O. Marchenko

DEVELOPMENT OF LEXICAL-SYNTACTIC MODEL OF NATURAL LANGUAGE BY USING MODERN METHODS OF LARGE TEXT CORPORA PROCESSING

Paper considers the algorithm for building the model of lexical and syntactic structural relations of natural language based on frequency-syntactic analysis of sentences from large text corpora. During the analysis the obtained data are accumulated in large arrays. To record syntactic structures of unlimited complexity, depth and length the natural language syntactic structures control spaces were used. The obtained arrays are huge and sparse. For effective and compact storage of the data the arrays were transformed by using the methods of non-negative matrix and tensor factorization.

Keywords: natural language text processing, syntactic structures control spaces, non-negative tensor factorization.

Матеріал надійшов 31.05.2013

УДК 510.6

Касьянюк В. С., Малютенко Л. М.

КЛАСТЕРИЗАЦІЯ ДАНИХ З ВИКОРИСТАННЯМ ТЕОРІЇ МОЖЛИВОСТЕЙ

У роботі розглянуто підхід до кластеризації, що базується на теорії можливостей. Такий підхід дозволяє врахувати неоднозначність вибору параметрів алгоритму кластеризації. На цій основі запропоновано методи оцінювання можливості та необхідності низки тверджень щодо точок та нечітких кластерів.

Ключові слова: множина даних, кластеризація, теорія можливостей.

Задача ефективного розбиття множини даних на класи еквівалентності за допомогою алгоритмів кластеризації має велике прикладне значення. При розв'язанні такої задачі часто виникає проблема адекватного вибору параметрів алгоритму кластеризації. При невдалому виборі цих параметрів розбиття не буде відповідати вимогам експерименту, проте навіть при вдалому їх виборі часто виникають питання щодо коректного віднесення певних підмножин множини даних до того чи іншого кластера. Пропонований в роботі підхід до кластеризації, що базується на теорії можливостей, дозволяє врахувати неоднозначність та нечіткість вибору параметрів алгоритму кластеризації.

1. Нечіткі розбиття. Кластери

Нехай $(X, 2^X, P)$ – деякий можливісний простір [1], D – певна скінчена множина даних. Позначимо як $\mathcal{D}$, де $\mathcal{D} \subseteq 2^D$, множину допустимих розбиттів D на класи еквівалентності, тобто кожне $\mathcal{A} \in \mathcal{D}$ тут таке, що $\bigcup_{i=1}^{n} A_i = D$, $A_i \in \mathcal{A}$, $|\mathcal{A}| = n$, $A_i \cap A_j = \varnothing$ для всіх $i,j \in \{1,\ldots,n\} : 1 \le i < j \le n$.

Нечітким розбиттям D на класи еквівалентності (скор. нечітким розбиттям) назвемо нечітку множину на D [2] – відображення вигляду $R : X \to \mathcal{D}$.

Водночас можна дати еквівалентне визначення нечіткого розбиття через канонічний нечіткий елемент [3] можливісного простору.

Нехай ξ – канонічний нечіткий елемент можливісного простору $(X, 2^X, P)$ з розподілом $\varphi^{\xi}(x)$. Нехай $\alpha : X \to \mathcal{D}$ – деякий параметризований алгоритм (алгоритм кластеризації), що будує певне розбиття D на класи еквівалентності залежно від значення вектора параметрів $x \in X$.

Нечітким розбиттям множини даних D назвемо нечітку множину $\alpha(\xi)$.

Зазначене визначення нечіткого розбиття дозволяє інтерпретувати X як множину допустимих значень вектора параметрів x алгоритму α, а $\varphi^{\xi}(x)$ – як розподіл вектора параметрів. Зрозуміло, що при різних значеннях x можна отримати різні варіанти розбиття $\alpha(x)$ із різною кількістю множин в розбитті.

Під *кластером* будемо розуміти відображення $C : X \to D$.

Отже, кластер трактуємо як певну нечітку підмножину D.

З практичного погляду важливо отримати відповідь на питання належності точки до того чи іншого кластера, тобто належності точки до деякої множини – елемента отриманого розбиття. Однозначно визначити, чи належить точка до даного кластера, ми не можемо, адже кластер є нечіткою множиною. Проте ми можемо визначити *можливість* належності точки до певної підмножини D.

Нехай C – множина всіх можливих кластерів, що утворюються алгоритмом α при всіх допустимих значеннях вектора параметрів x: $C = \bigcup_{x \in X} \alpha(x)$.

Теорема 1. Можливість та необхідність потрапляння точки d у кластер C відповідно рівні:

$$P(d \in C) = \sup_{\{x \mid x \in X, d \in \alpha(x) \cap C\}} \varphi^{\xi}(x) \chi_C(d); \tag{1}$$

$$N(d \in C) = \inf_{\{x \mid x \in X, d \notin \alpha(x)\}} \varphi^{\xi}(x) \chi_{D \setminus C}(d). \tag{2}$$

Доведення. Можливість події $d \in C$, де $C \in C$, рівна

$$P(d \in C) = P\left(\bigcup_{x \in X} \{d \in \alpha(x) \cap C\}\right) = \sup_{\{x \mid x \in X, d \in \alpha(x) \cap C\}} \varphi^{\xi}(x) = \sup_{\{x \mid x \in X, d \in \alpha(x)\}} \varphi^{\xi}(x) \chi_C(d).$$

Тут $\chi_C(d) = \begin{cases} 0, \text{ якщо } d \in C, \\ 1, \text{ якщо } d \notin C, \end{cases}$ – характеристична функція множини C.

Аналогічно можна визначити необхідність $d \in C$, де $C \in C$:

$$N(d \in C) = N\left(\bigcup_{x \in X} \{d \in \alpha(x) \cap C\}\right) = \theta\left(P\left(D \setminus \bigcup_{x \in X} \{d \in \alpha(x) \cap C\}\right)\right) =$$

$$= \theta\left(\sup_{\{D \setminus \bigcup_{x \in X} \{d \in \alpha(x) \cap C\}\}} \varphi^{\xi}(x)\right) = \inf_{\{x \mid x \in X, d \notin \alpha(x) \cup C\}} \varphi^{\xi}(x) = \inf_{\{x \mid x \in X, d \notin \alpha(x)\}} \varphi^{\xi}(x) \chi_{D \setminus C}(d).$$

Тут $\theta : [0, 1] \to [0, 1]$ – довільна строго монотонно спадна бієкція, причому $\theta(0) = 1$, $\theta(1) = 0$.

Цей результат добре узгоджується з природним трактуванням можливості та необхідності: подія A вважається необхідною, якщо протилежна до неї подія $D \setminus A$ є неможливою. Отже, найбільші значення необхідності потрапляння до кластера C мають такі точки, можливість потрапляння яких до $D \setminus C$ найменша.

Наведене вище визначення кластера має істотний недолік – не враховується поведінка кожного конкретного розбиття при зміні вектора параметрів $x \in X$. Для врахування цієї обставини введемо поняття нечіткого кластера.

2. Нечіткі кластери

Нечітким кластером називають [4] довільну нечітку множину з системи відображень $C_i : X \to D$, де $i \in \{1, \ldots, n\}$, таку, що для всіх $x \in X$ маємо:

1) $C_i(x) \in \alpha(x)$, де $i \in \{1, \ldots, n\}$;

2) $\bigcup_{x \in X} C_i(x) = D$;

3) $C_i(x) \cap C_j(x) = \varnothing \in$ для всіх $i, j \in \{1, \ldots, n\} : 1 \le i < j \le n$.

Систему відображень $\{C_i\}_{i\in\{1,\dots,n\}}$ назвемо нечітким розбиттям.

З іншого боку, нечіткий кластер C_i можна розглядати як нечітку величину $C_i(\xi)$ на множині $\mathcal{D}$, яка породжується канонічним нечітким елементом ξ простору $(X, 2^X, P)$. При цьому розподіл C_i – відображення $\varphi^{C_i}: C \to [0,1]$, – має вигляд $\varphi^{C_i}(C) = \sup\limits_{\{x|C_i(x)=C\}} \varphi^{\xi}(x)$.

Розглянемо подію потрапляння точки d до нечіткого кластера C_i та визначимо її основні характеристики: можливість та необхідність.

Теорема 2. Можливість потрапляння точки d в нечіткий кластер C_i дорівнює

$$P(d \in C_i(\xi)) = \sup_{\{C|C=C_i(x),x\in X\}} \sup_{\{x|x\in X,\ d\in\alpha(x)\}} \varphi^{\xi}(x)\chi_C(d); \tag{3}$$

необхідність потрапляння d в нечіткий кластер C_i, за умови N-незалежності подій $d\in C_i(x_1)$ та $d\in C_i(x_2)$, дорівнює

$$N(d \in C_i(\xi)) = \sup_{\{C|C=C_i(x),x\in X\}} \inf_{\{x|x\in X,\ d\notin\alpha(x)\}} \varphi^{\xi}(x)\chi_{D\setminus C}(d). \tag{4}$$

Доведення. Подію $d\in C_i(\xi)$ потрапляння точки d в нечіткий кластер C_i подамо як об'єднання подій: $d \in C_i(\xi) = \bigcup\limits_{x\in X}\{d \in C_i(x)\}$.

Враховуючи співвідношення (1) теореми 1, можливість події $d\in C_i(\xi)$ можна визначити так: $P(d \in C_i(\xi)) = P(\bigcup\limits_{x\in X}\{d \in C_i(x)\}) = \sup\limits_{\{C|C=C_i(x),x\in X\}} P(d \in C) = \sup\limits_{\{C|C=C_i(x),x\in X\}} \sup\limits_{\{x|x\in X,d\in\alpha(x)\}} \varphi^{\xi}(x)\chi_C(d)$.

Враховуючи співвідношення (2) теореми 1, отримаємо значення необхідності події $d\in C_i(\xi)$: $N(d \in C_i(\xi)) = N(\bigcup\limits_{x\in X}\{d \in C_i(x)\}) \geq \sup\limits_{\{C|C=C_i(x),x\in X\}} N(d \in C) = \sup\limits_{\{C|C=C_i(x),x\in X\}} \inf\limits_{\{x|x\in X,d\notin\alpha(x)\}} \varphi^{\xi}(x)\chi_{D\setminus C}(d)$.

У випадку, коли для довільних $x_1, x_2 \in X$ події $d\in C_i(x_1)$ та $d\in C_i(x_2)$ є N-незалежними [3], наведена вище нерівність набуває вигляду строгої рівності.

Розглянемо подію $\delta \cap C_i(\xi) \neq \varnothing$ перетину нечіткого кластера з деякою множиною D, де $\delta \subset D$, та виразимо можливість і необхідність цієї події. Ця подія може бути виражена як об'єднання подій: $\delta \cap C_i(\xi) \neq \varnothing = \bigcup\limits_{d\in\delta}\{d \in C_i(x)\}$.

Тому можливість події $\delta \cap C_i(\xi) \neq \varnothing$ може бути подана так:

$$P(\delta \cap C_i(\xi) \neq \varnothing) = \sup_{d\in\delta} P(d \in C_i(\xi)) = \sup_{d\in\delta} \sup_{\{C|C=C_i(x),x\in X\}} \inf_{\{x|x\in X,d\in\alpha(x)\}} \varphi^{\xi}(x)\chi_C(d).$$

Необхідність цієї події, за дотримання умови теореми 2, подається так:

$$N(\delta \cap C_i(\xi) \neq \varnothing) = \sup_{d\in\delta} N(d \in C_i(\xi)) = \sup_{d\in\delta} \sup_{\{C|C=C_i(x),x\in X\}} \inf_{\{x|x\in X,d\notin\alpha(x)\}} \varphi^{\xi}(x)\chi_{D\setminus C}(d).$$

3. Нечітке відношення еквівалентності

Алгоритм $\alpha(x)$ породжує на D відношення еквівалентності. Виникає питання, чи потраплять дві різні точки d_1 і d_2 до одного кластера, тобто чи будуть вони еквівалентними відносно отриманого розбиття. У випадку нечітких кластерів однозначної відповіді не існує, проте можна визначити можливість та необхідність події $d_1 = d_2$ за допомогою теорем 1 і 2.

Теорема 3. Нехай події $d_1\in C_i(\xi))$ та $d_2\in C_i(\xi))$ є P- та N-незалежними [3], тобто повністю незалежними. Тоді значення можливості та необхідності події $d_1 = d_2$ визначаються так:

$$P(d_1 = d_2) = \sup_{i\in\{1,\dots,n\}} \min(P(d_1 \in C_i(\xi)), P(d_2 \in C_i(\xi))); \tag{5}$$

$$N(d_1 = d_2) = \sup_{i\in\{1,\dots,n\}} \min(N(d_1 \in C_i(\xi)), N(d_2 \in C_i(\xi))). \tag{6}$$

Доведення. Виразимо подію $d_1 = d_2$ через ті події, можливість і необхідність яких ми вже вміємо визначати: $d_1 = d_2 \triangleq \bigcup\limits_{i=1}^{n}(\{d_1 \in C_i(\xi)\} \cap \{d_2 \in C_i(\xi)\})$.

Маємо такі співвідношення для значення можливості події $d_1 = d_2$:

$$P(d_1 = d_2) = P(\bigcup_{i=1}^{n}(\{d_1 \in C_i(\xi)\} \cap \{d_2 \in C_i(\xi)\})) =$$

$$= \sup_{i\in\{1,\ldots,n\}} P(\{d_1 \in C_i(\xi)\} \cap \{d_2 \in C_i(\xi)\}) \le \sup_{i\in\{1,\ldots,n\}} \min(P(d_1 \in C_i(\xi)), P(d_2 \in C_i(\xi))).$$

За умови *P*-незалежності подій $d_1 \in C_i(\xi))$ та $d_2 \in C_i(\xi))$ це співвідношення набуває вигляду строгої рівності.

Маємо співвідношення для отримання значення необхідності події $d_1 = d_2$:

$$N(d_1 = d_2) = N(\bigcup_{i=1}^{n}(\{d_1 \in C_i(\xi)\} \cap \{d_2 \in C_i(\xi)\})) \ge$$

$$\ge \sup_{i\in\{1,\ldots,n\}} N(\{d_1 \in C_i(\xi)\} \cap \{d_2 \in C_i(\xi)\}) = \sup_{i\in\{1,\ldots,n\}} \min(N(d_1 \in C_i(\xi)), N(d_2 \in C_i(\xi))).$$

За умови *N*-незалежності подій $d_1 \in C_i(\xi))$ та $d_2 \in C_i(\xi))$ це співвідношення набуває вигляду строгої рівності.

4. Оцінювання належності до кластера точок та пар точок

Розглянемо задачу вибору таких точок заданої підмножини $D' \subset D$, які трактуємо як найбільш гарантовані представники деякого кластера C_i.

Міри можливості та необхідності події $d \in C_i(\xi)$ (див. теореми 1 і 2) можуть бути використані для отримання відповіді на запитання, яка з точок $d_1 \in D$ чи $d_2 \in D$ найвірогідніше потрапить до кластера C_i.

Міра можливості (3) породжує на *D* відношення часткового порядку $\prec$:

$d_1 \prec d_2 \Leftrightarrow P(d_1 \in C_i(\xi)) \le P(d_2 \in C_i(\xi))$.

Міра необхідності (4) породжує на *D* відношення часткового порядку $\prec'$:

$d_1 \prec' d_2 \Leftrightarrow N(d_1 \in C_i(\xi)) \le N(d_2 \in C_i(\xi))$.

Визначимо *P*-оптимальну (оптимальну за можливістю) та *N*-оптимальну (оптимальну за необхідністю) стратегії вибору точок таким чином:

$$D_P = \sup_{\prec} D'; \quad D_N = \sup_{\prec'} D'.$$

Тобто, множину D_P можна отримати як розв'язок задачі оптимізації $P(d \in C_i(\xi)) \to \max_{d\in D'}$.

Множину D_N при дотриманні умов теореми 2 можна отримати як розв'язок задачі оптимізації $N(d \in C_i(\xi)) \to \max_{d\in D'}$.

Далі природно розглянути двокритеріальну парето-оптимальну стратегію вибору точок, в якій будуть враховані як можливість, так і необхідність. Для її визначення введемо ще одне відношення часткового порядку на *D*:

$d_1 \ll d_2 \Leftrightarrow d_1 \prec d_2$ та $d_1 \prec' d_2$.

Тоді множина парето-оптимальних розв'язків матиме такий вигляд:

$D_{PN} = \{d \in D' \mid$ не існує $b \in D'$ таких, що $d \ll b\}$.

Розглянемо тепер задачу вибору таких пар точок заданої підмножини $D' \subset D$, які можна трактувати як найбільш гарантовано рівні, тобто таких пар, які потраплять до одного кластера. Для отримання відповіді на питання, яка з пар точок (d_1, d_2) чи (b_1, b_2) найвірогідніше потрапить до одного кластера, використаємо міри можливості та необхідності події $d_1 = d_2$.

Міра можливості (5) та міра необхідності (6) породжують на $D \times D$ відношення часткового порядку $\prec$ та $\prec'$:

$(d_1, d_2) \prec (b_1, b_2) \Leftrightarrow P(d_1 = d_2) \le P(b_1 = b_2)$.

$(d_1, d_2) \prec' (b_1, b_2) \Leftrightarrow N(d_1 = d_2) \le N(b_1 = b_2)$.

Визначимо *P*-оптимальну (за можливістю) та *N*-оптимальну (за необхідністю) стратегії вибору пар точок:

$$D_P^2 = \sup_{\prec} D' \times D'; \quad D_N^2 = \sup_{\prec'} D' \times D'.$$

Отже, множину D_P^2 можна отримати як розв'язок задачі оптимізації $P(d_1 = d_2) \to \max\limits_{(d_1,d_2)\in D'\times D'}$,

а множину D_N^2 при виконанні умов теореми 2 – як розв'язок задачі оптимізації $N(d_1 = d_2) \to \max\limits_{(d_1,d_2)\in D'\times D'}$.

Визначимо двокритеріальну парето-оптимальну стратегію вибору пар точок. Така стратегія враховує як можливість, так і необхідність.

Для цього введемо таке відношення часткового порядку на $D \times D$:

$(d_1, d_2) \ll (b_1, b_2) \Leftrightarrow (d_1, d_2) \prec (b_1, b_2)$ та $(d_1, d_2) \prec' (b_1, b_2)$.

Це дає таку множину парето-оптимальних розв'язків:

$D_{PN}^2 = \{(d_1, d_2) \in D' \times D' \mid$ не існує $(b_1, b_2) \in D' \times D'$ таких: $(d_1, d_2) \ll (b_1, b_2)\}$.

Висновки

У роботі запропоновано базований на апараті теорії можливостей підхід до розв'язку задачі кластеризації у випадку, коли параметри алгоритму кластеризації подаються як нечітка величина або вектор нечітких величин. Побудована математична модель опирається на поняття нечіткого кластера та нечіткого розбиття. На цій основі запропоновано методи оцінювання базових теоретико-можливісних характеристик – можливості та необхідності, – для тверджень щодо належності точки до нечіткого кластера, непорожнього перетину нечіткого кластера з певною множиною та еквівалентності двох точок відносно нечіткого розбиття. Розглянуто задачі вибору в певному розумінні найкращих точок та пар еквівалентних точок заданої множини даних.

Список літератури

1. Пытьев Ю. П. Основы теории возможностей. Методы оптимального оценивания и принятия решений / Ю. П. Пытьев // Вестник Московского ун-та. Серия 3 : Физика. Астрономия. – 1997. – № 3, № 4.
2. Пытьев Ю. П. Возможность. Элементы теории и применения / Ю. П. Пытьев. – М. : Едиториал УРСС, 2000. – 192 с.
3. Пытьев Ю. П. Основы теории возможностей. Методы оптимального оценивания и принятия решений. Нечёткие элементы, независимость, условные распределения / Ю. П. Пытьев // Вестник Московского ун-та. Серия 3 : Физика. Астрономия. – 1998. – № 2.
4. Касьянюк В. С. Застосування теорії можливостей в задачах кластеризації даних / В. С. Касьянюк, М. В. Польща // Вісник Київського ун-ту. Серія : фіз. – мат. науки. – 2007. – Вип. 2. – С. 155–159.

V. Kasyanuk, L. Malutenko

DATA CLUSTERING BASED ON POSSIBILITY THEORY

In this paper we consider a possibility theory approach to data clustering. Possibility theory allows taking into account an ambiguity of parameters' selection in a clustering algorithm. We propose methods of possibility and necessity evaluations of various assertions about points and fuzzy clusters.

Keywords: data set, clustering, possibility theory.

Матеріал надійшов 07.06.2013