

МІНІСТЕРСТВО ВНУТРІШНІХ СПРАВ УКРАЇНИ
Харківський національний університет внутрішніх справ

Факультет № 4

Кафедра інформаційних технологій

ТЕКСТ ЛЕКЦІЇ

**з дисципліни «Моделі, методи та засоби аналітичної обробки
великих масивів даних»**

за темою «Методи і алгоритми нечіткої кластеризації»

Галузь знань: 12 "Інформаційні технології "

Спеціальність: 125 "Кібербезпека"

Ступінь вищої освіти - магістр

**м. Харків
2017 р.**

Передмова

СХВАЛЕНО

Науково-методичною радою ХНУВС

_____ Протокол № _____

(дата, місяць, рік)

ЗАТВЕРДЖЕНО

Вченою радою факультету № 4

ХНУВС

_____ Протокол № _____

(дата, місяць, рік)

_____ (підпис)

_____ (П.І.Б.)

ПОГОДЖЕНО

Секцією Науково-методичної ради
ХНУВС з технічних дисциплін

_____ Протокол № _____

(дата, місяць, рік)

_____ (підпис)

_____ (П.І.Б.)

ЗАТВЕРДЖЕНО

На засіданні кафедри інформаційних
технологій

_____ Протокол № _____

(дата, місяць, рік)

_____ (підпис)

_____ (П.І.Б.)

Рецензент:

Зацеркляний М.М., доктор технічних наук, професор;

Розробники: Струков В.М. – Харків: Харківський національний університет
внутрішніх справ, 2017

© Струков В.М., 2017

© Харківський національний
університет внутрішніх справ

План лекції

1. Нечіткий алгоритм K-means.
2. Нечіткий алгоритм C-means.
3. Нейромережеві алгоритми.

Література:

Основна:

1. Чубукова, И.А. Data Mining [Текст] / И.А. Чубукова // Интернет-университет информационных технологий, Бином, Лаборатория знаний, 2008. - 382 с.
2. Aggarwal C.C. Data Mining. – Cham: Springer Ltd. Publ. Switzerland, 2015. – 734p.
3. Aggarwal C.C., Reddy C.K. Data Clustering. Algorithms and Applications.- New York: CRC Press, Taylor & Francik Group, 2014. – 648p.
4. Конспект лекцій.

Додаткова:

5. Люгер, Д. Искусственный интеллект: стратегии и методы решения сложных проблем [Текст] / Д. Люгер. - Издательский дом «Вильяме», 4е изд. М.: - 2003. - 864 с.

Текст лекції

Нечіткі алгоритми

Багато плоскі чіткі алгоритми мають нечіткі варіації. Найбільш популярними алгоритмами нечіткої кластеризації є нечіткий алгоритм с-середніх (нечітка варіація алгоритму с-середніх) і його модифікації. Цільовою функцією алгоритму с-середніх є функція:

$$J_{\alpha} = \sum_{i=1}^{N_C} \sum_{j=1}^{N_D} \mu_{ij}^{\alpha} \left\| \vec{d}_j - \vec{c}_i \right\|^2,$$

Знаходження матриці нечіткого розбиття з мінімальним значенням цільової функції являє собою задачу нелінійної оптимізації. В основу алгоритму с-середніх для вирішення цього завдання покладено алгоритм невизначених множників Лагранжа. Він дозволяє знайти локальний оптимум, тому виконання алгоритму з різних початкових точок може призвести до різних результатів.

Перевагою даного алгоритму є те, що отримані кластери є пересічними.

1. Нечіткий алгоритм K-means.

Багаторазові спроби класифікації методів кластерного аналізу призводять до десятків, а то і сотень різноманітних класів. Таке різноманіття породжується великою кількістю можливих способів обчислення відстані між окремими спостереженнями, такою ж кількістю методів обчислення відстані між окремими кластерами в процесі кластеризації і різноманітними оцінками оптимальності кінцевої кластерної структури. Найбільшого поширення в популярних статистичних пакетах отримали два групи алгоритмів кластерного аналізу: ієрархічні агломеративні методи і ітеративні методи угруповання. Тим, хто намагається в своїй дослідницькій практиці застосовувати ті чи інші методи багатовимірної статистики, в тому числі і кластерний аналіз, слід пам'ятати, що отримуються при цьому результати не є єдиними, унікальними. Їм необхідно розуміти, що отриманий результат є породженням одного з багатьох можливих варіантів. І перевагу цього результату, а отже і методу аналізу, перед іншими ще слід оцінити, а можливо і довести своїм колегам або іншим зацікавленим особам.

У агломеративного-ієрархічних методах (agglomerative hierarchical algorithms), які більш часто використовуються в реальних біомедичних дослідженнях, спочатку всі об'єкти (спостереження) розглядаються як окремі, самостійні кластери складаються всього лише з одного елемента. Якщо прийняти, що обсяг вибірки дорівнює N , то в цьому випадку можна використовуючи ту чи іншу метрику, обчислити відстані між усіма можливими парами об'єктів. Таких відстаней буде $N * N$. Наприклад, для 105 пацієнтів буде обчислено $105 * 105 = 11025$ взаємних парних відстаней. Це будуть відстані для наступних пар:

1-1; 1-2; 1-3; 1-4; 1-103; 1-104; 1-105;
2-1; 2-2; 2-3; 2-4; 2-103; 2-104; 2-105;
.....;
.....;
103-1; 103-2;103-103; 103-104; 103-105;
104-1; 104-2;104-103; 104-104; 104-105;
105-1; 105-2;105-103; 105-104; 105-105.

З урахуванням того, що $d_{ii} = 0$, і що $d_{ij} = d_{ji}$ загальна кількість різних обчислюваних відстаней дорівнюватиме $N * (N-1) / 2$, що при $N = 105$ дорівнюватиме 5460, що приблизно вдвічі менше початкового числа. Однак це значно більше 10 відстаней, які були наведені в навчальному прикладі вище. Далі, з урахуванням того, що в реальних даних використовуються не дві ознаки, як в тому ж навчальному прикладі, а десятки, а іноді і сотні, можна уявити який великий обсяг обчислень необхідно виконати навіть для цієї найпростішої операції. Очевидно, що без використання потужної обчислювальної техніки реалізація кластерного аналізу даних вельми

проблематична.

Нагадаємо, що ця матриця відстаней може бути отримана за допомогою різноманітних метрик: евклідової, Махаланобіса, сімейства метрик Маньківського і т.д. Вибір метрики здійснюється самим дослідником. Після обчислення матриці відстаней починається процес агломерації (від латинського *agglomerare* - приєдную, нагромаджую), що проходить послідовно крок за кроком. На першому кроці цього процесу два вихідних спостереження (монокластера), між якими саме мінімальну відстань, об'єднуються в один кластер, що складається вже з двох об'єктів (спостережень). Таким чином, замість колишніх N монокластерів (кластерів, що складаються з одного об'єкта) після першого кроку залишиться $N-1$ кластерів, з яких один кластер буде містити два об'єкти (спостереження), а $N-2$ кластерів будуть як і раніше складатися всього лише з одного об'єкта. Відзначимо, що на другому етапі можливі різні методи об'єднання між собою $N-2$ кластерів. Це викликано тим, що один з цих кластерів вже містить два об'єкти. З цієї причини виникає два основних питання:

- як обчислювати координати такого кластера з двох (а далі і більше двох) об'єктів;
- як обчислювати відстань до таких "поліоб'єктних" кластерів від "монокластерів" і між "поліоб'єктними" кластерами.

Ці зовсім не риторичні питання, в кінцевому рахунку, і визначають остаточну структуру підсумкових кластерів (під структурою кластерів мається на увазі склад окремих кластерів і їх взаємне розташування в багатовимірному просторі). Різноманітні комбінації метрик і методів обчислення координат і взаємних відстаней кластерів і породжують ту різноманітність методів кластерного аналізу, про який було сказано вище. На другому кроці в залежності від обраних методів обчислення координат кластера що складається з декількох об'єктів і способу обчислення межкластерних відстаней можливе або повторне об'єднання двох окремих спостережень в новий кластер, або приєднання одного нового спостереження до кластеру, що складається з двох об'єктів. Для зручності більшість програм агломеративного-ієрархічних методів після закінчення роботи можуть надати для перегляду два основних графіка. Перший графік називається Дендрограма (від грецького *dendron* - дерево), що відображає процес агломерації, злиття окремих спостережень в єдиний остаточний кластер. Цей графік схематично нагадує дерево, за що і отримав таку назву. Нижче наведено малюнок з такою Дендрограма для нашого навчального прикладу складається з 5 спостережень за двома змінним.

Вертикальна вісь такого графіка являє собою вісь межкластерної відстані, а по горизонтальній осі відзначені номери об'єктів - випадків (cases), використаних в аналізі. З цієї дендрограми видно, що спочатку об'єднуються в один кластер об'єкти №1 і №2, оскільки відстань між ними сама мінімальна і дорівнює 1. Це злиття відображається на графіку горизонтальною лінією що з'єднує вертикальні відрізки виходять з точок

помічених як C_1 і C_2 . Звернемо увагу на те, що сама горизонтальна лінія проходить точно на рівні межкластерної відстані рівної 1. Далі на другому кроці до цього кластеру, що включає в себе вже два об'єкти, приєднується об'єкт №3, позначений як C_3 . На наступному кроці відбувається об'єднання об'єктів №4 та №5, відстань між якими дорівнює 1,41. І на останньому кроці відбувається об'єднання кластера з об'єктів 1, 2 і 3 з кластером з об'єктів 4 і 5. На графіку видно, що відстань між цими двома передостанніми кластерами (останній кластер включає в себе всі 5 об'єктів) більше 5, але менше 6, оскільки верхня горизонтальна лінія з'єднує два передостанніх кластера проходить на рівні приблизно рівному 7, а рівень з'єднання об'єктів 4 і 5 дорівнює 1,41.

Розташована нижче дендрограма отримана при аналізі реального масиву даних складається з 70 об'єктів, кожен з яких характеризувався 12 ознаками - електронномікроскопічними характеристиками еритроцитів дітей з хворобою щитовидною залозою.

З графіка видно, що на останньому кроці, коли відбулося злиття двох останніх кластерів, відстань між ними близько 200 одиниць. Видно, що перший кластер (домовимося, що він розташований зліва) включає в себе набагато менше об'єктів (9), ніж другий кластер (розташований праворуч). Оскільки всього в аналізі використано 70 об'єктів, то в другому кластері 61 об'єкт.

Другий графік, який будується в таких процедурах - це графік зміни межкластерних відстаней на кожному кроці об'єднання. Нижче наведено подібний графік для наведеної вище дендрограми.

Завершуючи знайомство з ієрархічними методами, відзначимо, що агломеративні (об'єднують) методи на останньому кроці об'єднують всі спостереження в одні кластер. Тому використовувати побудовану Дендрограму для виділення тієї чи іншої кількості окремих кластерів можна шляхом "розрізання" цієї дендрограми на певному значенні межкластерної відстані. Фактично це означає, що ми проводимо горизонтальну лінію, розсікаючи дерево зв'язків в тому місці, де спостерігається максимальний стрибок у зміні межкластерної відстані. Досить зручним сервісом, який надається при цьому в ряді статистичних пакетів, є обчислення основних статистичних характеристик кластерів, утворених шляхом розрізання дендрограми, таких як кількість об'єктів в кластері, середні значення ознак в кожному кластері, дисперсії і т.д. В інших пакетах є можливість "трасування" входження окремих спостережень в кластери, шляхом виділення кольором тих ділянок дендрограми, які відповідають проміжним кластерам містить це спостереження. Не менш зручно і масштабування і виділення в окремі вікна конкретних ділянок дендрограми і т.д. У деяких статистичних пакетах в ієрархічних процедурах задається кінцеве число кластерів, при досягненні якого подальше побудова дендрограми припиняється.

Крім об'єднуючих методів ієрархічної кластеризації існують і

протилежні методи - дівізімніе, в яких на початковому етапі вся вибірка розглядається як єдиний кластер, а потім вже починається процес його розподілу на складові частини. Процес поділу триває до тих пір, поки кожне спостереження чи не перетвориться в окремий кластер. У свою чергу дівізімніе алгоритми діляться на монотетичні і політетичні. У монотетичній класифікації розподіл проводиться на підставі єдиної ознаки, що має максимальну інформативність. У політетичних же алгоритмах враховуються всі ознаки. Оскільки дані алгоритми оперують відстанями між спостереженнями, то в деяких програмах передбачена можливість роботи не з вихідної матрицею "об'єкт - ознака", а з симетричною матрицею відстаней між спостереженнями.

2. Нечіткий алгоритм C-means.

Серед ітераційних методів найбільш популярним методом є метод Мак-Кіна. На відміну від ієрархічних методів в більшості реалізацій цього методу сам користувач повинен задати шукане число кінцевих кластерів, яке зазвичай позначається як "k". Як і в ієрархічних методах кластеризації, користувач при цьому може вибрати той чи інший тип метрики. Різні алгоритми методу k-середніх відрізняються і способом вибору початкових центрів кластерів. У деяких варіантах методу сам користувач може (або повинен) задати такі початкові точки, або вибравши їх із реальних спостережень, або задавши координати цих точок по кожній із змінних. В інших реалізаціях цього методу вибір заданого числа k початкових точок проводиться випадковим чином, причому ці початкові точки (зерна кластерів) можуть в подальшому уточнюватися в кілька етапів. Є і інші способи задання початкових центрів кластерів. Особливо це відіграє важливу роль у випадках, коли кількість спостережень нараховує десятки тисяч і більше. В цих випадках вибір заданого числа k початкових точок проводиться випадковим чином заздалегідь не є ефективним. Тому різні евристичні підходи – за принципом найбільш віддалених точок, мінімальної дисперсії відстаней між початковими точками.

Можна виділити 4 основних етапи таких методів:

- вибираються або призначаються k спостережень, які будуть первинними центрами кластерів;
- при необхідності формуються проміжні кластери приписуванням кожного спостереження до найближчих заданим кластерним центрам;
- після призначення всіх спостережень окремим кластерам проводиться заміна первинних кластерних центрів на кластерні середні;
- попередня ітерація повторюється до тих пір, поки зміни координат кластерних центрів не стануть мінімальними.

У деяких варіантах цього методу користувач може задати числове значення критерію, трактують як мінімальна відстань для відбору нових центрів кластерів. Спостереження не розглядатиметься як претендент на новий центр кластера, якщо його відстань до замінного центру кластера

перевищує вказану кількість. Такий параметр в ряді програм називається "радіусом". Крім цього параметра можливе завдання і максимального числа ітерацій або досягнення певного, зазвичай досить малого, числа, з яким порівнюється зміна відстані для всіх кластерних центрів. Цей параметр зазвичай називається "конвергенцією", тому що відображає збіжність ітераційного процесу кластеризації. Нижче ми наведемо частину результатів, які отримані при використанні методу k-середніх Мак-Кіна до попередніх даних. Число шуканих кластерів задавалося спочатку рівним 3, а потім - 2. Перша їх частина містить результати однофакторного дисперсійного аналізу (10, 18), в якому в якості групуючого фактора виступає номер кластера. У першому стовпчику - список 12 змінних, далі йдуть суми квадратів (SS) і ступені свободи (df), потім F-критерій Фішера і в останньому стовпчику - досягнутий рівень значимості "p".