МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ОДЕСЬКИЙ ДЕРЖАВНИЙ ЕКОЛОГІЧНИЙ УНІВЕРСИТЕТ

Факультет комп'ютерних наук, управління та адміністрування
Кафедра інформаційних технологій

Кваліфікаційна робота бакалавра

на тему: Розробка алгоритмів діагностування на основі нечіткої кластеризації даних

Виконав студент групи КН-20
спеціальності 122 Комп'ютерні науки
Бояринцев Олександр
Олександрович

Керівник д.т.н., проф.,
Мещеряков В.І.

Консультант

Рецензент д.т.н., проф.,
Казакова Н.Ф.

Одеса 2023

ЗМІСТ

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАКИ

ПМР – прийняття медичних рішень.

СППР – систем підтримки ухвалення медичних рішень.

ППМР – підтримки прийняття медичних рішень.

ПрО – предметна область.

ЛПМР – особа, яка приймає медичне рішення.

АМК – автоматизовані медичні комплекси.

АМС – автоматизовані медичні системи.

ТНМ – Теорія нечітких множин.

УМР – Управлінські медичним рішенням.

МБД – Медичні бази даних .

НАС – нечіткі аналітичні системи.

КП – класифікаційні ознаки.

FPC – Нечіткий коефіцієнт розподілу.

ВСТУП

Сучасні медичні установи функціонують в епоху технологічних, що динамічно розвиваються, і інформаційних процесів, що характеризуються стрімкими та широкомасштабними змінами зовнішнього середовища, економічних та соціальних відносин. Продуктивним інструментом дослідження проблем у галузі управління та створення інформаційних систем медичного призначення по даному напрямку є методи моделювання.

У сформованих умовах аналіз діяльності медичних організацій, незалежно від сфери діяльності та форми власності, є науковою основою прийняття управлінські медичні рішення.

Отже, виникає проблема автоматизації аналізу діяльності медичних установ, що дозволяє точно оцінювати за допомогою нових сучасних методів дослідження невизначеність медичної ситуації. Вирішення цієї проблеми може бути знайдено у розвитку методології проектування інтелектуальних систем підтримки ухвалення медичних рішень (СППМР).

Подання медико-технологічних процесів у вигляді моделей є основою проектування СППМР та сприяє підвищенню ефективності як прийняття медичних рішень (ПМР), а також управління медичним установою в цілому.

Загалом проблеми підтримки прийняття медичних рішень (ППМР) в умовах невизначеності часто бувають слабоструктурованими або погано формалізованими, у зв'язку з чим застосування традиційних методів моделювання складних систем є малоефективним, що у свою чергу веде до використання спеціально розроблених механізмів ППМР на базі нечітких множин разом із методами теорії алгебри логіки, семантичних мереж та теорії когнітивного аналізу .

У класичних СППМР у процесі формування груп виникає проблема вибору так званих групувальних ознак, кількість яких залежить від деякої метрики, кількість цих параметрів невпинно зростає. Залежно від цілей питання про вибір масштабів та метрики має різний зміст.

Таким чином, оптимальним механізмом для автоматизованого рішення є ефективна обробка статистичної інформації та комплексний аналіз отриманих результатів засобами інтелектуального аналізу із застосуванням технологій нечіткої кластеризації та нечітких множин.

При цьому етапи підтримки прийняття рішень та вибір результатів моделювання визначаються поточним станом системи, точніше, станом основи знань, що входить до її складу. До подібних методів, насамперед, можна віднести методи, що базуються на застосуванні нечіткої кластеризації з використанням нечіткого відношення рівнозначності, побудови предметної області з використанням семантичної мережі із застосуванням нечітко множинного підходу, математичного апарату нечітких когнітивних карт.

1 ОСНОВНІ ПРИНЦИПИ ПОБУДУВАННЯ СИСТЕМ ПІДТРИМКИ ПРИЙНЯТТЯ МЕДИЧНИХ РІШЕНЬ В УМОВАХ НЕВИЗНАЧЕННОСТІ

1.1 Невизначеність у медичній сфері та завдання підтримки прийняття медичних рішень

В даний час медико-технологічні процеси, що супроводжуються інформаційними потоками, є активними силами, що зв'язують систему та об'єкт управління між собою, а також із зовнішнім середовищем. Функціонування таких систем можливе за рахунок використання різних рішень, які мають технологічни характер.

Медико-технологічний процес – це сукупність дій та (або) взаємодій медичного, технічного, адміністративного персоналу медичного закладу та пацієнта, необхідних для реалізації заходів як лікувально-діагностичного, так і організаційно-управлінського характеру, здійснюваних у певній послідовності, взаємозв'язку та тимчасових режими з метою ефективного надання медичної допомоги.

Наприклад, для реалізації дій організаційно-управлінського характеру можуть використовуватися автоматизовані системи управління, а лікувально-діагностичного автоматизовані медичні системи (АМС) та (або) автоматизовані медичні комплекси (АМК).

Взаємозв'язок довкілля та організації визначається простором показників діяльності організації. Безліч показників діяльності медичної організації, що складають зазначене простір, що характеризуються значними потужностями.

Вони формують функції цілей діяльності господарюючого суб'єкта (траєкторні, робітники та ситуаційні) [1]. Характер зміни цих показників діяльності часто непередбачуваний.

При цьому основними завданнями управління соціально-економічною діяльністю є оцінка та діагностика діяльності. організації, визначення раціональних траєкторій її функціонування.

Надзвичайно глибокі та швидкі зміни економічних та соціальних явищ ускладнюють процес прогнозування їх розвитку та розвиток об'єктів усередині медичної організації в майбутньому на базі формальної логіки та традиційних математичних методів, що базуються на точних даних, при цьому застосування систем підтримки ухвалення рішень (СППР) може стати ефективним інструментом для вирішення перелічених завдань.

Кожна СППР має суто індивідуальний характер, оскільки відрізняється особливостями процедури прийняття рішень та конкретним змістом розв'язуваної управлінської проблеми у тій чи іншій області.

Усі процедури, що стосуються процесу прийняття рішень, укрупнено можна розділити на регулярні та періодичні. Як регулярні процедури ухвалення рішень можуть виступати планування діяльності медичної організації, контроль виконання, оперативне управління, ведення електронної історії хвороби (ЕІБ) тощо. Послідовність функціонування та склад СППР у цьому випадку закріплюються як нормативні методики, що застосовуються, як правило, як формальні моделі та методи із незначним використанням діалогових процедур. Однак велику групу складають завдання, які не мають повних аналогів у минулому, іншими словами, це процедури, зумовлені періодично виникають проблемними ситуаціями з досить високим ступенем невизначеності. У цьому випадку необхідно розробляти СППР індивідуально для кожної медичної предметної галузі. До складу таких систем включаються переважно експертні та логіко-евристичні методи та моделі, при цьому значна увага приділяється діалоговим процедурам.

У медичній сфері ППМР – систематизований процес, що пов'язано з великою відповідальністю ЛПМР. У медичному закладі всі рішення можна класифікувати як організаційно-управлінські та лікувально-діагностичні, але й ті

та інші обумовлені обійманою посадою ЛПМР та спрямовані на ефективне управління медичною установою в цілому (рис. 1.1).

Рисунок. 1.1 - Два класи рішень у медичній установі

Розглянемо підходи до підтримки прийняття медичних рішень, що стосуються обох класів , а саме: інформаційна підтримка при постановці первинного діагнозу, вибір варіанта перебігу хвороби, вибір схеми лікування, його коригування та оцінка СЗП, а також управління медичним матеріальним потоком - персоніфікований розподіл лікарських засобів коштів. У перерахованих задачах невизначеність виникає з низки причин, в тому числі через такі фактори:

- недостатність знань – відсутність достатніх знань про стан здоров'я пацієнта, про особливості організму пацієнта, про конкретне захворювання пацієнта та ін;
- суперечливості наявної інформації – суперечливості різних медичних аналізів; суперечливості медичних аналізів та анамнезу пацієнта; неможливості застосування певної схеми лікування супутніх захворювань чи протипоказань та ін;

- відсутності можливості залучення компетентних лікарів-експертів –неможливості здійснення прийому лікарем високого рівня, особливо в медичних установах, що у віддалених населених пунктах; відсутності лікаря через хворобу або відпустку; складних та рідкісних випадків захворювань та ін.
- обмеженості тимчасових ресурсів - не завжди є можливість одержання результатів медичних аналізів у короткий проміжок часу; регламентованості за часом на прийом одного пацієнта; потреби в негайне прийняття медичного рішення в деяких питаннях;необхідності термінових хірургічних втручань та ін;
- неповноти чи неточності інформації про стан пацієнта – помилки чи неточності у медичних аналізах; інтервальні значення деяких даних; інформація, отримана від пацієнта (навмисне або ненавмисне приховання фактів); відсутність результатів медичних аналізів на даний момент прийняття рішень та ін.

Медичні рішення кваліфікують як запрограмовані, тобто є результатом виконання конкретної послідовності дій, і незапрограмовані, коли умови, в яких приймаються рішення, певною мірою нові, пов'язані з невідомими факторами або внутрішньо не структуровані.

Будь-яке медичне рішення, зокрема організаційне, - це компроміс, а процес ПМР – процес психологічний. Рішення, які приймаються особою, яка приймає медичне рішення, може бути нелогічними та змінюватися від спонтанних до високологічних. Прийняте медичного рішення тільки на основі розуміння ЛПМР, що воно правильне, при цьому не розглядаються всі наявні можливі варіанти, враховуються їх переваги та недоліки, а також не проводиться аналіз проблемної ситуації, є інтуїтивним рішенням.

Вибір, зумовлений накопиченим досвідом чи знаннями, характеризує рішення, що ґрунтуються на судженнях. ЛПМР у проблемній ситуації використовує наявне знання про те, що відбувалося в схожих ситуаціях раніше, прогнозування результатів альтернативних рішень Цей підхід дозволяє ефективно

витрачати тимчасові, трудові та фінансові ресурси, оскільки ППМР здійснюється порівняно швидко і не вимагає збору та аналізу додаткової інформації, що характеризує його з позитивного боку. Однак за такого підходу рішення приймаються, так би мовити, на базі здорового сенсу, що в реальних умовах зустрічається дуже рідко, оскільки, з одної сторони, дані, на основі яких приймається медичне рішення, можуть бути спотворені з різних причин, а з іншого - в абсолютно нових або унікальних ситуаціях приймати правильні медичні рішення не дозволяють міркування, оскільки необхідним досвідом для обґрунтованого вибору ЛПМР не має, а судження, з урахуванням наявного досвіду, орієнтують ЛПМР на рішення, знайоме з попередніх ситуацій, при цьому упускаються нові альтернативи.

Рішення, ухвалене в результаті аналітичного процесу, що характеризується своєю об'єктивністю і не залежить від наявного досвіду, називається раціональним. Прийняття такого рішення полягає у послідовне виконання діагностики проблеми, формулювання обмежень та критеріїв прийняття медичного рішення, визначення альтернатив, оцінки альтернатив, вибору альтернативи, що дозволяє говорити про ньому як про структурований процес . Якщо ЛПМР з точністю може визначити результат будь-якого альтернативного рішення, можливого в ситуації, що розглядається, – це рішення, ухвалене за умов визначеності. Однак у реальних умовах рішень (як організаційних, так і медичних), що приймаються в таких умовах досить мало. Як певні можна розглядати частини більш складні рішення. При прийнятті рішень рівень визначеності залежить від предметної галузі. За наявності обмеження кількості альтернатив та зниження ризику рівень підвищується.

Якщо ймовірність кожного можливого результату рішення можна визначити в проміжку від 0 до 1 $\left(p_{\text{рез}} = \{0,1\}\right)$, то це рішення, яке приймається в умовах ризику, при цьому одиниці повинна дорівнювати сума ймовірностей всіх альтернатив ($\sum_{i=1}^{N_{\text{альт}}} p_i = 1$, де $N_{\text{альт}}$ - кількість альтернатив) .У цьому ви-

падку слід говорити про об'єктивності визначення ймовірності, що знаходиться за допомогою математичних методів або з використанням аналізу накопичених статистичних даних [2].

Якщо надійде достатня кількість релевантної інформації, то ймовірність може бути об'єктивно визначена і тоді прогноз виявиться статистично достовірним. Інакше використовуються судження про можливість виконання альтернатив із суб'єктивною (передбачуваною) ймовірністю.

Рішення (медичні, управлінські чи організаційні) приймаються за умов невизначеності у ситуаціях, у яких оцінити ймовірність очікуваних результатів неможлива. Це відбувається при прийнятті рішень, що вимагають обліку нових та складних факторів, за якими отримання релевантної інформації у достатньому обсязі для об'єктивного визначення ймовірності неможливо. Також може бути ситуація, яка не підкоряється відомим закономірностям, або рішення приймається у швидкозмінних умовах. У всіх випадках неможливо з достатньою мірою достовірності передбачити ймовірність певного результату ситуації ($p_{\text{посл}}$).Є кілька можливостей для позбавлення від невизначеності.

Одна з них - отриманняпроблемноїситуації додаткової релевантної інформації(використання накопичених досвіду або статистики, судження або (інтуїції), тим самим складність проблеми зменшується.

У результаті є можливість отримати як альтернативу передбачувану (суб'єктивну) ймовірність ($p_{\text{пред}}$). Іншим способом позбавлення від невизначеності, коли немає ресурсів для збору додаткової інформації, є дії ЛПМР, вчинені у повній відповідності до накопиченого досвіду.

У медичній предметній галузі виділяють невизначеність, пов'язану: з неповнотою наявних знань про проблемну ситуацію, за якою приймається медичне рішення (відсутність медичних аналізів, неповні дані анамнезу та ін); непередбачуваністю реакції зовнішнього середовища (або реакції організму пацієнта) на дії ЛПМР, що виробляються; неточністю розуміння мети ЛПМР (вибір схеми лікування). При цьому можуть виникати такі утруднення, як: неможливість перетворення завдань з невизначеністю до формалізованим, що

відповідно призводить до необхідності застосування поправки на суб'єктивність експерта; зростання кількості факторів, що, як слідство призводить до скорочення часу на аналіз.

Підтримка прийняття медичних рішень полягає в наступному:

- надання допомоги ЛПМР при здійсненні аналізу об'єктивної складової проблеми (складна медична ситуація);
- виявлення переваг ЛПМР (орієнтація на конкретного пацієнта
- Врахування невизначеності в оцінках ЛПМР (оцінка рекомендації, запропонованою системою);
- створення набору можливих рішень проблемної медичної ситуації;
- оцінки можливих рішень, що відповідають перевагам ЛПМР, та обмежень, що накладаються зовнішніми факторами (оцінка рекомендацій щодо релевантності для конкретної ситуації);
- аналізі наслідків прийнятих медичних рішень;
- виборі найкращого, з погляду ЛПМР, рішення (ухвалення рішення здійснює лікар-користувач на основі даних, наданих системою).

На медичне управлінське рішення може впливати безліч факторів, таких як: соціальні, політичні та економічні аспекти (Ситуація на ділянці, в районі або області – територіальна ситуація), компетентність ЛПМР (кваліфікація лікаря, досвід роботи), забезпеченість необхідною інформацією (наявність накопиченої статистики чи анамнезу), обґрунтованість поставлених цілей, завдань та ефективність обраного рішення (порівняння зі стандартними схемами лікування).

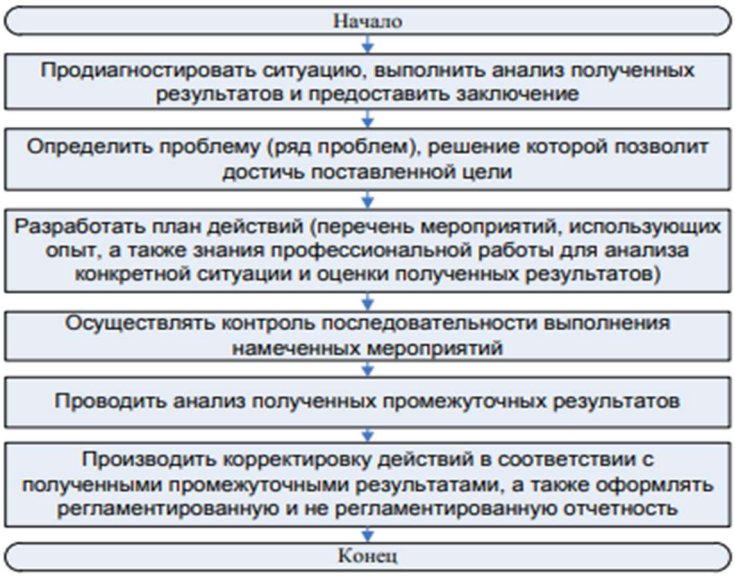

Рисунок 1.2 – Алгоритм дій ЛПМР

Особа, яка приймає медичне рішення – особа (медичний працівник,лікар) або група осіб (колектив відділення, консиліум лікарів), у межах власних повноважень, які мають право з питань, що стосуються їх сфери занять (діяльності медичної організації чи медичного процесу), приймати рішення, нести відповідальність за отримані результати та стежити за ходом виконання набору дій. ЛПМР самостійно для цілей реалізації управлінського медичного рішення може приймати рішення та розпоряджатися ресурсами медичного закладу. Для прийняття управлінського медичного рішення ЛПМР має діяти за алгоритму, представленому на рис. 1.2.

1.2 Управлінські медичні рішення та методи отримання медичної інформації в умовах невизначеності

Швидкість удосконалення цивілізації, процесів, способів та технологій обміну та управління інформацією, заснованих на тісному взаємодії із зовнішнім середовищем, – це ті причини, які нині призвели до появи нових трудно-

щів, що стосуються прийняття управлінських медичних рішень. На жаль, фактичні результати прийнятих медичних рішень рідко збігаються із запланованими. Разом з критеріями прийняття медичних рішень, які існували раніше, виникли та нові: здоров'я нації, вплив на довкілля, становище на внутрішньому та світовому ринках, індекс здоров'я, корпоративний пристрій організації, природний приріст населення та багато інших. До управлінським медичним рішенням (УМР) висуваються вимоги, якими система (або системні компоненти) повинна володіти для задоволення стандартів, специфікацій тощо.

Розробка управлінського медичного рішення в умовах визначеності має суттєву відмінність від розробки управлінського медичного рішення в умовах невизначеності і полягає в тому, що в першому випадку ймовірності наслідків альтернатив однозначно визначено, а в другий випадок безліч значень (можливих) результатів (з їх ймовірностями) відповідає кожному варіанту. Фахівці розглядають управлінське медичне рішення як перелік процедур та дій, орієнтованих на вирішення проблемної ситуації у формі нотацій, наказів, специфікацій в усному чи письмовому вигляді [3].

Будь-яка проблемна ситуація, пов'язана з прийняттям медичних рішень, характеризується кількома варіантами дій, у тому числі потрібно вибрати найкращий. При цьому необхідна формалізація постановки задачі розроблення управлінських медичних рішень.

Для зниження рівня невизначеності потрібна інформація про те, яким має бути результат її вирішення, а також де здійснюватиметься процес розв'язання (з метою утворення єдиного інформаційного простору).

Однією із проблем розробки управлінського медичного рішення є необхідність надання мети (цілям) кількісних (обсяг лікарського засобу, частота його застосування) та якісних (покращення стану пацієнта, погіршення епідеміологічної ситуації) характеристик.

Причому останні – найбільш пріоритетні, оскільки сприяють формалізації завдання вибору. Обмеженість ресурсів - інша проблема, що зводиться не

лише до необхідності їх розподілу, а й до здійснення вибору способів використання. Досягнення однієї й тієї ж мети може здійснюватися різними альтернативними способами, причому варіант, що дозволив досягти максимальної ефективності, є оптимальним (за певним критерію або групі критеріїв), а сам процес пошуку такого рішення називається оптимізацією. Близькі (за ефективністю) до оптимальних варіантів дії називають прийнятними. Процес пошуку найкращого управлінського медичного рішення доцільно розбити на два етапи: відбір раціонального варіанта з множини варіантів (перший етап) та вибір оптимального варіанта з представлених раціональних варіантів (другий етап).

Найкраще медичне рішення – компроміс тим часом, як у стислі терміни досягти заданої мети (тобто тимчасова цільова функція) і при цьому максимально покращити стан здоров'я пацієнта (значення медичних показників повинні відповідати нормам, регулювати кількість, дозування та частоту застосування лікарських засобів, проведених процедур тощо) і завдати найменшої шкоди організму пацієнта з погляду інтенсивності застосовуваної терапії (побічні реакції, супутні захворювання), а також з мінімальними витратами на конкретного пацієнта зі сторони медичного закладу (раціональне застосування лікарських засобів, розподіл навантаження медичного персоналу та ефективне застосування медичного устаткування).

Все це повинно враховуватися при складанні схем лікування пацієнта та при формуванні ранжованих наборів рекомендацій для ситуацій, що склалися. Під вибором «оптимального» варіанта розуміється рішення, яке буде найбільше відповідати поняттю «найкраще медичне рішення». Якщо через F позначити функцію оцінки медичного рішення [4]:

$$F(t,h,b,z), \tag{1.1}$$

де t - час, що витрачається на вирішення проблемної ситуації (досягнення мети), h – стан здоров'я пацієнта; b – побічні реакції організму пацієнта; z – загальні витрати медичного закладу на пацієнта, то:

$$\text{при } t \to min, h \to max, b \to min, z \to \min: F(t, h, b, z) \to max, \qquad (1.2)$$

таким чином, рішення, яке має функцію оцінки медичного рішення буде мати максимальне значення можна вважати найкращим медичним рішенням.

На практиці прийняте медичне рішення може зачіпати інтереси, зокрема об'єктивно важливі цілі та завдання, кількох підрозділів (відділень) медичного закладу, що може спричинити виникнення конфліктів переваг та цілей.

На етапі підготовки управлінського медичного рішення можливий розбаланс цілей підрозділів (оскільки кожен із зацікавлених підрозділів може орієнтуватися рішення власної проблеми). Рішення зазначеної проблеми можливо через роботу в групі з визначенням пріоритетної мети організації та відповідно з підпорядкуванням їй підцілей підрозділу. На етапі реалізації управлінського медичного рішення можуть виникнути конфлікти, пов'язані з розподілом прав, обов'язків, відповідальності та ресурсів.

У зв'язку з цим необхідно здійснити конкретизацію термінів та повноважень. Остаточне управлінське медичне рішення може бути прийнято однією людиною чи групою осіб з урахуванням особистих чи групових переваг відповідно. У зв'язку з тим, що прийняті управлінські медичні рішення завжди орієнтовані на майбутнє, у момент ухвалення такого рішення ЛПМР часто не має абсолютних знань про можливий розвиток ситуації. То є елемент невизначеності та ризику носить значний характер у момент ухвалення управлінського медичного рішення . Передбачення ризику, прагнення знизити його до нижчого рівня, а чи не уникнення його - це найголовніше правило медичної діяльності, яке виконується при грамотному управлінні ризиками (воно включає в себе завчасне передбачення, своєчасне виявлення невизначеностей та їх наслідків при розробці та реалізації управлінських медичних рішень).

Для аналізу ризику необхідна перш за все оперативна, актуальна, достовірна та адекватна інформація (у медичних системах може йтися про життя пацієнта). Для того щоб оцінити виявлені ризики та прийняти відповідне управлінське медичне рішення щодо їх зниження, потрібно зібрати вихідну інформацію про об'єкт – носій ризику (для медичних систем це анамнез або статистичні дані щодо схожих ситуацій). Для цього необхідно здійснити відбір інформації про структуру об'єкта та виявити інциденти чи небезпеки. Найважливішою характеристикою сучасного розвитку суспільства є той факт, що в інформаційну складову роблять свій внесок всі професійні групи медичних працівників – від медсестер та лаборантів до керуючих та лікарів. Порушення такого ланцюга призводить до втрати інформації та до погіршення результатів роботи загалом.

Медичні інформаційні системи, що застосовуються на сучасному етапі розвитку інформаційного товариства , допомагають ЛПМР приймати оперативні, адекватні та грамотні управлінські медичні рішення та припускають побудову наступного: медичних баз даних (МБД), медичних експертних систем, систем підтримки прийняття медичних рішень. У цих системах керуючі медичні рішення мають рекомендаційний характер, як правило, з кількісною оцінкою відповідності проблемної ситуації, а слідувати чи не дотримуватися цих рекомендацій -вибір залишається за медичним користувачем.

Алгоритм формування управлінського медичного рішення подано на рис. 1.3. Побудова таких систем тісно пов'язана з досягненнями ІІ, крім того, при автоматизації медико-технологічних процесів досить часто виникають задачі, пов'язані з невизначеністю або неповнотою вихідних даних, вирішення цих завдань можливе застосування нечітких об'єктів.

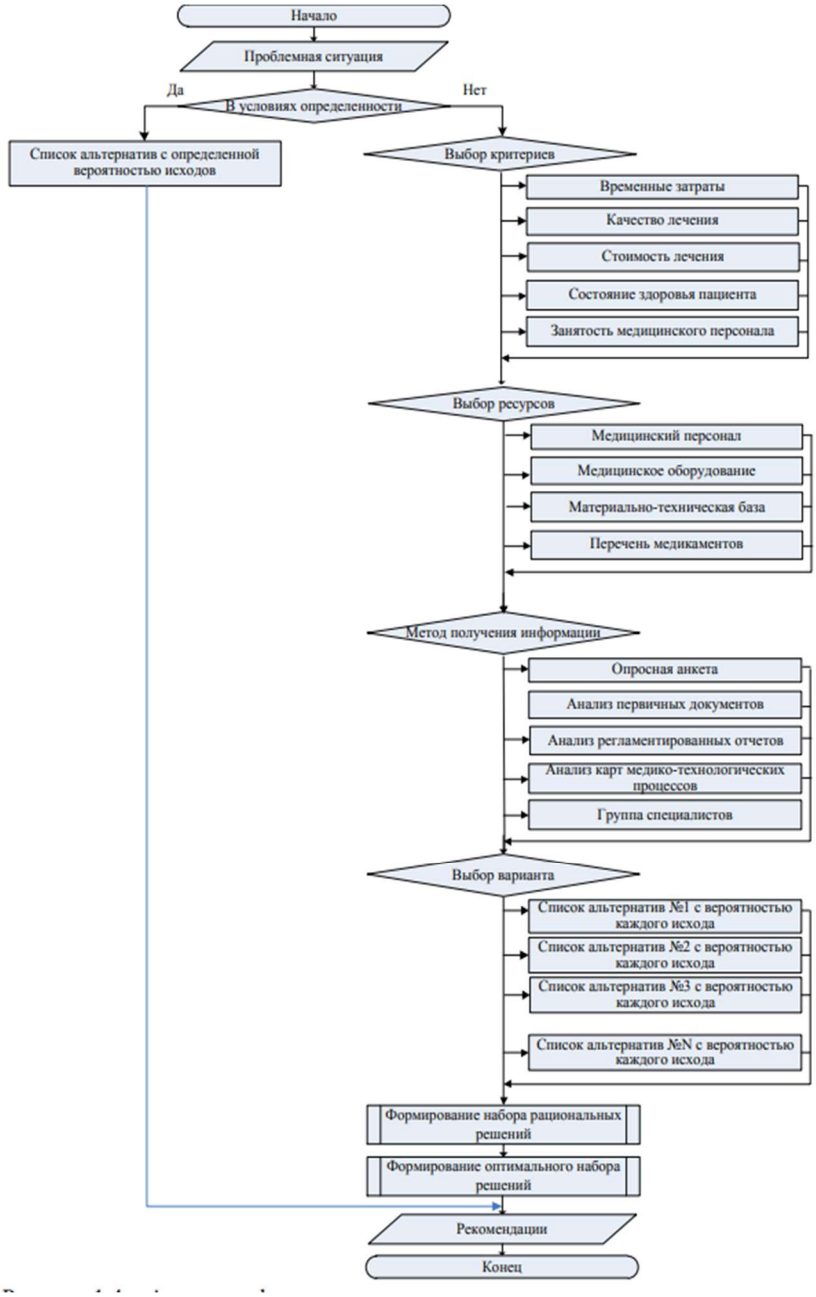

Рисунок 1.3 – Алгоритм формування управлінського медичного рішення

1.3 Подання знань на основі теорії нечітких множин у медичних предметних галузях

Неповноту інформації у системах медичного призначення характеризують трьома аспектами: нечіткістю, неточністю, невизначеністю.

У медицині поняття «невизначеність» нерозривно пов'язане з поняттями «медична інформація» та «можливість вибору», оскільки будь-яка невизначеність передбачає можливість альтернативи, а наявність будь-якої інформації зменшує невизначеність, скорочує можливість вибору. У разі повноти медичної інформації альтернативи немає.

Концепція невизначеності використовується при описі неповноти інформації, застосовується в теорії штучного інтелекту і служить для позначення ступеня істинності утвердження.

Відсутність точних меж множин об'єктів характеризує нечіткість даних, для формалізації яких Л. Заде запропонував застосовувати нечіткі множини [1], які потім склали основу нечіткої логіки [5]. Узагальнення поняття власності – основа запропонованого підходу до формалізації поняття нечіткості даних Теоретично нечітких множин функцією власності є характеристична функція. Висновок за умов невизначеності, класифікація даних, аналіз даних, проблеми прийняття медичних рішень – ось неповний перелік областей, в яких нечіткі множини знаходять широке застосування [3],кожному з яких використовується відповідний вид семантики ступенів приналежності в термінах невизначеності, подібності та переваги.

Представивши вираз $\mu_A(x)$ як ступінь належності нечіткому Безлічі A елемента x, заданого на множині X, можна $\mu_A(x)$ назвати ступенем близькості x до прототипу A або подібністю при інтерпретації приладдя:

$$A = \{x, \mu_A(x)\}. \tag{1.3}$$

Це один із перших підходів, який характерний для таких предметних областей (ПрО), у яких ставляться завдання вибірки даних (завдання абстракції). Потім був запропонований наступний підхід, в якому безліч A представлялося як безліч найбільш або найменш переважних об'єктів. При такому підході $\mu_A(x)$ розглядається як ступінь переваги та представляє придатність вибору x

як значення змінної b. Використання цього підходу здійснюється у ПрО, де для аналізу рішень потрібна нечітка оптимізація. Завдання оптимізації має бути перенесене з класу завдань «чіткою» оптимізації до класу завдань нечіткої оптимізації, якщо хоча б один її елемент є нечітким (випадковий параметр «невизначеність знання про нього» не може бути описаний за допомогою теорії математичної статистики або теорії ймовірностей). Нечітка оптимізація - багатокритеріальна оптимізація при нечіткої інформації .

Пізніше було вирішено розглядати інтерпретацію власності як ступінь невизначеності, тоді $\mu_A(x)$ - це ступінь можливості того, що параметр b має значення x: « b - це A ». При такому підході, представляючи в як ступінь невизначеності $\mu_A(x)$, можна розглядати тільки епістемологічну інтерпретацію. При цьому фізичний зміст поняття більше підходить до терміна «перевага» або «придатність». Теорія нечітких множин отримала подальший розвиток [3], в якій функція приналежності представляється в як відношення впорядкування, пов'язаного з предикатом A, і записується наступним виразом:

$$x \geq_A x' \tag{1.4}$$

де вираз означає, що x більше відповідає A, ніж x' . Наприклад, в медичній сфері як x може виступати значення будь-якого з медичних показників (x'- інше значення цього ж параметра), що бере участь в аналізі проблемної ситуації, що склалася A. При такому підході нечітке відношення Q може розглядатися на X × X як тернальнеставлення (x^3) (набір бінарних відносин):

$$\{\geq_x, x \in X\}, \tag{1.5}$$

представлене як повне упорядкування. Тоді нерівність

$$\mu_Q(x, x') \geq \mu_Q(x, x''), \tag{1.6}$$

описує ситуацію, в якій вираз означає, що x'ближче до x ,ніж x'':

$$x' \geq_x x'', \tag{1.7}$$

Теорія нечітких множин (ТНМ) була створена для поганого моделювання певних процесів, для опису яких використовуються судження людини (лікаря-користувача), які мають якісний характер; внутрішнє протиріччя цієї теорії полягає в тому, що для цих процесів складно побудувати точні моделі, оскільки про них немає точної інформації. Незважаючи на це, при використання ТНМ необхідно точно задавати значення приналежності (числові) в інтервалі [0, 1] (тобто зазначені значення вимірюються за абсолютною шкалою), над якими згодом виконуються кількісні операції. Оскільки завдання значень приналежності має суб'єктивний характер (група лікарів-експертів), який завжди можна задати їх абсолютно. Семантичні мережі – найбільш потужна математична модель для уявлення знань про предметну область, один з найважливіших напрямів штучного інтелекту [6].

За допомогою семантичної мережі, що має вигляд орієнтованого графа,представлятиметься інформаційна модель медичної предметної галузі, вершини графа відповідають об'єктам предметної області, яке ребра (дуги) ставлять відносини з-поміж них (ступеня зв'язку). Спосіб подання якісних суджень лікаря-експерта про ступінь приналежності володіння об'єктами як якісних величин (значеннями абсолютної шкали) дозволяє здійснювати побудову нечітких моделей і, як наслідок, некоректно моделюваних процесів. Вирішення зазначеної проблеми лежить у моделюванні об'єктивних процесів медичної ПрО, при цьому необхідно проводити налаштування моделі при налагодженні шляхом експериментування на моделюваному процесі. Однак досить гостро стоїть проблема підвищення коректності нечітких моделей під час моделювання суб'єктивних процесів.

Існує кілька критеріїв класифікації нечіткої інформації, один з яких – область визначення нечітких множин. Відповідно до неї виділяють два типи [6].

Перший тип нечіткої інформації - це множини, які визначені на числовому множині X (інтервалі дійсних чисел) . В цьому випадку безліч X описується за допомогою числової шкали, а нечіткі множини – це нечіткі величини, подані на цій шкалі. Нечіткі числа і інтервали можна навести як приклад нечітких величин. Другий тип нечіткої інформації - нечіткі множини, задані на нечисловій шкалі (фактів, правил експертної системи, альтернатив та цілей на елементах бінарного відношення об'єктів між собою тощо), за такого підходу нечітка безліч $B = \{(b, B(b))\}$, де $(b, B(b))$ це безліч «нечітких об'єктів». Нечіткі множини зазначених типів в залежності від області застосування можуть суттєво різнитися між собою за способом інтерпретації, визначення та обробки.

У роботі [3] також розглянуті та інші типи нечітких множин, які відображають уявлення ЛПМР про деякої розпливчастої категорії. Наприклад, у [3] «високий» - поняття А визначено на об'єктивному вимірі числової шкали (числа – зростання людини) або визначено на безлічі об'єктів, які якісно описуються за допомогою таких категорій (безліч людей) - $\mu_A(x)$ - величина, що виражає ступінь сумісності значення x з поняттям А.

Наведено розгляд експертної нечіткої інформації,визначеної як формальна система нечітких висловлювань, базисних понять ТНМ, розроблено схему побудови нечітких моделей прийняття рішень на етапах проектування, які носять важкоформалізований характер, заснованих на правилах індуктивного та дедуктивного висновків . Формальні моделі для створення систем ухвалення рішень, що носять нечіткий характер, розглянуті як нечіткі орієнтовані та неорієнтовані гіперграфи. Авторами запропоновані методики для здійснення вирішення завдань у предметній галузі за наявності нечіткої експертної інформації, пов'язаної з вибором як аналогів проектованого об'єкта, і варіанти проектування. Авторами [3] представлений набір методів для проведення формалізації та обробки інформації, що носить нечіткий характер.

Метою роботи є знайдення способів полегшення прийняття медичних рішень та груповань даних в умовах нечіткості.

Задачами роботи є пошук та розробка моделей діагностування даних , обрання найбільш відповідного алгоритму та програмна реалізація цього алгоритму

2 МЕТОД НЕЧІТКОЇ КЛАСТЕРИЗАЦІЇ ДЛЯ ЕФЕКТИВНОГО ДОСЛІДЖЕННЯ ДАНИХ МЕДИКО-ТЕХНОЛОГІЧНОГО ПРОЦЕСУ

2.1 Алгоритми кластерізації в умовах нечіткості

За останні роки в результаті автоматизації своєї діяльності при використання баз даних у багатьох організаціях накопичилися великі обсяги даних, в яких міститься величезна кількість додаткової невиявленої та потенційно корисної інформації. Зокрема, у медичних установ робота з даними зводиться до накопичення статистики та формування звітів по ній [6]. Ефективний моніторинг накопиченої статистична інформація дозволяє, наприклад, готувати мотиваційну базу для прийняття медичних управлінських рішень та визначати статистичні показники для оцінки та виявлення існуючих та потенційних загроз різних несприятливих епідеміологічних ситуацій, для підвищення якості та ефективності заходів щодо виключення цих загроз [6]. Ефективний моніторинг даних досягається шляхом застосування методів інтелектуального аналізу, особливе місце у яких займають методи класифікації та кластеризації. Аналіз медичних даних [7] в умовах нечіткості, неповноти вихідних даних має нечіткий характер. При цьому необхідно використовувати апарат нечіткої логіки та теорії нечітких множин для формалізації таких даних [5]. Для роботи з такими даними слід застосовувати нечіткі аналітичні системи (НАС), в яких нечіткість дій, що відбуваються в системі в момент прийняття рішення можна уявити нечіткими алгоритмами. Існують основні підходи до інтелектуального аналізу даних які потребують уточнень для медичної сфери.

Класифікувати можлива більшість об'єктів [4], у тому числі і в медичних ПрВ. Для цього застосовують класифікаційні ознаки, ще їх називають основою поділу. Для побудови класифікаційного поділу застосовується набір правил логічного поділу [7] з незначними коригуваннями для медичної предметної галузі.

У медичній сфері у СППМР завдання класифікаціЇ застосовна у питаннях постановки медичного діагнозу.

Виділяють кілька підходів до побудови класифікацій [4,5], яких є ряд переваг та недоліків. Найбільш прийнятно у завданнях медичного призначення застосовувати фасетно-ієрархічний підхід. На підставі обраного підходу необхідно розробити алгоритм постановки медичного діагнозу та побудувати технологічну схему, яка буде застосовуватись у роботі автоматизованої системи медичного призначення. Інший підхід до інтелектуального аналізу даних – це кластеризація. Існує кілька методів кластеризації [4].

Загальноприйнята систематизація методів кластеризації відсутня. Однак, узагальнюючи всілякі об'єднання методів кластеризації в класи, допустимо виділити набір груп, причому, оскільки деякі методи можливо визначити одночасно у кілька груп, цю типізацію можна вважати Досить умовною [9]. Алгоритми, що застосовуються у кластерному аналізі, можливо розділити на неієрархічні, ієрархічні та методи класифікації з урахуванням навчання. Неієрархічні методи дозволяють знаходити і ідентифікувати у просторі вихідних змінних так звані "згущення" об'єктів.

Ієрархічні методи засновані на послідовному об'єднанні в кластери об'єктів відповідно до їхнього рівня близькості один до одного або, навпаки, на послідовному розбиття набору об'єктів ПРО на дедалі більше маленькі кластери. Кластерне рішення у разі можна уявити ієрархічною структурою кластерів, вкладених один в одного (за допомогою цього методу проводилася оцінка факторів ризику, що впливають на результати лікування вперше виявлених хворих на туберкульоз легень). Оскільки у медичній сфері досить великий обсяг інформації з медико-технологічних процесам носить описовий характер, то окремо необхідно виділити кластеризацію в завданнях Data Mining [2, 8], яка набуває значущості, якщо є одним з етапів інтелектуального аналізу медичних даних та бере участь у побудові закінченого аналітичного рішення. Технології Data Mining [6] дозволяють виявляти в медичних даних шаблони, складові основу логічних правил.

Закономірності між взаємозалежними подіями можна шукати з допомогою асоціативних правил [1]. Принципи побудови класичних асоціативних правил добре відомі, але потребують додаткових уточнень для медичної предметної галузі Аналіз на основі кластерного підходу в медичній предметній галузі буває корисним , коли необхідно класифікувати великі обсяги інформації (кластеризація захворювань, варіантів перебігу хвороби, схем лікування захворювань, симптомів хвороб, таксономія препаратів, груп пацієнтів тощо). Існують різні способи реалізації кластеризації , таким чином, необхідно вибрати найбільш підходящий і зробити його уточнення та коригування для медичної предметної галузі. Для швидкого та якісного вирішення завдання кластеризації потрібні методики отримання найкращих рішень .

Провівши класифікацію основних найбільш відомих алгоритмів кластеризації [7] і проаналізувавши їх переваги та недоліки, можна зробити висновок, що для медичної предметної області для обраної задачі найбільше підійде алгоритм нечіткої кластеризації, але потрібно його модифікація. Основна проблема при використанні будь-якого методу – це оцінити результати кластеризації, вибрати її оптимальність та коефіцієнти з метою оцінки результату. Подолати недоліки можливо за допомогою застосування апарату нечітких відносин, а зв'язок атрибутів досліджуваних даних будемо розглядати як нечіткі об'єктні зв'язки. Необхідно розробити модифікований метод нечіткої кластеризації, в якому слід врахувати недоліки існуючих методів; адаптувати його для вирішення задачі вибору варіанта перебігу хвороби згідно набір медичних показників конкретного пацієнта.

Слід розробити структурну схему модифікованого методу нечіткої кластеризації Для здійснення нечіткої кластеризації необхідно визначити формулу розрахунку нормальної міри подібності на відстані. При аналізі кількох відомих метрик [4] було зроблено висновок, що у медичній предметної області найбільш коректно застосовуватиме зважену метрику Евкліда. Для отримання

нечіткого відношення рівнозначності будемо (згідно підходам, описаним у роботах Кофмана [5]) обчислювати транзитивне замикання нечіткого відношення.

На підставі обраного підходу необхідно сформувати етапи методу нечіткий кластеризації та побудувати технологічну схему розбиття на кластери, яка застосовуватиметься у роботі автоматизованої системи медичне призначення.

Проблема у використанні розробленого методу (як і у відомих методах) полягає в оцінці результату кластеризації. У модифікації методу в залежності від обраного рівня нечіткого відношення будується розбиття на кластери із застосуванням класів рівнозначності, причому змінюються кількість та склад кластерів. Вирішити це питання допомагає оцінна функція критерію якості, що використовується при вирішенні задачі кластеризації . Критерії якості кластеризації – коефіцієнт розбиття, ентропія розбиття, ефективність розбиття – вимагають модифікації для застосування в розробленої методики, оскільки при малих значеннях числа кластерів при обчисленні коефіцієнта розбиття результат виходить некоректним, а використовувати ентропію розбиття, якщо діапазон значень критерію буде різним для кожної з проведених кластеризації, порівнювати різні рішення з його допомогою некоректно.

Оскільки кількість рівнів у градації велика та близька до кількості елементів у досліджуваній множині, отже, максимально можна отримати саме стільки і розбиття, але не всі ці розбиття можуть бути практично корисними. Інший критерій, що оцінює якість розбиття вихідного множини, використовує поняття «потужний кластер», який входить до групи найбільш значних кластерів. Це поняття як один із критеріїв було використано при виборі відповідного рішення. Необхідно розробити алгоритм знаходження потужних кластерів та метод застосування їх для оцінки результатів кластеризації. Для формалізації методу нечіткої кластеризації для вирішення задачі вибору варіанта перебігу хвороби згідно з набором медичних показників конкретного пацієнта

необхідно розробити алгоритм, у якому врахувати все особливості. Частина алгоритму реалізовуватиме вибір варіанта перебігу хвороби.

При розробці технології нечіткої кластеризації, враховуючи особливості медичної предметної області, особлива увага слідує приділити нормуванню даних медико-технологічного процесу. Важливе місце у структурній схемі автоматизованої системи медичне призначення займають медичні експертні системи. У роботах розглянуто стандартну схему побудови експертних систем. Однак при побудові медичних експертних систем необхідно враховувати особливості медичної галузі. Є два підходи до розуміння сутності оцінки медичного рішення: розроблений на засадах теорії штучного інтелекту і той, що формується на основі емпіричних даних.

2.2 Дослідження даних у нечіткій аналітичній системі медичного призначення

У медичних установах під час здійснення моніторингу накопиченої статистичної інформації активно використовуються методи на основі нечіткої кластеризації, де на підставі значення нечіткої функції належності значення вхідної множини даних відносяться до певного кластера, але на слабоструктурованої вихідної інформації традиційні підходи до нечіткої кластеризації неможливо отримати адекватні рішення. В ці методи закладається ряд припущень: кластери мають центр кластера (особливу внутрішню точку) та певну форму; спосіб розбиття визначається з урахуванням взаємозв'язку між даними та центральною частиною кластерів. В загальному у разі форма кластерів може бути довільною, а центри відсутні, тому був розроблений метод кластеризації, вільний від зазначених припущень і забезпечує розбиття лише з урахуванням відносин у існуючих статистичних даних. Для швидкого та якісного вирішення задачі кластеризації необхідна розробка методики вибору релевантних

рішень у зв'язку з цим особливої актуальності набуває розробка методики нечіткої кластеризації, в якій отримання якісного рішення здійснюється за заданим критеріям з використанням нечіткої функції власності.

Незаперечна перевага теорії нечітких множин над ймовірнісними підходами полягає в тому, що побудовані на її основі експертні системи медичного призначення характеризуються підвищеною ступенем спроможності прийнятих рішень, це з тим, що у розрахунок приймаються різні варіанти розвитку подій медико-технологічного процесу, що імовірнісним методам невластиво, оскільки вони спочатку розраховані на дискретний (кінцевий) набір сценаріїв.

Аналіз даних [5] за умов нечіткості, неповноти вихідних даних має нечіткий характер. При цьому успішно використовується апарат нечіткий логіки та теорії нечітких множин для формалізації таких даних. За такого підходу нечіткі вихідні дані формалізуються як лінгвістичних та нечітких змінних. Нечіткість дій, що відбуваються в системі в момент прийняття рішення, можна уявити нечіткими алгоритмами [2].

Аналітичні системи, здатні формалізувати та обробляти нечіткі дані в рамках нечітких алгоритмів, називаються нечіткими аналітичними системами. Інтелектуальний аналіз даних – галузь знань, що відноситься до області обробки даних, вивчає в досліджуваних даних опис та здійснення пошуку нетривіальних, прихованих та практично корисних різних закономірностей.

В основі сучасної технології інтелектуального аналізу даних використовується концепція шаблонів (патернів), що дають фрагментальний опис багатоаспектних взаємовідносин, що існують у вихідних даних .Закономірності, властиві підвиборкам даних, що надаються цими шаблонами, вони виражені у зручній формі для розуміння людиною та мають компактний вигляд.

Способами, що не мають обмежень рамками апріорних припущень про вид розподілів значень показників, що вивчаються, і про структурі самої вибірки здійснюється пошук шаблонів. При вказаному підході вирішуються такі завдання, як класифікація, регресія, кластеризація, виявлення асоціативних

правил, і навіть питання прогнозування. Основні підходи до інтелектуального аналізу медичних даних зображені на схемою (рис. 2.1).

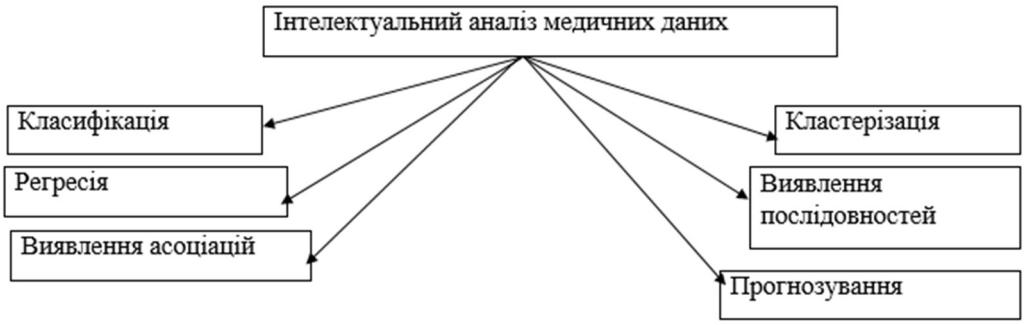

Рисунок 2.1 - Основні підходи до інтелектуального аналізу медичних даних

Класифікація дозволяє здійснити угруповання об'єктів (медичних показників) та позначити певні класи (діагнози), що характеризуються набором загальних якостей. Процедура угруповання з метою виділення загальних однорідних властивостей об'єктів на якісному рівні – це класифікація об'єктів. Класифікувати можливо більшість об'єктів [4], зокрема й у медичних ПрО. Ознаки, які дозволяють відокремити різні класи понять – це класифікаційні ознаки (КП), вони закладено основою класифікації. Ще КП називають основою розподілу. Перелічимо кілька правил логічного поділу [7]:

- правило альтернативності, що полягає у тому, що сформовані класи повинні бути несумісними, таким чином, кожен об'єкт з вихідна вибірка повинна належати тільки до єдиного класу;
- правило єдиного підстави поділу у тому, що логічний поділ може проводитися тільки по одній підставі, інакше для сформованих класів буде характерна різнорідність, які ряд буде безглуздим;
- правило однопорядковості поділу, згідно з яким неприпустимо здійснювати розподіл частини поняття ПрО на класи, а частини, що залишилася на підкласи;

- правило кінцівки поділу говорить, що кількість об'єктів, що належать ділимому поняттю, має дорівнювати числу всіх об'єктів, включених в класи. Необхідно відзначити два винятки з правил:
- розподіл за умови, що не всі класи наперед відомі. При такому підході до отриманих класів додають клас з умовною назвою "Інші";
- -дихотомія – розподіл поняття ПрО кілька класів. За такого поділу до першому класу належать об'єкти, що характеризуються деяким загальним властивістю, а другий клас містить інші об'єкти.

Виділяють три підходи побудови класифікацій – фасетний, ієрархічний та змішаний фасетно-ієрархічний. Недоліки ієрархічного підходу: неможливість урахування всіх існуючих аспектів об'єкта ПрО; труднощі при внесенні змін до структури через підпорядкованість КП однієї ієрархічної гілки, що призводять до необхідності вносити зміни в усі класифікаційні групи, розташовані нижче за змінений рівень ієрархії. Недоліки фасетного підходу: при побудові класифікації в рамках одного фасета відсутня деталізація. Їхнє спільне застосування дозволяє частково усунути перелічені недоліки. У задачі класифікації діагнозу будемо застосовувати змішаний фасетно-ієрархічний підхід. Набір однорідних значень необхідної класифікаційної ознаки включає кожен фасет. При цьому значення всередині фасету можуть розміщуватися як впорядковано (такий підхід краще), так і довільно. Побудова класифікації з урахуванням фасетного підходу організується у вигляді таблиці за наведеним алгоритмом:

- Визначити для кожного поняття $F = (F_1, F_2, \dots, F_n)$ фасетну формулу, що включає до свого складу конкретний набір класифікаційних ознак поняття.
- Побудувати таблицю, назви стовпців цієї таблиці будуть містити назви виділених у фасетній формулі класифікаційних ознак концепції.

– Поля таблиці заповнити значеннями відповідних фасетів $F_1^1, F_1^2, F_1^3, \dots, F_n^1, \dots, F_n^m$.Число значень різних фасетів може відрізнятись один від одного.

При ієрархічному підході можна описати лише один аспект і при цьому досить докладно деталізувати його. При фасетному – описується безліч аспектів поняття і навіть не деталізуються. Спільне використання цих підходів дозволяє побудувати вичерпну класифікацію понять, здатну вивчати поняття "вглиб" і "вшир". Отримана у результаті класифікація (при комбінованому підході) буде складатися з набору вкладених фасетних таблиць, аналіз яких вкрай утруднений для сприйняття, і можлива в основному при використанні програмних засобів. У класифікації даних на підставі значень змінних, які характеризують об'єкт, необхідно визначити значення залежної змінної. Припустимо, що задана множина об'єктів (досліджуваних даних) ПрО звичайно:

$$G = \{g_1, g_2, \dots, g_i, \dots, g_n\}, \tag{2.1}$$

Де g_i, - і-й об'єкт ПрО, n – кількість об'єктів у ПрО. Кожен із об'єктів ПрО характеризується певним набором атрибутів:

$$g_i(x_{i1}, x_{i2}, \dots, x_{ik}, \dots, x_{im}, x_{im+1}), \tag{2.2}$$

Нехай відомі значення m атрибутів об'єкта ПрО вказані в (2.2), тоді завдання полягає у знаходженні невідомого атрибуту x_{im+1}.

При класифікації безліч значень x_{im+1} звичайно, а при регресії безліч має потужність континууму або є лічильним.

Для отримання інформаційної підтримки при постановці медичного діагнозу в системі необхідно задати значення всіх відомих атрибутів, а система видасть ранжований список найбільш підходящих діагнозів, якого лікар-кори-

стувач вибирає (якщо він згоден) потрібний варіант. При відсутності потрібного діагнозу у списку можливе визначення значення додаткових атрибутів чи уточнення значення деяких атрибутів, інакше є можливість додати до списку новий опис діагнозу.

Завдання кластеризації в СППМР полягає в розбиття певного набору об'єктів (ситуацій з ПрО) на підмножини (кластери) таким чином, щоб отриманий кластер складався з подібних об'єктів ПрО, при цьому різних кластерах об'єкти значно відрізнялися б один від одного [4]. Завдання кластеризації належить до великого набору завдань навчання без вчителі, а також до завдань статистичної обробки.Шляхом виявлення кластерної структури відбувається розуміння даних – це одна з цілей кластеризації. Інша її мета – розбиття вихідного набору на плеяду подібних об'єктів, це дозволяє полегшити подальшу обробку інформації, а як наслідок, та процес прийняття рішення. До кожного кластера застосовується свій метод аналізу, наприклад: виявлення новизни у даних здійснюється за допомогою виділення нетипових об'єктів, які неможливо віднести до жодного з певних кластерів; стиск даних проводиться, якщо вихідна вибірка має надмірно великий обсяг, тоді можна здійснити її скорочення, залишивши від кожного кластера по одному представника, найбільш типового для цього набору. Кластеризація на основі навчання означає, що існує навчальна вибірка та кількість кластерів спочатку відома.

Навчальна вибірка – це група об'єктів, у яких визначено, яких груп вони ставляться. Інші об'єкти з аналізованої ПрО класифікуються відповідно до ступеня їх близькості до елементів із навчальної вибірки. Результати аналізу, проведеного за допомогою кластерного підходу, частіше всього видаються у вигляді дендрограми (тобто графічним способом), яка відображає порядок угруповання до кластерів об'єктів ПрО. Інтерпретація кластерної структури є досить творчим завданням, яка найчастіше починається з визначення кількості кластерів. Для ефективного вирішення цього завдання необхідно мати досить

повну інформацію про об'єкти, над якими проводиться кластеризація. Результати при кластеризації "з навчанням", можливо, представлені у вигляді набору об'єктів, віднесених до різних кластерів.

Цей підхід буде застосовуватися в задачі вибору варіанта перебігу хвороби, коли результат кластеризації порівнюється з еталонним набором, який відповідає шаблону БЗ. Основна перевага кластеризації: у використовуваних у кластерному аналіз змінних відсутній обмеження на розподіл; універсальність кластеризації - вона може застосовуватися як в аналізі сукупності об'єктів ПрО, так і до плеядів змінних із ПрО або працювати з будь-якими іншими одиницями аналізу із ПрО; кластеризацію можна проводити за відсутності апріорної інформації про характер класів та/або про кількість таких класів. При аналізі часто буває легше відзначити групи схожих за своїми характеристикам об'єктів, вивчити їх особливості та сформувати для кожної групи самостійну модель, ніж для всіх даних формувати одну загальну модель.

Закономірності між взаємозалежними подіями можливо знаходити з допомогою асоціативних правил. Визначення асоціативного правила в класичному сенсі можна представити так. Припустимо, що $I = \{i_1, i_2, \dots, i_{nd}\}$ – безліч атрибутів, у медичній предметної області найчастіше використовують атрибути так званих медичних показників. Dt – безліч з nd транзакції (етапи медико-технологічного процесу) таке, що

$$Dt = \{T_1, T_2, \dots, T_{nd}\}, \tag{2.3}$$

Де T_i– набір елементів (медичні показники у конкретний моментчасу) з I. Асоціативним правилом називається імплікація виду:

$$X \Rightarrow Y \text{ , где } X \subset I, Y \subset I, X \cap Y = \varnothing \tag{2.4}$$

Правило має підтримку (support), де sup - відсоток транзакцій з Dt, які містять $X \cup Y$:

$$sup(X \Rightarrow Y)=sup(X \cup Y) \quad (2.5)$$

Це правило вважається справедливим з певною достовірністю (confidence), де conf - відсоток транзакцій з Dt таких, що якщо транзакція містить X , то вона також містить Y :

$$conf(X \Rightarrow Y) = sup(X \cup Y)/sup(X). \quad (2.6)$$

Нестача класичних асоціативних правил у тому, що вони дозволяють працювати лише з атрибутами категоріального типу, тоді як дані медико-технологічного процесу найчастіше не обмежуються тільки бінарними значеннями. Усунути зазначений недолік можна за рахунок виявлення закономірностей найпоширеніших наборів (груп) об'єктів ПрО у великій плеяді таких груп.

2.3 Кластеризація у дослідженнях даних медико-технологічний процесу

Аналіз на основі кластерного підходу використовується в різних предметних областях. У медичній предметній галузі кластерний аналіз буває корисним [7], коли необхідно класифікувати великі обсяги інформації, наприклад застосовується кластеризація захворювань, варіантів перебігу хвороби, схем лікування захворювань, симптомів хвороб, таксономія препаратів, груп пацієнтів тощо. Кластеризація має різні способи реалізацій ,але за будь-якого підходу проблема полягає в недоступності будь-якої додаткової корисної інформації про дані на початковий момент аналізу, при цьому можлива безліч рішень щодо кардинального числу можна порівняти з вхідним безліччю, що не

можна здійснити практично. Необхідні методики підбору найкращого рішення для швидкого та якісного отримання результату у завданнях кластеризації, при цьому формально за певними критеріями здійснюється підбір найкращого рішення та попередньої інформації про кластери не потрібно. Елемент даних g (Об'єкт ПрО, що описується вектором характеристик) є елементом m-мірного простору: $g = (x_1, .., x_m)$.

Атрибут (Характеристика) x_i – числова компонента вектора g . Розмірність m – число характеристик об'єкта g.

Безліч об'єктів $G = (g_1, g_2, .., g_n)$. - масив вхідних даних, і-й об'єкт з G визначається як $g_i = (x_{i1}, \dots, x_{im})$.

Підмножина «близьких один до одного» (за своїми характеристиками) об'єктів ПрО з X називають кластером. Відстань d(g_i, g_j)між об'єктами g_i і g_j – результат використання певної метрики (квазиметрики) просторі певних показників об'єктів ПрО.

Формування оптимального розбиття об'єктів ПрО на групи метою кластеризації: розбити n об'єктів ПрО на кластери або розбити n об'єктів ПрО на k кластерів, де n – кількість об'єктів у ПрО, k – кількість кластерів, на які необхідно зробити розбиття. Оптимальність кластеризації можна визначити через середньоквадратичну помилку розбиття, пред'явивши до неї вимоги мінімізації:

$$e^2(X, L) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left\| x_i^{(i)} - c_j \right\|^2. \qquad (2.7)$$

Побудова кластеризації у класичному вигляді складається з чотирьох етапів.

- 1-й етап. Виділення показників об'єктів ПрО.
- 2-й етап. Вибір метрики визначення характеристик.
- 3-й етап. Розбиття об'єктів ПрО на групи.
- 4-й етап. Подання одержаних результатів.

Для виділення характеристик необхідно здійснити визначення властивостей, що найбільш точно характеризують об'єкти. Об'єкт може бути характеристики, що вимірюються кількісними значеннями (розміри, координати, інтервали, вимірювання медичних показників…), та характеристики, що вимірюються якісними значеннями (як правило, це значення, які вибираються зі списку: статус пацієнта, номер дієти ...). Потім треба виконати нормалізацію характеристик, для цього необхідно зменшити розмірність простору. Відображення об'єктів у вигляді характеристичних векторів відбувається на кінцевому етапі При визначенні метрики при побудові кластеризації необхідно вибрати метрику в залежності від простору, де розташовуються об'єкти ПрО, та від неявних характеристик кластерів.

У разі, коли є необхідність, щоб кластери представлялися областю у вигляді гіперсфер (що використовується досить часто), а всі координати об'єкти були речові і безперервні, використовується метрика Евкліда:

$$d(x_i, x_j) = \left(\sum_{z=1}^{|x|}(x_{i,z} - x_{j,z})^2\right)^{1/2} = \|x_i - x_j\|_2, \quad (2.8)$$

Вважена метрика Евкліда:

$$d_{ij} = d(x_i, x_j) = \left(\sum_{z=1}^{|x|} w_z(x_{i,z} - x_{j,z})^2\right)^{1/2}, \quad (2.9)$$

Де d_{ij} – відстань між i-м та j-м об'єктами; $x_{i,z}$, $x_{j,z}$ – значення z-ї змінної відповідно у i-го та j-го об'єктів; w_z - вага, що приписується z-ї змінної в моделі медичної ПрО.

Подання результатів розбиття має бути організоване в наочному вигляді, у якому зручно здійснювати оцінку результатів. Найчастіше всього кластери є центроїдами, набором характерних точок або їх обмеженнями . Для оцінки якості кластеризації використовують різні процедури, приклад такі, як: ручна

перевірка; перевірка контрольних точок на сформованих кластерах; процедура визначення стабільності кластеризації при додаванні нових змінних у модель; порівняння кластерів, отриманих різними методами.

При здійсненні ефективного моніторингу накопиченої статистичної інформації знайшли своє використання методи нечіткої кластеризації, в них до того чи іншого кластера відносять елементи вхідного множини на підставі отриманих значень нечіткої функції приналежності, але на слабоструктурованій вихідній інформації класичні способи нечіткої кластеризації не отримують у результаті адекватні рішення.

У ці методи закладається ряд припущень: кластери мають певного виду внутрішню точку – кластерний центр та задану форму; розбиття визначається з урахуванням взаємозв'язків між центрами кластерів та даними. Загалом кластери можуть не мати центрів і бути абсолютно довільної форми, тому був розроблений метод кластеризації, вільний від зазначених припущень і забезпечує розбиття лише з урахуванням відносин у наявних статистичних даних [4]. Таким чином, для швидкого та якісного розв'язання задач кластеризації потрібні методики отримання кращих рішень . На рис. 2.2 наведено класифікацію алгоритмів кластеризації.

Основними та найбільш використовуваними є такі: на основі мінімального покриваючого дерева; різновиди ієрархічних алгоритмів; алгоритм загартування; алгоритм, використовує метод найближчого сусіда; алгоритми на основі нечіткої кластеризації; генетичні алгоритми; алгоритми з урахуванням нейронних мереж; алгоритм k-середніх (k-means алгоритм) та ін.

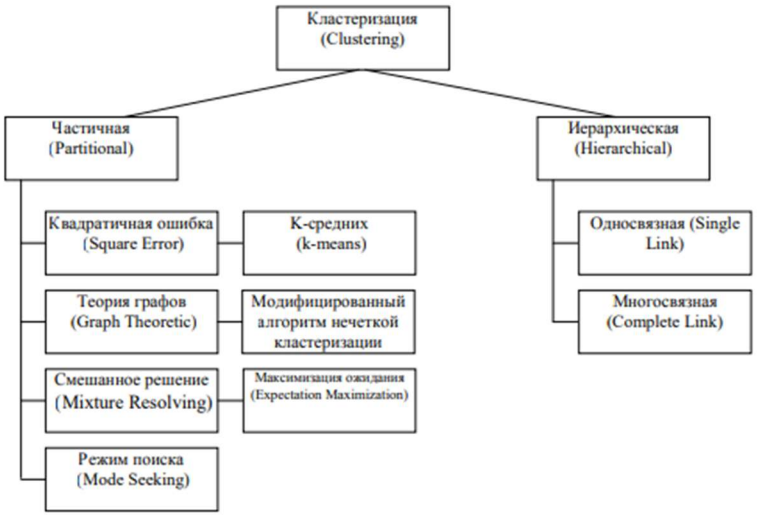

Рисунок 2.2 – Класифікація алгоритмів кластеризації

Методи та алгоритми вирішення різних завдань інтелектуального аналізу вихідної інформації використовують математичний апарат теорії нечітких множин, різні класичні підходи теорії множин, методи теорії семантичних мереж та математичної статистики, а також апарат універсальної алгебри тощо.

Для формалізованого завдання алгоритмічне рішення пов'язане із знаходженням екстремального значення оціночної функції. Ефективний аналіз даних за умови нечіткості та неповноти вихідної інформації має нечіткий характер. В даний час для їх формалізації успішно використовується апарат теорії нечіткої логіки та нечітких множин. Над імовірнісними підходами теорія нечітких множин має явною перевагою, яка полягає в тому, що МЕС, спроектовані на її основі, мають підвищений рівень обґрунтованості прийнятих рішень, оскільки розглядаються всі схеми розвитку подій, які можуть статися, що не властиво імовірнісним методам, визначеним на дискретному (кінцевому) безлічі сценаріїв.

2.4 Метод та алгоритм нечіткої кластеризації

В інтелектуальній аналітичній системі моніторингу пацієнтів на основі нечіткої кластеризації для медичних установ за допомогою кластеризації можливе вирішення задачі визначення варіанта перебігу хвороби з використанням аналізу статистичної інформації, у якій нічого невідомо про внутрішні взаємозалежності даних. Завдання знаходження залежностей у вихідній множині, що впливають на угруповання даних, можна формулювати більш детально завдяки кластеризації та проводити ефективний моніторинг інформації.

У відомих алгоритмах є недоліки: застосування у рішенні поняття кластерного центру (хоча може бути відсутнім); вилучення кластерів лише формою, визначеною алгоритмом (частина кластерів може бути пропущено); отримання кластерів з урахуванням відносин між центрами кластерів та елементами даних.

Подолати недоліки відомих алгоритмів можливо з допомогою застосування апарату нечітких відносин, а зв'язок атрибутів досліджуваних даних розглядатимемо як нечіткі об'єктні зв'язки. Укрупнена структурна схема модифікованого методу нечіткої кластеризації представлено рис. 2.3. Модифікація методу полягає у додаванні циклу, в тілі якого здійснюється розрахунок критеріїв якості результатів проведеної кластеризації, і основі цього вибирається рівень нечіткого відносини.

Рисунок 2.3 – Укрупнена структурна схема методу нечіткої кластеризації

Медичні показники в системі представлені у вигляді множини $G = \{g_1, g_2, \dots, g_n\}$. Для кожного медичного показника визначено кваліфікаційні ознаки$g_i(x_{i1}, x_{i2}, \dots, x_{ik})$, а також визначено основні характеристики у моделі медичної ПрО на основі семантичної мережі (Опис якої наводиться в розділі 3). Кожен медичний показник є об'єктом моделі медичної ПрЗ. Виділено п'ять основних груп медичних показників, представлених рис. 2.4.

Рисунок 2.4 – Основні групи медичних показників

Таким чином, на початковому етапі кожного пацієнта в системі зберігається 200 медичних показників. Кожен показник характеризується п'ятьма ознаками. Всі дані щодо пацієнта зберігаються в системі в нормалізованому вигляді. У результаті для нечіткої кластеризації на основі значень медичних показників за пацієнтом формується безліч X. Чітка (неперетинна) кластеризація визначає атрибут x_i з ПрВ тільки до певного кластера, тоді як нечітка кластеризація обчислює кожному атрибуту x_{ik} з ПрО ступінь його належності u_{ij}до кожного з k кластерів, що визначається функцією приналежності f_{ij},яка показує ступінь належності x_i до кластера kl_j.Для здійснення нечіткої кластеризації необхідно виконати перелічені дії.

Схема нечіткої кластеризації:

– Задати початкове нечітке розбиття n об'єктів на k кластерів шляхом вказівки матриці приладдя U розмірністю n × k . Значення елементів u_{ij} Матриці знаходиться в межах[0,1].
– Використовуючи елементи приладдя матриці U , обчислити критерій нечіткої помилки:

$$E^2(X,U) = \sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}\left\|x_i - c_j\right\|^2, \quad (2.10)$$

де c_j- центральна частина нечіткого кластера kl_j,

$$c_j = \sum_{i=1}^{n} u_{ij}\, x_i, \quad (2.11)$$

– Для зменшення величини критерію нечіткої помилки можна перегруповувати об'єкти вихідної множини, доки не буде отримано результат, що задовольняє задану похибку.
– Виконувати дії з пункту 2 до досягнення незначних (Певна величина, що задається в налаштуваннях) змін елементів матриці U.

Гібридні підходи найкраще зарекомендували себе у вирішенні практичних завдань, у яких доналаштування кластерів виконується методом k means, а одним із найбільш підходящих методів здійснюється первісне розбиття. Метод k-means простий у реалізації та досить швидко працює, але при цьому здатний створювати схожі на гіперсфери кластери.

В алгоритмах апріорне застосування природи кластерів бувають такими як неявне використання характеристик кластерів:

– вибір метрики для аналізу (гіперсферичні кластери зазвичай виходять за метрики Евкліда);
– вибір з усіх наявних характеристик у ПрО відповідних характеристик об'єктів;

або явне використання характеристик кластерів:

- підрахунок схожості [застосовується поняття нескінченності для відображення відстані між об'єктами ПрО фактично різних кластерів];
- формування подання результатів (необхідно враховувати явні обмеження).

Для здійснення нечіткої кластеризації визначимо такі концепції. Нормальний захід подібності на відстані μ_y(x) породжує близькі до у нечіткі безлічі точок:

$$\mu_y(x) = 1\frac{d(y,x)}{\max\limits_{z\in X}(d(y,z))}, \tag{2.12}$$

где x, y,z∈X, d(y, x) – відстань між у и х . При цьому $\mu_y(x) = 0$, якщо атрибут максимально відрізняється від х , та $\mu_y(x) = 1$, якщо атрибут абсолютно подібний х для х∈ X . Визначимо відносний захід подібності 2-х атрибутів по відношенню до третьому атрибуту як $\tau_y(x,z)$:

$$\tau_y(x,z) = 1 - \left|\mu_y(x) - \mu_y(z)\right|, \tag{2.13}$$

Де x, y,z∈ X , а μ_y – нормальний захід подібності (2.17). У цьому сімействі кожне ставлення є нечітким ставленням. Через $\tau_y(x,z)$ можна визначити міру подоби на безлічі X 2-х атрибутів як:

$$\tau(x,z) = T\left(\tau_{y1}(x,z), \tau_{y2}(x,z), \dots, \tau_{y|X|}(x,z)\right), \tag{2.14}$$

Де $\tau_{yi}(x,z)$ - відносний захід подоби $y_i \in X$, $i = 1, \dots, |X|, x, z \in X, T -$ Операція t-норма.

Трикутною нормою (t -нормою) визначається бінарна операція T , що відповідає перерахованим аксіомам, задана на одиничному інтервалі $[0,1]\times[0,1]\rightarrow[0,1]$ і вірна для різних a, b, $c\in[0,1]$

Досить часто застосовується t - норма об'єднання по Заді, яку і будемо використовувати в даному випадку:

$$T(a,b)=\min(a,b), \tag{2.15}$$

Використовуючи зв'язки t -норми по Заді, мірі подібності 2-х атрибутів на безлічі X (2.19) відобразимо у вигляді:

$$\tau(x,z)=\min\left(\tau_{y1}(x,z),\tau_{y2}(x,z),\dots,\tau_{y|X|}(x,z)\right), \tag{2.16}$$

Таким чином, якщо пара атрибутів подібна до y_1 и y_2 … і подібна щодо $y_{|X|}$, то пара атрибутів подібна до всього безлічі X. Отримане вираз об'єктивно відображає схожість атрибутів безлічі X і є нечітким ставленням. При обчисленні транзитивного замикання нечіткого відношення виходить нечітке ставлення до рівнозначності. Для підтвердження цього існує низка тверджень та положень О.Кофмана , які можна застосувати у разі.

- Твердження 1. Завдання рівня нечіткої рівнозначності породжує розбиття множини X на групи рівнозначних атрибутів так, що кожен з атрибутів X належить до однієї групи рівнозначності.
- Твердження 2. При транзитивному замиканні відносини нечіткі толерантності виникає відношення нечіткої рівнозначності на множині X.
- Твердження 3. Операція об'єднання відносин нечіткої толерантності є ставленням нечіткої толерантності.

З урахуванням усіх розглянутих аспектів [діагностика при ранніх формах захворювань навіть за відсутності клінічної картини; аналіз динаміки розвитку патологічного процесу із припущенням можливих несприятливих ситуацій (включаючи терапію та побічні ефекти, що проводиться) медикаментів); облік супутніх захворювань при доборі пацієнта курсу лікування; оцінка стану здоров'я пацієнта «в режимі реального часу» з використанням даних, що надходять з моніторно-приладових комплексів з актуалізацією логіко-обчислювальних систем тощо] запропоновано алгоритм нечіткої кластеризації, що використовує нечітке відношення рівнозначності.

Етапи нечіткої кластеризації представлені на рис. 2.5 і дозволяють оброблюваних даних медико-технологічного процесу ефективно виявляти кластери (для визначення варіанта перебігу хвороби). Градація відношення нечіткої рівнозначності породжує сімейство відносин рівнозначності, які розбивають на класи рівнозначності вихідна множина досліджуваних вихідних даних. Чим більший рівень відносини, тим більше детально розбиття множини X (рис. 2.6). Градація S_τ нечіткого відношення рівнозначності у класичному сенсі групує сімейство відносин рівнозначності, кожне з яких розбиває вихідну множину на класи рівнозначності. Більш детальне розбиття множини X характерно для відношення вищого рівня. Кожен рівень породжує відповідне розбиття на класи рівнозначності. Для побудови S_τ упорядкуємо отримані елементи матриці $\tau(x, z)$ по зростанню.

Рисунок 2.5 – Етапи нечіткої кластеризації

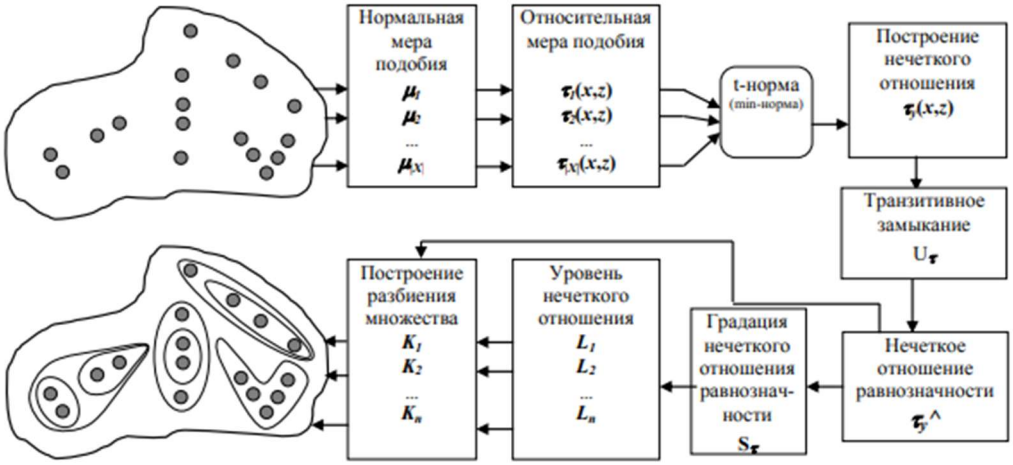

Рисунок 2.6 – Технологія нечіткого розбиття на кластери

Залежно від обраного рівня нечіткого відношення Li будується розбиття на кластери, при цьому змінюються кількість та склад кластерів. Результат цього розбиття необхідно оцінювати за допомогою різних характеристик, які

будуть представлені далі. На підставі їх аналізу робиться висновок про прийняття результатів кластеризації. При використанні алгоритмів кластеризації виникають певні складності, які найчастіше пов'язані з необхідністю визначати число кластерів на початок аналізу, яке завжди можна призначити обгрунтовано. Вирішити це питання допомагає оцінна функція критерію якості, використовувана під час вирішення завдання кластеризації [2].Для оцінки якості кластеризації існує кілька відомих критеріїв, які застосовуються для нечіткої кластеризації. Першим критерієм розглянемо коефіцієнт розбиття:

$$Kp = \frac{1}{|X|} \times \sum_{i=1}^{|X|} \sum_{j=1}^{|X|} u_{ij}^2, \tag{2.17}$$

Де u_{ij} $(\tau(x, z))$ відповідає елементу матриці приладдя U , K – безліч кластерів, X – вхідна множина атрибутів. Значення цього критерію приймають із діапазону $[|K|^{-1}, 1]$, причому значенню 1 відповідає максимально чітке розбиття, що відповідає найбільшій невизначеності і є найгіршим випадком розбиття. При такому підході за малих значень числа кластерів при обчисленні коефіцієнта розбиття результат виходить некоректним - це пов'язано з його областю значень. Залишивши колишнім характер критерію, область значень перемістимо так, щоб залежність від числа кластерів була пов'язана з його закінченням, а не з початком вибраного відрізка. Отримати такий результат можливо, виконавши віднімання з коефіцієнта розбиття $|K|^{-1}$. Перетворений критерій обчислюватиметься за такою формулою:

$$Kp_y = \frac{1}{|X|} \times \sum_{i=1}^{|X|} \sum_{j=1}^{|k|} u_{ij}^2 - \frac{1}{|K|}, \tag{2.18}$$

Kp_y - Поліпшений коефіцієнт розбиття. При цьому область значень покращеного коефіцієнта належатиме відрізку [0,(K –1)/ K] і проблема, що виникла, буде вирішена.

Другим критерієм розглянемо ентропію розбиття, яку можна представити у такому вигляді:

$$Эр = \frac{1}{|X|} \times \sum_{i=1}^{|X|} \sum_{j=1}^{|K|} u_{ij} \ln(u_{ij}), \quad (2.19)$$

Де u_{ij} відповідає елементу матриці приладдя U , K – безліч кластерів, X – вхідна множина атрибутів. Значення цього критерію приймають з діапазону [0,ln K], причому найгіршому розбиття відповідає ln K , а найкращому - 0. Так як діапазон значень критерію буде різним для кожної з проведених кластеризацій, порівнювати різні рішення з його допомогою некоректно, тому коректнішим буде використовувати покращену ентропію розбиття:

$$Эр_{У} = \frac{1}{|X|} \times \sum_{i=1}^{|X|} \sum_{j=1}^{|K|} u_{ij} \ln(u_{ij}) \times \frac{1}{\ln|K|}, \quad (2.20)$$

При такому підході діапазон значень лежить у відрізку [0,1] і не пов'язаний з кількістю кластерів. За допомогою коефіцієнта покращеної ентропії розбиття можливо здійснювати порівняння кластеризації з різним числом кластерів. Третім критерієм у групі кластерних критеріїв розглянемо ефективність розбиття, яку можна представити формулою:

$$ЭФр = \sum_{i=1}^{|X|} \sum_{j=1}^{|K|} u_{ij}^2 \, (d^2\left(c_j, \underset{x}{\rightarrow}\right) - d^2(x_i, c_j), \quad (2.21)$$

Після перетворень цю формулу можна представити у вигляді:

$$ЭФр = \sum_{i=1}^{|X|} \sum_{j=1}^{|K|} u_{ij}^2 \, d^2\left(c_j, \underset{x}{\rightarrow}\right) - \sum_{i=1}^{|X|} \sum_{j=1}^{|K|} u_{ij}^2 \, d^2(x_i, c_j), \quad (2.22)$$

Де $\underset{x}{\rightarrow}$ - середнє значення елементів, що належать вхідній множині, c_j – центр кластера kl_j [Розрахунок за формулою (2.16)] , d(x_i, c_j) представляє

відстань між двома об'єктами ПрО x_i та c_j - значення результату, отриманого при застосування підібраної квазиметрики (або метрики) у просторі відомих характеристик ПрО [розрахунок за формулою (2.13)]. Ефективність розбиття умовно поділяється на дві частини: перша частина відображає внутрішньокластерні відмінності, чим ці відмінності менші, тим кластеризація виконана краще, друга частина коефіцієнта показує міжкластерні відмінності, чим рівень відмінності вищий, тим кластеризація виконано краще. Отже, чим краще виконано кластеризацію, тим значення критерію ефективності розбиття більше.

У роботі при створенні системи були використані перелічені вище критерії, потім зроблено висновок про корисність застосування практично покращеного коефіцієнта розбиття, покращеної ентропії розбиття та покращеного критерію ефективності розбиття. Опис цих коефіцієнтів представлено на рис. 2.7.

Покращені коефіцієнти дозволяють оцінювати якість кластеризації як у великій кількості кластерів, і на малому, у своїй отримувати результати оцінок в інтервалі [0,1], що зручно і при ручній та при автоматизованої оцінки. За допомогою розробленого нечіткого відношення рівнозначності з набутого сімейства відносин рівнозначності необхідно вибрати найкраще рішення. При цьому виникає низка труднощів, подолати які можна за допомогою розроблених критеріїв, що оцінюють якість розбиття вихідної множини.

Назва	Формула розрахунку
1.Поліпшений коефіцієнт розбиття	$Kp_y = \frac{1}{\|X\|} \times \sum_{i=1}^{\|X\|} \sum_{j=1}^{\|k\|} u_{ij}^2 - \frac{1}{\|K\|}$ (2.23)
2. Поліпшена ентропія розбиття	$Эр_y = \frac{1}{\|X\|} \times \sum_{i=1}^{\|X\|} \sum_{j=1}^{\|K\|} u_{ij} \ln(u_{ij}) \times \frac{1}{\ln\|K\|}$ (2.25)
3.Ефективність розбиття	$ЭФр = \sum_{i=1}^{\|X\|} \sum_{j=1}^{\|K\|} u_{ij}^2 d^2\left(c_j, \underset{x}{\rightarrow}\right) - \sum_{i=1}^{\|X\|} \sum_{j=1}^{\|K\|} u_{ij}^2 d^2(x_i, c_j)$ (2.27)

Рисунок 2.7 – Критерії якості кластеризації

Кількість рівнів у градації S_τ велике і близько до кількості елементів у досліджуваній множині, значить, максимально можна отримати саме стільки і розбиття, але не всі ці розбиття можуть бути практично корисними. Зазвичай класи рівнозначності значно відрізняються за потужністю, особливо у першій половині градації S_τ.

Один із критеріїв, що оцінюють якість розбиття вихідної множини, використовує поняття «потужний кластер», який входить до групи найбільш значних кластерів. Введення цього поняття стало необхідним через те, що класи рівнозначності суттєво розрізняються за кардинальним числом (значимістю), особливо в початкових значення градації S_τ нечіткої рівнозначності, коли найбільша кількість кластерів містить невелику кількість елементів об'єктів ПрО. Набір потужних кластерів можна визначити за формулою:

$$K_{MK} = \{\forall k_i' \in K'\}, \quad (2.23)$$

де k_i' - клас – клас рівнозначності такий, що $|k_i'| \geq$ ПКЧ, K' – безліч класів рівнозначності, отримане для вибраного рівня еквівалентності, а ПКЧ - поріг кардинального числа (мінімальна потужність класу рівнозначності) при

якому кластер вважається практично корисним. ПКЧ розраховується рекурсивним способом: кластери – класи рівнозначності впорядковуються згідно з кількістю елементів, включених до них (від більшого до меншого) він обчислюється зважене ставлення кожної пари класів по формуле:

$$\text{ПКЧ}_{\text{ВО}} = \text{ПКЧ}_{\text{ВК}} \times \text{ПКЧ}_{\text{ОК}}, \tag{2.24}$$

Де $\text{ПКЧ}_{\text{ВК}}$ - ваговий коефіцієнт, який зміщує максимум $\text{ПКЧ}_{\text{ВО}}$ ближче до початку послідовності, $\text{ПКЧ}_{\text{ОК}}$ – це відношення двох сусідніх класів.

Для отримання набору потужних кластерів наведену вище процедуру необхідно зробити для кожного рівня нечіткого відношення L_i , потім обчислюємо проміжний коефіцієнт розбиття Крп за такою формулою:

$$K_{pn} = \frac{1}{|X|} \times \sum_{k_j' \in K_{MK}} |k_j'|, \tag{2.25}$$

Де $|k_j'|$ – кардинальне число сильних кластерів, |X| - загальне кардинальне число множини. Сформулюємо критерій якості з метою оцінки розбиття, застосовуючи поняття рівня відношення рівнозначності, потужних кластерів та проміжного коефіцієнта розбиття:

$$KP_i = L_i \cdot |K_{MK}| \cdot K_{pm}, \tag{2.26}$$

Де L_i – рівень нечіткої рівнозначності $|K_{MK}|$ – кардинальна кількість множини потужні кластери. Найкращим розбиттям вважатимемо таке, за якого:

$$KP = \max_i \left(L_i \cdot |K_{MK}| \cdot K_{pm}\right) \tag{2.27}$$

Критерій дозволяє враховувати такі фактори:

- Високий рівень нечіткої рівнозначності (L_i) говорить про те, що всередині класів рівнозначності перебувають найбільш схожі об'єкти ПрО.
- Найкраща якість розбиття досягається за великого кардинального числі безлічі потужних кластерів $|K_{MK}|$.
- Чим вище значення коефіцієнта розбиття K_{pm}, тим більше об'єктів Про включено в результат остаточного розбиття.

У міру збільшення рівня нечіткої рівнозначності безліч потужних кластерів збільшується, але зменшується проміжний коефіцієнт розбиття. формула (2.27) та процедура вибору потужних кластерів виведені емпіричним шляхом на підставі загальних понять про якість розбиття, згідно проведеним тестам на вже наявних даних статистики та консультацій з експертами з медичної ПрВ. Алгоритм модифікованого методу нечіткої кластеризації представлений рис.2.8.У системі зберігаються шаблони варіантів перебігу хвороб по кожному діагнозу, отримані шляхом аналізу медичних статистичних данихгрупою експертів та інженером зі знань. За кожним шаблоном зберігається набір медичних показників з певними значеннями ваг. Якщо діагноз у пацієнта не встановлено, можливо отримати варіант перебігу хвороби із загального списку, без прив'язки до діагнозу. У цьому випадку список варіантів перебігу хвороби буде набагато більше, а точність пропонованого рішення нижче.

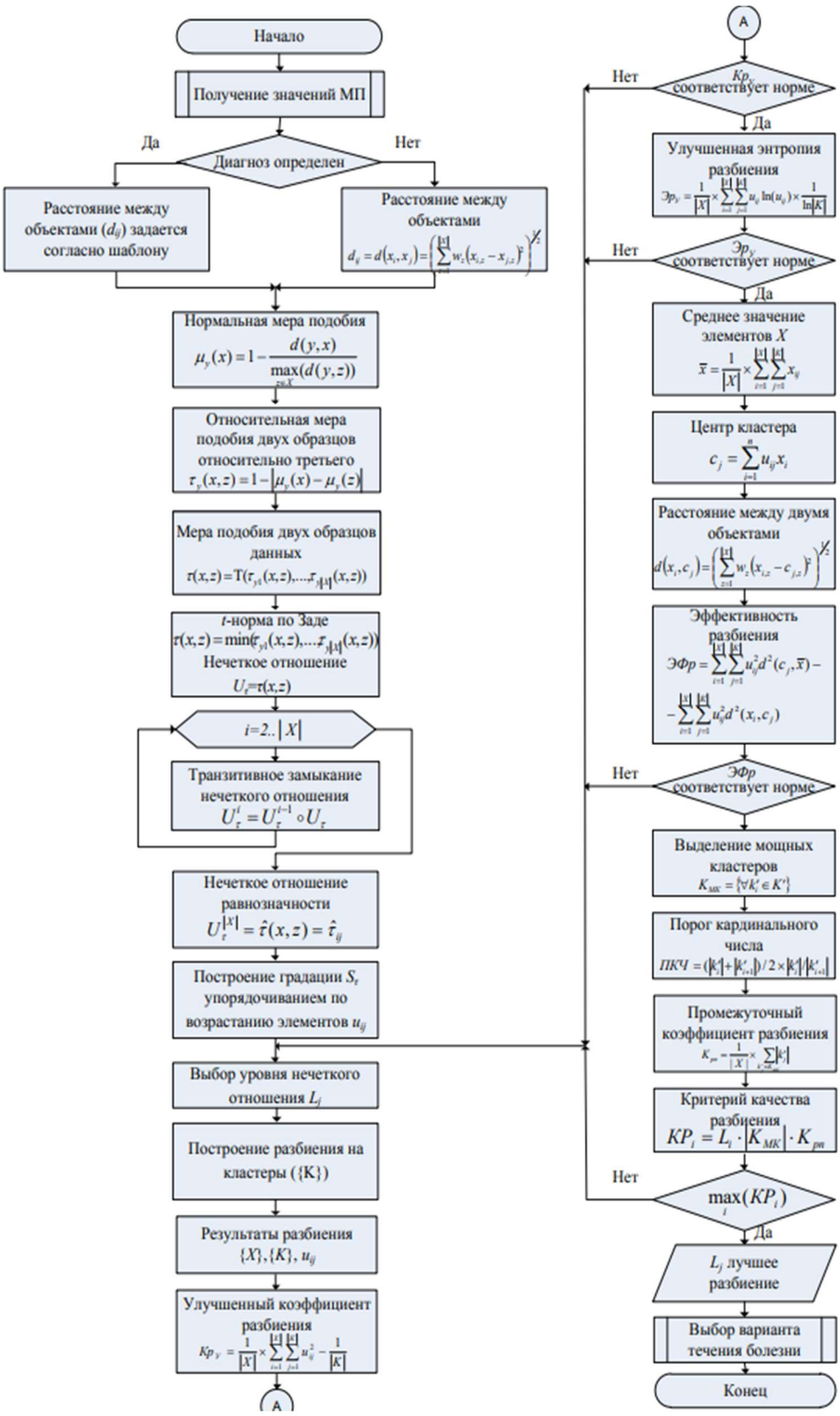

Рисунок 2.8 – Алгоритм модифікованого методу нечіткої кластеризації медичних показників при виборі варіанта перебігу хвороби.

Значення ваг (u_{ij}) медичних показників пацієнта (результат нечіткої кластеризації) необхідно порівняти з вагами шаблонів і на підставі цього порівняння побудувати ранжований перелік можливих варіантів. Для складання ранжованого списку та усічення його необхідно провести оцінку міри близькості МП пацієнта та МП шаблонів. Оцінка $\varphi_{\text{МПП}}(H_k)$ встановлює міру близькості варіанта перебігу хвороби, що склалася ситуації щодо СЗП:

$$\varphi_{\text{МПП}}(H_k) = \sum_{j=1}^{N_{\text{МПП}}} u_j \sum_{i=1}^{M_{Hk}} w_i \, \mu_S(P_i, P_j), \tag{2.28}$$

Де $N_{\text{МП}}$ – число понять, що належать моделі СЗП, M_{Hk} – число понять, що належать шаблону H_k, u_i – значимість поняття у ситуації пацієнта, w_i – значимість поняття у шаблоні, $\mu_S(P_i, P_j)$- близькість і-го поняття до j-го. Чим більше значення $\varphi_{\text{МПП}}(H_k)$ тим шаблон ближче до ситуації пацієнта і тим більше значимість набору процесів для пацієнта. Приписане шаблону H_k число $\varphi_{\text{МПП}}(H_k)$ визначимо як критеріальну оцінку, а сформовану шкалу назвемо критерієвою. Таким чином, шуканим шаблоном буде безліч шаблонів, які відповідають умові:

$$\max_{H_k \in H} \varphi\,(H_k), \tag{2.29}$$

Шаблони, які при порівнянні з іншими шаблонами, що належать моделі ПрО, що мають максимальне значення критеріальної оцінки, включаються в список вибраних шаблонів у ранжованій послідовності від більшого до меншому. Лікар-користувач із запропонованого списку вибирає найбільше придатний відповідно до ситуації пацієнта. Алгоритм вибору варіанта перебігу хвороби представлений рис. 2.9.

Потім згідно з вибраним варіантом перебігу хвороби формується схема лікування пацієнта. Розроблена методика нечіткої кластеризації складається з

10 кроків. Схема послідовності кроків зображена рис.2.10. Попередня підготовка даних для аналізу та виділення атрибутів полягає у виборі безлічі об'єктів ПрО (медичні показники), які повинні повно, але при цьому коротко представляти досліджувану множину (решта не приймаються до розгляду, оскільки не несуть корисної для аналізу інформації), та в оцінці класифікаційних ознак медичних показників (якщо вони були отримані чи задані у вигляді). На цьому етапі всі масиви інформації про медико-технологічний процес формалізуються та наводяться до числових значень.

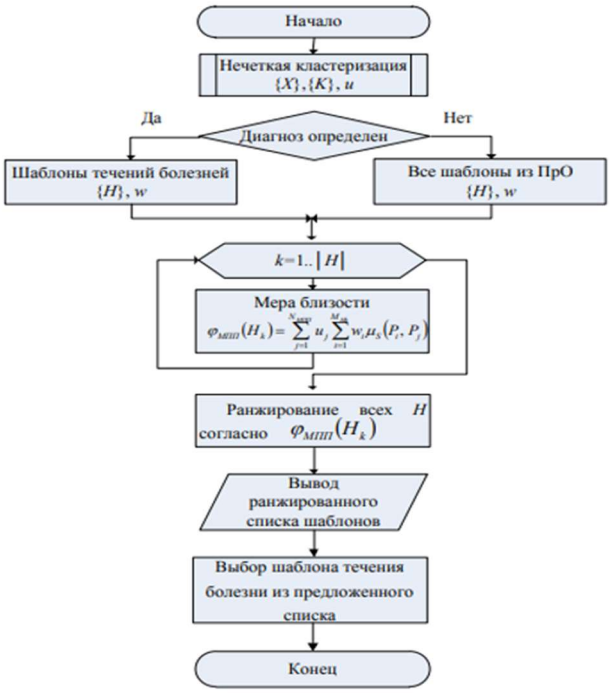

Рисунок 2.9 – Алгоритм вибору варіанта перебігу хвороби

Рисунок 2.10 – Технологія нечіткої кластеризації

Визначення мети аналізу здійснюється з представленого набору:

- Визначення кількості та складу кластерів, при цьому виявляється кластерний склад даних. Цей етап є основною метою, заради якої вирішується завдання кластеризації.
- Пошук елементів безлічі об'єктів ПрО, які не входитимуть в один із кластерів. На цьому етапі виявляються відхилення (що вказують на патологію в процесі),що створили досліджувані дані з ПрО.
- На додаток до проведення кластерного аналізу даних необхідно ще підготуватися до вирішення задачі класифікації для постобробки отриманих результатів, провести опис та дати назву сформованим кластерів.

Тоді класами назвемо пойменовані та описані кластери, а вирішенням задачі класифікації будуть значення атрибутів елементів безлічі кластерів та встановлені залежності між ними.

Визначення засобів аналізу та методу відображення результатів. Різні програмні системи можуть бути засобами аналізу. Дляпроведення дослідження критеріїв може використовуватися середовище Matlab та Mathcad

або інші системи такого типу. При розв'язанні задачі аналізу статистичних даних використовувалися власна розроблена система та бібліотека алгоритмів, які у ній.

Спосіб відображення результатів залежить від типу досліджуваної множини об'єктів ПрО і може бути представлений в такий спосіб:

- Просте перерахування. За такого підходу кожен сформований кластер описується представленням елементів, у тому числі він складається. Цей спосіб підходить при застосуванні будь-якого методу кластеризації.
- дендрограма. Побудова здійснюється природним чином при застосування методу нечіткого відношення рівнозначності та ієрархічних методів. Цей спосіб відображення результатів у загальному випадку не підходить для нечіткої кластеризації.
- Матриця приладдя - спосіб відображення результатів, який є найбільш відповідним для нечіткої кластеризації. Результати кластеризації відображаються у вигляді таблиці, де стовпці відповідають кластерам, а рядки елементам. У осередках таблиці зберігаються значення, відповідні функції власності.

Нормування даних - технологічний етап підготовки даних, полягає в наступному:

Переклад ординальних даних та категоріальних у числові. Так як числовий ряд спочатку є впорядкованим, то ординальні змінніможна вважати найближчими до числової форми. Отже, для кодування таких змінних достатньо визначити у відповідність до номерів категорій числові значення, що зберігають існуючу впорядкованість. Одиничний відрізок розбивається за кількістю класів на n відрізків із довжинами, пропорційними кількості зразків кожного класу у вибірці для навчання:

$$\Delta a_k = \frac{N'_k}{N}, \tag{2.30}$$

Де a_k - змінна з даними з ПрО, N_k - Число зразків класу k , а N - загальна кількість прикладів. Чисельне значення для відповідного ординального класу – це центр кожного представленого відрізка. Категоріальні змінні медичних показників є іменами категорій та своїм значенням зазвичай позначають один із класів.

Наприклад, це можуть бути назви лікарських препаратів або назви категорій соціального статусу пацієнта Упорядкуванням або зважуванням здійснюється переведення в числові дані категоріальних даних. Експерт ПрО приписує числові значення категоріальним атрибутам, цим здійснюється зважування. Упорядкування можна здійснити без експерта, але це менш ефективне Стратегія. Значення категоріального атрибута надається порядковий номер, така операція провадиться для кожного значення. Щоб уникнути помилок розбиття, атрибут краще виключити з розгляду у разі сумніву у коректності застосування стратегії упорядкування чи неможливості застосувати знання експерта.

Нормування числових даних у діапазоні [0,1] необхідно для того, щоб кожен атрибут мав однакове вагове значення у порівнянні даних. При нормуванні необхідно враховувати випадок, коли вагове значення атрибутів різне. Нормування змінної ПрО на інтервал розкидання її значень здійснюється зведенням даних медико-технологічного процесу до одиничного масштабу. У класичному варіанті можна використовувати лінійне перетворення:

$$\tilde{a}_k = \frac{a_k - a_{kmin}}{1}(a_{k\,max} - a_{kmin}), \qquad (2.31)$$

яке дозволяє перейти до одиничного відрізку: $\tilde{a}_k \in [0,1]$.

Аналогічно за такому ж принципу здійснюється відображення даних в інтервал [-1,1], рекомендований для вхідних даних. У разі щільного заповнення певного інтервалу значенням змінної a_k оптимальне застосування лінійного нормування.

Але далеко не завжди подібний "лінійний" підхід можна застосувати. Існують випадки, коли в даних є порівняно рідкісні викиди, значення яких набагато перевищують типовий розкид даних. Відповідно до попередньої формули нормувальний масштаб визначатимуть значення саме цих викидів. Невірно отриманий масштаб спричинить ситуацію, за якої основний набір значень нормованої змінної $\tilde{a}_k$буде зосереджений по близу нуля:$|\tilde{a}_k| \ll 1$.

Згідно з наведеними вище зауваженнями можна зробити висновок, що більше коректно при нормуванні вихідних даних медико-технологічного процесу при кластеризації орієнтуватися на такі типові значення (статистичні характеристики вихідних даних), як дисперсія та середня, а не використовувати екстремальні значення:

$$\tilde{a}_k = \frac{a_k - \bar{a}_k}{\sigma_k}, \tag{2.32}$$

$$\bar{a}_k = \frac{1}{N}\sum_{b=1}^{N} a_k^b, \tag{2.33}$$

$$\sigma_k^2 = \frac{1}{N-1}\sum_{b=1}^{N}\left(a_k^b - \bar{a}_k\right)^2, \tag{2.34}$$

Де N - загальна кількість прикладів, a_k^b - b-е значення k-й змінної з ПрО. За такого підходу в основному наборі даних типові значення будуть можна порівняти у всіх змінних, тобто. набір матиме одиничний масштаб.

Формулювання обмежень та визначення критеріїв оцінки кластеризації. Для скорочення ресурсів, що витрачаються на кластеризацію, необхідно ввести обмеження на можливе значення кількості кластерів, що формуються в внаслідок проведеної кластеризації. В результаті кластеризації отримуємо набір рішень, якість яких можна оцінити, використовуючи критерії, підбір яких можна здійснити на основі викладених вище рекомендацій. Основною частиною нечіткої кластеризації за ресурсами, що витрачаються. є виконання

алгоритму, за допомогою якого здійснюється кластеризація, послідовно здійснюється перебір значень із заданого діапазону кількості кластерів та обчислюються значення критеріїв, взятих для аналізу. Потім здійснюється вибір найкращого розбиття за допомогою аналізу набору екстремумів кожного з критеріїв, це буде основний результат нечіткої кластеризації - розбиття на кластери.

Залежно від поставлених цілей далі може бути виконана оцінка відхилень та проведена підготовка результати до класифікації. При задовільних підсумках аналізу, коли лікар-користувач підтверджує результат, отриманий згідно з цілями, поставленим для вирішення завдання, проведений аналіз завершується (при необхідності повторюється з певного етапу, після чого здійснюється ще одна ітерація). За допомогою представленого модифікованого методу нечіткої кластеризації у розробленій системі медичного призначення вирішується завдання визначення варіанта перебігу хвороби (опис якої більш детально буде наведено далі).

3 АНАЛІЗ РОЗВИТКУ ПРОЦЕСУ ЛІКУВАННЯ ПАЦІЄНТА НА ОСНОВІ НЕЧІТКИХ СЕМАНТИЧНИХ МЕРЕЖ

3.1 Вимоги до семантичних мереж в умовах нечіткості

Головною проблемою при побудові СППМР є проблема подання та використання медичних знань, якими володіють лікарі експерти (інженери знань), тобто. люди, які мають суттєвий і позитивний досвід під час вирішення завдань певного класу. Визначатимемо медичні знання як набір відомостей, що створюють цілісний опис, який відповідає деякому певному рівню поінформованості про описуваної медичної ситуації, медичної рекомендації, проблеми, питанні, методології лікування і т.д.[6].

На подання знань накладаються певні вимоги реальними медичними предметними областями [4], які включають безліч об'єктів, що мають велику різноманітність зв'язків (відносин); об'єкти і відносини мають композиційність і різнорідність.

Реальні Про є динамічними структурами. Ці особливості потребують вирішення проблем щодо засобів подання знань. По-перше, при виборі зовнішньої мови опису знань необхідно підвищити ступінь непроцедурності та одночасно забезпечити достатню ефективність при інтерпретації. По друге, при виборі внутрішньої мови уявлення знань про ПрО, що включає розробку способів представлення знань та засобів реалізації процесів, пов'язаних з формуванням та використанням знань, необхідно організувати зберігання знань, їх корекцію, аналіз, узагальнення та об'єднання [6].

По-третє,необхідно вирішити проблему розуміння, пов'язану з неповнотою знань, неточністю чи багатозначністю висловлювань, прихованістю структури та змісту знань, правом на помилку.

У роботі [1] наведено опис деяких моделей уявлень знань, таких як: логічні моделі [2,3], кадри [3], продукційні системи [4,], нейронні мережі [6],

когнітивні карти , семантична мережа, проаналізовано їх переваги та недоліки. Для медичної ПрО зручніше використовувати визначення семантичної мережі як довільного графа із зазначеними ребрами та вершинами. Ребра відповідають відносинам між поняттями, а вершини – це поняття із ПрО. І для зв'язків, і для понять можливі не лише опис кількох типів їх застосування, а й різні види уявлень. Переваги моделей, побудованих на основі такого підходу для медичної ПрО,наступні:

- орієнтація на проблему пошуку вирішення медичної ситуації – взаємозв'язку між об'єктами, що дозволяють визначати шлях доступу; підбір щодо об'єкта всієї наявної інформації медико-технологічного процесу, що дозволяє виділяти взаємовідносини, зміст, рівні деталізації та внутрішню структуру;
- наявність організаційних принципів – асоціативності та ієрархічності – дозволяє забезпечити підвищену гнучкість під час побудови моделей;
- узгоджене поєднання семантичного (що стосується цієї ПрО) та синтаксичного (структурного) описів медичних знань дає можливість порівняно однорідної структурі досить легко оновлювати ці знання
- семантичну мережу можна перетворити під час виконання певних умов у моделі подання знань іншого виду , це буде корисно при побудові нечітких когнітивних карток;
- природність поєднання процедурного та декларативного знання дає можливість розбивати на незалежні частини процеси обробки при представлення даних медико-технологічного процесу всередині системи .

Структура семантичної мережі для представлення об'єктів усередині інформаційна система є універсальною і впливає на складність реалізації математики всередині системи . Концентрація знань навколо понять та процедур,

пов'язаних з цими поняттями, є головною особливістю моделей для представлення медичних знань (рис. 3.1), таким чином, важливий принцип представлення знань у системах інтелектуального аналізу статистичних даних медико-технологічного процесу - це подання інформації у вигляді мережі, багаторівневої ієрархічно упорядкованої системи чи довільного графа, а не у формі масиву, як і звичайних системах .

Рисунок 3.1 - Склад моделі подання медичних знань

На основі аналізу способів моделювання уявлень знань був зроблено висновок, що найбільш універсальною для вирішення задачі інтелектуального аналізу статистичної інформації медико-технологічного процесу є модель подання знань на основі семантичної мережі, так як при невеликих змінах її легко перетворити на інші типи моделей. Для опису моделі медичних знань пропонується застосовувати універсальну алгебру [8].

У роботах [3] наведено опис блокової структури інтелектуальної системи підтримки прийняття медичних рішень, яка складається з наступних елементів: база даних, база знань, інтерфейс експерта, інтерфейс користувача, модуль набуття знань та модуль формування рекомендацій. Однак у медичній

сфері цю структуру необхідно розширити, увімкнувши два додаткові блоки: модель стану здоров'я пацієнта; модель медичних дій, які проводяться з пацієнтом. Також слід зазначити, що стандартні елементи слід скоригувати згідно з особливостями медичної предметної галузі. Для відображення знань про стан здоров'я пацієнта в інтелектуальній медичній системі необхідно спроектувати модель подання цієї інформації.

У роботах [6] запропоновано підходи до вирішення зазначеної проблеми. Однак для медичної сфери необхідно використовувати модель стану здоров'я пацієнта як модель актуального на даний момент стану знань про пацієнта та ситуацію для аналізу, являє собою «ідеальну» модель знань про пацієнта, що включає знання про медичну предметну область, когнітивні механізми та типові помилки. Відповідно до проблемної ситуації медичної інтелектуальної системи необхідно надати набір рекомендацій лікарю-користувачу, для цього необхідно побудувати модель медичних дій, що виробляються з пацієнтом [7]. У медичній сфері наочно її відобразити упорядкованою послідовністю, у якій кожен елемент є модель стану здоров'я пацієнта, що змінюється в залежності від дій, вироблених із нею (вибір ситуації чи знаходження набору рекомендацій). Завдяки такому підходу можна говорити, що відбувається прогнозування розвитку медичної ситуації, лікар-користувач може отримати рекомендації згідно з обраною ситуацією, подальший прогноз розвитку ситуації здійснюється за такою ж схемою.

Загальна схема послідовності дій, необхідні реалізації аналізу розвитку процесу лікування пацієнта на основі нечітких семантичних мереж в автоматизованій медичній інформаційній системі, представлена рис. 3.2. При нумерації блоків використовувався багаторівневий підхід, у якому перша цифра позначає номер блоку в алгоритмі вищого рівня, друга - порядковий номер блоку аналізованому алгоритмі.

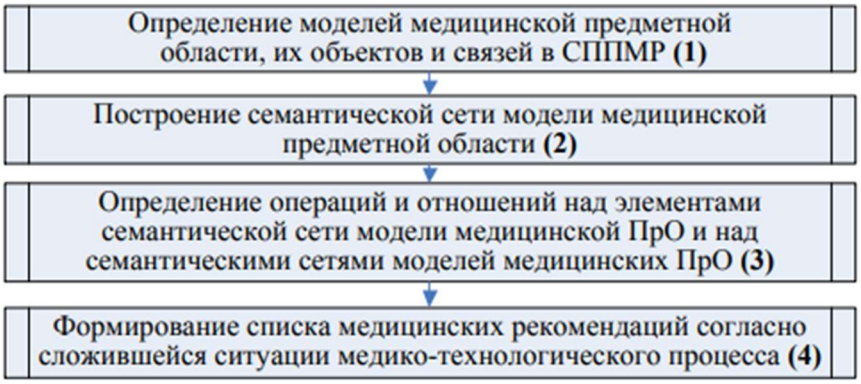

Рисунок 3.2 - Схема послідовності дій, необхідних для реалізації аналізу розвитку процесу лікування пацієнта на основі семантичних мереж у МІС.

3.2 Універсальна алгебра опису медичної предметної області на основі семантичної мережі

Найважливішим завданням при побудові моделі медичної ПрО є визначення об'єктів (елементів). Наприклад, для моделі медичної ПрО «Надходження пацієнта» об'єктами всередині моделі є «Медичні показники» - це узагальнене поняття, яке може представляти як результат конкретного медичного аналізу пацієнта, так і опис фізичного стану пацієнта. Для об'єктів має бути визначено спосіб їх уявлення у системі. Це різні параметри об'єкта, які зберігаються в моделі, наприклад, такі параметри, як «Найменування об'єкта» (обов'язковий атрибут), «Фізичний (або логічний) зміст», «Значення», «Одиниці виміру», «Клінічна важливість», «Норми» та ін. допускається відсутність деяких параметрів, якщо їх значення визначити неможливо або їх наявність не має значення для конкретного об'єкта. Коли список об'єктів у моделі медичної ПрО визначено, необхідно встановити зв'язок між об'єктами всередині моделі, визначити способи угруповання об'єктів, виділити головні об'єкти і т.д., таким чином, задати деяку структуру моделі медичної ПрО. Узагальнений алгоритм придбання медичного знання у системі представлений рис 3.3. Для побудови моделі медичної ПрО необхідно визначити, що буде об'єктом моделювання.

Для цього спочатку визначається список завдань, які вирішуватимуться в СППМР, що розробляється.

Рисунок 3.3 – Алгоритм набуття медичного знання у системі

Алгоритм визначення моделей медичних ПрО, їх об'єктів та зв'язків у СППМР представлений в рис. 3.4

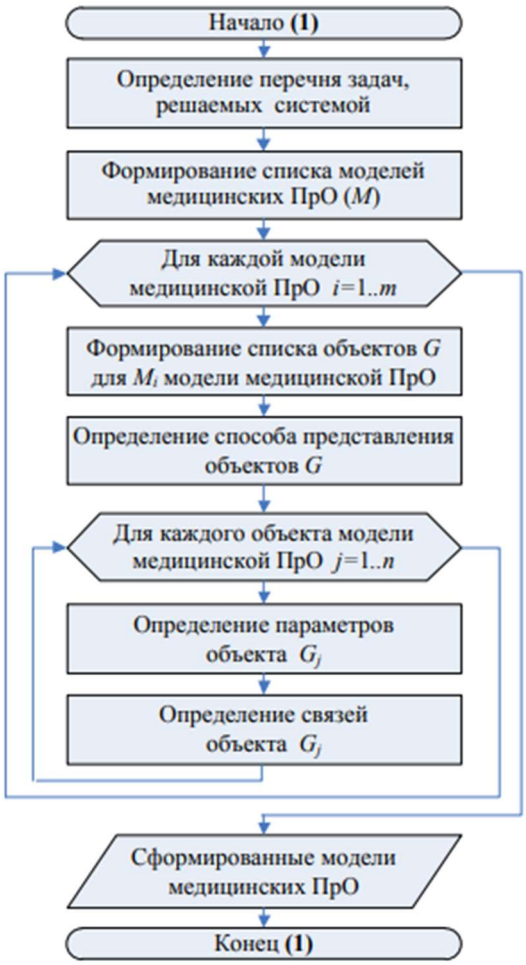

Рисунок 3.4 – Алгоритм визначення моделей медичної предметної області, їх об'єктів та зв'язків у СППМРУ проектованій системі є можливість створювати та редагувати кілька медичних ПрО.

Перелічимо назви деяких із них: модель ПрО «Надходження пацієнта», модель ПрО «Стан здоров'я пацієнта», модель ПрО «Медичні ситуації та рекомендації», модель ПрО «Схема лікування», модель ПрО «Медичний потік лікарських засобів» та ін. процесі функціонування система може одночасно працювати з декількома ПрО. Для моделювання медичної предметної галузі будемо застосовувати семантичну мережу з урахуванням універсальної алгебри. Розкриємо поняття універсальної алгебри, яка використовується для описи медичної предметної галузі Нехай частково визначена функція типу φ : $M^n \rightarrow M$ називається n -арною частковою операцією на безлічі M , де як безліч

М у медичній предметній галузі найчастіше використовуються різні масиви медичних даних (показники медичних приладів, біометричних датчиків, результати медичних аналізів, анамнез пацієнтів та ін.).

Іншим джерелом медичних масивів даних служить накопичена статистика як у роботі лікувального закладу цілому (регламентована звітність), а також результатів лікування конкретних пацієнтів (електронні історії хвороб).Також при описі моделі медичної предметної області є важливим джерелом масивів медичних даних служать регламентовані довідкові дані (довідник МКХ-Х, методи та способи лікування захворювань). У разі коли функція типу $\varphi : M^n \rightarrow M$ всюди визначено, говорять про n-арної операції на безлічі М ; n носить назву арності операції φ . Значення арності операцій у медичних предметних областях прийнято брати рівним двом, оскільки бінарних операцій буває достатньо для опису семантичної мережі медичної предметної галузі. Збільшення арності ускладнює побудову моделей та процес їх формалізації та обробки в системі. Безліч М спільно з сукупністю часткових операцій $\Omega = \{\varphi_1, \dots, \varphi_m, \dots\}$, заданою на ньому, тобто. система A ={M;Ω} носить назву часткової універсальної алгебри М називається несучою, або основною, безліччю алгебри А (або просто носієм алгебри). Якщо кожна з операцій, що належить сигнатурі Ω, визначена на всій множині М , кажуть просто про універсальну алгебру A ={M;Ω} . Типом алгебри називається вектор арностей її операцій, сигнатурою Ω - сукупність операцій . Безліч $M' \subset M$ вважається замкнутим по відношенню до n – арною операції φ на безлічі М , якщо $\varphi (M'^n) \subseteq M'$, іншими словами, коли значення φ на аргументах з M' належать M' . Якщо M' замкнуто щодо всіх операцій$\varphi_1, \dots, \varphi_m, \dots$ алгебри А , то система $A' = \{M', \varphi_1, \dots, \varphi_m, \dots\}$ називається подалгеброю А (при цьому $\varphi_1, \dots, \varphi_m, \dots$ розглядаються як операції на M').

Як сигнатура операцій Ω у медичній предметній галузі найчастіше використовуються стандартні операції, прийняті в універсальній алгебри з

урахуванням специфіки медичної галузі. Основна складність полягає у формалізації даних (медичне інформаційне простір є слабо структурованим, як наслідок складно піддається формалізації та подальшій обробці), що вимагає введення додаткові операції. Для побудови моделі медичної ПрО на основі нечітких об'єктів необхідно визначити функцію приналежності кожного об'єкта. Існують два підходи для вирішення цього завдання.

Перший підхід – завдання функції власності інженером зі знань спільно з лікарями-експертами, що є досить трудомісткою операцією, вимагає хороших знань і досвіду в кожній моделюваній ПрО. Розбиття ПрО на ряд суміжних ПрО полегшує вирішення задачі. Залучення кількох лікарів-експертів зменшує суб'єктивний чинник у оцінках. Кількість експертів впливає якість оцінок. Крім того, при розрахунку загального показника можна враховувати ранг кожного експерта та здійснювати зважування його оцінки. За такого підходу розрахунок характеристичної функції власності здійснюється за такою формулою:

$$\mu_{\tilde{A}}(x_i) = \frac{1}{m_э}\sum_{j=1}^{m_э} w_j \mu_j(x_i), \qquad (3.1)$$

де $\mu_j(x_i)$– оцінка характеристичної функції приналежності j-го лікаря експерта для x_i параметра об'єкта моделі медичної ПрО, вид та параметри; w_j - ранг j-го лікаря-експерта, тобто. число, яке характеризує значимість елемента у формуванні властивості, що описується нечітким термом.

Допустимо, що виконується правило: що більше ранг елемента, то більший ступінь приладдя. Ранг лікаря залежить від таких параметрів: кваліфікації, наявності наукового ступеня, стажу роботи та ін. Вигляд та параметри функції приладдя визначається на основі досвіду та різних додаткових припущень. Їхній вид докладно розглянутий у [5]. Вибір функції приладдя здійснено у розділі 1. Інженер зі знань при цьому перетворює отриману інформацію до певного формату та здійснює введення даних у систему.

Алгоритм визначення $\mu_{\widetilde{A}}(x_i)$ на основі знань лікаря експерта та інженера з знань представлений на рис. 3.6, а. Другий підхід - завдання функції належності на основі статистичних даних медико-технологічного процесу, що характеризують обрану медичну ПрЗ. Для реалізації цього підходу необхідно вибрати медичну ПрО та з медичної бази даних вибрати дані медико-технологічного процесу, що належать лише до обраної ПрО. Оскільки за рік у диспансері проходять лікування щонайменше 2500 пацієнтів, то вибірка для кожної ПрО області за 2005 – 2015 роки. становила 10 тис. випадків. Для кожного об'єкту x_i моделі медичної ПрО з кожного випадку статистики проводиться розрахунок N_{x_i} (вважається той об'єкт, у якому зберігаються дані), потім сума ділиться на потужність безлічі випадків. Тоді розрахунок характеристичної функції власності здійснюється за формулою :

$$\mu_{\widetilde{A}}(x_i) = \frac{1}{m_{сл}} \sum_{j=1}^{m_{сл}} a \text{ при а} \begin{cases} 1, \text{якщо } x_i \text{ заповнено}, \\ 0, \text{якщо } x_i \text{ не заповнено} \end{cases}, \qquad (3.2)$$

Де $m_{сл}$ – кількість випадків з статистики.

Алгоритм визначення $\mu_{\widetilde{A}}(x_i)$ на основі статистичних даних медико-технологічного процесу представлений на малюнку 3.6,б. Можливе застосування змішаного підходу. В цьому випадку розраховані статистичним способом характеристичні функції приналежності $\mu_{\widetilde{A}}(x_i)$ як підказка (для вибору метрики оцінки) можна надавати лікарям-експертам. Потім два отриманих безлічі функцій приналежності (першим та другим способом) усереднюються для формування результуючої множини, яка використовується під час побудови семантичної мережі обраної медичної ПрО.

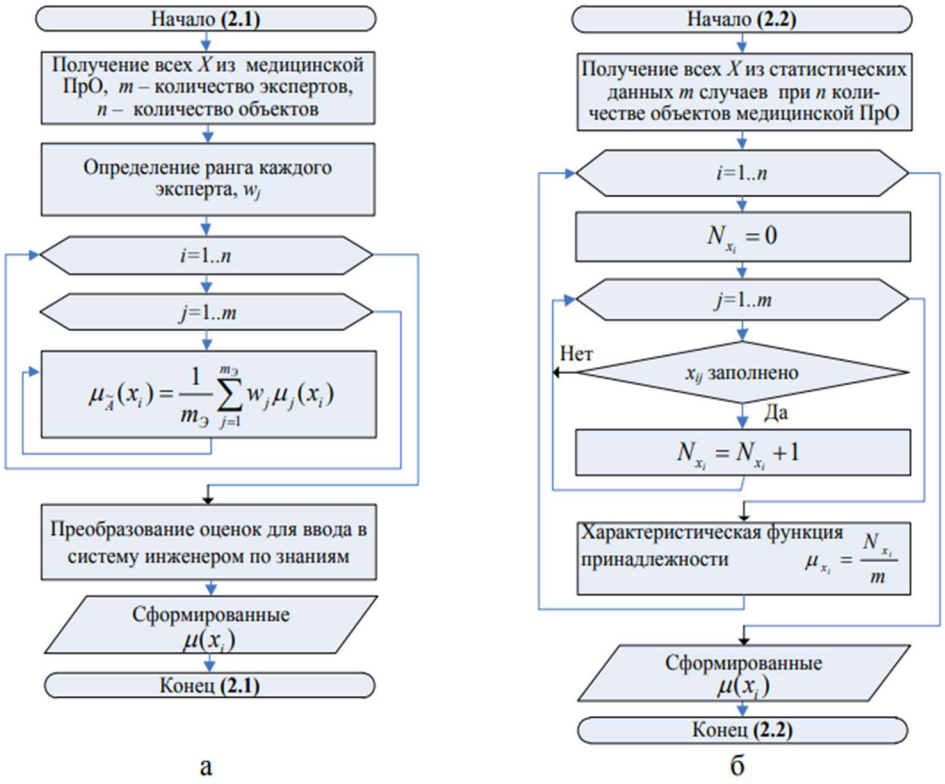

Рисунок 3.6 – Алгоритм визначення характеристичної функції приладдя $\mu_{\tilde{A}}(x_i)$: а – на основі знань лікарів-експертів та інженера з знанням, б – на основі статистичних даних медико-технологічного процесу.

За такого підходу розрахунок підсумкової характеристичної функції приладдя здійснюється за такою формулою

$$\mu_{\tilde{A}}(x_i) = \frac{\left(\mu_{\tilde{A}}^{Cm}(x_i)+\mu_{\tilde{A}}^{э}(x_i)\right)}{2}, \qquad (3.3)$$

Де $\mu_{\tilde{A}}^{Cm}(x_i)$ - Функція приналежності, розрахована першим способом, $\mu_{\tilde{A}}^{э}(x_i)$ – функція власності, розрахована другим способом.

При змішаному підході у разі недостатнього обсягу вибірки статистики можливий розрахунок, за якого статистика виступає в «ролі» одного з лікарів-експертів та використовується при обчисленні з певним рангом (wj). З понят-

тям множини пов'язане поняття нечіткої підмножини, яке дає можливість досліджувати певні поняття, застосовуючи математичні структури Нечітка логіка дозволяє ширше використовувати логічні операції, включаючи розширені оператори нечіткого безлічі, які використовуються для маніпулювання лінгвістичними змінними [2], що є необхідним при побудові складних ПрО, до яких належать медичні ПрВ. Використовуючи нечіткі множини, можна описати нечіткі об'єкти медичної ПрО.

Для створення інтелектуальної СППМР найкращі результати показує об'єктно-орієнтований підхід [2], згідно якому всі процедури обробки інформації медико-технологічного процесу та всі дані містяться в одному об'єкті . високого ступеня точності під час формулювання вимагають об'єкти. Зокрема, діапазон значень та Значення атрибутів визначаються з високим ступенем впевненості.

Для проектування найбільш адекватної моделі медичної ПРО потрібні об'єкти, що характеризуються невизначеністю та неточністю, особливо в тих випадках, коли проводиться моделювання СППМР. Проблему моделювання систем, що характеризуються невизначеним описом значень атрибутів можна вирішити за допомогою нечітких об'єктів. Нечіткий об'єкт у медичній предметній галузі – це стандартний об'єкт, який розширено для невизначених моделей та даних. Нечітким об'єктом у медичній сфері можна уявити такі об'єкти моделі предметної галузі, як: результати медичних аналізів, анамнез пацієнта та інші, значення яких можуть бути отримані з деякою похибкою, так і мати випадкові навмисні чи ненавмисні помилки. Нечітке уявлення значення необхідного атрибуту об'єкта – це розширення, уявлення використовує нечітку множину там, де це необхідно та можливо. Атрибути об'єктів (медичні показники) представляються як композиції чітких і нечітких уявлень. Нечіткі об'єкти інтегрують можливості нечітких та об'єктно-орієнтованих понять, що забезпечують механізм проектування та моделювання інтелектуальних медичних систем Базові принципи, що застосовуються при об'єктно-орієнтованому

підході, зберігаються в нечітких об'єктах, зокрема узагальнення, успадкування та інкапсуляція .

3.3 Застосування семантичної мережі під час опису медичної предметної області

У загальному вигляді семантичну мережу для опису медичної ПрО можна подати набором концептуальних графів , кожен з яких формується з урахуванням логічної формули. Аргументи та імена предикатів представляються у ньому парою типів вузлів. Дуги графа (логічні зв'язки між медичними показниками, що визначаються експертним шляхом та специфікою медичної предметної області) виробляють поєднання аргументів предикатів зі своїми іменами. Покладена основою концептуального графа логіка предикатів - це мова, що інтерпретується в термінах медичної ПрО міркувань (фрази метамови представляють логічні формули). Аргументи логічних функцій та предикатів використовуються для представлення подій, станів та атрибутів. Спосіб об'єднання цих понять вказують імена предикатів. При відображенні концептуального графа кругами видаються імена предикатів, а прямокутниками – аргументи (наприклад: «Аргумент 1») результат загального аналізу крові, «Аргумент 2» – результат ФЛГ).

Якщо прямокутник з'єднується стрілкою з колом, то вони представляють аргумент і ім'я того самого предикату (рис. 3.7). У предикатів може бути кілька аргументів, за такого підходу в імені предикату може бути кілька вихідних та/або вхідних стрілок (рис. 3.7,а), а предикат буде представлятися логічною формулою. Для представлення знань можна використовувати бінарні предикати, які володіють парою аргументів. Тоді предикат матиме дві стрілки: вихідну та вхідну (рис. 3.7,б). Побудована на основі концептуальних графів семантична мережа буде використана для семантичного та синтаксичних досліджень даних медико-технологічного процесу [14].

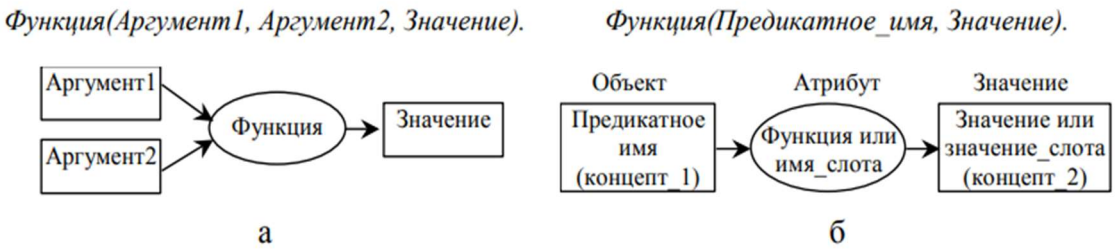

Рисунок 3.7 – Приклади концептуальних графів: а – загальний вигляд концептуального графа; б – концептуальний граф бінарного предикату.

Як структурну модель довготривалої пам'яті будемо застосовувати модель, запропоновану в роботах Куілліана (уявлення про сенс знання, розуміння мови), - Teachable Language Comprehender [4]. =

Як спосіб відображення семантичних відносин між різними концептами ПрО він використовував мережну структуру для опису довгострокової пам'яті. Завдання цієї моделі – імітувати природне розуміння та використання мови людиною. Основу цієї моделі становить уявлення значень класу, до якого відносяться об'єкт, його зв'язки (встановлені зі словами) та відображення властивостей об'єкта та прототипу цього об'єкта. Наприклад, результат проведеного лікування пацієнта, який безпосередньо залежить від знань, досвіду, уміння та кваліфікації лікаря, може бути чисельно оцінений [0,1] як «біль пройшла = 1»; «Біль зменшився = 0,5»; "Біль пройшов = 0" і т.д.

Асоціативними мережами, що включають до свого складу вершини, які представляють концепти (ситуації або рекомендації), та дуги, які показують відносини між концептами, у моделі видаються концептуальні об'єкти медичної ПрВ. Розглядати медичну ПрО як сукупність об'єктів ПрО та зв'язків між ними можна за допомогою апарату семантичних мереж на основі ієрархічної структури. Ієрархічну семантичну мережу можливо розділити на підмережі та визначити відносини як між вершинами всередині мережі, так і між підмережами.

Оскільки при вирішенні певних медичних задач виникає потреба використовувати кілька моделей ПрО (наприклад, при виконанні кластеризації використовуються одразу дві моделі медичних ПрО: «Надходження пацієнта» та «Стан здоров'я пацієнта»), зв'язок між семантичними мережами є важливою особливістю та необхідним функціоналом при побудові СППМР. Для опису медичної ПрЗ в інтелектуальній медичній системі аналізу статистичної інформації застосовується семантична мережа, заснована на універсальній алгебрі, яка описана трійкою: A = S,O,R де S – безліч семантичних мереж, які репрезентують моделі реальних ПрВ; O - безліч операцій семантичної мережі на S ; R – безліч відносин семантичної мережі на S. У проектованій інтелектуальній медичній системі аналізу статистичної інформації семантична мережа, яка відповідає моделі медичної ПрО, є двійкою наступного виду:

$$S_{\text{ПрО}} = \{G, U\}, \tag{3.4}$$

де G – безліч об'єктів ПрО (ситуація для аналізу та набір рекомендацій); U – безліч дуг, що здійснюють зв'язок об'єктів ПрО. Кожна з дуг являє собою відносини між ситуаціями або взаємний зв'язок між ситуаціями (вказується ступінь залежності однієї ситуації від іншої), а також взаємозв'язок ситуацій та дій із медичної ПрО (рис. 3.8). Рекомендації можна інтерпретувати як процедуру прогнозу перебігу хвороби.

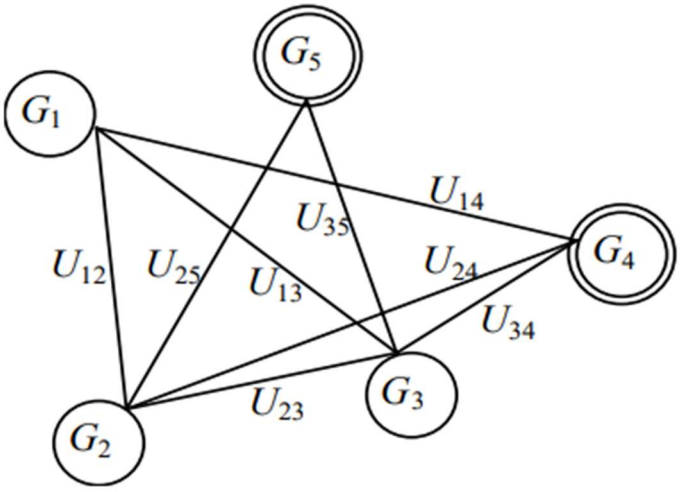

Рисунок 3.8 – Семантична мережа моделі медичної ПрО

Наприклад, на рис. 3.8 як G_1, G_2, G_3 виступають ситуації для аналізу (виявлено стійкість до 1-ї групи антибіотиків, пацієнту поставлено первинний діагноз, підвищення температури), G_4, G_5 – рекомендації згідно ситуації для аналізу (призначення антибіотиків 2-ї групи, призначення антибіотиків 1-ї групи). Об'єктна асоціація, яка встановлює зв'язки між об'єктами, складає базис об'єктно-орієнтованого моделювання. Стандартна об'єктно-орієнтована методологія використовує чіткі зв'язки між об'єктами, але вона не може бути використана для опису відносин між об'єктами в медичній предметній галузі, які встановлюються з застосуванням різних ступенів залежності. Як нечіткі зв'язки між об'єктами медичної ПрО (наприклад, підвищення температури, призначення антибіотиків 1-го ряду, призначення антибіотиків 2-го ряду тощо) розглядаються типи градуйованих зв'язків. Задамо пару нечітких об'єктів:

$$\widetilde{A} = \{x_i, \mu_{\widetilde{A}}(x_i) | x_i \in \psi_1, 1 \le i \le n\}, \quad (3.5)$$

$$\widetilde{B} = \{y_i, \mu_B(y_i) | y_i \in \psi_2, 1 \le i \le m\}, \quad (3.6)$$

Тоді ставлення нечітких об'єктів можна записати як:

$$R_{f(\widetilde{A},\widetilde{B})} = \left\{(x_i, y_i), \mu_{f(\widetilde{A},\widetilde{B})}(x_i, y_i)\right\}, \quad (3.7)$$

Де $(x_i, y_i \in \psi_1 \times \psi_2)$($\psi_1$модель медичної ПрО). Нечіткі вирази відносин цього типу широко використовуються в нечіткому міркуванні і часто згадуються як композиційні правила логічного висновку.

Для побудови моделі медичної ПрО на основі семантичної мережі для кожного нечіткого об'єкта необхідно визначити тип об'єкта та ступінь залежності між поняттям, що входить до його складу, та об'єктом. Алгоритм визначення ступеня залежності понять всередині об'єкта та типу об'єкта для семантичної мережі моделі медичної ПрО представлено рис. 3.10. Визначення списку поняття відбувається на момент розробки семантичної мережі моделі предметної області (рис. 3.4) виходячи з думки групи експертів та на основі аналізу накопиченої статистики в інформаційної системи (приклад деяких ситуацій та рекомендацій, а також списку понять, що входять до них, наведено на малюнку 3.11).

Рисунок 3.10 – Алгоритм визначення ступеня залежності понять усередині об'єкта для семантичної мережі моделі медичної ПрО

Для цього лікарі-експерти та інженер із знань встановлюють ці залежності для кожного об'єкта з моделі медичної ПрО та визначають його тип. Тип об'єкта та їх кількість залежать від конкретної ПрО, наприклад для моделі медичної ПрО «Медичні ситуації та рекомендації» визначено два типи об'єкта («медична ситуація» та «рекомендація»).

Якщо виходити із запропонованої схеми побудови вузла семантичної мережі інтелектуальної медичної системи аналізу статистичної інформації, залежність між вузлами будемо будувати на основі зв'язків між поняттями, належать об'єктам моделі ПрО.

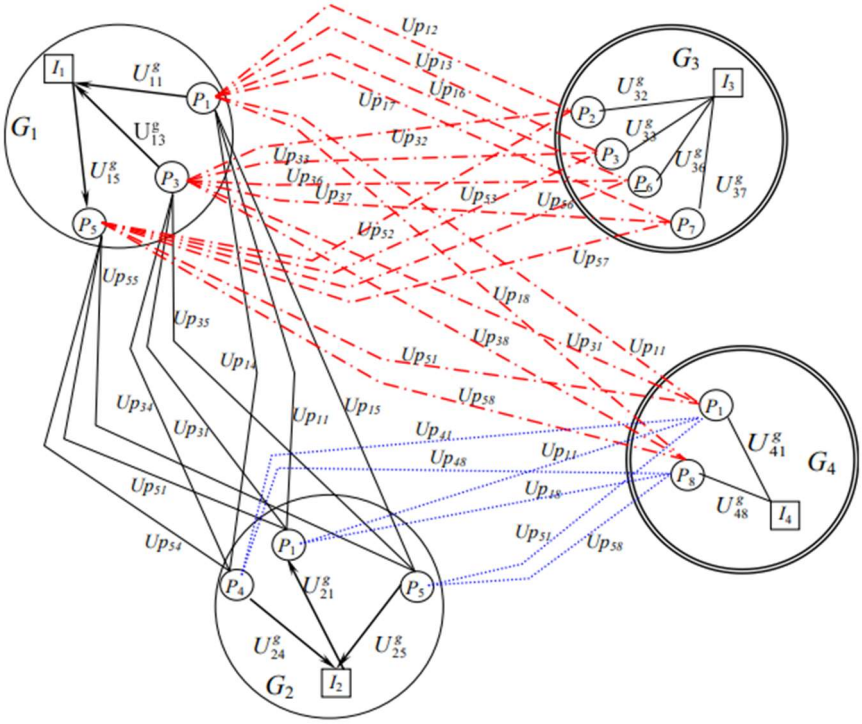

Рисунок 3.11 – Фрагмент представлення семантичної мережі моделі ПрО

Представимо нечітке відношення:

$$\mathrm{Up_{ij}} = \mu_s(\mathrm{P_i}, \mathrm{P_j}), \tag{3.8}$$

Це ставлення визначає міру близькості між поняттями. На його базі утворюється нечітка підмножина Up :

$$\mathrm{Up} = \{\mathrm{P_i}, \mathrm{P_j}, \mu_s(\mathrm{P_i}, \mathrm{P_j}) | \mathrm{P_i}, \mathrm{P_j} \in \mathrm{P}; \mathrm{i}, \mathrm{j} = 1..\mathrm{N}\}, \tag{3.9}$$

Де N – кількість у кожній моделі медичної предметної області інтелектуальної СППМР обмежено 70 понять, оскільки велике кількість призводить до ускладнення моделі та роботи з нею. Фрагмент отриманої семантичної мережі наведено рис. 3.9. Наприклад, розглянемо дві ситуації для аналізу (рис. 3.11), де інформаційною частиною є опис ситуації та список ключових понять [3], характерних для цієї ситуації, і два набори дій з ПрО, яких інформаційною

частиною є назва набору та список ключових понять, що характеризують цей набір.

Інтелектуальна медична система аналізу статистичної інформації орієнтована на роботу з кількома медичними моделями ПрО, які не пов'язані або пов'язані між собою. Це продиктовано особливістю завдань, на які орієнтована СППМР, оскільки кожна з моделей медичних ПрО будується на базі окремої семантичної мережі, які потім можна поєднувати в єдину модель медичної ПрО. Таке об'єднання може здійснюватися для двох та більше моделей медичних ПрО.Для об'єднання моделей медичних ПрО необхідно визначити безліч операцій та безліч відносин між семантичними мережами та між елементами семантичної мережі Перш ніж визначити безліч операцій (O) і безліч відносин (R) між семантичними мережами, необхідно визначити безліч операцій та безліч відносин між елементами семантичної мережі. Визначення операцій між елементами обґрунтованої семантичної мережі включає низку базових операцій.

4.АНАЛІЗ МОДЕЛІ ТА АЛГОРИТМА НЕЧІТКОЇ КЛАСТЕРИЗАЦІЇ

4.1 Обрання основних методів кластерізації

В якості основи розроблених алгоритмів були обрані процедури прямої нечіткої кластеризації, що дозволяють працювати з класами, що перетинаються, використовують апріорні відомості про класи і еталони. Як міру близькості використані метричні методи: евклідова відстань і діагностика по відстані до еталона як найбільш загальний випадок.

Одним з основних методів виділення кластерів є метод, заснований на мінімальній відстані від об'єкта (точки) до всіх точок класу. Нехай K - число класів станів об'єкта класифікації і кожен j-клас (j = 1, ..., K) характеризується в n-мірному діагностичному просторі навчальною вибіркою – набором векторів $\{X^j\}, X^j = (x^j, \dots x_i^j, \dots, x^i)$.

Результатом застосування методу кластеризації до навчальних вибірок є еталонні кластери – області в ознаковому просторі.

Точка X* належить j -му кластеру, якщо відстань від точки X * до точок кластера Dj менше, ніж відстань до точок інших кластерів.

Цей метод був покладений в основу одного з алгоритмів нечіткої кластеризації та діагностики. При його реалізації вхідними даними служить набір точок навчальних послідовностей K класів, при цьому як вхідні дані можуть використовуватися набори раніше класифікованих точок.

Кожен клас характеризується кількістю точок у класі.

Якщо ініціалізація раніше не проводилася, то виконується ініціалізація, яка полягає в додаванні в кожен вихідний клас по одному елементу з вихідної вибірки. У цьому кожному за класу величина d встановлюється рівної нулю, величина Count – одиниці.

Якщо ініціалізація раніше була проведена, виконуються такі дії:

- для кожного класу обчислюється максимальна відстань між усіма парами точок у кластері, а для кожної поточної точки з вихідних даних обчислюються відстані від цієї точки до всіх точок кожного класу;
- з отриманого набору відстаней вибирається номер класу, для якого ця відстань мінімально, і проводиться порівняння цієї відстані з максимальною відстанню в класі d, і якщо вона більша, то d необхідно перевизначити (установити d дорівнює максимальному відстані) ;
- число елементів у вибраному класі збільшується на одиницю і підраховується нову максимальну відстань між елементами класу.

Розв'язання задачі діагностики з використанням даного алгоритму кластеризації здійснюється в такий спосіб.

Вхідними даними можуть бути кластери, сформовані шляхом обчислень, проведених відповідно до алгоритму кластеризації, або одна або більше точок, що підлягають діагностиці. Вихідними даними є масив з K елементів, де K – кількість кластерів. У процесі діагностування для кожної діагностованої точки обчислюється максимальна відстань від неї до всіх точок кожного кластера і далі визначається номер класу, для якого ця відстань є мінімальною, на основі чого робиться висновок про належність точки даного класу

2. Інший спосіб побудови кластера, що є окремим випадком методу динамічних згущень, полягає в наступному. Точка X* належить j-му кластеру, якщо відстань від точки X* до центра Cj кластера Dj менше, ніж відстань до центрів інших кластерів:

$$l(X^*, C_j) < l(X^*, C_k), k = 1, \dots, K, j \neq k, \quad (4.1)$$

При цьому кластеризація полягає в такому розбитті безлічі об'єктів на заздалегідь задану кількість класів K, щоб мінімізувався функціонал

$$W = \sum_{j=1}^{K} W_i \, , \; W_i = \sum_{X^* \epsilon D_j} l^2(X^*, C_j), \tag{4.2}$$

Відмінністю алгоритму кластеризації, що реалізує метод формування кластера за формулою (4.2), від попереднього є тільки дії, що випливають з його особливостей і виконуються за наявності раніше проведеної ініціалізації. І тут першому кроці кожної поточної точки з вихідних даних обчислюються відстані від цієї точки до центрів кожного класу, причому початковими центрами вважаються перші точки кожного класу. Далі з отриманого набору відстаней вибирається номер класу, котрого це відстань мінімально. Число елементів у вибраному класі збільшується на одиницю, а величина d перевизначається за формулою (4.2). На етапі вирішення завдання діагностики для кожної точки, що діагностується, обчислюються відстані до центрів кожного кластера і далі визначається номер класу, для якого відстань є мінімальним, на підставі чого робиться висновок про належність точки даного класу. 3. Третій спосіб побудови кластера заснований на методі формування кластера, при якому точка X * належить j-му кластеру, якщо середня відстань від точки X * до точок кластера Dj менше серед- його відстані до точок інших кластерів:

$$l(X^*, D_j) < l(X^*, D_k), k = 1, \dots, K \, , j \neq k, \tag{4.1}$$

При вирішенні діагностичної задачі вхідними даними є кластери, сформовані відповідно до алгоритму (4.4), а також точки, що підлягають діагностиці. У процесі обчислень для кожної точки, що діагностується, обчислюються середні відстані до всіх точок кожного кластера, на підставі чого визначається клас, для якого відстань є мінімальною, і таким чином робиться висновок про

належність точки даного класу. Метод статистичного моделювання з використанням навчальної та тестової (контрольної) вибірок при різних варіантах взаємного розташування еталонних класів і класифікованого об'єкта дозволяє оцінити ефективність розроблених алгоритмів класифікації. Достатнім є проведення обчислювальних експериментів для випадків апріорної незалежності та апріорної залежності класів (в останньому випадку класи мають перетин, діагностований об'єкт потрапляє в область їх перетину). При організації обчислювального експерименту генерувалися штучні вибірки з різними математичними очікуваннями, дисперсією, кількістю сигналів у вибірці, формою класу, тобто. математична модель передбачала зміну обсягу, розмірності навчальних вибірок, форми та дисперсії («розмитості») класів. Очевидно, що на якість класифікації значно впливає ступінь перетину класів.

Для оцінки області перетину класів інформативним параметром є обсяг перетину нечіткої множини класів Vпер. Нехай визначено функцію приналежності класу (нечіткої множини) μ(x). Тоді обсяг цього множини визначається величиною $V = \iint\ldots\int \mu(x)dx$, *D де D* - область, яка охоплює всі точки даної множини. Враховуючи що

$$\mu(x) = \mu^2\left(x^{(2)}\right)\mu_3\left(x^{(3)}\right)\ldots\mu_n\left(x^{(n)}\right), \tag{4.5}$$

де $i = 2,3,\ldots,n$ – число діагностичних ознак, то обсяг, який займає клас, буде визначатися величиною

$$V = \iint\ldots\int \mu_2\left(x^{(2)}\right)\mu_3\left(x^{(3)}\right)\ldots\mu_n\left(x^{(n)}\right), dx^{(2)}\, dx^{(3)}\ldots\, dx^{(n)}, \tag{4.6}$$

Обсяг перетину K класів визначається за формулою

$$V_{пер} = \iint\ldots\int \mu_2\left(x^{(2)}\right)\mu_3\left(x^{(3)}\right)\ldots\mu_n\left(x^{(n)}\right), \tag{4.7}$$

де D – загальна область, яку займають K класи. Припустимо, що кожен із класів D_j , j 1,...,K характеризується навчальною вибіркою, що складається з N_j об'єктів (точок), і при використанні визначенного алгоритму кластеризації для кожного з класів визначено M j об'єктів.Отримані результати розпізнавання можна подати у вигляді матриці розміром $K \times K$, елементи якої aij являють собою число точок класу Di, віднесених до класу Dj. Природно, що a_{jj} – це число правильно розпізнаних точок класу Dj. Загальне число точок класу D дорівнює $\mathrm{N_i} = \sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{a_{ij}}$тоді коефіцієнт правильного розпізнавання для Dj класу визначається формулою $\mathrm{k_j} = \frac{\mathrm{a_{jj}}}{\mathrm{N}}, 0 \le \mathrm{k_j} \le 1$, в ідеальному випадку k_j 1 Якщо в процесі розпізнавання використаний лише один алгоритм класифікації, то якість алгоритму можна визначити як $\eta = \sum_{j=1}^{K} K_j \Big/ K$.Визначивши обсяг Vj, що займає клас Dj , по формулі $V_j = \int \mu_j(x) d_x$, де $\mathrm{D_j'}$ - область n-мірного простору ознак; $_j(x)$ - n -вимірна функція приналежності Dj класу, відмі- тим, що певна за цією формулою величина об'єму Vj є мірою класу Dj . Загальний обсяг, займаючий усіма класами .Тоді оцінка якості розпізнавання визначається як $\eta \le V \Big/ \sum_{J=1}^{K} V_j$,величина $V \Big/ \sum_{J=1}^{K} V_j = 1$ тіки якщо усі класи D , j =1,2,...,K , не перетинаються між собою. Якщо значення різниці $V \Big/ \sum_{J=1}^{K} V_j - \eta$ достатньо велике то це свідок того що функція власності обрана невдало і відображає адекватно реальних даних.

Приклад результатів кластеризації, проведеної за другим алгоритмом за умови, що число сигналів у кожному класі N 500, число діагностичних ознак n 2, наведено на рис.4.1

Аналіз отриманих результатів дослідження алгоритмів класифікації на модельних даних дозволив зробити такі висновки.

Алгоритми класифікації, засновані на різних критеріях відстані, за результатами моделювання практично мало відрізняються один від одного, але

при цьому мають різну швидкість: мінімальний машинний час у другого алгоритму, максимальне - у третього алгоритму. Виходячи з цього при виборі з цих алгоритмів віддано перевагу другому алгоритму.

Показники ефективності всіх алгоритмів покращуються зі зростанням обсягу N навчальної вибірки. Зі зростанням N спостерігається тенденція зниження частки недостовірно розпізнаних точок. Така поведінка помилки розпізнавання обумовлена наступним: чим повніше представлена навчальна вибірка, тим вище щільність точок, які відповідають об'єктам різних класів у діагностичному просторі. Межі класів стають більш вираженими, а отже, класифікація об'єкта проходить успішніше.

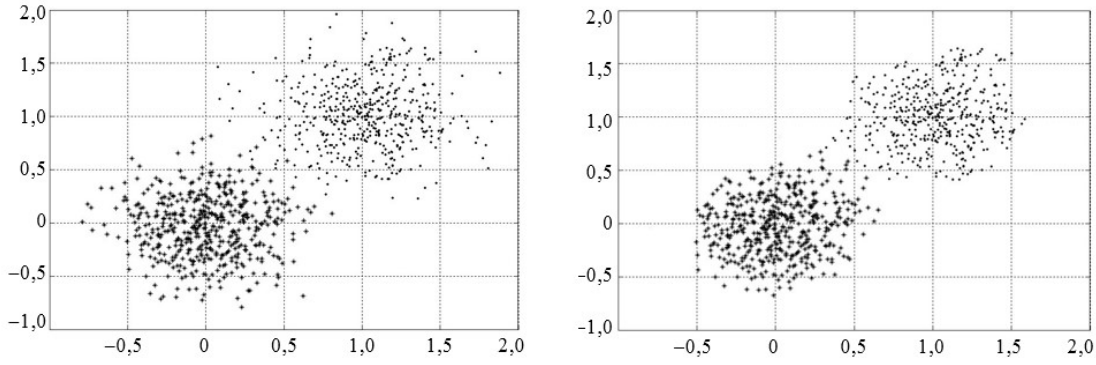

Рисунок 4.1 – початкова множина точок та результат кластерізації

Зазначена закономірність зберігається незалежно від величини дисперсії еталонних класів. При меншій дисперсії класів якість розпізнавання закономірно вища, але розроблений алгоритм нечіткої класифікації передбачає його застосування в умовах часткового перетину класів.

4.2 Вибір програмних засобів

Для напису алгорітму буде використана мова програмування Python, його дистрибутиви, бібліотеки та програмні забезпечення

Python - це інтерпретована мова програмування високого рівня з акцентом на читабельність коду і простоту використання. Вона була розроблена у 1991 році Гвідо ван Россумом і широко використовується у різних сферах, включаючи веб-розробку, наукові обчислення, штучний інтелект, аналіз даних і багато іншого

Anaconda - це платформа та дистрибутив Python, який призначений для розробки, тестування та розгортання програмного забезпечення. Вона включає в себе Python-інтерпретатор, багато популярних бібліотек для аналізу даних та наукових обчислень, а також набір інструментів для управління пакетами і віртуальними середовищами.

Scikit-learn (також відомий як sklearn) - це відкрите програмне забезпечення для машинного навчання, яке надає широкий спектр алгоритмів і інструментів для розробки моделей машинного навчання. Він побудований на основі популярних мов програмування Python і NumPy, SciPy та matplotlib, і надає простий та зручний інтерфейс для використання і реалізації алгоритмів машинного навчання.

Jupyter Notebook є інтерактивним середовищем для розробки та виконання коду, аналізу даних, візуалізації та спільної роботи у мові програмування Python та інших мовах програмування. Він дозволяє поєднати код, текст, графіки та інші елементи у одному документі, що дозволяє легко створювати динамічні звіти, дослідження та презентації.

4.3 Розробка алгоритму з використанням згенерованих даних

Приклад з використанням генеріруємих данних. Нечітка с-означає, що кластеризація виконується за допомогою skfuzzy.cmeans, а вихідні дані цієї функції можна перепрофілювати для класифікації нових даних відповідно

до обчислених кластерів (також відомих як передбачення) за допомогою skfuzzy.cmeans_predict.Приведена генерація тестуємих даних (додаток А) на рис 4.2 приведена візуалізація згенерованих даних.

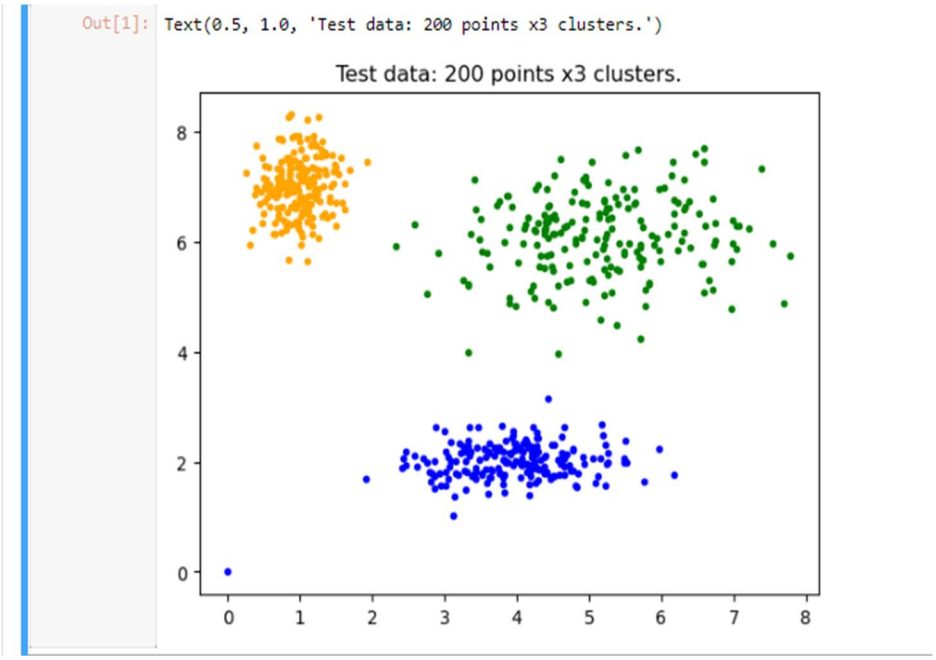

Рисунок 4.2 – візуалізація згенерованих даних

Наступним кроком буде налаштування цикла і побудува графіка для кластерізування даних кілька разів від 2 до 9 кластерів. Налаштування циклу приведене в додатку Б. Результат кластерізування даних кілька разів від 2 до 9 кластерів показаний на рис. 4.3.

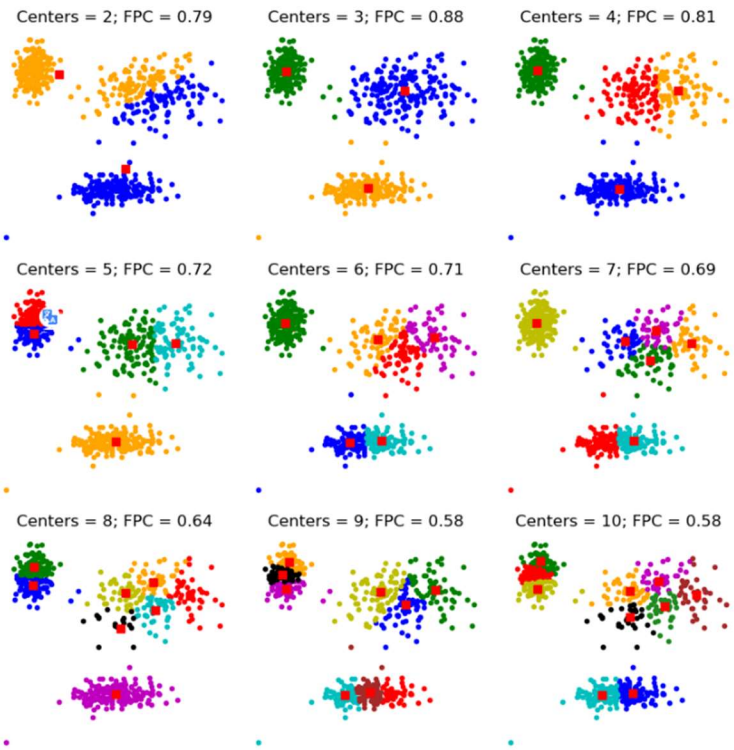

Рисунок 4.3 – кластерізування даних

Наступним кроком йде Нечіткий коефіцієнт розподілу(FPC),визначається в діапазоні від 0 до 1, де 1 є найкращим.

Це показник, який говорить нам, наскільки точно наші дані описує певна модель. Далі ми кластеризуємо наш набір даних, який, як ми знаємо, має три кластери, кілька разів, від 2 до 9 кластерів. Потім ми покажемо результати кластеризації та побудуємо графік нечіткого коефіцієнта розподілу. Коли FPC максимізовано, наші дані описуються найкраще.

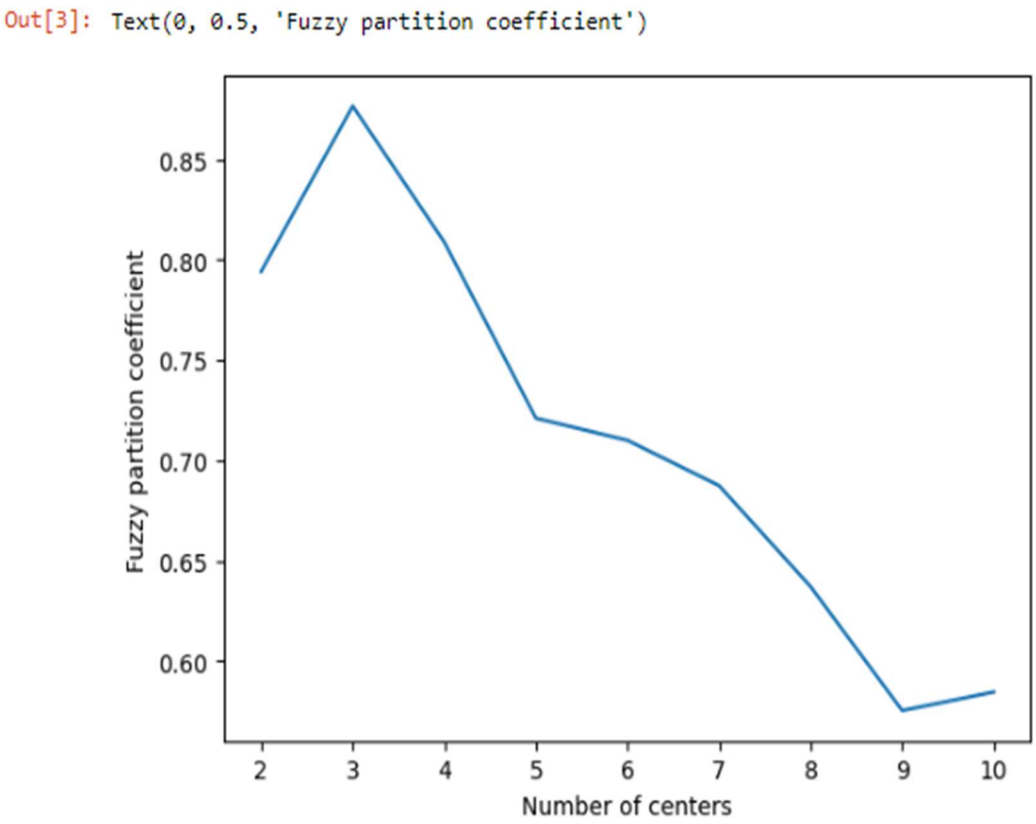

Рисунок 4.4 - графік нечіткого коефіцієнта розподілу

Як ми бачимо на рис.4.4, ідеальна кількість центрів — 3. Наявність FPC може бути дуже корисною, коли структура ваших даних незрозуміла.

Зверніть увагу, що ми почали з *двох* центрів, а не з одного; кластеризація набору даних лише з одним центром кластера є тривіальним рішенням і за визначенням поверне FPC == 1.

Класифікація нових даних та підгонка нових точок до існуючої моделі. Це відомо як передбачення. Для класифікації потрібна як існуюча модель, так і нові дані.

Для побудови ,ми знаємо, що наша найкраща модель має три центри кластерів. Ми перебудуємо 3-кластерну модель (рис.4.5) для використання в прогнозуванні, створимо нові уніфіковані дані та передбачимо, до якого кла-

стера належить кожна нова точка даних. Регенерація нечіткої модель із 3 центрами кластерів - зауважте, що порядок центрів є випадковим у цьому алгоритмі кластеризації, тому центри можуть мінятися місцями.

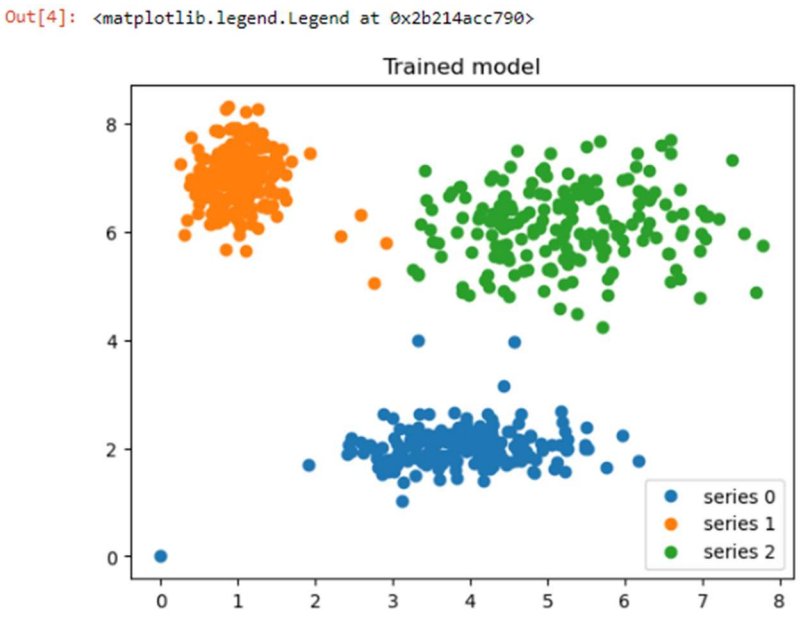

Рисунок 4.5 - 3-кластерна модель

В прогнозуванні ми генеруємо однорідні дані для цього поля та класифікуємо їх за допомогою cmeans.predict включаючи їх у вже існуючу модель яке приведене в додатку В.

Генерування даних з рівномірную вибіркою в діапазоні [0,10] у x і y : newdata = np.random.uniform(0, 1, (1100, 2)) * 10

Передбачення нового членства в кластері за допомогую cmeans.predict , а також ‘cntr’ з 3 – кластерної моделі

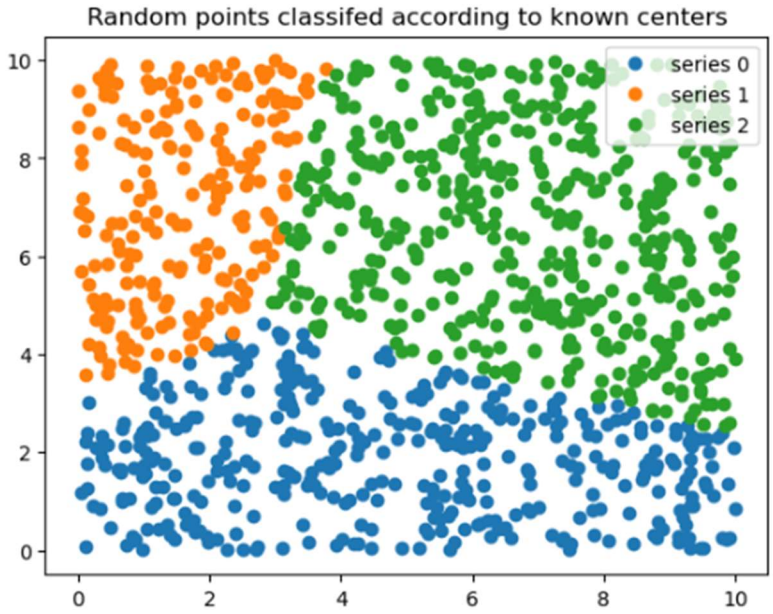

Рисунок 4.6 – графік класифікованих уніфікованих даних

В результаті був побудован графік класифікованих уніфікованих даних (рис.4.6).Для візуалізації максимальне значення приналежності було взято в кожній точці (тобто вони посилені, не візуалізуються нечіткі результати), але повний нечіткий результат є виходом cmeans_predict.

ВИСНОВОК

В результаті виконання роботи розроблено інформаційно-програмне забезпечення яке використовує алгоритм нечіткої кластерізації для роботи з різними об’ємами інформації .

Сформульовано та проанализовано проблеми що виникають при проектуванні моделей та алогоритмів в умовах нечіткості а також запропоновано шляхи їх вирішеннь.

Проведено когнітивний аналіз необхідності розробки інформаційно-програмного забезпечення і його переваги в порівнянні з ручною роботою.

Зроблено детальний огляд і аналіз можливостей та моделей нечітких алгоритмів , розглянут і обран кращий з них.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Заде, Л. Понятие лингвистической переменной и ее применение к принятию приближенных решений / Л. Заде, 1976. – 168 с
2. Дюран, Б. Кластерный анализ / Б. Дюран, П. Одел, 1977. – 128 с
3. Дюбуа, Д. Теория возможностей. Приложения к представлению знаний в информатике / Д. Дюбуа, А. Прад;, 1990. – 288 с.
4. Асаи, К. Прикладные нечеткие системы: пер. с япон. /К. Асаи, Д. Ватада, С. Иваи., 1993. – 386 с
5. Zadeh, L.A. Outline of a New Approach to the Analysis of Complex Systems and Decision Processes / L.A. Zadeh // IEEE Transactions on Systems, Man and Cybernetics. –1973. – Vol. SMC–3. – P. 28–44.
6. Кини, Р. Л. Принятие решений при многих критериях: предпочтения и замещения / Р.Л. Кини, Х. Райфа, 1981. – 560 с
7. Bothe, H.H. Fuzzy Neural Networks. Tutorium / H.H. Bothe // IFSA’97, Prague, 1997, 37 p

ДОДАТКИ

ДОДАТОК А
Генерація рандомізованих даних

```
from __future__ import division, print_function
import numpy as np
import matplotlib.pyplot as plt
import skfuzzy as fuzz
colors = ['b', 'orange', 'g', 'r', 'c', 'm', 'y', 'k', 'Brown', 'ForestGreen']
# Визначити три центри кластерів
centers = [[4, ],2
           [1, 7],
           [5, 6]]
# Визначити три кластерні сигми для x і y, відповідно
sigmas = [[0.8, 0.3],
          [0.3, 0.5],
          [1.1, 0.7]]
# Створити тестові дані
np.random.seed(42)  # Встановити початкове значення для відтворюваності


xpts = np.zeros(1)
ypts = np.zeros(1)
labels = np.zeros(1)
```

```
    for i, ((xmu, ymu), (xsigma, ysigma)) in enumerate(zip(centers, sig-
mas)):
        xpts = np.hstack((xpts, np.random.standard_normal(200) *
xsigma + xmu))
        ypts = np.hstack((ypts, np.random.standard_normal(200) *
ysigma + ymu))
        labels = np.hstack((labels, np.ones(200) * i))
    # візуалізація
    fig0, ax0 = plt.subplots()
    for label in range(3):
        ax0.plot(xpts[labels == label], ypts[labels == label], '.',
                 color=colors[label])
    ax0.set_title('Test data: 200 points x3 clusters.')
```

ДОДАТОК Б
Налаштування циклу кластеризування

```
fig1, axes1 = plt.subplots(3, 3, figsize=(8, 8))
alldata = np.vstack((xpts, ypts))
fpcs = []


for ncenters, ax in enumerate(axes1.reshape(-1), 2):
    cntr, u, u0, d, jm, p, fpc = fuzz.cluster.cmeans(
        alldata, ncenters, 2, error=0.005, maxiter=1000, init=None)


Зберігання значення fpc для наступних
fpcs.append(fpc)
```

Побудува призначенних кластерів для кожної точки даних у навчальному наборі

```
cluster_membership = np.argmax(u, axis=0)
    for j in range(ncenters):
        ax.plot(xpts[cluster_membership == j],
                ypts[cluster_membership == j], '.', color=colors[j])
```

Позначення центра кожного нечіткого кластера

```
for pt in cntr:
        ax.plot(pt[0], pt[1], 'rs')
    ax.set_title('Centers = {0}; FPC = {1:.2f}'.format(ncenters, fpc))
    ax.axis('off')
```

ДОДАТОК В
Налаштування циклу кластеризування

```
u, u0, d, jm, p, fpc = fuzz.cluster.cmeans_predict(
    newdata.T, cntr, 2, error=0.005, maxiter=1000)
cluster_membership = np.argmax(u, axis=0)
fig3, ax3 = plt.subplots()
ax3.set_title('Random points classifed according to known centers')
for j in range(3):
    ax3.plot(newdata[cluster_membership == j, 0],
             newdata[cluster_membership == j, 1], 'o',
             label='series ' + str(j))
ax3.legend()
plt.show()
```