**Соколов Артем Михайлович. Методи нейромережевого розподіленого представлення та пошуку схожих символьних послідовностей в задачах класифікації на основі міркувань за прикладами : Дис... канд. наук: 05.13.23 – 2008**

|  |  |
| --- | --- |
| |  | | --- | | Соколов А. М. Методи нейромережевого розподіленого представлення та пошуку схожих символьних послідовностей в задачах класифікації на основі міркувань за прикладами. – Рукопис.  Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.26 – системи та засоби штучного інтелекту. – Міжнародний науково-навчальний центр інформаційних технологій та систем НАН України і МОН України, Київ, 2008.  Дисертаційна робота присвячена розробці та дослідженню методів розподіленого представлення символьних послідовностей на основі розробленого *q*-грамного методу вкладення простору із класичною метрикою редагування в векторний простір з метрикою *l*1.  Розроблено детермінований *q*-грамний метод вкладення простору символьних послідовностей фіксованої довжини над скінченним алфавітом з класичною метрикою редагування в векторний простір з метрикою *l*1. Завдяки використанню підрядків змінної довжини поліпшено якість апроксимації відстані редагування порівняно з відомими методами, що продемонстровано аналітично шляхом застосування апарату графів де Брейна і чисельно на штучних даних.  На основі розробленого детермінованого методу запропоновано локально-чуттєву функцію для класичної відстані редагування, що продукує розподілене представлення послідовностей, яке забезпеченує малу ресурсоємність і можливість створення ефективної процедури пошуку приблизних найближчих послідовностей за сублінійний до розміру бази час – базової операції підходу на основі міркувань за прикладами. Чисельні експерименти показали можливість використання менших, ніж отриманих теоретично, значень параметрів процедури. Розроблені методи пошуку схожих послідовностей за допомогою кластеризації за довжиною послідовностей та вирівнювання довжини, що дало змогу виконувати пошук приблизних найближчих послідовностей різної довжини у реальних базах даних.  Метод застосовано у ряді прикладних задач, де отримано результати кращі відомих або результати на рівні відомих, але за значно менший час. Усі методи реалізовано як програмні засоби, які можуть використовуватися в системах штучного інтелекту. | |
| |  | | --- | | Отримані в дисертаційній роботі результати забезпечують розв'язання актуальної наукової задачі розробки методів нейромережевого розподіленого представлення послідовностей, а також їх пошуку та класифікації, для ефективної оцінки схожості та використання інформації про послідовності в системах штучного інтелекту, які застосовують моделі міркувань людини на основі прикладів. Розроблено, аналітично досліджено, а також програмно реалізовано методи розподіленого представлення та пошуку послідовностей. Ефективність розроблених методів підтверджено експериментальними дослідженнями на тестових і реальних даних при розв'язанні задач пошуку схожих послідовностей і класифікації інформації різного роду (тексти, ДНК, аудит-послідовності).  За результатами проведеного дослідження зроблено такі висновки:  1. Розроблений метод векторного представлення послідовностей забезпечує збереження їх схожості, лінійну (за довжиною вектора) складність апроксимації, можливість аналізу за допомогою теорії метричних вкладень. Аналітично і шляхом чисельних експериментів на штучних даних показана більш висока, порівняно з відомими результатами, точність апроксимації відстані редагування.  2. Розроблені, проаналізовані та реалізовані методи розподіленого представлення послідовностей, які за рахунок використання локально-чуттєвого хешування забезпечують малу ресурсоємність та сублінійний час пошуку приблизних найближчих послідовностей відносно розміру бази прикладів. Експериментальне дослідження якості пошуку на штучних даних показало достатність використання на практиці менших, ніж визначено аналітично, значень параметрів методу, що дозволяє зменшити ресурсоємність пошуку.  3. Запропонований метод нейромережевого розподіленого представлення послідовностей, який використовує рандомізацію векторних представлень і зв’язування елементів послідовності з їхніми позиціями, забезпечує уніфікацію формату представлення і можливість використання мір схожості векторних представлень для оцінки схожості послідовностей.  4. Розроблені методи пошуку схожих послідовностей за допомогою кластеризації за довжиною послідовностей та їх вирівнювання забезпечують пошук послідовностей різної довжини в реальних базах даних і розв'язання прикладних задач пошуку дублікатів і спаму на основі міркувань за прикладами за рахунок використання розподілених представлень і локально-чуттєвого хешування.  Ефективність і практична значимість розроблених методів підтверджені порівнянням отриманих результатів з відомими. Так, при пошуку дублікатів у базі РОМІП результат покращено на 20–40%, на базі *Reuters*-21578, – на рівні відомих. Перспективність застосування цих методів для виявлення спаму в великих поштових серверах показано на прикладі оцінки кількості спаму в колекціях електронних листів *TREC Spam Track* 2006 і 2005, де виявлено до 80% спаму при рівні неправильно класифікованих легальних повідомлень 5–10%.  5. Розроблені методи представлення і пошуку послідовностей забезпечують розв'язання прикладних задач класифікації ділянок ДНК і виявлення вторгнень, що підтверджує ефективність використання міркувань на основі прикладів для обробки послідовностей в реальних базах даних.  У задачі класифікації ділянок ДНК пошук екзонів з використанням підходу на основі міркувань за прикладами пришвидшено в 750 раз при збереженні якості на рівні відомих у цій області результатів. Розроблений метод пошуку послідовностей може застосовуватися при більш широкій області значень параметрів, ніж випливає з теоретичного аналізу, що експериментально показано на прикладі задачі пошуку некодуючих ділянок бета-глобіну при обробці коротких рядків. Запропонований метод є перспективним також для застосування в реальних системах виявлення вторгнень до комп’ютерних систем, що підтверджується результатом класифікації аудит-послідовностей, де отримано точність класифікації на рівні понад 90%.  6. Створені програмні засоби, що реалізують розроблені методи представлення і пошуку приблизних найближчих послідовностей, можуть застосовуватися як компоненти інформаційних технологій, або як самостійні модулі в системах класифікації й пошуку послідовностей. Практична значимість розробок підтверджується 3 актами впровадження. | |