МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
СХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМ. В. ДАЛЯ
ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА ЕЛЕКТРОНІКИ
КАФЕДРА КОМП'ЮТЕРНИХ НАУК ТА ІНЖЕНЕРІЇ

До захисту допускається
Завідувач кафедри
_____ Скарга-Бандурова І.С.
«____»_____ 20__ р.

МАГІСТЕРСЬКА РОБОТА

НА ТЕМУ:

Комп'ютерна система прогнозування ходу лікування

Освітній ступінь"Магістр"
Спеціальність 123 "Комп'ютерна інженерія"

Науковий керівник роботи:	_____ (підпис)	О.І.Рязанцев (ініціали, прізвище)
Консультант з охорони праці:	_____ (підпис)	Я.О.Критська (ініціали, прізвище)
Студент:	_____ (підпис)	Є.С.Зуєв (ініціали, прізвище)
Група:		КІ-17змі

Сєвєродонецьк 2019

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
СХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВОЛОДИМИРА ДАЛЯ

Факультет Інформаційних технологій та електроніки
Кафедра Комп'ютерних наук та інженерії
Освітній ступінь магістр
Напрям підготовки _____
(шифр і назва)
Спеціальність 123 "Комп'ютерна інженерія"
(шифр і назва)

ЗАТВЕРДЖУЮ:

Завідувач кафедри _____
_____ І.С. Скарга-Бандурова
«_____» _____20____р.

З А В Д А Н Н Я
НА МАГІСТЕРСЬКУ РОБОТУ СТУДЕНТУ

Зуєву Євгенію Сергійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Комп'ютерна система прогнозування ходу лікування

керівник проекту (роботи) Рязанцев Олександр Іванович, д.т.н., проф.
(прізвище, м.'я, по батькові, науковий ступінь, вчене звання)
затверджені наказом вищого навчального закладу від «18» 10 2018 р. № 221/48
2. Строк подання студентом роботи 10.01.2018
3. Вихідні дані до роботи Матеріали науково-дослідної практики, медичні дані у вигляді бази даних форм CMS-1500, методи інтелектуального аналізу даних, теоретичні відомості про методи кластеризації даних, бібліотека SPMF , середа розробки Eclipse Kepler

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити) Сучасний стан інформаційних технологій у медицині, аналіз даних та побудова прогнозної моделі, програмна реалізація й дослідження алгоритмів, охорона праці та безпека в надзвичайних ситуаціях, висновки

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень)
Електронні плакати

6. Консультанти розділів проекту (роботи)

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Охорона праці та безпека в надзвичайних ситуаціях	Критська Я.О. ст. викл. кафедри КНІ		

7. Дата видачі завдання 18.10.2018

Керівник _____
(підпис)

Завдання прийняв до виконання _____
(підпис)

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів дипломного проекту (роботи)	Строк виконання етапів проекту (роботи)	Примітка
1	Отримання завдання на дипломування	10.09.2018-15.09.2018	
2	Аналіз завдання, робота з літературою	16.09.2018-22.09.2018	
3	Розробка програмної системи	26.09.2018-06.10.2018	
4	Тестування програмної системи	07.10.2018-25.11.2018	
5	Розробка частини проекту "Охорона праці та безпеки в надзвичайних ситуаціях"	26.11.2018-1.12.2018	
6	Оформлення пояснювальної записки, автореферату та презентації	2.12.2018-09.01.2019	
7			

Студент _____ Зуєв Є.С.
(підпис) (прізвище та ініціали)

Науковий керівник _____ Рязанцев О.І.
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Зуєв Є.С. Комп'ютерна система прогнозування ходу лікування.

Робота присвячена дослідженню алгоритмів аналізу медичних даних пацієнтів. Аналіз проводиться з метою побудови прогнозної моделі для виявлення подальших потреб пацієнтів в медикаментах і різних медичних процедурах. Аналіз даних полягає у формуванні множин з пацієнтів, які перебувають у зоні ризику, і виявлення факторів, які призвели до даного результату. Для подібного роду аналізу в роботі використовується кластеризація. У роботі розглянуті і проаналізовані найбільш відомі алгоритми кластеризації, такі як K-Means, Fuzzy C-Means, PAM, CLOPE, DBSCAN. Виявлено алгоритми з найбільшою точністю з різними параметрами.

Ключові слова: кластеризація, алгоритм, моделювання, прогнозування, machine learning, data mining, медицина

ABSTRACT

Zuev E.S. Computer system for predicting the course of treatment.

Master’s degree thesis is devoted to the study of algorithms analysis of medical data of patients. The analysis is done in order to build a predictive model to identify further needs of patients in a variety of medicines and medical procedures. Analysis of the data is to form sets of patients who are in the risk zone, and identify the factors that led to this result. For this type of analysis used in the clustering. The paper discusses and analyzes the most well-known clustering algorithms such as K-Means, Fuzzy C-Means, PAM, CLOPE, DBSCAN. Identified with the greatest accuracy algorithms with different parameters.

Keywords: clustering, algorithms, modeling, forecasting, machine learning, data mining, medicine

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ПМ – прогнозна модель

АК – алгоритм кластеризації

ІТ – інформаційні технології

DM – data mining

ВСТУП

Сучасну медицину неможливо уявити без інформаційних технологій. За час використання різноманітних медичних інформаційних систем накопичилася величезна кількість даних, у тому числі про пацієнтів, їх захворюваннях, методи лікування цих захворювань, результати та відхиленнях від прописаного лікування, препарати, які використовувалися, і їх дозуваннях, а також аналізах і процедурах, коли-небудь пройдених пацієнтами, їх результати та багатьох інших. Крім цього, в базах даних зберігається різна інформація, надана самим пацієнтом про протікання хвороби і так далі. Актуальним представляється формування на основі аналітики цих даних систем знань, необхідних спеціалістам для різних аспектів їхньої діяльності, таких як: виявлення причин, прогнозування та діагностики захворювань; попередньої оцінки ризиків ускладнень в ході лікування; оцінки ефективності використовуваних методів лікування і лікарських засобів; планування кількості медичних установ, обладнання та персоналу і так далі.

Традиційно для класифікації та діагностики різних захворювань використовуються інтелектуальні системи . Але тільки зараз переваги прогнозної аналітики починають реалізовуватися в медичній сфері. У міру перекладу даних в цифровий формат в майбутньому з'являться все більше прогнозних рішень, покликаних контролювати стан пацієнтів у реанімації , виявляти шахрайство та зловживання та вирішувати інші завдання.

Аналіз існуючих аналітичних медичних систем показав, що обробка інформації про низькою або високою небезпеки виникнення певного захворювання або стану в групі пацієнтів за допомогою засобів інтелектуального аналізу або прогнозної аналітики допомагає лікарям створювати протоколи лікування з урахуванням особливостей різних груп. Таким чином, лікарі можуть створювати різні стратегії для контролю стану пацієнтів з групи невисокого ризику і скорочення ризику для життя пацієнтів з групи високого ризику.

При вирішенні подібного класу задач використовується прогнозне моделювання. Прогнозне моделювання являє собою набір методологій і технологій аналізу даних, що базуються на методах штучного інтелекту та інструментах управління прийняття рішень. У процесі моделювання будується прогнозна модель. Прогнозна модель - це модель явища, яке матиме місце в майбутньому. Її дослідження дозволяє отримати інформацію про можливі стани об'єкта в перспективі або шляхи досягнення цих станів. Складність

розробки прогнозної моделі, на відміну від моделі, відповідної явищу, яке існує в даний час, полягає в необхідності врахування всіх можливих змін, які можуть статися з об'єктом прогнозування в майбутньому, передбачити дії різних факторів (зовнішніх і внутрішніх), здатних кардинально змінити його розвиток і функціонування.

1 СУЧАСНИЙ СТАН ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ У МЕДИЦИНІ

1.1 Інтеграція інформаційних технологій та медицини

Інформаційні технології - це процеси, що використовують сукупність засобів і методів збору, обробки і передачі даних (первинної інформації) для отримання інформації нової якості про стан об'єкта, процесу або явища (інформаційного продукту). І ці процеси не оминули й питання сучасної медицини. ІТ значною мірою вплинули на якість та швидкість більшості процесів.

Інформатизація суспільства - це глобальний соціальний процес, особливість якого полягає в тому, що домінуючим видом діяльності в сфері суспільного виробництва є збір, накопичення, обробка, зберігання, передача, використання, продукування інформації, здійснювані на основі сучасних засобів мікропроцесорної та обчислювальної техніки, а також різноманітних засобів інформаційної взаємодії та обміну.

Інформаційні технології можна розглядати як елемент і функцію інформаційного суспільства, спрямовану на регулювання, збереження, підтримку та вдосконалення системи управління нового мережевого суспільства. Якщо протягом століть інформація і знання передавалися на основі правил і приписів, традицій і звичаїв, культурних зразків і стереотипів, то сьогодні головна роль відводиться технологіям.

Основні парадигми розвитку технологій:

– швидше: всі медичні процеси прискорюються. Наприклад, ЕКГ можна зробити вдома, просто приклавши до себе мобільний телефон із датчиком і переслати дані до лікаря за лічені секунди;

– менше: пристрої, за допомогою яких проводиться діагностування та моніторинг, зменшуватися в розмірі;

– дешевше: чим більше точність тестів, тим менше їх потрібно робити. Чим раніше можна зловити ознаки хвороби, тим менше грошей піде на її лікування. По суті, медицина стане в більшості превентивної (ця тенденція спостерігається вже з початку минулого століття);

– якісніше: чим дешевше і швидше все стає, тим більше можливостей для впровадження персоналізованої медицини, заснованої на генокоді людини.

Сучасні проблеми:

– висока вартість охорони здоров'я (медицина може бути в рази дешевше);

– низька доступність медичної допомоги (вона повинна бути швидше і вже на місці);

– різниця медичних підходів.;

– неефективне використання інформації про пацієнта.;

– роздробленість серед самих докторів: часто не спілкуються між собою;

– будь-який рух вперед обмежена регулятивними огранами і недоліком в коштах;

– будова системи охорони здоров'я заплутано і непродумано.

Завдяки ІТ ми зараз маємо набагато більше можливостей:

– персоналізована медицина: вибір методів лікування та препаратів заснований на індивідуальних особливостях пацієнта;

– віддалена медицина: роботи телеприсутності, віддалені хірургічні роботи, додатки для діагностування (наприклад, ScinScan). Це дозволить медицині розвинених країн стати доступною для віддалених куточків планети;

– молекулярна і генетична терапія для запобігання і лікування хвороб.

Також очікується швидке зростання інформаційної бази - це дає можливість аналізувати великі обсяги інформації (від окремих пацієнтів + краудсорс-джерел для виявлення взаємозв'язків). Один із прикладів - зниження вартості і доступність генетичних досліджень для населення в цілому, що призведе до різкого збільшення кількості практичної інформації і до серйозних проривів у медицині.

Основними принципами створення інформаційної системи охорони здоров'я:

– одноразовий введення і багаторазове використання первинної інформації;

– забезпечення сумісності різних медичних інформаційних систем;

– забезпечення інформаційної безпеки та захисту персональних даних;

– застосування єдиної технологічної політики з урахуванням галузевих державних, національних та адаптованих до вітчизняних умов міжнародних стандартів.

За підсумками ряду клінічних досліджень, в тому числі, в медичних організаціях Канади, Італії, Австралії, Англії та Німеччини (2004-2009 р), були виявлені наступні позитивні тенденції впровадження ІТ в практичній охороні здоров'я:

– поліпшення прихильності до лікування, зокрема за даними частка пацієнтів, які активно використовують методи домашнього самоконтролю, збільшується до 90%;

– зниження частоти госпіталізацій пацієнтів;

– поліпшення якості життя, психологічного та соціального стояння пацієнт;

– зниження смертності серед хворих з серцево-судинними захворюваннями на 20-25% порівняно з рутинної технологією організації медичної допомоги, тобто без використання ІТ систем;

– підвищення задоволеності хворих медичними послугами та поліпшення якості життя;

– підвищення інформованості пацієнтів про своє захворювання;

– поліпшення якості обслуговування, своєчасна корекція лікарської терапії, висока ефективність медикаментозного лікування;

– підвищення економічної ефективності медичної допомоги.

Усе це можливо завдяки переходу на новий рівень запису та зберігання інформації, тобто інформатизації медичних структур.

Інформаційні технології в охороні здоров'я оцінюються за двома критеріями, а саме: рішення конкретних соціально-економічних проблем та підвищення якості діагностики та лікування при зниженні собівартості. Це можливо, коли основою концепції медицини є орієнтація на пацієнта. У цьому випадку найбільш прийнятними є 3 напрямки в області інформаційних технологій: електронна історія хвороби, інформаційна структура здоров'я та телемедицина.

1.1.1 Електронна медична карта

Електронні медичні записи є однією з найбільш затребуваних сучасних ІТ інструментів, що дозволяють в одному місці концентрувати всю необхідну інформацію медичного характеру в одній базі даних.

Такий підхід дозволяє:

– лікарям отримувати вибірки по цікавлять критеріям, щоб виявляти оптимальні схеми лікування;

– швидко і оперативно освіжати інформацію з історії хвороби конкретного пацієнта;

– здійснювати підбір індивідуальних дозувань лікарських препаратів, що сприяє підвищенню ефективності лікування;

– знизити витрати на папір;

– унеможливити втрату медичних карт пацієнтів;

– автоматизувати передачу результатів аналізів з лабораторій лікуючим лікарям.

У квітні 2008 року офіс національного координатора медичних інформаційних технологій США опублікував документ «Defining Key Health Information Technology Terms». Цей документ став результатом широкої суспільної дискусії, спеціально проведеної напередодні нового етапу масового впровадження інформаційних систем в практику охорони здоров'я США. У цьому документі визначені 6 ключових понять, у тому числі:

– Electronic Medical Record, EMR - електронна інформація, пов'язана зі здоров'ям суб'єкта (пацієнта), яка створюється, зберігається, ведеться і використовується сертифікованими медичними фахівцями і персоналом в одній медичної організації;

– Electronic Health Record, EHR - електронна інформація, пов'язана зі здоров'ям суб'єкта (пацієнта), відповідна національним стандартам сумісності (інтероперабельності), яка створюється, ведеться і використовується сертифікованими медичними фахівцями і персоналом більш ніж однієї медичної організації;

– Personal Health Record, PHR - електронна інформація, пов'язана зі здоров'ям суб'єкта (пацієнта), відповідна національним стандартам сумісності (інтероперабельності), отримана з різних джерел, ведення, управління та надання доступу до яких здійснює сам суб'єкт (пацієнт).

Таким чином, і EMR, і EHR, і PHR - це набори інформації про конкретну людину, пов'язаної з його здоров'ям. Ці інформаційні набори відрізняються один від одного місцем збору інформації та способом управління нею. Важливо, що в сферу цих визначень входить інформація, пов'язана зі здоров'ям (а не тільки історія хвороби). До неї можуть відноситися і рахунки за медичні послуги, і питання дієтології, здорового способу життя та диспансеризації, фізіологічні параметри та особливості розвитку здорової людини.

Найважливіші 3 напрямки інформатизації охорони здоров'я:

– інформатизацію конкретних медичних організацій;

– створення інтеграційних проектів для обміну інформацією між медичними організаціями, аж до проектів національного і навіть міжнародного масштабу. Це завдання створення єдиного інформаційного простору для медиків (професіоналів);

– єдиний інформаційний простір для непрофесіоналів (пацієнтів і навіть тих, хто поки себе пацієнтом не відчуває), сервіси та засоби для ведення медичної інформації і електронної взаємодії лікарів та пацієнтів.

Всі три напрямки дуже важливі, повинні розвиватися паралельно і доповнювати один одного, але якщо перші два широко обговорюються, то про 3-му шляху в Росії практично не говорять. У той же час, саме третій шлях зараз стає найбільш обговорюваною темою на міжнародних конференціях з медичної інформатики.

1.1.2 Переваги залучення сучасних ІТ в медицину

Завдяки ефективному впровадженню сучасних інформаційних технологій в медичну сферу лікарі та медсестри перестають «мучити» серйозні обсяги паперу на ведення історій хвороби пацієнтів, формування звітів тощо. У керівної ланки медичних закладів з'являється можливість оптимізувати розподіл всіляких ресурсів, що знаходяться в їх розпорядженні. Завдяки організації медичних карток пацієнтів у формі конфіденційних медичних записів, у лікарів є можливість оперативно отримувати необхідні дані, знання яких дозволять швидко прийняти рішення щодо подальшого лікування, варіантів надання допомоги, організації ефективної профілактики та ін.

Зіставляючи витрати на перенесення паперової медичної інформації в електронний формат, витрати на розробку і експлуатацію спеціального програмного забезпечення для ефективної роботи медиків з даними в цифровому форматі, істотно нижче, ніж на аналогічні дії з паперовою документацією. Крім того, ефективність роботи медичного персоналу, коли вся цікавить інформація може бути доступна в лічені хвилини, досить істотна.

Сучасний рівень розробки спеціального програмного забезпечення для роботи медиків відповідають найвищим стандартам безпеки даних, розміщених у всесвітній павутині, що дозволяє здійснювати онлайн доступ до баз даних, що містять конфіденційні дані пацієнтів.

Ще однією істотною перевагою впровадження сучасних інформаційних технологій в медичних установах, особливо невеликих (районних, сільських і т.д.) є зменшення витрат на штат співробітників, в чиї функціональні обов'язки якраз входить робота з паперовими документами.

Не менш важливим позитивним наслідком впровадження ІТ в медицину є можливість взаємодіяти з іншими зовнішніми джерелами інформації завдяки онлайн-конференцій, симпозіумів та ін., Що дозволяє, не залишаючи пацієнта, вирішити складні питання за допомогою більш досвідчених колег, почути думки інших професіоналів на складну проблему . Це суттєва допомога для невеликих лікарень, розташованих на територіях віддалених від центру країни.

Однак розробка і впровадження сучасних ІТ йде не тільки в області полегшення і вдосконалення роботи медичного персоналу та медичного закладу в цілому. Не менш істотні розробки для пацієнтів. Так, сьогодні абоненти спеціальних медичних систем, мають можливість отримати допомогу кваліфікованого медика з питань здоров'я

практично 24 години на добу, не залишаючи дому. Працюючи у взаємодії зі страховими організаціями, пацієнти мають можливість замовити собі страховий поліс за допомогою ресурсів, під'єднаних до Інтернету, отримати роз'яснення за страховими програмами від фахівців, також, не залишаючи дому, викликати лікаря та ін.

Цікавим напрямком розвитку спеціального програмного забезпечення для медиків - здійснення тісної взаємодії з аптечними мережами, яке дозволять не виписувати паперовий рецепт, а безпосередньо посилати його в ту чи іншу аптеку, куди пацієнт прийде і просто викупить необхідні ліки. Такий підхід дозволяє не тільки знизити витрати медичних установ, а й знизити ймовірність придбання не того ліки внаслідок нерозбірливого почерку на рецептах, створюються умови для перевірки лікарських взаємодій і алергії. Крім того, час очікування пацієнтами в аптеках істотно знижується, так як співробітники завчасно можуть подбати про наявність необхідних препаратів.

Суттєву допомогу наявність єдиної бази даних для медиків при наданні допомоги пацієнтам в рамках подолання наслідків стихійних лих, усунення наслідків надзвичайних ситуацій та ін., Так як у медичних працівників є можливість працювати об'єктивної медичною інформацією про кожного постраждалого. Бездротовий інтернет, мікро комп'ютери допомагають вчасно і оперативно отримувати інформацію про кожного, хто потребує допомоги, вести актуальний список постраждалих і т.д.

1.1.3 Недоліки залучення сучасних IT в медицину

Незважаючи на явні переваги, надані засобами медичної інформатизації, є і питання, які викликають негативні емоції у пацієнтів. Йдеться про людей, які борються за збереження таємниці стану здоров'я кожного пацієнта. Піддаються серйозним сумнівам можливість збереження конфіденційності інформації, що стосується опису захворювань, результатів аналізів тощо. З причини крадіжки баз даних в результаті діяльності хакерів. На жаль, від шкідливого впливу хакерів не застрахована жодна організація. Однак, застосування сучасних IT в медичній сфері дозволяє отримати суттєві переваги, а при належному рівні забезпечення безпеки варіант розголошення конфіденційної інформації пацієнтів зводиться до мінімуму.

Ще одним суттєвим недоліком впровадження інформаційних технологій в медичну сферу є людський фактор, який проявляється в помилках, пов'язаних з введенням даних. Можна віднести до недоліків затребуваність в людях зі спеціальними навичками для

підтримки працездатності та ефективної роботи ІТ в медицині, що вимагає деяких фінансових витрат.

1.2 Прогнозування

Розвиток сучасної медицини неможливо без впровадження в клінічну практику процесу прогнозування. Прогнозування результатів лікування дає можливість об'єктивного вибору лікувальної тактики, оцінки ефективності та економічного обґрунтування доцільності того чи іншого методу терапії, а також підвищує надійність планування ресурсів охорони здоров'я.

Прогноз (грец. Prognosis - передбачення, прогноз) медичний - передбачення виникнення, розвитку та результату захворювання, засноване на знанні закономірностей патологічних процесів і перебігу хвороб; прогноз визначають також як діагноз майбутнього.

Прогнозування - це розробка прогнозу; у вузькому значенні - спеціальне наукове дослідження конкретних перспектив розвитку будь-якого процесу. В цілому прогнозування може бути представлено як деякої системи підходів і методів, використовуваних для досягнення найбільш точного прогнозу.

Будь який прогноз заснований на вивченні деякого минулого безлічі спостережень. Цей проміжок часу, на підставі якого будується прогноз, отримав назву періоду заснування прогнозу.

Будь який прогноз має властивими йому характеристиками. Такими характеристиками є:

– точність прогнозу - оцінка довірчого інтервалу прогнозу для певної довірчої ймовірності його здійснення (в тому випадку, коли прогноз має імовірнісний характер);

– достовірність прогнозу - оцінка ймовірності здійснення прогнозу для заданого довірчого інтервалу (в тому випадку, коли прогноз має імовірнісний характер);

– помилка прогнозу - фактична величина відхилення прогнозу від дійсного стану об'єкта прогнозування.

У тому випадку, коли імовірнісні оцінки прогнозу не можуть бути дані, точність прогнозу і його достовірність визначаються якісними, а не кількісними характеристиками або задаються кордонами без вказівки ймовірності попадання прогнозованої величини в ці межі.

Необхідність прогнозу обумовлена бажанням знати події майбутнього, що неможливо на 100% в принципі, виходячи зі статистичних, імовірнісних, емпіричних, філософських принципів. Точність будь-якого прогнозу обумовлена:

– об'ємом істинних (верифікованих) вихідних даних і періодом їх збору;

– об'ємом неверифікованих вихідних даних, періодом їх збору;

– властивостями системи, об'єкта, що піддаються прогнозуванню;

– методиками та підходами прогнозування.

При зростанні сукупності факторів, що впливають на точність прогнозу він практично заміщається рутинним розрахунком з деякою сталою похибкою.

Прогнози діляться (умовно):

– за термінами: короткострокові, середньострокові, довгострокові, далекостроковій;

– за масштабом: приватні, місцеві, регіональні, галузеві, країнові, світові (глобальні).

– по відповідальності (авторству): особисті, на рівні підприємства (організації), на рівні державних органів.

До основних методів прогнозування відносять:

– статистичні методи;

– експертні оцінки (наприклад, метод Дельфі);

– методи моделювання;

– інтуїтивні.

1.2.1 Класифікація методів і моделей прогнозування

Метод прогнозування являє собою послідовність дій, які потрібно зробити для отримання моделі прогнозування. За аналогією з кулінарією метод є послідовність дій, згідно з якою готується страва - тобто зробиться прогноз.

Модель прогнозування є функціональне уявлення, адекватно описує досліджуваний процес і є основою для отримання його майбутніх значень. У тій же кулінарної аналогією модель є список інгредієнтів і їх співвідношення, необхідний для нашого блюда - прогнозу.

Поняття «метод прогнозування» набагато ширше поняття «модель прогнозування». Один з найбільш важливих ознак методів прогнозування - ступінь формалізації, яка досить повно охоплює прогностичні методи.

Особливості об'єктів прогнозування такі, що точність прогнозу змінюється не тільки залежно від того, який природи об'єкт прогнозується, але й залежно від того, на якій період попередження виконується прогноз. Дійсно, якщо необхідний прогноз на саму найближчу перспективу, коли об'єкт, в силу притаманної йому інерційності, не встигає змінити свої характеристики, то прогнозується не так стан об'єкта, скільки відхилення від цього стану. А от якщо необхідно виконати прогноз на далеку перспективу, то тут виникає задача розрахувати довготривалу тенденцію руху самого об'єкта, а відхилення від нього оцінюються як деякий прогнозний фон.

Вся сукупність методів прогнозування групується за такими ознаками:

– за способом отримання та обробки інформації: статистичні методи, методи аналогій, що випереджають методи;

– за ступенем формалізації: формалізовані та інтуїтивні;

– за загальним принципом дії;

– за напрямами і призначенням прогнозування;

– по процедурі отримання параметрів прогнозної моделі та ін.

Статистичні методи – система прийомів, способів обробки інформації, спрямованих на отримання кількісних закономірностей, які в структурі, динаміці та взаємозв'язках прогнозованих масових соціально-економічних явищ.

Метод аналогій – побудований на отриманні прогнозів побудованих на логічному висновку, з якого знання про прогнозовані процесах виникає на основі відомого подібності закономірностей розвитку одних процесів з іншими. Ця властивість дозволяє після дослідження робити висновки, хоча й не остаточні, і не доказові в повному розумінні цього слова. Існують наступні три види аналогій: аналогія властивостей, аналогія відносин і ізоморфізму. Принцип ізоморфізму покладений в основу розробки економіко-математичних моделей прогнозування соціально-економічного розвитку.

Випереджаючі методи прогнозування базуються на певних принципах спеціальної обробки науково-технічної інформації, що реалізують у прогнозі її властивість відбивати нові тенденції закономірностей розвитку об'єкта прогнозування. У свою чергу їх можна розділити на методи дослідження динаміки розвитку об'єкта і методи дослідження та оцінки рівня розвитку об'єкта.

Формалізовані методи - описані в літературі методи прогнозування, в результаті яких будують моделі прогнозування, тобто визначають таку математичну залежність, яка

дозволяє обчислити майбутнє значення процесу, тобто зробити прогноз. Формалізовані методи прогнозування використовують математичний опис виявлених закономірностей у розвитку об'єкта для отримання прогнозу. Інтуїтивні методи прогнозування використовуються в тих випадках, коли неможливо врахувати вплив багатьох факторів через значну складність об'єкта прогнозування. Інтуїтивні методи прогнозування мають справу з судженнями і оцінками експертів. На сьогоднішній день вони часто застосовуються в маркетингу, економіці, політиці, так як система, поведінка якої необхідно спрогнозувати, або дуже складна й не піддається математичному опису, або дуже проста і в такому описі не потребує. Інтуїтивні і формалізовані методи схожі за своїм складом з експертними та фактографічними методами. Фактографічні методи засновані на фактично наявної інформації про об'єкт прогнозування і його минулий розвиток, експертні базуються на інформації, отриманої за оцінками фахівців-експертів.

Моделі прогнозування бувають двох типів: статистичні та структурні.

У статистичних моделях прогнозування функціональна залежність між майбутніми та фактичними значеннями часового ряду, а також зовнішніми факторами, якщо такі враховуються, задана аналітично, тобто формулою. До статистичних моделям прогнозування відносяться наступні групи:

- регресивні моделі;
- авторегресивні моделі;
- моделі експоненціального згладжування.

У структурних моделях прогнозування функціональна залежність між майбутніми та фактичними значеннями часового ряду, а також зовнішніми чинниками задана структурно, наприклад, у вигляді графа. До структурних моделям прогнозування відносяться наступні групи:

- нейромережеві моделі;
- моделі на базі ланцюгів Маркова;
- моделі на базі класифікаційно-регресивних дерев.

Існують цілий клас непопулярних моделей прогнозування на основі, наприклад, методу опорних векторів, генетичного алгоритму, нечіткої логіки і багатьох інших, однак застосування таких моделей скупо описано в літературі. Більшість такого роду моделей було створено для вирішення інших завдань, проте в наслідок знайшло вузьке застосування в прогнозуванні часових рядів. Так наприклад, генетичний алгоритм є алгоритм для вирішення завдань оптимізації, тобто знаходження екстремуму, і лише деякі дослідники зуміли застосувати його для прогнозування часових рядів.

1.2.2 Прогнозна аналітика

Прогнозна аналітика являє собою процес аналізу статистичних та фактичних даних, а також створення статистичної моделі, що дозволяє прогнозувати можливі наслідки різних дій. В основі концепції лежить поняття «прогнозних факторів», які з найбільшою ймовірністю вплинуть на подальші події і які можуть бути оцінені або враховані при прогнозуванні результатів.

У рішеннях прогнозної аналітики використовуються різні методики, що дозволяють які виявляти закономірності в накопичених даних - штучні нейронні мережі, деревовидні схеми рішень, а також безліч інших статистичних підходів. Отримана інформація може потім використовуватися для визначення або прогнозування тенденцій в нових даних.

Перекладені в цифрову форму дані легко аналізувати. Засоби інтелектуального аналізу і прогнозної аналітики дозволяють виявляти в накопичених даних закономірності, які використовуються для прогнозування тенденцій. Традиційно прогнозна аналітика разом з експертними знаннями застосовуються в діагностиці та лікуванні багатьох хвороб. У числі перших прикладів таких систем були Next і HYCONES. У регіонах, де кваліфікованих медичних фахівців недостатньо або немає зовсім, прогнозні рішення можуть принести величезну користь. Широке поширення он-лайн систем даних і прогнозування призводить до появи більш швидких і точних засобів, що допомагають медичним працівникам приймати рішення. Останнім часом прогнозні системи стають все досконалішою. Як я вже писав в торішній статті про прогнозної аналітики і стандартах (див. Ресурси), IBM та Інститут технології Університету Онтаріо в даний час працюють над впровадженням рішення з аналізу даних і прогнозування з метою спостереження за недоношеними дітьми. При використанні цього рішення небезпечні інфекції вдається виявляти на основі біомедичних показників на 24 години раніше, ніж традиційними способами.

Обробка інформації про низьку або високу небезпеку виникнення певного захворювання або стану в групі пацієнтів за допомогою засобів інтелектуального аналізу або прогнозної аналітики також допомагає лікарям створювати протоколи лікування з урахуванням особливостей різних груп. Наприклад, у разі серцево-судинних захворювань пацієнтам, віднесених за допомогою прогнозного рішення в групу високого ризику, можна запропонувати прості профілактичні заходи, які можуть істотно знизити ризик серцевого нападу - наприклад, скорочення трансжирів в раціоні харчування, дотримання

дієти і відмова від куріння. Таким чином, лікарі можуть створювати різні стратегії для контролю стану пацієнтів з групи невисокого ризику і скорочення ризику для життя пацієнтів з групи високого ризику.

Відповідно до прийнятого в США федеральним законом лікарні, де рівень повторної госпіталізації перевищує очікуваний, тепер будуть отримувати менше бюджетних коштів за програмою Medicare. За оцінками Консультативного комітету з виплат Medicare в 2005 році, витрати на повторну госпіталізацію, оплачені з фондів Medicare, склали 15 млрд. доларів, з яких 12 млрд. можна було б уникнути. Враховуючи, що в більшості випадків повторної госпіталізації можна уникнути, вже зараз лікарні використовують засоби прогнозної аналітики для зниження числа повторних госпіталізацій. Хоча для запобігання повторної госпіталізації часто досить повторного огляду у лікаря, прогнозна аналітика дозволяє точно визначити пацієнтів, які потребують контрольного спостереження. З її допомогою лікарям простіше визначати пацієнтів, яким можуть знадобитися додаткові лікувальні рекомендації, аж до роз'яснення правил проходження дієті.

Прогнозні системи вже багато років використовуються фінансовими установами для виявлення шахрайства. Сьогодні для оцінки ризику шахрайства по більшості транзакцій з кредитними картами використовується прогнозне рішення, що працює в режимі реального часу. Якщо рішення вважатиме певну транзакцію ризикованою, можуть бути вжиті заходи аж до її відхилення для запобігання шахрайства. Оскільки витрати, пов'язані з шахрайством у системі Medicare, набагато перевищують витрати, пов'язані з повторною госпіталізацією, запобігання шахрайства стане основним завданням прогнозних рішень. Прогнозні технології, які з успіхом застосовуються для виявлення шахрайства у фінансовій сфері, наприклад, нейронні мережі, можуть і повинні використовуватися для виявлення шахрайства та зловживань в охороні здоров'я.

Велику користь для прогнозної аналітики в охороні здоров'я принесе об'єднання різних сховищ даних. Доступність більшого обсягу інформації про конкретний пацієнті або групі пацієнтів дозволить отримати більш детальну картину і, тим самим, підвищити якість прогнозів. Чим більше точок даних включається в модель, тим ефективніше її можна адаптувати під конкретного пацієнта або групи пацієнтів. Це дозволить лікарям приймати більш точні та ефективні заходи боротьби з хворобою, що, з одного боку, підвищує загальну ефективність системи охорони здоров'я, а з іншого - скорочує витрати.

Враховуючи, що дослідження в області прогнозної аналітики та охорони здоров'я ведуться вже багато років, дивно, чому досягнення в цій галузі починають застосовуватися в повсякденному житті тільки зараз. Втім, пояснення досить просте -

система охорони здоров'я з великим небажанням впроваджує цифрові технології. Навіть сьогодні багато лікарів в США досі заповнюють медичні карти від руки, роздруковують рентгенівські знімки і вставляють їх у карти пацієнтів. Таким чином, навіть сьогодні доступ до цих даних для глибокого аналізу та прогнозної аналітики залишається серйозною проблемою. Але також відомо, що все більше інформації про пацієнтів і лікувальних установах зберігається в цифровій формі. У США на передньому краї переходу на електронну медичну документацію знаходяться такі великі організації системи охорони здоров'я. У країнах з перехідною економікою та країнах, що розвиваються впровадження цих технологій є ще більш актуальним завданням.

1.3 Постановка завдання дослідження

Метою даної атестаційної роботи є дослідження методів аналізу медичних даних для подальшої побудови прогнозної моделі. Для реалізації поставленої задачі треба дослідити технологію інтелектуального аналізу даних(Data Mining). Для цього необхідно виконати наступні завдання:

- дослідити етапи та задачі DM;
- обрати необхідну стратегію аналізу;
- обрати алгоритми для аналізу;
- безпосередньо реалізувати досліджувані алгоритми чи застосувати готові програмні рішення;
- провести порівняльний аналіз досліджуваних алгоритмів.

У дослідженні основним джерелом даних є форми CMS-1500, в яких прибрані всі персональні дані про пацієнтів, крім дати народження. Таким чином, ми маємо можливість працювати з реальними даними, але при цьому зберігається повна конфіденційність. Попередній перегляд даних, які є в формах CMS-1500 показав, що найчастіше дані не є повними. Таким чином всі дані форм CMS-1500 вимагають попередньої обробки. Приклад форми CMS-1500 наведено у додатку А.

2 АНАЛІЗ ДАНИХ ТА ПОБУДОВА ПРОГНОЗНОЇ МОДЕЛІ

2.1 Інтелектуальний аналіз даних

Інтелектуальний аналіз даних можна описати як процес визначення нових, коректних і потенційно корисних знань на основі великих масивів даних. В англомовній літературі замість терміна «інтелектуальний аналіз даних» зазвичай використовується термін Data Mining (дослівний переклад - «видобуток даних»), а також близький термін Knowledge Discovery in Databases (KDD) - «Виявлення знань у великих базах даних».

Технологію Data Mining досить точно визначає Григорій Піатецкій-Шапіро (Gregory Piatetsky-Shapiro) - один із засновників цього напрямку: Data Mining - це процес виявлення в сирих даних раніше невідомих, нетривіальних, практично корисних і доступних інтерпретації знань, необхідних для прийняття рішень у різних сферах людської діяльності.

Суть і мета технології Data Mining можна охарактеризувати так: це технологія, яка призначена для пошуку у великих обсягах даних неочевидних, об'єктивних і корисних на практиці закономірностей. Неочевидних - це означає, що знайдені закономірності не виявляються стандартними методами обробки інформації або експертним шляхом. Об'єктивних - це означає, що виявлені закономірності будуть повністю відповідати дійсності, на відміну від експертної думки, яке завжди є суб'єктивним. Практично корисних - це означає, що висновки мають конкретне значення, якому можна знайти практичне застосування.

Сучасні технології Data Mining переробляють інформацію з метою автоматичного пошуку шаблонів (патернів), характерних для будь-яких фрагментів неоднорідних багатомірних даних. На відміну від оперативної аналітичної обробки даних (OLAP) в Data Mining тягар формулювання гіпотез і виявлення незвичайних (unexpected) шаблонів перекладено з людини на комп'ютер. Data Mining - це не один, а сукупність великої кількості різних методів виявлення знань. Вибір методу часто залежить від типу наявних даних і від того, яку інформацію ви намагаєтеся отримати.

Основна особливість Data Mining - це поєднання широкого математичного інструментарію (від класичного статистичного аналізу до нових кібернетичних методів) і останніх досягнень у сфері інформаційних технологій. У технології Data Mining гармонійно поєдналися строго формалізовані методи і методи неформального аналізу, тобто кількісний і якісний аналіз даних.

2.1.1 Стадії аналізу даних

Процес Data Mining складається із трьох стадій, але третя несе суто коригуючий характер:

– вільний пошук (Discovery);

– прогностичне моделювання (Predictive Modeling);

– аналіз винятків (Forensic Analysis).

На стадії вільного пошуку здійснюється дослідження набору даних з метою пошуку прихованих закономірностей. Попередні гіпотези щодо виду закономірностей тут не визначаються.

Закономірність (law) - істотна і постійно повторювана взаємозв'язок, визначальна етапи і форми процесу становлення, розвитку різних явищ або процесів. Система Data Mining на цій стадії визначає шаблони, для отримання яких в системах OLAP, наприклад, аналітику необхідно обдумувати і створювати безліч запитів. Тут же аналітик звільняється від такої роботи - шаблони шукає за нього система. Особливо корисно застосування даного підходу в надвеликих базах даних, де вловити закономірність шляхом створення запитів досить складно, для цього потрібно перепробувати безліч різноманітних варіантів. Вільний пошук представлений такими діями:

– виявлення закономірностей умовної логіки (conditional logic);

– виявлення закономірностей асоціативної логіки (associations and affinities);

– виявлення трендів і коливань (trends and variations).

Описані дії, в рамках стадії вільного пошуку, виконуються за допомогою:

– індукції правил умовної логіки (задачі класифікації і кластеризації, опис в компактній формі близьких або схожих груп об'єктів);

– індукції правил асоціативної логіки (задачі асоціації та послідовності і витягувана за їх допомогою інформація);

– визначення трендів і коливань (вихідний етап задачі прогнозування).

На стадії вільного пошуку також повинна здійснюватися валідація закономірностей, тобто перевірка їх достовірності на частини даних, які не брали участь у формуванні закономірностей. Такий прийом розділення даних на навчальне та перевірочне безліч часто використовується в методах нейронних мереж і дерев рішень.

Друга стадія Data Mining - прогностичне моделювання - використовує результати роботи першої стадії. Тут виявлені закономірності використовуються безпосередньо для прогнозування. Прогностичне моделювання включає такі дії:

– пророкування невідомих значень (outcome prediction);

– прогнозування розвитку процесів (forecasting).

У процесі прогностичного моделювання вирішуються завдання класифікації і прогнозування.

При вирішенні задачі класифікації результати роботи першої стадії (індукції правил) використовуються для віднесення нового об'єкта, з певною впевненістю, до одного з відомих, визначених класів на підставі відомих значень. При вирішенні задачі прогнозування результати першої стадії (визначення тренда або коливань) використовуються для передбачення невідомих (пропущених або ж майбутніх) значень цільової змінної (змінних).

Вільний пошук розкриває загальні закономірності. Він за своєю природою індуктований. Закономірності, отримані на цій стадії, формуються від часткового до загального. В результаті ми одержуємо деяке загальне знання про деякий клас об'єктів на підставі дослідження окремих представників цього класу. Прогностичне моделювання, навпаки, дедуктивно. Закономірності, отримані на цій стадії, формуються від загального до приватного і одиничного. Тут ми отримуємо нове знання про деякий об'єкт або ж групи об'єктів на підставі:

– знання класу, до якого належать досліджувані об'єкти;

– знання загального правила, діючого в межах даного класу об'єктів.

Слід зазначити, що отримані закономірності, а точніше, їх конструкції, можуть бути прозорими, тобто допускають тлумачення аналітика (розглянуті вище правила), і непрозорими, так званими "чорними ящиками". Типовий приклад останньої конструкції - нейронна мережа.

На третій стадії Data Mining аналізуються виключення або аномалії, виявлені в знайдених закономірностях. Дія, що виконується на цій стадії, - виявлення відхилень (deviation detection). для виявлення відхилень необхідно визначити норму, яка розраховується на стадії вільного пошуку.

2.1.2 Класифікація задач аналізу даних

Різні методи Data Mining характеризуються певними властивостями, які можуть бути визначальними при виборі методу аналізу даних. методи можна порівнювати між собою, оцінюючи характеристики їх властивостей.

Серед основних властивостей і характеристик методів Data Mining розглянемо наступні: точність, масштабованість, інтерпретованість, перевіряємість, трудомісткість, гнучкість, швидкість і популярність.

Більшість аналітичних методів, що використовуються в технології Data Mining – це відомі математичні алгоритми і методи. Новим в їх застосуванні є можливість їх використання при вирішенні тих чи інших конкретних проблем, обумовлена можливостями, що з'явилися технічних і програмних засобів. Слід зазначити, що більшість методів Data Mining були розроблені в рамках теорії штучного інтелекту. З них основними є: класифікація, регресія, пошук асоціативних правил і кластеризація. Нижче наведено короткий опис основних завдань аналізу даних:

– класифікація (Classification) - найбільш проста і поширена задача Data Mining. В результаті рішення задачі класифікації виявляються ознаки, які характеризують групи об'єктів досліджуваного набору даних - класи; за цими ознаками новий об'єкт можна віднести до того чи іншого класу. Методи рішення. Для вирішення завдання класифікації можуть використовуватися методи: найближчого сусіда (Nearest Neighbor); k-найближчого сусіда (k-Nearest Neighbor); байєсовські мережі (Bayesian Networks); індукція дерев рішень; нейронні мережі (neural networks);

– кластеризація (Clustering) - кластеризація є логічним продовженням ідеї класифікації. Це завдання більш складне, особливість кластеризації полягає в тому, що класи об'єктів спочатку не визначені. Результатом кластеризації є розбиття об'єктів на групи. Приклад методу розв'язання задачі кластеризації: навчання "без вчителя" особливого виду нейронних мереж - самоорганізованих карт Кохонена;

– асоціація (Associations) - у результаті виконання завдання пошуку асоціативних правил відшукуються закономірності між пов'язаними подіями в наборі даних. Відмінність асоціації від двох попередніх завдань Data Mining: пошук закономірностей здійснюється не на основі властивостей аналізованого об'єкта, а між кількома подіями, які відбуваються одночасно. Найбільш відомий алгоритм вирішення задачі пошуку асоціативних правил - алгоритм Apriori;

– послідовність (Sequence), або послідовна асоціація (sequential association) - послідовність дозволяє знайти тимчасові закономірності між транзакціями. Завдання послідовності подібна асоціації, але її метою є встановлення закономірностей не між одночасно наступаючими подіями, а між подіями, пов'язаними в часі (тобто відбуваються з деяким певним інтервалом у часі). Іншими словами, послідовність визначається високою ймовірністю ланцюжка пов'язаних у часі подій. Фактично, асоціація є окремим випадком послідовності з тимчасовим лагом, рівним нулю. Це завдання Data Mining також

називають завданням знаходження послідовних шаблонів (sequential pattern). Правило послідовності: після події X через певний час відбудеться подія Y. Приклад. Після покупки квартири мешканці в 60% випадків протягом двох тижнів набувають холодильник, а протягом двох місяців в 50% випадків купується телевізор. Рішення даного завдання широко застосовується в маркетингу та менеджменті, наприклад, при управлінні циклом роботи з клієнтом (Customer Lifecycle Management) ;

– прогнозування (Forecasting) - в результаті рішення задачі прогнозування на основі особливостей історичних даних оцінюються пропущені або ж майбутні значення цільових чисельних показників. Для вирішення таких завдань широко застосовуються методи математичної статистики, нейронні мережі та ін.;

– визначення відхилень або викидів (Deviation Detection), аналіз відхилень або викидів Мета рішення даної задачі - виявлення та аналіз даних, найбільш відрізняються від загальної множини даних, виявлення так званих нехарактерних шаблонів;

– оцінювання (Estimation) - завдання оцінювання зводиться до передбачення безперервних значень ознаки;

– аналіз зв'язків (Link Analysis) - задача знаходження залежностей в наборі даних;

– візуалізація (Visualization, Graph Mining) - в результаті візуалізації створюється графічний образ аналізованих даних. Для вирішення завдання візуалізації використовуються графічні методи, що показують наявність закономірностей в даних. Приклад методів візуалізації - подання даних в 2-D і 3-D вимірах;

– підведення підсумків (Summarization) - завдання, мета якої - опис конкретних груп об'єктів з аналізованого набору даних.

Перераховані завдання за призначенням поділяються на описові та передбачувльні.

Описові (descriptive) завдання приділяють увагу поліпшенню розуміння аналізованих даних. Ключовий момент у таких моделях - легкість і прозорість результатів для сприйняття людиною. Можливо, виявлені закономірності будуть специфічною рисою саме конкретних досліджуваних даних і більше ніде не зустрінуться, але це все одно може бути корисно і тому має бути відомо. До такого виду завдань відносяться кластеризація і пошук асоціативних правил.

Рішення передбачувльних (predictive) завдань розбивається на два етапи. На першому етапі на підставі набору даних з відомими результатами будується модель. На другому етапі вона використовується для передбачення результатів на підставі нових наборів даних. При цьому, природно, потрібно, щоб побудовані моделі працювали максимально точно. До даного виду завдань відносять задачі класифікації і регресії. Сюди

можна віднести і завдання пошуку асоціативних правил, якщо результати її рішення можуть бути використані для передбачення появи деяких подій.

За способами вирішення завдання поділяють на supervised learning (навчання з учителем) і unsupervised learning (навчання без учителя). Така назва походить від терміна Machine Learning (машинне навчання), часто використовуваного в англомовній літературі і позначає всі технології Data Mining.

Unsupervised learning об'єднує завдання, що виявляють описові моделі, наприклад закономірності в покупках, чинених клієнтами великого магазину. Очевидно, що якщо ці закономірності є, то модель повинна їх представити і недоречно говорити про її навчанні. Звідси й назва - unsupervised learning. Перевагою таких завдань є можливість їх вирішення без будь-яких попередніх знань про аналізованих даних. До них відносяться кластеризація і пошук асоціативних правил.

У разі supervised learning завдання аналізу даних вирішується в кілька етапів. Спочатку за допомогою якого-небудь алгоритму Data Mining будується модель аналізованих даних - класифікатор. Потім класифікатор піддається навчанню. Іншими словами, перевіряється якість його роботи і, якщо воно незадовільно, відбувається додаткове навчання класифікатора. Так продовжується до тих пір, поки не буде досягнутий необхідний рівень якості або не стане ясно, що обраний алгоритм не працює коректно з даними, або ж самі дані не мають структури, яку можна виявити. До цього типу завдань відносять задачі класифікації і регресії.

Таким чином задача поставленна у данній атестаційній роботі відноситься до завдань supervised learning та для побудови моделі аналізованих даних використовуються алгоритми автоматичної класифікації, тобто алгоритми кластерізації.

2.2 Завдання кластеризації та кластерний аналіз

Одним з найважливіших завдань Data Mining є кластеризація - об'єднання об'єктів в групи на основі подібності їх ознак. Такі групи називаються кластерами. Попадання двох об'єктів в один кластер дозволяє припустити високу ступінь схожості їх властивостей, і навпаки, якщо об'єкти в результати кластеризації потрапили в різні кластери, то вони істотно відрізняються один від одного за своїми ознаками.

В результати кластеризації деякого безлічі даних формується певна кількість кластерів, що виражається в підсумковій моделі даних, яка є рішенням задачі кластеризації.

Завдання кластеризації полягає в пошуку незалежних груп (кластерів) та їх характеристик у всьому безлічі аналізованих даних. Вирішення цього завдання допомагає нам краще зрозуміти дані. Крім того, угрупування однорідних об'єктів дозволяє скоротити їх число, а, отже, і полегшити аналіз.

Завдання кластеризації схожа з завданням класифікації, є її логічним продовженням, але її відмінність в тому, що класи досліджуваного набору даних заздалегідь не визначені. Синонімами терміну "кластеризація" є "автоматична класифікація", "навчання без вчителя" і "таксономія".

Кластеризація призначена для розбиття сукупності об'єктів на однорідні групи (кластери або класи). Якщо дані вибірки представити як точки в просторі ознак, то завдання кластеризації зводиться до визначення "згущення точок". Мета кластеризації - пошук існуючих структур.

Кластеризація є описовою процедурою, вона не робить ніяких статистичних висновків, але дає можливість провести розвідувальний аналіз і вивчити "структуру даних".

Термін кластерний аналіз (вперше ввів Tryon, 1939) насправді включає в себе набір різних алгоритмів класифікації. Загальне питання, що задається дослідниками в багатьох областях, полягає в тому, як організувати спостережувані дані в наочні структури.

Потреба в кластерному аналізі виникає в тих областях чи етапах діяльності, де є необхідність в розбитті об'єктів (ситуацій) на непересічні підмножини, званими кластерами, так, щоб кожен кластер складався з схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися. Чіткий поділ на кластери можливо тільки в ідеальних умовах і при сильно розрізняються параметрах об'єктів кластеризації. Тому для вирішення реальних завдань все частіше застосовуються нечіткі методи, в яких розбиття об'єктів (ситуацій) виконується на частково пересічні підмножини.

Важливою передумовою застосування нечітких методик кластеризації в реальних умовах є те, що характеристики об'єктів не завжди є вимірними і тому в ряді випадків присутні експертні оцінки характеристик об'єктів, які є суб'єктивними і можуть бути суперечливими. Для завдання кластеризації характерна відсутність будь-яких відмінностей, як між змінними, так і між об'єктами. Навпаки, шукаються групи найбільш близьких, схожих об'єктів. Методи автоматичного розбиття на кластери рідко використовуються самі по собі, а тільки для одержання груп схожих об'єктів. Після

визначення кластерів використовуються інші методи Data Mining, щоб спробувати встановити, що означає таке розбиття і чим воно викликане.

Відзначимо ряд особливостей, властивих завданню кластеризації.

По-перше, рішення сильно залежить від природи об'єктів даних (і їх атрибутів). Так, з одного боку, це можуть бути однозначно визначені, кількісно окреслені об'єкти, а з іншого - об'єкти, що мають імовірнісний або нечіткий опис.

По-друге, рішення значною мірою залежить і від подання кластерів і передбачуваних відносин об'єктів даних і кластерів. Так, необхідно враховувати такі властивості, як можливість чи неможливість приналежності об'єктів до кільком кластерам. Необхідно визначення самого поняття приналежності кластеру: однозначна (належить, не належить), імовірнісна (ймовірність приналежності), нечітка (ступінь приналежності).

Цілі кластеризації в Data Mining можуть бути різними і залежать від конкретної розв'язуваної задачі. Розглянемо ці завдання:

– вивчення даних. Розбиття множини об'єктів на групи допомагає виявити внутрішні закономірності, збільшити наочність представлення даних, висунути нові гіпотези, зрозуміти, наскільки інформативні властивості об'єктів;

– полегшення аналізу. За допомогою кластеризації можна спростити подальшу обробку даних та побудова моделей: кожен кластер обробляється індивідуально, і модель створюється для кожного кластера окремо. У цьому сенсі кластеризація може розглядатися як підготовчий етап перед вирішенням інших завдань Data Mining: класифікації, регресії, асоціації, послідовних шаблонів;

– стиснення даних. У випадку, коли дані мають великий обсяг, кластеризація дозволяє скоротити обсяг збережених даних, залишивши по одному найбільш типовому представнику від кожного кластера;

– прогнозування. Кластери використовуються не тільки для компактного представлення об'єктів, а й для розпізнавання нових. Кожен новий об'єкт відноситься до того кластеру, приєднання до якого найкращим чином відповідає критерію якості кластеризації. Значить, можна прогнозувати поведінку об'єкта, припустивши, що воно буде схожим з поведінкою інших об'єктів кластера;

– виявлення аномалій. Кластеризація застосовується для виділення нетипових об'єктів. Це завдання також називають виявленням аномалій (outlier detection). Інтерес тут представляють кластери (групи), в які потрапляє вкрай мало, скажімо один-три, об'єктів.

Саме ціль прогнозування є метою даної атестаційної роботи, а оскільки вхідними даними є об’ємний масив структурованих медичних даних, то вирішення задачі кластеризації інтелектуального аналізу є найбільш простим та ефективним інструментом.

Методи кластеризації можна класифікувати на наступними критеріями:

- за способом обробки даних:
    1) ієрархічні методи;
    2) неієрархічні методи або плоскі.
- за способом аналізу даних:
    1) чіткі;
    2) нечіткі.
- за кількістю застосувань алгоритмів кластеризації;
    1) з одноетапною кластеризацією;
    2) з багатоетапною кластеризацією.
- по можливості розширення обсягу оброблюваних даних;
    1) здатні до масштабування;
    2) не здатні до масштабування.
- за часом виконання кластеризації.
    1) потокові (on-line);
    2) не потокові (off-line).

Але саме класифікація за способом обробки даних є найбільш розповсюдженою. Ієрархічні методи відповідно до класифікації діляться на агломеративні і дивізимні. Агломеративна група методів характеризується послідовним об'єднанням вихідних елементів і відповідним зменшенням кількості кластерів. Дивізимна група методів характеризується послідовним поділом вихідних елементів і відповідним збільшенням кількості кластерів. Найбільш значущу частина неієрархічних методів представляють ітеративні методи, засновані на поділі набору даних на деяку кількість окремих кластерів. Існують два підходи для розділення даних. Перший полягає у визначенні меж кластерів як найбільш щільних ділянок в багатовимірному просторі характеристик об'єктів, тобто визначення кластера там, де є велика «згущення» об'єктів. Другий підхід полягає в мінімізації заходи відмінності об'єктів.

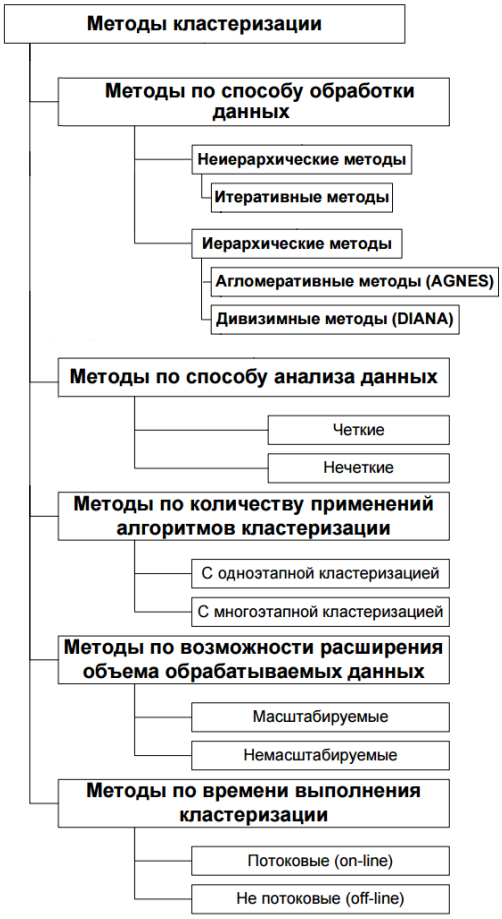


Рисунок 2.1 – Класифікація методів кластеризації

Класичні ієрархічні алгоритми працюють тільки з категорійними атрибутами, коли будується повне дерево вкладених кластерів. Тут поширені агломеративні методи побудови ієрархій кластерів - в них проводиться послідовне об'єднання вихідних об'єктів і відповідне зменшення числа кластерів. Ієрархічні алгоритми забезпечують порівняно високу якість кластеризації і не вимагають попереднього завдання кількості кластерів. Більшість з них мають складність.

Неієрархічні алгоритми засновані на оптимізації деякої цільової функції, що визначає оптимальне в певному сенсі розбиття множини об'єктів на кластери. У цій групі популярні алгоритми сімейства K-середніх (K-засоби, нечіткі засоби, Густафсон-Кесселя), які в якості цільової функції використовують суму квадратів зважених відхилень координат об'єктів від центрів шуканих кластерів. Кластери шукаються сферичної або еліпсоїдної форми. У канонічній реалізації мінімізація функції проводиться на основі методу множників Лагранжа і дозволяє знайти тільки найближчий локальний мінімум. Використання методів глобального пошуку (генетичні алгоритми) значно збільшить обчислювальну складність алгоритму.

Серед неієрархічних алгоритмів, які не ґрунтуються на відстані, слід виділити EM-алгоритм (Очікування-Максимізація). У ньому замість центрів кластерів передбачається наявність функції щільності ймовірності для кожного кластеру з відповідним значенням математичного очікування і дисперсією.

До недоліку ієрархічних алгоритмів можна віднести систему повних розбитть, яка може бути зайвої в контексті розв'язуваної задачі. Тому атестаційна робота здебільшого спрямована на аналіз неієрархічних методів кластерного аналізу.

2.3 Метрики та досліджувані алгоритми кластерного аналізу

Для дослідження алгоритмів необхідно визначити критерії схожості об’єктів. Під час кластерного аналізу для визначення схожості об’єктів вводиться поняття метрики. Подібність чи відмінність між об'єктами, що класифікуються, встановлюється залежно від метричної відстані між ними. Якщо кожен об'єкт має k різних ознак, то він може бути представлений у вигляді точки у k-вимірному просторі. Схожість з іншими об'єктами буде визначатися як відповідну відстань між такими точками. Алгоритми кластерного аналізу можуть використовувати різні метрики. Основні з них:

1) Евклідова відстань - найбільш поширена функція відстані. Являє собою геометричним відстанню в багатовимірному просторі:

$$p(x, x') = \sqrt{\sum_{i}^{n} (x_i - x_i')^2} \quad ,$$

де x, x’ – координати об’єктів у n-мірному просторі,

x_i, x_i’ – значення і-го параметру відповідних об’єктів.

2) Квадрат евклідової відстані - використовується, щоб надати великі ваги більш віддаленим один від одного об'єктам:

$$p(x, x') = \sum_{i}^{n} (x_i - x_i')^2 \quad ,$$

де x, x’ – координати об’єктів у n-мірному просторі,

x_i, x_i’ – значення і-го параметру відповідних об’єктів.

3) Виважена евклідова відстань - використовується при завданні довільних ваг для тих чи інших ознак:

$$p(x, x') = \sum_{i}^{n} w_i \times (x_i - x_i')^2 ,$$

де x, x’ – координати об’єктів у n-мірному просторі,

x_i, x_i’ – значення і-го параметру відповідних об’єктів,

w_i – ваговий коефіцієнт і-го параметру.

4) Відстань міських кварталів. Ця відстань є середнім різниць по координатах. У більшості випадків ця міра відстані призводить до таких же результатів, як і для звичайного відстані Евкліда. Однак для цього заходу вплив окремих великих різниць (викидів) зменшується. Формула для розрахунку манхеттенської відстані:

$$p(x, x') = \sum_{i}^{n} |x_i - x_i'| ,$$

де x, x’ – координати об’єктів у n-мірному просторі,

x_i, x_i’ – значення і-го параметру відповідних об’єктів.

5) Відстань Махаланобіса - застосовують у разі залежних компонент x_1, x_2, ..., x_n вектора спостережень і їх різної значущості у вирішенні питання класифікації:

$$p(x, x') = (x - x')^T \times C^{-1} \times (x - x') ,$$

де x, x’ – координати об’єктів у n-мірному просторі,

C^{-1} – коваріаційна матриця генеральної сукупності.

6) Відстань Чебишева. Ця відстань може виявитися корисним, коли потрібно визначити два об'єкти як «різні», якщо вони розрізняються за якоюсь однією координаті. Відстань Чебишева обчислюється за формулою:

$$p(x, x') = \max(|x_i - x_i'|) \ ,$$

де x, x’ – координати об’єктів у n-мірному просторі.

7) Статечна відстань. Застосовується у випадку, коли необхідно збільшити або зменшити вагу, що відноситься до розмірності, для якої відповідні об'єкти сильно відрізняються. Статечне відстань обчислюється за наступною формулою:

$$p(x, x') = \sqrt[r]{\sum_{i}^{n} (x_i - x_i')^p} \ ,$$

де x, x’ – координати об’єктів у n-мірному просторі,

r, p – параметри, які визначаються користувачем. Параметр p відповідальний за поступове зважування різниць по окремих координатах, параметр r відповідальний за прогресивне зважування великих відстаней між об'єктами. Якщо обидва параметри – r і p – рівні двом, то ця відстань збігається з відстанню Евкліда.

2.3.1 Алгоритми K-Means та PAM

Перші застосування алгоритму k-середніх були описані в роботі Джеймса Маккуїна в 1967 році. При заздалегідь відомому числі кластерів k алгоритм k-середніх починає з деякого початкового розбиття об’єктів і уточнює його, оптимізуючи цільову функцію – середньоквадратичне помилку кластеризації як середньоквадратичне відстань між об’єктами і центрами їх кластерів:

$$e(D, C) = \sum_{j=1}^{|D|} \sum_{i:d_i \in C_j}^{|C|} \left\| \overrightarrow{d_i} - \overrightarrow{\mu_j} \right\|^2 \ ,$$

де μ_j - центр, або центроїд, кластера C_j, котрий підраховується за наступною формулою:

$$\overrightarrow{\mu_j} = \frac{1}{|C_j|} \sum_{i:d_i \in C_j}^{|C|} \overrightarrow{d_i} \quad ,$$

де $|C_j|$ - кількість об’єктів в C_j.

Ідеальним кластером алгоритм k-середніх вважає сферу з центроїдом в центрі сфери.

Дія алгоритму починається з вибору k початкових центрів кластерів. Зазвичай вихідні центри кластерів вибираються випадковим чином. Потім кожен об’єкт присвоюється того кластеру, чий центр є найбільш близьким об’єктом, і виконується повторне обчислення центру кожного кластера як центроїда, або середнього своїх членів. Таке переміщення об’єктів і повторне обчислення Центроїд кластерів продовжується до тих пір, поки не буде досягнуто умова зупинки. Умовою зупинки може служити наступне: досягнуто порогове число ітерацій, центроїди кластерів більше не змінюються та досягнуто порогове значення помилки кластеризації. На практиці використовують комбінацію критеріїв зупинки, щоб одночасно обмежити час роботи алгоритму і отримати прийнятну якість. У загальному випадку алгоритм k-середніх досягає локального мінімуму цільової функції, що призводить до субоптимального розбиттю об’єктів. Тому важливий спосіб вибору початкових значень Центроїд. Для цього відомі різні евристичні правила, наприклад, отримати початкові центри за допомогою іншого алгоритму - детермінованого, наприклад, ієрархічного агломеративного.

Алгоритм PAM аналогічний алгоритму K-Means, тільки при роботі алгоритму перерозподіляються об'єкти щодо медіани кластера, а не його центру

2.3.2 Алгоритм Fuzzy C-Means

Нечіткий алгоритм Fuzzy C-Means був запропонований Джоном С. Даному в 1973 році (пізніше вдосконалений Дж. Беждеком в 1981 році) як вирішення проблеми м'якої кластеризації, тобто присвоєння кожного об’єкта більш ніж одному кластеру. Як і його

чіткий варіант - алгоритм k-середніх - даний алгоритм, починаючи з деякого початкового розбиття даних, ітеративно мінімізує цільову функцію, якою є такий вираз:

$$e_m(D,C) = \sum_{j=1}^{|D|} \sum_{i:d_i \in C_j}^{|C|} u_{ij}^m \left\| \overrightarrow{d_i} - \overrightarrow{\mu_j} \right\|^2 \quad ,$$

де m – ступінь нечіткості, $1 < m < \infty$;

u_{ij} – ступінь належності i-го об'єкту до j-го кластеру,

$$u_{ij} = \frac{1}{\sum_{k=1}^{|C|} \left(\frac{\left\| \overrightarrow{d_i} - \overrightarrow{c_j} \right\|}{\left\| \overrightarrow{d_i} - \overrightarrow{c_k} \right\|} \right)^{\frac{2}{m-1}}} \quad ,$$

μ_j - центроїд кластера C_j, котрий підраховується за наступною формулою:

$$\overrightarrow{\mu_j} = \frac{\sum_{i=1}^{D} u_{ij}^m \times \overrightarrow{d_i}}{\sum_{i=1}^{D} u_{ij}^m} \quad .$$

2.3.3 Алгоритм CLOPE

Алгоритм CLOPE. Призначений для кластеризації величезних наборів категорійних даних. До достоїнств відносяться високі масштабованість і швидкість роботи і якість кластеризації, що досягається використанням глобального критерію оптимізації на основі максимізації градієнта висоти гістограми кластеру. Відрізняється простотою програмної реалізації. Під час роботи алгоритм зберігає в пам'яті невелику кількість інформації по кожному кластеру і вимагає мінімальне число сканувань набору даних. CLOPE автоматично підбирає кількість кластерів, причому це регулюється одним єдиним параметром коефіцієнтом відштовхування. CLOPE запропонований в 2002 році групою китайських учених. При цьому він забезпечує більш високу продуктивність і кращу якість кластеризації в порівнянні з багатьма ієрархічними алгоритмами.

Алгоритм CLOPE є алгоритмом кластеризації транзакційних даних. Транзакція в даному контексті - це довільний набір об'єктів. Тобто в конкретному випадку транзакція схожа на кортеж в якому можуть знаходити нетипізовані дані, що не мають ніякого зв'язку між собою, і, не вибудовуються в просторі. Завдання кластеризації подібних даних полягає в отриманні такого розбиття, щоб схожі транзакції перебували в одному кластері, а відмінні - в одному з інших. Для цього в алгоритмі застосовується принцип максимізації глобальної функції вартості, зближуючої транзакції в кластерах, збільшуючи параметр кластерної гістограми.

Формула глобального критерію:

$$\text{Profit}(C) = \frac{\sum_{i=1}^{k} G(C_i) \times |C_i|}{\sum_{i=1}^{k} C_i} = \frac{\sum_{i=1}^{k} \frac{S(C_i)}{W(C_i)^r} \times |C_i|}{\sum_{i=1}^{k} C_i} \quad ,$$

де $|C_i|$ - кількість об'єктів в і-му кластері,

k - кількість кластерів,

r - коефіцієнт відштовхування (repulsion), позитивно дійсне число перевершує 1.

За допомогою коефіцієнта відштовхування визначається ступінь подібності транзакцій в кластері, причому залежність обернено пропорційна, тобто з підвищенням значення коефіцієнта, знижується показник схожості транзакцій при порівнянні, і тим самим збільшується кількість кластерів.

Таким чином, постановка задачі кластеризації для алгоритму CLOPE виглядає так робота алгоритму проходить методом ітеративного перебору записів бази даних, а глобальність критерію оптимізації, заснованому на розрахунку параметрів кластерів, дозволяє обробляти дані значно швидше, ніж в порівнянні транзакцій між собою.

2.3.4 Алгоритм DBSCAN

Алгоритм DBSCAN (Density Based Spatial Clustering of Applications with Noise – щільнісний алгоритм для кластеризації просторових даних з присутністю шуму) – алгоритм з автоматичним вибором кількості кластерів. Він заснований на припущенні про тому, що щільність точок всередині кластерів більше, ніж поза кластерів. цей алгоритм дозволяє знаходити кластери довільної форми. Вперше був запропонований Мартіном

Естер, Гансом-Пітером Крігель і колегами в 1996 році як вирішення проблеми розбиття (спочатку просторових) даних на кластери довільної форми. Більшість алгоритмів, які виробляють плоске розбиття, створюють кластери за формою близькі до сферичних, так як мінімізують відстань об’єктів до центру кластера. Автори DBSCAN експериментально показали, що їх алгоритм здатний розпізнати кластери різної форми.

Ідея, покладена в основу алгоритму, полягає в тому, що всередині кожного кластера спостерігається типова щільність точок (об'єктів), яка помітно вище, ніж щільність зовні кластера, а також щільність в областях з шумом нижче щільності будь-якого з кластерів. Ще точніше, що для кожної точки кластера її сусідство заданого радіуса повинно містити не менше деякого числа точок, це число точок задається граничним значенням.

Неформально кластер це множина, всередині якої щільність більше, ніж поза нею. Точки кластера діляться на два типи внутрішні і граничні. Внутрішні точки ті точки, навколо яких досить щільно розташовуються інші точки. Граничні точки, що знаходяться на не більше, ніж на певній відстані від внутрішніх. Безліч точок, які не належать ні одного кластеру називається шумом.

3 ПРОГРАМНА РЕАЛІЗАЦІЯ Й ДОСЛІДЖЕННЯ АЛГОРИТМІВ

3.1 Вибір середовища розробки

На сьогоднішній день є багато готових рішень для реалізації Data Mining таких як RapidMiner, Statistica by DELL, Statistica bt StatSoft, GMDH Shell, Orange, KNIME, Weka та інші. Але оскільки це повноцінні додатки із певними обмеженнями та форматами, тоді вони не досить гнучкі для повноцінного дослідження.

Таким чином середовищем для дослідження обрано вільне модульне інтегроване середовище розробки програмного забезпечення Eclipse, а отже мовою програмування було обрано Java. Java відрізняється швидкістю, високим рівнем захисту і надійністю.

Для прискорення процесу дослідження та виключення можливих помилок при реалізації алгоритмів використовується зовнішня бібліотека. Із трьох най поширених варіантів таких як Apache Mahout, Weka та SPMF, була обрана найменш відома SPMF. А саме тому, що вона розвивається і зараз, на відміну від конкурентів.

SPMF - це бібліотека інтелектуального аналізу даних з відкритим вихідним кодом написана на Java, спеціалізується на видобутку шаблонів. Поширюється під ліцензією GPL v3. Вона пропонує реалізацію 93 даних алгоритмів інтелектуального аналізу для:

- видобутку послідовних шаблонів;
- видобутку послідовних правил;
- видобутку асоціативних правил;
- видобутку наборів даних;
- кластеризації;
- класифікації.

Вихідний код кожного алгоритму може бути інтегрована в інше програмне забезпечення Java. Крім того, SPMF може бути використана як окрема програма з простим для користувача інтерфейсом або з командного рядка.

3.2 Опис вихідних даних

Дана база містить медичні форми CMS-1500 за п'ять років, від 2008 до 2012. Цього достатньо для ґрунтовного аналізу.

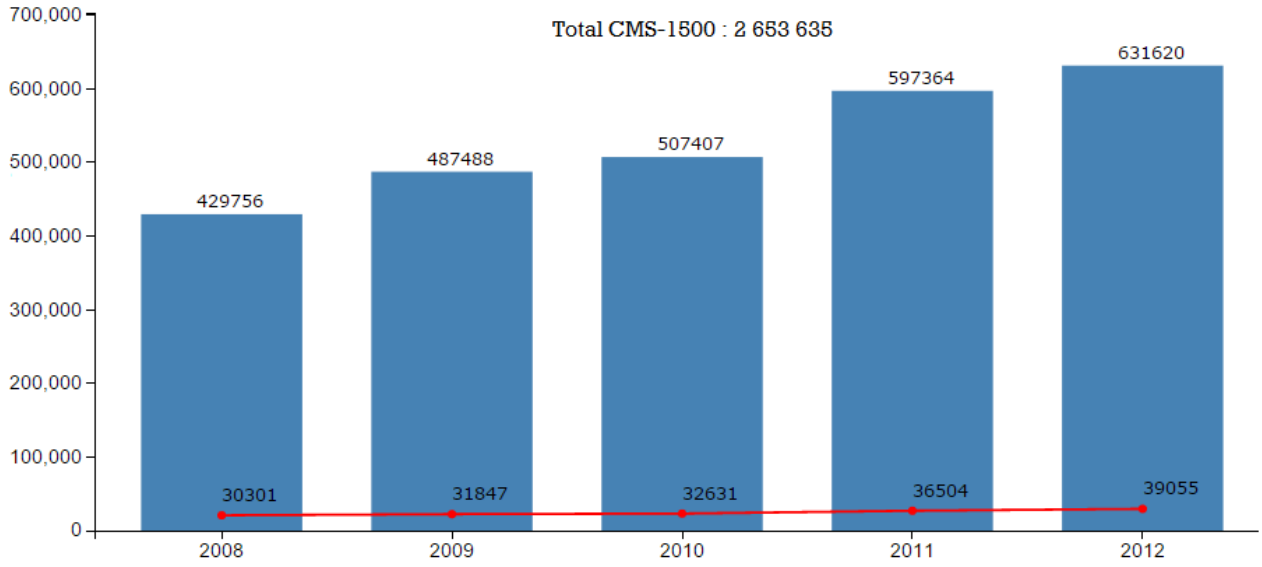


Рисунок 3.1 – Кількість CMS-1500 та унікальних пацієнтів по роках

З метою конфіденційності прибрані всі особисті дані пацієнтів. Специфіка наявної бази даних пов'язана із запальними захворюваннями кишечнику. У загальному випадку ми маємо можливість аналізувати таку інформацію:

– вік;
– стать;
– паління
– алергії;
– хронічні захворювання;
– діагноз, визначається набором ICD 9 кодів;
– процедури, пов'язані з кожним діагнозом визначається набором кодів CPT;
– рецепти препаратів, процедур;
– операції;
– результати лікування, визначається набором аналізу та суб'єктивного почуття пацієнта - показника якості лікування;
– тривалість лікування цього час, а в пацієнта призначається відповідних процедур та приписів.

Ґрунтуючись на обширній інформації, яка доступна в формах CMS 1500 можна проводити найрізноманітніші дослідження: починаючи від простого аналізу груп пацієнтів і закінчуючи складними завданнями прогнозування, аналізу методів лікування, аналізу результатів лабораторних досліджень з метою оптимізації методів лікування та побудови експертних систем. Але для дослідження даної атестаційної роботи обрана мета виявлення пацієнтів у яких висока ймовірність операцій пов'язаних із запальним захворюванням кишечнику.

Основними полями для аналізу в даному випадку виступають діагнози і процедури, а також період часу проведення процедур. Оскільки мета аналізу прогнозування операцій, то не обхідно визначити які саме процедури пов'язані з досліджуваними операціями:

- CPT коди від 44005 до 44346;
- CPT коди від 44602 до 44701;
- CPT коди від 45000 до 45190;
- CPT коди від 45395 до 45999;
- CPT коди від 46020 до 46060;
- CPT коди від 46270 до 46320;
- CPT коди від 49000 до 49084.

3.3 Результати роботи алгоритмів та їх аналіз

Для аналізу даних алгоритмами кластеризації були обрані параметри діагнози, процедури та час. Але якщо час має кореляцію стосовно заданих параметрів, то коди діагнозів та процедур ніяк не корелюють між собою. Тому необхідно визначити коефіцієнти для кодів відносно мети прогнозування.

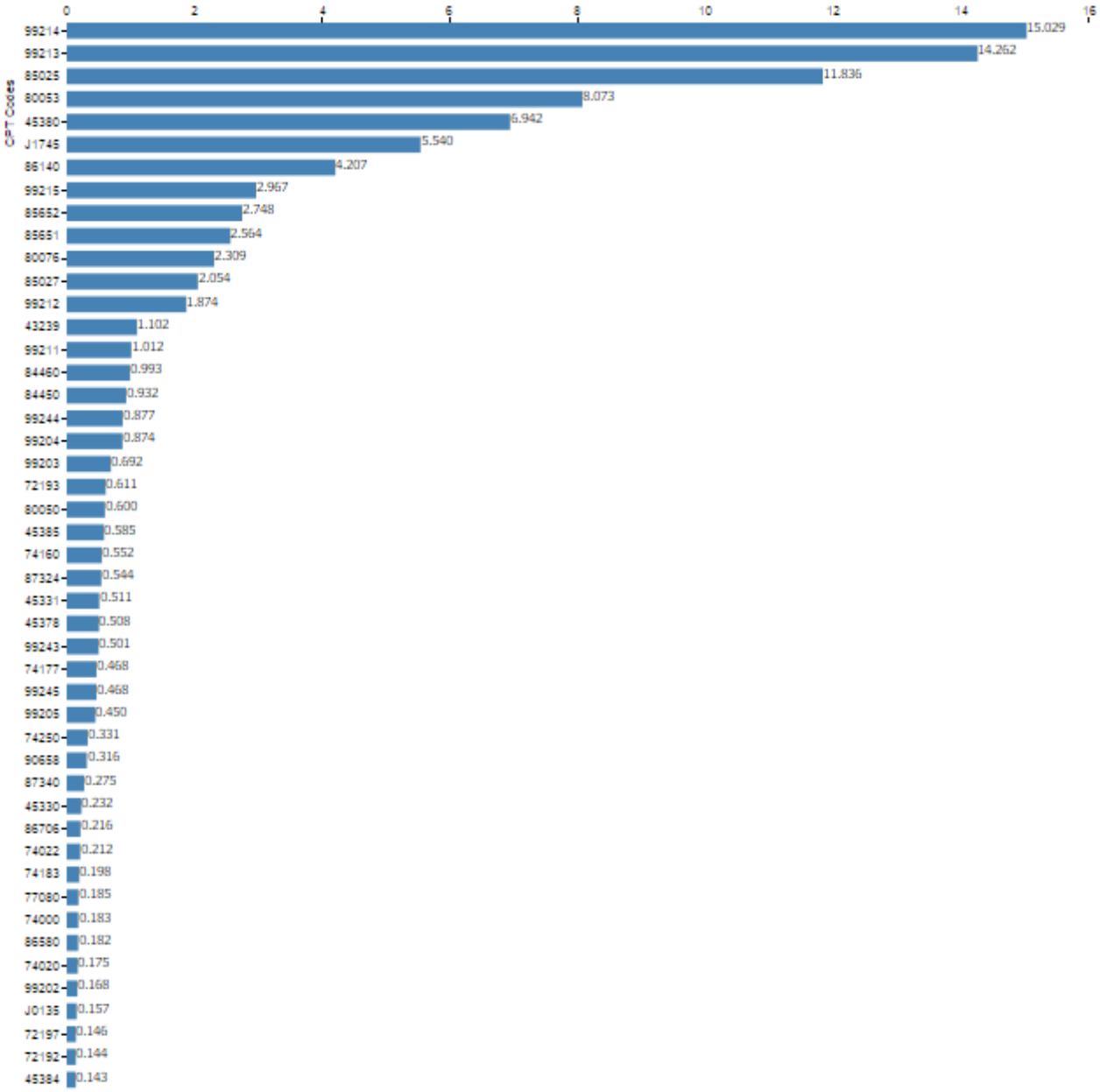


Рисунок 3.2 – Основні коефіцієнти CPT кодів

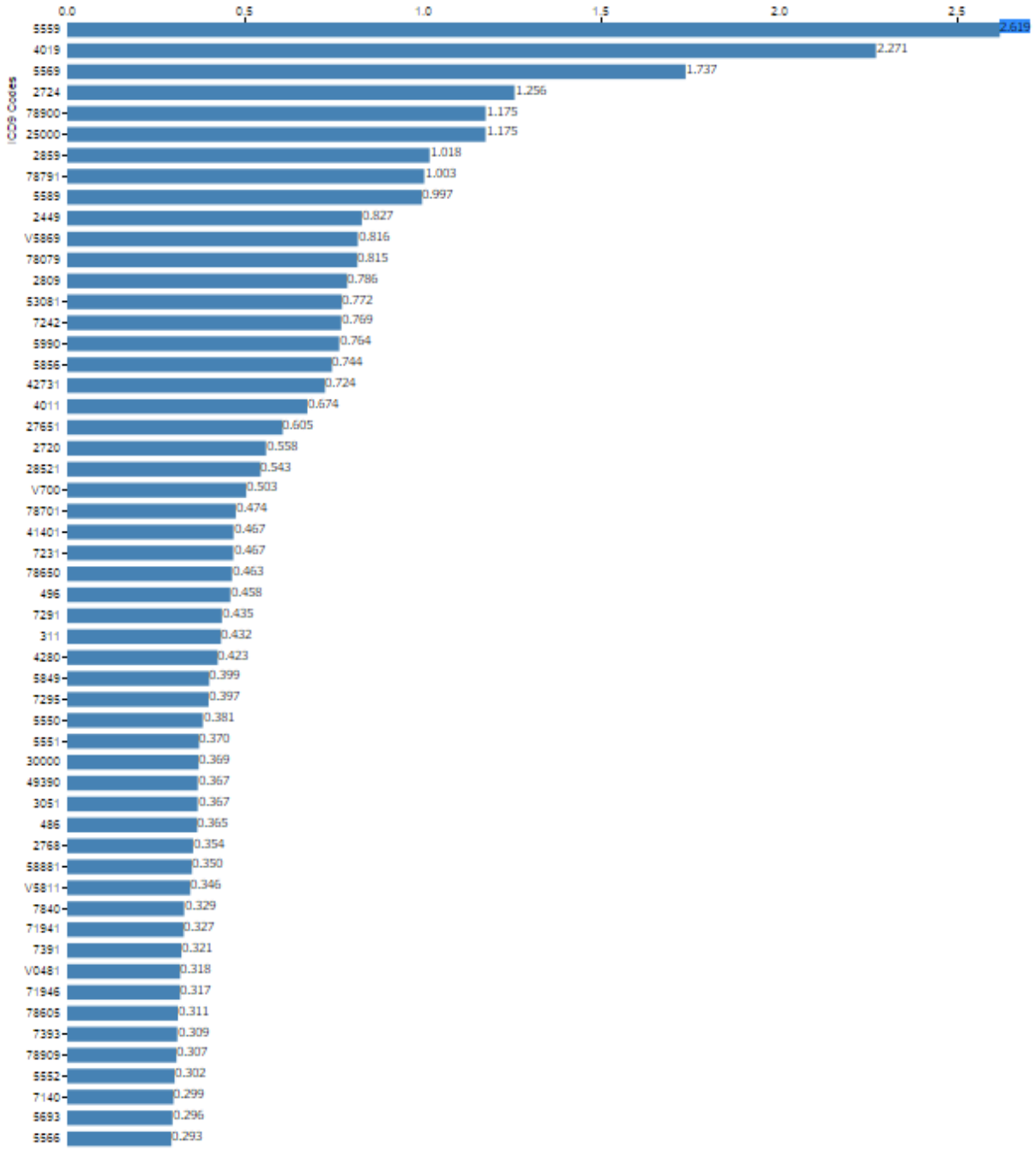


Рисунок 3.2 – Основні коефіцієнти ICD 9 кодів

Для зручності усі дані розбиті по роках. Оскільки дані можуть бути не повними, тому для побудови прогнозної моделі може використовуватись пацієнт, у котрого була операція пов'язана із запальними захворюваннями кишечнику, але за рік він не має жодного ICD 9 чи CPT коду із коефіцієнтом відміннім від нуля, а отже дані не повні чи не статистичні. Таким чином необхідно попередньо обробити усі вихідні дані на випадок таких аномалій.

Алгоритм прогнозування розпочинається із знаходження усіх пацієнтів, що мали будь яку операцію пов'язану із запальними захворюваннями кишечнику. Далі вони

розбиваються на дві групи: група для подальшого аналізу та побудови прогнозної моделі, та контрольна група для перевірки прогнозування. На наступному кроці виконується аналіз даних передуючих операції. Використовуються дані передуючих на рік, а отже пацієнти 2008 року не мають достатньої інформації.

Далі надані графіки результатів прогнозування, котрі відображають по роках скільки операцій було прогнозовано, скільки операцій відбулися, а також перетин цих двох множин.

Результати прогнозування на основі алгоритму K-Means представлені на рисунку 3.3.

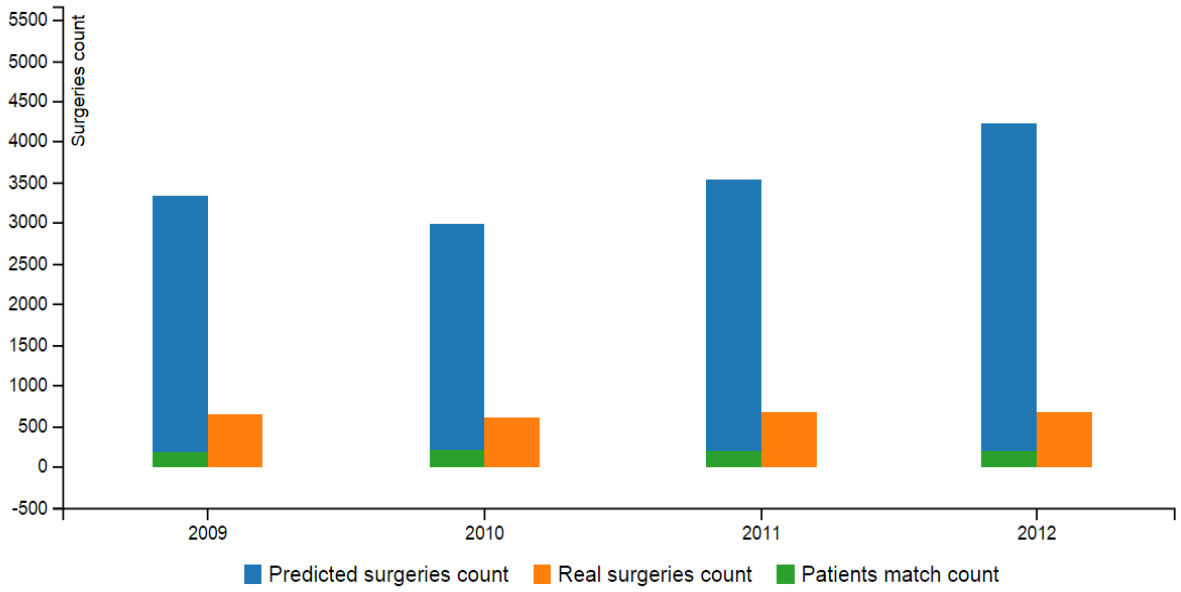


Рисунок 3.3 – Прогнозовані та дійсні операції, алгоритм K-Means

Обмеження: розрахований на невеликий обсяг даних.

Переваги: простота використання; швидкість використання; зрозумілість і прозорість алгоритму.

Недоліки: алгоритм занадто чутливий до викидів, які можуть спотворювати середнє; повільна робота на великих базах даних; необхідно задавати кількість кластерів самостійно.

Результати прогнозування на основі алгоритму Fuzzy C-Means представлені на рисунку 3.4.

Призначення: також добре пристосований для кластеризації великих наборів числових даних.

Переваги: нечіткість при визначенні об'єкта в кластер дозволяє визначати об'єкти, які знаходяться на кордоні, в кластери.

Недоліки: обчислювальна складність, завдання кількості кластерів, виникає невизначеність з об'єктами, які віддалені від центрів всіх кластерів.

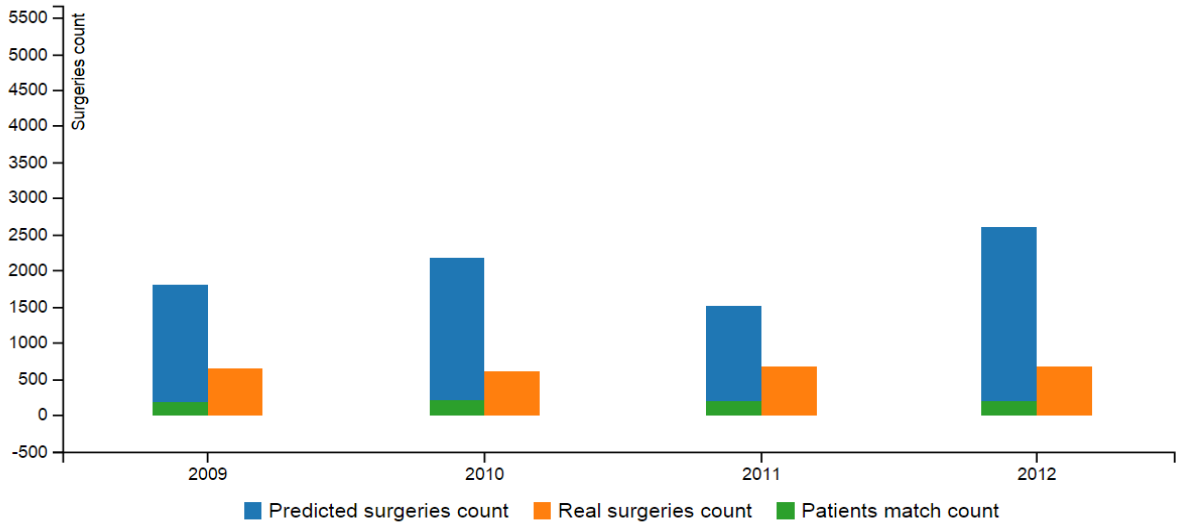


Рисунок 3.4 – Прогнозовані та дійсні операції, алгоритм Fuzzy C-Means

Результати прогнозування на основі алгоритму PAM представлені на рисунку 3.5.

Обмеження: оскільки PAM подібний до K-Means, то має ті ж обмеження – розрахований на невеликий обсяг даних.

Переваги: простота використання; швидкість використання; зрозумілість і прозорість алгоритму, алгоритм менш чутливий до викидів в порівнянні з K-Means.

Недоліки: необхідно задавати кількість кластерів; повільна робота на великих базах даних.

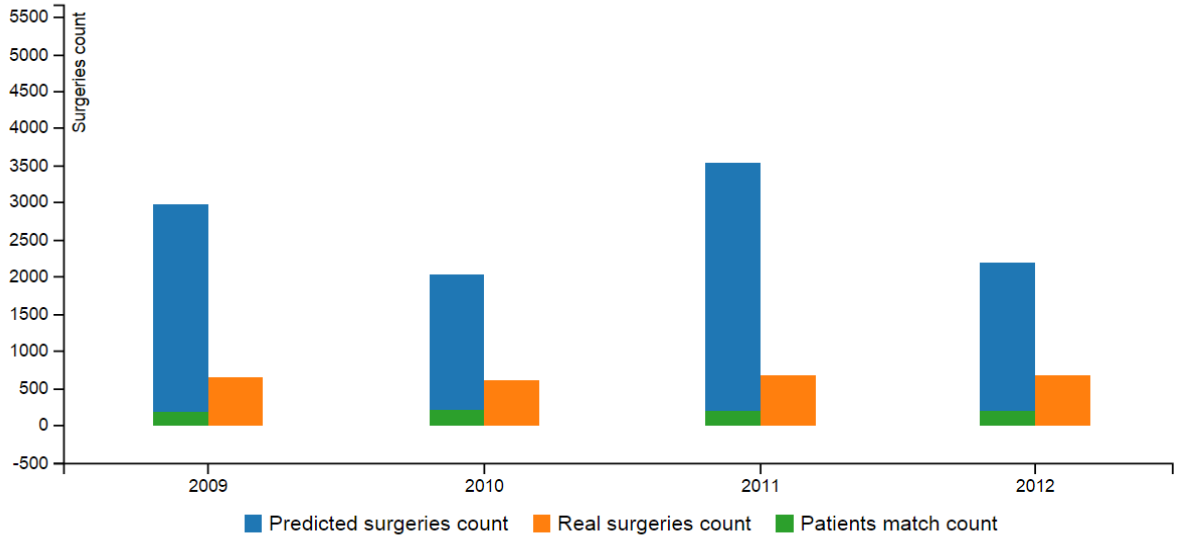


Рисунок 3.5 – Прогнозовані та дійсні операції, алгоритм PAM

Результати прогнозування на основі алгоритму CLOPE представлені на рисунку 3.6.

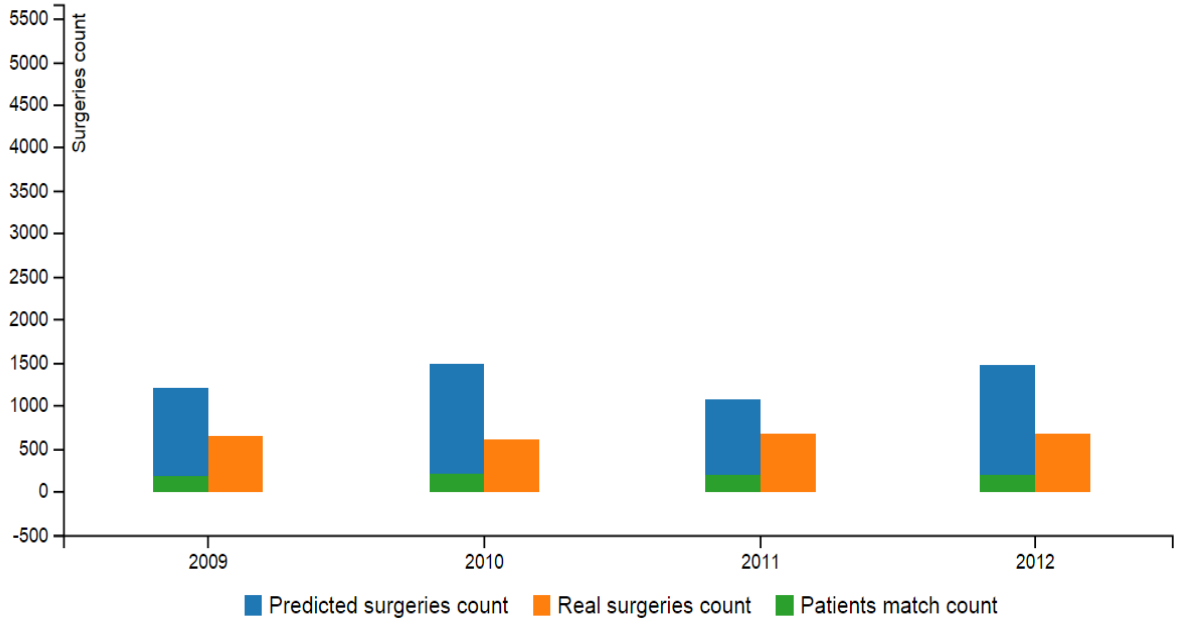


Рисунок 3.6 – Прогнозовані та дійсні операції, алгоритм CLOPE

Призначення: кластеризація величезних наборів категорійних даних.

Переваги: висока масштабованість і швидкість роботи, а так само якість кластеризації, що досягається використанням глобального критерію оптимізації на основі максимізації градієнта висоти гістограми кластеру. Він легко розраховується і

інтерпретується. Під час роботи алгоритм зберігає в RAM невелику кількість інформації по кожному кластеру і вимагає мінімальне число сканувань набору даних. CLOPE автоматично підбирає кількість кластерів, причому це регулюється одним єдиним параметром - коефіцієнтом відштовхування.

Результати прогнозування на основі алгоритму DBSCAN представлені на рисунку 3.7.

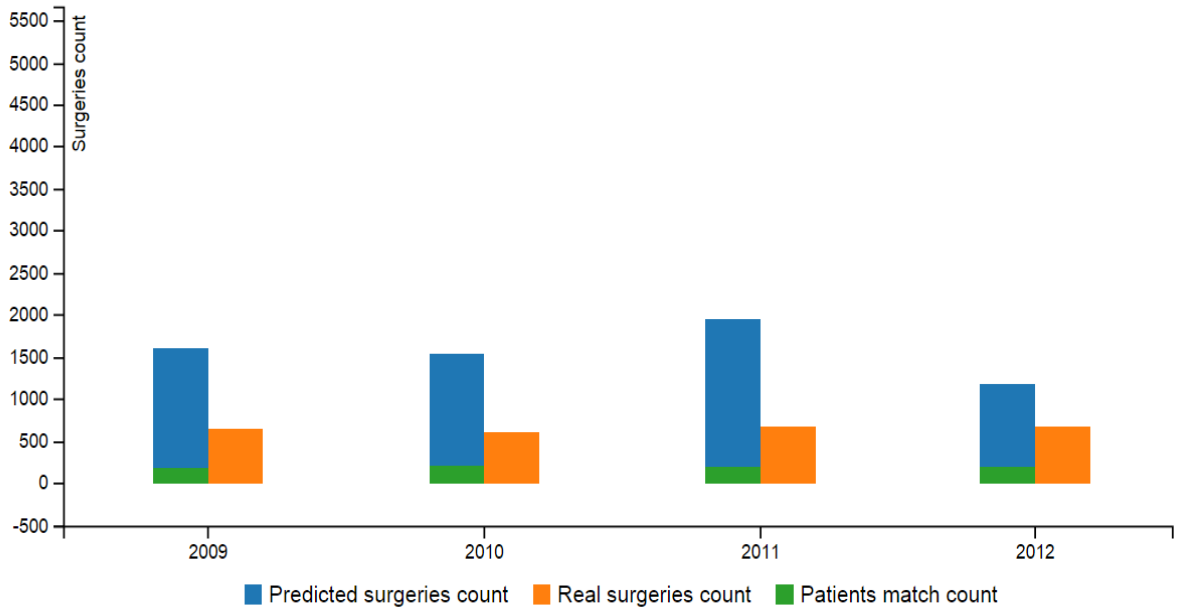


Рисунок 3.7 – Прогнозовані та дійсні операції, алгоритм DBSCAN

Призначення: кластеризація величезних наборів категорійних даних.

Переваги: не потребує завдання кількості кластерів, не чутливий до шумів, кластери можуть бути будь-якої форми.

Недоліки: можуть виникнути проблеми, якщо щільність різних кластерів значно відрізняється.

На рисунку 3.8 загреговані параметри точності усіх алгоритмів відносно медіан по роках попередніх значень.

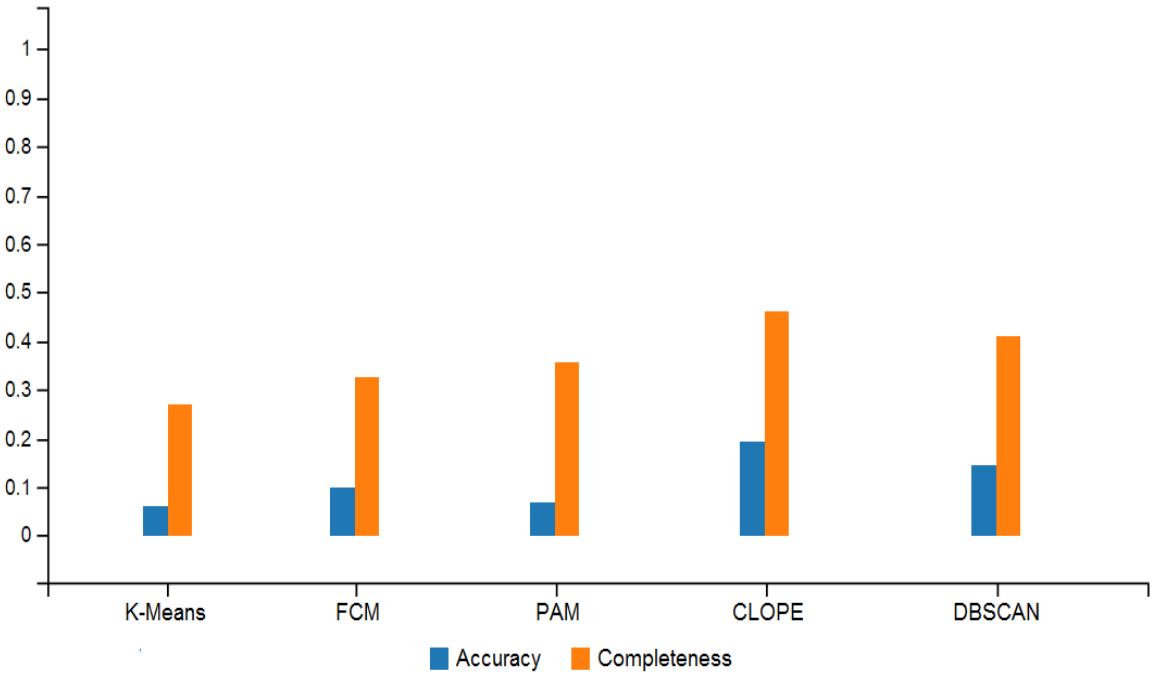


Рисунок 3.8 – Точність алгоритмів

Таким чином можна зазначити що алгоритм CLOPE має найбільші показники, а отже в нашому досліджені він виявився найбільш ефективним.

Але взагалі параметри точності занадто низькі для якогось практичного застосування. Для підвищення точності необхідно більше параметрів для порівняння і більш точні експерті оцінки. Також для поліпшення результатів можна застосувати для кластерного аналізу метрику із коефіцієнтами, але для цього також необхідні експертні дані.

Загалом, для отримання більш точних результатів кластерного аналізу замало для таких складних даних. Для подальшого дослідження необхідно використовувати більш складні методики такі, як нейронна мережа чи лінійна регресія.

4 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

4.1 Аналіз потенційно небезпечних і шкідливих виробничих факторів , що впливають на персонал

Персональні ЕОМ типу IBM PC AT має наступні характеристики:

- споживана потужність 350 Вт;
- робоча напруга 220 В;
- напруга джерел живлення +12 В, -12 В, 5 В;
- робоча частота 50 Гц.

Виходячи з приведених характеристик, очевидно, що для користувача існує небезпека поразки електричним струмом у разі недбалого поводження з комп'ютером і порушення правил експлуатації (невиконання огляду відкритих частин ПЕВМ, що знаходяться під напругою або знятих для ремонту вузлів і т. д.).

Джерелами підвищеної небезпеки можуть служити наступні елементи:

- розподільний щит;
- джерела живлення;

У відповідності з [11] до легкої фізичної роботи відносяться всі види діяльності, вироблювані сидячи і не вимагаючи фізичної напруги. Робота користувача розробленого пакету програм відноситься до категорії Іа.

Згідно з [11] приміщення для ПЕОМ по ступеню небезпеки поразки людини електричним струмом відноситься до приміщень без підвищеної небезпеки (немає струмопровідної полови, вогкості, підвищеної температури, можливості одночасного дотику до корпусів устаткування з “землею” і до струмонесучих частин).

У відповідності з [12] при обслуговуванні ПЕВМ мають місце фізичні і психофізичні небезпечні, а також шкідливі виробничі чинники:

- підвищене значення напруги в електричному ланцюзі, замикання якого може відбутися через тіло людини;
- підвищений рівень статичної електрики;
- підвищений рівень електромагнітних випромінювань;
- підвищена або знижена температура повітря робочої зони;
- підвищена або знижена рухливість повітря;
- підвищена або знижена вогкість повітря;
- відсутність або недолік природного світла;

- підвищена пульсація світлового потоку;
- недостатня освітленість робочого місця;
- підвищений рівень шуму на робочому місці;
- розумове перенапруження;
- емоційні навантаження;
- монотонність праці.

Щодо до впливу на довкілля, то програмний засіб, який було розроблено під час дипломного проекту на довкілля ніяк не впливає.

Діяльність за темою магістерської роботи в процесі її виконання впливає на навколишнє природнє середовище і регламентується нормами діючого законодавства: [11-26].

Основним екологічним аспектом в процесі діяльності за даними спеціальностями є процеси впливу на атмосферне повітря та процеси поводження з відходами, які утворюються, збираються, розміщуються, передаються на відалення (знешкодження), утилізацію, тощо в ІТ галузі.

Вплив на атмосферне повітря при нормальних умовах праці не оказує, бо не має в приміщенні сканерів, принтерів та інших джерел викиду забруднюючих речовин в повітря робочої зони.

В процесі діяльності виникають процеси поводження з відходами ІТ галузі. Нижче надано перелік відходів, що утворюються в процесі роботи:

- батарейки та акумулятори (малі) - ІІІ клас небезпеки;
- макулатура - IV клас небезпеки;
- матеріали пакувальні, що не вміщують целюлозу - IV клас небезпеки;
- матеріали пакувальні, що вміщують п/ет, п/пр - IV клас небезпеки;
- змінні носії інформації - IV клас небезпеки.

Наводяться вимоги зберігання виявлених за своєю роботою відходів відповідно до вимог Державних санітарних правил і норм [26].

Відходи в міру їх накопичення збирають у тару, відповідну класу небезпеки, з дотриманням правил безпеки, після чого доставляють до місця тимчасового зберігання відходів відповідно до затвердженої схеми їх розміщення Зазначені для зберігання відходів місця чи об'єкти повинні використовуватися лише для заявлених відходів.

Не допускається зберігання відходів у невстановлених схемою місцях, а також перевищення норм тимчасового зберігання відходів.

Способи тимчасового зберігання відходів визначаються видом, агрегатним станом і класом небезпеки відходів:

- Відходи III класу небезпеки зберігаються в тарі, яка забезпечує локалізацію зберігання, дозволяє виконувати вантажно-розвантажувальні і транспортні роботи і виключає поширення в ОС шкідливих речовин;

- Відходи IV класу небезпеки можуть зберігатися відкрито на промисловому майданчику у вигляді конусоподібної купи, звідки їх автонавантажувачем перевантажують у самоскид і доставляють на місце утилізації або захоронення;

В разі тимчасового зберігання відходів у стаціонарних складах або промислових приміщеннях повинні бути забезпечені санітарно-гігієнічними етичні вимоги до повітря робочої зони згідно з [25].

Не допускається змішування відходів різних видів і класів небезпеки з будівельними і побутовими відходами, відходами дерев'яної, металевої, синтетичної тари, відходами текстильних матеріалів (старий спецодяг, ганчірки) і інш.

Проведення заготовки, здачі, переробки та реалізації металобрухту встановлені окремо [27].

Всі відходи, що утворюються в процесі діяльності/роботи, підлягають обліку.

Вимоги безпеки при поводженні з відходами:

Під час роботи з відходами (прибирання виробничих приміщень, збір і сортування, навантаження, транспортування, розвантаження та ін.) працівники та обслуговуючий персонал підприємства повинні бути забезпечені засобами індивідуального захисту та дотримуватися вимог інструкцій з охорони праці, що діють на підприємстві.

Наведено перелік деяких відходів, які передаються на утилізацію організаціям, які мають ліцензію на поводження з відходами як вторинної сировини:

- лом і кускові відходи міді, бронзи, латуні, алюмінію, свинцю;
- брухт чорних металів;
- макулатура;
- склобій;
- матеріали текстильні вторинні;
- відходи деревини кускові
- відпрацьовані фільтрувальні засоби індивідуального захисту
- відпрацьовані вогнегасники
- матеріали пакувальні вторинні

Відвантаження таких відходів здійснюється відповідно до договору (контракту).

Побутові та будівельні відходи вивозяться на полігон твердих побутових відходів міста, також відповідно до договору з комунальним дорожньо-експлуатаційним управлінням.

Особи, винні в порушенні встановленого порядку поводження з відходами (порушення правил обліку відходів, самовільне складування і видалення відходів, передача відходів в інші підприємства/організації з порушенням встановлених правил), згідно законодавства несуть дисциплінарну, адміністративну або кримінальну відповідальність.

4.2 Заходи щодо техніки безпеки

Основним небезпечним чинником при роботі з ЕОМ є небезпека поразки людини електричним струмом, яка усугубляє тим, що органи чуття людини не можуть на відстані знайти наявності електричної напруги на устаткуванні.

Проходячи через тіло людини, електричний струм надає на нього складну дію, що є сукупністю термічної (нагрів тканин і біологічних середовищ), електролітичної (розкладання крові і плазми) і біологічної (роздратування і збудження нервових волокон і інших органів тканин організму) дій.

Ступінь ураження людини електричним струмом залежить від наступних факторів:

- значення сили струму;
- електричного опору тіла людини і тривалості протікання через нього струму;
- роду і частоти струму;
- індивідуальних властивостей людини і навколишнього середовища.

Даним проектом передбачаються наступні технічні способи і засоби, застережливі поразки людини електричним струмом:

- заземлення електроустановок;
- занулення;
- захисне відключення;
- електричне розділення сітей;
- використовування малої напруги;
- ізоляція струмоведучих частин;
- огорожа електроустановок.

Проведемо розрахунок заземлюючого пристрою.

Початкові дані для розрахунку заземлюючого пристрою:

– напруга установки, що заземляється, - 220В;
– режим нейтралу мережі - з ізольованою нейтралою;
– питомий опір ґрунту – 100 Ом·м(суглинок);
– гранично допустимий опір заземлюючого пристрою - 4 Ом;
– характеристика кліматичної зони (III):

а) середня багаторічна низька температура, ^{o}C - від –14 до -10;

б) тривалість замерзання вод, дні - 150;

в) коефіцієнт сезонності для вертикального електроду завдовжки 3м -1,5.

Визначимо розрахунковий опір ґрунту (Ом·м) по формулі (4.1).

$$\rho_{расч} = \psi \cdot \rho = 1{,}5 \cdot 100 = 150 Ом \cdot м \qquad (4.1)$$

де ρ - питомий опір ґрунту;

ψ_i – кліматичний коефіцієнт, що враховує стан ґрунту під час вимірювань (таблиця 4 [12]).

Розрахуємо опір розтіканню одиночного трубчастого заземлювача по формулі (4.2).

$$R_{з.1} = (\frac{\rho_{расч}}{2 \cdot \pi \cdot l}) \cdot \ln(4 \cdot \frac{l}{d}) \qquad (4.2)$$

де l – довжина заземлювача (l=5м);

d – діаметр труби і стрижня (d=0,05м);

$$R_{з.1} = (\frac{\rho_{расч}}{2 \cdot \pi \cdot l}) \cdot \ln(4 \cdot \frac{l}{d}) = (\frac{150}{2 \cdot 3{,}14 \cdot 5}) \cdot \ln(4 \cdot \frac{5}{0{,}05}) = 28{,}6 \text{ Ом}$$

Розрахуємо кількість паралельно сполучених одиночних заземлювачей по формулі (4.3).

$$n = \frac{R_{з.1}}{R_{доп} \cdot \eta} = \frac{28{,}6}{4 \cdot 0.47} = 15{,}2 \qquad (4.3)$$

де Rдоп=4. – самий допустимий опір заземлюючого пристрою;

η - коефіцієнт використовування ґрунтового заземлення (для шістки заземлювачей η=0,47).

Округлятимемо отримане значення у більшу сторону n=[15,2]=16.

Розрахуємо довжину горизонтальної сполучної смуги по формулі (5.4).

$$L = a \cdot (n-1) = 3 \cdot (16-1) = 45м \tag{4.4}$$

де a – відстань між вертикальними заземлювачами (a=3м);

n – кількість вертикальних заземлювачей (n=16).

Розрахуємо опір сполучної смуги по формулі (4.5).

$$R_n = \frac{\rho_{расч}}{2 \cdot \pi \cdot l} \cdot \ln(\frac{L^2}{d \cdot h}) \tag{4.5}$$

де d – еквівалентний діаметр смуги шириною l=5 (d=0,05м);

h – глибина заставляння смуги (h=0,8м).

$$R_n = \frac{\rho_{расч}}{2 \cdot \pi \cdot l} \cdot \ln(\frac{L^2}{d \cdot h}) = \frac{150}{2 \cdot 3{,}14 \cdot 5} \cdot \ln(\frac{45^2}{0{,}05 \cdot 0{,}8}) = 51{,}7 \text{ Ом}$$

Розрахуємо результуючий опір заземлюючого електроду з урахуванням сполучної смуги по формулі (4.6).

$$R_{гр} = \frac{R_{з.1} \cdot R_n}{R_{з.1} \cdot \eta_n + R_n \cdot n \cdot \eta_з} \le R_{доп} \tag{4.6}$$

де η_n – коефіцієнт використовування сполучної смуги (для 6-і заземлювачей $\eta_{n=0,27}$).

$$R_{гр} = \frac{R_{з.1} \cdot R_n}{R_{з.1} \cdot \eta_n + R_n \cdot n \cdot \eta_з} = \frac{26{,}6 \cdot 51{,}7}{26{,}6 \cdot 0{,}27 + 51{,}7 \cdot 16 \cdot 0{,}47} = 3{,}47 \text{ Ом}$$

3,47<4 $\Rightarrow$ умова забезпечення електробезпеки персоналу виконується.

Таким чином, остаточна кількість заземлювачей 15 шт.

4.3 Заходи, що забезпечують виробничу санітарію і гігієну праці

Підвищення працездатності людини і збереження його здоров'я забезпечується стабільними метеорологічними умовами.

Мікроклімат виробничих приміщень – це поєднання температури, вогкості і швидкості руху повітря, а також температури навколишніх поверхонь. Значне коливання параметрів мікроклімату приводить до порушення систем кровообігу, нервової і пітовидільної, що може викликати підвищення або пониження температури тіла, слабкість, запаморочення і навіть непритомність.

В приміщенні для виконання робіт операторського типу, пов'язаних з нервово-емоційною напругою, проектом передбачається дотримання наступних нормованих величин параметрів мікроклімату (див. таблицю 4.1).

Таблиця 4.1 - Оптимальні параметри мікроклімату в робочій зоні виробничого приміщення для категорії робіт 1

Період року	Температура, оС	Відносна вогкість %	Швидкість руху повітря, м/с
Холодний	22.24	40.60	0,1
Теплий	23.25	40.60	0,1

Оскільки в приміщенні немає джерел виділення шкідливих речовин, можна використовувати природну вентиляцію. Площа приміщення складає 32 м2 . Для забезпечення прийнятних параметрів мікроклімату в приміщенні з такою площею можна використовувати 1 кондиціонер типу БК-2000.

Спектр випромінювання монітора комп'ютера включає рентгенівську, ультрафіолетову, інфрачервону області, а також широкий діапазон хвиль інших частот. Небезпека рентгенівського проміння нехтує мала, оскільки цей вид випромінювання поглинається речовиною екрану.

Для зниження дії електромагнітного випромінювання пропонується захист часом і відстанню. Захист часом передбачає обмеження часу перебування людини в зоні дії полів. Тривалість роботи на ПЕОМ повинна складати не більше 3.5–4.5 години.

Також необхідно забезпечити раціональне освітлення в робочому приміщенні. В проекті, що розробляється, передбачається використовувати суміщене освітлення. В світлий час доби приміщення освітлюватиметься через віконні отвори, в решту часу використовуватиметься штучне освітлення.

Штучне освітлення в робочому приміщенні передбачається здійснювати з використанням люмінесцентних джерел світла в світильниках загального освітлення, оскільки люмінесцентні лампи володіють високою світловою віддачею до 75 Лам/Вт і більш, тривалим терміном служби до 10000 годин, спектральним складом випромінюваного світла, близьким до сонячного.

Зорова робота оператора ПЕВМ відповідно до [15] відноситься до розряду Va з світловим потоком Фл=3120 кожна. Нормована освітленість на робочому місці (Ен) при загальному освітленні складає 200 лк.

Проведемо розрахунок кількості світильників в робочому приміщенні завдовжки a=6 м, шириною b=3 м, заввишки c=4 м. Формула розрахунку штучного освітлення при горизонтальній робочій поверхні методом світлового потоку (4.7):

$$Фл = \frac{Ен \cdot S \cdot Z \cdot K}{N \cdot U \cdot M} \qquad (4.7)$$

де *Фл* – світловий потік, Лм;

Ен – нормована освітленість;

S – площа підлоги, кв.м;

Z=1.1–1.3 - поправочний коефіцієнт світильника (для стандартних світильників);

K – коефіцієнт запасу, що враховує зниження освітленості в процесі експлуатації світильників;

N – число світильників;

U=0.55–0.6 – коефіцієнт використовування, залежний від типу світильника, показника індексу приміщення і др.;

M – число ламп в світильнику.

З формули (4.7) виразимо N і визначимо кількість світильників для даного приміщення:

$$N = \frac{200 \cdot 18 \cdot 1,2 \cdot 1,5}{3120 \cdot 0,6 \cdot 2} = 1,7$$

Виходячи з цього, рекомендується використовувати 2 світильники. Світильники слід розміщувати рядами, бажано паралельно стіні з вікнами. Схема розташування світильників зображена на рис. (4.1).

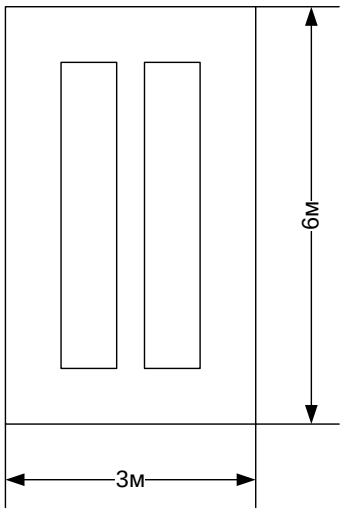


Рисунок 4.1 – Схема розташування світильників

Зниження шуму можна добитися раціонально розпланувавши приміщення, установкою устаткування на спеціальні амортизуючи прокладки. Згідно вимогам [9] рівні звуку не повинні перевищувати 50 дБ.

Для зниження стомлюваності обслуговуючого персоналу в приміщеннях, де розташовані обчислювальні засоби передбачаються використовувати спокійні колірні поєднання і покриття, що не дають відблисків. Від електромагнітного випромінювання, витікаючого від ПЕОМ, використовуються захисні екрани.

Для забезпечення чистоти повітря і відповідних мікрокліматичних умов пропонується застосувати приточування-витяжну вентиляцію. Для зменшення дії шкідливих речовин і загазованості для роботи з розплавленими матеріалами робоче місце забезпечується примусовою витяжною вентиляцією. Цей метод забезпечує притоку потрібної кількості свіжого повітря (30 мм^3 /ч на одного працюючого).

Кількість повітря, яка необхідна подавати в приміщення для забезпечення необхідних параметрів повітряного середовища, визначається на підставі кількість тепла, вологи і шкідливих речовин, що поступають в приміщення, а також враховуючи видалення повітря місцевими відсмоктуваннями від устаткування, загальнообмінною вентиляцією.

4.4 Рекомендації по пожежній профілактиці

Пожежі представляють небезпеку для життя людини і зв'язані як з матеріальними втратами, так і з відмовою засобів обчислювальної техніки, що спричиняє за собою порушення ходу технологічного процесу.

Горючими матеріалами в приміщенні, де розташовані ПЕОМ, є:

– поліамід - матеріал корпусу мікросхеми. Горюча речовина. Температура самозапалення 420 ^{o}С, енергія запалення 2мДж;

– полівінілхлорид - ізоляційний матеріал. Горюча речовина. Температура самозаймання 480 єС, енергія запалення 50мДж;

– склостоліт ДЦ - матеріал друкарської платні. Складногорючий матеріал;

– пластикат кабельний No.489 - матеріал ізоляції кабелю. Складногорючий матеріал. Температура самозаймання 1500 єС;

– плита деревостружкова - будівельний і обробний матеріал, матеріал з якого виготовлені меблі. Складнозапалений матеріал. Показник горючості 1.8;

– папір – довідкова і робоча документація, література. Горючий матеріал. Показник горючості більше 2.1.

Відповідно до [16] приміщення відноситься до категорії В (пожежовибухонебезпечної).

Джерелами запалення можуть бути:

– іскри при замиканні і розмиканні ланцюгів;

– іскри і дуги коротких замикань;

– перегріви від тривалого перевантаження і наявності перехідного опору;

– розряди статичної електрики.

Для того, щоб зупинити реакцію горіння, порушують умови її виникнення і підтримки. Звичайно для гасіння використовуються порушення двох основних умов сталого стану – пониження температури і режим руху газів. Пониження температури може бути досягнутий шляхом введення речовин, які поглинають багато тепла в результаті випаровування і дисоціації (наприклад, вода, порошки).

При повному тому, що згоряє органічних сполук утворюються С, SO, Н Про, N, а при тому, що згоряє неорганічних з'єднань – оксиди. Залежно від температури плавлення і тривалості реакції можуть знаходиться або у вигляді розплавів (Al O, Ti O), або підійматися в повітря у вигляді диму (P O, Na Про, MgO).

Склад продуктів неповного згоряє горючих речовин складений і різноманітний. Це можуть бути горючі речовини:

– H, C, CH;

– атомарний водень і кисень;

– різні радикали – OH, CH .

Продуктами неповного згоряє можуть бути також оксиди азоту, спирти, альдегіди, кетони і високотоксичні з'єднання, наприклад, синильна кислота.

Для захисту персоналу від дій небезпечних і шкідливих чинників пожежі проектом передбачено застосування промислового фільтруючого протигаза з коробкою марки В (жовтий).

До системи запобігання пожежі відносяться: запобігання утворення горючого середовища і освіти в горючому середовищі джерел запалення, забезпечення пожежебеспеки устаткування.

Щоб запобігти пожежі в обчислювальних центрах, проектом пропонується виконання наступних вимог:

– електроживлення ЕОМ має автоматичне блокування відключення електроенергії на випадок перегріву системи, що може бути результатом зупинки системи охолоджування і кондиціонування;

– система вентиляції обчислювальних центрів обладнується блокуючими пристроями, що забезпечують її відключення на випадок пожежі. Система обладнується вогнеперегороджуючими клапанами;

– застосування устаткування, що задовольняє вимогам електростатичної іскробезпеки [12];

– після закінчення роботи, перед закриттям приміщення, всі електроустановки і персональні комп'ютери відключаються від сіті електроживлення;

– в приміщеннях обчислювальних центрів забороняється:

– влаштовувати електророзетки на основах, що згоряють;

– використовувати синтетичні доріжки і килими;

– користуватися побутовими електронагрівальними приладами;

– захаращувати евакуаційні виходи і проходи;

– влаштовувати на вікнах глухі грати;

– залишати без нагляду включену в електромережу апаратуру, що використовується для вимірювань і нагляду.

Для протипожежного захисту проектом пропонується обладнати приміщення площею 18 м2, яке відноситься до категорії В, автоматичною протипожежною сигналізацією із застосуванням датчиків сповіщення РІД-1 (оповіщувач димовий іонізаційний) в кількості 1 штуки і застосовується в первинних засобах пожежегасінні. Площа контрольована оповіщувачем 150 м2.

Крім того, необхідно проводити навчання робочого персоналу правилам пожежної безпеки.

Розрахуємо вірогідність виникнення пожежі у виробничому приміщенні у разі запалювання транзистора:

$$Q = л \cdot T \cdot Ркз/отк \cdot Qвоспл \cdot Рзащ \qquad (4.8)$$

де л – інтенсивність відмов пожежеопасних ЕРІ;

Т – час роботи пожежеопасного ЕРІ за оцінюваний інтервал часу;

Ркз/отк - умовна вірогідність виходу ЕРІ в стан короткого замикання при його відмові;

Qвоспл - вірогідність запалювання ЕРІ, що знаходиться в стані короткого замикання;

Рзащ – вірогідність відмови захисту пожежеопасного ЕРІ. Якщо захист відсутній, Рзащ приймається рівній 1.

Вірогідність виникнення пожежі у разі запалювання транзистора:

$$Q = 1 \cdot 10^{-6} \cdot 1 \cdot 10^{-4} \cdot 0.1 \cdot 1 \cdot 10^{-4} = 1 \cdot 10^{-15}$$

Розрахована вірогідність виникнення пожежі значно менше допустимої, яка складає $1 \cdot 10^{-6}$.

ВИСНОВКИ

Сучасні медичні системи вже накопичили дуже великий об’єм статистичних даних по всьому світі. Накопичені дані мають величезний потенціал у вигляді неочевидних, об'єктивних і практично корисних закономірностей.

У результаті виконання атестаційної роботи були досліджені основні алгоритми кластеризації у рамках завдання кластеризації технології інтелектуального аналізу даних з метою побудови прогнозної моделі. Для виконання роботи була досліджена технологія інтелектуального аналізу даних, її етапи та задачі. Був проведений порівняльний аналіз досліджуваних алгоритмів як з точки зору побудови прогнозної моделі, так і з точки зору придатності до роботи з вихідними даними, а також їх об’єму.

У ході дослідження виявлено що даний підхід до прогнозування має низьку ефективність у зв’язку із недостатньою кількістю експертних даних. Запропоновано рішення для підвищення точності методу.

Дана робота заснована на кластерному аналізі та не є вичерпною. В майбутньому планується також дослідити поставлене завдання ґрунтуючись на нейронних мережах або регресійному аналізі для отримання більш точних результатів прогнозування.

В розділі «ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ» були проаналізовані небезпечні і шкідливі виробничі чинники, що роблять вплив на персонал, розроблені заходи щодо техніки безпеки, заходу, забезпечуючи виробничу санітарію і гігієну праці, а також заходи щодо пожежної профілактики.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1) Тарасова С.А. Математические методы прогнозирования в медицине / С.А. Тарасова // Актуальные проблемы и перспективы преподавания по математике [Текст] : сб. ст. / Юго-Зап. гос. ун-т. Курск, 2013.

2) Прикладная статистика: классификация и снижение размерности. [Текст] / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин — М.: Финансы и статистика, 1989.

3) Мандель И. Д. Кластерный анализ. [Текст] — М.: Финансы и Статистика, 1988.

4) Айвазян С. А., Бухштабер В. М., Юнюков И. С., Мешалкин Л. Д. Прикладная статистика: Классификация и снижение размерности. [Текст] — М.: Финансы и статистика, 1989.

5) Уиллиамс У.Т., Ланс Д.Н. Методы иерархической классификации [Текст] // Статистические методы для ЭВМ / под ред. М. Б. Малютова. - М.: Наука, 1986.

6) Кулаичев А.П. Методы и средства комплексного анализа данных. [Текст] – М.: ИНФРА- М, 2006.

7) Котов А., Красильников И. Кластеризация данных. [Текст] 2006. 16с.

8) Чубукова И. А. Data Mining: учебное пособие. [Текст] М.: Интернет-Университет Информационных Технологий; БИНОМ; Лаборатория знаний, 2006.

9) Дюк В., Самойленко А. Data Mining: Учебный курс. [Текст] СПб.: Изд-во «Питер», 2001. 368 с.

10) J. Frank Wharam, Jonathan P. Weiner, “The Promise and Peril of Healthcare Forecasting.” [Текст] Am J Manag Care. 2012

11) 12.1.005–88. ССБТ. Общие санитарно–гигиенические требования к воздуху рабочей зоны.

12) 12.0.003–74. ССБТ. Опасные и вредные производственные факторы. Классификация.

13) 12.1.009–76. ССБТ. Электробезопасность. Термины и определения

14) 12.1.003-83. ССБТ. Шум. Общие требования безопасности

15) ДБН В.2.5-28-2006. Природне і штучне освітлення

16) НАПБ Б.03.002-2007. Нормы определения категорий помещений, зданий и наружных установок по взрывопожарной и пожарной опасности

17) ДСТУ Б А.3.2-13:2011 Система стандартів безпеки праці. Будівництво. Електробезпечність. Загальні вимоги

18) Закон України «Про охорону навколишнього природного середовища»

19) Закон України «Про забезпечення санітарного та епідемічного благополуччя населення»

20) Закон України «Про відходи»

21) Закон України «Про охорону атмосферного повітря»

22) Закон України «Про захист населення і територій від надзвичайних ситуацій техногенного та природного характеру»

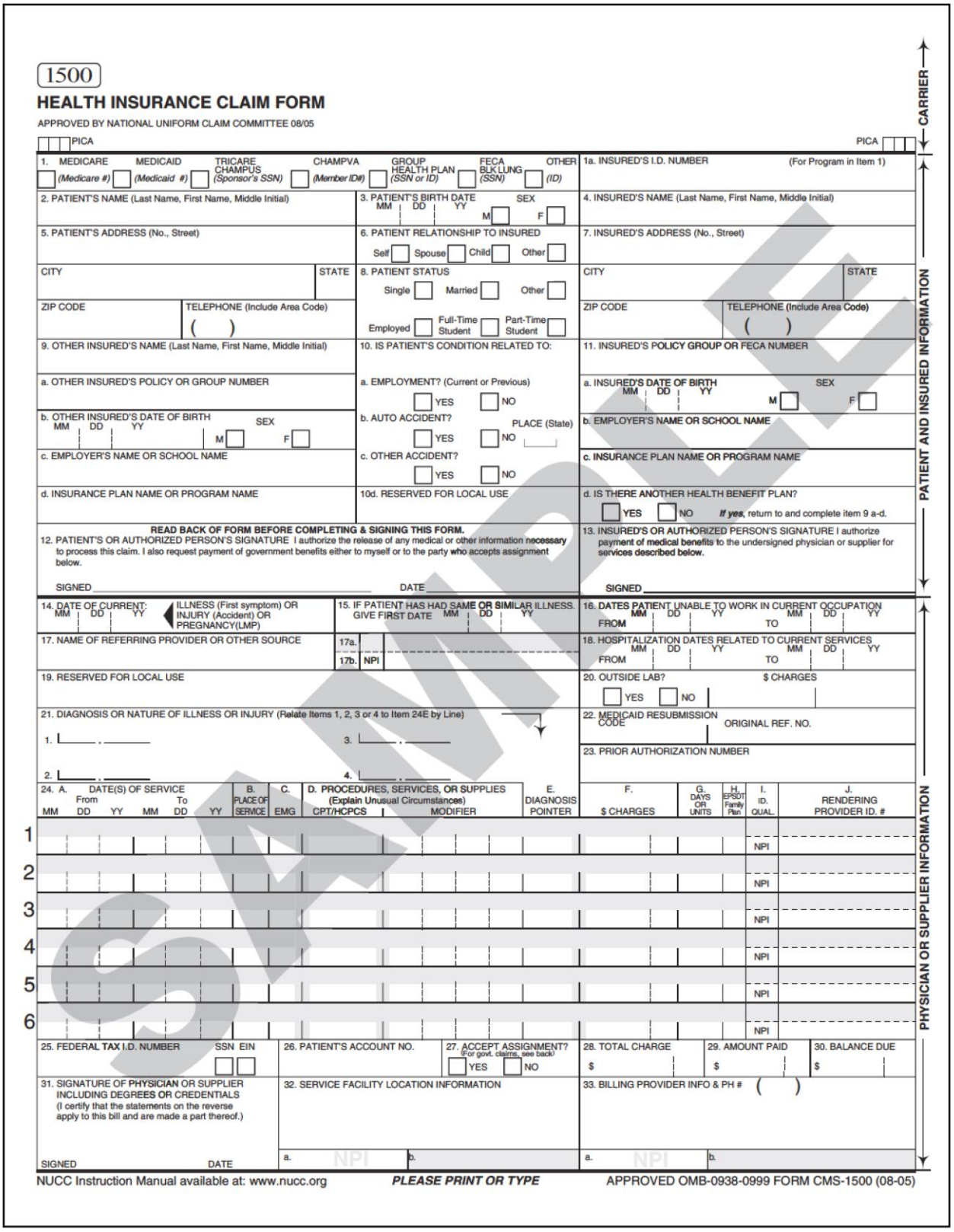
23) Водний кодекс України

24) ДСанПіН 2.2.7.029-99. Гігієнічні вимоги щодо поводження з промисловими відходами та визначення їх класу небезпеки для здоров'я населення.

25) ГОСТ 12.1.005-88 Система стандартов безопасности труда. Общие санитарно-гигиенические требования к воздуху рабочей зоны.

26) Законом України «Про металобрухт»

ДОДАТОК А.
Медична форма CMS-1500



1500

HEALTH INSURANCE CLAIM FORM

APPROVED BY NATIONAL UNIFORM CLAIM COMMITTEE 08/05

PICA

1. MEDICARE (Medicare #) MEDICAID (Medicaid #) TRICARE CHAMPUS (Sponsor's SSN) CHAMPVA (Member ID#) GROUP HEALTH PLAN (SSN or ID) FECA BLK LUNG (SSN) OTHER (ID)

1a. INSURED'S I.D. NUMBER (For Program in Item 1)

2. PATIENT'S NAME (Last Name, First Name, Middle Initial)

3. PATIENT'S BIRTH DATE MM DD YY SEX M F

4. INSURED'S NAME (Last Name, First Name, Middle Initial)

5. PATIENT'S ADDRESS (No., Street)

6. PATIENT RELATIONSHIP TO INSURED Self Spouse Child Other

7. INSURED'S ADDRESS (No., Street)

CITY STATE

8. PATIENT STATUS Single Married Other Employed Full-Time Student Part-Time Student

CITY STATE

ZIP CODE TELEPHONE (Include Area Code)

ZIP CODE TELEPHONE (Include Area Code)

9. OTHER INSURED'S NAME (Last Name, First Name, Middle Initial)

10. IS PATIENT'S CONDITION RELATED TO:

11. INSURED'S POLICY GROUP OR FECA NUMBER

a. OTHER INSURED'S POLICY OR GROUP NUMBER

a. EMPLOYMENT? (Current or Previous) YES NO

a. INSURED'S DATE OF BIRTH MM DD YY SEX M F

b. OTHER INSURED'S DATE OF BIRTH MM DD YY SEX M F

b. AUTO ACCIDENT? YES NO PLACE (State)

b. EMPLOYER'S NAME OR SCHOOL NAME

c. EMPLOYER'S NAME OR SCHOOL NAME

c. OTHER ACCIDENT? YES NO

c. INSURANCE PLAN NAME OR PROGRAM NAME

d. INSURANCE PLAN NAME OR PROGRAM NAME

10d. RESERVED FOR LOCAL USE

d. IS THERE ANOTHER HEALTH BENEFIT PLAN? YES NO If yes, return to and complete item 9 a-d.

READ BACK OF FORM BEFORE COMPLETING & SIGNING THIS FORM.

12. PATIENT'S OR AUTHORIZED PERSON'S SIGNATURE I authorize the release of any medical or other information necessary to process this claim. I also request payment of government benefits either to myself or to the party who accepts assignment below.

SIGNED DATE

13. INSURED'S OR AUTHORIZED PERSON'S SIGNATURE I authorize payment of medical benefits to the undersigned physician or supplier for services described below.

SIGNED

14. DATE OF CURRENT: MM DD YY ILLNESS (First symptom) OR INJURY (Accident) OR PREGNANCY(LMP)

15. IF PATIENT HAS HAD SAME OR SIMILAR ILLNESS. GIVE FIRST DATE MM DD YY

16. DATES PATIENT UNABLE TO WORK IN CURRENT OCCUPATION MM DD YY FROM MM DD YY TO

17. NAME OF REFERRING PROVIDER OR OTHER SOURCE

17a.

17b. NPI

18. HOSPITALIZATION DATES RELATED TO CURRENT SERVICES MM DD YY FROM MM DD YY TO

19. RESERVED FOR LOCAL USE

20. OUTSIDE LAB? YES NO $ CHARGES

21. DIAGNOSIS OR NATURE OF ILLNESS OR INJURY (Relate Items 1, 2, 3 or 4 to Item 24E by Line)

1. 3.

2. 4.

22. MEDICAID RESUBMISSION CODE ORIGINAL REF. NO.

23. PRIOR AUTHORIZATION NUMBER

24. A. DATE(S) OF SERVICE From MM DD YY To MM DD YY	B. PLACE OF SERVICE	C. EMG	D. PROCEDURES, SERVICES, OR SUPPLIES (Explain Unusual Circumstances) CPT/HCPCS	MODIFIER	E. DIAGNOSIS POINTER	F. $ CHARGES	G. DAYS OR UNITS	H. EPSDT Family Plan	I. ID. QUAL.	J. RENDERING PROVIDER ID. #
1									NPI	
2									NPI	
3									NPI	
4									NPI	
5									NPI	
6									NPI	

25. FEDERAL TAX I.D. NUMBER SSN EIN

26. PATIENT'S ACCOUNT NO.

27. ACCEPT ASSIGNMENT? (For govt. claims, see back) YES NO

28. TOTAL CHARGE $

29. AMOUNT PAID $

30. BALANCE DUE $

31. SIGNATURE OF PHYSICIAN OR SUPPLIER INCLUDING DEGREES OR CREDENTIALS (I certify that the statements on the reverse apply to this bill and are made a part thereof.)

SIGNED DATE

32. SERVICE FACILITY LOCATION INFORMATION

a. NPI b.

33. BILLING PROVIDER INFO & PH # ( )

a. NPI b.

NUCC Instruction Manual available at: www.nucc.org

PLEASE PRINT OR TYPE

APPROVED OMB-0938-0999 FORM CMS-1500 (08-05)

CARRIER

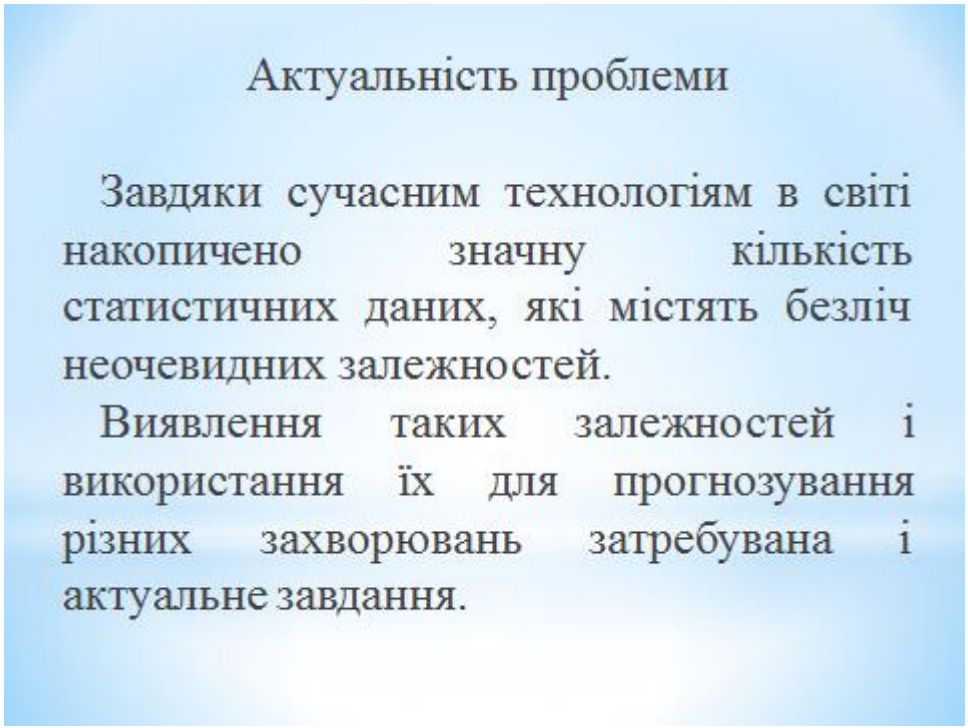
PATIENT AND INSURED INFORMATION

PHYSICIAN OR SUPPLIER INFORMATION

SAMPLE

Рисунок. А1 – форма CMS-1500

ДОДАТОК Б.
Електронні плакати



Постановка задачі

Метою атестаційної роботи є дослідження методів аналізу медичних даних. Завдання включає в себе:

- дослідити етапи і завдання Data Mining;
- вибрати необхідну стратегію аналізу;
- вибрати алгоритми для аналізу;
- безпосередньо реалізувати досліджувані алгоритми або застосувати готові програмні рішення;
- провести порівняльний аналіз досліджуваних алгоритмів.

Опис тестового матеріалу

У дослідженні основним джерелом даних є форми CMS-1500, в яких прибрані всі персональні дані про пацієнтів, крім дати народження. Таким чином, ми маємо можливість працювати з реальними даними, але при цьому зберігається повна конфіденційність.

Інтелектуальний аналіз даних

Інтелектуальний аналіз даних (Data Mining) - це процес виявлення в сирих даних раніше невідомих, нетривіальних, практично корисних і доступних інтерпретації знань, необхідних для прийняття рішень в різних сферах людської діяльності.

Стадії Data Mining

Процес Data Mining складається з трьох стадій, але третя несе суто коригувальний характер:

- вільний пошук (Discovery);
- прогностичне моделювання (Predictive Modeling);
- аналіз винятків (Forensic Analysis).

Кластерний аналіз

Завдання кластеризації полягає в пошуку незалежних груп (кластерів) і їх характеристик у всьому безлічі аналізованих даних. Вирішення цього завдання допомагає нам краще зрозуміти дані.

K-means та PAM

K-середніх починає з деякого початкового розбиття об'єктів і уточнює його, оптимізуючи цільову функцію – середньоквадратичне помилку кластеризації як середньоквадратичне відстань між об'єктами і центрами їх кластерів:

$$e(D,C) = \sum_{j=1}^{|D|} \sum_{i:d_i \in C_j} \|\overrightarrow{d_i} - \overrightarrow{\mu_j}\|^2 \ ,$$

Алгоритм PAM аналогічний алгоритму K-Means, тільки при роботі алгоритму перерозподіляються об'єкти щодо медіани кластера, а не його центру

Fuzzy C-Means

Нечіткий алгоритм Fuzzy C-Means був запропонований в 1973 році як рішення проблеми м'якою кластеризації, тобто привласнення кожному об'єкту більш ніж одного кластеру. Як і його чіткий варіант - алгоритм k-середніх - даний алгоритм, починаючи з деякого початкового розбиття даних, итеративно мінімізує цільову функцію, як

$$e_m(D,C) = \sum_{j=1}^{|D|} \sum_{i:\vec{d}_i \in C_j}^{|C|} u_{ij}^m \left\| \vec{d}_i - \vec{\mu}_j \right\|^2 ,$$

CLOPE

До переваг відносяться високі масштабованість і швидкість роботи і якість кластеризації, що досягається використанням глобального критерію оптимізації на основі максимізації градієнта висоти гістограми кластера. Формула глобального критерію:

$$\text{Profit}(C) = \frac{\sum_{i=1}^{k} G(C_i) \times |C_i|}{\sum_{i=1}^{k} C_i} = \frac{\sum_{i=1}^{k} \frac{S(C_i)}{W(C_i)^r} \times |C_i|}{\sum_{i=1}^{k} C_i} ,$$

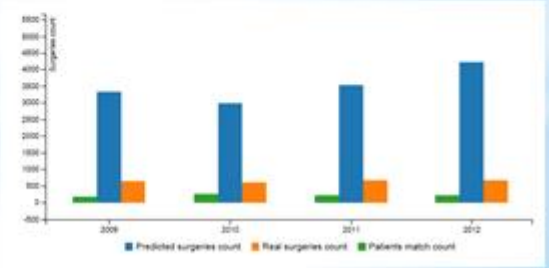
DBSCAN

Алгоритм DBSCAN (Density Based Spatial Clustering of Applications with Noise - плотностной алгоритм для кластеризації просторових даних з присутністю шуму) - алгоритм з автоматичним вибором кількості кластерів. Він заснований на припущенні про те, що щільність точок всередині кластерів більше, ніж за кластерів. цей алгоритм дозволяє знаходити кластери довільної форми.

Результати роботи алгоритмів

Далі на графіках зображені результати прогнозування, тобто скільки було спрогнозовано, скільки було реально операцій і скільки з них збіглося

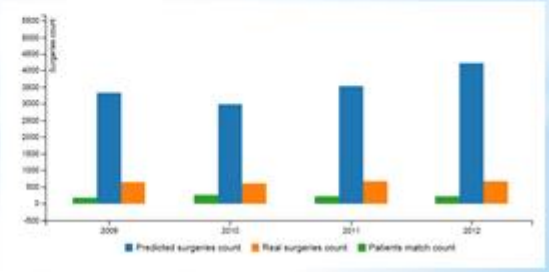
K-Means



Результати роботи алгоритмів

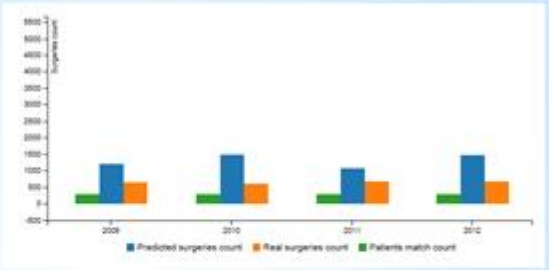
Далі на графіках зображені результати прогнозування, тобто скільки було спрогнозовано, скільки було реально операцій і скільки з них збіглося

K-Means

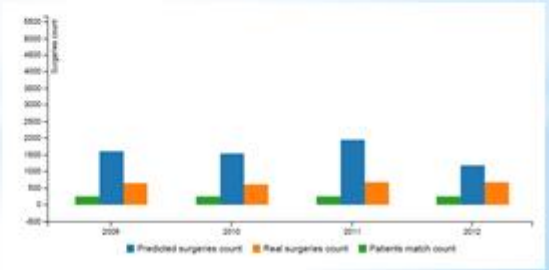


Результати роботи алгоритмів

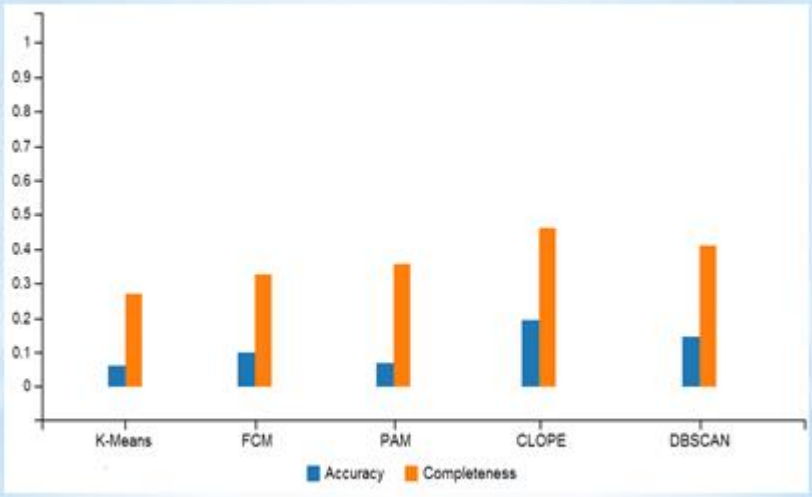
CLOPE



DBSCAN



Точність алгоритмів



Аналіз результатів

CLOPE має найбільші показники, а значить в нашому дослідженні він виявився найбільш ефективним.

Але взагалі параметри точності занадто низькі для якого практичного застосування. Для підвищення точності необхідно більше параметрів для порівняння і більш точні експертні оцінки. Також для поліпшення результатів можна застосувати для кластерного аналізу метрику з коефіцієнтами, але для цього також необхідні експертні дані.

Загалом, для отримання більш точних результатів кластерного аналізу недостатньо для таких складних даних. Для подальшого дослідження необхідно використовувати більш складні методики такі, як нейронна мережа або лінійна регресія.

Висновки

В результаті виконання атестаційної роботи були досліджені основні алгоритми кластеризації в рамках завдання кластеризації технології інтелектуального аналізу даних. В ході дослідження виявлено що даний підхід до прогнозування має низьку ефективність у зв'язку з недостатньою кількістю експертних даних.

ДЯКУЮ ЗА УВАГУ!