УДК 519

І.В. Баклан

ІНТЕРВАЛЬНИЙ ПІДХІД ДО ПОБУДОВИ ЛІНГВІСТИЧНОЇ МОДЕЛІ

Анотація. В статті розглянуті питання побудови розбиття множини значень часового ряду на інтервали для побудови лінгвістичних моделей.

Ключові слова: лінгвістичне моделювання, інтервальна математика, інтервальний підхід, лінгвістизація.

Вступ

Процес побудови лінгвістичних моделей в тому числі включає застосування інтервального підходу для розбиття множини значень часового ряду. Для формування розбиття пропонується використання елементів інтервальної математики.

Постановка проблеми

Перетворення чисельних образів до символьного вигляду, який можна використовувати для вирішення певного кола складних задач, й є основною метою лінгвістичного моделювання.

Лінгвістичне моделювання — комплекс методів, методик та алгоритмів, які використовують процес перетворення числових масивів інформації до лінгвістичних послідовностей, на основі яких відновлюється формальна граматика.

Лінгвістична модель — побудована на основі лінгвістичного моделювання сукупність символьних (лінгвістичних) послідовностей за обраними параметрами лінгвістизації та відновлена на її основі формальна граматика.

Лінгвістичне моделювання треба розглядати як специфічний вид математичного моделювання для обробки даних у символьному (не чисельному) вигляді [1].

Головна ідея переходу від множини чисельних значень до певного символьного алфавіту є розбиття цієї множини на чисельні інтервали.

© Баклан І.В., 2013

Аналіз досліджень і публікацій

Першим науковцем, який звернувся до інтервальних обчислень прийнято вважати Архімеда, що використовував двосторонні наближення для обчислення числа «пі» [2].

Інтервальний підхід пов'язаний з основами інтервальної математики, перші принципи обчислення двосторонніх границь були викладені в роботах Сунаге Г. та Канторовича Л.В. [3,4].

Мета роботи

Метою даної статті є розгляд інтервального підходу для розбиття множини чисельних даних часового ряду у межах підготовки до відображення чисельних даних до певного символьного алфавіту. Такий процес надалі будемо називати інтервалізацією.

Деякі елементи формалізації інтервального підходу

Зрозуміло, що при найпростішій схемі лінгвістизації потужність відповідної абетки повинен бути не набагато меншою, ніж сама послідовність часового ряду.

Нехай X та Y - дві частково впорядкованих множини. Позначимо через B(X) та B(Y) множини усіх підмножин множин X та Y .

Множина, на якій задано відношення порядку, має назву впорядкованої , якщо відношення порядку визначено для будь-яких двох його елементів, а частково впорядкованою у протилежному випадку.

Частково впорядкована множина називається структурою, якщо будь яка її двоелементна підмножина має точну верхню та нижню межу, а повною структуру, якщо кожна її непуста підмножина має такі точні межі.

Кожну з множин X та Y будемо вважати умовно повними структурами та позначати $S(X)$ та $S(Y)$. Відношення порядку будемо позначати через $\leq$.

Якщо $a, b \in S(X)$ та $a \leq b$, то множину $I(a,b) = [a,b] = \{x \in X, a \leq x \leq b\}$, будемо називати інтервалом на $S(X)$. Множину усіх інтервалів на структурі $S(X)$ будемо позначати $\mathcal{I}_{S(X)}$. При цьому, якщо $X = R^1$ – множина дійсних чисел, то $\mathcal{I}_{R^1}$ – множина закритих

 ISSN 1562-9945

інтервалів на прямій дійсних чисел. В такому випадку $\mathcal{I}_{R^1}$ називають інтервальних числом.

З основ інтервальної математики нам відомі співвідношення $X \subseteq \mathcal{I}_{S(X)} \subseteq \mathbf{B}(\mathbf{X})$ та $Y \subseteq \mathcal{I}_{S(Y)} \subseteq \mathbf{B}(\mathbf{Y})$.

Якщо підмножина $A \subseteq \mathbf{X}$ - обмежена, то інтервал $I(A)$, який визначається за правилом $I(A) = \left[\inf_{S(X)} A, \sup_{S(X)} A\right]$, будемо називати поданням зовнішнього інтервалу множини A.

Якщо наша множина $\mathbf{X}$ утворює поле, то в $\mathcal{I}_{S(X)}$ можна ввести інтервальні арифметику:

$[a,b] * [c,d] = \{x * y | a \le x \le b, c \le y \le d\}$, де $*\in \{+,-,\times,:\}$. Для кожної з приведених операцій маємо наступні співвідношення:

$$[a,b] + [c,d] = [a + c, b + d],$$

$$[a,b] - [c,d] = [a - c, b - d],$$

$$[a,b] \times [c,d] = [min(a \times c, a \times d, b \times c, b \times d), max(a \times c, a \times d, b \times c, b \times d)]$$

$$[a,b] : [c,d] = [a,b] \times \left[\frac{1}{d}, \frac{1}{c}\right], \mathbf{0} \notin [c,d].$$

Крім цих операцій для вирішення наших задач буде мати зиск операція «конкатенація», яка визначається для суміжних інтервалів. Суміжними інтервалами будемо називати інтервали вигляду

$$[a,b], [c,d] \subseteq X,\ b = c,\ a < b,\ c < d.$$

Операцією конкатенації двох суміжних інтервалів будемо називати інтервал

$$[a,b] \oplus [c,d] = [a,d]\{x | a \le x \le b \vee c \le x \le d\},\ a < d,\ b = c.$$

На множині усіх інтервалів також можна ввести відношення порядку та відношення тотожності. Якщо маємо інтервали $I[a,b]$ та $I[c,d]$, то між ними є відношення порядку $I[a,b] < I[c,d]$, якщо $b \le c$. Інтервали $I[a,b]$ та $I[c,d]$ будуть тотожними $I[a,b] = I[c,d]$, якщо $a = c$, $b = d$.

Нехай $X = \{x_i | x_i \geq 0, x_i \leq m\}_{i=1,\dots,M}$, $x_1 \leq x_2 \leq \dots \leq x_i \leq \dots \leq x_M$. Потужність цієї множини $dim\{X\} = M$. Назвемо інтервалізацією рівня n множини X представлення у вигляді $X = \bigcup_{i=1}^{n} I[a_i, b_i]$, де $I[a_1, b_1] \prec I[a_2, b_2] \prec \dots \prec I[a_n, b_n]$, при цьому

$$dim\{X\} = \sum_{i=1}^{n} dim\{I[a_i, b_i]\}.$$

Нагадаємо деякі співвідношення, які вводять для інтервалів [4].

Шириною інтервалу $I[a_i, b_i]$ будемо називати величину $\omega(a_i, b_i) = b_i - a_i$.

Серединою інтервалу $I[a_i, b_i]$ є полусума меж інтервалу $m(a_i, b_i) = \frac{b_i - a_i}{2}$. Медіана інтервалу розраховується за формулою $med(a_i, b_i) = a_i + m(a_i, b_i)$.

Абсолютна величина інтервалу $I[a_i, b_i]$ знаходиться за відношенням

$$|I[a, b]| = max\{|a_i|, |b_i|\}.$$

Крім того, $\mu(I[a, b]) = min\{|a_i|, |b_i|\}$, $\sigma(I[a, b]) = \frac{|a_i| + |b_i|}{2}$.

Відстанню між інтервалами $I[a_i, b_i]$ та $I[a_j, b_j]$ називається величина

$$\rho(I[a_i, b_i], I[a_j, b_j]) = max\{|a_i - a_j|, |b_i - b_j|\}.$$

Виродженим інтервалом будемо називати інтервал із співпадаючими верхньою та нижньою межами, який за традицією обожнюють з дійсним числом.

Повертаючись до процедури інтервалізації, зауважимо, що в основному ми будемо розглядати інтервали, які не є виродженими. При цьому у найпростішому методу лінгвістизації можна було б звести усе до того, що усі значення часового ряду (або його різниць) є виродженими інтервалами та інших інтервалів немає.

Нас будуть цікавити певні випадки, які відображують наступні типи інтервалізації:

– коли інтервали рівнозначні;

 ISSN 1562-9945

– логарифмічні інтервали;

– коли інтервали рівно ймовірнісні;

– інтервали за певним розподілом ймовірностей (нормальним, бета-розподілом, Пуасона, Дирихле та ін.).

При рівнозначній інтервалізації N-того рівня множини X маємо:

$$\omega(a_1, b_1) = \omega(a_2, b_2) = \cdots = \omega(a_N, b_N).$$

При рівноймовірній (або рівночастотній) інтервалізації маємо

$$dim\{I[a_1, b_1]\} = dim\{I[a_2, b_2]\} = \cdots = dim\{I[a_i, b_i]\} = \cdots = dim\{I[a_N, b_N]\}$$

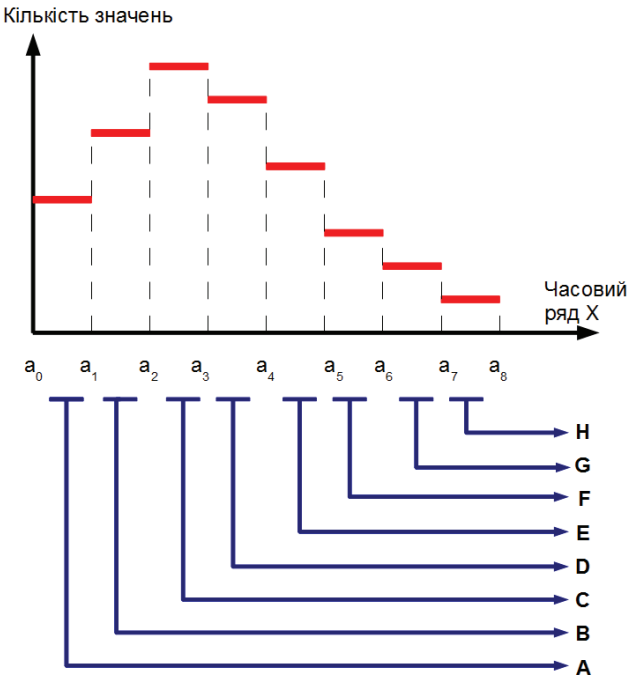

Рисунок 1 - Схема перетворення чисельних даних до символьного вигляду

На рисунку 1 представлена умовна схема лінгвістизації при рівнозначній інтервалізації. Інтервалізація дає множину інтервалів $I[a_0, a_1], I[a_1, a_2], \ldots, I[a_7, a_8]$. За віссю абсцис розташовані впорядковані значення часового ряду та інтервали. За віссю ординат ранжується кількість значень часового ряду, що потрапили до відповідного інтервалу. Кількість елементів, які потрапляють до певного інтервалу (значення часового ряду можуть повторюватися) будемо позначати через $D\{I[a_i, a_{i+1}]\}$. Кількість інтервалів – $M < N$. Легко бачити, що сума

$$\sum_{i=0}^{M-1} D\{I[a_i, a_{i+1}]\} = N.$$

Тепер ми можемо ввести поняття частотності інтервалу $I[a_i, a_{i+1}]$ на часовому ряду потужності N:

$$\nu_{i,i+1} = \nu[a_i, a_{i+1}] = \frac{D\{I[a_i, a_{i+1}]\}}{N}.$$

Легко довести, що $\sum_{i=0}^{M-1} \nu_{i,i+1} = 1$, а враховуючи, що для усіх i $\nu_{i,i+1} > \emptyset$ будемо мати аналогію аксіоматики теорії ймовірностей.

Зауважимо, що для рівноймовірнісного випадку будемо мати

$$\nu_{0,1} = \nu_{1,2} = \cdots = \nu_{i,i+1} = \cdots = \nu_{M-1,M} = \frac{1}{M}.$$

Висновки

Були розглянуті більш докладніше класичні методи в переломлюванні до маніпулювання інтервалами. Стаття буде присвячена заполученнями з сучасних досягнень математичної статистики інтервальних даних, тобто коли статистичні дані не числа, а інтервали, як у нашому випадку інтервалізації часових рядів.

Тут маємо найпростіший підхід щодо моделі групування даних, згідно якої деяке значення x замінюється на найближче з множини I. Але виникає проблема у тому випадку, коли відстані від x до двох найближчих елементів множини I десь рівні, то природним було би ввести рандомізації при виборі замінюю чого числа.

Таким чином ми звернулися до того випадку, коли наші дані замінюються інтервалами, де працює інтервальна математика, про арифметичні операції якої було докладно викладено у статті.

ЛІТЕРАТУРА

1. Баклан И.В. Лингвистическое моделирование: основы, методы, некоторые прикладные аспекты // Системные технологии. Региональный межвузовский сборник научных работ. - Выпуск 3 (74). - Днепропетровск, 2011. - с.10 — 19.

2. Moore R.E. Interval analysis. – Englewood Cliffs: Prentice Hall, 1966.

3. Sunaga T. Theory of an interval algebra and its application to numerical analysis // RAAG Memoirs. – 1958. – Vol. 2, Misc. II. – P. 547-564.

4. Канторович Л.В. О некоторых новых подходах к вычислительным методам и обработке наблюдений // Сибирский Математический Журнал. – 1962. – Т. 3, No. 5. – С. 701-709.