Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _інформаційних радіотехнологій та технічного захисту інформації_
(повна назва)

Кафедра __медіаінженерії та інформаційних радіоелектронних систем_____
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти _____другий (магістерський)_____

_____Дослідження методу автоматичної постановки медичних діагнозів
(тема)

Виконав:
студент __2_ курсу, групи ___МІм-20-1___
_____Прокопюк П.Р.
(прізвище, ініціали)

Спеціальність 172 Телекомунікації
та радіотехніка
(код і повна назва спеціальності)
Тип програми ____освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма _Медіаінженерія_____
(повна назва освітньої програми)

Керівник проф. Тихонов В.А.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ ___Карташов В.М.__
(підпис) (прізвище, ініціали)

2021р.

Харківський національний університет радіоелектроніки

Факультет Інформаційних радіотехнологій та технічного захисту інформації
Кафедра Медіаінженерії та інформаційних радіоелектронних систем

Рівень вищої освіти другий (магістерський)
Спеціальність 172 Телекомунікації та радіотехніка
(код і повна назва)
Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма "Медіаінженерія"
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____»_____ 2021 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Прокопюку Павлу Романовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження методу автоматичної постановки медичних діагнозів затверджена наказом по університету від "08"11 2021 р. № 1676 Ст
2. Термін подання студентом роботи 08.12.2021 р.
3. Вихідні дані до роботи:
1. Розробка та дослідження методу автоматичної постановки діагнозу в медицині.
2. Моделювання постановки діагнозу за симптоми захворювання.
3. Розробка регресійної багатофакторної моделі, методу прийняття рішення по коефіцієнтам детермінації, генерація послідовностей випадкових подій.
4. Перелік питань, що потрібно опрацювати в роботі
Вступ
1. Аналітичний огляд літературних джерел.
2. Статистична модель об'єктів нечислової природи і метод прийняття рішення при постановці діагнозу в медицині.
3. Події і операції над ними в теорії ймовірності. Регресійний аналіз процесів.
4. Регресійні моделі подій і процесів
5. Прийняття рішення при постановці діагнозів.
Висновки
Перелік посилань
ДОДАТКИ
5. Перелік графічного матеріалу із зазначенням обов'язкових креслеників, схем, плакатів, комп'ютерних ілюстрацій:

1. Постановка задачі (1аркуш)
2. Адаптивна суміш синусоїди та гаусового шуму (1 аркуш)
3. Графіки випадкового процесу та корельованого з ним потоку подій (1 аркуш)
4. Найбільш характерні симптоми для обраних хвороб (1 аркуш)
5. Результати навчання при розпізнаванні хвороб за 16 симптомами (1 аркуш)

6.	Етап розпізнавання (постановка первинного діагнозу) (1 аркуш)
7.	Результати розпізнавання хвороб за коефіцієнтом детермінації (1 аркуш)
8.	Висновки (1 аркуш)

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Аналітичний огляд літературних джерел	08.11.21–11.11.21	
2	Діагностичні рішення	12.11.21–14.11.21	
3	Перехід від міркувань до гіпотез	14.11.21–19.11.21	
4	ДСМ-підхід	20.11.21–23.12.21	
5	Статистичні моделі об'єктів нечислової природи та метод прийняття рішення при постановці діагнозу в медицині	24.11.21–26.11.21	
6	Статистика об'єктів нечислової природи	26.11.21–29.12.21	
7	Байєсівський класифікатор об'єктів нечислової природи	26.11.21–29.12.21	
8	Події та операції з них у теорії ймовірності. Регресійний аналіз процесів	26.11.21-29.12.21	
9	Моделювання статистичних зв'язків подій	29.12.21-04.12.21	
10	Регресійні моделі подій та процесів	05.12.21	
11	Ухвалення рішення при постановці діагнозу	06.12.21–07.12.21	

Дата видачі завдання ____08.11.2021 р._____

Студент _____
(підпис)

Керівник роботи (проекту) _____проф. Тихонов В.А._____
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка кваліфікаційної роботи має: 65 с., 3 рис., 7 табл., 2 додатки, 20 джерел.

СИМПТОМИ ЗАХВОРЮВАННЯ, ДІАГНОЗ ЗАХВОРЮВАННЯ, КОРЕЛЯЦІЙНА ФУНКЦІЯ МОДЕЛІ СИМПТОМІВ, АНАЛІЗ ПОДІЙ, РЕГРЕСІЙНИЙ АНАЛІЗ ОБ'ЄКТІВ НЕЧИСЛОВОЇ ПРИРОДИ.

Об`єкт дослідження – автоматична система постановки діагнозів з метою збільшення якості лікування та зменшення помилок діагнозів.

Мета наукової роботи є розробка та дослідження методу автоматичної постановки діагнозу в медицині.

Методи дослідження. У роботі використані методи теорії множин, лінійних систем, різницевих лінійних рівнянь, методи статистичного моделювання, прикладної аналіз випадкових процесів.

В кваліфікаційній роботі виконано аналітичний огляд процедури постановки діагнозів лікарями, якщо відомі симптоми захворювання. У роботі застосовувався регресійний аналіз, що включає симптоми, як ознаки розпізнавання хвороб і діагноз у числовому поданні. Показано, що отримані коефіцієнти регресії дають змогу з високою точністю діагностувати захворювання. Як вирішальне правило оцінювалася величина коефіцієнта детермінації та F-статистика Фішера. У роботі були проведені методом статистичного моделювання експериментальні дослідження постановки діагнозу за симптомами, в яких припускалися помилки. Основними результатами роботи є розроблений метод прийняття рішення при статистичному аналізу симптомів.

THE ABSTRACT

Explanatory note: 65 p., 3 fig., 7 tabl., 20 sources, 2 app.

SYMPTOMS OF THE DISEASE, DIAGNOSIS OF THE DISEASE, CORRELATION FUNCTION OF THE SYMPTOM MODEL, ANALYSIS OF EVENTS, REGRESSION ANALYSIS OF OBJECTS NON-NUMERICAL NATURE.

Object of research - an automatic system of diagnosis in order to increase the quality of treatment and reduce errors of diagnosis.

The purpose of scientific work is to develop and study the method of automatic diagnosis in medicine.

Research methods. Methods of set theory, linear systems, difference linear equations, methods of statistical modeling, applied analysis of random processes are used in the work.

In the qualification work, an analytical review of the procedure of diagnosis by doctors, if the symptoms of the disease are known. Regression analysis was used in the work, which includes symptoms as signs of disease recognition and diagnosis in numerical representation. It is shown that the obtained regression coefficients allow to diagnose the disease with high accuracy. As a decisive rule, the value of the coefficient of determination and Fisher's F-statistics were evaluated. Experimental studies of the diagnosis of symptoms based on symptoms in which errors were assumed were carried out by the method of statistical modeling. The main results of the work are the developed method of decision making in the statistical analysis of symptoms.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ДСМ	-	Джона Стюарта Мілля експертна система;
БД	-	бази даних;
ППВ	-	правила правдоподібного висновку;
PrTFIDF	-	алгоритм.

ВСТУП

Медицина являє собою слабо структуровану область знань, що створює серйозні труднощі при побудові систем процесу прийняття рішень. В той же час, у практичній діяльності лікар вибудовує послідовність висновків, що спираються на уявлення про зв'язок спостережуваних у хворого ознак з певним діагнозом. Такий підхід до постановки діагнозу дозволяє формалізувати це завдання під статистичну теорію прийняття рішень. Для цього, перш за все, необхідно перейти від нечислових даних до їх числового подання.

У одних випадках, що характеризуються класичними проявами хвороби, гіпотеза і навіть остаточне рішення виникає у процесі огляду, за іншими - лише після обстеження. Слід зазначити, що послідовність діагностичних досліджень може піддаватися корекції, а іноді корінній трансформації, залежно від одержуваних у процесі обстеження результатів. Швидкість прийняття рішення залежить як від кваліфікації та діагностичного "чуття" лікаря, так і від особливостей прояву захворювання у конкретного хворого. При створенні експертної системи слід враховувати, що у лікарів існують індивідуальні неявні переваги до порядку обстеження хворого та ролі симптомів, що виявляються, тобто різна ступінь уваги до фактів, а іноді нехтування деякими із них через ймовірну незначності, іноді помилкової, для аналізованої гіпотези.

Існуючі програми автоматичної постановки діагнозу, зазвичай, мають справу з даними числової природи (кардіограми, ритмограми, енцефалограми та інших.). При аналізі даних нечислової природи виникає ряд проблем, які вирішувалися в кваліфікаційній роботі. До них відносяться, зокрема, чисельне подання наявності та відсутності симптомів, правильний та неправильний діагноз, вибір моделі даних, визначення вирішального правила.

Таким чином, тема досліджень щодо вирішення завдання створення системи автоматичної постановки діагнозу в медицині, проведених у кваліфікаційній роботі, є **актуальною**.

Метою кваліфікаційної роботи є розробка та дослідження методу автоматичної постановки діагнозу в медицині.

1 АНАЛІТИЧНИЙ ОГЛЯД ЛІТЕРАТУРНИХ ДЖЕРЕЛ

1.1 Діагностичні рішення

Одне з основних положень диференціального діагнозу - виявлення характерних патологічних симптомів - є обов'язковим, але не завжди достатнім фактором для ідентифікації захворювання, що пояснюється як змінюється, в процесі прогресування хвороби, діагностичною цінністю тих самих ознак, так і необхідністю засновувати діагностичний висновок на відтінках симптомів та ознак шуканих хвороб. З іншого боку, зустрічаються ситуації, у яких як симптоми, і навіть ознаки хвороби мають менше діагностичне значення, ніж клінічний фон, у якому вони виявляються [1, 2]. Це те, що носить назву анамнезу життя та анамнезу хвороби, включаючи попередні захворювання. Ще одна проблема - атипові симптоми, які зустрічаються так часто, що виправдовують, як підкреслює Р. Рігельман [3], існування наступної максими: "Атипові симптоми частих хвороб бувають частіше, ніж типові рідкісні симптоми".

Метою диференціальної діагностики, як метафорично висловив це Г. Евербек [4], є визначення найкоротшого шляху від найяскравішого симптому до діагнозу. Більш строго слід було б сказати, що процес диференціальної діагностики спрямований на ідентифікацію стану хворого в сенсі розпізнавання хвороби, що вразила його. Формально це може звучати наступним чином: необхідно визначити хворобу (нозологічну форму) при якій матимуть місце несуперечливі відносини між ознаками, що спостерігаються, і інтегруючим їх поняттям діагнозу. При цьому слід мати на увазі, що ці відносини можуть бути неповними внаслідок відсутності будь-яких із відповідних даному захворюванню ознак.

Аргументація лікаря-діагноста спрямована, з одного боку, на виявлення ознак, що є характерними для передбачуваного ним діагнозу, а з іншого боку, на пошук альтернативних ознак, що заперечують інші захворювання

(наприклад, високий зріст є однозначно заперечуючим хворобі, при яких обов'язкове значне зниження зросту), тобто використовуються аргументи та контраргументи - факти "за" та "проти". У самому загальному вигляді можна говорити, що одночасно з винятком одного діагнозу має місце підтвердження іншого (або інших) діагнозів (діагностичних гіпотез). Але аргументи розрізняються також за ступенем їхньої "важливості" - на "сильні" або "слабкі" - і один "сильний" аргумент може змінити рішення, скасувавши дію безлічі "слабких" аргументів. Результат міркування і рішення може змінити поява нового аргументу, тобто має місце немонотонність аргументації, що особливо суттєво в умовах неповноти та недостовірності вихідної інформації.

Виявлення специфічних для одного діагнозу або відсутність специфічних для іншого діагнозу симптомів дозволяє відносно швидко дійти остаточного висновку (наприклад, при діабетичній комі в анамнезі в більшості випадків діабету, нещодавно перенесена інфекція і низький артеріальний тиск, у той час як при уремічній комі в анамнезі нирок або гіпертонічна хвороба, а артеріальний тиск підвищений або в нормі). Цей відомий факт знайшов цікаве підтвердження при аналізі "діагностичних" протоколів [5], який показав, що фахівець практично завжди цікавиться скаргами та анамнезом захворювання, після чого ставить попередній діагноз або найбільш ймовірні діагнози, а потім ставить ще одне питання для остаточного уточнення. З цього виходить, що лікар високої кваліфікації не користується повним обсягом інформації, представленим в історії хвороби, а лише цілком певним мінімумом, який забезпечує вирішення завдання. При цьому він подумки здійснює процес диференціальної діагностики з аналізом усіх фактів, що підтверджують і заперечують. Це зумовлено тим, що досвід та професійна підготовка навчили більшість лікарів скептично ставитися до своїх початкових гіпотез та шукати додаткові дані, які могли б підтвердити чи спростувати перші підозри [3]. Альтернативні версії змушують лікаря зосередити увагу інших хворобах, можливих при даної симптоматиці.

Диференціальна діагностика допомагає сформулювати питання, на які слід відповісти, як прийняти чи відкинути початкове припущення.

Викладене дозволяє стверджувати, що лікарі в неявній формі використовують апарат аргументації для підтвердження однієї з гіпотез - одного або, у поодиноких випадках, поєднання двох діагнозів (наприклад, при диференціальній діагностиці ішемічної хвороби серця та міокардиту вказівкою на користь першої гіпотези є зменшення болю від прийому нітрогліцерину). У складних випадках, коли має місце широкий диференціально-діагностичний ряд і велика кількість збігаються ознак, іноді одна характерна але відсутня ознака відразу призводить до відхилення тієї чи іншої діагностичної гіпотези, навіть за наявності ряду ознак, що підкріплювали до того цю гіпотезу. Наприклад, при диференціації між міокардіодистрофією та дефіцитом карні тину, відсутність характерної зміни зубця Т на електрокардіограмі дозволяє однозначно відхилити другий із названих діагнозів. Такий вплив контраргументів, як виступають специфічні прояви хвороби. У той же час, можуть виявлятися не альтернативні асоціативні ознаки, що мають майже однакову диференціальну значущість, облік яких практично ніяк не впливає на вирішення питання про відхилення або підтримку певної діагностичної гіпотези.

Звернемося до понять нецеситарного та евентуального причинного зв'язку [6] для характеристики патологічного процесу в організмі. Можна вважати, що нецеситарний зв'язок, що має місце тоді, коли при настає А подія (або факт) В є обов'язковою, відноситься до ряду захворювань, що характеризуються етіопатогенетичними зв'язками, що однозначно виявляються (тобто впливом конкретного агенту або фактору і подальшим механізмом змін в організмі, Прикладом чого може служити СНІД при попаданні в організм людини вірусу імунодефіциту - ВІЛ). Що стосується евентуальних причин, коли при настанні А подія (або факт) В стає можливим, то це відноситься до розвитку ускладнень в процесі захворювання на певному етапі.

1.2 Етапність діагностичного процесу

У процесі диференціальної діагностики здійснюється, прямо чи опосередковано, облік як безпосередніх, і опосередкованих відносин між ознаками і хворобами (нозологічними формами). Чим вище кваліфікація лікаря, тим більш глибинні стосунки, зумовлені механізмом патологічного процесу, він може розглядати. У більшості випадків здійснюється аналіз за типом причинно-наслідкових залежностей між групою ознак і найбільш ймовірним з точки зору лікаря діагнозом, який визначається низкою факторів - частотою даної патології, прочитаної останнім часом статтею, тим, що запам'ятався випадком. З цього видно, що вибір аналізованих ознак носить відносно суб'єктивний чи умовно-об'єктивний характер, оскільки відбираються хоч і безумовно суттєві в діагностичному плані ознаки, але ті, які є доцільними конкретному лікарю. Природно, що при цьому значне місце має кваліфікація лікаря (як сума знань) та його практичний досвід (як “архів” випадків особистого спостереження – прецедентів).

Діагностичний процес можна поділити на три взаємопов'язані етапи: постановка первинного діагнозу (попередня гіпотеза), побудова диференційно-діагностичного ряду (висунення додаткових гіпотез), остаточний діагноз (обґрунтування остаточної гіпотези).

У першому - в процесі " міркування " на кшталт логічного висновку (доказові міркування по Е.В. Левенцю [7]) - лікар рухається від скарг, аналізу анамнестичних даних (історії виникнення перших проявів захворювання) і патологічних проявів до встановлення попереднього (первинного)) діагнозу, тобто. до побудови вихідної діагностичної гіпотези шляхом "не аргументованого" міркування, за принципом "оскільки є ознаки ... може бути діагноз ...", не використовуючи систему доказів, а лише викладаючи думки в логічно послідовній формі [8]. При побудові гіпотези враховується також можливість фонових захворювань, симптоми яких можуть деформувати картину основного захворювання. Так цукровий діабет як фонове

захворювання обтяжує перебіг ішемічної хвороби серця і може бути причиною розвитку інших патологічних станів, як, наприклад, ретинопатія.

Використання терміну "міркування" передбачає у цьому випадку його відмінність від доказів, розуміючи відкритість безлічі можливих аргументів [9]. Фактично лікар оцінює ознаки, маючи в своєму розпорядженні їх у мисленій шкалі важливості (значимості) шляхом не аргументованого міркування і аналогій, тобто спираючись на власний досвід (включаючи пам'ять про аналогічні випадки) та літературні дані, але не використовуючи формальні логічні процедури.

На наступному етапі здійснюється аргументація "за" , що забезпечує залучення додаткових діагностичних гіпотез. Формується диференційно-діагностичний ряд, тобто коло захворювань, у яких зустрічаються зумовлені у хворого ознаки й у яких можуть бути характерні подібні початкові прояви. Іншими словами, здійснюється розширення потенційно діагностичної послідовності для подальшого ухвалення остаточного рішення. Цей етап вкрай важливий, тому що оберігає від помилкового рішення на користь першої діагностичної гіпотези, тому що певні стани можуть імітувати далекі один від одного захворювання. У медицині такі ситуації прийнято визначати як прояв одного захворювання під маскою іншого.

На третьому етапі здійснюється процес послідовного виключення нозологічних форм, включених раніше у диференціальний ряд, тобто критична порівняльна оцінка виявлених симптомів, результатів досліджень та їх сукупностей - аргументація "за" та "проти" (контраргументи).

Слід мати на увазі, що процес діагностичного міркування може повертатися до попередніх кроків і станів, видаляючи деякі припущення та аргументи з відповідних множин і розмірковуючи знову з урахуванням нових фактів (аргументів) та прийняття нових припущень. У цих випадках проблема управління вибором правил виведення перетворюється на проблему реалізації рефлексивної поведінки системи [10].

1.3 Аргументація та аналогія

Під міркуванням зазвичай мають на увазі і процес дедуктивного виведення з деякої множини вихідних суджень (висновків), і міркування за аналогією, і міркування, що спираються на приховані асоціації [11]. Будь-яке міркування складається з деякої сукупності суджень. Під міркуванням, в контексті цієї роботи, будемо розуміти деяке обґрунтоване, доведене чи просто передбачуване співвідношення між деякими сутностями (множинами, об'єктами, ознаками, подіями тощо), як це викладено у [12].

Особливостями міркування, що відрізняють його від логічного висновку, є [13]: 1) відкритість безлічі можливих аргументів; 2) використання метатеоретичних, і зокрема металогічних засобів, за допомогою яких здійснюється управління логічними висновками, що застосовуються в процесі міркування; 3) використання як правил достовірного висновку, а й правил правдоподібного висновку, апроксимуючих застосовані евристики, аналогії тощо.

Аргументація, заснована загалом на доводах розуму, що лежить на власному досвіді та знанні літератури, обов'язково кореспондує з науковою школою, до якої явно чи неявно звертається лікар, і спирається на систему переконань, що включають індивідуальні психологічні характеристики людини, її рефлексію, що визначає здатність самооцінки і корекції гіпотез, що виникають у нього. Виходячи з цього ясно, що аргументація та міркування можуть бути і помилкові, наприклад, при відсталості та догматичності переконань, при однозначному відхиленні будь-яких інших думок, наприклад, помилкова посилка і псевдо аргументація на кшталт використовуваної Т.Д. Лисенко та його послідовниками щодо наслідуваних і ознак. На жаль, іноді це має місце і в медицині, хоча не в такій різко вираженій формі.

Можливий варіант, коли здійснюється аргументація за подібністю до безпосередніх проявів або аналогічної реакції на лікування з випадком-

прецедентом. Як відомо, висновки за аналогією поділяються на аналогію ознак та аналогію відносин [14]. Аналогія ознак сприймається як порівняння окремих ознак предметів. Така аналогія має менш правдоподібний характер, оскільки ігнорує зв'язок ознак один з одним. Аналогія відносин будується якраз з урахуванням такого зв'язку. Діагностика або вибір терапевтичної стратегії за аналогією (за прецедентом) якраз передбачає облік лікарем сукупності факторів, які йому важко побудувати у вигляді логічної послідовності, але він "збирає" їх як ціле (уявний образ) при згадці про аналогічний випадок [15]. У цьому одна із сильних сторін висококваліфікованого (досвідченого) лікаря-експерта.

Подібність може бути представлена як наявність деяких загальних властивостей чи відносин. Якщо два (або більше) стани подібні в якійсь частині їх ознак, тобто підстава припускати, що вони можуть мати й інші однакові ознаки, про які нам нічого не відомо. Тоді висновком за аналогією є перенесення властивостей чи відносин, які мають місце одного стану, в інший стан з урахуванням подібності з-поміж них. Висновки за аналогією мають правдоподібний характер, оскільки ми спостерігаємо схожість між станами лише у частині їх ознак. Виділяються такі чинники, що впливають на ступінь правдоподібності висновків за аналогією: кількість однакових ознак, значимість однакових ознак, врахування відмінностей між ознаками. Ступінь правдоподібності укладання за аналогією збільшується зі збільшенням числа однакових ознак [7]. Механізм виведення, заснований на методах теорії аналогій, може вбудовуватися в загальний процес аргументаційного прийняття рішення як одну з його ланок (як правило, на перших етапах діагностики або при виборі лікування). Рішення за аналогією має в основі опору на лікарський досвід і, таким чином, пов'язане з накопиченням випадків у пам'яті, а подібним чином воно може реалізовуватися і в базах даних систем, що використовують даний принцип аналізу станів.

Синтезом пізнавальних процедур, що поєднує правила правдоподібного висновку, що породжують гіпотези про причини, правила висновку за аналогією, абдуктивний висновок та індуктивні узагальнення є ДСМ (John Stuart Mill)-міркування. Як і в логіці аргументації, в ДСМ-міркування висловлювання оцінюються за допомогою зіставлення аргументів "за" та "проти" [16]. На розширеній логіці аргументації реалізовані такі ДСМ-міркування, що [13]: 1) умови їх застосування можуть бути точно визначені (і навіть аксіоматизовані); 2) складаються з послідовної рекурентної реалізації двох типів правил правдоподібного висновку, які застосовуються до початкового стану даних та до наступних станів даних, породжених застосуванням цих правил; 3) поділяються на правила породження гіпотез про причини ефектів (множин властивостей об'єктів) та на правила прогнозування наявності або відсутності властивостей у об'єктів (правила висновку за аналогією); 4) безліч породжених гіпотез приймається лише за виконання критерію достатньої підстави, формульованого як спеціальна аксіома, регулююча і прийняття гіпотез, і розширення вихідної вибірки (початкового стану бази даних - БД); 5) заключним етапом є породження індуктивних узагальнень.

При проведенні консиліуму аргументи відображають окремі думки фахівців і можуть мати різну оцінку істинності для його учасників. Тому формалізм для побудови логічного висновку на основі аргументації повинен враховувати структуру безлічі аргументів, зокрема, можливість існування відношення порядку на безлічі аргументів [10], що визначається діагностичною значимістю ознак, що залучаються як аргументи. Це можуть бути патогномонічні (однозначно характеризують захворювання), обов'язкові (що зустрічаються з частотою 80 - 90% при даній патології), головні (що зустрічаються з частотою 50 - 60%) та супутні або другорядні ознаки.

1.4 Перехід від міркувань до гіпотез

Відносно чіткі залежності в системі відносин "слідство - результат" або " ознаки, що спостерігаються - ідентифікація стану" справедливі для певних, більш менш стандартних ситуацій, в яких аргументація на користь певного діагнозу будується на пошуку класичних проявів захворювання і використанні добре відомих ознак диференціювання. Нерідко попередній діагноз формується безпосередньо в процесі збору анамнезу та послідовного "сканування ознак" (огляді хворого), наприклад, больові відчуття під час прийому їжі, і різні диспепсичні явища викликають у лікаря думки про виразкову хворобу шлунка. А подальше обстеження (або спостереження) пацієнта є необхідною умовою отримання інформації для аргументації в процесі диференціальної діагностики та підтвердження чи відхилення первинної діагностичної гіпотези. Інакше кажучи, має місце те, що В.К. Фін [13] визначає як пізнавальний цикл продуктивного мислення. Щодо аналізованої ситуації для медичної діагностики він може бути представлений у наступному вигляді: аналіз результатів огляду - міркування та аргументація, включаючи аналогії - гіпотеза або альтернативні гіпотези - верифікація або фальсифікація - поповнення даних та знань - повторний цикл міркування та аргументації - корекція гіпотези. Можливий ряд таких ітерацій у процесі отримання нової медичної інформації, що відноситься насамперед до особливо важких для діагностики випадків, що характеризуються високим рівнем подібності клінічних проявів. Цьому можна поставити у відповідність наявність для деяких ситуацій багатьох приписів без вказівки того, яким з них слід слідувати – буриданові ситуації [17].

Діагностичний процес у лікаря високої кваліфікації ґрунтується на особистісних уявленнях, що є сплавом досвіду (пам'яті про хворих, що спостерігалися) і накопичених знань, що поєднуються з перетвореними ("пропущеними через себе") даними медичної літератури. Цьому відповідають вимоги до вирішувачів інтелектуальних партнерських систем

[18], які використовують логіки об'єктивного та суб'єктивного (експертного) знання.

Досягнення мети в інтелектуальній системі [13] та в медичній діагностиці вимагають для свого здійснення певних дій та здібностей (табл. 1.1).

Таблиця 1.1 - Цілі в інтелектуальній системі та в медичній діагностиці

В інтелектуальних системах	У медичній діагностиці
Упорядкування інформації щодо ступеня суттєвості	Підрозділ ознак, залежно від їхньої діагностичної цінності, на: патогномонічні (характерні тільки для певного захворювання), обов'язкові (що зустрічаються в переважній більшості випадків), головні (часто зустрічаються), другорядні
Усунення невизначеності за допомогою використання інформації, впорядкованої за рівнем релевантності для цієї ситуації	Зменшення діагностичної невизначеності шляхом спрямованого пошуку ідентифікуючих ознак
Рефлексивне управління - здатність як до оцінки отриманих результатів та обраних засобів отримання цих результатів, так і до корекції даних (поповнення даних, відмови від деяких даних, перегляду результатів тощо)	Уявна самооцінка спостережуваних ознак і гіпотез, що висуваються, їх відхилення або підтвердження за допомогою додатково знайденої інформації
Вибір стратегій, адекватних розв'язуваній задачі	Пошук аргументів та контраргументів (в анамнезі та у вигляді специфічних змін) або прецеденту

В інтелектуальних системах	У медичній діагностиці
Виведення логічних наслідків	У розпізнаванні станів загальноприйнята логіка "якщо... то"
Пошук подібності фактів та генерування припущень	Наявність фактів, притаманних низці захворювань, дозволяє будувати диференціально-діагностичні ряди
Верифікація та фальсифікація одержуваних результатів	Відхилення контраргументів та подання фактів, однозначно характерних для певної нозологічної форми

1.5 Висновки до розділу та постановка задачі

Медична діагностика, що базується на принципі аргументації та контраргументації у поєднанні з використанням аналогів (прецедентів), включає етапи формування гіпотез з подальшим обґрунтуванням або відхиленням шляхом залучення додаткових фактів. Серед факторів, які можуть вести до неправильних діагнозів, трапляються помилки у судженнях як наслідок недостатньо конструктивного мислення чи нелогічності висновків [19]. Талант діагноста передбачає вміння швидко виділяти провідні симптоми і аналізувати нечіткі уявлення. Виходячи з цього, стає ясною доцільність побудови інтелектуальної системи, що спирається на систему аргументів, що враховує відносини ознак і включає способи ефективної обробки нечітких даних. Формалізація міркувань у рамках квазіаксіоматичної теорії дає можливість уточнити феномен правильності (або коректності) міркування за допомогою критерію достатньої підстави для прийняття висновку правдоподібного міркування. Використання як засобів формалізації, логіки об'єктивного знання і логіки суб'єктивного знання (логіки аргументації) може бути основою для відображення принципу міркувань і доказів лікаря, роблячи інтелектуальні системи більш зрозумілими для користувачів.

Проведений огляд наукової літератури на тему кваліфікаційної роботи дозволяє сформулювати такі завдання:

1. Формалізація задачі встановлення діагнозу.
2. Імітаційне моделювання потоку подій.
3. Вибір вирішального правила з урахуванням рівняння регресії.
4. Обґрунтування критерію якості рівняння регресії.
5. Імітаційне моделювання процедури встановлення діагнозу.

2 СТАТИСТИЧНІ МОДЕЛІ ОБ'ЄКТІВ НЕЧИСЛОВОЇ ПРИРОДИ І МЕТОД ПРИЙНЯТТЯ РІШЕННЯ ПРИ ВСТАНОВЛЕННІ ДІАГНОЗУ В МЕДИЦИНІ

2.1 Статистика об'єктів нечислової природи

Для опису даних, що є об'єктами нечислової природи, застосовують, зокрема, таблиці спряженості, а середніх величин – рішення оптимізаційних завдань. Як вибіркові середні для вимірювань у порядковій шкалі використовують медіану і моду, а в шкалі найменувань – лише моду.

Для вирішення параметричних завдань оцінювання використовують оптимізаційний підхід, метод однокрокових оцінок, метод максимальної правдоподібності, метод стійких оцінок. Для вирішення непараметричних задач оцінювання поряд з оптимізаційними підходами до оцінювання характеристик використовують непараметричні оцінки розподілу випадкового елемента, густини розподілу.

Як приклад методів перевірки статистичних гіпотез для об'єктів нечислової природи розглянемо критерій «хі-квадрат» (позначають χ^2) розроблений К. Пірсоном для перевірки гіпотези однорідності (іншими словами, збігу) розподілів, що відповідають двом незалежним вибіркам. Розглядаються дві вибірки об’ємів n_1 і n_2, що складаються з результатів спостережень якісної ознаки, що має k градацій. Нехай m_{1j} і m_{2j} – кількості елементів першої та другої вибірок відповідно, для яких спостерігається j–та градація, а p_{1j} і p_{2j} – ймовірності того, що ця градація буде прийнята, для елементів першої та другої вибірок, $j = 1, 2, \ldots, k$.

Для перевірки гіпотези однорідності розподілів, що відповідають двом незалежним вибіркам

$$H_0: p_{1j} = p_{2j}, j = 1,2, \ldots, k,$$

застосовують критерій χ^2 (хі-квадрат) зі статистикою

$$x^2 = n_1 n_2 \sum_{i-1}^{k} \frac{1}{m_{1i} + m_{2i}} \left(\frac{m_{1i}}{n_1} - \frac{m_{2i}}{n_2} \right)^2 .$$

Встановлено, що статистика χ^2 при великих обсягах вибірок n_1 і n_2 має асимптотичний розподіл хі-квадрат з $(k - 1)$ ступенем свободи.

В табл. 2.1 наведено дані про вміст сірки у вуглецевій сталі, що виплавляється двома металургійними заводами. Перевіримо, чи можна вважати розподіл домішки сірки в плавках сталі цих двох заводів однаковими. Розрахунок за даними табл. 2.1 дає $X^2 = 3{,}39$. Квантиль порядку 0,95 розподілу хі-квадрат з $k - 1 = 3$ ступенями свободи дорівнює $\chi^2_{0.95}(3) = 7.8$, тому гіпотезу про збіг функцій розподілу вмісту сірки в плавках двох заводів не можна відхилити, тобто, її слід прийняти (на рівні значущості $\alpha = 0{,}05$).

Таблиця 2.1-Розподіл плавок сталі за відсотковим вмістом сірки

Вміст сірки, в %	Число плавок	
	Завод А	Завод Б
0,00 ÷ 0,02	82	63
0,02 ÷ 0,04	535	429
0,04 ÷ 0,06	1173	995
0,06 ÷ 0,08	1714	1307

Вище дано лише короткий опис змісту прикладної статистики на сучасному етапі. Детальний виклад конкретних методів міститься у спеціальній літературі.

2.2 Об'єкти нечислової природи як результат статистичної обробки даних

Об'єкти нечислової природи виникають не тільки на "вході" статистичної процедури, а й у процесі обробки даних на "виході" як результат статистичного аналізу. Розглянемо найпростішу прикладну постановку задачі регресії. Початкові дані мають вигляд $(x_i, y_i) \in R^2, i = 1,2, \dots, n$. Мета у тому, щоб із достатньою точністю описати *y* як поліном від *x*, тобто модель має вигляд

$$y_i = \sum_{k=0}^{m} a_j x_i^k + \varepsilon_i, \qquad (2.1)$$

де m - невідомий ступінь полінома; $a_0, a_1, a_2, \dots, a_m$ - невідомі коефіцієнти багаточлена; $\varepsilon_i, i = 1,2, \dots, n$ - похибки, які для простоти приймемо незалежними і мають один і той нормальний розподіл. Тут наочно проявляється одна із причин поширеності статистичних моделей на основі нормального розподілу. Такі моделі, хоч і, як правило, неадекватні реальній ситуації, з математичної точки зору дозволяють проникнути глибше в суть явища, що вивчаються. Тому вони придатні для початкового аналізу ситуації, як і в цьому випадку. Подальші наукові дослідження мають бути спрямовані на зняття нереалістичного припущення нормальності та переходу до непараметричних моделей похибки.

Поширена процедура така: спочатку намагаються застосувати модель (2.1) для лінійної функції (m = 1), при невдачі (неадекватності моделі) переходять до багаточлена другого порядку (m = 2), якщо знову невдача, то беруть модель (2.1) з m=3 і т.д. (адекватність моделі перевіряють за F-критерієм Фішера).

Обговоримо властивості цієї процедури у термінах прикладної статистики. Якщо ступінь полінома задана (m=m_0), то його коефіцієнти

оцінюють методом найменших квадратів, властивості цих оцінок добре відомі [20]. Однак в описаній вище реальній постановці задачі m теж є невідомим параметром та підлягає оцінці. Таким чином, потрібно оцінити об'єкт (m, a_0, a_1, a_2, …, a_m), безліч значень якого можна описати як $R^1UR^2UR^3U$ .... Це об'єкт нечислової природи, звичайні методи оцінювання йому непридатні, оскільки m - дискретний параметр. У розглянутій постановці завдання, розроблені на сьогодні методи оцінювання ступеня полінома носять переважно евристичний характер.

У загальному випадку лінійної регресії дані мають вигляд

$(y_i X_i), = 1,2, \dots, n,$ де $X_i = (x_{i1}, x_{i2}, \dots, x_{iN})$ - вектор предикторів (чинників, які пояснюють змінні), а модель така

$$y_i = \sum_{jeK} a_j x_i + \mathcal{E}_i, i = 1,2, \dots, n. \qquad (2.2)$$

Тут K - деяка підмножина множини {1,2,…,n}; a_j - невідомі коефіцієнти при предикторах із номерами з K. Модель (2.1) зводиться до моделі (2.2), якщо

$$x_{i1} = 1, x_{i1} = x_i, x_{i2} = x_i^2, x_{i3} = x_i^3, \dots, x_{ij} = x_i^{j-1},$$

У моделі (2.1) є природний порядок введення предикторів на розгляд - відповідно до зростання ступеня, а в моделі (2.2) природного порядку немає, тому тут стоїть довільна підмножина безлічі предикторів. Є тільки частковий порядок - чим потужність підмножини менша, тим краще. Модель (2.2) особливо актуальна у технічних дослідженнях. Вона застосовується в завданнях управління якістю продукції та інших техніко-економічних дослідженнях, в економіці, маркетингу та соціології, коли з великої кількості факторів, що імовірно впливають на змінну, що досліджується, треба відібрати по можливості найменшу кількість значущих факторів і за їх допомогою сконструювати прогнозуючу формулу (2.2).

Завдання оцінювання моделі (2.2) розбивається на дві послідовні задачі: оцінювання множини K - підмножини безлічі всіх предикторів, а потім - невідомих параметрів a_j. Методи вирішення другого завдання добре відомі та докладно вивчені. Набагато гірша ситуація з оцінюванням об'єкта нечислової природи K. Як зазначалося, існуючі методи - переважно евристичні, часто не є навіть заможними. Навіть саме поняття спроможності у разі вимагає спеціального визначення. Нехай K_0 - істинна підмножина предикторів, тобто підмножина, для якої справедлива модель (2.2), а підмножина предикторів K_n - його оцінка. Оцінка K_n називається спроможною, якщо

$$\lim_{n\to\infty} Card(K_n \Delta K_0) = 0,$$

де Δ - символ симетричної різниці множин; $Card(K)$ означає кількість елементів у множині K, а межа розуміється у сенсі збіжності за ймовірністю.

Завдання оцінювання у моделях регресії, таким чином, розбивається на дві - оцінювання структури моделі та оцінювання параметрів при заданій структурі. У моделі (2.1) структура описується невід'ємним цілим числом m, у моделі (2.2) - множиною K. Структура – об'єкт нечислової природи. Завдання її оцінювання складне, тоді як завдання оцінювання чисельних параметрів при заданій структурі добре вивчено, розроблені ефективні (у сенсі прикладної математичної статистики) методи.

Така сама ситуація і в інших методах багатовимірного статистичного аналізу - у факторному аналізі (включаючи метод головних компонентів) та в багатовимірному шкалюванні, в інших оптимізаційних постановках проблем прикладного багатовимірного статистичного аналізу.

Перейдемо до об'єктів нечислової природи на "виході" статистичної процедури. Приклади численні. Розбиття - результат роботи багатьох алгоритмів класифікації, зокрема, алгоритмів кластер-аналізу. Ранжування – результат, наприклад, упорядкування професій із привабливості чи

автоматизованої обробки думок експертів - членів комісії з підбиття підсумків конкурсу наукових праць. В останньому випадку використовується ранжування зі зв'язками. Так, до однієї групи, найбільш численної, потрапляють роботи, які не отримали нагород. З усіх об'єктів нечислової природи, мабуть, найчастіші на "виході" дихотомічні дані - прийняти чи не прийняти гіпотезу, зокрема, прийняти чи забракувати партію продукції. Результатом статистичної обробки даних може бути безліч, наприклад, зона найбільшого ураження при аварії, або послідовність множин, наприклад, "середньомірний" опис поширення пожежі. Нечітким безліччю Е. Борель ще на початку XX століття, пропонував описувати уявлення людей про кількість зерен, що утворює "купу". За допомогою нечітких множин формуються значення лінгвістичних змінних, що виступають як підсумкова оцінка якості систем автоматизованого проектування, сільськогосподарських машин, побутових газових плит, надійності програмного забезпечення або систем управління. Можна констатувати, що це види об'єктів нечислової природи можуть з'являтися " на виході " статистичного дослідження.

2.3 Байєсівський класифікатор об'єктів нечислової природи

Наївна Байєсівська модель є імовірнісним методом навчання та класифікації об'єктів нечислової природи. Імовірність того, що документ d потрапить до класу c записується як $P(c|d)$. Оскільки мета класифікації - знайти відповідний клас для даного документа, то в наївній байєсівський класифікації завдання полягає в знаходженні найбільш ймовірного класу c_m

$$c_m = \underset{c \in C}{argmax} P\left(c/d\right).$$

Вирахувати значення цієї ймовірності безпосередньо неможливо, оскільки для цього потрібно, щоб навчальна множина містила всі (або майже

всі) можливі комбінації класів та документів. Однак, використовуючи формулу Байєса, можна переписати вираз для $P(c / d)$

$$c_m = \underset{c \in C}{argmax} \frac{P\left(c/d\right)}{P(d)} = \underset{c \in C}{argmax} P\left(d/c\right)P(c),$$

де знаменник $P(d)$ опущений, тому що не залежить від c і, отже, не впливає на знаходження максимуму; $P(c)$ - ймовірність того, що зустрінеться клас c, незалежно від документа, що розглядається; $P(d/c)$ - ймовірність зустріти документ d серед документів класу c.

Використовуючи навчальну множину, ймовірність $P(c)$ можемо оцінити як

$$\hat{P}(c) = \frac{N_c}{N}, \qquad (2.3)$$

де N_c - кількість документів у класі c, N - загальна кількість документів у навчальній множині. Тут використаний інший знак для ймовірності, оскільки за допомогою навчальної множини можна лише оцінити ймовірність, але не знайти її точне значення.

Щоб оцінити ймовірність $P\left(d/c\right) = P(t_1, t_2, \dots, t_{nd}/c)$, де t_k - терм із документа d, n_d - загальну кількість термів у документі (включаючи повторення), необхідно ввести полегшення припущення про умовну незалежність термів і про незалежність позицій термів. Іншими словами, ми нехтуємо, по-перше, тим фактом, що в тексті природною мовою поява одного слова часто тісно пов'язана з появою інших слів (наприклад, ймовірніше, що слово інтеграл зустрінеться в одному тексті зі словом рівняння, ніж зі словом бактерія) . По-друге, ймовірність зустріти одне й те саме слово різна для різних позицій у тексті. Саме через ці грубі спрощення аналізована модель природної мови називається наївною (проте, вона є досить ефективною в

задачі класифікації). Отже, у світлі зроблених припущень, використовуючи правило множення ймовірностей незалежних подій, можна записати

$$P(d/c) = P\left(\frac{t_1, t_2, \dots, t_{nd}}{c}\right) = P\left(\frac{t_1}{c}\right) P\left(\frac{t_2}{c}\right) \dots P\left(\frac{t_{nd}}{c}\right) = \prod_{k=1}^{n_d} P\left(\frac{t_k}{c}\right).$$

Оцінка ймовірностей $P(t \mid c)$ за допомогою навчальної множини буде

$$\hat{P}\left(\frac{t}{c}\right) = \frac{T_{ct}}{T_c},$$

де T_{ct} - кількість входження терму t у всіх документах класу c (і на будь-яких позиціях - тут суттєво використовується друге спрощуюче припущення, інакше довелося б обчислити ці ймовірності для кожної позиції в документі, що неможливо зробити досить точно через розрідженість навчальних даних - важко очікувати, щоб кожен терм зустрівся в кожній позиції достатню кількість разів); T_c - загальна кількість термів у документах класу c. При підрахунку враховуються всі повторні входження.

Так як класифікатор " навчений " , тобто знайдено величини $\hat{P}(c)$ і $\hat{P}(t/c)$, уже можливо знайти клас документу

$$c_m = \underset{c \in C}{argmax}\, P\left(\frac{d}{c}\right) P(c) = \underset{c \in C}{argmax}\, P(c) \prod_{k=1}^{nd} P\left(\frac{t_k}{c}\right).$$

Щоб уникнути в останній формулі переповнення знизу через велику кількість співмножників, на практиці замість множення зазвичай використовують суму логарифмів. Логарифмування не впливає знаходження максимуму, оскільки логарифм є монотонно зростаючою функцією. Тому в більшості реалізацій замість останньої формули використовується

$$c_m = \underset{c \in C}{argmax} \left[\log \hat{P}(c) + \sum_{k=1}^{n_d} \log \hat{P}(t_k/c) \right].$$

Ця формула має просту інтерпретацію. Шанси класифікувати документ класом, що часто зустрічається вище, доданок $\log P(c)$ вносить у загальну суму відповідний внесок. Величини ж $\log P(t/c)$ тим більше, чим важливіше терм *t* для ідентифікації класу *c*, і, відповідно, тим вагоміший їхній внесок у загальну суму.

Застосування класичного методу Байєса виявляється скрутним при великих обсягах даних. Замість нього пропонується використати метод на основі алгоритму PrTFIDF, який так само полягає у обчисленні умовних ймовірностей приналежності документа до рубрики. Експерименти проводились на основі даних інтернет-каталогу. Результати експериментів дозволяють зробити висновок про можливість успішного застосування методу.

Завдання методів класифікації полягає в тому, щоб найкраще вибрати відмінні ознаки та сформулювати правила, на основі яких прийматиметься рішення про віднесення тексту до рубрики. Існує два протилежні підходи до вибору ознак та побудови правил: машинне навчання та експертний метод. Машинне навчання передбачає наявність набору відрубрицированих документів, у якому навчається алгоритм. Експертний метод передбачає, що виділення ознак та складання правил проводиться експертами.

У загальному випадку завдання автоматичної рубрикації, заснованої на машинному навчанні, можна поставити так: є безліч документів D (навчальна вибірка) і безліч рубрик R організованих в ієрархію, кожному документу з D приписана одна або кілька рубрик з R. Потрібно на основі цих даних побудувати процедуру рубрикації, яка полягає у знаходженні найбільш ймовірної рубрики з R для довільного документа, що досліджується.

Якість отриманої процедури оцінюється з урахуванням тестової вибірки, куди також входять документи, котрим заздалегідь відомі відповідні їм рубрики. Документи із текстової вибірки проходять автоматичну класифікацію. Результати цієї класифікації порівнюються із заздалегідь відомими значеннями. Для оцінки якості застосовуються кілька метрик, які будуть описані нижче.

У серії робіт школи бізнесу Університету штату Канзас описані байєсові методики оцінки ризику та прогнозу дохідності портфелів фінансових інструментів. Основними перевагами байєсових мереж у фінансових завданнях є можливість спільного обліку кількісних та якісних ринкових показників, динамічне надходження нової інформації, а також явні залежності між суттєвими факторами, що впливають на фінансові показники.

Результати моделювання подаються у формі гістограм розподілу ймовірностей, що дозволяє зробити детальний аналіз співвідношень «ризик-прибутковість». Дуже ефективними є такі широкі можливості для ігрового моделювання.

3 ПОДІЇ ТА ОПЕРАЦІЇ НАД НИМИ В ТЕОРІЇ ЙМОВІРНОСТІ. РЕГРЕСІЙНИЙ АНАЛІЗ ПРОЦЕСІВ

3.1 Класичні операції над подіями

Поняття “подія” є ключовим в теорії ймовірності. Подія в теорії ймовірності сприймається як будь-який факт, що у результаті досвіду може статися чи не статися. Щоб порівнювати події між собою за рівнем ймовірності, потрібно з кожною подією пов'язати певну число (характеристику) ймовірності появи. Це число є ймовірністю події.

При створенні способу постановки діагнозу, поняття “подія” є також ключовим. Проте на відміну теорії ймовірності, в кваліфікаційній роботі події описуються не на основі ймовірності, а вводиться визначення кореляції подій нечислової природи. Нижче показано, що операції над подіями не використовують кореляційні зв'язки. Наведено також елементи регресійного аналізу, що використовується для побудови вирішального правила при постановці діагнозу.

В теорії ймовірності розглядаються такі види подій:

Достовірна подія - це така подія, яка в результаті досліду обов'язково станеться.

Неможлива подія - це така подія, яка в результаті досліду не може статися.

Повна група подій – це такі події, одна з яких у результаті досліду обов'язково станеться.

Несумісні події – це такі події, які не можуть статися одночасно.

Рівноможливі події – жодна з подій не є більш можливою, ніж інша.

Протилежними подіями називаються дві несумісні події, що утворюють повну групу.

Випадок сприятливий для певної події, якщо поява випадку тягне за собою появу події.

Для дослідів, у яких можливі наслідки симетричні та однаково можливі, застосовується спосіб безпосереднього підрахунку ймовірностей. Події, що утворюють повну групу, несумісні та рівноможливі, характеризуються *«схемою випадків»* для підрахунку ймовірностей. Відповідно до цього способу ймовірності рівні

$$P(A) = \frac{m}{n},$$

де P(A) - ймовірність події A, n -загальна кількість випадків, m - кількість випадків, сприятливих для події A $0 \leq P(A) \leq 1$.

Якщо події не зводяться до подій, що описуються схемою випадків, для них при розрахунку ймовірності використовують метод частоти події або статистичної ймовірності. Вона виконується згідно з формулою

$$P(A) = \frac{m}{n},$$

де m - кількість випадків події A, n - загальна кількість зроблених дослідів. При використанні цього методу необхідно, щоб події належали до категорії масових явищ і виконувалася властивість стійкості частот на інтервалі часу дослідження. Тільки в цьому випадку можна говорити про ймовірність подій, маючи на увазі не математичні функції, а реальні характеристики випадкових явищ.

Випадковою величиною називається величина, що приймає в результаті досвіду те чи інше значення, причому наперед невідоме. Вони можуть бути як безперервними, так і дискретними. Часто буває зручніше працювати з подіями, а не з випадковими величинами. Кожне значення в теорії ймовірності грає роль практично достовірної і недостовірної події, що дозволяє передбачати події.

3.2 Теорема теорії ймовірності

У багатьох випадках не можна застосовувати частоту подій для отримання ймовірностей. Тому застосовуються непрямі методи, коли за відомими ймовірностями одних подій знаходять ймовірність іншої, пов'язаної з ним події. Теореми, що використовуються для цього, доводяться для подій, що зводяться до схеми випадків, а для інших схем приймаються як аксіоми.

Сума двох подій А і В є подія С, що полягає у виконанні подій А або В, або двох разом. Для кількох подій їхня сума дорівнює події появи хоча б одного з них.

Добуток кількох подій дорівнює події спільної появи всіх цих подій.

Ці операції дозволяють отримати складні події як комбінації елементарних, наприклад

$$B = A_1\overline{A_2A_3} + \overline{A_1}A_2\overline{A_3} + \overline{A_1A_2}A_3,$$

де $\overline{A}$ подія протилежна А.

Теорема додавання ймовірностей

Імовірність суми двох несумісних подій дорівнює сумі ймовірностей цих подій

$$P(A + B) = P(A) + P(B).$$

Для декількох подій

$$P(A_1 + A_2 + \cdots + A_{n+1}) = P(A_1) + P(A_2) + \cdots + P(A_{n+1}).$$

Наслідок 1:

Якщо події $A_1,\ldots,A_n$ утворюють повну групу подій, то сума їх ймовірностей дорівнює 1

$$\sum_{i-1}^{n} P(A_i) = 1.$$

Наслідок 2:

Сума ймовірностей протилежних подій дорівнює 1

$$P(A) + P\left(\overline{A}\right) = 1.$$

У разі спільних подій

$$P(\sum_i^n A_i) = \sum_i P(A_i) - \sum_{i,j} P\left(A_i A_j\right) + \sum_{i,j,k} P\left(A_i A_j A_k\right) \ldots + (-1)^{n-1} P(A_1, A_2, \ldots, A_n).$$

З цих формул випливають формули для ймовірності твору спільних подій

$$P(AB) = P(A) + P(B) - P(A + B),$$

$$P(ABC) = P(A) + P(B) + P(C) - P(A + B) - P(A + C) - P(B + C) + P(A + B + C),$$

$$P(A_1, A_2, \ldots, A_N) = \sum_i P(A_i) - \sum_{i,j} P\left(A_i + A_j\right) + \sum_{i,j,k} P\left(A_i + A_j + A_k\right) \ldots + (-1)^{n-1} P(A_1 + \cdots + A_n)$$

Теорема множення ймовірностей

Подія А незалежна від події В, якщо ймовірність А не залежить від того чи відбулась подія В. Подія А залежить від події В, якщо ймовірність А залежить від того, чи відбулась подія В. Ймовірність події А, обчислюваної

за умови, що має місце подія B, називається умовною ймовірністю події A, тобто P(A/B).

Для залежних подій

$$P(AB) = P(A)P(B/A),\ P(A_1 \dots A_n) = P(A_1)P\left(\frac{A_2}{A_1}\right) \dots P\left(\frac{A_n}{A_1 A_2 \dots A_n}\right).$$

Для незалежних подій

$$P(AB) = P(A)P(B), \text{ так як } P(B) = P\left(\frac{B}{A}\right),$$
$$P(A_1 \dots A_n) = P(A_1)P(A_2) \dots P(A_n).$$

Іноді корисно шукати ймовірність протилежних подій.

3.3 Формула повної ймовірності

Формула повної ймовірності є наслідком теорем додавання та множення. Нехай потрібно визначити ймовірність певної події A, яке може статися спільно з однією з подій (гіпотез) $H_1, H_2, \dots, H_n$, утворюють повну групу подій. Тоді формула повної ймовірності

$$P(A) = \sum_{i=1}^{n} P(H_i)P(A/H_i).$$

Для A, H_i справедлива формула

$$A = H_1A + H_2A + \dots + H_nA.$$

Якщо $H_1, H_2, \dots, H_n$ несумісні, то несумісні і H_iA. Тоді випливає, що

$$P(A) = \sum_{i=1}^{n} P(H_i A).$$

Теорема Байєса визначає можливість незалежних гіпотез за умови появи події A

$$P\left(\frac{H_i}{A}\right) = \frac{P(H_i)P(A/H_i)}{P(A)} = \frac{P(H_i)P(A/H_i)}{\sum_{i=1}^{n} P(H_i)P(A/H_i)}.$$

Як показав аналіз операцій над подіями теорії ймовірності поняття кореляції подій, формул для оцінювання кореляції подій ще не запропоновано. Разом з тим з таким поняттям оперують, коли йдеться про зв'язок подій на рівні кореляцій, їх взаємний вплив один на одного. Тому при створенні методу автоматичної постановки діагнозу, визначається та використовується поняття кореляції подій, їхня лінійна залежність на рівні регресійного аналізу. Під подіями ми розумітимемо симптоми хвороб та подію, що полягає у поставлені діагнозу. Зв'язок між цими подіями визначено через багатофакторну регресію.

3.4 Регресійний аналіз корельованих процесів

Регресійний аналіз є ефективним методом аналізу корельованих випадкових процесів. У наступному розділі класичний регресійний аналіз буде узагальнено на кореляційні події. У зв'язку з цим доцільно зупинитись на основних положеннях регресійного аналізу корельованих випадкових процесів.

Розглянемо завдання множинної регресії, коли залежний випадковий процес може бути представлений виваженою сумою випадкових процесів, що його пояснюють. Рівняння регресії має вигляд

$$Y(t) = \sum_{l=1}^{n} h_l X_l(t) + a(t), \qquad (3.1)$$

де $Y(t)$ - залежний випадковий процес, $X_l(t)$ - пояснюючі випадкові процеси h_l – коефіцієнти ваги пояснювальних процесів, $a(t)$ - помилка регресії. Для визначення невідомих коефіцієнтів регресії необхідно, щоб пояснюючі процеси були корельовані між собою та із залежним випадковим процесом. При цьому помилка регресії має бути некорельованим випадковим процесом.

Для отримання системи рівнянь для розрахунку коефіцієнтів регресії використовують два оптимальні критерії, еквівалентних у рамках статистик другого порядку. Перший критерій передбачає, що оптимальні коефіцієнти ваги мінімізують дисперсію помилки регресії. Згідно з другим критерієм, процеси, що пояснюють, некорельовані з помилкою регресії $E[X_l(t)a(t)] = 0$.

Взаємна кореляція залежного процесу та пояснювальних процесів визначається співвідношенням

$$R_{YXi} = \frac{1}{N}\sum_{t=1}^{N} Y(t)X_i(t)\ , \tag{3.2}$$

Взаємна кореляція пояснювальних процесів $X_i(t)$ знаходиться за звичайною формулою

$$R_{X_iX_l} = \frac{1}{N}\sum_{t=1}^{N} X_i(t)X_l(t). \tag{3.3}$$

Щоб отримати рівняння для обчислення коефіцієнтів h_l необхідно помножити ліву та праву частини (3.1) на $X_i(t), i = 1, n$, а потім знайти середнє. В результаті маємо систему рівнянь

$$R_{YX_i} = \sum_{j=1}^{n} h_j R_{X_j,X_i}. \tag{3.4}$$

В матричному вигляді це рівняння має вигляд

$$R_{XX} \times h = R_{YX},$$

де

$$R_{YX} = \begin{bmatrix} R_{YX_1} \\ R_{YX_2} \\ \dots \\ R_{YX_n} \end{bmatrix}, R_{xx} = \begin{bmatrix} R_{X_1X_1} & R_{X_2X_1} & \dots & R_{X_nX_1} \\ R_{X_1X_2} & R_{X_2X_2} & \dots & R_{X_nX_2} \\ & & \dots & \\ R_{X_1X_n} & R_{X_2X_n} & \dots & R_{X_nX_n} \end{bmatrix}, h = \begin{bmatrix} h_1 \\ h_2 \\ \dots \\ h_n \end{bmatrix}.$$

Рішення матричного рівняння можна записати у вигляді

$$h = R^{-1}{}_{XX} R_{YX}.$$

Як приклад розглянемо рівняння регресії з двома процесами, що пояснюють їх залежність. Множинне рівняння регресії має вигляд

$$Y(t) = h_1 X_1(t) + h_2 X_2(t) + a(t). \qquad (3.5)$$

Згідно з (3.5), система рівнянь для обчислення коефіцієнтів h_l має вигляд

$$R_{YX_1} = h_1 R_{X_1X_1} + h_2 R_{X_2X_1},$$

$$R_{YX_2} = h_1 R_{X_1X_2} + h_2 R_{X_2X_2}.$$

Мета регресійного аналізу полягає у поясненні поведінки залежної змінної $Y(t)$. У будь-якій даній вибірці розкид $Y(t)$ виявляється порівняно низьким в одних спостереженнях та порівняно високим - в інших. Розкид значень $Y(t)$ у будь-якій вибірці можна сумарно описати за допомогою вибіркової дисперсії Var($Y(t)$). Для цього потрібно розраховувати величину цієї дисперсії. У багатофакторному регресійному аналізі пояснюють поведінку $Y(t)$ шляхом визначення регресійної залежності $Y(t)$ від

відповідно обраних незалежних змінних $X_i(t)$. Після побудови рівняння регресії можна розбити значення $Y(t)$ у кожному спостереженні на дві складові: $\hat{Y}(t)$ і $a(t)$.

$$Y(t) = \hat{Y}(t) + a(t), \tag{3.6}$$

де величина $\hat{Y}(t)$ — розрахункове оцінкове значення $Y(t)$ у спостереженні, тобто це те значення, яке мав би $Y(t)$ за умови, що рівняння регресії було правильним і при відсутності випадкового чинника. Це, іншими словами, величина $Y(t)$, спрогнозована за значенням $X(t)$ у цьому спостереженні. Тоді залишок $a(t)$ є розбіжність між фактичним та спрогнозованим значеннями величини $Y(t)$. Це та частина $Y(t)$, яку не можна пояснити за допомогою рівняння регресії. Використовуючи (3.6), розкладемо дисперсію $Y(t)$

$$\text{Var}(Y(t)) = \text{Var}(\hat{Y}(t) + a(t)) = \text{Var}(\hat{Y}(t)) + \text{Var}(a(t)) + 2\text{Cov}(\hat{Y}(t) + a(t)). \tag{3.7}$$

Далі, виявляється, що $Cov(\hat{Y}(t) + a(t))$ повинна дорівнювати нулю, в силу статистичної незалежності помилок регресії. Отже, отримаємо:

$$\text{Var}(Y(t)) = \text{Var}(\hat{Y}(t)) + \text{Var}(a(t)). \tag{3.8}$$

Це означає, що ми можемо розкласти Var $(Y(t))$ на дві частини: $\text{Var}(\hat{Y}(t))$ — частина, яка «пояснюється» рівнянням регресії у вищеописаному сенсі, та $\text{Var}(a(t))$ — "непояснену" частину. Згідно (3.8), $\text{Var}(\hat{Y}(t))/\text{Var}(Y(t))$ — це частина дисперсії $Y(t)$, пояснена рівнянням регресії. Це відношення відоме як коефіцієнт детермінації, і його зазвичай позначають

$$R^2 = \text{Var}(\hat{Y}(t))/\text{Var}(Y(t)),$$

що рівносильно

$$R^2 = 1\text{-}Var(e(t))/Var(Y(t)). \quad (3.9)$$

Максимальне значення коефіцієнта R^2 дорівнює одиниці. Це відбувається в тому випадку, коли лінія регресії точно відповідає всім спостереженням, тому всі залишки дорівнюють нулю. Тоді

$$Var\ (Y(t)) = Var\ (\hat{Y}(t)),\ Var\ (e(t)) = 0$$

і

$$R^2 = 1.$$

Якщо у вибірці відсутній видимий зв'язок між $Y(t)$ та незалежними змінними $X_i(t)$, то коефіцієнт детермінації R^2 буде близьким до нуля.

За інших рівних умов бажано, щоб коефіцієнт R^2 був якнайбільше. Можна показати, що принцип мінімізації суми квадратів залишків еквівалентний мінімізації дисперсії помилок регресії. Якщо ми мінімізуємо дисперсії помилок регресії, то відповідно (3.9) автоматично максимізується коефіцієнт детермінації R^2.

Якщо залежною змінною є обраний діагноз, а в якості пояснюючих змінних використовуються симптоми хвороби, то коефіцієнт детермінації показує, наскільки добре симптоми відповідають обраному діагнозу. Тобто максимальне значення коефіцієнта детермінації із усіх можливих рівнянь регресії може бути вирішальним правилом у виборі діагнозу.

4 РЕГРЕСІЙНІ МОДЕЛІ ПОДІЙ І ПРОЦЕСІВ

4.1 Регресійний аналіз подій та процесів

Розглянуті вище традиційні теорії ймовірності уявлення подій та операції з них дають можливість вирішувати ряд задач і мають практичне значення під час аналізу подій. Нижче розглядатимуться події у представленому раніше класичному розумінні. Однак надалі нас цікавитимуть не ймовірність їхньої появи, а їхній взаємозв'язок.

У теорії ймовірності для вирішення завдань застосовуються два підходи: перший використовує поняття ймовірності чи щільності розподілу, а другий виходить з аналізу статистичних зв'язків, описуваних кореляційними або моментними функціями. З наведеного вище аналізу випливає, що до випадкових подій зазвичай застосовується перший підхід, а аналіз статистичних зв'язків застосовується для випадкових величин. В теорії ймовірності, зазвичай, не аналізуються статистичні зв'язки між подіями, між подіями і випадковими процесами. Немає також математичного поняття "кореляції подій". Разом з тим, людство оперує такими поняттями як "вплив подій на процеси", "кореляція подій", оскільки це дозволяє глибше розуміти причини та сутність багатьох явищ. Для вирішення завдань, поставлених у кваліфікаційній роботі, виникає необхідність формалізувати ці поняття та визначити для них математичний апарат. Хоча під час постановки діагнозу використовується регресійний аналіз подій, у цьому розділі було також розглянуто загальніший випадок спільної регресії подій та процесів. Він пов'язаний із важливою проблемою обліку впливу на події та процеси деяких незалежних пояснюючих подій та процесів.

Розглянемо завдання множинної регресії, коли випадкові події впливають випадкові процеси. У більш загальному вигляді можна розглядати завдання впливу пояснюючих процесів або подій на залежний процес. Представимо рівняння регресії у загальному вигляді, використовуючи

позначення $Y(t)$ - аналізований залежний процес, що приймає довільні значення або залежна подія, що приймає значення -1, 0, +1 або інші комбінації з цих чисел. $Y(t)$ залежить від пояснюючих подій $A_i(t)$, або від пояснюючих процесів $X_i(t)$, або від їхньої сукупності. Тоді рівняння регресії має вигляд

$$Y(t) = \sum_{i=1}^{n1} k_i A_i(t) + \sum_{l=1}^{n2} h_l X_l(t) + a(t). \qquad (4.1)$$

Взаємна кореляція процесу та подій визначається співвідношенням

$$R_{YA_i} = \frac{1}{N}\sum_{t=1}^{N} Y(t)A_i(t). \qquad (4.2)$$

Якщо знаки відліків процесу та знаки подій збігаються, то кореляція буде великою. В іншому випадку, коли події та процес не пов'язані, то $R_{YA_i} \to 0$. Кореляція процесів $Y(t)$ і $X_l(t)$знаходиться за звичайною формулою

$$R_{YX} = \frac{1}{N}\sum_{t=1}^{N} Y(t)X_i(t). \qquad (4.3)$$

При знаходженні взаємних кореляцій необхідно враховувати лише зв'язок подій та процесів чи зв'язок подій та подій. Отже, в кореляції враховуватимуться лише ті відліки процесу, які мають тимчасові інтервали, що збігаються, з відповідними подіями.

Щоб отримати рівняння для обчислення коефіцієнтів k_i і h_l необхідно помножити ліву та праву частини (4.1) на $A_j(t), j = 1, n_1$ або $X_m(t), m = 1, n_2$, а потім усереднити. В результаті маємо систему рівнянь

$$R_{YAj} \text{или } X_m = \sum_{l=1}^{n2} h_l R_{A_j,A_j} \text{ или } X_m + \sum_{l=1}^{n2} h_l R_{Xl,Aj} \text{ или } X_m \text{ ,} \qquad (4.4)$$

де для взаємних кореляцій введено позначення типу R_{A_j,A_j} или X_m, що означає, що взаємна кореляція визначається для $A_i(t)$ і $A_j(t)$ або для $A_j(t)$ і $X_m(t)$.

Відповідно до (4.2), система рівнянь для обчислення коефіцієнтів k_i і h_l має вигляд

$$\begin{aligned} R_{YA_1} &= \sum_{i=1}^{2} k_i R_{A_i,A_1} + \sum_{l=1}^{2} h_l R_{X_l,A_1}, \\ R_{YA_2} &= \sum_{i=1}^{2} k_i R_{A_i,A_2} + \sum_{i=1}^{2} h_l R_{X_l,A_2}, \\ R_{YX_1} &= \sum_{i=1}^{2} k_i R_{A_i,X_1} + \sum_{l=1}^{2} h_l R_{X_l,X_1}, \\ R_{YX_2} &= \sum_{i=1}^{2} k_i R_{A_i,X_2} + \sum_{l=1}^{2} h_l R_{X_l,X_2}. \end{aligned} \tag{4.5}$$

Після визначення коефіцієнтів регресії, рівняння (4.5) дозволяє прогнозувати майбутні значення процесу. Для цього в це рівняння необхідно підставити відповідні значення процесів та подій та зважити їх із знайденими коефіцієнтами регресії.

Рівняння (4.1) дозволяє враховувати вплив тих чи інших незалежних процесів та подій на залежний процес чи подію $Y(t)$. В теорії ймовірності такі рівняння та кореляційні зв'язки між подіями не аналізуються, а аналітики це роблять умоглядно або використовуючи свій досвід. При коректній постановці завдання, що полягає у правильному виборі кількості процесів,

що враховуються $X_l(t)$ і подій $A_i(t)$, достатньої точності оцінки взаємних кореляцій, помилка $a(t)$ буде некорельованою з $X_l(t)$ і $A_i(t)$.

Враховуючи важливість регресійного аналізу у дослідженнях та прогнозах, можна сподіватися, що наведений метод розширить коло розв'язуваних завдань та дозволить підвищити точність вже отриманих рішень регресійних рівнянь для випадкових процесів.

4.2 Моделювання статистичних зв'язків подій

Під час створення нових алгоритмів важливо використовувати імітаційне моделювання обробки статистичних зв'язків подій. Воно дозволяє як більш тонко зрозуміти особливості алгоритму обробки і перевірити потенційні можливості застосовуваних методів. У регресійному аналізі дуже рідко зустрічаються імітаційні експерименти. Тому певний інтерес представляє запропонований спосіб імітаційного моделювання регресійного аналізу процесів і, зокрема, регресійного аналізу подій.

З точки зору математичної статистики, не існує способу довести чи існує у корельованих процесів фізичний зв'язок, який з них є причиною іншого. Такий аналіз має робитися з урахуванням нематематичних доказів чи припущень. Тому, наприклад, якщо є три корельовані випадкові процеси або потік подій, то можна припустити, що будь-який з них був отриманий як зважена сума двох інших. Використовуючи ці міркування, проводився експеримент з оцінки коефіцієнтів регресії, якщо припустити, що з трьох корельованих процесів, один із них становить виважену суму двох інших, тобто описується рівнянням

$$A(t) = k_1 A_1(t) + k_2 A_2(t) + a(t).$$

Потоки імітаційних подій отримані описаним вище способом. Події $A_2(t)$ були отримані з використанням протифазного синусоїдального

процесу. Знайдені оцінки складали k_1=0,96; k_2=-0,21. У регресії, складеної з процесів, що породжують потоки подій, оцінки коефіцієнтів були k_1=1,14; k_2=-0,39.

Для характеристики якості рівняння регресії або обраної моделі зв'язку симптомів та захворювання оцінюють коефіцієнт детермінації, рівний

$$R^2 = \frac{var(\widehat{D})}{var(D)}.$$

Тут дисперсія, розрахованого значення діагнозу $\widehat{D}$, дорівнює

$$var(\widehat{D}) = \frac{1}{n}\sum_{j=1}^{n}(\widehat{D}_j - \overline{D})^2,$$

а var(D) - дисперсія діагнозу дорівнює

$$var(D) = \frac{1}{n}\sum_{j=1}^{n}(D_j - \overline{D})^2.$$

Середнє значення діагнозу становить

$$\overline{D} = \frac{1}{n}\sum_{j=1}^{n} D_j.$$

Коефіцієнт детермінації R2 набуває значення в діапазоні

$$0 \le R^2 \le 1.$$

Він показує, яка частина дисперсії результативної ознаки D пояснена рівнянням регресії. Наприклад, значення R2=0,9 свідчить, що відповідне рівняння регресії пояснює 90% дисперсії результативного ознаки. Чим більше R2, тим краще в регресії результуючого діагнозу пояснюється симптомами з урахуванням яких складено рівняння регресії. Тобто даний набір симптомів визначає діагноз.

Для оцінки якості регресії, тобто за яких значень R2, рівняння вважатимуться статично значимим, що доводить його використання у регресійному аналізі. Для цього застосовується F-критерій Фішера. При розрахунку величини F, вважає, число ступенів свободи дорівнює

$$h1=p, \qquad h2= n-p-1,$$

де p - число симптомів, що відповідають даній хворобі, n-кількість хворих з даним діагнозом. В імітаційному експерименті, представленому в кваліфікаційній роботі, ці дані становили p=16, h= 83.

Величину F можна виразити через R^2

$$F = \frac{R^2}{1-R^2} \times \frac{h_2}{h_1}.$$

Для визначення ймовірності, що виправдовує цю гіпотезу, вибирають рівень значущості α. Рівень значущості зазвичай вибирають рівним 0,05 або 0,01, що відповідає ймовірності помилки 5% і 1%. Розраховані F - статично були більше F критичного які для h1=16 и h2=83 складали 2 для α=5% і 2,9% для α=1%.

Відповідно до F-критерію Фішера, при F>Fкрит, F-статистики визначаються за таблицями F-критерія Фішера при вибраному рівні значущості та розрахованих величинах ступенів свободи. Для розрахованих вище ступенів свободи та критеріях значимості 5% та 1%, отримані значення

F-статистик значно перевищують F-критичні. Таким чином, регресійні рівняння значущі як за коефіцієнтом детермінації, так і за F-статистиками.

4.3 Генерація пояснюючих подій та випадкових процесів

Для генерації випадкового процесу, що складається з подій (потоку подій) A(t)=(-1,0,+1) пропонується спочатку отримати два періодичні випадкові процеси $x_{1,2}(t)$з рівними частотами. Періодичні процеси обрані для отримання простим способом стаціонарного процесу. Для цього використовуються адитивні суміші однієї і тієї ж синусоїди з одиничною амплітудою, але з різними вибірками білого гаусового шуму. Отримані суміші можуть мати також різні відносини с/ш. На рис. 4.1 представлені моделі двох періодичних випадкових процесів, отриманих таким способом.

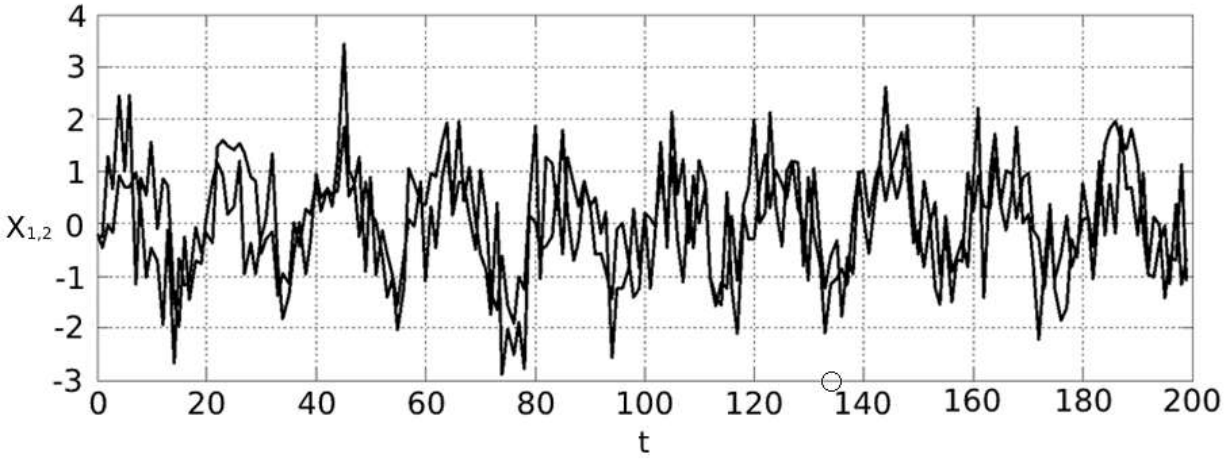

Рисунок 4.1 – Графіки двох вибірок, що являють собою адитивну суміш синусоїди та гаусового білого шуму, з відношенням с/ш=0,8

Взаємна кореляція цих процесів дорівнює 0,41. При відношенні с/ш=0,95 взаємна кореляція цих процесів дорівнює 0,46, а при с/ш=2 вона становить 0,65. Таким чином, змінюючи відношення с/ш, можна підібрати процеси з різним рівнем взаємної кореляції. Зсунувши синусоїду на половину періоду, легко отримати протифазний випадковий процес із негативною кореляцією.

Для генерації потоку подій із випадкового процесу використовувалося правило:

$$\begin{aligned} &\text{Якщо } -0.3 \le x(t) \le 0.3 \text{ , то } A(t) = 0; \\ &\text{Якщо } 0.3 < x(t), \text{ то } A(t) = 1; \\ &\text{Якщо } x(t) < -0.3, \text{ то } A(t) = -1. \end{aligned} \qquad (4.6)$$

Результати моделювання потоку подій випадкового процесу, показані на рис. 4.2. Взаємна кореляція випадкового процесу і породжуваного ним потоку подій досить висока і становить 0,87.

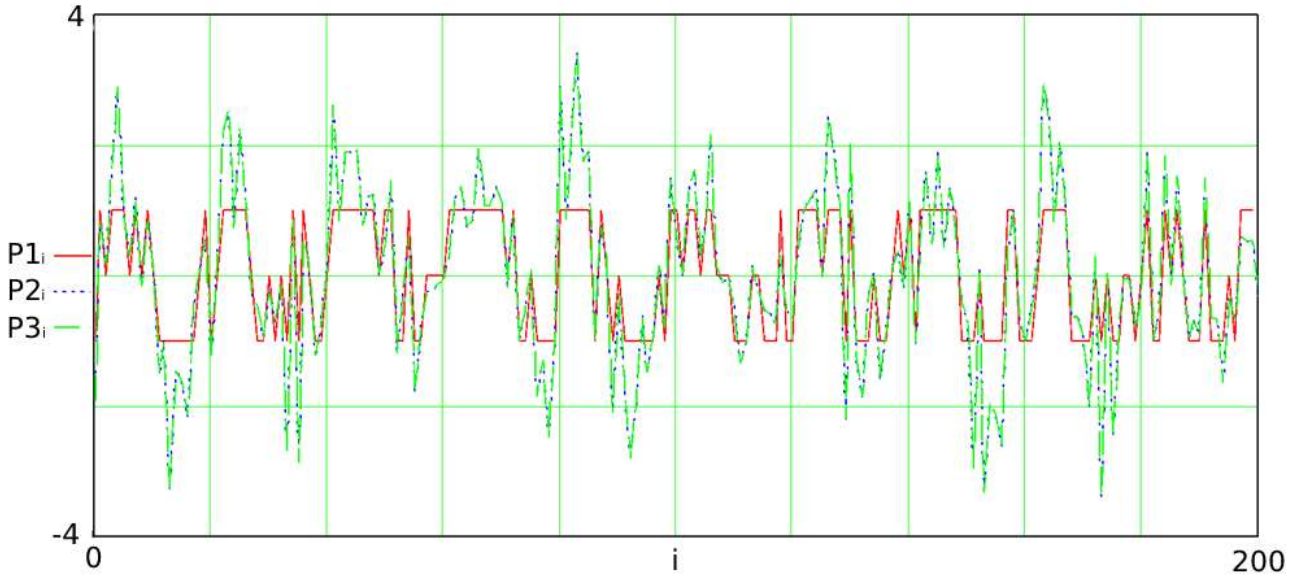

Рисунок 4.2 – Графіки випадкового процесу та отримані з нього потоки подій при с/ш=0,8

З графіків на рис. 4.3 видно, що автокореляційні функції випадкового процесу та породжуваного ним потоку подій, представлених на рис. 4.2, досить близькі. Це говорить про те, що їхня інформативність у рамках статистик другого порядку практично однакова.

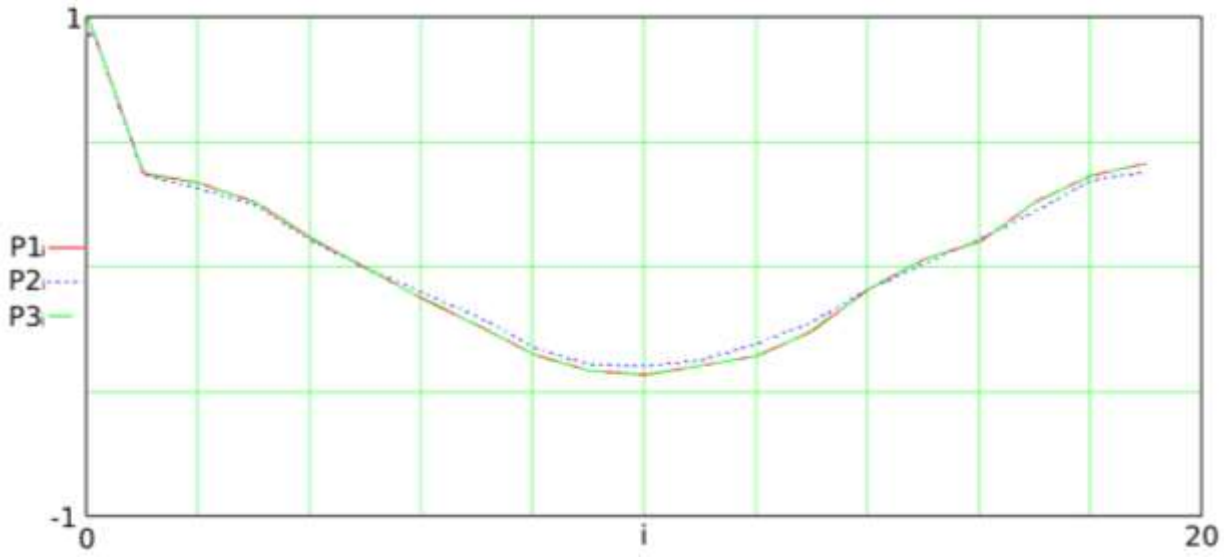

Рисунок 4.3 – Автокореляційні функції випадкового процесу та породжуваного ним потоку подій, представлених на рисунку 4.2

5 ПРИЙНЯТТЯ РІШЕННЯ ПРИ ПОСТАНОВЦІ ДІАГНОЗУ

5.1 Особливості завдання ухвалення рішення

Постановка діагнозу - це неоднозначне рішення, що залежить від ряду факторів - подій, що відбулися з хворим, які називаються симптомами. При постановці діагнозу можна переплутати симптоми. Наприклад, гепатит може супроводжуватися як захворюванням печінки, так і гастритом, хворобами жовчного міхура, панкреатитом тощо. Тому часто доводиться встановлювати діагноз та призначати відповідне лікування за відсутності повної визначеності.

Існує також проблема, пов'язана з необхідністю поставити складний для лікаря діагноз при захворюваннях, рідкісних для даної місцевості або країни. У цьому випадку неоціненна допомога є досвід інших лікарів. Під досвідом тут розуміємо ефективне використання зв'язків симптомів та діагнозу, наведені у різних джерелах, майстерність найкращих лікарів, використання передових досліджень.

На етапі призначення лікування виникає неоднозначність у виборі оптимального поєднання лікувальних призначень. Іноді доводиться ставити діагноз на ранній стадії захворювання, коли хвороба ще не характеризується усіма симптомами, а у хворого на розвиток хвороби багато в чому впливають випадкові фактори. При цьому імовірнісна складова задачі посилюється.

Метою роботи є розробка автоматичної системи встановлення діагнозу, що дозволяє максимально повно використовувати накопичений колективний лікарський досвід у лікуванні аналогічних захворювань. І цим знизити ймовірність лікарських помилок, підвищити ефективність лікування. Ця мета може бути досягнута за рахунок застосування при обробці наявних даних симптоматики нових інформаційних методів аналізу даних, що базуються на оригінальних розробках.

Наведені вище міркування показують, що у кожному з етапів диференціальної діагностики та лікування захворювань виникають проблеми, на вирішення яких корисно використовувати методи математичної статистики. Необхідні статистичні дані одержують під час аналізу медичних карток хворих. Вони включають: відомості про хворого, правильно встановлені діагнози, симптоми, що спостерігаються при захворюванні, елементи успішного лікування хворого. Після створення бази даних (постійно поповнюваної) постановка діагнозу та призначення лікування зводиться з формально математичної точки зору до вирішення задачі розпізнавання та оптимального управління.

Існуючі нині численні експертні системи діагностики не мають, як правило, математичного алгоритму, адекватного базі використовуваних даних. У них не проводиться оцінки правдоподібності аналізованих можливих діагнозів (гіпотез), а якщо й проводиться, то на підставі найпростіших кореляційних співвідношень без перевірки їхньої статистичної достовірності. У той самий час, відомо, що у разі малих вибірок на вирішення зворотних задач розпізнавання, потрібні додаткові методи аналізу логічної структури даних. При цьому через неповноту даних та їх неповну достовірність (суб'єктивність), завдання розпізнавання переходить у область нечіткої логіки. Ці та інші особливості роблять малоефективним застосування існуючих експертних систем у складних випадках захворювань та за браком даних. При цьому лікарю доводиться покладатися в основному на свою інтуїцію і досвід. Таким чином, задача підвищення інформативності та достовірності автоматичної системи диференціальної діагностики та лікування захворювань є надзвичайно актуальним.

5.2 Вибір вирішальних правил під час діагностики захворювань

Вважатимемо, що в загальному випадку діагноз і симптоми є випадковими подіями. При використанні регресійного аналізу, діагноз є

залежним, а симптоми є незалежними подіями (це лише терміни регресійного аналізу, що означають що від чого залежить). Нехай діагноз визначається багатофакторним рівнянням регресії деяких подій, що пояснюють діагноз – симптомів:

$$D_s = k_1 A_1 + k_2 A_2 + \cdots + k_p A_p + a, \qquad (5.1)$$

де k_i – коефіцієнти ваги подій, що пояснюють вплив симптому, A_i – подія – симптоми, а - помилка регресії. Сутність розв'язування задачі полягає в тому, що, визначивши за історичними даними k_i, знаючи події A_i можливо передбачити оптимальним способом методом найменших квадратів залежну подію D_s - діагноз із похибкою а, що має мінімальну дисперсію для всіх захворювань.

Розглянемо процедуру встановлення діагнозу. Нехай відомі всі можливі симптоми хвороби A_i, де $i = 1, \dots, p$. Природно вважати, що у пацієнтів можуть спостерігатися не всі симптоми A_i, а тільки їх частина. Крім того, у пацієнтів можуть бути симптоми, не характерні для деякої хвороби. Частина симптомів може бути загальними для різних хвороб. За наявності або відсутності симптому, значення A_i (5.1) приймає значення 1 або 0. Нижче буде показано, що правильно поставлений діагноз при наявності всіх симптомів приймає максимальне значення з усіх інших. Якщо діагноз не відповідає симптомам, його значення (5.1) істотно менше.

За історичними даними розглядаються вибірки симптомів, одержані для достовірного діагнозу. При цьому симптоми мають вигляд вибірок типу 1, 1, 1, 0, 1, …, 0, 0, 1, 1, 1. Після цього визначаються взаємні кореляції між симптомами, а також між симптомами та діагнозом, дисперсії симптомів, необхідні для вирішення системи рівнянь

$$R_{AA_j} = k_1 R_{A_1 A_j} + k_2 R_{A_2 A_j} + k_n R_{A_n A_j}. \qquad (5.2)$$

Якщо при вирішенні цього рівняння деякі з k_i занадто малі, їх можна вважати рівними нулю. Вирішуючи цю систему рівнянь можна оцінити коефіцієнти ваги k_i, які відрізнятимуться для різних хвороб. Зауважимо, що ці коефіцієнти, навіть за сильної кореляції, можуть мати різні знаки.

По суті, встановлення діагнозу, у такій постановці завдання, це ухвалення рішення про належність ознак (симптомів) до деякого класу документів (діагнозу). У кваліфікаційній роботі досліджувалися такі методи ухвалення рішення:

1. За коефіцієнтом детермінації на етапі навчання. Чим точніше регресія описує дані, тим менше помилка регресії і ближче коефіцієнт детермінації до 1. Якщо діагноз не описується набором коефіцієнтів регресії симптомів, лінійна апроксимація діагнозу симптомами не підходить. Тоді коефіцієнт детермінації менший.

2. За величиною залежної змінної – діагнозу. Теоретично правильний діагноз дорівнює одиниці. Його оцінне значення перебуває виваженим підсумовуванням симптомів. Коефіцієнти ваги визначаються на етапі навчання. Якщо симптоми відповідають діагнозу, їх зважене підсумовування, що визначає діагноз, має бути близьким до 1. Якщо вони не відповідають діагнозу, то результат має бути близький до нулю. Коефіцієнти ваги (регресії) розраховуються таким чином, щоб діагноз максимально був близьким до 1 на етапі навчання за правильними симптомами. Коефіцієнтами враховується випадковий характер прояву деяких симптомів, їх необов'язковий прояв, скритність, помилки у їх виявленні.

Крім цих вирішальних правил, для встановлення діагнозу можна приймати рішення за величиною взаємної кореляції характерних симптомів хвороби і виявлених симптомів на етапі розпізнавання хвороби. Також можна використовувати мінімум евклідової відстані між характерними симптомами хвороби та виявленими симптомами на етапі розпізнавання.

Перевага кожного вирішального правила визначається шляхом розпізнавання за умов зростання випадкових помилок. При моделюванні

навчання та розпізнавання необхідно підрахувати кількість помилок при постановці діагнозу та підрахувати загальну кількість поставлених діагнозів. При визначенні ефективного вирішального правила, слід підвищувати кількість помилок і визначати середнє помилковим діагнозам. Якісним показником вирішального правила може бути ймовірність помилкового діагнозу імітаційного експерименту.

5.3 Моделювання експерименту з автоматичної діагностики захворювання

Моделювання автоматичної системи встановлення діагнозу повинно включати такі етапи:

1. Введення даних пацієнта.
2. Етап навчання.
3. Етап розпізнавання (постановка первинного діагнозу).
4. Після встановлення первинного діагнозу, передбачається навчання і постановка більш точного вторинного діагнозу.
5. Вибір медикаментозного лікування.
6. При необхідності підбирається фізіотерапія, трави, дієта, вітаміни і т.д. у імітаційному експерименті ці етапи не розглядаються.

Моделювання проводилося у багатофункціональному редакторі електронних таблиць Microsoft Office Excel. Для моделювання постановки діагнозу вибиралися 4 захворювання печінки: Гострий гепатит, Хронічний гепатит, Гострий вірусний гепатит, Цироз. За літературними джерелами, кожної хвороби обрані супутні симптоми (всього 16). Нижче наведено 16 симптомів, які враховувалися при постановці 4 діагнозів захворювання, зазначених вище. Деякі симптоми загальні для різних діагнозів, тому розмальовка слів симптомів вказує, яким захворюванням (діагнозу) може відповідати даний симптом.

Таблиця 5.1 – Найбільш характерні симптоми для обраних хвороб

Симптоми	Гостр. гепатит	Хроніч. гепатит	Гостр. вірус. гепатит	Цироз
диспепсичний синдром	+	+	-	+
Збільшена печінка	+	+	+	-
болі	+	+	+	+
непереносимість продуктів дієти №5	+	-	-	-
відчуття тяжкості в правому підребер'ї	-	+	-	-
схуднення	-	+	-	+
дратівливість	-	+	-	-
жовтуха	-	+	+	-
шкірний зуд	-	+	-	+
збільшення печінки	-	+	-	-
астеничний синдром	-	+	-	+
гепатолієнальний синдром	-	+	-	-
синдром холестаза	-	+	-	-
синдром цитоліза	-	+	-	-
підвищення температури	-	-	-	+

Симптоми	Гостр. гепатит	Хроніч. гепатит	Гостр. вірус. гепатит	Цироз
блідість, що супроводжує анемію	-	-	-	+

Для навчання випадковим чином у 100 хворих ставився один із цих можливих діагнозів і розставлялися характерні симптоми. При виявленні симптомів використовувалося несуттєве переплутування (допускалися помилки) для ускладнення процедури розпізнавання. Для цих хвороб деякі симптоми загальні, що також робить помилки на етапі навчання і призводить до статистичного завдання. Кольори симптомів вказують на приналежність до захворювання. Зауважимо, що при моделюванні допускалася деяка неточність відповідно до діагнозу – симптом через некомпетентність авторів системи діагностики. Але це не вплинуло на математичну сторону вирішення завдання. Дані наведені в припущенні, що було 100 історій захворювання з відповідними симптомами і була точна (як виявилося згодом в процесі одужання) постановка діагнозу.

На етапі навчання отримано: коефіцієнти ваги регресійних рівнянь для кожної хвороби, їхнє середньоквадратичне відхилення, постійний член, коефіцієнт детермінації та деякі інші стандартні дані. Нижче наведено приклади розрахованих статистичних характеристик подій та перші 2 коефіцієнти шістнадцяти-факторної регресії. У першому рядку представлені коефіцієнти регресії, у другому рядку їх середньоквадратичні помилки. У третьому рядку наведено коефіцієнт детермінованості та стандартна помилка діагнозу. Для оцінювання коефіцієнта детермінованості порівнюються фактичні значення діагнозу та значення, що отримуються з рівняння прямої; за результатами порівняння обчислюється нормований коефіцієнт детермінованості, що лежить у межах від 0 до 1. Якщо він дорівнює 1, то має

місце повна кореляція з моделлю, тобто відмінностей між фактичним та оцінним значеннями діагнозу немає. У протилежному випадку, якщо коефіцієнт детермінованості дорівнює 0, використовувати рівняння регресії для передбачення значень діагнозу немає сенсу. У четвертому рядку наведено значення F-статистики та число ступенів свободи. Вони застосовуються з метою оцінки значимості використовуваної регресії.

Таблиця 5.2-Результати навчання при розпізнаванні хвороб за 16 симптомами

Гострий гепатит	-0,305259	-0,089504	-0,069228	-0,158168
	0,0884145	0,0465795	0,0620343	0,0665276
	0,9142635	0,1551725	#Н/Д	#Н/Д
	55,317672	83	#Н/Д	#Н/Д
	21,311482	1,9985171	#Н/Д	#Н/Д
Хронічний гепатит	-0,007077	0,1131511	-0,01524	-0,019590
	0,0577542	0,0304267	0,0405221	0,0434572
	0,9638046	0,1013619	#Н/Д	#Н/Д
	138,13208	83	#Н/Д	#Н/Д
	22,707238	0,8527620	#Н/Д	#Н/Д
Гострий вірусний гепатит	-0,311729	-0,021169	-0,020926	0,0056002
	0,0885360	0,0466435	0,0621195	0,0666190
	0,7773319	0,1553857	#Н/Д	#Н/Д
	18,109511	83	#Н/Д	#Н/Д
	6,9959875	2,0040124	#Н/Д	#Н/Д
Цироз	0,8306236	-0,033110	0,0084328	0,0887634
	0,0541559	0,0285310	0,0379974	0,0407496
	0,9377233	0,0950466	#Н/Д	#Н/Д
	78,110148	83	#Н/Д	#Н/Д
	11,290188	0,7498110	#Н/Д	#Н/Д

Великий коефіцієнт детермінації (близький до 1) в отриманих рівняннях регресії вказує на високу якість факторів, що пояснюють. Він становив для відповідних діагнозів хвороб 0,98, 0,95, 0,89, 0,94 для поставленого

експерименту. Такі значення свідчать, що регресія досить точно визначає діагноз.

Етап розпізнавання (постановка первинного діагнозу) моделювався для чотирьох хворих. Нижче наведено імітацію постановки діагнозу для 4-х хворих. Симптоми хвороб наведено у таблиці. Дані підбиралися таким чином, щоб у першого хворого був діагноз Гострий гепатит, у другого – Хронічний гепатит, у третього – Гострий вірусний гепатит, у четвертого – Цироз.

Таблиця 5.3- Етап розпізнавання (постановка первинного діагнозу)

Симптоми	1-й хворий	2-й хворий	3-й хворий	4-й хворий
блідість, що супроводжує анемію	0	0	0	1
підвищення температури	0	0	0	0
синдром цитоліза	0	1	0	1
синдром холестаза	0	0	0	0
гепатолієнальний синдром	0	1	0	0
астенічний синдром	1	1	0	1
збільшення печінки	0	1	0	0
шкірний зуд	0	1	1	1
жовтуха	1	0	1	1
дратівливість	0	1	0	0
Схуднення	0	1	0	1
відчуття тяжкості в правому підребер'ї	0	0	1	0
непереносимість продуктів з дієти №5	1	0	0	0
болі	1	1	1	0
збільшення печінки	1	0	1	0
диспепсичений синдром	1	0	0	1

Нижче наведено результати моделювання постановки діагнозу для 4-х хворих. У таблиці вказані значення діагнозу, розраховані за коефіцієнтами регресії, наведеними вище, отриманих на етапі навчання.

Таблиця 5.4-Результати розпізнавання хвороб за коефіцієнтом детермінації (1 приклад)

Діагноз	1 хворий	2 хворий	3 хворий	4 хворий
Гострий гепатит	0,92269866	0,10347161	-0,10351067	-0,0442538
Хронічний гепатит	-0,0456837	0,43937557	0,07403036	0,08896102
Гостр. вір. гепатит	0,11890148	0,19069169	1,00932809	-0,0142356
Цироз	-0,020186	0,12937275	0,05329434	0,95934841

Як і передбачалося, для відповідної хвороби сукупність симптомів давала максимальне значення за залежною змінною. Вони розташовані на діагоналі наведеної матриці. Аналогічні результати отримано і для іншого імітаційного експерименту.

Таблиця 5.5-Результати розпізнавання хвороб за коефіцієнтом детермінації (2-й приклад)

Діагноз	1 хворий	2 хворий	3 хворий	4 хворий
Гострий гепатит	0,85706634	0,09039669	0,07307219	-0,1296761
Хронічний гепатит	-0,0365935	0,5101952	0,10537681	0,09688167

Діагноз	1 хворий	2 хворий	3 хворий	4 хворий
Гостр. вір. гепатит	0,08088801	0,15227462	1,00451581	0,01626034
Цироз	0,02599127	0,24772315	0,05918802	0,98598644

ВИСНОВКИ

У кваліфікаційній роботі розглянуто питання створення системи розпізнавання хвороби щодо її симптомів на основі теорії прийняття рішень. Складність поставленого завдання полягає в нечисловій природі даних, що аналізуються. Запропонований підхід значною мірою імітує роботу лікаря, що приймає рішення при постановці діагнозу за сукупністю симптомів. Водночас облік статистичних даних, наукової літератури та досліджень дозволить більш точно поставити діагноз.

Було розроблено алгоритм імітаційного моделювання обліку впливу незалежних подій та випадкових процесів на випадкову подію чи процес. Для цього використовувався багатофакторний регресійний аналіз. Розроблено підхід до генерації випадкових подій та їх кореляційний аналіз.

Методом статистичного моделювання у програмі Excel було проведено імітаційні експерименти щодо постановки завдання діагностики та запропоновано критерій прийняття рішення. Проведені експериментальні дослідження підтвердили запропонований спосіб ухвалення рішення.

ПЕРЕЛІК ПОСИЛАНЬ

1. Кобринский Б.А., Казанцева Л.З., Фельдман А.Е. Автоматизированные системы дифференциальной диагностики наследственных заболеваний // Наследственная патология человека / Под общ. ред. Ю.Е. Вельтищева и Н.П. Бочкова. Т.II. - М., 1992. - С. 229-239.
2. Виноградов А.В. Дифференциальный диагноз внутренних болезней: справочное руководство для врачей. - М.: Медицина, 1987. – С. 496.
3. Эвербек Г. Дифференциальная диагностика болезней в детском возрасте: Пер. с нем. - М.: Медицина, 1980. – С. 623.
4. Ригельман Р. Как избежать врачебных ошибок. Книга практикующего врача: Пер. с англ. - М.: Практика, 1994. – С. 142.
5. Лукашевич И.П., Сыркин А.Л. Проблема получения и передачи медицинских знаний // Компьютерная хроника. - 1994. - №8-9. - С. 39-43.
6. Есенин-Вольпин А.С. Об антитрадиционной (ультраинтуиционистской) программе оснований математики и естественнонаучном мышлении // Семиотика и информатика. - 1993. - Вып.33. - С. 13-67.
7. Левенец Е.В. Рассуждения по аналогии // Логика и компьютер. 2: Логические языки, содержательные рассуждения и методы поиска доказательств. - М.: Наука, 1995. - С. 99-112.
8. Кобринский Б.А. Логика и интуиция специалиста в медицинских системах искусственного интеллекта // Научная сессия МИФИ-2000: Сб. науч. тр. Т.3. - М., 2000. - С. 64-65.
9. Клини С.К. Введение в метаматематику. - М.: Изд-во иностр. лит., 1957.
10. Таран Т.А. Формализация рассуждений на основе аргументации при принятии решений в конфликтных ситуациях // НТИ. Сер. 2. - 1998. - №9. - С.23-33.
11. Поспелов Д.А. Моделирование рассуждений. Опыт анализа мыслительных актов. - М.: Радио и связь, 1989. – С. 114.

12. Кулик Б.А. Основные принципы философии здравого смысла (познавательный аспект) // Новости искусственного интеллекта. - 1996. - № 3. - С. 7-91.
13. Финн В.К. Интеллектуальные системы: проблемы их развития и социальные последствия // Будущее искусственного интеллекта. - М.: Наука, 1991. - С.157-177.
14. Воробьев Н.В. Умозаключения по аналогии. - М., 1963. – С. 92.
15. Кобринский Б.А. К вопросу о формальном отражении образного мышления и интуиции специалиста в слабо структурированной предметной области // Новости искусственного интеллекта. - 1998. - №3. - С. 64-76.
16. Финн В.К. JSM-рассуждение как синтез познавательных процедур // 3-я Междунар. конф. "Информационные ресурсы. Интеграция. Технологии": Матер. конф. - М., 1997. - С. 207-208.
17. Есенин-Вольпин А.С. О теории модальностей // Философия. Логика. Поэзия. Защита прав человека: Избранное. - М.: Рос. гос. гуманит. ун-т, 1999. - С.165-177.
18. Финн В.К. Об интеллектуальных системах автоматизированной поддержки научных исследований // НТИ. Сер.2. - 1996. - №5-6. - С. 1-2.
19. Есенин-Вольпин А.С. О теории диспутов и логике доверия // Философия. Логика. Поэзия. Защита прав человека: Избранное. - М.: Рос. гос. гуманит. ун-т, 1999. - С. 178-192.
20. Справочник по прикладной статистике: Пер.с англ. / Под ред. Э. Ллойда, У. Ледермана, С.А. Айвазяна, Ю.Н. Тюрина. - М.: Финансы и статистика, 1990. - Т.2. – С. 526.