Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА
МАГІСТРА

на тему *Метод агрегативної кластеризації на базі послідовного підходу*

Галузь знань *12 – Інформаційні технології*
Шифр і назва галузі знань

Спеціальність *122 – Комп'ютерні науки*
Шифр і назва спеціальності

Виконав: студент *2* курсу, група *КНм-20-2* — Підпис — *Е.А. Манзюк* (Ініціали, прізвище)

Керівник: *д.т.н., професор кафедри КН* — Підпис — *О.В. Бармак* (Ініціали, прізвище)

Нормоконтроль: *к.т.н., доцент кафедри КН* — Підпис — *Р.О. Багрій* (Ініціали, прізвище)

До захисту допускаю:
Зав. кафедри КН, д.т.н., професор — Підпис — *О.В. Бармак* (Ініціали, прізвище)
25 листопада 2021 р.

Хмельницький 2021

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет *інформаційних технологій*
Кафедра *комп'ютерних наук*
Освітній ступінь *магістр*
Галузь знань *12 – Інформаційні технології*
Спеціальність *122 – Комп'ютерні науки*

ЗАТВЕРДЖУЮ
Завідувач кафедри *комп'ютерних наук*

_____
(підпис)
д.т.н., професор О.В. Бармак
«__1__» ____вересня__ *2021* року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ МАГІСТРА

1. Тема кваліфікаційної роботи магістра: *«Метод агрегативної кластеризації на базі послідовного підходу»*

2. Завдання видано студенту *Манзюк Едуард Андрійович*
(прізвище, ім'я, по батькові)

3. Керівник роботи *д.т.н., професор Бармак Олександр Володимирович*
(прізвище, ім'я, по батькові)

4. Затверджені наказом університету від «__25__» серпня *2021* р. № _102_

5. Зміст пояснювальної записки (перелік задач) та вихідні дані:

Мета роботи – є розробка методу реалізація процесу машинного навчання, який здатний агрегатувати структурні послідовні дані. Для досягнення поставленої мети визначенні такі основні завдання дослідження: Проаналізувати сучасний стан практичних рішень систем формування кластерних структур для формування завдань в предметній області. Удосконалити методи формування кластерних структур із застосуванням цільової кластеризації. Удосконалити системи кластеризації послідовних структур даних та провести порівняльний аналіз із векторними структурами. Розробити метод машинного навчання цільової кластеризації послідовних структур даних. Провести дослідження ефективності практичного застосування запропонованих методів на практичних задачах. Результатом є метод цільової кластеризації для забезпечення визначення ефективних моделей навчання.

Реферат

Кваліфікаційна робота магістра присвячена розробці методу агрегативної кластеризації на базі послідовного підходу.

Актуальність теми. В магістерській роботі було розроблено та набуло практичної реалізації метод агрегативної кластеризації на базі послідовного підходу.

Застосування методів машинного навчання та інтелектуального аналізу даних до реальних даних і реальних проблем часто вимагає значного втручання і досвіду від дослідників. Зокрема, застосування дослідницьких методів часто вимагає повторного пошуку двох моделей: формального пошуку моделі, вбудованого в алгоритм машинного навчання, і неформального пошуку моделі, здійснюваного практиком машинного навчання. Наприклад, дослідник в області розпізнавання мовлення може спочатку спробувати розібрати людську мову за допомогою підходу динамічного викривлення, потім прихованих марківських моделей, а потім лінійних динамічних систем. Цей процес ітеративного тестування різних можливих алгоритмів і оцінки отриманих моделей схожий на стохастичний алгоритм “підняття вгору”. Цей процес необхідний з кількох причин. По-перше, в порівнянні з теоретичним моделюванням і типовими наборами даних, практичні завдання часто включають дані з великою кількістю шуму, великою кількістю спотворень, більш складною структурою і меншою відповідністю загальноприйнятим допущенням (наприклад, незалежності). Особливо з урахуванням того, що в реальних наборах даних часто зустрічаються нові структури, пошук алгоритму, придатного для конкретного набору даних, є в рівній мірі мистецтвом і наукою. По-друге, реальні завдання часто мають цілі, що не обмежуються високою точністю класифікації або щільними, чітко визначеними кластерами. Наприклад, у сфері освіти метою машинного навчання та інтелектуального аналізу даних зазвичай є пошук моделей, що передбачають результати успішності студентів.

Існують алгоритми, які можуть включати в себе різні типи інформації без міток, такі як обмеження "повинен бути пов'язаний" і "не може бути пов'язаний", які зустрічаються в напівконтрольованої кластеризації, або, в більш загальному випадку, імовірнісні відносини, характерні для структурованих графічних моделей. Більшість алгоритмів, що працюють на рівні трасування даних освітнього програмного забезпечення, не можуть прямолінійно включати такі показники, як навчання студентів, а саме, не в контексті кластеризації послідовностей. Це обмеження пов'язане з відстанню між різними рівнями даних: вхідними даними можуть бути серії взаємодій студентів з частотою клацання або набору тексту для кожного завдання; мітками для навчання можуть бути конструктивні і неконструктивні типи взаємодій для кожного завдання; а підсумковим показником може бути кількість засвоєних знань між попереднім і наступним тестом, проведеним і оціненим для кожного студента. Крім того, взаємозв'язки між рівнями даних можуть бути досить складними (наприклад, тестові бали можуть бути пов'язані з розподілом різних типів дій студентів, як буде вказано пізніше). В результаті, ці високорівневі показники результатів служать орієнтиром або метою, корисною для оцінки обчислювальних моделей, але непридатною для виведення параметрів або структури. Обговорювані тут алгоритми розроблені в рамках нової парадигми, яка безпосередньо включає ці високорівневі цілі в процес навчання послідовності, дозволяючи цим додатковим даними поліпшити як навчання моделі, так і її оцінку.

Метою дослідження є розробка методу реалізація процесу машинного навчання, який здатний агрегатувати структурні послідовні дані.

Для досягнення поставленої мети визначенні такі основні завдання дослідження:

– Проаналізувати сучасний стан практичних рішень систем формування кластерних структур для формування завдань в предметній області.

– Удосконалити методи формування кластерних структур із застосуванням цільової кластеризації.

– Удосконалити системи кластеризації послідовних структур даних та провести порівняльний аналіз із векторними структурами.

– Розробити метод машинного навчання цільової кластеризації послідовних структур даних.

– Провести дослідження застосування методів машинного навчання на предметній області.

– Провести дослідження ефективності практичного застосування запропонованих методів на практичних задачах.

Об'єктом дослідження є процес отримування ефективної кластеризації даних для отримання інформативності щодо структури даних.

Предметом дослідження моделі, методи та алгоритми автоматизації побудови цільових кластерів на основі послідовних даних.

Наукова новизна одержаних результатів. В результаті проведеної роботи були отримані такі результати.

Набула подальшого розвитку система формування кластерних структур які залежні від цілі агрегатування даних.

Запропоновано інноваційний метод розробки систем цільової кластризації.

Запропоновано метод формування кластерів на множинах результатів моделей прихованих Марківських процесах.

Відкриття нових психологічних і освітніх конструктів було в основному долею дослідників, а не алгоритмів. Чисто обчислювальні підходи, навіть при наявності хороших даних, в більшості випадків не можуть знайти нові конструкти, які застосовуються до цікавих освітніх проблем. Ця трудність частково пов'язана з тим, що традиційні проблеми машинного навчання спираються на вхідні дані *x* і пов'язані з ними мітки *y* - рамки, відірвані від реальності освітніх досліджень, в яких існують складні взаємозв'язки між конструктами, предметами і завданнями. Стандартні алгоритми просто не враховують багатство наборів освітніх даних, таких як наявність окремих

студентів, окремих завдань і загального навчання. У машинному навчанні існує кілька підходів, що дозволяють включати додаткові дані і обмеження, наприклад, обмеження "необхідність-зв'язок" в кластеризації, але адаптація цих алгоритмів за освітніми даними є нетривіальним завданням. Цільова кластеризація, запропонована в цьому документі, є підхід, створений з нуля, щоб генерувати нових конструктів з освітніх даних. Основний внесок цієї роботі включає:

– новий клас алгоритмів, простих для розуміння і реалізації, але досить потужних для роботи з освітніми даними з широким діапазоном складності і структури;

– різноманітність нових конструкцій для навчання людини, створених на основі існуючих даних. Ці конструкції можуть дати поштовх для додаткових досліджень або надати докази, що підтверджують існуючі результати.

– додаткові алгоритми і підходи для вирішення інших завдань дослідження в галузі освіти.

Практична значимість дослідження полягає в тому, що отримані практичні результати досліджень можуть бути застосовні для побудови цільової кластеризації даних послідовної структури.

Алгоритми кластеризації є потужними методами для розуміння структури даних. Однак у багатьох прикладних областях алгоритми кластеризації корисні тільки в тому випадку, якщо вони дають кластери, які пояснюють зовнішні змінні. Наприклад, в дослідженнях в галузі освіти кластери поведінки студентів корисні тільки в тому випадку, якщо вони допомагають передбачити навчання. Як правило, алгоритми кластеризації явно не призначені для обробки цього додаткового обмеження.

В даній роботі показано, що включення цих зовнішніх змінних безпосередньо в алгоритми кластеризації може поліпшити як релевантність кластерів, так і узагальнення моделі. Цей підхід, названий кластеризацією цільових послідовностей, дав багатообіцяючі результати на множині наборів даних. Хоча продемонстровано ефективність цього підходу на освітніх наборах

даних, також показано, як цей метод може застосовуватися в більш широкому сенсі. Основний внесок цієї роботі включає в себе:

– впровадження та розвиток парадигми прикладного машинного навчання;

– дослідження, розроблені спеціально для вирішення ряду практичних, які раніше не розв'язувалися проблем;

– кілька цільових алгоритмів кластеризації і варіантів алгоритмів;

– зв'язки між кластеризацією цілей і іншими підходами, а також пов'язані з цим можливості для майбутньої адаптації і розробки алгоритмів.

Апробація кваліфікаційної роботи.

Основні положення і результати роботи опубліковані в збірнику наукових праць – Манзюк Е.А. Система цільової кластеризації на послідових даних / Манзюк Е.А., Скрипник Т. К. // Збірник наукових праць за матеріалами Всеукраїнської науково-практичної конференції «Актуальні проблеми комп’ютерних наук - 2021» Хмельницький, 2021, – С. 364 - 366 .

Структура та обсяг роботи. Кваліфікаційна робота магістра складається з завдання, реферату, змісту, вступу, 4 розділів, висновків, переліку посилань із 21 найменувань та додатків. Загальний обсяг кваліфікаційної роботи магістра становить 82 сторінок, з них 79 сторінки основного тексту та 2 сторінки додатків. в роботі наведено 14 рисунків та 3 таблиць.

Ключові слова: кластеризація, приховані моделі Маркова, машинне навчання.

Зміст

Вступ 4
Розділ 1 Аналіз застосування систем машинного навчання 9
1.1 Опис предметної області 9
1.2 Інтелектуальні навчальні системи 10
1.3 Послідовне і векторне представлення 14
1.4 Постановка задачі 19
Висновки до розділу 23
Розділ 2 Розробка кластерної моделі 24
2.1 Структуровані графічні моделі 24
2.2 Напівконтрольована кластеризація 25
2.3 Навчання за допомогою шумових міток 30
2.4 Навчання за декількома фактами 31
2.5 Аналіз цільових факторів 32
2.6 Алгоритм мінімізації помилки на прихованих Марківських моделях 34
2.7 Кластеризація максимізації очікування для прихованих Марківських моделей 35
Висновки до розділу 39
Розділ 3 Розробка системи кластеризації послідовностей 41
3.1 Модель системи кластеризації послідовностей 41
3.2 Структура системи кластеризації 44
3.3 Опис моделювання 48
3.4 Обмеження на моделі 50
3.5 Максимізація очікування дискримінантної і зваженої ПММ 51
3.5 Результати моделювання 54
Висновки до розділу 58
Розділ 4 Дослідження ефективності методу кластеризація на базі Марківських процесів 60
4.1 Чисельні результати моделювання 60

4.2 Цільва кластеризація 62
4.3 Параметри попередньої обробки 65
4.4 Експериментальні моделі 70
Висновки до розділу 73
Загальні висновки 75
Перелік посилань 77

Додатки

Вступ

Кваліфікаційна робота магістра присвячена розробці методу агрегативної кластеризації на базі послідовного підходу.

Актуальність теми. В магістерській роботі було розроблено та набуло практичної реалізації метод агрегативної кластеризації на базі послідовного підходу.

Застосування методів машинного навчання та інтелектуального аналізу даних до реальних даних і реальних проблем часто вимагає значного втручання і досвіду від дослідників. Зокрема, застосування дослідницьких методів часто вимагає повторного пошуку двох моделей: формального пошуку моделі, вбудованого в алгоритм машинного навчання, і неформального пошуку моделі, здійснюваного практиком машинного навчання. Наприклад, дослідник в області розпізнавання мовлення може спочатку спробувати розібрати людську мову за допомогою підходу динамічного викривлення часу, потім прихованих марківських моделей, а потім лінійних динамічних систем. Цей процес ітеративного тестування різних можливих алгоритмів і оцінки отриманих моделей схожий на стохастичний алгоритм “підняття вгору”. Цей процес необхідний з кількох причин. По-перше, в порівнянні з теоретичним моделюванням і типовими наборами даних, практичні завдання часто включають дані з великою кількістю шуму, великою кількістю спотворень, більш складною структурою і меншою відповідністю загальноприйнятим допущенням (наприклад, незалежності). Особливо з урахуванням того, що в реальних наборах даних часто зустрічаються нові структури, пошук алгоритму, придатного для конкретного набору даних, є в рівній мірі мистецтвом і наукою. По-друге, реальні завдання часто мають цілі, що не обмежуються високою точністю класифікації або щільними, чітко визначеними кластерами. Наприклад, у сфері освіти метою машинного навчання та інтелектуального аналізу даних зазвичай є пошук моделей, що передбачають результати успішності студентів,

Існують алгоритми, які можуть включати в себе різні типи інформації без міток, такі як обмеження "повинен бути пов'язаний" і "не може бути пов'язаний", які зустрічаються в напівконтрольованої кластеризації, або, в більш загальному випадку, імовірнісні відносини, характерні для структурованих графічних моделей. Більшість алгоритмів, що працюють на рівні трасування даних освітнього програмного забезпечення, не можуть прямолінійно включати такі показники, як навчання студентів, а саме, не в контексті кластеризації послідовностей. Це обмеження пов'язане з відстанню між різними рівнями даних: вхідними даними можуть бути серії взаємодій студентів з частотою клацання або набору тексту для кожного завдання; мітками для навчання можуть бути конструктивні і неконструктивні типи взаємодій для кожного завдання; а підсумковим показником може бути кількість засвоєних знань між попереднім і наступним тестом, проведеним і оціненим для кожного студента. Крім того, взаємозв'язки між рівнями даних можуть бути досить складними (наприклад, тестові бали можуть бути пов'язані з розподілом різних типів дій студентів, як буде вказано пізніше). В результаті, ці високорівневі показники результатів служать орієнтиром або метою, корисною для оцінки обчислювальних моделей, але непридатною для виведення параметрів або структури. Обговорювані тут алгоритми розроблені в рамках нової парадигми, яка безпосередньо включає ці високорівневі цілі в процес навчання послідовності, дозволяючи цим додатковим даними поліпшити як навчання моделі, так і її оцінку.

Метою дослідження є розробка методу реалізація процесу машинного навчання, який здатний агрегатувати структурні послідовні дані.

Для досягнення поставленої мети визначенні такі основні завдання дослідження:

– Проаналізувати сучасний стан практичних рішень систем формування кластерних структур для формування завдань в предметній області.

– Удосконалити методи формування кластерних структур із застосуванням цільової кластеризації.

– Удосконалити системи кластеризації послідовних структур даних та провести порівняльний аналіз із векторними структурами.

– Розробити метод машинного навчання цільової кластеризації послідовних структур даних.

– Провести дослідження застосування методів машинного навчання на предметній області.

– Провести дослідження ефективності практичного застосування запропонованих методів на практичних задачах.

Об'єктом дослідження є процес отримування ефективної кластеризації даних для отримання інформативності щодо структури даних.

Предметом дослідження моделі, методи та алгоритми автоматизації побудови цільових кластерів на основі послідовних даних.

Наукова новизна одержаних результатів. В результаті проведеної роботи були отримані такі результати.

Набула подальшого розвитку система формування кластер них структур які залежні від цілі агрегатування даних.

Запропоновано інноваційний метод розробки систем цільової кластеризації.

Запропоновано метод формування кластерів на множинах результатів моделей прихованих Марківських процесах.

Відкриття нових психологічних і освітніх конструктів було в основному долею дослідників, а не алгоритмів. Чисто обчислювальні підходи, навіть при наявності хороших даних, в більшості випадків не можуть знайти нові конструкти, які застосовуються до цікавих освітніх проблем. Ця трудність частково пов'язана з тим, що традиційні проблеми машинного навчання спираються на вхідні дані *x* і пов'язані з ними мітки *y* - рамки, відірвані від реальності освітніх досліджень, в яких існують складні взаємозв'язки між конструктами, предметами і завданнями. Стандартні алгоритми просто не враховують багатство наборів освітніх даних, таких як наявність окремих

студентів, окремих завдань і загального навчання. У машинному навчанні існує кілька підходів, що дозволяють включати додаткові дані і обмеження, наприклад, обмеження "необхідність-зв'язок" в кластеризації, але адаптація цих алгоритмів за освітніми даними є нетривіальним завданням. Цільова кластеризація, запропонована в цьому документі, є підхід, створений з нуля, щоб генерувати нових конструктів з освітніх даних. Основний внесок цієї роботі включає:

– новий клас алгоритмів, простих для розуміння і реалізації, але досить потужних для роботи з освітніми даними з широким діапазоном складності і структури;

– різноманітність нових конструкцій для навчання людини, створених на основі існуючих даних, ці конструкції можуть дати поштовх для додаткових досліджень або надати докази, що підтверджують існуючі результати;

– додаткові алгоритми і підходи для вирішення інших завдань дослідження в галузі освіти.

Практична значимість дослідження полягає в тому, що отримані практичні результати досліджень можуть бути застосовні для побудови цільової кластеризації даних послідовної структури.

Алгоритми кластеризації є потужними методами для розуміння структури даних. Однак у багатьох прикладних областях алгоритми кластеризації корисні тільки в тому випадку, якщо вони дають кластери, які пояснюють зовнішні змінні. Наприклад, в дослідженнях в галузі освіти кластери поведінки студентів корисні тільки в тому випадку, якщо вони допомагають передбачити навчання. Як правило, алгоритми кластеризації явно не призначені для обробки цього додаткового обмеження.

В даній роботі показано, що включення цих зовнішніх змінних безпосередньо в алгоритми кластеризації може поліпшити як релевантність кластерів, так і узагальнення моделі. Цей підхід, названий кластеризацією цільових послідовностей, дав багатообіцяючі результати на множині наборів

даних. Хоча продемонстровано ефективність цього підходу на освітніх наборах даних, також показано, як цей метод може застосовуватися в більш широкому сенсі. Основний внесок цієї роботі включає в себе:

– впровадження та розвиток парадигми прикладного машинного навчання;

– дослідження, розроблені спеціально для вирішення ряду практичних, які раніше не розв'язувалися проблем;

– кілька цільових алгоритмів кластеризації і варіантів алгоритмів;

– зв'язки між кластерізацією цілей і іншими підходами, а також пов'язані з цим можливості для майбутньої адаптації і розробки алгоритмів.

Апробація кваліфікаційної роботи.

Основні положення і результати роботи опубліковані в збірнику наукових праць – Манзюк Е.А. Система цільової кластеризації на послідових даних / Манзюк Е.А., Скрипник Т. К. // Збірник наукових праць за матеріалами Всеукраїнської науково-практичної конференції «Актуальні проблеми комп’ютерних наук - 2021» Хмельницький, 2021, – С. 364 - 366.

Структура та обсяг роботи. Кваліфікаційна робота магістра складається з завдання, реферату, змісту, вступу, 4 розділів, висновків, переліку посилань із 21 найменувань та додатків. Загальний обсяг кваліфікаційної роботи магістра становить 82 сторінок, з них 79 сторінки основного тексту та 2 сторінки додатків. в роботі наведено 14 рисунків та 3 таблиць.

Розділ 1
Аналіз застосування систем машинного навчання

1.1 Опис предметної області

Дослідження в галузі освіти мають довгу і багату історію. У дослідженнях в галузі освіти існує множина тем для вивчення, починаючи від когнітивних моделей і закінчуючи соціальною динамікою в класі [1-4]. Дана робота присвячена області метакогнітівної стратегій і поведінки. Метакогніція відрізняється від інших тим, що не залежить від області: метакогнітівного стратегія повинна бути застосовна в різних дисциплінах і включає в себе завдання самоконтролю та саморегуляції [5, 6].

Методи машинного навчання використовуються для аналізу даних освітнього моніторингу вже кілька десятиліть [7-9]. Зокрема, для метакогнітівного стратегій і поведінки, деякі теми, які були досліджені за допомогою машинного навчання, включають:

- припущення;
- гра з системою / зловживання допомогою ;
- стратегії викладання / репетиторства;
- загальні стратегії високого рівня.

Насправді, багато хто з цих тем були вивчені за допомогою прихованих марківських моделей (ПММ), включаючи кожен із прикладів, наведених вище. Наприклад. використовували нейронні мережі і ПММ для аналізу взаємодії студентів при спробі вирішити нечітко сформульовані завдання. Нейронні мережі знаходять закономірності у виборі студентів, які потім формують стратегії [10, 11]. Використовували ПММ для моделювання переходів між цими стратегіями [12,13]. Використовували ПММ для аналізу кроків, які студент використовував для "навчання" комп'ютерного агента по темі. Розбили ПММ на окремі категорії. Наприклад, в ПММ побудови карти студенти зазвичай

редагували карту, відправляли карту і іноді використовували ресурси для читання.

Такі дослідження відрізняється від даного за кількома параметрами: вони визначають структуру ПММ вручну, вони використовують результати іншого алгоритму в якості вхідних даних для своїх ПММ, вивчають одну ПММ для кожного студента, і вони проводять кластеризацію студентів (не тактик) тільки після вивчення параметрів ПММ.

Проте, ключовий результат як і раніше актуальний: ПММ можуть працювати як описові та прогностичні моделі поведінки студентів в процесі навчання і можуть знаходити корисні закономірності без використання когнітивних моделей або знань про зміст предмета.

1.2 Інтелектуальні навчальні системи

Дана робота буде присвячена одній конкретній подмножині освітніх досліджень, сфокусованих на інтелектуальних навчальних системах. Інтелектуальні навчальні системи дозволяють студентам взаємодіяти з програмним забезпеченням, яке адаптується до когнітивних навичок (і, можливо, метакогнітівним навичкам). Така адаптація може бути простою - не пропонувати проблеми, засновані на навичках, якими студент вже опанував, або складною - додаткова оцінка і навчання, сфокусоване на навичках домену та метакогнітівного зворотного зв'язку [14-17]. Приклади інтелектуальних навчальних систем включають:

1. Система репетиторства на базі Інтернету, спрямована на вивчення концепцій і навичок, пов'язаних з навчальною програмою і тестами, в середніх школах.

2. Авторепетитор: Навчає критичного мислення і метакогнітівного навичкам в контексті фізики.

3. Когнітивні репетитори: Варіанти охоплюють ряд навчальних програм з

математики, природничих наук і мови, включаючи основні алгебру і геометрію, з повним набором допоміжних навчальних програм.

4. Мультимедійна навчальна система, орієнтована на стандартні тести з математики.

Дослідження всіх вищезазначених навчальних систем включали роботу над метакогніфікацією, іноді у формі прогностичних моделей, а іноді у формі активного втручання. В даній роботі зосередимося на прогностичних моделях, але варто розглянути в контексті ті форми, які може приймати освітнє втручання. Іноді втручання приймає форму заохочення або заборони поганого поводження, іноді - заохочення або моделювання позитивної поведінки, а іноді воно включає в себе значну перебудову, спрямовану на запобігання поганого поводження до його виникнення.

Стандартний підхід дослідження, який використовується в інтелектуальних навчальних системах, полягає в наступному:

1. Запропонувати студентам пройти попереднє тестування, щоб визначити початковий рівень засвоєння матеріалу.

2. Дозволити студентам взаємодіяти з навчальною системою.

3. Запропонуйте студентам пройти пост-тест, щоб визначити остаточне засвоєння матеріалу.

На другому етапі цього процесу студенти генерують велику кількість даних журналу. Ці дані можуть бути такими докладними, як клацання мишею і стеження за очима, або такими грубими, як час, проведений за монітором. Алгоритми машинного навчання можуть бути застосовані до даних журналу, створюючи різні типи моделей поведінки студентів. Потім ці моделі можуть бути використані для прогнозування успішності студентів. Саме на цьому етапі аналізу може виявитися корисною кластеризація цілей.

В основній частині прикладів і аналізу будуть використовуватися дані, які були використана в якості канонічного прикладу в попередніх роботах по кластеризації цілей [18-22].

У цих наборах даних студенти виконують ряд дій, наприклад, запитують підказку. Кожна дія виконується для певного етапу (наприклад, знаходження кута в трикутнику). Завдання складається з серії пов'язаних кроків, а завдання, в свою чергу, діляться на блоки.

На будь-якому заданому кроці дії студента можуть бути представлені послідовністю $s \in \Sigma n_s$, Де Σ - множина можливих дій, а n_s - випадкова змінна, що представляє довжину кроку або завдання. Нехай Σ також відомо як алфавіт. У найпростішому випадку спроба студента вирішити етап представлена послідовністю С (спроб) і П (підказок). Наприклад, довга серія підказок, за якою слідує спроба, матиме вигляд ППППС. У цьому прикладі Σ= {*С, П*} і n_s дорівнює 4.

Як правило, найбільш важливими діями будуть С і П. Однак студенти можуть зробити як правильну, так і неправильну спробу, а не тільки загальну. Таким чином, в деяких експериментах представлятимуть правильні і неправильні спроби, замінюючи С.

Послідовності дій з декількох кроків можуть бути об'єднані в послідовності дій для вирішення завдання. Наприклад, якщо студент з першої спроби правильно виконує перші два кроки, але потім припускається помилки і потребує підказку для вирішення останнього кроку, то окремі послідовності кроків будуть виглядати наступним чином:

- **С**
- **С**
- **П**
- **ППППС**

У вигляді послідовності одного завдання це буде: СППССППС. Крім того, студенти можуть перемикатися з одного етапу на інший. Таким чином, для

позначення перемикачів можна додати додатковий символ - Х. З символом Х приклад послідовності завдань буде виглядати так: СХПСХСХПППСХ.

Як правило, ці різні представлення вирішують різні питання. Зокрема, моделі поведінки, які спостерігаються в послідовності кроків, можуть бути наступними: повторне вгадування кроку, звернення за підказкою і т.д. Однак моделі поведінки, які спостерігаються в послідовності завдань, можуть забезпечити додатковий контекст і розглянути більш складні моделі поведінки, наприклад, стратегії вирішення завдань, наприклад, рішення спочатку легких кроків, а не проходження кроків по порядку. На жаль, репрезентація на рівні проблеми має і недоліки. Розмірність є проблемою, і робити висновки складніше для великих алфавітів символів, що збільшує кількість параметрів, які необхідно вивчити, а також для проблем в порівнянні з кроками, так як в поданні на рівні проблеми менше проблем і, отже, менше прикладів.

Кожна дія також має відповідну тривалість. Наприклад, одному студенту може знадобитися 7 секунд, щоб прочитати підказку, або іншого студента може знадобитися 3,3 секунди, щоб ввести спробу рішення. Одним із способів включення цієї тимчасової інформації в алфавіт є поділ за пороговим значенням таким чином, що швидкі дії позначаються одним символом, а повільні - іншим.

Загалом, всі символи в верхньому регістрі представляють дії великої тривалості (або дії, які не мають порога), а всі символи в нижньому регістрі - дії малої тривалості. Практично, це обмежується спробами (С, с), підказками (П, п), правильними спробами (О, о) і неправильними спробами (Н, н). Інші дії відбуваються досить рідко, і немає ніяких доказів поліпшення роботи моделі або поліпшення інтерпретується, коли вони обмежуються граничними значеннями.

Використовуємо просте скорочення для ідентифікації набору даних, і використовуваних порогових значень. Наприклад, С8П8 вказує, що набір даних відноситься використовує репрезентацію на рівні кроків і використовує поріг в 8 секунд для спроб і підказок. Нарешті, щось на зразок 8 секунд - це часто використовуваний поріг, оскільки він виявився ефективним вибором.

1.3 Послідовне і векторне представлення

Як було показано раніше, основне представлення цих журнальних файлів - послідовне. На жаль, більшість існуючих алгоритмів, що відносяться до кластеризації цілей, визначені для векторних просторових представлень даних; існують алгоритми, краще підходять для навчання послідовності, але ці алгоритми, як правило, вимагають більше даних, ніж є в наявності.

Однак багато існуючих алгоритмів теоретично можуть бути адаптовані для цільової кластеризації. Наприклад, послідовні дані можна розглядати як мішок слів, де кожен можливий тип символу перетворюється в один вимір у векторному поданні.

В принципі, як тільки це перетворення завершено, будь-який алгоритм векторного простору є потенційний підхід до аналізу даних. На практиці це не так, оскільки багато алгоритми не працюють з векторними даними з невеликими наборами ознак. Крім того, як буде показано далі, це перетворення (яке жертвує всією інформацією про впорядковування) має тенденцію приводити до поганих результатів у реальних наборах даних, що представляють інтерес.

Існують і інші можливі перетворення. Наприклад, якщо вибирати верхню межу довжини будь-якій послідовності, то всі послідовності можуть бути стандартизовані до цієї довжини (або шляхом усічення, або шляхом додавання порожніх символів). Тоді кожен індекс в послідовності можна розглядати як ознаку.

Тривіальний аналог проблеми кластеризації цілей. Кожен можливий символ представляє номінальне значення. Цей випадок не буде розглядатися тут, але, швидше за все, він буде погано працювати з даними інтелектуальної навчальної системи, де мінливість довжини послідовності містить життєво важливу інформацію.

Однак є кілька переваг вибору векторного представлення простору і відповідних алгоритмів. По-перше, більшість алгоритмів, пов'язаних з цільовою кластерізацією, є векторними алгоритмами. По-друге, ці алгоритми можна легко протестувати на векторних даних, щоб визначити життєздатність після адаптації до цільової кластеризації, що неможливо зробити з більшістю послідовних алгоритмів через відсутність стандартних наборів даних з відомою базової істиною. Нарешті, більшість цих алгоритмів є варіантами класичних, надійних рішень проблем навчання без нагляду або з наглядом, тому для пошуку рішень можна використовувати результати десятиліття досліджень.

На жаль, ці методи також погано працюють, коли до послідовних даними застосовуються необхідні перетворення з втратами. Для ілюстрації розглянемо *k*-середніх і спектральну кластеризацію (з k-середніми):

К-середніх - це стандартний рандомізований алгоритм максимізації очікування для пошуку кластерів в евклідовому просторі. Однак для послідовних даних, перетворених в мішки слів, k-means навряд чи буде працювати добре, внаслідок втрати послідовної інформації, так і за наявності точок перекриття даних.

Спектральна кластеризація, як правило, полягає у виконанні кластеризації k-means після використання розкладання за сингулярним значенням для зменшення розмірності. Вона здатна знаходити кластери уздовж різноманітності в даних, але все ще страждає від втрати даних.

Щоб застосувати ці методи, необхідно виконати кілька кроків:

– перетворення даних в дискретне, граничне, символьне представлення;

– перетворення кожної послідовності дій в представлення у вигляді мішка слів, в результаті чого виходить $m \times |\Sigma|$ матриця X;

– нормалізування кожного рядка;

– розділення навчальних та тестових даних;

– застосування к-середніх або спектральну кластеризацію до навчальних даних, отримуючи c кластерів і привласнення кожного рядка кластеру.

– створення $n \times c$ матрицю D, де $D_{i,j}$ - це ймовірність того, що випадкова послідовність від *i-го* студента буде належати кластеру j.

– застосувати регресію до навчальних кластерів, намагаючись передбачити навчання студентів (вектор $n \times 1$).

– зберігши коефіцієнти регресії і центри кластерів з навчальних даних, розрахувати кластери на тестових даних. Використовуючи ті ж коефіцієнти, що і раніше, розрахувати R2 для оцінки успішності студентів.

Рисунок 1.1 - Експериментальна процедура, яка використовується для векторних алгоритмів.

Таблиця 1.1 - Продуктивність векторного алгоритму

алгоритм	**середній тренувальний** *R2*	**кращий тест** *R2*
K-Means	0.4269	-28.6939
спектральний	0.5124	-15.3498

Перші 80% кроків кожного студента стають доступними в якості навчальних даних. У першому стовпці показана продуктивність навчального набору, усереднена за 10 ітераціями. У другому стовпці показана низька продуктивність тестового набору, що говорить про надмірну підгонку. Варто відмітити що *R2*, хоча і обмежений в межах від 0 до 1 на тренувальних даних, при застосуванні до інших даних не має обмежень по спадаючій.

Є дві ймовірні причини того, що ці алгоритми не генерують моделі, що передбачають навчання студентів. По-перше, самі алгоритми можуть бути поганими, наприклад, k-сердніх чутливих до початкових стартових центрів і вибору k. По-друге, може бути принципово занадто велика втрата даних при перетворенні з послідовностей в моделі.

Щоб краще зрозуміти, чому традиційні методи (наприклад, K-Means) можуть зазнати невдачі при вирішенні подібних завдань, корисно розглянути кластеризацію цільових послідовностей за аналогією з пошуком інформації в Інтернеті. Кожен студент відповідає веб-сторінку, кожна послідовність дій відповідає слову, а результат навчання є певною оцінкою якості веб-сторінки. Коли людина читає слово на веб-сторінці, в гру вступає контекст. На веб-сторінці слова формують пропозиції, пропозиції формують абзаци, а абзаци розвивають аргументи і лінії міркувань. Аналогічним чином, для студента, що взаємодіє з навчальною системою, дії на кроках об'єднуються в поведінку при вирішенні завдань, поведінка при вирішенні завдань об'єднується в поведінку протягом сесії.

У цій аналогії корисно думати про мету інформаційного пошуку як про релевантну конструкції: слова, релевантні для пошуку спортивних автомобілів, можуть не бути доречними для пошуку рецептів яблучних пирогів. Цей приклад є ілюстрацією того, чому кластеризація цілей важлива, особливо для освіти: немає апріорної причини вважати, що поведінка, релевантна для однієї оцінки, обов'язково має бути релевантною для іншою.

Класичним рішенням інформаційного пошуку, менш популярним в даний час, було ігнорування контексту і розгляд кожної веб-сторінки як пакета слів. Таке спрощення було можливим (і необхідним) частково через неймовірну кількості доступних веб-сторінок; властивості будь-якого слова можна було визначити, використовуючи сотні тисяч веб-сторінок, на яких зустрічалося навіть маловідоме слово. Ця залежність від розміру набору даних відрізняє проблеми кластеризації цільових послідовностей від проблем інформаційного

пошуку: в Інтернеті можуть бути мільярди веб-сторінок, але більшість наборів освітніх даних з потрібними характеристиками (наприклад, послідовності дій, показники для кожного студента, окремі кроки і т.д .) містять лише кілька студентів.

Через брак даних успішний підхід до вирішення проблем кластеризації цільових послідовностей в освіті не може повністю спиратися на існуючі методи. Проте, існує багато спільних рис, і варто пам'ятати про аналогію з інформаційним пошуком. У наступних розділах будуть представлені методи попередньої обробки, майже ідентичні стоп-листами, а також алгоритми, що використовують метрики інформаційного пошуку.

В цілому, буде показано, що, включивши інформацію про мету, алгоритми кластеризації послідовностей можуть бути адаптовані для навчання кластерам, що належать до обраних тем. Зокрема, ці алгоритми будуть добре працювати на симульованих даних, отриманих на основі розроблених експертами моделей, будуть добре працювати на випадково згенерували моделях при певних умовах і будуть добре працювати на реальних даних. Однак найбільш ефективні алгоритми будуть повільніше і більш схильні до локальних максимумів, в той час як більш швидкі, глобально оптимальні алгоритми будуть менш стійкими. Крім того, буде показано, що ці результати залежать як від алгоритмів, які адаптуються до інформації про мету, так і від наявності високоякісної інформації про мету.

Цільова кластеризація тісно пов'язана з наступними класами завдань: структуровані графові моделі, напівконтрольована кластеризація, навчання з зашумленими мітками, навчання з декількома мітками і навчання з множинами обставин. Багато алгоритмів з кожного класу можуть бути адаптовані для вирішення завдань кластеризації. Кожна область попередніх досліджень також надає ключові мотиви і рамки для цільової кластеризації. Існує також кілька окремих алгоритмів, розроблених для вирішення схожих проблем або особливих

випадків. Конкретні алгоритми, які мають найтісніший контакт цільової кластеризації, будуть розглянуті в кінці цього розділу.

Для цілей даного розділу припустимо, що всі $\lambda_j \in \Lambda$ дискретні, оскільки в більшості попередньої даних λ_j розглядається як дискретна змінна. Прикладами дискретних λ є залік / незалік, випускник / невипускник, добре вчиться / погано вчиться, освоїв / не освоїв і A / B / C / D / F. Ці дискретні λ відрізняються від безперервних λ, таких як оцінки до і після тестування, відсоток правильно вирішених завдань і рівень відвідуваності. Хоча більшість λ, використовуваних в даній роботі, насправді неперервні, припущення про те, що вони дискретні, спрощує багато прикладів в цьому розділі.

1.4 Постановка задачі

Нехай $X = \{x_i \mid i = 1 \ldots n\}$ - це дані, що представляють інтерес. x_i може бути точкою в евклідовому просторі високої розмірності, послідовністю символів, кутом гіперкуба або будь-яким іншим типом даних. Для цілей даної роботі, x_i буде послідовністю символів, кожен з яких представляє дію студента в освітній програмному середовищі. Нехай $Y = \{y_i \mid i = 1 \ldots n\}$ - це (зазвичай неспостережувані) мітки для x_i. Кожен y_i можна розглядати як контрольне значення відношення *i-ї* точки до кластеру, інтерпретованих у конкретній галузі. Наприклад, кожен кластер може представляти хороші або погані дії, оцінку уважності студента або будь-який інший вид загальної категоризації. Для цілей даної роботи $y_i \in Y$ завжди буде дискретним. Нарешті, нехай $\Lambda = \{\lambda_j\}$ - це множина цілей більш високого рівня, де $j < n$. Наприклад, якщо кожна λ_j є оцінкою тесту студентів, то $1 \leq j \leq m$. Нехай кожна λ_j визначена таким чином, що вона відображається на підмножині *X*. Наприклад, оцінка тесту кожного студента відображається на його або її низькорівневі дії в освітньому програмному середовищі. У загальному

випадку, ці мітки λ_j можуть бути будь-якими - від успіхів в навчанні до кількості прогулів.

Як більш докладної ілюстрацією розглянемо автоматизоване оцінювання контрольної роботи. Для автоматизованого оцінювання кожен x_i - це контрольна робота студента, можливо, представлена у вигляді вектора слів або *n*-грам. Кожен y_i - це мітка для контрольної роботи, наприклад, "залік" або "незалік". Ця частина кластеризації цілей така ж, як і в стандартних підходах до оцінки контрольної роботи.

Цільова кластеризація додає цілі (тобто значення λ), які представляють собою бажані класифікації студентів. Наприклад, якщо автоматизована система оцінювання контрольної роботи призначена для вступних іспитів до коледжу, то значення λ можуть бути показниками закінчення коледжу. В цьому випадку y_i будуть не просто представлення "залікових" або "незданих" контрольних робіт, а "заліковими" або "незданими" контрольними роботами, класифікованими таким чином, щоб y_i передбачали рівень закінчення коледжу.

Мета алгоритму цільової кластеризації полягає в тому, щоб, поставивши X і Λ, визначити Y таким чином, щоб X відображався на Y, а Y - на Λ. Це завдання передбачає оптимізацію в двох просторах пошуку: відображення між X і Y і відображення між Y і Λ. Ці співвідношення можуть бути задані параметрично або непараметричні, як і в звичайних задачах навчання.

Наприклад, якщо повернутися до прикладу автоматизованого оцінювання контрольної роботи, то один з простих способів класифікації контрольної роботи - за обсягом словникового запасу. Якщо вступник використовує понад 100 окремих слів, позначте його як "залік", а якщо менше 100 слів, позначте його як "незалік". Цей алгоритм є простий спосіб оцінки іспитів і може навіть добре поєднуватися з навчальними мітками, такими як людська оцінка контрольної роботи. Однак якщо мета полягає в тому, щоб оцінювати контрольної роботи таким чином, щоб оцінки за контрольної роботи допомагали передбачити

успішність студента в коледжі, то обсяг словникового запасу, ймовірно, є важливим фактором.

Таким чином, існує постійний компроміс між поліпшенням відображення між X і Y (іспит і оцінка) і відображенням між Y і Λ (оцінка і майбутня успішність). Цей компроміс є основним компонентом цільової кластеризації і може бути оптимізований або ітеративно (оптимізація одного відображення, а потім іншого), або одночасно.

Інший погляд на цільову кластеризацію полягає в тому, щоб розглядати проблему як завдання адаптувати систему. При традиційній кластеризації метою є навчання функції f така, що $f(X) = Y$ мінімізує деяку об'єктивну функцію. При цільовій кластеризації вводиться друга функція g, яка повинна бути оптимізована так, щоб $g(Y) \approx \Lambda$.

Одним із прикладів цільового алгоритму кластеризації є ПММ, який буде описаний пізніше. Однак, для ілюстрації, в ПММ, f - це набір прихованих моделей Маркова (ПММ), а g - лінійна регресія. Оптимізація f вимагає вивчення параметрів набору ПММ, а оптимізація g включає покрокову регресію, побічним ефектом якої є вибір моделі для ПММ. Цей процес створює петлю зворотного зв'язку, яка дозволяє цільовим значенням λ впливати на навчання f.

Сучасний погляд на цільову кластеризацію полягає в тому, щоб розглядати кластеризацію як надання відповіді на питання; проте в прикладних областях несамостійна кластеризація часто зводиться до пошуку відповіді і подальшим спробам пов'язати відповідь з вихідним питанням. Це можна назвати "кластерізацією з підкріпленням", і прикладом тому служить поширене використання алгоритмів кластеризації, в яких групові послідовності лише послідовно співвідносяться зі змінними, що представляють реальний інтерес. Тут немає особливих причин вважати, що на кожне питання про набір даних має бути надана відповідь за допомогою одного і того ж розбиття даних. Крім того, напівконтрольоване рішення працює тільки при наявності високоякісних міток.

Робота починається з обговорення досліджень в галузі освіти і машинного навчання, особливо сфокусованих на інтелектуальних навчальних системах. Далі слідує обговорення даних і представлення даних, а також короткий екскурс в алгоритми векторного простору. Основна частина документа розділена на два класи алгоритмів: алгоритми, засновані на прихованих моделях Маркова (ПММ), і алгоритми, засновані на навчанні з використанням метрики дисбалансу. Ці розділи складаються з опису алгоритмів, результатів моделювання і емпіричних результатів. У заключних розділах докладно описуються практичні результати для сфери освіти і загальні висновки.

Перший клас алгоритмів передбачає вивчення прихованих марківських моделей для всіх або деяких даних, а потім коригування, поліпшення або адаптацію моделей для отримання кластерів, які передбачають мету (Λ). Другий клас алгоритмів передбачає використання (або навчання) метрики відстані між послідовностями, перетворюючи проблему з послідовної в проблему кластеризації з метрикою відстані.

Ці метрики зберігають деяку інформацію про порядок в своїх вихначеннях відстані, а у випадку навчання метрика інформації про цілі може бути легко включена.

Відповідно метою дослідження є розробка методу реалізація процесу машинного навчання, який здатний агрегатувати структурні послідовні дані.

Для досягнення поставленої мети визначенні такі основні завдання дослідження:

– Проаналізувати сучасний стан практичних рішень систем формування кластерних структур для формування завдань в предметній області.

– Удосконалити методи формування кластерних структур із застосуванням цільової кластеризації.

– Удосконалити системи кластеризації послідовних структур даних та провести порівняльний аналіз із векторними структурами.

– Розробити метод машинного навчання цільової кластеризації послідовних структур даних.

– Провести дослідження застосування методів машинного навчання на предметній області.

– Провести дослідження ефективності практичного застосування запропонованих методів на практичних задачах.

Висновки до розділу

Таким чином в цьому розділі досліджено предметну область. Аналіз відомих досліджень та проблематики показав на актуальність поставленої мета та досить широке коло пропонованих методів, які були застосовні із сумісними зданим предметної області. Проведено детальний огляд методів машинного навчання та аналіз структур даних. Це дозволило визначити сукупність завдання, практична реалізація яких дасть змогу розробити нові підходи щодо застосування методів машинного навчання з врахуванням специфіки даних предметної області.

Розділ 2
Розробка кластерної моделі

2.1 Структуровані графічні моделі

Основна ідея структурованих графічних моделей полягає в розширенні традиційних графічних моделей за межі плоских даних (тобто однієї таблиці реляційної бази даних) на дані, в яких змінні можуть мати ймовірні батьківські відносини (наприклад, правильність або неправильність кроку частково передбачається студентом, а частково - самим кроком). Одним з основних підходів є використання імовірнісних реляційних моделей (IPM), які, як випливає з назви, об'єднують імовірнісні моделі зі схемами реляційних баз даних. Сила IPM полягає в універсальності: вони можуть моделювати невизначеність в змінних, залежностях між змінними і навіть в реляційної схемою. Однак IPM та інші структуровані графічні моделі раніше не застосовувалися до освітніх даних, і неясно, як вони будуть працювати. Зокрема, освітні дані часто мають незвичайну структуру. Наприклад, як показано в попередньому прикладі з мішком слів, виходили з того, що поведінка студента відноситься до кластерів і що розподіл дій студента по кластерам прогнозує навчання; структуровані графічні моделі, звичайно, можуть враховувати цей взаємозв'язок, але необхідні розширення нетривіальні.

Однак, крім того, що структуровані графічні моделі є потенційним методом прямого аналізу даних про освіту, вони також можуть представляти собою шлях для розширення алгоритмів кластеризації. Наприклад, основний алгоритм кластеризації цілей в цій роботі спирається на приховані моделі Маркова (ПММ) для моделювання спостережуваної поведінки студентів. Є робота по адаптації IPM для роботи з ПММ. Використовують ПММ для моделювання спостережуваної поведінки, а потім відділяють ПММ від складних зв'язків з іншими змінними за допомогою шару абстракції.

2.2 Напівконтрольована кластеризація

Напівконтрольоване навчання, в загальному випадку, являє собою задачу навчання на мітках з суміші непомічених і помічених навчальних даних, плюс обмеження. Як формалізованих задач, вона сходиться до задач про спільне навчання. Однак завдання напівконтрольованого навчання мають тісний зв'язок з проблемами відсутності даних. Напівконтрольована кластеризація - це особливий випадок, коли мітки вивчаються на основі суміші немічених навчальних даних, мічених навчальних даних і обмежень, але коли повний набір міток невідомий. Таким чином, напівконтрольований алгоритм кластеризації може викликати нові мітки, не присутні в навчальних даних. Це робить кластеризацію подібною до цільової кластеризації.

Існує кілька парадигм для напівконтрольованої кластеризації, але найбільш поширені або припускають парні обмеження на навчання, наприклад, $(x_i, x_j) \rightarrow y_i = y_j$, або припускають марковані точки з неповним набором міток. Назвемо перший варіант напівконтрольована кластерізацією з обмеженнями, а другий - напівконтрольована метричної кластерізацією.

Алгоритми кластеризації з обмеженнями зазвичай засновані на алгоритмах кластеризації без спостереження з модифікованою метрикою відстані, яка включає деяку форму чистоти кластера, які іноді описуються як методи адаптації подібності. Алгоритми метричної кластеризації зазвичай засновані на графічних методах, які розглядають зв'язки як присутні або відсутні ребра, які іноді описуються як методи, засновані на пошуку. Для навчання модифікованих метрик відстані з використанням парних обмежень. Ця категорія навчання метрик за допомогою обмежень з тих пір стала активною областю досліджень.

Цільова кластеризація і напівконтрольована кластеризація не ідентичні. Напівконтрольована кластеризація передбачає, що, хоча у нас немає повної

інформації про *f*, у нас є часткова інформація у вигляді неповного набору міток або обмежень на мітки. В результаті напівконтрольована кластеризація передбачає зменшений простір пошуку (деякі моделі-кандидати відсіюються зазначеними обмеженнями) і відсутність функції *g*. Однак при цільової кластеризації пошук ведеться як за *f*, так і по за *g*.

З двох основних класів алгоритмів напівконтрольованої кластеризації (виключаючи гібриди) алгоритми напівконтрольованої метричної кластеризації є більш перспективними для алгоритмів цільової кластеризації. Ці методи зазвичай залежать від нової метрики відстані, яка поєднує стандартну метрику відстані чистоти кластера. Наприклад, стандартно. метрикою відстані є евклідова відстань між центроїдами двох кластерів; при розгляді присвоєння точки *x* з міткою *y* в кластер C, евклідова відстань може бути оштрафована коефіцієнтом |C|, де C*y* - множина всіх точок в C з міткою *y*. Методи визначення чистоти кластера (включаючи такі методи, як ентропія кластера і взаємна інформація) можуть бути адаптовані для цільової кластеризації, дозволяючи використовувати кілька міток і як дискретний, так і безперервний Λ. Наприклад, розглянемо найпростішу міру чистоти кластера - кількість примірників мажоритарного класу, поділене на розмір кластера (як показано в попередньому прикладі). При дискретному Λ цей показник може бути обчислений точно так, як і для кожного рівня ієрархії міток, тобто для однієї точки *x* з *j* мітками y_j, n_j.

Деякі напівконтрольовані алгоритми кластеризації з обмеженнями також застосовні до цільової кластеризації. Головними серед них є алгоритми з ймовірними обмеженнями, оскільки вони послаблюють вимогу строгих парних обмежень і замінюють з нечіткою схожістю. Прикладами цього є CVQE. Однак навіть парні алгоритми теоретично можуть бути адаптовані для цільової кластеризації. Особливо перспективною є ієрархічна агломеративна кластеризація (ІАК). ІАК, яка будує ієрархію кластерів з використанням парних відстаней, є перспективною, оскільки рівні ієрархії ІАК теоретично можуть бути обмежені набором міток Λ.

Ієрархічна агломеративна кластеризація починається з того, що кожна точка даних знаходиться в своєму власному кластері. Під час кожної ітерації групується два найближчих кластера разом. Протягом багатьох ітерацій ІАК створює дендрограмму, яка представляє собою граф з деревовидної структурою, що показує ієрархію кластерів. ІАК є агломеративною, оскільки використовує підхід "від низу до верху", об'єднуючи менші кластери в великі, поки не буде створено єдиний кластер розміру n.

Як і у всіх майбутніх дендрограммах, ідентичні послідовності згортаються, і утворюється тільки одна точка даних, що представляє всі послідовності виду. В цілому, існує занадто багато унікальних послідовностей.

На практиці, враховуючи вхідну матрицю X, ІАК передбачає існування трьох функцій: функції відстані $d(x_i, x_j)$; матриці відстаней D, де $D_{ij} = d(x_i, x_j)$; і функції зв'язку $L(c_i, c_j)$, де c_i і c_j - кластери (набори) точок в X. Потім ІАК виводить ієрархію кластерів, представлену історією злиттів кластерів. ІАК реалізується таким чином:

дані: $n \times n$ матриця D

результат: Послідовність злиттів $F = \{\}$;

$C = \{C_i \mid c_i = \{X_i\}, Y \leq n\}$;

поки $\mid F \mid < n$ **здійснити**

Знайти a і b з умови $L(a, b) = \min L(c_i, c_j)$;

Додати (a, b) в F;

Нехай $a = a \cup b$;

кінець

повернути F;

Алгоритм: Ієрархічна агломеративного кластеризація

Мета функції зв'язку - це визначення відстані між двома кластерами (наборами точок даних) з урахуванням існуючої функції відстані для окремих

точок даних. На практиці існує множина можливостей. Два найпростіших випадків - це повний зв'язок і одиночний зв'язок. Повний зв'язок визначає $L(c_i, c_j) = \max d(x_i, x_j), x_i \in c_i, x_j \in c_j$; одиночний зв'язок визначає $L(c_i, c_j) = \min d(x_i, x_j), x_i \in c_i, x_j \in c_j$. Для метрик і наборів даних, що обговорюються тут, не було помічено емпіричної різниці між цими двома параметрами, тому буде використовуватися простіший випадок одиночного зв’язку.

Подібно ІАК, спектральна кластеризація залежить тільки від матриці подібності. Спектр матриці подібності використовується для зменшення розмірності. Існує множина методів зменшення розмірності. Основна схема підходу показана в описі алгоритму.

Як і ІАК, так і спектральна кластеризація дозволяють використовувати різні метрики відстані. На практиці цільова кластеризація вимагає коригування або штрафування метрики відстані для обліку Λ-інформації. Однак як доказ концепції розглянемо відстань.

дані: Матриця подібності *S*, *k* областей
результат: Кластери C_1, ..., C_k
L - лапласіан *S*;
$e_1, \ldots, e_k$ - перші *k* власних векторів *L*;
кластеризуються точки y_i в $_k$ кластерів *C*;
повернути *C*;
Алгоритм: спектральна кластеризація

Відстань редагування, також відома як відстань Левенштейна, - це просто кількість правок (вставок, замін і вилучень), необхідних для перетворення однієї послідовності в іншу. Важливо відзначити, що на відміну, наприклад, від використання евклідової відстані в поданні мішка слов, відстань редагування зберігає інформацію про порядок. Це дозволяє ІАК і спектральній кластеризації з

відстанню редагування служити перевіркою попередніх припущень, а саме, чи є цільова кластеризація строго необхідноа для цих наборів даних або досить збереження інформації про послідовність.

Незбалансованість дендрограмми говорить про те, що дані можуть бути більш сильно зміщені на користь одного типу. В Набір 1 більшість послідовностей короткі і складаються в основному або з коротких спроб або довгі спроби, що створює дисбаланс. Використаємо нормалізовану відстань редагування, яка врівноважує кластери шляхом ділення відстані редагування для пари послідовностей на довжину найдовшої послідовності.

Однак, незважаючи на збалансовану кластеризацію, ІАК як і раніше показує не дуже хороші результати на тестових даних.

Ключовим моментом є те, що результати для тестового набору весь час менше нуля. Проте, відстань редагування є досить простою метрикою відстані. Зокрема було розроблено множину більш складних метрик відстаней і ядер послідовності.

Цілком ймовірно, що інші метрики відстані, що включають інформацію про послідовність, будуть настільки ж корисні, як і відстань редагування, а можливо, і краще.

Однак тут основна увага приділяється цільової кластеризації. При використанні парадигми метричного навчання алгоритми цільової кластеризації припускають, що існує метрика для зміни, наприклад, відстань редагування.

Потім вивчається новий метричний простір, отриманий з цієї вихідної метрики. Хоча в прикладах в якості початкової метрики буде використовуватися відстань редагування, що не є строго необхідним. Проте, відстань редагування має зручну властивістю бути L1-нормою над своїми операціями, будуть використовувати лінійні оператори, які призводять до прямих інтерпретацій для відстані редагування.

2.3 Навчання за допомогою шумових міток

Навчання з зашумленими мітками (також відоме як навчання з шумом класифікації) має на увазі навчання в присутності випадкового або систематичного шуму в наданих мітках і застосовується як до напівконтрольованого, так і до повністю контрольованого навчання. В цілому, це досить реалістична парадигма, оскільки як людські, так і автоматичні джерела міток схильні до помилок. Наприклад, в даних про освіту більшість міток дається або вчителями, або спостерігачами, які, незважаючи на значну підготовку, все одно схильні до людських помилок. Навчання з шумом класифікації, навіть випадковим, несистематичним шумом, є важким завданням. Наприклад, жоден алгоритм не може вивчити клас понять при достатньому рівні шуму, і навіть навчання напівпросторів в присутності шуму є NP-важкою задачею. Навіть невелика кількість шуму може привести до швидкого погіршення продуктивності в прикладних випадках. Це погіршення відбувається, незважаючи на те, що вдається уникнути в значній мірі виродженого випадку систематичного шуму.

Використовуючи парадигму навчання, цільові кластери λ можна розглядати як мітки. Наприклад, для даних про освіту коефіцієнт корисної дії кожного студента можна розглядати як мітку для всіх тактик навчання, які використовує студент. Тактикою навчання може бути, наприклад, багаторазове вгадування або запит всіх доступних підказок перед відповіддю. Звичайно, результати навчання є надзвичайно вагомими мітками для тактик навчання: хороші студенти, ймовірно, іноді використовують погані тактики, а погані студенти, ймовірно, іноді використовують хороші тактики. Таким чином, відображення значень λ на мітки не може бути ефективним рішенням для загальних задач кластеризації цілей.

Однак є кілька цікавих ідей, які можна почерпнути з літератури за мітками. Прикладом того, як взаємодіють ці різні класи задач, є дослідження з

навчання з зашумленими обмеженнями, які відносяться як до напівконтрольованої кластеризації, так і до навчання з зашумленими мітками.

2.4 Навчання за декількома фактами

Багатофакторне навчання визначається як навчання на пакетах прикладів, де якщо хоча б один приклад в пакеті позитивний, то весь пакет отримує позитивну мітку; в іншому випадку пакет отримує негативну мітку. Багатофакторне навчання являє собою цікаву структуру і відображає багато реальних проблем, починаючи від логічної атрибуції і закінчуючи рекомендаціями веб-сторінок . Для прив'язки навчання за декількома фактами до цільової кластеризації можна уявити, що кожен Λ є мітка для мішка прикладів. Наприклад, "позитивний" студент, який має високий попередній результат, буде представлений у вигляді пакету дій, але тільки деякі з цих дій будуть дійсно "позитивними", незважаючи на те, що студент позначений як "позитивний". На жаль, це занадто спрощене представлення можливих цілей кластеризації: наприклад, безумовно, існують відмінності між студентом, який використовує одну хорошу тактику навчання, але в іншому є поганим студентом, і студентом, який використовує тільки хорошу тактику навчання. Однак алгоритми навчання на кількох примірниках працюють в припущенні, що досить одного позитивного прикладу. Крім того, менше роботи було виконано на навчанні на багатьох примірниках з реальними мітками, ніж з дискретними мітками. Однак алгоритми навчання на кількох примірниках працюють в припущенні, що досить одного позитивного прикладу. Крім того, менше роботи було виконано на навчанні на багатьох примірниках з реальними мітками, ніж з дискретними мітками.

Однак один клас алгоритмічних рішень особливо цікавий як можливе рішення для цільової кластеризації. Різноманіття щільності намагається знайти точку *t*, яка близька хоча б до одного прикладу з кожного позитивного мішка.

Мінімізація помилки при різноманітті щільності розширює цю ідею, використовуючи точку t в якості початкової точки. На етапі очікування вибирається точка t_i з кожного позитивного мішка і, яка найбільш близька до точки t_i до центральної точки t. На етапі максимізації, він повторно оцінює t, використовуючи новий m_u. Аналогічний алгоритм може працювати для кластеризації цільових послідовностей. Наприклад, знайти одну ПММ M_t для класифікації всього набору послідовностей нескладно. Потім завдання навчання перетворюється в традиційний алгоритм кластеризації МО для ПММ, за винятком того, що центри кожного кластера визначаються близькістю до M_t, а не структурою самого кластера.

Варто зазначити, що для прикладів з незалежною вибіркою навчання за декількома обставинами є окремим випадком навчання з зашумленими мітками. Однак в цільових завданнях кластеризації вибірка навряд чи може бути незалежною: студенти діляться проблемами, подіями в класі і навіть погодою, і все це може вплинути на поведінку студентів.

2.5 Аналіз цільових факторів

Один конкретний метод заслуговує на особливу увагу: цільовий факторний аналіз (також відомий як цільове тестування). Розроблений для хімічних досліджень в кінці 70-х - початку 80-х років, цільовий факторний аналіз, по суті, використовує відомі емпіричні результати для перевірки та інтерпретації результатів факторного аналіз. Загалом, він включає в себе взяття цільового вектора t і вимір його відстані до підпростору, знайденого за допомогою розкладання за сингулярними значеннями. У хімії t зазвичай являє собою чистий спектр деякого з'єднання, де з'єднання має бажані властивостями при експериментальному спектральному досліджені. Наявність або відсутність різних цільових сполук t потім використовується для перевірки та інтерпретації даних хімічних датчиків.

Одним з варіантів цільового факторного аналізу є ітеративний цільовий факторний аналіз. В ітеративному цільовому факторному аналізі замість тестування методу збору даних за допомогою тестового вектора *t*, вихідні дані використовуються для поліпшення або пошуку відсутніх значень *t*. Ітеративне поліпшення виконується шляхом поновлення *t* новими прогнозами і використання передбачених значень в якості нових входів для алгоритму.

Кілька алгоритмів, розроблених для біологічних даних, схожі на алгоритми кластеризації цільових послідовностей або вирішують аналогічні завдання. Є підхід який використовує ієрархії дискримінаційних ПММ для виявлення високодискримінантних мотивів. Загалом, мотиви - це набір послідовностей з деякою загальною ознакою або походженням, наприклад, білкові послідовності, співставлені з певними біологічними функціями. Вони зазвичай представлені деяким розподілом ймовірності по символам на кожному індексі в серії, причому точні деталі представлення мотивів залежать від припущень дослідника. Підхід може знаходити високодискримінантні мотиви в наборах білкових послідовностей. Однак, на відміну від алгоритмів кластеризації цілей, описаних в даному документі, підхід вимагає заздалегідь визначених наборів різних класів.

Наприклад, в просторі наборів освітніх даних такий підхід може бути використаний для пошуку однієї ПММ, що представляє мотив, характерний для хороших студентів, і однією ПММ, що представляє мотив, характерний для поганих студентів, причому ці дві ПММ забезпечують високу ступінь дискримінації між двома класами. З деякими змінами підхід може бути адаптований для визначення набору високодискримінантних ПММ для кожного класу. Однак це показує два ключові відмінності між цим підходом і алгоритмами кластеризації цілей: по-перше, такий підхід базується на існуванні дискретних класів, а алгоритми кластеризації цілей призначені для роботи без окремих класів; по-друге, алгоритми кластеризації цілей працюють з

можливістю, відкинути шаблони. Але корисні патерни можуть, насправді, бути загальними як для хороших, так і для поганих студентів.

Розглянемо студента, який з першої спроби правильно вирішує завдання без використання підказок: така поведінка зазвичай вважається хорошим і свідчить про розуміння і майстерність. Однак як хороші, так і погані студенти зроблять кілька кроків правильно з першої спроби. В цілому, обидва підходи призначені для вирішення принципово різних завдань, але описаний далі діскрімінативний алгоритм ПММ подібний до кластеризації цілей і є загальний як для хороших, так і для поганих студентів.

2.6 Алгоритм мінімізації помилки на прихованих Марківських моделях

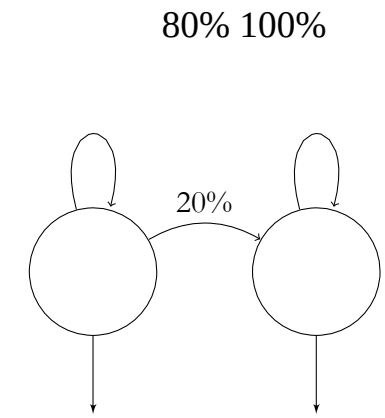
Приховані Марківські моделі (ПММ) представляють собою основну генеративну модель, яка використовується в даній роботі. ПММ були вибрані тому, що вони забезпечують хороше представлення тактик навчання. Розглянемо наступний простий приклад тактики навчання:

Студент запитує підказки швидко, багаторазово, поки репетитор не надасть рішення. Потім студент вводить рішення. Виходячи з цього прикладу, тактику навчання можна узагальнити до спостережуваної, передбачуваною і повторюваної моделі поведінки, яка досить абстрактна, щоб включати кілька спостережуваних інстанцій. ПММ є особливо вдалим поданням тактики навчання, оскільки вона включає в себе набір неспостережуваних станів, кожний з яких пов'язаний зі спостереженнями за допомогою розподілу ймовірностей.

Для даних "студент-викладач" спостереженнями є дії студента. На рисунку 7 показаний приклад ПММ, який аналізує концепцію студента, швидко переглядає доступні підказки перед відповіддю.

Кожний не спостережуваний стан представлено областю; кожна стрілка між станами або повернення до стану являє собою перехід; число над стрілкою -

це ймовірність переходу. У таблицях під станами показані ймовірності спостереження дії. Коли ПММ виробляє символ дії, говоримо, що він випускає цей символ.



с	**С**	п	**П**
0	0	8	2

с	**С**	п	**П**
5	5	0	0

Рисунок 2.1 - Приклад ПММ – для випадку спроба, підказка.

Нехай ряд спостережуваних дій є послідовністю. Послідовності можуть бути визначені або для всіх дій в завданні, або для всіх дій на кроці.

З огляду на набір послідовностей студентів, пов'язаних з ПММ, алгоритм Баума-Уелча може заново вивчити параметри цієї ПММ, щоб краще відповідати спостережуваним даними. Баума-Уелча - це стандартний метод навчання параметрів однієї ПММ.

2.7 Кластеризація максимізації очікування для прихованих Марківських моделей

Нехай кожна окрема ПММ представляє одну тактику навчання. Для виявлення тактик навчання необхідно виявити набори ПММ. Нехай набір ПММ називається колекцією. У колекції спостерігається послідовність дій класифікується за найбільш вірогідною ПММ, яка її генерує.

Наприклад, розглянемо ПММ, які називаються відповідно "Вгадування" і "Повторні спроби". Імовірність того, що послідовність довжини 2, згенерувала "Повторними спробами", буде Сс, дорівнює50%, Що є дуже високою ймовірністю; однак імовірність того, що така послідовність буде згенеровано "вгадування", дорівнює100%.

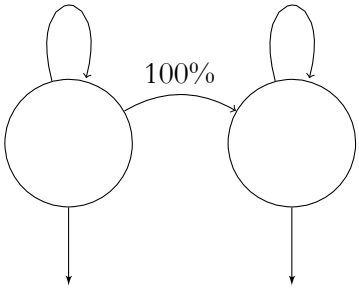
Таким чином, Сс буде від контрольної роботи на до "вгадування", а не до "Повторні спроби". Цей процес класифікації призводить до розбиття множини послідовностей, причому кожне розбиття відповідає одній ПММ. Таким чином, кожен розділ включає всі спостережувані приклади такої тактики.

Алгоритм Баума-Уелча може дізнатися параметри тільки для одного ПММ, але алгоритми кластеризації можуть дізнатися набори ПММ, а значить, і набори тактик. Звичайна мета алгоритму кластеризації ПММ полягає в максимізації загальної правдоподібності для генерації спостережуваних послідовностей.

Цей тип проблеми історично вирішувалося за допомогою алгоритмів Максимізації очікування (МО), і для кластеризації ПММ, враховуючи початковий набір ПММ, одна ітерація алгоритму кластеризації МО є:

- очікування: Призначити кожну послідовність ПММ, яка з найбільшою ймовірністю її згенерує;
- максимізація: Для кожного ПММ заново обчислити його параметри за допомогою методу Баума-Уелча.

0% 100%

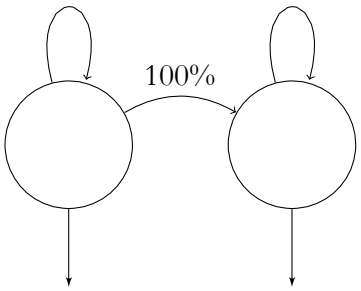


с	**С**	п	**П**
0	10	0	0

с	**С**	п	**П**
10	0	0	0

Рисунок 2.2 - Приклад ПММ - "Вгадування"

0% 100%



с	**С**	п	**П**
0	10	0	0

с	**С**	п	**П**
5	5	0	0

Рисунок 2.3 - Приклад ПММ - "Повторні спроби"

Цей процес починається з початкових ПММ і повторюється до тих пір, поки не буде досягнута умова припинення, наприклад, коли в результаті ітерації буде перекваліфікували менше 10 послідовностей. Колекція, вивчена цим алгоритмом, добре підходить до даних, якщо ймовірність генерації спостережуваних послідовностей висока. Цей алгоритм, далі в тексті МО-ПММ, гарантовано сходиться до локального мінімуму. Крім того, МО-ПММ ніколи не змінить кількість ПММ в колекції (k) або кількість станів на ПММ (n); зміняться

тільки параметри (і, отже, розбиття даних). Більш формальний опис наведено в алгоритмі 1.

Алгоритми кластеризації на основі МО ПММ використовувалися раніше в багатьох випадках, для розпізнавання слів. Хоча існують більш нові варіанти, більшість кластеризації ПММ все ще виконується за допомогою оригінального алгоритму Рабінера. Особливо показовим було дослідження, проведене Шліп і ін. Для аналізу експресії генів.

Оскільки алгоритм Баума-Уелча сам по собі є алгоритмом МО, кластеризація МО-ПММ є вкладені алгоритми МО.

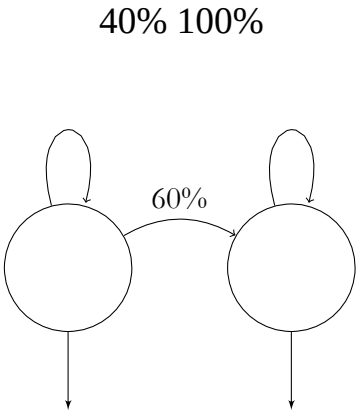
дані: Початкові ПММ і дані X
результат: ПММ
поки що *критерії завершення не задоволені,* **зробити**
$Mt = \{\}$;
$\forall i \leq |M|, Xi = \{X \mid i = \text{argmin}_j\, l(x \mid m_j), m_j \in M\}$;
для $mi \in M$ **зробити**
mi = Baum-Welch(m_i, X_i); $M_t = M_t \cup \{M_t\}$;
кінець
$M = M_t$;
кінець
return (M);

Алгоритм: МО ПММ

ПММ, описані досі, не створюють послідовності невідомої довжини; скоріше, довжина послідовності є параметром для генерації послідовності ПММ. Наприклад, можна запитати у ПММ випадкову послідовність довжини 20, але щоб отримати випадкову послідовність довжини менше або дорівнює 20, необхідно встановити пріоритет для ймовірності кожної довжини. Однак замість

того, щоб застосовувати перетворення для стандартизації довжини всіх послідовностей дій студентів, ПММ буде дозволено проводити спеціальні термінальні символи (позначаються тут ω). Наприклад, розглянемо ПММ, показана на рисунку 2.4.

Примітно, що на рисунку 2.4 відсутність термінального символу означає, що навіть після переходу в стан-пастку (стан 2, з якого не відбувається ніяких переходів).



с	**С**	п	**П**
0	10	0	0

с	**С**	п	**П**
10	0	0	0

Рисунок 2.4 - Приклад ПММ без термінального символу

Висновки до розділу

Система ПММ може бути протестована за допомогою тієї ж експериментальної процедури, яка використовувалася для мішка слів, тільки без перетворення в мішку слів. Результати навчання знаходяться в розумному діапазоні 0,3-0,4 для скоригованого R2, але результати тестування часто гірші, ніж простий вибір середнього значення. В певних випадках прихований тестовий

набір знаходиться всередині даних студентів, що робить результати навчання недостатніми для використання. Цей результат справедливий для всіх наборів даних. МО ПММ створює моделі, які добре працюють на навчальних даних, але не завжди працюють на прихованих тестових даних. Однак той факт, що наївний алгоритм МО ПММ може досить добре підходити до навчальних даних, є багатообіцяючим. В якості пропозиції виступає модифікація алгоритму, який може навчати і вибирати ПММ з урахуванням цільової інформації, наприклад, результатів навчання.

Розділ 3
Розробка системи кластеризації послідовностей

3.1 Модель системи кластеризації послідовностей

Інтуїтивно основна ідея, що лежить в основі напрямних, проста. Алгоритми МО страждають від локальних мінімумів, і тому для ефективного дослідження простору пошуку необхідно перезапускати зі випадкових початкових точок. Напрямні пропонують альтернативу. У кожному локальному мінімумі, замість випадкового перезапуску алгоритму МО, алгоритми використовують модель-кандидат плюс інформацію про мету для розумного вибору нової початкової точки в просторі пошуку.

На практиці точна реалізація орієнтування буде залежати від доступної інформації про мету Λ, деяких параметричних припущень про функції $f(X) \rightarrow Y$ і $g(Y) \rightarrow \Lambda$, а також від розміру і структури потенційного простору пошуку. Однак для видів цілей, які використовуються в освіті, і для простору моделей, включаючи ПММ, і припускаючи, що МОПМ реалізує стандартний алгоритм кластеризації Е- М ПММ, а Select () реалізує алгоритм вибору моделі, така реалізація показана в алгоритмі.

дані: Початкові ПММ, дані X і цілі Λ

результат: ПММ

поки що *критерії завершення не задоволені,* **зробити**

Mt = МОПМ (M, X);

M = Select (Mt, Λ);

кінець

return (M);

Алгоритм : МО ПММ

Підпрограма Select () може бути реалізована кількома способами. Найбільш простий підхід полягає у використанні покрокової лінійної регресії. Швидкість і простота покрокової регресії дозволяє витратити більше часу і обчислень на відносно більш цікаві базові ПММ.

Крім того, хоча лінійна регресія є досить суворим параметричних поданням, в самих ПММ існує значна гнучкість моделювання, і лінійна регресія виявилася досить ефективною в різних освітніх контекстах.

Також корисно передбачити додаткове зміщення в бік більш простих моделей і більш простих колекцій. Це допомагає уникнути надмірної підгонки і робить результуючі колекції більш інтерпретованими.

Для цього можна починати з простих моделей і замінювати тільки тоді, коли нові моделі виявляються значно краще. Таким чином, на практиці відбувається ітерація ПММ:

1. Початок з набору ПММ M, набору послідовностей X і набору високорівневих цілей Λ.

2. Призначити кожну послідовність $x \in X$ на ПММ $M_i = \text{argmax}_j\, l(x \mid M_j)$.

3. Перенавчання параметрів кожної ПММ для максимізації правдоподібності на відповідному розбитті.

4. Створити матрицю S так, щоб кожен стовпець S_i був розподілом для студента i всіх послідовностей цього студента по ПММ. Таким чином, S_{ij} буде ймовірність того, що студент i створив послідовність дій, зв'язану з j-м ПММ.

5. Вирішити регресію $\Lambda = S\beta + E$.

6. Нехай C - це множина всіх стовпців S, які вважаються значущими (є статистично значущими предикторами виграшу в навчанні). Тоді визначимо $M_t = \{M_i \mid m_i \in M \wedge i \in C\}$.

Після певної кількості ітерацій ПММ або після досягнення деяких критеріїв збіжності, збільшити кількість ПММ і станів. Тут застосовні

стандартні критерії збіжності: Δ належність до кластеру, Δ логарифмічна правдоподібність і т.д.

Нарешті, існує проблема вибору найкращої колекції. Найбільш ефективний спосіб зробити це - за допомогою прихованого валідаційного набору. Теоретично перехресна валідація краще, але приховування декількох останніх послідовностей досить добре працює на практиці.

Загальний протокол для всіх інших алгоритмів, про які піде мова, виглядає наступним чином:

– Ітеративне навчання колекцій на основі навчальних даних.

– Вибрати колекцію з найвищою скоригованої метрикою R2 на основі прихованих валідаціонних даних.

– Протестувати обрану колекцію на окремих тестових даних, оцінюючи колекцію, показники діяльності компанії з використанням нескоректованого R2.

Використання скоригованого R2 в якості метрики відбору створює зсув в бік більш простих колекцій. Інтуїтивно зрозуміло, що МО-ПММ має два способи зміщення результатів у бік простоти.

По-перше, параметри покрокової регресії можуть і повинні бути встановлені так, щоб приймати колекції з певною ймовірністю (наприклад, $p \leq 0{,}1$), але відкидати колекції тільки в тому випадку, якщо p-значення досить велике (наприклад, $p \geq 0{,}3$).

Така параметризація зміщує ПММ в бік збереження старіших ПММ, які, як правило, простіші (двоскладові ПММ пробуються раніше, ніж трискладові і т.д.). Крім того, використання скоригованого R2 для вибору кращої колекції зміщує в сторону колекцій з меншою кількістю ПММ в цілому.

Комбінація цих двох упереджень, особливо для більш пізніх алгоритмів (з деякими коригуваннями), призведе алгоритми до простих колекцій з декількома ПММ, кожна з яких має всього кілька станів.

3.2 Структура системи кластеризації

Симулювання даних дозволяє порівнювати моделі і алгоритми в умовах відомого істинного розподілу. Для кластеризації цільових послідовностей за допомогою ПММ це дозволяє досліджувати кілька ключових питань:

– Для яких типів колекцій ПММ алгоритми цільової кластеризації можуть відновити вихідні (породжують) моделі.

– Чи можуть алгоритми кластеризації цілей відновити типи ПММ, про які йшлося в гіпотезі експертами в галузі освіти.

– Які алгоритми кластеризації цілей відновлюють які типи ПММ та наскільки ефективно.

– Коли алгоритми кластеризації цілей не відновлюють вихідні ПММ, як вони себе ведуть.

Питання порівняння алгоритмів (наприклад, третє питання вище), а також основні результати моделювання для кожного алгоритму будуть включені в обговорення окремих алгоритмів. Теми цього розділу такі: питання моделювання даних з ПММ, результати відновлення колекцій, заданих людиною, особливо створених експертами в цій галузі або на основі існуючих моделей студентів; дослідження умов, необхідних для відновлення випадкових колекцій ПММ; спостерігаються відмінності в поведінці алгоритмів цільової кластеризації, особливо МО, що породжують колекції.

На рисунку 3.1 показана експериментальна процедура, яка використовується з моделювати даними. Кілька кроків є новими в порівнянні з попередніми експериментальними процедурами. По-перше існує неявна вимога визначити скільки породжуючих ПММ використовувати, скільки студентів моделювати і скільки дій вибрати для кожного студента. По-друге розподіл ПММ (вектор ймовірностей для кожного симулятора) є неявним.

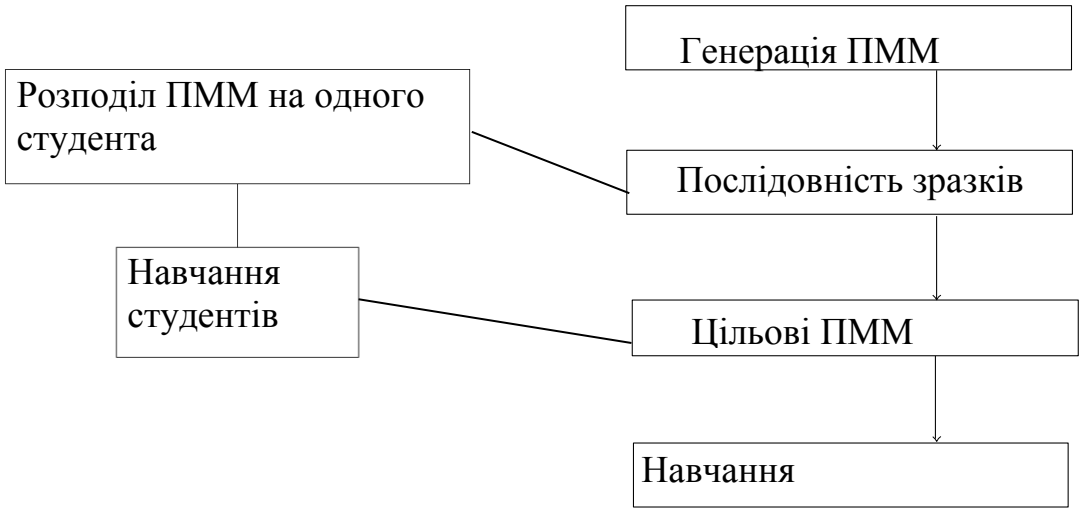


Рисунок 3.1 - Експериментальна процедура, використана для моделювання даних

На практиці і те, і інше може бути згенеровано випадковим чином. По-третє, для генерації фактичних покрокових даних потрібно як розподіл по ПММ (щоб визначити, коли робити вибірку з кожної ПММ), так і самі ПММ. Виграш в навчанні виводиться з розподілу по ПММ, таким чином, умовно незалежний від обраних послідовностей. Такий поділ між процесами генерації ПММ і посилення навчання, а також шум, властивий вибірці з ПММ, призводить до того, що найкраща можлива ефективність прогнозування на змодельованих даних виявляється не такою ефективною, навіть при використанні оригінальних генеруючих моделей.

Пряме порівняння структур ПММ не тільки складно, але і той факт, що кожна ПММ існує в зв'язку з іншими ПММ, означає, що структура однієї індивідуальної ПММ не може однозначно визначити послідовності в асоційованим з ними кластері. Наприклад, візьмемо пару ПММ, одна з яких (*M1*) генерує послідовності, що складаються з A, а інша (M2) - з a M1 може мати будь-яку ймовірність генерації a і A за умови, що P (A | M1) > P (A | M2) і P (a | M1) <P (a | M2). Це означає, що навіть якщо ймовірність того, що M1 має

49%, то генеруючі a послідовності все одно можуть бути присвоєні *M2*, а не *M1*. Цей же аргумент можна поширити на довільно складні структури ПММ і алфавіти.

Проста колекція легко піддається машинному навчанню ПММ, яка визначається людиною, повинна включати окремі, які незначно перекриваються моделі, що представляють найбільш гіпотетично ймовірні моделі поведінки студентів. Колекція включає в себе:

– Модель "Правильно" ця ПММ генерує короткі послідовності в основному довгих правильних спроб.

– Модель “Вгадування” ця ПММ генерує довгі послідовності в основному коротких неправильних спроб, за якими слідує правильна спроба.

– Модель “Підказка” ця ПММ генерує довгі послідовності в основному коротких запитів підказки, за якими слід правильна спроба.

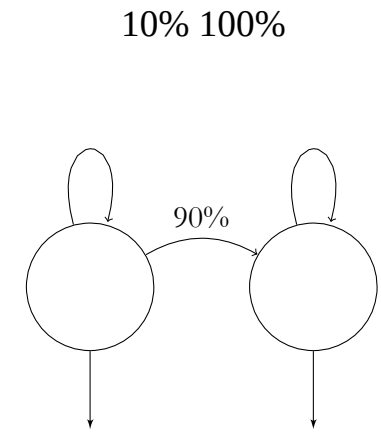


Рисунок 3.2 - Базова колекція - Проста

Це породжує колекцію, яка не містить моделі правильного використання підказок, оскільки не існує спільної угоди про правильну поведінку підказок. Ця

колекція може бути тривіально адаптована до будь-якого набору емісійних символів (наприклад, {A, H}, {I, C, H}).

Колекції відсортовані так, що "кращі" колекції ПММ розташовані зліва, і для цих колекцій коефіцієнт помилкової класифікації близький до нуля, показуючи, що всі колекції близько відображають породжуючу колекцію. Це говорить про те, що для простих колекцій МО надійно працює на рівні або близькому до верхньої границі. Варто також відзначити, що МО знаходить клас схожих колекцій, які є дзеркальним відображенням породжуючих колекцій, але за межами цього класу продуктивність знижується.

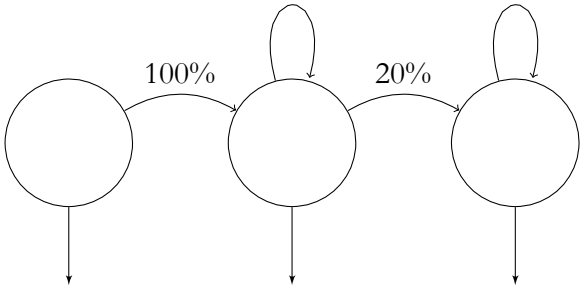
На відміну від простої колекції, колекція базова має значне перекриття між послідовностями, представленими кожною з її моделей. Це перекриття відбувається в двох формах: внутрішнє перекриття, обумовлене природою конструкцій, і шум, неконтрольною заміною структурних нулів на малі ймовірності очікування. Характерне перекриття являє собою фактичну невизначеність про предмету область, в той час як шум є просто природним побічним продуктом реальних явищ. Базова колекція включає в себе:

– Базова правильна модель: ПММ генерує короткі послідовності переважно довгих правильних спроб. Вона відрізняється від простої моделі тим, що в ній присутній шум. Це дає можливість внутрішнього перекриття, оскільки тепер вона може генерувати послідовності, схожі на послідовність вгадування. Тим не менш, вона як і раніше сильно зміщена в бік генерації A.

– Модель базового вгадування: ПММ генерує довгі послідовності в основному коротких неправильних спроб, за якими слідує правильна спроба. Вона відрізняється від моделі "вгадування" тим, що існує ймовірність видачі послідовностей, подібних до тієї, що була в моделі "Базова правильна", і у всіх станах без завершення присутній шум.

– Базова модель підказка: ПММ генерує довгі послідовності переважно коротких запитів підказки, за якими слідує правильна спроба. Вона відрізняється від простої моделі тим, що в її викиди вноситься шум.

90% 100%



с	**С**	п	**П**	ω
5	8	5	2	0

с	**С**	п	**П**	ω
3	3	2	2	0

с	**С**	п	**П**	ω
0	0	0	0	10

Рисунок 3.3 - Базова колекція - Вгадування

3.3 Опис моделювання

Наведені вище результати моделювання призводять до деяких важливих наслідків. По-перше, існують породжуючі колекції ПММ, для яких МО-ПММ є ефективним алгоритмом, здатним відновити наближення вихідного розподілу. По-друге, також існують породжуючі колекції (наприклад рандомні), цільова ПММ буде знаходити колекції, які явно не відповідають даним.

Таблиця 3.1 - Підгонка Набір 1 - дані випробувань

τ	**Рівень проблем (**R2**)**	**Ступінчастий рівень (**R2**)**	**Кількість ПММ**	**Максимальна кількість станів**
5	0.21	0.17	6	3
7	0.20	0.22	5	5
10	0.24	0.23	4	4

Обмеження випадкових ПММ для отримання ПММ, які піддаються навчанню за цілями, є складним завданням. Наприклад, недостатньо обмежити тільки матриці викидів. Ці проблеми дозволяють вважати, що якщо ПММ знаходить хорошу колекцію в емпіричних даних, значить, в цих даних є реальна структура, яка ідентифікується.

Важливо щоб цільові алгоритми кластеризації могли знаходити моделі передбачення в реальних наборах даних. У першому застосуванні МО-ПММ, виявили, що метод МО для ПММ був набагато більш ефективний в пошуку передбачення (і в деякій мірі інтерпретованих моделях) моделей студентів, ніж проста кластеризація МО ПММ. У таблиці 3.1 наведені результати первинного тестового набору. У таблиці τ є порогові значення. Примітно, що в тому дослідженні використовувався однопороговий метод, де поріг був один для всіх дій, а дії були згруповані тільки в дві категорії: спроби та підказки.

Не менш важливо і те, що результати виявилися хоча б в деякій мірі інтерпретованими. По-перше, 3, виявлені колекції є простими, що містять 4-5 ПММ не більше 3-4 станів кожна. Крім того, фактичні ПММ легко інтерпретувати. Це, очевидно, модель для вгадування, так як вона зазвичай генерує одне С, перш ніж перейти в другий стан і генерувати в основному с. Це відповідає тому, що студенти швидко вводять багато неправильних відповідей поспіль. Вгадування є елементом азартних ігор і невдалого пошуку допомоги, які також негативно пов'язані з навчанням, тому зв'язок моделі з навчанням також обгрунтовано.

У зв'язку з цим доцільно також визначити, чи є скоригована статистика R2, розрахована на даних валідації, хорошим підґрунтям ефективності тестових даних. Для цього кореляція між скоригованими R2 на валідних даних і результатом R2 на тестових даних становить 0,93, усереднена за багатьма ітераціями. Незважаючи на те, що іноді допускаються помилки, в довгостроковій перспективі, протягом багатьох перезапусків, скоригований R2, розрахований на даних валідації, є хорошим підґрунтям продуктивності тестового набору.

Первинний аналіз був надмірно спрощений по ряду причин. Він не брав до уваги коректність або перемикання кроків. Не взято до уваги проблеми стабільності результатів, МО відомий тим, що для досягнення гарної продуктивності потрібно багато випадкових перезапусків. Ігнорується можливість того, що різні дії повинні мати різні порогові значення. Не враховувалася проблема відмінностей між високопродуктивними колекціями. Дві ПММ часто повторювалися у всіх колекціях, але інші ПММ відрізнялися між собою.

3.4 Обмеження на моделі

ПММ створює прості, інтерпретовані моделі, які як передбачають навчання студентів, так і групують поведінку. Однак алгоритм має ряд обмежень: він знаходить локальні мінімуми, повільно сходиться і залежить від випадкових насіння. Перше обмеження, що полягає в тому, що знаходить локальні мінімуми, є спільною проблемою для МО алгоритмів. ПММ гірше, ніж більшість алгоритмів МО, через складність простору пошуку та через те, що він використовує вкладені алгоритми.

При виборі "кращої" колекції з використанням скоригованого показника R2 на перевірочних даних, "краща" колекція має R2 на тестовому наборі 0,30. Однак мінливість результатів окремих запусків говорить про те, що для впевненого знаходження гарного рішення необхідно багато перезапусків.

Для оптимізації можна зменшити кількість поганих кандидатів, ввівши інформацію про ціль для параметрів ПММ. ПММ сходиться повільно, що є результатом вкладених циклів МО. Однак швидкість збіжності МО можна поліпшити, замінивши всього один цикл. Наприклад, якщо для навчання ПММ використовувати інший алгоритм, то можна уникнути значної частини ітеративной оптимізації.

Останнє обмеження, що полягає в тому, що існує сильна залежить від випадкових початкових даних і тому вимагає багаторазових перезапусків, є прямим наслідком процесу вибору моделі . Кожна ітерація зовнішнього циклу кластеризації призводить до кроку вибору моделі, де деякі ПММ видаляються. Після цього вони замінюються новими, випадковими ПММ. Ця повторювана рандомізація робить визначення потрібної кількості ітерацій дуже складним.

Алгоритм МО-ПММ перебирає mxk-матрицю можливих чисел ПММ (m) і можливих чисел станів (k). Однією простою зміною оригінального алгоритму є обмеження діапазону можливих ПММ і станів нижньої трикутною матрицею, що запобігає випадкам, коли окремі ПММ мають більше станів, ніж ПММ.

Крім того, версія МО-ПММ з обмеженням за станом більш ніж удвічі скорочує обчислювальні витрати. Для всіх наступних варіантів ця модифікація буде прийнята.

3.5 Максимізація очікування дискримінантної і зваженої ПММ

Стандартні приховані марківські моделі є генеративними: вони вивчають модель з високою ймовірністю генерації спостережуваних позитивних прикладів. Цей підхід ефективний для окремих ПММ, але для колекцій ПММ стандартне генеративне рішення тільки досягає оптимального рішення, якщо:

– Існує нескінченна кількість навчальних даних.

– Алгоритми локальної оптимізації для кожної ПММ об'єднуються для знаходження спільного рішення максимального правдоподібності.

– ПММ є точним поданням вихідної породжуючої моделі.

Жоден з цих критеріїв не відповідає дійсності на практиці. Перша проблема, пов'язана з навчальними даними, є стандартною для оптимізації параметрів за допомогою МО. Однак дві інші проблеми складні для МО-ПММ: пошук кластерів ПММ включає в себе вбудовування пошуку МО в інший пошук

МО, що посилює другу проблему, і припущення ПММ не є ідеальним представленням людського пізнання і навчання, що посилює останню проблему.

Основна причина цього обмеження полягає в тому, що при використанні в алгоритмі кластеризації МО кожен ПММ знає тільки про послідовності в своєму розділі і не має ніякого іншого значення.

Для кожної ПММ, припускаючи, що $x \in P$ являє собою послідовність у разі необхідності розділення для k-го ПММ, об'єктивна функція має вигляд:

$$\sum_i^N lnP(x_i|M_k), \tag{3.1}$$

де $N = |P_k|$

Однак існує альтернативний підхід в тому, що дискримінантні ПММ використовують в навчанні як позитивні, так і негативні приклади. Одним з простих алгоритмів дискримінативность навчання є алгоритм оцінки максимальної взаємної інформації (MBI).

У загальному випадку, для n послідовностей x_i і n відповідних ПММ-завдань m_i, MBI оптимізує об'єктивну функцію.

$$\sum_i^n ln\frac{P\left(x_i|M_{m_i}\right)P\left(M_{m_i}\right)}{\sum_m P(x_i|M_m)P(M_m)} \tag{3.2}$$

Оптимізація об'єктивної функції MBI максимізує ймовірність послідовності за умови, що вона породжена найбільш імовірною ПММ, але також мінімізує ймовірність послідовності за умови, що вона породжена всіма іншими ПММ. Це дискримінує процес навчання, в якому негативні приклади можуть сприяти навчанню параметрів. На відміну від базового алгоритму МОПММ, який максимізує ймовірність для P (x_i | M_k), MBI максимізує апостеорну ймовірність того, що колекція буде зібрана, враховуючи дані.

На практиці значення P (M) є критичною точкою. Хоча можна оптимізувати функцію MBI, де $P(M)$ замінюється емпіричної оцінкою $P(M)$, На практиці це працює погано. Натомість $P(M)$ зазвичай оцінюється з яких-небудь зовнішніх даних. Наприклад, MBI найбільш часто використовується в розпізнаванні мови, де $P(M)$ часто оцінюється за базовою мовної моделі.

Якщо P(M) замінити на $\hat{P}$(M), то об'єктивна функція MBI буде прямою заміною стандартної функції правдоподібності. Однак можна включити Λ-інформацію в об'єктивну функцію безпосередньо, що дозволяє вбудувати мету в алгоритм на більш низькому рівні. Визначимо для кожної послідовності x_i, $G(x_i)$ як нормалізований виграш в навчанні для студента, пов'язаного з послідовністю i. Далі, визначимо $G(M, x_i)$:

$$G(M, x_i) = \begin{cases} G(x_i) \text{ якщо } M \text{ позитивно асоціюється з навчанням} \\ 1 - G(x_i) \text{ якщо } M \text{ негативно асоціюється з навчанням} \end{cases} \quad (3.3)$$

Основна ідея полягає в тому, щоб замінити P(M), яка зазвичай є апріорною ймовірністю, подібною, знайденою в даних про освіту. $G(M, x_i)$, являє собою зашумлену оцінку P(M). На відміну від розпізнавання мови, де існує відома ймовірність для кожної моделі (слова), набори даних про освіту не мають еквівалентної апріорної ймовірності. Найближчим еквівалентом в цих наборах даних є оцінка ймовірності кожної тактики навчання з урахуванням того, що відомо про те, як ця тактика впливає на навчання. Ця нова цільова функція дозволяє надати більшої ваги спостережуваним послідовностям, що відбувається з більш екстремальних ситуацій.

Узагальнимо цей підхід таким чином - візьмемо базовий алгоритм МО ПММ, замінимо звичайну функцію цілі. Нехай цей алгоритм буде відомий як зважений МО ПММ. Коефіцієнт об'єктивної функції зваженого ПММ можна взяти так само, як і коефіцієнт в об'єктивної функції діскримінативного ПММ, але тут більше немає знаменника, який дозволяє оптимізувати весь набір

моделей. Однак обидва алгоритми дозволяють безпосередньо включати значення Λ в алгоритм пошуку шляхом перезважування кожної ймовірності на ймовірність того, що відповідний студент виявить здатність до навчання. Таким чином, ПММ, позитивно пов'язані з навчанням, будуть виділяти послідовності в розділі, які часто використовуються студентами з високими результатами навчання, і навпаки для ПММ, негативно пов'язаних з навчанням.

Перевага зваженої версії над повністю дискримінаційної версією полягає в наступному. Зважені ПММ як і раніше розрізняє позитивні і негативні приклади, але на відміну від МО ПММ дискримінативного, вона розрізняє (прогнозує) поведінку навчання, а не по тому, чи належить вона до певної моделі.

3.5 Результати моделювання

На рисунку 3.4 показана продуктивність зважених і простих ПММ на базовій колекції для різних обсягів даних. Тут y - тестовий R2, а вісь x чергується між набором результатів для простого і набором результатів для зваженого. Суттєвої різниці немає, іноді зважений показник трохи перевершує незважений, і навпаки.

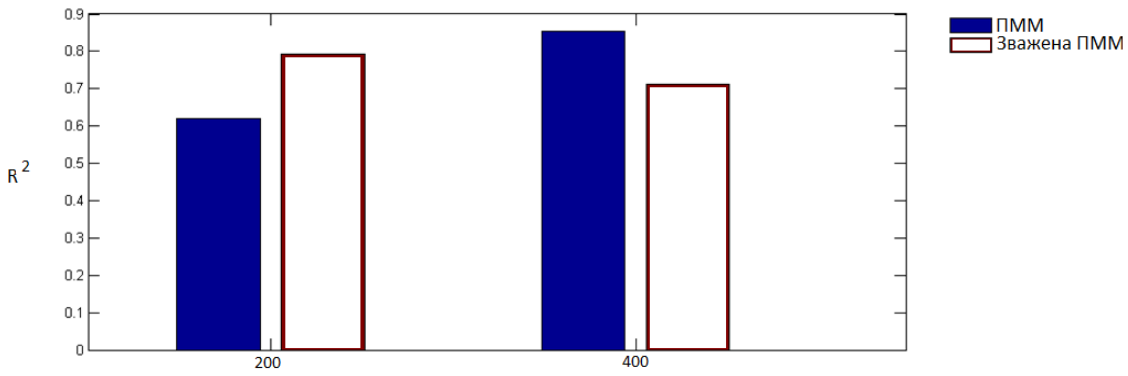


Рисунок 3.4 - Порівняння максимальної відповідності умовам для зважених і простих ПММ на різних наборах змодельованих даних

Навіть якби це був остаточний результат, зважена ПММ була б поліпшенням, так як зважене рішення обмежена станом, що дозволяє йому навчатися швидше, а саме зважування змушує окремі ПММ сходитися ще швидше.

Пряме порівняння структур ПММ може бути непродуктивним, оскільки ці порівняння не тільки важко визначити і обчислити, але і той факт, що кожна ПММ існує по відношенню до інших кластерів ПММ, означає, що структура однієї окремої ПММ неоднозначно визначає послідовності, які вона класифікує. Наприклад, візьмемо пару ПММ, одна з яких (*M1*) генерує послідовності, що складаються з C, а інша (*M2*) генерує послідовності, що складаються з c. *M1* може мати будь-яку довільну ймовірність генерації c і C до тих пір, поки оскільки P (*C* | *M1*)> P (*C* | *M2*) і P (*c* | *M1*) <P (*c* | *M2*). Цей же аргумент може поширюються на довільно складні структури ПММ.

Розглянемо наступне визначення неправильної класифікації: якщо дві ПММ мають великі розбіжності при застосуванні міток до спостережуваних послідовностей, то вони несхожі; однак якщо дві ПММ згодні з мітками для всіх спостережуваних послідовностей, то будь-які структурні відмінності між ними несуттєві для практичних цілей.

На жаль, відмінності в кількості ПММ в кожній колекції ускладнюють це порівняння. Наприклад, можливо, що колекція з 5 ПММ і колекція з 2 ПММ майже ідентичні.

Назвемо ці колекції A і B. ПММ A1 може відповідати ПММ B1, а інші чотири ПММ A2, A3, A4 і A5 можуть бути підпорядковані B2. Тоді, в межах якої складності моделей, A і B є однією і тією ж колекцією. Цей приклад ілюструє дві основні труднощі при обчисленні коефіцієнта помилкової класифікації: мітки для ПММ можуть бути довільно переміщені між колекціями, і в одній колекції може бути більше ПММ, ніж в інший.

Без втрати загальності припустимо, що A - це колекція з великою кількістю ПММ, а B - колекція з меншою кількістю. Нехай S_i - це набір послідовностей, призначених і-й ПММ в колекції A, і нехай L (B, S_i) - це найбільш поширена мітка, що застосовується колекцією.

Таблиця 3.2 - Рівень взаємної помилки класифікації для 5 кращих колекцій

	колекція 1	**колекція 2**	**колекція 3**	**колекція 4**	**колекція 5**
колекція 1	0	0.0800	0.0930	0.0589	0.0603
колекція 2	0.0800	0	0.0640	0.0845	0.0690
колекція 3	0.0930	0.0640	0	0.0895	0.0954
колекція 4	0.0589	0.0845	0.0895	0	0.0943
колекція 5	0.0603	0.0690	0.0943	0.0724	0

Якщо говорити про інтерпретацію, то рішення формули регресії означає, що більше значення надається парам послідовностей, створених більш різними типами студентів, і, таким чином, формула надає більшої ваги операції, якщо її наявність передбачає різницю в навчанні студентів.

Необхідно відзначити два важливих застереження. По-перше, можливо, що коефіцієнти β будуть негативними, тоді послідовності C і Ссссс матимуть негативний відстань. Негативні відстані особливо проблематичні, тому що відстань між C і C все одно дорівнюватиме нулю, а значить, більше, ніж негативна відстань. Одне з можливих рішень - взяти абсолютне значення відстані.

Це рішення суперечливе, велика негативна вага β передбачає, що операція редагування насправді покращує шанси того, що дві послідовності належать схожим типам студентів, в той час як абсолютне значення передбачає зворотне; проте це рішення добре працює на практиці і перемогло інші варіанти в серії випробувань.

Абсолютне значення, ймовірно, ефективно, тому що більшість негативних β мають невеликі абсолютні значення.

Алгоритм: Переоцінка лінійної відстані

дані: Послідовності X, Дані G

Обчислити $V_{i,j}$ для кожної пари різних послідовностей (X_i, X_j);

Обчислити $Y_{i,j}$ для кожної пари різних послідовностей (X_i, X_j);

Вирішити $Y = \alpha + \beta V + E$;

Обчислити D, де $D_{i,j}$ - відстань між різними послідовностями (X_i, X_j);

Обчислити кластери C з D, використовуючи будь-який алгоритм метричної кластеризації;

Повернути (C);

результат: Кластери C

Високий, позитивний коефіцієнт β значить, що пов'язана дія має тенденцію вказувати на різку зміну в інтерпретації послідовності. Наприклад, П часто отримує високий коефіцієнт β, що має на увазі те, що спостереження додаткового швидкого запиту підказки значно змінює інтерпретацію послідовності.

Це відповідає спостереженням поведінки і результатів (в цих наборах даних): запит підказки, ймовірно, є гіршим варіантом, ніж відсутність підказки (за інших рівних умов), а оскільки в більшості систем існує всього декілька фактичних рівнів підказок, кожен запит підказки має велику вагу.

І навпаки, низький або негативний коефіцієнт β означає, що дія практично не впливає на інтерпретацію послідовності. Наприклад, швидкі спроби часто отримують низький коефіцієнт β. Це узгоджується із спостережуваними поведінками: хоча повторне вгадування здається поганою

поведінкою в цих наборах даних, вгадування ще раз, якщо студент вже вгадав раніше, навряд чи призведе до різкої різниці в інтерпретації.

Існує множина варіантів набору можливих ваг. Наприклад, простим випадком може бути одна вага для всіх замін і одна вага для всіх вставок і видалення. Цей варіант, однак, не дозволить вагам адаптуватися до важливості різних типів дій. Крайнім варіантом в бік ускладнення було б дозволити вагу для кожної можливої вставки, видалення і заміни. Це відповідає значенню V, показаному раніше.

Є ще кілька важливих варіантів, коли справа доходить до реалізації лінійного перезважування відстаней. Результати навчання можна агрегувати за допомогою середнього значення, медіани або будь-якої іншої агрегатованної статистики, але, крім того, ці результати навчання можуть бути зважені по ряду показників: кількість спостережень послідовності, ймовірність послідовності і т.д. Наприклад, якщо студент використовує послідовність *i* сім разів з 70 загальних послідовностей, то при обчисленні Y_i, виграш в навчанні студента може бути або семиразовим, або 0,14-кратним.

По-друге, існує пов'язане з цим питання нормалізації для *V*. Візуально кожна операція вносить вклад в відстань редагування тільки один раз, незалежно від довжини послідовності. Однак, як було показано раніше, кластери мають тенденцію бути більш збалансованими, коли відстані нормалізовані; для нормалізації по довжині послідовності відповідає нормалізації по рядках *V*.

Висновки до розділу

Таким чином розроблена система має ряд переваг. Збіжність системи, яка побудована на основі лінійної регресії і алгоритму навчання з увігнутою метрикою (в даному випадку ієрархічна агломеративного кластеризація), і тому гарантовано сходиться до одного і того ж рішення при однаковому наборі даних. Домінуючою обчислювальної витратою при вирішенні є обчислення відстаней

між різними послідовностями, що вимагає часу $O(m^2)$.

Для великих наборів даних це робить більш ефективним, оскільки масштабується в поліноміальній залежності від загального числа послідовностей n (а не від числа окремих послідовностей m). Багато послідовності, такі як Cc, повторюються багато разів в будь-якому даному наборі даних, що означає m n. Коефіцієнти β, розраховані на етапі регресії, легко інтерпретуються. Великі значення β говорять про те, що додавання дії певного типу значно прогнозує зміну результатів. Таким чином, значення β можна використовувати для побудови інтервенцій (наприклад, втрутитися, якщо хороший студент виконує дії з великими значеннями β).

Розділ 4
Дослідження ефективності методу кластеризація на базі Марківських процесів

4.1 Чисельні результати моделювання

Емпірично в даній роботі використовуються три набору даних. Більшість результатів, наведених вище, були зосереджені на Набір 1. До теперішнього часу Набір 1 показав, що вибір порога і гранулярності має значення для визначення ефективності різних алгоритмів кластеризації цілей. В цілому, найбільш перспективним варіантом виявилися ПММ з зваженими напрямними, за якими слідують ПММ з редагуванням.

Кожен з цих наборів даних пропонує щось унікальне: Набір 2 схожий на Набір 3, але також має помітні відмінності в даних і деякі суттєві відмінності, і тому певна поведінка буде відмінною.

Важливо коротко розглянути результати, отримані за допомогою імітаційних наборів даних, щоб забезпечити базові показники для порівняння. До тепер розглядалися три типи імітаційних наборів даних: Наївний, Базовий і Випадковий. Модель Наївний відносно легко вивчається за допомогою окремих, рідко перетинающихся ПММ. Базова модель заснована на моделях, раніше вивчених на емпіричних наборах даних; вона генерує більш реалістичні дані. Модель Випадковий відноситься до будь-якого набору випадково згенерували ПММ.

Для наївною моделі МО-ПММ може знаходити колекції ПММ в середньому з R2 тестового набору 0,65 і, як мінімум, з R2 тестового набору 0,5. Для базової моделі МО-ПММ може знайти колекції ПММ в середньому з R2 тестового набору 0,6. Для випадкової моделі МО-ПММ може іноді знаходити колекції ПММ з позитивним R2 тестового набору; для більш обмежених випадкових ПММ R2 тестового набору зростає до 0,25. В цілому, чим більш

обмежені вихідні ПММ і чим чіткіше розділені пов'язані з ними викиди, тим легше МО-ПММ вивчити колекції.

В цілому, результати моделювання показують, що існують колекції ПММ, які можуть бути відновлені алгоритмами, що це не той випадок, коли будь-яка випадкова колекція ПММ призведе до виявлення корисної моделі, то ймовірно, що модель описує деяку внутрішню структуру даних.

Крім результатів моделювання, більшість результатів до цього моменту стосувалися Набір 1. Основними результатами щодо Набір 1 є наступні:

– МО-ПММ може знаходити моделі поведінки студентів, які передбачають (R2 = 0,26) результати навчання до і після.

– МО-ПММ працює (для Набір 1) як на рівні кроку, так і на рівні завдання.

– МО-ПММ в значній мірі стійка до різних варіантів граничних значень в Набір 1.

– З ваговими коефіцієнтами ПММ досягає аналогічних або кращих результатів (R2 = 0,36), але вимагає менше часу на обчислення.

– Обидва алгоритми зазвичай створюють кілька колекцій кандидатів з високою продуктивністю, але ці колекції також мають тенденцію бути схожими.

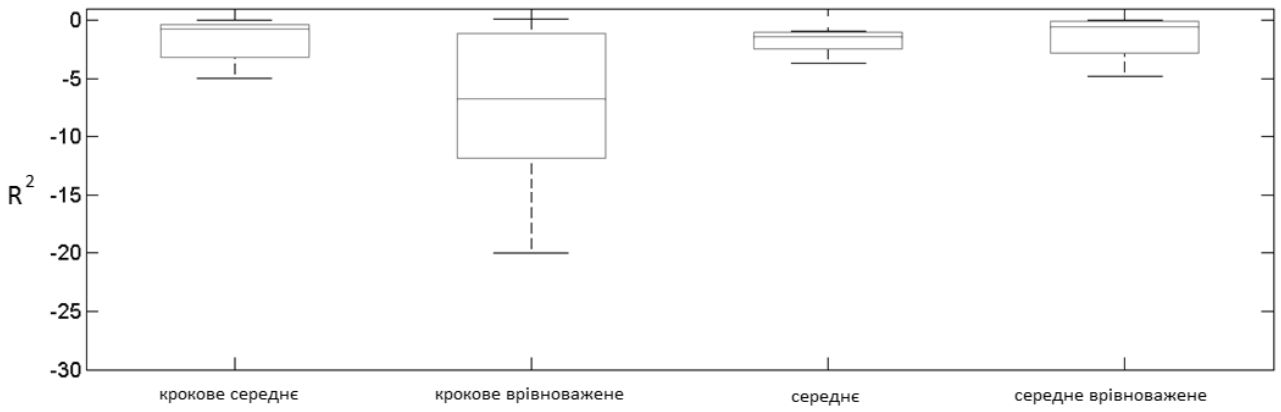


Рисунок 4.1 - Продуктивність R2 на тестовому наборі для зважених ПММ на чотирьох різних версіях даних Набір 2 з загальними поганими результатами

Наступним найбільш схожим набором даних є Набір 2. Набір даних Набір 2 заснований на іншій версії когнітивного репетитора. Проте, в принципі, це відносно схожий набір даних, який забезпечує хорошу перевірку емпіричних властивостей МО-ПММ. При майже постійно негативному R2 тестового набору ці результати показують, що зважений ПММ не знайшов жодної колекції ПММ, яка б розумно моделювала поведінку і навчання студентів.

Існує кілька можливостей пояснити низьку продуктивність алгоритмів на цьому наборі даних. По-перше, цільові дані (бали до і після тестування) можуть бути абсолютно не пов'язані з поведінкою студента, що порушує основне припущення кластеризації цілей. По-друге, студенти можуть вести себе так, що поведінка не може бути описана простими прихованими моделями Маркова з декількома станами. Поведінка студентів також може вимагати додаткової складності, що виходить за рамки кластерів ПММ; наприклад, поведінка студентів на одному кроці може умовно залежати від поведінки на попередніх кроках, яке не може бути представлено поточними колекціями. Нарешті, низька продуктивність може бути просто проблемою виведення, яка повинна бути вирішена за рахунок більшої кількості обчислень,

Існують переконливі докази на користь першої можливості, що полягає в тому, що бали до і після тестування не пов'язані зі спостережуваною поведінкою.

4.2 Цільва кластеризація

Одне із припущень, що пояснює низьку продуктивність алгоритмів цільової кластеризації на деяких з цих наборів даних, полягає в тому, що оцінки можуть бути відірвані від реальної поведінки студентів. Цю теорію легко перевірити як на емпіричних, так і на змодельованих наборах даних. Просто рандомізуя присвоєння студентам значень λ і потім намагаючись вчитися на цих значеннях, можна визначити ступінь, при якому шумова інформація про цілі може заплутати або спотворити алгоритм кластеризації цілей.

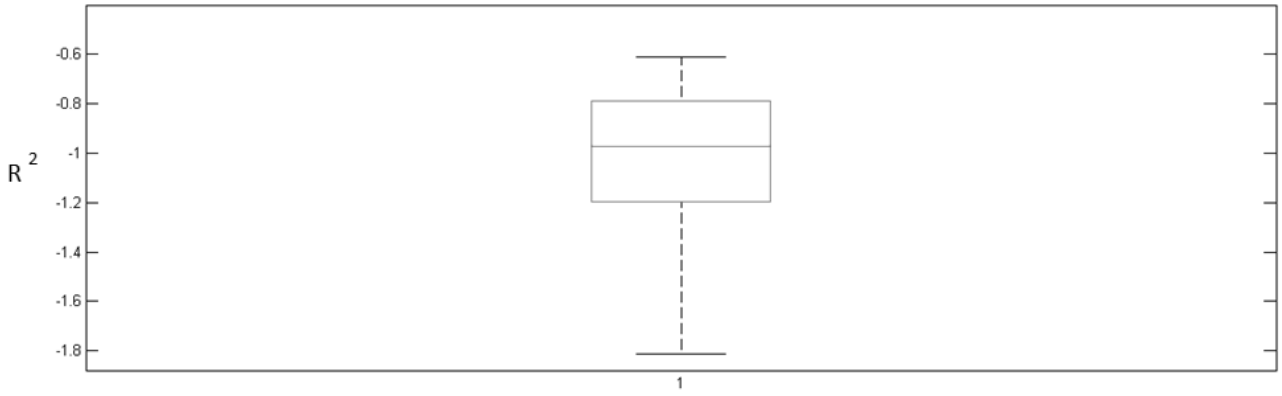


Рисунок 4.2 - Продуктивність тестового набору для Набір 1 з випадковими цілями

Оскільки Набір 1 - це набір даних, на якому алгоритми цільової кластеризації працюють добре, він є хорошим кандидатом для рандомізації значень λ. Інформація про мету може бути рандомізовані шляхом простого перепризначення існуючих значень λ випадковим студентам. Результуюча продуктивність тестового набору показана на рисунку 4.2. Очевидно, що пророкування вкрай погане.

Ключове питання полягає в тому, чи можна виявити, коли цільові λ містять мало або взагалі не містять релевантної інформації. Одна з можливостей полягає в якості підгонки навчального набору. На рисунку 4.3 показані гістограми продуктивності навчального набору для колекцій ПММ, протестованих зважених-ПММ під час ітераційного пошуку.

У разі випадкових λ, продуктивність навчального набору є поганою, при цьому повторні сеанси зазвичай не знаходять значимого предиктора.

У разі невипадкових λ продуктивність навчального набору іноді буває низькою, що відображає пошук набору розумних кандидатів.

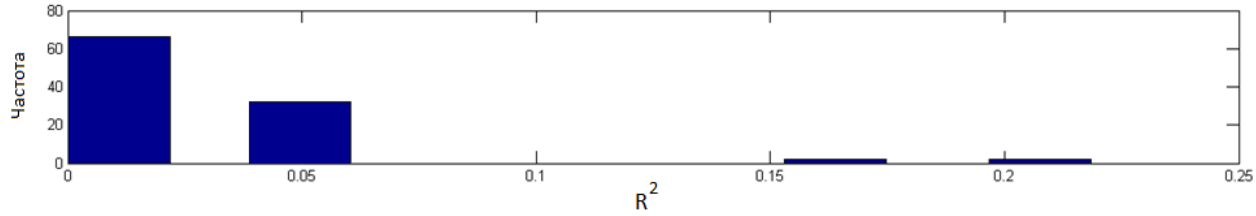


Рисунок 4.3 - Гістограми для Набір 1 з випадковими цілями

Ця різниця пропонує можливу евристику для визначення потенціалу цільової кластеризації на наборі даних: якщо колекції, вивчені на навчальних даних, настільки погані, що не мають значущих предикторів, цільова кластеризація, ймовірно, незастосовна; проте, якщо колекції мають діапазон підгонки на навчальних даних, але продуктивність тестового набору погана, то, ймовірно, варто подивитися на основні припущення моделювання алгоритму і внести коригування, якщо це необхідно.

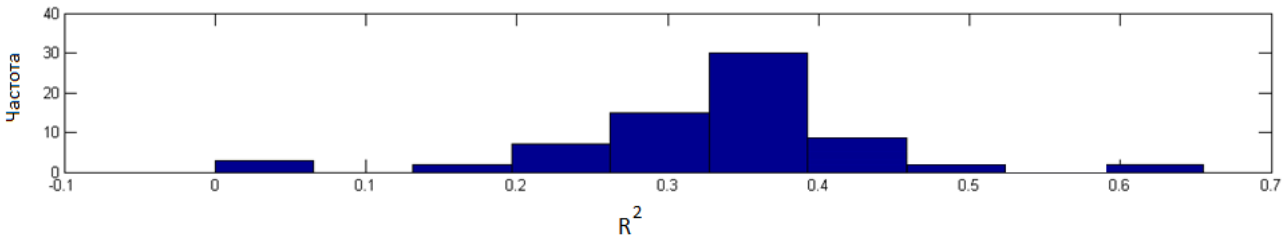


Рисунок 4.4 - Гістограма для Набір 1 з і невипадковими цілями

Один із способів імітації невдалого вибору порога - почати з даних, згенерованих наївною моделлю, і випадковим чином замінити половину "C" на "c". Це імітує випадок, коли поріг для спроб встановлений занадто високо, в результаті чого навіть тривалі спроби класифікуються як швидкі. Продуктивність тестового набору хороша, але що більш важливо, продуктивність навчального набору демонструє моделі, що більше нагадують нормальні цілі, ніж випадкові цілі. Цей факт дає корисну евристику для визначення того, чи є інформація про ціль, наявна в наборі даних, або проблема криється десь в передумовах попередньої обробки або моделювання.

4.3 Параметри попередньої обробки

У Набір 2 хвіст довших послідовностей все ще існує, але послідовностей довжини 2 і 3 значно менше. Це ускладнює алгоритму пошук виразних ПММ, оскільки a і A домінують. У Набір 2, зокрема, послідовності довжини 1 складають більше 75% даних.

Одним з можливих рішень є видалення послідовностей довжини 1 з набору даних.

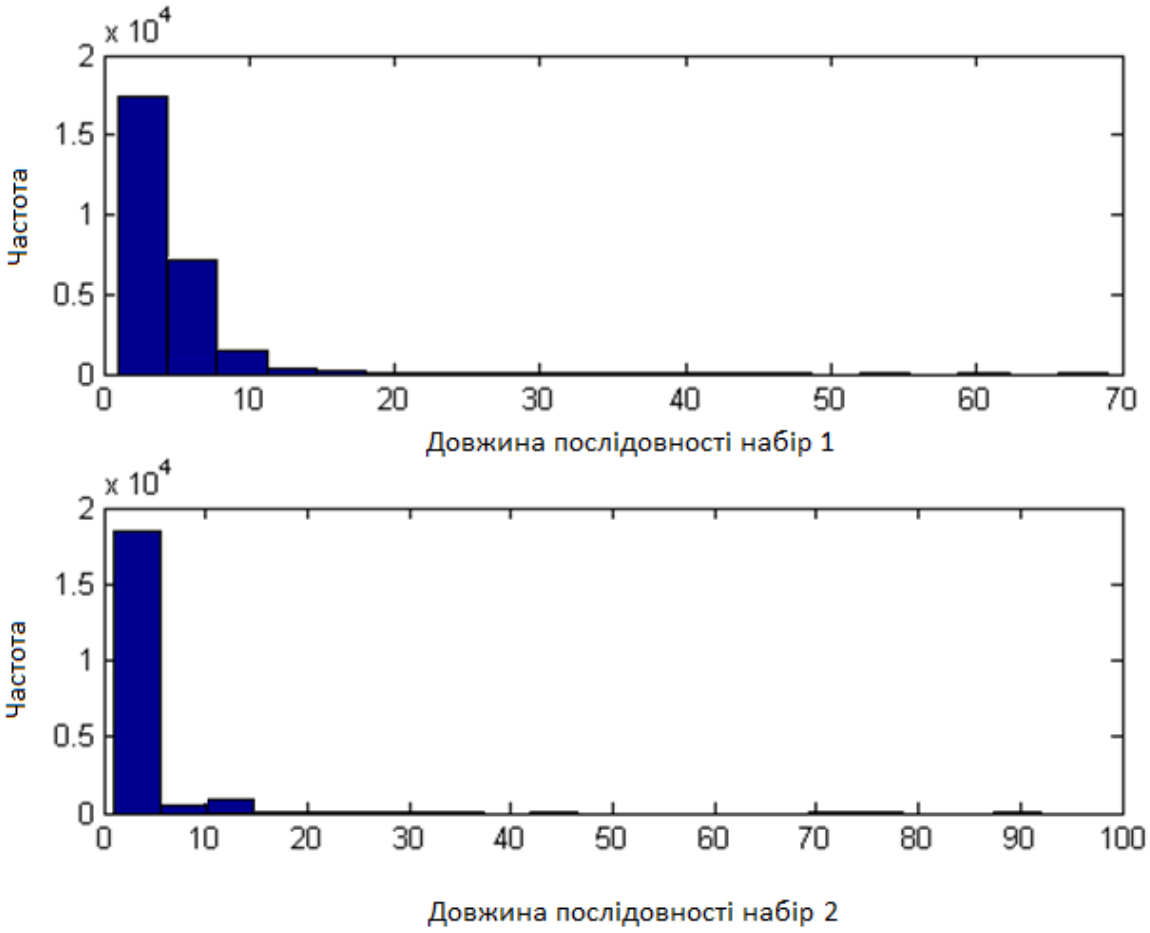


Рисунок 4.5 - Порівняння розподілу довжин послідовностей за розділами даних

Переважна більшість - це поодинокі спроби, які призводять до правильних відповідей. Оскільки метою є пошук моделей, які передбачають

навчання, а не обов'язково тих, які передбачають знання, ці дії потенційно малокорисні. Однак присутність приводить до зміщення параметрів ПММ.

Існують і інші методи попередньої обробки даних для інших проблемних парадигм, які можуть бути корисні при цільової кластеризації. Наприклад, одним з основних підходів до вирішення проблеми зашумленних міток є попередня обробка даних.

Зазвичай це включає або очищення (видалення), або виправлення прикладів, які вважаються зашумленими. Аналогічна ідея може бути застосована до проблем кластеризації цілей.

Наприклад, в освіті, якщо з першої спроби зробили правильний крок у вирішенні завдання, це означає, що отримали цю навичку; така поведінка зазвичай позитивно корелює з навчанням.

Це означає, що різко позитивні послідовності, такі як перша спроба-правильна, можуть бути виявлені за допомогою прямих кореляцій і видалені з числа погано навчаються; аналогічно, різко негативні послідовності можуть бути виявлені і вилучені з числа добре навчаються. Все це може бути виконано без будь-яких невід'ємних припущень про моделювання функції *f*.

Також рекомендується забезпечити кращу ініціалізацію для алгоритмів. В даний час початкові ПММ визначаються випадковим чином. Замість цього можна форматувати ПММ, використовуючи цільові λ.

Наприклад, перша колекція ПММ може складатися з двох ПММ, одна з яких навчена на всіх послідовностях, отриманих від студентів з позитивною успішністю, а друга ПММ навчена на всіх послідовностях, отриманих від студентів з негативною або поганою успішністю.

Далі, під час кожної ітерації, коли ПММ видаляється з колекції, він може бути замінений двома ПММ, як описано вище, за винятком використання тільки послідовностей в ПММ.

Велика частина навчання, що визначається за кластеризацією цілей, відбувається в момент першого застосування, а не в момент першого пояснення.

По суті, як описано вище, спроби є потужними предикторами навчання; підказки ж практично не показують, що вони корисні для навчання або передбачають його. Кращим позитивним предиктором навчання, виявленим за допомогою кластеризації цілей, є відсоток кроків, на яких студент або правильно виконує першу спробу, або робить повторні, повільні, наполегливі спроби.

Це спостереження добре узгоджується з теоріями навчання, які в першу чергу зосереджені на застосуванні навичок, а не на читанні інструкцій або пояснень.

Велика частина сучасної освітньої літератури з навчання присвячена поведінці при пошуку допомоги та самоосвіті. В основі цих теорій лежить аргумент, що студенти вчаться або в присутності декларативних інструкцій (наприклад, підказок), або на відпрацьованих прикладах (наприклад, підказки знизу). Альтернативою є більш традиційний аргумент, який фокусується на кожному застосуванні навички як на можливості навчання. Наприклад, в навчання за виробничими правилами (навичками) відбувається при правильному чи неправильному застосуванні, а не тоді, коли студент читає інструкцію про виробничі правила.

Використання посібника в колекціях ділиться на дві категорії: негативно пов'язані з навчанням і не пов'язані з навчанням. ПММ з великою кількістю підказок, негативно пов'язані з навчанням, схильні моделювати послідовності з великою кількістю спроб, зазвичай одну за одною в довгій послідовності. Ймовірно, це вказує на те, що вони пробираються крізь шари підказок, шукаючи або нижню (кінцеву) підказку, або іншу підказку низького рівня. Більш нейтральні ПММ з великою кількістю підказок, як правило, включають всі інші повторювані моделі поведінки підказок, такі як прості послідовності або послідовності, що змішують підказки і спроби. У будь-якому випадку, ці моделі свідчать проти ефективності підказок в вищевказаних наборах даних. Вони свідчать про те, що корреляція з навчанням приблизно дорівнює нулю.

Хоча виявлені колекції свідчать про те, що використання підказок не є корисним предиктором навчання, неясно, студенти взагалі не вчаться на підказках або підказки просто не так корисні, коли студенти отримують пряме навчання за допомогою інших засобів, таких як пряме керівництво і підручники. Також незрозуміло, чи пов'язана основна проблема з вмістом підказок, подачею або порядком розташування. Можливо, наприклад, демонстрація відпрацьованого прикладу замість багатошарових підказок була б більш ефективною.

Крім того, багато теорій про саморегульоване навчання припускають, що студентам потрібна більш активна підтримка, щоб бути продуктивними та саморегульована. Наприклад, самопояснення, хоча фізично вимагає від студентів лише нагляду за прикладами, когнітивно вимагає активного рівня залученості в матеріал. Таким чином, відсутність спостережуваного ефекту підказки може бути результатом незалученості студентів. Підказки самі по собі можуть бути корисні, але тільки в тому випадку, якщо студенти отримують активність, що сприяє більш активному залученню.

Можуть існувати корисні поведінкові моделі підказок, які важко виявити за допомогою цих моделей. Одним із прикладів, як і раніше пов'язаним з самопоясненням, є попередня робота, в якій детально описана модель продуктивного використання підказки знизу. В основі цієї моделі лежать відмінності в часі між підказками з нижнього і не нижнього рівнів, з деякими поправками на індивідуальні відмінності в часі набору тексту і часу читання. Ця модель передбачала навчання, незважаючи на те, що фокусувалася на послідовностях завдання, що включають множина швидких запитів підказок.

Однак алгоритми ПММ не завжди можуть знайти таку модель з кількох причин:

– Набір не відрізняє підказки з нижнім виходом від підказок без нижнього виходу. Для коригування цього параметра буде потрібно більше даних, щоб компенсувати відносну рідкість підказок знизу.

– Пороги маскують тонкі відмінності в термінах виконання окремих дій.

– Алгоритми ПММ моделюють розподіл дій на студента, а не середній час на одну дію на студента.

– Коригування індивідуальних відмінностей в моделі підказок вимагає доступу до першої дії в наступних кроках, що недоступно для цільової кластеризації в її нинішньому дизайні.

Всі ці аргументи говорять про те, що можуть існувати корисні поведінки, пов'язані з підказками, які просто залишилися непоміченими через обмеження в припущеннях моделювання, в алгоритмах або в даних. Останнє здається особливо можливим, так як в цих наборах недостатньо даних, з яких можна вивчити параметри для певної кількості ПММ.

При всіх хороших результатах існують практичні проблеми з цільовою кластерізацією для даних про освіту. Найбільш ефективні алгоритми кластеризації цілей, протестовані на сьогоднішній день, засновані на алгоритмі максимізації очікування (МО). Алгоритми МО схильні до локального оптимуму, що ускладнює інтерпретацію моделей як відображають фундаментальні цілі.

Найбільш ефективні на сьогоднішній день алгоритми також спираються на приховані моделі Маркова. Хоча ПММ є потужними і інтерпретуються, для використання в моделюванні поведінки студентів не вистачає теоретичної основи. Наприклад, ПММ припускають, що вся необхідна інформація представлена в стані ПММ і не потрібно ніякої іншої минулої історії. Для досить великих просторів станів це, ймовірно, вірно, але для невеликого числа станів, що використовуються в кластеризації цілей, це припущення стає недостатнім.

Кластеризація цілей як парадигма, незалежно від реалізації, створює додаткову складність і чутливість до процесу машинного навчання. Більшість алгоритмів залежать тільки від якості вихідних даних; алгоритми кластеризації цілей залежать від якості вихідних даних і якості цілей. Якщо цілі погані (наприклад, ненадійна оцінка), то і результати кластеризації цілей будуть поганими. Фактично, можливо, що вихідні дані можуть бути високої якості, а

оцінка - високої якості, в той час як між ними, на перетині цих двох параметрів, знаходяться дані трасування, які не корелюють з оцінкою. Це робить мету кластеризації потенційно більш чутливою до недоліків наявних даних.

У кожного з цих аргументів є й контраргумент. Хоча МО гіпотетично проблематична, на практиці алгоритми ПММ виявилися стійкими до змін початкових умов, параметрів і навіть наборів даних. По-друге, хоча ПММ можуть включати в себе деякі теоретично необґрунтовані припущення щодо даних про освіту, вони також є високоефективними моделями не тільки поведінки студентів, а також поведінки в інших областях. По-третє, хоча взаємодія між якістю даних і якістю цілей є потенційно проблемною, з попередніх результатів ясно, що відсутність цілей є набагато більш серйозною проблемою.

Таким чином, незважаючи на свої обмеження, цільова кластеризація виконує ряд корисних функцій. Алгоритми цільової кластеризації продемонстрували ефективність в освітніх даних для таких цілей:

– Створення прогностичних моделей.

– Дослідницький аналіз наборів даних.

– Оцінка послідовного інтерфейсу.

– Адаптація до нових наборів даних і систем.

Багато з них будуть включають конкретний, хоча і гіпотетичний, приклад того, як кластеризація цілей може бути застосована дослідником в галузі освіти в ході проекту.

4.4 Експериментальні моделі

Для наборів існують колекції, знайдені за допомогою зваженого-ПММ, які добре працюють як на тренувальних, так і на тестових даних. Однак, щоб бути корисними не тільки для простого прогнозування навчання студентів, ПММ в цих колекціях повинні бути інтерпретовані експертами.

Набір 1 дає стабільні результати, які виявляються ПММ при різних порогових значеннях і гранулярності. В результаті важливо розглянути загальні риси між окремими колекціями ПММ, виявленими за допомогою цільової кластеризації на наборі 1. В наведених нижче прикладах є п'ять типів викидів. Існує чотири основних типи ПММ, які зазвичай з'являються в колекціях на покрокових рівнях, отриманих на наборі 1. До них відносяться:

– Домінуюча модель, яка включає в себе як послідовності правильних спроб, так і більш тривалі послідовності спроб. Ця модель зазвичай позитивно пов'язана з с - швидка спроба, С - тривала спроба, п - швидка підказка, П - тривала підказка, Н - запит на доступ до допомоги наставника. Ця модель є "домінуючою", тому що вона містить найбільш передбачувану послідовність - С. Ця послідовність, яка вказує на одну правильну спробу і, таким чином, на те, що студент, ймовірно, знає навик, необхідний на даному етапі, часто зустрічається і є важливою послідовністю. Крім того, модель містить довші послідовності повторних спроб. Ці довші послідовності зазвичай включають всі послідовності спроб, що не містяться в "Моделі вгадування".

– Модель вгадування, яка зазвичай генерує одне "С", за яким слідує ряд "с", представляючи таким чином важливий тип послідовності. Ця модель негативно асоціюється з навчанням. Тут вона називається "Вгадування", хоча невідомо, що студент справді вгадує, проте повторювані швидкі спроби дозволяють припустити, що студент не продумує свої дії.

– Модель "Підказки", яка негативно пов'язана з навчанням і містить множину типів поведінки з підказками. Ця модель іноді об'єднується з "Різною моделлю".

– Модель "Різне", яка просто поглинає всі інші послідовності, не розглянуті вище. Оскільки модель "Різне" практично не пов'язана з навчанням, послідовності, які їй присвоюються, розглядаються як шум.

Як для емпіричних, так і для змодельованих наборів даних ПММ можуть вивчати колекції, які піддаються інтерпретації і попередньо формують

продуктивність як на навчальних, так і на тестових даних. Алгоритми знаходять прогнозні моделі при роботі з простими породжуючими моделями, більш реалістичними (базовими) імітаційними моделями і навіть обмеженими, рандомізованими моделями.

Алгоритми знаходять прогнозні моделі в декількох варіаціях на двох наборах даних R2 тестового набору варіюється від 0,2 до високих 0,32 (Набір 1). Однак, зважені ПММ не дали корисних моделей принаймні на одному наборі даних (Набір 2), а інші алгоритми були ще менш послідовними (давали корисні моделі тільки для певних варіантів).

Важливо відмітити, що можна провести відмінність між різними типами невдач. Як для емпіричних, так і для модельованих даних можна показати (емпірично), що рандомізація інформації про цілі призводить до інших очікуваних результатів навчання, ніж поганий набір (наприклад, тому що пороги між швидкими і повільними діями невірні).

Для освітніх даних це означає, що існує помітна різниця між випадками, коли інформація про результати навчання ненадійна або не пов'язана з модельований поведінкою, і випадками, коли фундаментальні припущення про набір даних невірні. Так, для всіх протестованих комбінацій порогів і для всіх форм попередньої обробки жоден алгоритм цільової кластеризації не зміг знайти прогностичні моделі для набору даних Набір 2.

Цей результат був цілком передбачуваний, оскільки алгоритми цільової кластеризації не змогли знайти навіть відповідні моделі на навчальних даних. Це дозволяє відокремити перспективні випадки від випадків, коли розбіжності неможливо розв'язати.

На змодельованих даних алгоритми ПММ створюють колекції, які не тільки узгоджуються між експериментами, але і узгоджуються з породжуючими моделями. На емпіричних даних окремі експерименти з алгоритмами ПММ, що використовують різні початкові колекції (а іноді і різні параметризації), дають схожі колекції ПММ.

Фактично, при різних порогових значеннях для набору, виявлені колекції практично ідентичні. Крім того, коли послідовності довжини 1 видаляються, виявлені колекції адаптуються і виглядають як колекції, виявлені з включеними послідовностями довжини 1, за винятком того, що ці послідовності відсутні; структурно ПММ мало змінилися. Це спостереження досить очевидно при порівнянні колекцій.

Послідовність є більш серйозною проблемою для лінійного перерозподілу відстаней, яке добре працює тільки для певних наборів даних з певними граничними значеннями.

Хоча можливо, що сам алгоритм потребує вдосконалення, імовірно, що проблема полягає в цільовому вирішенні. Наприклад, після того, як перерозподіл відстаней отримує коефіцієнти для різних операцій, він передбачає, що всі попередні розрахунки відстані залишаються в силі, тільки тепер використовуються нові коефіцієнти. Однак нові коефіцієнти можуть змінити шлях мінімальної ваги між двома послідовностями, а значить, і відстань між ними.

Цільова кластеризація може відновити вихідні породжуючі моделі з симульованих даних, якщо вихідні моделі добре розділені. Ці добре розділені моделі залишаються виразними, відображаючи як існуючі моделі, так і моделі, фактично виявлені цільової кластерізацією в емпіричних наборах даних. Хоча іноді існують значні структурні відмінності між моделями в генеруючої і виявленої колекціях, при розгляді як цілої колекції.

Висновки до розділу

В цілому, алгоритми цільової кластеризації показали хороші результати на наборах даних, змодельованих і емпіричних, навіть коли дані були складними для алгоритмів нецільової кластеризації. Однак кожен окремий аналіз, представлений раніше, є частиною більш цілісної картини, що ілюструє, чому

цільова кластеризація необхідна, може стати реальністю, стійка, послідовна і надійна. Таким чином, основні результати роботі охоплюють:

– Необхідність кластеризації цілей, як існування прикладної проблеми, яку неможливо вирішити традиційними методами.

– Доцільність і надійність цільової кластеризації що проявляється в здатності цільової кластеризації до навчання прогностичних моделей на реальних і змодельованих наборах даних.

– Алгоритми цільової кластеризації створюють схожі моделі, незважаючи на різні початкові умови, параметри і навіть набори даних.

– Зважені ПММ з направляючими, засновані на МО, є надійним, ефективним алгоритмом, який створює інтерпретовані моделі.

– Запропоновані методи показали перспективність як засобу зворотного зв'язку щодо змін в дизайні, новими освітніми теоріям та інше. Це також цілком прийнятне рішення, яке може бути використано багатьма дослідниками для вирішення проблем послідовності освоєння та конструювання інтерфейсу.

Загальні висновки

Більшість традиційних алгоритмів кластеризації розроблені для максимізації відповідності вхідних даних, але не призначені для адаптації до додаткових, зовнішніх показників. Таким чином, існує принаймні два способи, якими алгоритми цільової кластеризації відрізняються від більшості традиційних алгоритмів кластеризації. По-перше, вони навчаються, використовуючи послідовності, а не вектори. По-друге, вони враховують цільову інформацію, наприклад, результати навчання студентів. Як показано на Набір 1 для різних векторних алгоритмів, без цільової інформації важко (або неможливо) знайти моделі, що передбачають успішність. Наприклад, і базовий k-means, і спектральна кластеризація (обидва з яких навчаються на векторних даних) знайшли кластери, які добре працювали на навчальних даних. Насправді, навіть МО-ПММ, базовий алгоритм, на якому засновані ПММ алгоритми, які були створені, аналогічним чином не змогли передбачити поліпшення результатів навчання на тестових даних, незважаючи на навчання на повних послідовних даних. Однак, коли в алгоритм була додана функціональність тобто інформація про цілі, продуктивність значно зросла.

Крім того, коли інформація про цілі є випадковою, алгоритми кластеризації цілей перестають працювати добре. При роботі з випадковими цілями очікується низька продуктивність, якщо інформація про мету важлива для процесу машинного навчання на цих наборах даних. Цей контраст з випадковими цілями ще більше підкреслює, що кластеризація цілей не тільки важлива для навчання корисних моделей на прикладних наборах даних, але і що без якісної інформації про цілі навіть алгоритми кластеризації цілей працюватимуть погано.

Алгоритми цільової кластеризації показали багатообіцяючі результати на кількох наборах даних, змодельованих і емпіричних. Алгоритми цільової кластеризації дали результати, що представляють інтерес в поліпшенні існуючих алгоритмів, в адаптації алгоритмів до парадигми цільової кластеризації, в поєднанні цільової кластеризації з іншими підходами, такими як імовірнісні реляційні моделі, в дослідженні повністю ієрархічної форми цільової кластеризації, в розширенні існуючих алгоритмів для підтримки моделей навичок та безперервних змінних, і, звичайно, в дослідженні нових наборів даних і областей.

Перелік посилань

1. Billig M. Arguing and thinking: A rhetorical approach to social psychology. – Cambridge University Press, 1996.

2. Nespor J. The role of beliefs in the practice of teaching //Journal of curriculum studies. – 1987. – Т. 19. – №. 4. – С. 317-328.

3. Hofer B. K., Pintrich P. R. Personal epistemology: The psychology of beliefs about knowledge and knowing. – Routledge, 2012.

4. Hofer B. K. Dimensionality and disciplinary differences in personal epistemology //Contemporary educational psychology. – 2000. – Т. 25. – №. 4. – С. 378-405.

5. Chew M. S. F., Shahrill M., Li H. C. The Integration of a Problem-Solving Framework for Brunei High School Mathematics Curriculum in Increasing Student's Affective Competency //Journal on Mathematics Education. – 2019. – Т. 10. – №. 2. – С. 215-228.

6. Schommer M. An emerging conceptualization of epistemological beliefs and their role in learning //Beliefs about text and instruction with text. – Routledge, 2019. – С. 25-40.

7. Geigle C., Zhai C. X. Modeling MOOC student behavior with two-layer hidden Markov models //Proceedings of the fourth (2017) ACM conference on learning@ scale. – 2017. – С. 205-208.

8. Wang G. et al. Modeling student learning Behaviors in ALEKS: A two-layer hidden Markov modeling approach //International Conference on Artificial Intelligence in Education. – Springer, Cham, 2018. – С. 374-378.

9. Ali S., Bouguila N. Maximum a posteriori approximation of hidden markov models for proportional sequential data modeling with simultaneous feature selection //IEEE Transactions on Neural Networks and Learning Systems. – 2021.

10. Wang S. et al. Convolutional neural network-based hidden Markov models for rolling element bearing fault identification //Knowledge-Based Systems. – 2018. – Т. 144. – С. 65-76.

11. Glasser I. et al. Expressive power of tensor-network factorizations for probabilistic modeling, with applications from hidden Markov models to quantum machine learning //arXiv preprint arXiv:1907.03741. – 2019.

12. Sun D. et al. Comparing interaction activity patterns of different achievement learner groups in MPOCs //Learning. – 2019. – Т. 3. – С. 67.9.

13. Rajendran R. et al. A temporal model of learner behaviors in OELEs using process mining //Proceedings of ICCE. – 2018. – С. 276-285.

14. Perez R. S., Skinner A., Sottilare R. A. – A review of intelligent tutoring systems for science technology engineering and mathematics (STEM) //Assessment of Intelligent Tutoring Systems Technologies and Opportunities. – 2018. – С. 1.

15. Paladines J., Ramírez J. An Intelligent Tutoring System for Procedural Training with Natural Language Interaction //CSEDU (2). – 2019. – С. 307-314.

16. Kochmar E. et al. Automated personalized feedback improves learning gains in an intelligent tutoring system //International Conference on Artificial Intelligence in Education. – Springer, Cham, 2020. – С. 140-146.

17. Cai Z., Hu X. AutoTutor: An Intelligent Tutoring System and Its Authoring Tools //Deep Comprehension. – Routledge, 2018. – С. 140-153.

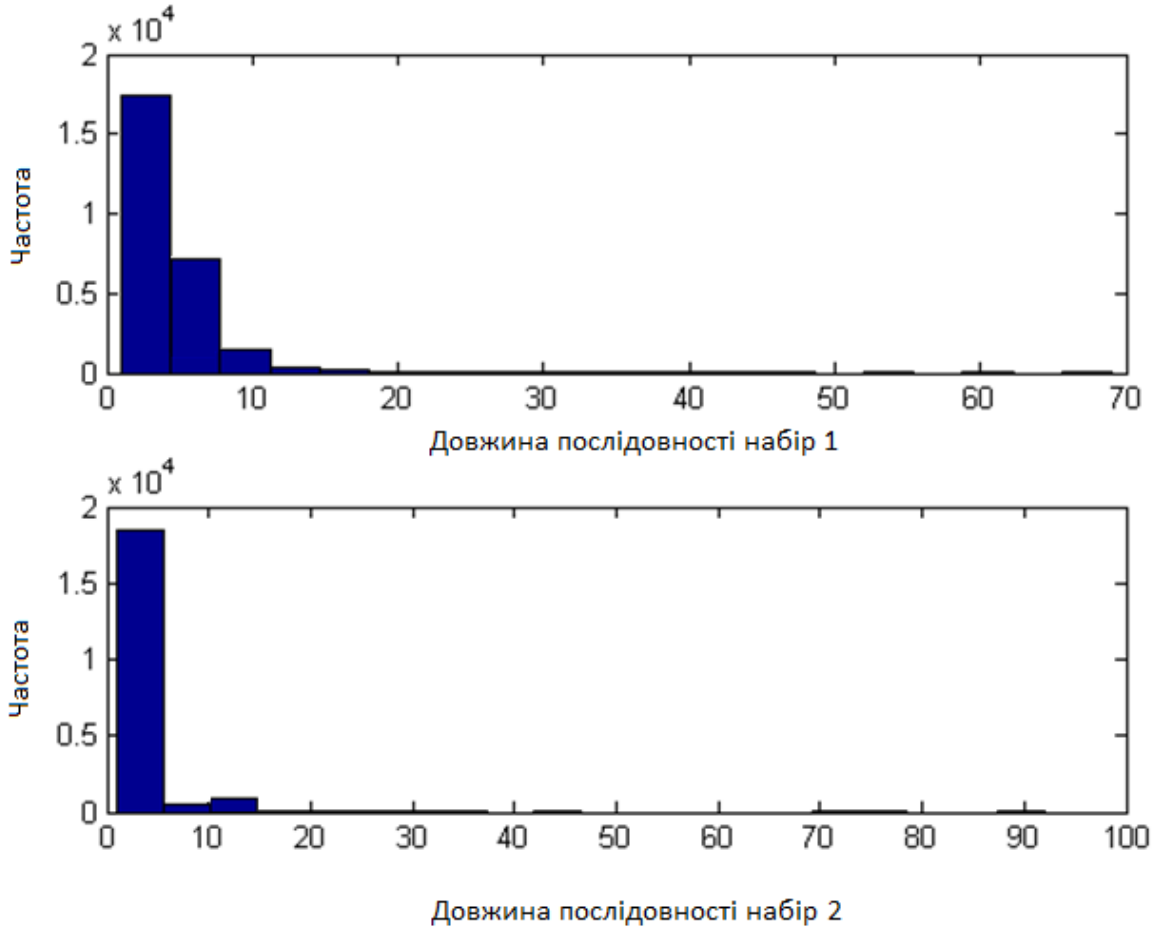
18. Liu J. et al. A discrete hidden Markov model fault diagnosis strategy based on K-means clustering dedicated to PEM fuel cell systems of tramways //International Journal of Hydrogen Energy. – 2018. – Т. 43. – №. 27. – С. 12428-12441.

19. Manogaran G. et al. Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering //Wireless personal communications. – 2018. – Т. 102. – №. 3. – С. 2099-2116.

20. Zhang F. et al. UD-HMM: An unsupervised method for shilling attack detection based on hidden Markov model and hierarchical clustering //Knowledge-Based Systems. – 2018. – Т. 148. – С. 146-166.

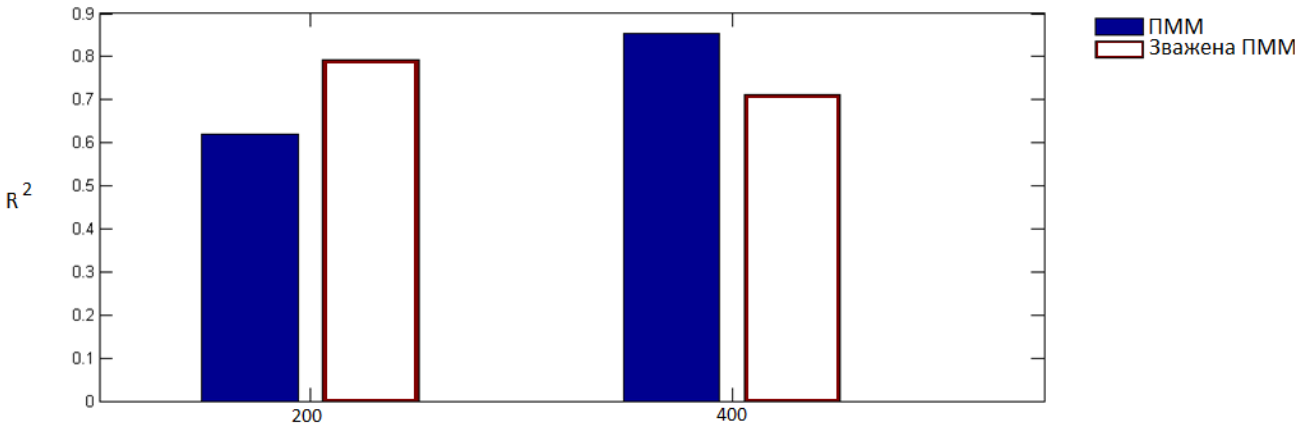
21. Zheng Y. et al. Student’s t-hidden markov model for unsupervised learning using localized feature selection //IEEE Transactions on Circuits and Systems for Video Technology. – 2017. – Т. 28. – №. 10. – С. 2586-2598.

ДОДАТКИ



Порівняння розподілу довжин послідовностей за розділами даних

Продуктивність R2 на тестовому наборі для зважених ПММ на чотирьох різних версіях даних Набір 2 з загальними поганими результатами



Порівняння максимальної відповідності умовам для зважених і простих ПММ на різних наборах змодельованих даних

Таблиця: Рівень взаємної помилки класифікації для 5 кращих колекцій

	колекція 1	**колекція 2**	**колекція 3**	**колекція 4**	**колекція 5**
колекція 1	0	0.0800	0.0930	0.0589	0.0603
колекція 2	0.0800	0	0.0640	0.0845	0.0690
колекція 3	0.0930	0.0640	0	0.0895	0.0954
колекція 4	0.0589	0.0845	0.0895	0	0.0943
колекція 5	0.0603	0.0690	0.0943	0.0724	0

УДК 004.4

Манзюк Е. А., Скрипник Т. К.

Хмельницький національний університет

СИСТЕМА ЦІЛЬОВОЇ КЛАСТЕРИЗАЦІЇ НА ПОСЛІДОВИХ ДАНИХ

Розроблено систему оцінки ефективності навчального процесу використовуючи методи машинного навчання, а саме цільову кластеризацію із максимізацією очікування на прихованих моделях Маркова. Проведені дослідження показали можливість застосовувати цільову кластеризацію для послідовностей на прикладі предметної області. Алгоритми цільової кластеризації показали багатообіцяючі результати на кількох наборах даних, змодельованих і емпіричних. Алгоритми цільової кластеризації дали результати, що мають значимість з освітньої точки зору.

A system for evaluating the effectiveness of the learning process using machine learning methods has been developed, namely targeted clustering with maximizing expectations based on hidden Markov models. Studies have shown the possibility of applying target clustering for sequences on the example of the subject area. Target clustering algorithms have shown promising results on several data sets, simulated and empirical. Target clustering algorithms have yielded results that are relevant from an educational point of view.

Алгоритми кластеризації є потужними методами для розуміння структури даних. Однак у багатьох прикладних областях алгоритми кластеризації корисні тільки в тому випадку, якщо вони дають кластери, які пояснюють зовнішні змінні [1, 2]. Наприклад, в дослідженнях в галузі освіти кластери поведінки студентів корисні тільки в тому випадку, якщо вони допомагають передбачити їх навчання. Як правило, алгоритми кластеризації явно не призначені для обробки цього додаткового обмеження [3]. В даній роботі показано, що включення цих зовнішніх змінних безпосередньо в алгоритми кластеризації може поліпшити як релевантність кластерів, так і узагальнення моделі. Цей підхід, названий кластеризацією цільових послідовностей, дав багатообіцяючі результати на множині наборів даних. Хоча продемонстровано ефективність цього підходу на освітніх наборах даних, також показано, як цей метод може застосовуватися в більш широкому сенсі.

Метою роботи є розробкам методів цільової кластеризації послідовностей для визначення показників ефективностей використання послідовностей в предметній області.

Відкриття нових психологічних і освітніх підходів є в основному завданням дослідників, а не алгоритмів. Обчислювальні підходи, навіть при наявності хороших

даних, в більшості випадків не можуть знайти нові конструкції, які застосовуються до освітніх проблем. Ця трудність частково пов'язана з тим, що традиційні проблеми машинного навчання спираються на вхідні дані і пов'язані з ними мітки, які відірвані від реальності освітніх досліджень, в яких існують складні взаємозв'язки між конструкціями, предметами і завданнями [4]. Стандартні алгоритми просто не враховують багатство наборів освітніх даних, таких як наявність окремих учнів, окремих завдань і загального навчання. У машинному навчанні існує кілька підходів, що дозволяють включати додаткові дані і обмеження, наприклад, обмеження "необхідність-зв'язок" в кластеризації, але адаптація цих алгоритмів за освітніми даними є нетривіальним завданням. Цільова кластеризація, запропонована в цій роботі, є підхід для генерування нових конструкцій з освітніх даних. Основний внесок цієї роботі включає:

- новий клас алгоритмів, простих для розуміння і реалізації, але досить ефективних для роботи з освітніми даними з широким діапазоном складності і структури;
- різноманітність нових конструкцій для навчання людини, створених на основі існуючих даних. Ці конструкції можуть дати поштовх для додаткових досліджень або надати докази, що підтверджують існуючі результати.
- додаткові алгоритми і підходи для вирішення інших завдань дослідження в галузі освіти.

Цільова кластеризація дозволяє вирішити фундаментальну проблему в освітніх дослідженнях: створення правдоподібних, заснованих на даних моделей (і гіпотез), які співвідносяться з фактичними показниками освіти. Неконтрольовані методи машинного навчання дозволяють отримати моделі, які описують дані, але не завжди добре пов'язані із зовнішніми показниками. Напівсамостійні методи створюють моделі, які передбачають зовнішні показники і описують навчання, але вимагають дорогих тонких міток і, крім того, не можуть генерувати нові конструкції. Наприклад, багато дослідників використовували алгоритми для навчання моделям для неуважних правильних відповідей, наприклад, вгадування, але їх моделі обмежені вихідними конструкціями дослідника; якщо їх не додати в якості особливого випадку, ці моделі не будуть включати, наприклад, кілька правильних вводів в комбіноване вікно, що, як було показано раніше, є високо передбачуваним в даних по стехіометрії. Кластеризація цілей здатна генерувати нові моделі і гіпотези, які також мають відношення до освітніх заходів, змішуючи аспекти як несамостійних, так і напівсамостійних методів, при цьому використовуючи тільки високорівневі вказівки цілей, а не покладаючись на трудомісткі мітки для окремих послідовностей.

Алгоритми цільової кластеризації показали багатообіцяючі результати на кількох наборах даних, змодельованих і емпіричних. Алгоритми цільової

кластеризації навіть дали результати, що представляють інтерес з освітньої точки зору.

Перелік посилань

1. Liu J. et al. A discrete hidden Markov model fault diagnosis strategy based on K-means clustering dedicated to PEM fuel cell systems of tramways //International Journal of Hydrogen Energy. – 2018. – Т. 43. – №. 27. – С. 12428-12441.
2. Zheng K., Li Y., Xu W. Regime switching model estimation: spectral clustering hidden Markov model //Annals of Operations Research. – 2021. – Т. 303. – №. 1. – С. 297-319.
3. Samir A., Pahl C. Detecting and predicting anomalies for edge cluster environments using hidden markov models //2019 Fourth International Conference on Fog and Mobile Edge Computing (FMEC). – IEEE, 2019. – С. 21-28.
4. Akogul S., Erisoglu M. An approach for determining the number of clusters in a model-based cluster analysis //Entropy. – 2017. – Т. 19. – №. 9. – С. 452.

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

Метод агрегативної кластеризації на базі послідовного підходу

Розробив ст. гр. КНм-20-2:
Манзюк Е.А.

Хмельницький - 2021

Актуальність дослідження

Застосування методів машинного навчання та інтелектуального аналізу даних до реальних даних і реальних проблем часто вимагає значного втручання і досвіду від дослідників. Зокрема, застосування дослідницьких методів часто вимагає повторного пошуку двох моделей: формального пошуку моделі, вбудованого в алгоритм машинного навчання, і неформального пошуку моделі, здійснюваного практиком машинного навчання. Наприклад, дослідник в області розпізнавання мовлення може спочатку спробувати розібрати людську мову за допомогою підходу динамічного викривлення, потім прихованих марківських моделей, а потім лінійних динамічних систем.

Об'єктом дослідження є процес отримування ефективної кластеризації даних для отримання інформативності щодо структури даних.

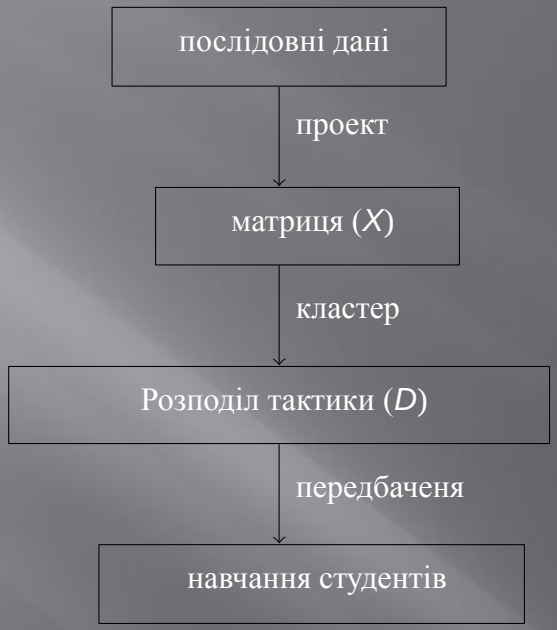
Предметом дослідження моделі, методи та алгоритми автоматизації побудови цільових кластерів на основі послідовних даних.

Метою дослідження є розробка методу реалізація процесу машинного навчання, який здатний агрегатувати структурні послідовні дані.

Для досягнення поставленої мети визначенні такі основні завдання дослідження:
1. Проаналізувати сучасний стан практичних рішень систем формування кластерних структур для формування завдань в предметній області.
2. Удосконалити методи формування кластерних структур із застосуванням цільової кластеризації.
3. Удосконалити системи кластеризації послідовних структур даних та провести порівняльний аналіз із векторними структурами.
4. Розробити метод машинного навчання цільової кластеризації послідовних структур даних.
5. Провести дослідження застосування методів машинного навчання на предметній області.
6. Провести дослідження ефективності практичного застосування запропонованих методів на практичних задачах.

Наукова новизна одержаних результатів. В результаті проведеної роботи були отримані такі результати.
Набула подальшого розвитку система формування кластерних структур які залежні від цілі агрегатування даних.
Запропоновано інноваційний метод розробки систем цільової кластеризації.
Запропоновано метод формування кластерів на множинах результатів моделей прихованих Марківських процесах.

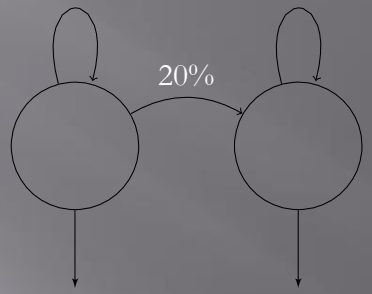
Основні положення і результати роботи опубліковані в збірнику наукових праць – Манзюк Е.А. Система цільової кластеризації на послідових даних // Збірник наукових праць за матеріалами Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук - 2021» Хмельницький, 2021.



Експериментальна процедура, яка використовується для векторних алгоритмів.

Вид вектора: СППсСпПС

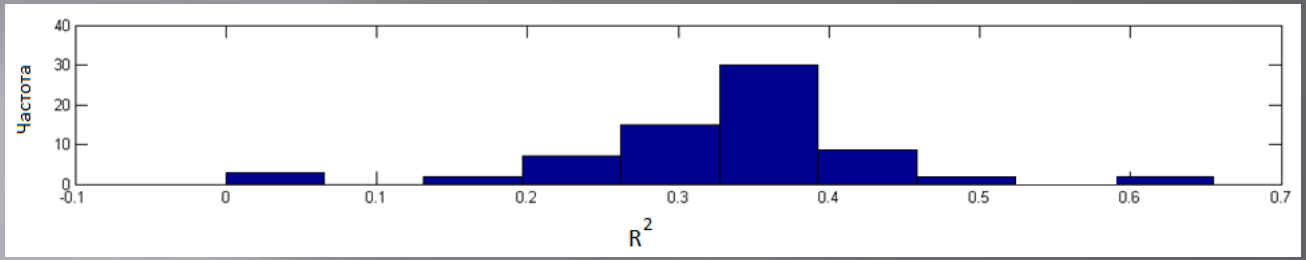
80% 100%



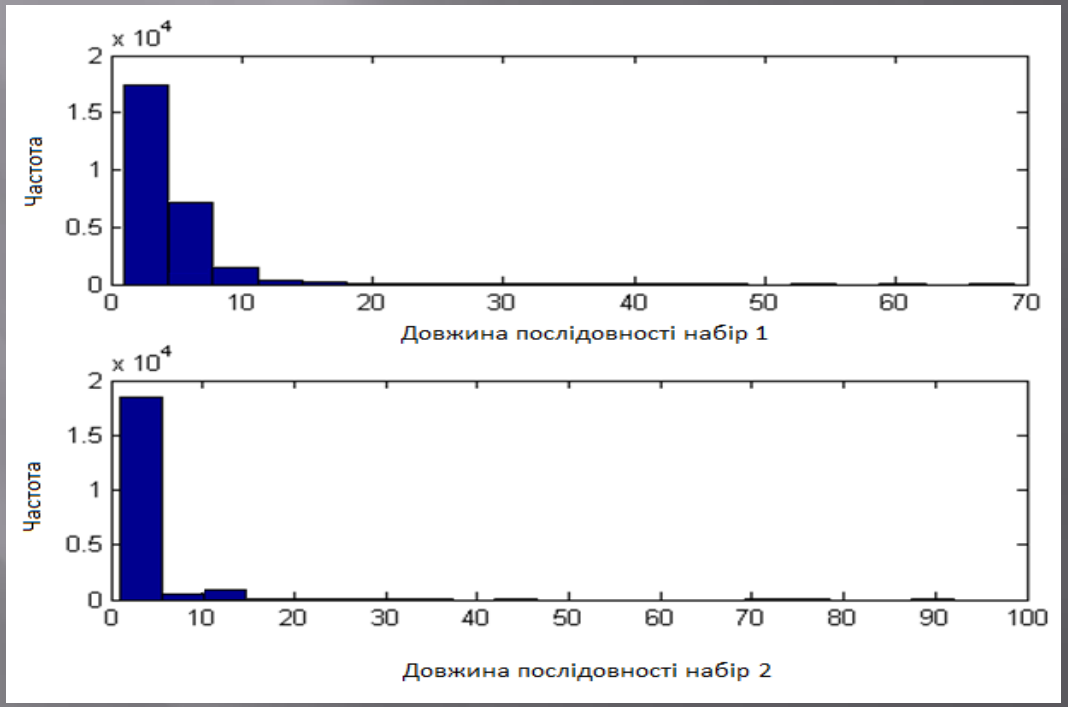
с	С	п	П
0	0	80	20

с	С	п	П
50	50	0	0

Приклад ПММ – для випадку спроба, підказка



Гістограма для Набір 1 з невипадковими цілями



Порівняння розподілу довжин послідовностей за розділами даних

Основні результати роботі :

Обгрунтовано кластеризацію цілей, як існування прикладної проблеми, яку неможливо вирішити традиційними методами.

Доведено доцільність і ефективність цільової кластеризації що проявляється в здатності цільової кластеризації до навчання прогностичних моделей на реальних і змодельованих наборах даних.

Алгоритми цільової кластеризації створюють схожі моделі, незважаючи на різні початкові умови, параметри і навіть набори даних.

Зважені ПММ, які засновані на функції максимізації, є надійним, ефективним алгоритмом, який створює інтерпретовані моделі.

Запропоновані методи показали перспективність як засобу зворотного зв'язку щодо змін в дизайні, новим освітнім теоріям та інше. Це також цілком прийнятне рішення, яке може бути використано багатьма дослідниками для вирішення проблем послідовності освоєння та конструювання інтерфейсу.

Дякую за увагу

Tue Nov 23 10:20:26 EET 2021, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism v-15.257

Максимальное совпадение с одним документом 46.0%

Словари проверки: en_US, ru_RU, ua_UA. **Ошибок в документах: 6%**

	Документ		Суммарное совпадение по Базе Данных	
ID: 97032 Название: Метод агрегативної кластеризації на базі послідовного підходу Добавлено в БД: 2021-11-23 Авторы: Е.А. Манзюк Руководители: О.В. Бармак Консультанты: Опоненты:	Символы	Лексемы	Символы	Лексемы
	108181	772	49972 (46%)	330 (43%)

Источник плагиата

ID	Описание	Наличие плагиата в документе	
		Символы	Лексемы
95909	Название: ЗВІТ з професійної практики Добавлено в БД: 2021-09-30 Авторы: Манзюк Е.А. Руководители: Скрипник Т.К. Консультанты: Опоненты:	49460 (46.0%)	358 (46.0%)

UNICHECK

Ім'я користувача:
Кафедра КН

ID перевірки:
1009309535

Дата перевірки:
23.11.2021 11:46:22 EET

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
23.11.2021 11:53:24 EET

ID користувача:
100005671

Назва документа: **Дип_Манзюк Е.А. Lite**

Кількість сторінок: **76** Кількість слів: **15862** Кількість символів: **121258** Розмір файлу: **494.90 KB** ID файлу: **1009334638**

3.45%
Схожість

Найбільша схожість: **1.9% з джерелом з Бібліотеки** (ID файлу: **1009334742**)

1.6% Джерела з Інтернету	114	Сторінка 78
2.03% Джерела з Бібліотеки	58	Сторінка 78

0% Цитат

Вилучення цитат вимкнене

Вилучення списку бібліографічних посилань вимкнене

0%
Вилучень

Немає вилучених джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Замінені символи	15

РІШЕННЯ ЕКСПЕРНОЇ КОМІСІЇ

КАФЕДРИ КОМП'ЮТЕРНИХ НАУК

ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ МАГІСТРА ДО ЗАХИСТУ ЗА РЕЗУЛЬТАТАМИ АНАЛІЗУ ЗВІТУ ПОДІБНОСТІ

Підтверджуємо ознайомлення з результатом звіту подібності щодо роботи, генерованого системою виявлення текстових збігів/ідентичності/схожості:

Назва: *Метод агрегативної кластеризації на базі послідовного підходу*

Автор: *Манзюк Едуард Андрійович*

Спеціальність: 122 – Компютерні науки

Освітня програма: освітньо-професійна

Науковий керівник: д.т.н., проф. Олександр Бармак

Після аналізу звіту подібності зроблено такий висновок:

№	Висновок	Позначка про відповідність
1	Запозичення, виявлені в роботі, є законними і не є плагіатом. Робота приймається до захисту.	відповідає
2	Виявлені запозичення не є плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи Робота приймається до захисту, але має бути відкоригована. Відкоригований варіант має бути поданий на кафедру за 2 дні до захисту, разом із заявою щодо самостійності виконання письмової роботи та ідентичності друкованої та електронної версії роботи	
3	Виявлені запозичення не є плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. В зв'язку з цим мета роботи та поставлені завдання не були досягнені. Робота може бути допущена до захисту (наступного року) після того як буде відкоригована та допрацьована і успішно пройде повторну перевірку на академічний плагіат.	
4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	

Підтвердження:

Запозичення, виявлені в роботі, є законними і не є плагіатом, оскільки:

1) за програмою Anti-Plagiarism виявлені 46% запозичень вказують на документ автора роботи Манзюка Е.А. та містять ЗВІТ з науково-дослідної практики.
2) За програмою UNICHECK виявлені 3.45% є фрагментарними – містять поширені конструкції, загальновідомі терміни, скорочення та визначення.

Сумарний обсяг всіх запозичень, визначений системою виявлення збігів/ ідентичності/схожості, складає 46% і 3.45% відповідно, що, з урахуванням наведених обґрунтувань, відповідає характеру наукового дослідження і свідчить на користь кваліфікаційної роботи.

Керівник роботи _____ Олександр Бармак

Гарант ОП _____ Руслан Багрій

Завідувач кафедри КН _____ Олександр Бармак

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ
кафедра комп'ютерних наук

ВІДГУК ОПОНЕНТА

на кваліфікаційну роботу магістра

гр. КНм-20-2 Манзюк Едуард Андрійович за темою: *Метод агрегативної кластеризації на базі послідовного підходу*

1. Актуальність обраної теми

Тема кваліфікаційної робота є актуальною та відповідає сучасному рівню досліджень предметної області. В роботі на належному рівні представлено обґрунтування та проведений огляд досліджень в напрямку обраної теми.

2. Відповідність роботи предметній області спеціальності 122 Комп'ютерні науки та загальним вимогам до наукових робіт

Тема кваліфікаційної роботи та її реалізація відповідає предметній області спеціальності 122 Комп'ютерні науки а також відповідає вимогам до наукових робіт освітньо-кваліфікаційного рівня магістри.

3. Повнота розкриття мети та завдань дослідження

Завдання досліджень розкривають поставлену мету кваліфікаційної роботи та повною мірою представлені в роботі.

4. Наявність наукової новизни

Запропоновані в роботі методи з цільової кластеризації та кластеризації послідовних структур мають наукову новизну та відповідають кваліфікаційному рівню магістра. Результати дослідження оприлюднені на науковій конференції.

5. Зміст кожного розділу роботи

Робота містить чотири розділи. В першому розділі подано обґрунтування актуальності вибраної теми, проведено дослідження сучасних близьких до теми наукових робіт, поставлено завдання дослідження. Наступний розділ присвячений розробці кластерної моделі даних. В третьому розділі представлена розробка метода кластеризації послідовних структур. Четвертий розділ містить дослідження ефективності запропонованих методів. Робота також містить висновки до кожного розділу та загальні висновки, список використаний джерел.

6. Ступінь розкриття теми роботи

Тема наукового дослідження належним чином розкрита в логічній та послідовній структурі представлення. Тема в достатній мірі обґрунтована та досліджено сучасний рівень

наукових робіт. Поставлені завдання реалізовані та проведено дослідження ефективності запропонованих методів.

7. Якість оформлення кваліфікаційної роботи

Оформлення кваліфікаційної роботи здійснено у відповідності до необхідних норм та правил.

8. Недоліки кваліфікаційної роботи

Доцільно було б розширити область даних для порівняння ефективності. Виявлені недоліки стосуються аспектів оформлення та не впливають на зміст роботи.

9. Загальний висновок (допускається чи не допускається до захисту), якої оцінки заслуговує кваліфікаційна робота.

Враховуючи рівень виконання та забезпечення усіх необхідних вимог робота може бути допущена до захисту. Рекомендована оцінка «відмінно».

Опонент д.т.н., проф. Тетяна Говорущенко

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ
кафедра комп'ютерних наук

ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу магістра

гр. КНм-20-2 Манзюк Едуард Андрійович за темою: *Метод агрегативної кластеризації на базі послідовного підходу*

1. Актуальність теми

В магістерській роботі було розроблено та набув практичної реалізації метод агрегативної кластеризації на базі послідовного підходу. В роботі достатньою мірою обґрунтована актуальність за проведеним аналізом сучасних наукових досліджень у відповідній предметній області. На основі чого визначено напрямок дослідження та поставлено задачі.

2. Відповідність роботи предметній області спеціальності 122 Комп'ютерні науки та загальним вимогам до наукових робіт

За мірою розкриття мети дослідження, рівнем запропонованих методів, логічною послідовністю, проведеними експериментальними дослідженнями робота відповідає вимога до наукових робіт освітньо-кваліфікаційного рівня магістр. Мета, завдання, об'єкт та предмет дослідження відповідають предметній області спеціальності 122 Комп'ютерні науки та вимогам до кваліфікаційної роботи магістра.

3. Професійні та особистісні якості магістранта

Рівень набутих компетенцій продемонстрований у кваліфікаційній роботі показує належний рівень у вирішенні наукових задач та за сукупністю ознак відповідає вимогам що ставляться до професійних якостей магістрів.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Кваліфікаційна робота виконана магістром самостійно. Визначенні завдання дослідження, розроблено моделі та методи а також проведені експериментальні дослідження для встановлення ефективності запропонованих методів.

5. Наукова новизна та оригінальність запропонованих підходів

В роботі присутня наукова новизна. Набула подальшого розвитку система формування кластерних структур, які залежні від цілі агрегатування даних. Запропоновано інноваційний метод розробки систем цільової кластеризації. Запропоновано метод формування кластерів на множинах результатів моделей за

прихованими Марківськими процесами. Результати досліджень оприлюдненні на науковій конференції.

6. Ступінь оволодіння методами дослідження

Під час виконання наукових досліджень продемонстровано належний рівень володіння методами наукового пізнання. Також продемонстровано практичне втілення набутих компетенції рівня магістра.

7. Повнота та якість розкриття теми роботи

Тема робота в повній мірі розкрита проведеними дослідженнями на належному науковому рівні, який відповідає освітньо-кваліфікаційному рівню магістра.

8. Логічність, послідовність, аргументованість, літературна грамотність викладу матеріалу

Кваліфікаційна робота магістра виконана з дотриманням та у відповідності до вимог щодо наукової складової, послідовності викладення матеріалу, рівню аргументованості, наукового обґрунтування, доведення та перевірки пропонованих методів. За стилістичним викладенням робота відповідає науковому рівню.

9. Можливість практичного застосування кваліфікаційної роботи, окремих її частин

Практичне значення роботи полягає в розроблені прикладних методів визначення якості навчальних процесів, які подані у вигляді послідовних даних та орієнтовані на визначення оптимальних послідовностей у вивчені матеріалу.

10. Висновок про можливість допуску кваліфікаційної роботи до захисту, на яку оцінку заслуговує робота

Кваліфікаційна робота виконана на належному рівні і може бути допущена до захисту та заслуговує на оцінку «відміно».

Науковий керівник _____ д.т.н., проф. Олександр Бармак