb. Задача кластеризации формулируется также в терминах нечеткой логики в широком смысле, вводится вспомогательная (непротиворечивая) нечеткая теория, в рамках которой определяется однозначная принадлежность объектов друг к другу и объектов к классу.

Ключевые слова: нечеткая логика в широком смысле (FLb), информативные сочетания, синтагма, предикат, условная клауза, лингвистическая переменная, интенсивность.

Введение

Известно, что термин “кластерный анализ” охватывает ряд различных алгоритмов классификации, которые группируют объекты в кластеры. Кластерный анализ является одной из парадигм технологии Soft Computing, на основе которой построены различные системы искусственного интеллекта (ИИ). Синтезируя определения ИИ из различных источников, в данной работе в качестве рабочего определения можно предположить следующее: искусственный интеллект – одно из направлений информационных технологий, целью которого является разработка средств, позволяющих пользователю или программисту ставить и решать свои традиционно считающиеся интеллектуальными задачи, общаясь с компьютером на ограниченном подмножестве естественного языка.

Среди множества направлений ИИ есть несколько ведущих, одним из которых является распознавание образов, выделившееся в самостоятельную науку. 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, разделяющие 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.

1. Постановка задачи

Понятие нечеткой логики обычно используется в двух смыслах - узком и широком. В узком смысле, нечеткая логика – это логическая система, являющаяся расширением многозначной логики. В широком смысле слова, который сегодня преобладает, нечеткая логика равнозначна теории нечетких множеств, т.е. классов с неточными, размытыми границами [12]. Таким образом, нечеткая логика, понимаемая в узком смысле, является разделом нечеткой логики в широком смысле. Здесь вопрос о принадлежности к множеству - вопрос степени принадлежности.

Нечеткая логика в широком смысле (FLb) позволяет 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 разделимости 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 [7], а также более сложная схема, целью которой является установление жесткой или функциональной зависимости между объектами и между объектами и кластерами.

Критерия для установления полноты пространства признаков по сей день не существует, 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 [13], требуется уточнение смысла признаков, которые привносят информативность. Для управления результатом при выполнении импликации в таких задачах используется представление данных триплетами “объект-атрибут-значение”, поскольку большее количество информативных признаков ведет к усложнению вычислений. Одной из главных проблем задачи распознавания является нахождение тех признаков, которые обуславливают различия объектов разных классов.

Для построения алгоритмов и программ в задачах кластерного анализа необходимо задать правила работы с единицами различных уровней естественного языка [3]. Обобщенные языковые единицы определяются понятием синтагма, а синтагматические описания могут отражать ту или иную специфику исследования. Синтагматические обороты обладают большой информативной нагруженностью: они содержат дополнительное сообщение, сопутствующее сообщению, содержащемуся в распространяемой части предложения, и характеризуются относительной информативной самостоятельностью. В данной работе описание объекта синтагмой с точки зрения объема и содержания объекта, обозначаемого словом, составляет основу кластеризации. Цель работы заключается в моделировании информационных сочетаний, посредством которых происходит распознавание образа, с помощью синтагм в рамках нечеткой логики в широком смысле [9].

2. Методы решения

Пусть дано множество нечетких допустимых объектов

$$\widetilde{X} = \{\widetilde{x}_1,...,\widetilde{x}_m\}, \qquad (1)$$

представляющих собой сложные однотипные системы. Под однотипными понимаются системы, описание состояний которых дано в одном и том же признаковом пространстве. Любой объект из (1) может находиться в одном из своих конечных, расплывчатых состояний. Состояния объектов описываются набором некоторых нечетких признаков:

$$\widetilde{T} = \left\| x_{ij}, \mu_{X_j}\left(x_{ij}\right) \right\|, \; i = 1,...,m, \; j = 1,...,n, \qquad (2)$$

где $\widetilde{T}$ - прямоугольная матрица, $\left(x_{ij}, \mu_{X_j}\left(x_{ij}\right)\right)$ - j-ый нечеткий признак i-го объекта, X_j - нечеткое подмножество изменения j-го признака.

Совокупность всевозможных наборов значений признаков (2) образует пространство признаков размерности n. Требуется разбить строки матрицы (2) на подмножества по некоторому критерию сходства.

Нечеткие подмножества $\widetilde{K}_1,...,\widetilde{K}_l$ множества (1) являются кластерами, если

1. $\bigcup_{i=1}^{l} \widetilde{K}_i = \widetilde{X}$;
2. $\widetilde{K}_i \bigcap \widetilde{K}_j = \emptyset$ при $i \neq j, \quad i, j = 1,...,l$,

где l - число кластеров.

Касательно данной работы алгоритм нечеткой кластеризации (Fuzzy c-Means) [4] позволяет получить для каждой переменной кластеры. На основе полученных кластеров определяются термы (значения) входных и выходных переменных, выраженных синтагмами, т.е. каждый кластер инициализирует определенный терм синтагмой.

Для построения алгоритма классификации задачи использованы разные условные клаузы, являющиеся импликациями, описанными на ес-

тественном языке [8,10]. Множество таких утверждений называется лингвистическим описанием. Далее использован двухэтапный алгоритм классификации [2]. В ходе первого этапа формируется предварительный набор кластеров, на втором этапе, т.е. этапе уточнения классификации, используется нечеткая аппроксимация отношений между объектами $\widetilde{X} = \{\widetilde{x}_1,...,\widetilde{x}_m\}$ и классами $\widetilde{K}_1,...,\widetilde{K}_l$.

Для отражения сути поставленной задачи приведем некоторые определения. Существуют три концепции в логическом анализе естественного языка: интенсивность, расширение и возможный мир.

Возможным миром называется категория модальной логики, используемая для установления истинности или ложности модальных высказываний. В общих чертах возможный мир можно интерпретировать как возможное положение дел, либо возможное развитие событий. Возможный мир может быть расширен за счет вовлечения тех языковых средств, которые предоставляют право выбора при расшифровке смысла сказанного [1].

Интенсивностью называется совокупность мыслимых признаков обозначаемого понятием предмета или явления, которая может привести к различным значениям истинности в различных возможных мирах [5]. В логике интенсивность представляет собой функцию, ставящую в соответствие значение истинности объекту в каждом возможном мире.

Расширением является множество элементов, определенное одной интенсивностью, которая входит в значение синтагм в данном возможном мире [5].

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

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

называется *оценочным предикатом*.

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 2. Пусть $\mathscr{A} \in S$ является синтагмой, а $\mathrm{A}(x_1,...,x_n) \in F_j$ является соответствующей ей формулой. Тогда множество

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

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

Определение 3. Пусть $\mathscr{A}_i \in S$, $i=1,...,m$ являются синтагмами с интенсивностями $\mathbf{A}_i$.

Формальная теория FLb есть

$$\mathcal{T} = \{\mathscr{A}_0[\mathbf{A}_0],\ldots,\mathscr{A}_m[\mathbf{A}_m]\}. \qquad (3)$$

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

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

где BT является вспомогательной нечеткой теорией [10]. Таким образом, все основные операции FLb могут трансформироваться в FLn.

Общую схему формальной теории $\mathcal{T}$ формируем, используя естественный язык, при этом она ставится в соответствие нечеткой теории T в FLn. Затем в FLn производятся выводы. Результатом будет некоторая мультиформула, которую можно рассматривать как наиболее точную интенсивность соответствующей синтагмы, являющуюся выводом в FLb.

Введем специальную синтагму

$\mathscr{R}$:= "<*существительное*>$_1$ *в отношении с* "<*существительное*>$_2$"

с интенсивностью

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

где R является некоторым бинарным предикатным символом.

Определение 4. Пусть $\mathcal{T}$ является теорией в FLb. Лингвистическое утверждение $\mathscr{A}$, которое может являться условной клаузой, верно в $\mathcal{T}$, если оно имеет интенсивность

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

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

"<*существительное*>$_1$ *есть* $\mathscr{A}$
и <*существительное*>$_2$ *есть* $\mathscr{B}$",

где <*существительное*>$_1$ есть имя первого элемента каждой пары, а <*существительное*>$_2$ - имя второго. Таким образом, каждая из рассматриваемых частей может быть описана на естественном языке с использованием условной клаузы вида

$\mathscr{P}$:=ЕСЛИ <*существительное*>$_1$ есть $\mathscr{A}$
и <*существительное*>$_2$ есть $\mathscr{B}$, ТО $\mathscr{R}$ (4)

что может быть интерпретировано как формирование набора кластеров.

Условная клауза (4) является лингвистическим утверждением и рассматривается как истинное. Это означает, что интенсивность всей клаузы $\mathscr{P}$ есть

$$\mathbf{P}_{\langle x,y\rangle} := \{1/((\mathrm{A}_x[\mathrm{t}] \wedge \mathrm{B}_y[\mathrm{s}]) \Rightarrow \mathrm{R}_{x,y}[\mathrm{t,s}]) | t \in M_{l_1}, s \in M_{l_2}\}$$

при том, что <*существительное*>$_1$ ставится в соответствие переменной x, а <*существительное*>$_2$ - переменной y.

Определение 5. Пусть $\mathscr{A}_j$, $\mathscr{B}_j$ 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

$$\mathcal{LD}^I = \{\mathscr{R}^I_1, \ldots, \mathscr{R}^I_m\},$$

где $\mathscr{R}^I_j$ =ЕСЛИ $\mathscr{A}_j$, ТО $\mathscr{B}_j$, j=1,…,m являются условными клаузами;

$$\text{и } \mathcal{LD}^A = \{\mathscr{R}^I_1, \ldots, \mathscr{R}^I_m\},$$

где $\mathscr{R}^A_j = \mathscr{A}_j$ И $\mathscr{B}_j$, $j=1,\ldots,m$ 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, которое, как было сказано выше, дает возможность формировать кластеры с использованием синтагм. Второй метод основывается на дополнительном предположении и работает с лингвистическим описанием $\mathcal{LD}^A$, состоящем из конъюнкций лингвистических предикатов. В данной работе поведение описания $\mathcal{LD}^A$ рассмотрено не будет, а решение задачи кластеризации будет дано именно с помощью описания $\mathcal{LD}^I$.

Теорема 1. Пусть $\mathcal{LD}^I$ является простым лингвистическим описанием, состоящим из m правил и $\mathscr{P}_j$, $j=1,\ldots,m$ являются истинными условными клаузами вида (4). Предположим, что <*существительное*>$_1$ ставится в соответствие переменной x, а <*существительное*>$_2$ - переменной y, и пусть $\mathscr{A}'$ является простым оценочным предикатом с интенсивностью $\mathbf{A}'_{\langle x\rangle}$. Тогда существует теория в FLb

$$\mathcal{T} = \{\mathcal{LD}^I, \mathscr{P}_j, \mathscr{A}' \mid j=1,\ldots,m\}$$

такая, что можно вывести следующее:

а) можно получить вывод $\mathscr{B}'$ в $\mathcal{T}$ с интенсивностью

$$\mathbf{B}'_{\langle y\rangle} = \left\{b'_s = \vee_{t\in M_{l1}}\left(a'_t \wedge \vee_{j=1}^m \left(a_{j,t} \wedge b_{j,s}\right)\right) / B'_y[s] \mid s \in M_{l2}\right\}$$

где $B'(y) := (\exists x)(A'(x) \wedge R(x,y))$ и все $b'_s = \vee_{t\in M_{l1}}\left(a'_t \wedge \vee_{j=1}^m \left(a_{j,t} \wedge b_{j,s}\right)\right)$ являются максимальными;

б) лингвистическое высказывание

$$\text{ЕСЛИ } \mathcal{LD}^I, \text{ ТО } \mathscr{R} \qquad (5)$$

верно в $\mathcal{T}$ и порождает кластеризацию между объектами $\tilde{X} = \{\tilde{x}_1,\ldots,\tilde{x}_m\}$ и классами $\tilde{K}_1,\ldots,\tilde{K}_l$.

Доказательство:

а) Пусть $A_1(x),\ldots,A_m(x)$ являются независимыми формулами, x переменная сорта l_1, $B_1(y),\ldots,B_m(y)$ являются независимыми формулами, y переменная сорта l_2, $R(x,y)$ является атомарной формулой. Кроме того, пусть $A'_1(x)$ является либо независимой формулой по отношению к $A_1(x),\ldots,A_m(x)$, либо является одной из этих формул. Пусть нечеткая теория T в FLn имеет вид

$$T = \{a_{j,t} \wedge b_{j,s} / A_{j,x}[t] \wedge B_{j,y}[s],\ 1/(A_{j,x}[t] \wedge B_{j,y}[s] \Rightarrow$$

$$\Rightarrow R_{x,y}[t,s]),\ a'_t / A'_x[t] \mid t \in M_{l_1}, s \in M_{l_2},\ j=1,\ldots,m\} \quad (6)$$

По условию теоремы

$$B'(y) := (\exists x)(A'(x) \wedge R(x,y)).$$

Тогда, если

$T \vdash_{a'_t} A'_x[t] \mid t \in M_{l_1}$, то $T \vdash_{b'_s} B'_y[s] \mid s \in M_{l_2}$,

где $b'_s = \vee_{t\in M_{l1}}\left(a'_t \wedge \vee_{j=1}^m \left(a_{j,t} \wedge b_{j,s}\right)\right)$.

Так как $\mathcal{LD}^I$ является простым лингвистическим описанием, а формула $R(x,y)$ с интенсивностью $\mathbf{R}_{\langle x,y\rangle}=\{r_{ts}/\mathrm{R}_{x,y}[t,s]|t\in M_{l_1}, s\in M_{l_2}\}$ является атомарной, мы можем построить нечеткую теорию T вида (6), присоединенную к $\mathcal{T}$.

б) Используя тавтологию $\mathrm{T}\vdash(A\Rightarrow C)\Rightarrow((B\Rightarrow C)\Rightarrow(A\vee B)\Rightarrow C)$ и правило modus ponens, получаем $\mathrm{T}\vdash\vee_{i=1}^{m}(A_x[t]\wedge B_y[s])\Rightarrow R_{x,y}[t,s]$ для всех $t\in M_{l1}, s\in M_{l2}$. Это означает, что лингвистическое описание имеет интенсивность

$$\{1/\ \vee_{i=1}^{m}(A_x[t]\wedge B_y[s])\Rightarrow R_{x,y}[t,s]|\ t\in M_{l1}, s\in M_{l2}\}$$

Видно, что формула (4) влечет определение лингвистического описания вида (5) отношением $\mathscr{R}$ всякий раз, когда связь между некоторыми объектами определяется конъюнкцией оценочных высказываний. Но имплицирование в обратном направлении не дает нужного результата. Поэтому определим $\mathscr{R}$ иначе, чем это было показано выше, а именно, используя условные клаузы вида

$$\mathscr{P}:=\text{ЕСЛИ } \mathscr{R} \text{ и } <\textit{существительное}>_1 \text{ есть } \mathscr{A},\\ \text{ТО } <\textit{существительное}>_2 \text{ есть } \mathscr{B} \qquad (7)$$

Теорема 2. Пусть $\mathscr{P}_j$, $j=1,...,m$ являются истинными условными клаузами вида (7), определяющими теорию FLb

$$\mathcal{T}=\{\mathcal{LD}^I, \mathscr{P}_j|\ j=1,...,m\}.$$

Тогда существует такое лингвистическое высказывание, что условная клауза

$$\text{ЕСЛИ } \mathscr{R}, \text{ ТО } \mathcal{LD}^I \qquad (8)$$

верна в $\mathcal{T}$.

Доказательство: Пусть нечеткая теория T, присоединенная к $\mathcal{T}$, имеет вид:

$$T=\{1/((R_{x,y}[t,s]\wedge A_{j,x}[t])\Rightarrow \\ \Rightarrow B_{j,y}[s])|\ j=1,...,m, t\in M_{l1}, s\in M_{l2}\}.$$

Тогда, используя тавтологии
$\mathrm{T}\vdash(A\Rightarrow C)\Rightarrow((B\Rightarrow C)\Rightarrow(A\vee B)\Rightarrow C)$
и $\mathrm{T}\vdash A\Rightarrow(B\Rightarrow(A\,\&\,B))$, заключаем, что

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

для всех $t\in M_{l1}, s\in M_{l2}$. Следовательно, мы получаем мультиформулу:

$$\mathbf{P}_{\langle x,y\rangle}:=\\ =\{1/R_{x,y}[t,s]\Rightarrow\wedge_{j=1}^{m}(A_{j,x}[t]\Rightarrow B_{j,y}[s])|\ t\in M_{l1}, s\in M_{l2}\},$$

которая является интенсивностью истинной условной клаузы (8) в $\mathcal{T}$.

Из этой теоремы следует, что лингвистическое описание $\mathcal{LD}^I$ делает возможным определение отношения $\mathscr{R}$: всякий раз, когда некоторые объекты состоят в рассматриваемом отношении. Они могут быть определены, используя выводы оценочных предикатов. Описание $\mathcal{LD}^I$ используется, когда необходимо вывести утверждения из некоторых фактов [4, 8].

Таким образом, с помощью вышеназванных лингвистических описаний можно выделить информативные сочетания признаков, представленных синтагмами.

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. Вардзелашвили Ж. «Возможные миры» текстуального пространства // Тр. Санкт-Петербургского гос. ун., № VII. 2003, с. 37-45.
2. Журавлев Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики, №33. 1978, с. 39-40.
3. Филиппович Ю.Н. Метафоры информационных технологий: рабочие материалы исследования. М.: Изд-во МГУП. 2002. 288 с.
4. Bezdek J. Fuzzy mathematics in pattern classification / Ph.D. dissertation, Cornell University, Ithaca, New York.
5. Gallin D. Intensional and higher-order modal logic. Amsterdam, North Holand Publishing Company, 1975

6. Hajek P. Mathematics of fuzzy logic. Dordrecht, Netherlands: Kluwer, 1998.
7. Mamdani E., Assilian S. An experiment in linguistic synthesis with a fuzzy logic controller // International Journal of Man-Machine Studies, vol. 7, № 1, 1975, pp. 1-13.
8. Kerimov A.K., Rzayeva U. Sh. Fuzzy interpolation of partial functions of membership, characterizing affinity of objects to each other and objects to the class / Proceedings ICAFS', Prague, 2010, pp. 229-234.
9. Novak V., Perfilieva I. On model theory in Fuzzy Logic in broader sense / Proceedings FUZZ-IEEE'97, Barcelona, 1997, pp. 693-698.
10. Novak V., Perfilieva I., Mockor J. Mathematical principles of fuzzy logic. Kluwer Academic Publisher, 1999.
11. Zadeh L.A. The concept of a linguistic variable and its application to approximate reasoning I, II, III. // Inf. Sci., № 8, pp. 199-257, 301-357; № 9, 1975, pp. 43-80.
12. Zadeh L.A. Toward a Theory of Fuzzy Systems// Aspect Network and System Theory. – New York: Rinehart and Winston, 1971, pp. 209-245.
13. Zagoruiko N.G., Borisova I.A., Dyubanov V.V., and Kutnenko O.A. Methods of Recognition Based on the Function of Rival Similarity // Pattern Recognition and Image Analysis, 2008, Vol. 18, №1, pp.1-6.

Керимов Адалят Керим оглы. Профессор кафедры «Информационная экономика и технологии» Азербайджанского государственного экономического университета. Окончил Азербайджанскую государственную нефтяную академию (АГНА) в 1976 году. Доктор физико-математических наук, профессор. Автор 50 печатных работ. Область научных интересов: распознавание образов и классификация, дискретная математика, оптимизация, оптимальное управление, нечеткая логика, теория принятия решений, искусственный интеллект, интеллектуальные системы.
E-mail: k_adalat@aseu.az.

Рзаева Ульвия Шахин гызы. Преподаватель кафедры «Информационная экономика и технологии» Азербайджанского государственного экономического университета. Окончила Бакинский государственный университет (БГУ) в 1995 году. Автор 7 печатных работ. Область научных интересов: распознавание образов и классификация, дискретная математика, искусственный интеллект, интеллектуальные системы. E-mail: r.ulviyye@aseu.az.