УДК 506+510

І.В. Баклан

ЛІНГВІСТИЧНЕ МОДЕЛЮВАННЯ: ОСНОВИ, МЕТОДИ, ДЕЯКІ ПРИКЛАДНІ АСПЕКТИ

Анотація. Розглядаються основи лінгвістичного моделювання – одного з нових напрямів нечисельного моделювання. Запропоновано новий підхід для представлення часових рядів, особливості лінгвістизації часових рядів, а також застосування лінгвістичного моделювання для вирішення задач прогнозування часових рядів, автентифікація користувача за його рухами мишкою, розпізнавання емоційного стану користувача та рання діагностика розладів руху. В основі лежить процес відновлення формальної граматики.

Ключові слова: структурний підхід, лінгвістичне моделювання, прогнозування часових рядів, автентифікація, розпізнавання образів

Постановка дослідження лінгвістичного моделювання

Головним завданням лінгвістичного моделювання є перетворення чисельних рядів, експериментальних даних, багатомірних даних до лінгвістичних послідовностей та відновлення за ними формальної граматики мови відповідного харатеру для вирішення наступного спектру проблем: аналіз та прогнозування часових рядів, розпізнавання образів різноманітної природи, автентифікація користувача за його рухами, розпізнавання емоційного стану оператора, діагностика хвороб опорно-рухового апарату операторів складних технічних систем на ранніх стадіях захворювання.

Лінгвістичне моделювання базується на трьох основних підходах: структурний підхід та математична лінгвістика, інтервальні обчислення та робастні методи, сучасні методи ймовірнісного моделювання [1,2,3].

В основі лінгвістичного моделювання лежить лема існування ізоморфізма відворення чисельних даних до лінгвістичних послідовностей, на основі яких може бути побудована мова. Як висновок існування унікальної мови, яка фактично уособлюється наборами чисельних даних.

 ISSN 1562-9945

Початок цієї роботи був покладений при створенні математичних основ автентифікації оператора (користувача) складною технічною системою. Автентифікація - шлях встановлення вірогідності інформації, пред'явленої користувачем у разі звернення його до системи та відкриття йому доступу, якщо він має на це право. Дамо загальну постановку завдання. Аналізується кінематика рухів оператора при керуванні складною технічною системою. При цьому стоїть необхідність автентифікації оператора – в цей момент керує складною технічною системою (СТС) оператор А чи не А. Якщо це не А, то хто керує СТС? Спектр складних технічних систем може бути від літака, космічного апарату до систем керування космічними станціями, атомними електростанціями, небезпечними військовими об'єктами, які можуть стати в руках терористів небезпекою для життя тисяч мирних мешканців. Орган керування – це технічний пристрій, який використовується для маніпулювання (керування) складною технічною системою (руль, штурвал, клавіатура, пульт та т. і.).

Слід зазначити, що на особливості динаміки поведінки оператора при керуванні СТС накладається вплив цілого ряду чинників, які можна класифікувати за такими напрямками, як постійні чинники та випадкові чинники. Під постійними (або фоновими) чинниками будемо мати на увазі чинники, які є постійною складовою впливу та не мають імовірнісний характер. У свою чергу постійні чинники розрізняють зовнішні та внутрішні. Конкретно, внутрішні чинники – це особисті властивості оператора СТС. Випадкові чинники мають чіткий імовірнісний характер. Враховуючи все вищезгадане можна сформулювати лему.

$O = \{o_1, o_2, \dots, o_i, \dots, o_M\}$- множину об'єктів назвемо *множиною автентифікації*. Нехай $R = \{r_1, r_2, \dots, r_k, \dots, r_N\}$– множина параметрів, які назвемо *параметрами автентифікації*, $N = dimR$. У деякі моменти t_jвідбуваються виміри значень параметрів r_k, набір яких позначимо через $Z_{ijk} = z_{1ijk}, z_{2ijk}, \dots, z_{lijk}, \dots, z_{Nijk}$, де z_{lijk}– це значення параметра r_kу момент t_j, $dimZ_{ijk} = N$. Таким чином можливо розглядати часові ряди побудовані на парах $\langle Z_{ijk}, t_j \rangle$. Множину всіх можливих часових рядів,

побудованих на множинах Rта Z, P^Nназвемо *простором автентифікації*.

Лема: Для кожного o_i, $o_i \in O$існує функція φ_iвизначена на множині P^Nтака, що $\varphi_i(z_{1ijk}, z_{2ijk}, \dots, z_{lijk}, \dots, z_{Nijk}) = 0$для всіх моментів часу t_j.

Звідси маємо, що існує відображення$\Phi: O \to P^N$.

Образ об'єкту o_iна просторі автентифікації $P^N: R_i = \Phi(o_i), R_i \subset P^N$.

Зворотне відображення $\Phi^{-1}: P^N \to O$, назвемо *функцією автентифікації*. Дійсно маючи таке відображення просто вирішуємо задачу автентифікації на деякій підмножині $R_j \subset P^N: O_j = \Phi^{-1}(R_j)$. З іншого боку, легко бачити, що $R_j = \bigcap_i \Phi(o_i)$.

Ідеальним варіантом побудови Φє побудова ізоморфізму (взаємно однозначної відповідності) на множині автентифікації Oта просторі автентифікації P^N, тобто якщо $R_i = \Phi(o_i)$, то $o_j = \Phi^{-1}(R_i)$.

Таким чином вирішення задачі автентифікації полягає у пошуку такого розбиття та побудові на ньому відповідного відображення. Для будь-якої структури простору автентифікації P^Nповинно було б існувати розбиття його на підмножини таким чином, що для будь-якої пари цих підмножин R_iта R_jє вірним, що$R_i \cap R_j = \emptyset$. При цьому $R_i = \Phi(o_i)$, $R_j = \Phi(o_j)$.

Слід відзначити , що такого розбиття може не існувати, тобто реально може мати місце факт, що $R_i \cap R_j \neq \emptyset$. Практичний момент будуть мати підмножини $R_i^0 = \Phi(o_i) \setminus \Phi(o_j)$та $R_j^0 = \Phi(o_j) \setminus \Phi(o_i)$є вірним, що $R_i^0 \cap R_j^0 = \emptyset$. При цьому $R_i^0 \subset R_i = \Phi(o_i)$, $R_i^0 \subset R_i = \Phi(o_i)$. R_i^0та R_j^0назвемо *зонами автентифікації*, а $\bar{R}_{i,j} = R_i \cap R_j$- *зоною зашумлення*. Якщо маємо $O = \{o_1, o_2, \dots, o_i, \dots, o_N\}$, $R_i = \Phi(o_i)$, $\forall o_i \in O$, то у загальному вигляді зона автентифікації об'єкта o_i- $R_i^0 = R_i \setminus \bigcup_{j \neq i} R_j$, а загальною зоною зашумлення буде підмножина $\bar{R} = \bigcup_{i=1, j=1}^{N} R_i \cap \bigcap_{j \neq i} R_j$.

 ISSN 1562-9945

Враховуючи той факт, що відображення може й не бути, поставимо завдання для більш м'якішої задачі, яка полягає в тому, що шукається відображення Φ^{-1} таке, що $\Phi^{-1}(R_i) = O_j$, $O_j = \{o_1^j, o_2^j, \ldots, o_L^j\}$, $L = dim O_j$. При цьому $R_j \subset \bigcup_{k=1}^{L} \Phi(o_k^j)$.

Для вирішення завдання побудови простору ідентифікації можливо використання багатьох відомих математичних формалізмів. Одним з них ми пропонуємо лінгвістичне моделювання.

Загальна схема лінгвістизації часових рядів

В поданому нижче тексті розглянуто один з можливих підходів до проблеми аналізу та моделювання часових рядів за допомогою структурних підходів.

Одним з методів прогнозу є структурний (синтаксичний) підхід. Він базується на деяких принципах розпізнавання образів, яка складається з трьох основних частин – блока попередньої обробки, блока опису об'єкта, блока синтаксичного аналізу.

Кожний попередньо опрацьований об'єкт сегментується на складові частини об'єкта на підставі наперед визначених операцій та операцій об'єднання; в свою чергу кожна складова частина об'єкту ідентифікується за допомогою заданого набору зразків. Тепер кожний об'єкт подається набором зразків із визначеними синтаксичними операціями, тобто правил їх поєднання . Наприклад, в термінах операції „конкатенації" кожний об'єкт подається рядком зразків, які утворюють ланцюжок.

Більш досконалі системи повинні були б мати здібністю визначати різні синтаксичні відношення усередині об'єкту.

На базі структурного підходу [2,5] розроблено алгоритм реалізації методу, який полягає в наступному: нехай ми маємо ряд спостережень $\{y(i)\}$, де $y(i)$, $i = \overline{1,N}$- це деякі значення, які отримано в ході спостереження з якимось кроком $\Delta t_i = const$, $i = \overline{1,N}$.

1. Порахуємо різниці $\Delta y(i) = y(i) - y(i+1), i = \overline{1,N}$ між сусідніми значеннями ряду.

2. Відсортуємо окремо додатні та від'ємні значення $\Delta y(i)$ за спаданням (чи за зростанням). Отримаємо дві послідовності $a(k)$ і $b(l), K + L = N - 1$.

3. Далі поставимо кожному члену послідовностей $a(k)$і $b(l)$, символи абетки a_iі $b_j i = \overline{1,K}$, $j = \overline{1,L}$відповідно.

4. Перепишемо послідовність $\Delta y(i)$символами a_iі b_jтакож будемо ставити між парами сусідніх символів $a_q b_p$символ **c**, а між парами сусідніх символів $a_m b_n$символ d, які будуть означати точки локальних екстремумів в послідовності $y(i)$, локальні максимуми і локальні мінімуми відповідно. Отримаємо послідовність e_i.

В послідовності e_iпроаналізуємо частоту існування пар символів $(e_i e_{i+1}) j = \overline{1, N-1}$і побудуємо таблицю ймовірностей виникнення символу $e_{i+1} P_{j+1}(e_{i+1} \mid e_i)$.

Далі в послідовності e_i аналізуємо частоту існування трійок $(e_{i-1} e_i e_{i+1}) j = \overline{1, N-2}$ будуємо таблицю ймовірностей $P_{j+1}(e_{i+1} \mid e_i e_{i-1})$. В загальному випадку аналізуємо частоту існування ланцюжків $P_{j+1}(e_{i+1} \mid e_{i-k} \dots e_i)$. Тобто обчислюємо ймовірність появи символу e_{i+1} за умови що попередні символи є $e_{i-k} \dots e_i$.

За допомогою обчислених ймовірностей можемо зробити імовірнісний прогноз виникнення символу e_{i+1}за умови що відомі ланцюжки попередніх символів.

Особливий інтерес визиває розрахунок рівноймовірнісних інтервалів.

Нехай часовий ряд має N елементів. Стоїть задача вибору оптимального розміру інтервалу, який фактично задається кількістю елементів m, які до нього потрапляють. Ймовірність потрапляння елементу часового ряду до інтервалу буде дорівнювати $\frac{m}{N}$. Кількість інтервалів буде дорівнювати $K = \frac{N}{m} + 2$. Легко перевірити, що $\sum_{i=1}^{K} P_i$.

Тепер будуємо правила перетворення часового рядо до лінгвістичного ланцюжка.

Відновлення граматики сигналу на основі лінгвістичних ланцюжків

Розглянемо проблему навчання в рамках формальних граматик.

 ISSN 1562-9945

Нагадаємо, що формальна граматика - це четвірка $G = (V_T, V_N, P, S)$, де V_T – алфавіт термінальних символів; V_N – алфавіт нетермінальних символів, причому $V_N \cap V_T = \emptyset; V_N \cup V_T = V$- алфавіт граматики G; P – множина правил підстановки; S– початковий символ, $S \in V_N$.

Нехай Γ_0 – деяка досліджувана мова. Введемо наступні визначення.

Лінгвістичною послідовністю (ланцюжком) $I(\Gamma_0)$мови Γ_0 будемо називати послідовність ланцюжків, кожний з яких належить однієї з множин $\{\alpha^+ \mid \alpha^+ \in \Gamma_0\}$або $\{\alpha^- \mid \alpha^- \in V_T^{\,0} \setminus \Gamma_0\}$ з вказанням того, до якої множини буде належати той або інший ланцюжок. Послідовність ланцюжків мови $\{\alpha^+ \mid \alpha^+ \in \Gamma_0\}$ будемо називати позитивною лінгвістичною послідовністю $I^+(\Gamma_0)$, а послідовність ланцюжків з додатка мови $\{\alpha^- \mid \alpha^- \in V_T^{\,0} \setminus \Gamma_0\}$ будемо називати від'ємною лінгвістичною послідовністю $I^-(\Gamma_0)$.

Лінгвістична послідовність $I(\Gamma_0)$ називається повною, якщо $I^+(\Gamma_0)$ має усі ланцюжки мови Γ_0, $I^-(\Gamma_0)$ має всі ланцюжки, що не належать мові Γ_0.

Граматика G узгоджена з граматикою G_0, якщо породжувані ними мови співпадають $\Gamma(G) = \Gamma(G_0)$.

Нехай $C = \{G_i\}$- клас граматики. Клас мов $\Gamma(C) = \{\Gamma(G) \mid G \in C\}$ називається ідентифікованим, якщо для будь-якої граматики $G \in C$та будь-якої лінгвістичної послідовності $I(\Gamma(G))$ існує деяке число N та алгоритм, який би, отримуючи на вході підпослідовність $I(\Gamma(G))$, що має у складі не меньш N ланцюжків, на виході дає граматику, узгоджену з граматикою G.

Поряд з поняттям лінгвістичної послідовності використовується поняття зразка (або вибірки) для мови Γ_0. Зразком S_t мови Γ_0 будемо називати послідовність ланцюжків $\{\alpha_i\}_{i=1}^{t}$, для кожного ланцюжка якої відомо, належить вона мові Γ_0 або його доповненню $V_T^{\,0} \setminus \Gamma_0$. Позитивним зразком будемо називати множину $S_t^- = S_t \cup (V_T^0 \setminus \Gamma_0)$.

Граматика G називається сумісною із зразком S_t, якщо вона породжує усі позитивні приклади цього зразка й не породжує жодного від'ємного прикладу.

Структурно повний зразок $S_t(\Gamma(G))$ мови $\Gamma(G)$- це зразок, маючий у складі такі ланцюжки, при побудові яких кожне правило підстановки граматики G було використане хоча по одному разу.

Існує два варіанти постановки завдання відновлення грамматик.

У першому формулюванні (на яку будемо посилатися як на проблему узгодження) передбачається, що є деяка щира граматика G_0, і потрібно по інформаційній послідовності побудувати таку граматику G, яка була б погоджена із граматикою G_0, тобто $\Gamma(G) = \Gamma(G_0)$.

У другому формулюванні (на яку будемо посилатися як на проблему граматичного висновку) уважається, що за зразком S_tнеобхідно побудувати таку граматику G, яка б породжувала всі ланцюжки позитивного зразка S_t^+(і, можливо, нескінченна безліч інших ланцюжків) і не породжувала ланцюжка негативного зразка S_t^-(і, можливо, нескінченна безліч інших ланцюжків), тобто була б сумісна із цим зразком.

Побудова гібридних моделей на основі лінгвістичного підходу

Основи гібридизації лінгвістичних та ймовірнісних моделей були викладені в ряді робіт [7,8]. Викладемо основні принципи гібридизації: принцип неоднорідності, принцип плюралізму, принцип системного аналізу неоднорідного завдання, принцип конструктора, принцип пріоритету знань, принцип поступовості, перший та другий принципи спадкування, принцип самоорганізації агрегованої моделі, принцип повноти, принцип зниження продуктивності агрегованої моделі.

Дані принципи ні в якому разі не претендують на істину останньої інстанції, або на строгий характер математичних або логічних правил. Ці принципи певним чином узагальнюють с погляду автора світову практику досліджень розробки гібридних інтелектуальних систем прийняття рішень. Використання цих принципів дозволить робити агреговані моделі там де це доцільно, такими які вони повинні бути, та функціонуючими таким чином, щоб не розчаровувати своїх творців.

Класичні гібридні системи комбінують аналогові та дискретні моделі. Агрегативні системи моделюють аналітико-статистичні закономірності біхаверістики складних систем. Методологія інтелектуа-

льних експертних систем дозволяє перебороти недоліки символьного підходу за рахунок комбінування із традиційними інформаційними технологіями та технологіями штучного інтелекту.

У кожному із цих випадків мова йде про становлення нової інформаційної методології моделювання складних процесів й явищ шляхом побудові й застосування моделей-гібридів. З розглянутого видно, що існують різні шляхи для досягнення поставленої мети.

Звичайно, що ці підходи не є вичерпними для усього різноманіття методологій розробки й використання агрегованих моделей. Залишаються відкритими питання розробки гібридних моделей, які комбінують більш ніж два автономних методи й інформаційні технології, питання комбінування, релевантного складності завдання, питання адаптивності складу й структури агрегату залежно від змін у складі й структурі розв'язуваного завдання та багато інших питань.

Практичне застосування лінгвістичного моделювання

Окреслимо ті прикладні галузі та задачі, вирішення яких пропонуєтьсявикористання лінгвістичного моделювання: аналіз та прогнозування часових рядів [9], задача розпізнавання користувача за його рухами (автентифікація [10], визначення емоційного стану користувача[11], рання діагностика хвороб операторів складних технічних систем на ранніх стадіях захворювання за рухами).

Людино-комп'ютерна взаємодія на емоційному рівня є одним з перспективних напрямків науки, орієнтованої на забезпечення природних шляхів використання комп'ютерної техніки. Не секрет, що для органічної взаємодії комп'ютера з людиною, він повинен мати здатність розпізнавати емоційний стан людини.

Відомі розробки щодо використання відображення людських емоцій за допомогою міміки. Такого роду роботи зосереджені на автоматичне розпізнавання емоцій особи з введеної відеоінформації, яка поступає з відповідних пристроїв спостереження.

Оскільки відображення певного емоційного стану людини представлена часовим рядом рухів мишкою, цілком природно кожному емоційному стану поставити у відповідність гібридну ЛМ-ПММ та перехідні до нього стани людини. Будемо розглядати шість основних емоційних станів людини: 1) радість, 2) злість, 3) подив, 4) відраза, 5) страх, 6) сум.

Новою прикладною галуззю є питання пов'язані з застосуванням гібридних моделей на основі лінгвістичних та прихованих Марковських моделей для ранньої діагностики діскинезії (порушень рухової діяльності).

Будь-якою осмисленою руховою активністю керує кора головного мозку. Саме тут виникає спонукання до дії, що по нервових волокнах надходить у спинний мозок, звідки й посилають імпульси в м'язи. Підтримувати центр ваги тіла й координувати руху, не особливо замислюючись, ми вміємо завдяки мозочку. А за м'язовий тонус, ритмічність і плавність рухів відповідають підкіркові ядра головного мозку. Завдяки їм ми здатні на безліч різноманітних дій: біг, стрибки, танцювальні па. Багатство емоційних реакцій (плач, сміх, міміка) теж сфера діяльності підкіркових ядер.

Порушення в роботі цієї складної системи проявляються всілякими мимовільними рухами. Така надмірна активність називається в медицині гримеркою. Для подібних станів характерне ослаблення тонусу мускулатури в сполученні з мимовільними безглуздими рухами й м'язовими спазмами.

Бувають наступні прояви нав'язливої рухової активності: тремор; міоклонія- моментальне скорочення одного м'яза або групи м'язів; тік - короткі, різкі, повторювані здригування голови, тулуба, м'язів особи, рук або ніг; хорея — швидкі хаотичні рухи; атетоз - хробакоподібні рухи; гемібалізм - мимовільні кидкові або обертальні випади одною рукою;

Багато хто з описаних порушень рухової активності свідчать про захворювання нервової системи. Їхні причини можуть бути найрізноманітнішими: невротичні розлади; токсичні поразки головного мозку (наприклад, алкоголем або з'єднаннями міді); спадкоємні або вроджені хвороби; прийом деяких ліків; травми головного мозку, у тому числі родові; уживання деяких наркотиків; пухлини головного мозку; розладу мозкового кровообігу (атеросклероз, інсульт і його наслідки).

Зрозуміло, що ранішня діагностика таких проявів можлива при аналізі рухів людини мишкою під час роботи за комп'ютером.

Застосовуючи лінгвістичне моделювання, ми можемо дослідити відповідні звіни у лінгвістичних ланцюжках. На допомогу тут нам приходить апарат стохастичних та темперальних формальних грама-

 ISSN 1562-9945

тик. Практична реалізація цих алгоритмів може стати в нагоді для служб безпеки особливо небезпечних об'єктів.

ЛІТЕРАТУРА

1. Баклан І.В. Структурний підхід до розпізнавання образів у системах безпеки / Баклан І.В., Селін Ю.М., Петренко О.О. // Національна безпека України: стан, кризові явища та шляхи їх подолання. Міжнародна науково-практична конференція (Київ, 7-8 грудня 2005 р.). Збірка наукових праць. - К.: Національна академія управління — Центр перспективних соціальних досліджень, 2005. - С.375-380.
2. Баклан І.В. Використання карт Кохонена в задачах структурного аналізу поведінки часових рядів / Баклан І.В., Селін Ю.М. // Інтелектуальні системи прийняття рішень та прикладні аспекти інформаційних технологій: зб. наук. праць за матеріалами міжнар. наук.-прак. конф., 15–18 трав. 2006 р. Євпаторія. — Херсон: ХМІ, 2006. — Т. 3. — С.54–58.
3. Баклан І.В. Аналіз поведінки економічних часових рядів з використанням структурних підходів / Баклан І.В., Селін Ю.М. // Сборник МКММ-2006. - Херсон: ХГТУ, 2006.
4. Fu K.S. Syntactic Pattern Recognition and Application. - N.J.: Prentice-Hall, Inc. Englewood Cliffs., 1982. - 596 p.
5. Баклан І.В. Гібридні моделі в статистичних методах розпізнавання образів / Баклан І.В., Рифа В.М. // Вісник ХДТУ. - 2003. - № 3(19). - С.26-28.
6. Баклан І.В. Гібридні технології в проектуванні інтелектуальних систем прийняття рішень / Баклан І.В. // Сучасні інформаційні та інноваційні технології на транспорті: Матеріали Міжнародної науково-практичної конференції. Том 1. – Херсон: Видавництво Херсонського державного морського інституту, 2009. – С.32 –37.
7. Баклан І.В. Імовірнісні моделі для аналізу та прогнозування часових рядів/ Баклан І.В., Степанкова Г.А. // Искусственный интеллект. - 2008. - N 3. - С. 505-515.
8. Баклан І.В. Використання ймовірнісних моделей для аутентифікації оператора складної технічної системи / Баклан І.В. // Національна безпека України: стан, кризові явища та шляхи їх подолання. Міжнародна науково-практична конференція (Київ, 7-8 грудня 2005 р.). Збірка наукових праць. - К.: Національна академія управління — Центр перспективних соціальних досліджень, 2005. - С.380-386.
9. Баклан І.В. Розпізнавання емоцій людини за рухами мишкою за допомогою багаторівневих прихованих Марковських моделей / Баклан І.В. // Искусственный интеллект. Интеллектуальные системы ИИ-2010: Материалы Международной научно-технической конференции. - Донецк: ИПИИ “Наука і освіта”. - 2010. - Т.2. - С.27-30.