**Лек.1 (Вводна)  
Введення в ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ**

**1. Початкові визначення і поняття. 2. Завдання, методи і засоби ИАД 3. Варіативне моделювання  
Вступ.**  
Характерною рисою останніх тенденцій комп'ютерного аналізу і інтерпретації даних і ухвалення рішень є бурхливий розвиток технологій і засобів "здобичі", "витягання" знань з даних, "інтелектуального аналізу даних" (ИАД) [1-5]. Для початкової стадії розвитку цього напряму, що природно, характерна наявність несталих понять цієї предметної області. У лекції зроблена спроба до певної міри заповнити цей пропуск.  
1. **Початкові визначення і поняття**  
Передусім, визначимо поняття "інформація". Визначимо його як "сукупність змістовних відомостей, які можуть бути вироблені, зібрані, передані, збережені, перероблені, відтворені

і використані".  
Наступними важливими поняттями є терміни: сигнал, дані і знання [7,8]. Їх визначення дані в [7] (см також [8]). Сигнал - матеріальний носій інформації - фізична речовина, поле, процес. Дані - набір символів і записів, що представляються ними, що розглядається безвідносно до змістовного сенсу. **Знання** - перевірені практикою результати пізнання дійсності, корисні відомості, які можуть багаторазово використовуватися людьми. Акцентуємо увагу на наступних аспектах приведених визначень.  
**Перше.** Схожість сигналів і даних - вони носії інформації, а не сама інформація. Їх відмінність зводиться до того, що інформація, що міститься в сигналі, недоступна безпосередньому сприйняттю суб'єктом. Дані ж це, по-перше, сукупність символів, що містять інформацію, по-друге, сукупність, доступна для сприйняття людиною. З точки зору рішення конкретної задачі усю інформацію про деякий об'єкт можна умовно вважати такою, що складається з трьох взаємозв'язаних

пересічних, нечітко роздільних частин: знання, протознания і інформаційне сміття. Знання на відміну від сигналів і даних це вже елемент інформації, що відрізняється логічною повнотою, обмеженістю набору, і, в завершеному варіанті, проверенностью практикою. Вони (знання) включають усю релевантну інформацію [8], тобто ту частину "корисного" для вирішення завдання інформації, яка повністю усуває (з точки зору вирішуваної задачі!) невизначеність про об'єкт у одержувача, що стоїть. Релевантна інформація, у свою чергу, є частина корисної (для одержувача) інформації - сукупності відомостей, що зменшують міру невизначеності про даний об'єкт у її одержувача. На відміну від знань корисна інформація може бути логічно не повною, суперечливою, не завжди достовірною і тому подібне. Знання ж це корисна інформація, представлена у виді, зручному для її інтелектуального "переварювання". "Протознания" - це та частина інформації, з якої можуть бути отримані нові знання.

Інша частина інформації, - це інформаційне сміття - відомості, що містяться в "даних", які не містять корисної для користувача інформації і/або протознаний, але багаторазово збільшують витрати користувача. У цьому сенсі сигнали, дані і знання, як носії корисної інформації і інформаційного сміття, надалі умовно означатимемо одним словом "дані".  
Наступні важливі узагальнювальні збиральні поняття - це збір, обробка, аналіз, інтерпретація і застосування "даних".  
Назвемо "збором "даних" дії з "даними", семантики, що не призводять до зміни.  
"**Обробкою "даних"** (сигналів, даних, знань) називатимемо дії, спрямовані на приведення що цікавить дослідника, інформації, що міститься в них, до зручнішого, компактному для аналізу виду. Помітимо, що при обробці "даних" кількість інформації, що міститься в них, не може бути збільшена. Тут відбувається лише перетворення інформації до виду, зручнішого для подальших операцій.

Під ним розумітимемо аналіз, пов'язаний, поперше, саме з неформальним витяганням знань про досліджуваний об'єкт, породжувачі "дані", безпосередньо з цих "даних" (АИД- 1), подруге, з отриманням нових знань про об'єкт на базі витягнутих знань, віртуальних\*) "даних" про об'єкт і природного і/або штучного інтелекту (ИАД- 2), по-третє, з пошуком, вибором, синтезом методів і засобів обробки і аналіз "дані" з урахуванням поставлений ціль дослідження об'єкт (ИАД- 3).  
Інтелектуальний аналіз "даних" пов'язаний з "додаванням інформації" про об'єкт і/або методи і засоби аналізу, з привнесенням її "інтелектом". Під "інтерпретацією результатів обробки і аналізу "даних" розумітимемо тлумачення, роз'яснення сенсу, значення, їх "переклад" мовою, в терміни, образи доступні і зрозумілі користувачеві.  
Під "застосуванням результатів обробки і аналізу "даних" розуміються дії, пов'язані з використанням їх (результатів) для вирішення теоретичних і практичних

завдань, з реалізацією технології, з досягненням поставленої користувачем кінцевої мети. Помітимо, що будь-яка формалізація, структуризація, обробка "даних" привносить до них елементи знань.  
Наступні важливі поняття це англомовні терміни "Data Mining", "Data - Based Knowledge", "Data Cleaning", "Data Warehouse", "Metadata".  
Зазвичай під технологією "Data Mining" (DM) розуміють пошук прихованих залежностей, що характеризують об'єкт, і взаємозв'язків, проявляються через дані про нього (см, наприклад [4,5,10]). "Data - Based Knowledge" (DBK) - знання, грунтовані на "даних", тобто знання, які "витягаються" (виводяться) за допомогою інтелектуальних інструментальних засобів аналізу (Business Intelligence Tools - BIT) з сховища "даних". "Data Cleaning" (DС) очищення "даних" - є усунення з даних або коригування тієї їх частини, яка містить легко визначуване "інформаційне сміття" : помилки, протиріччя, збої, явні викиди і тому подібне. Ця операція зазвичай використовується перед завантаженням "даних"

"Data Warehouse" (DW) - сховище даних - це база, в якій збираються "дані" для подальшого їх аналізу під конкретну вирішувану задачу. "Metadata" - "дані", що описують "дані" в DW.  
Нарешті, ще одним важливим поняттям є термін "управління знаннями". Під ним з подання фахівців Anderson Consalting зазвичай розуміють систематичне придбання, синтез, обмін і використання (мабуть, корисних?!) ідей і досвіду для досягнення успіху у бізнесі або в управлінні компанією. Іншими словами, це процес використання того, що відомо людям, на новому рівні, з метою підвищення потенціалу компанії через використання кращих рішень, інтелектуального капіталу або організаційного навчання, а також через "капіталізацію інтелекту, знань".

**2. Завдання, методи і засоби ИАД**  
Ясно, що, залежно від додатків, найбільш значимим, первинним для кожного застосування являється свій перелік завдань і найбільш ефективні методи його рішення.  
З найчастіше вирішуваних методами ИАД завдань в додатку до практики являються: аналіз значимих чинників, скорочення або збільшення їх числа, виявлення залежностей, асоціацій, виключень і закономірностей, у тому числі для зменшення розмірності факторного простору або для віртуального (збору, обробки) аналізу "даних"; класифікація; моделювання і прогноз; ранжирування; сегментація; профілізація найкращих досягнень. До найчастіше використовуваним методам, технологіям рішення цих завдань відносяться методи дерев рішень, нейронних мереж, математичної статистики, експертного аналізу і нечіткої логіки, візуалізації, генетичних алгоритмів, еволюційного програмування, прецедентів, варіативного моделювання, а також інтегровані методи і технології.  
З сучасного програмного забезпечення, використовуваного як засоби ИАД, відмітимо наступні.  
ПО сховищ даних [10]: Oracle (фірма Arbor), Platinum Technology (Business Objests), Praxis (Carleton), Prism (Cognos)

Pyramid (Hewlett - Packard), Red Brick (IBM), SAS Institute (Information Bulders), Sequent (Informix), Software AG (Intellidex), Sybase (Microsoft), Tandem (MSP).  
ПО для реалізації технологій "Data Mining" [4,5]: Poly Analyst, Scenario, 4 Thought, MineSet.  
Інструментарій (платформи і додатки) управління знаннями пропонується компаніями Glyphica (система Portalware), Autonomy (Portal - in - a - Box, Content Server та ін.), Plumtree Software (Plumtree Server), Hyperknowledge (Hyperknowledge Builder), Intraspect Software (Intraspect Knowledge Server 2.0), Documentum (Documentum Enterprise Document Management System - EDMS), Open Text (Livelink) та ін.  
**3. Варіативне моделювання**  
Варіативне (від англ. variety - різноманітність, різновид), або варіантне моделювання (VM) (об'єкту) є метод дослідження, грунтований на заміні досліджуваного об'єкту-оригіналу набором різноманітних моделей його і на роботі з ними

Таким чином, відмітна особливість варіативного моделювання від звичайного (класичного) полягає в тому, що тут обов'язковою є побудова і спільне застосування в процесі моделювання не менше двох різних моделей досліджуваного (модельованого) об'єкту. Це можуть бути моделі різних класів (см, наприклад [6]), одного класу, але різних типів, складнощів; описи об'єкту, що використовують різні рівні, різні засоби і технології їх побудови, інтерпретації і застосування і тому подібне. Як випливає з приведених визначень, якщо DM- технології є технології, реалізовуючі ИАД- 1, тобто перший варіант ИАД, DBK - ИАД- 2 (другий варіант), то VM- технології реалізують третій варіант ИАД - ИАД  
Технологія виявлення знань  
У поточній версії підсистеми реалізовані методи, що отримали найбільше комерційне поширення у світовій практиці, а саме:  
- кластеризація - реалізує угрупування об'єктів, максимізувавши внутрішньогрупову схожість і міжгрупові відмінності;

- дерево рішень - забезпечує побудову причинно-наслідкової ієрархії умов, що призводить до певних рішень;  
- пошук асоціацій - виконує пошук стійких комбінацій елементів в подіях або об'єктах;  
- розпізнавання класів  
**Методи виявлення знань**  
**Кластеризація**  
Мета кластеризації - виділення з безлічі об'єктів однієї природи деякої кількості відносно однорідних груп (сегментів або кластерів). Об'єкти розподіляються по групах так, щоб внутрішньогрупові відмінності були мінімальними, а міжгрупові - максимальними. Методи кластеризації дозволяють перейти від пообъектного до групового представлення сукупності довільних об'єктів, що істотно спрощує оперування ними. Нижче описані декілька можливих сценаріїв застосування кластеризації на практиці.

Сегментація клієнтів по певній сукупності параметрів дозволяє виділити серед них стійкі групи, що мають схожі купівельні переваги, рівні продажів і платоспроможності, що значно спрощує управління взаємовідносинами з клієнтами.  
При класифікації товарів дуже часто використовуються досить умовні принципи класифікації. Виділення сегментів на основі групи формальних критеріїв дозволяє визначити дійсно однорідні групи товарів. В умовах широкої і досить різнорідної номенклатури товарів управління асортиментом на рівні сегментів в порівнянні з управлінням на рівні номенклатури істотно підвищує ефективність просування, ціноутворення, мерчендайзинга, управління ланцюжками постачань.  
Сегментація менеджерів дозволяє ефективніше спланувати організаційні зміни, поліпшити мотиваційні схеми, скоректувати вимоги до найманого персоналу, що кінець кінцем дозволяє підвищити

керованість компанії і стабільність бізнесу в цілому.  
Спосіб виміру відстані залежить від метрики, яка вказує принцип визначення схожості/відмінності між об'єктами вибірки. Поточна реалізація підтримує наступні метрики:  
- "метрика Евкліда" - ця стандартна відстань між двома точками в N- мірному просторі Евкліда атрибутів;  
- метрика "Евкліда в квадраті" - посилює вплив відмінності (відстані) на результат кластеризації;  
- "метрика домінування" - визначає відмінність між об'єктами вибірки як максимальну з існуючих різницю між значеннями їх атрибутів, тому корисна для посилення відмінностей між об'єктами по одному атрибуту.  
Спосіб формування кластерів на основі інформації про відстань між об'єктами, що кластеризуються, визначається методом кластеризації. Наприклад, в поточній версії "1С :Предприятие 8.0" реалізовані наступні методи кластеризації :

- "ближньому зв'язку" - об'єкт приєднується до тієї групи, для якої відстань до найближчого об'єкту мінімально;  
- "телекомунікації" - об'єкт приєднується до тієї групи, для якої відстань до найбільш далекого об'єкту мінімально;  
- "центру тяжіння" - об'єкт приєднується до тієї групи, для якої відстань до центру кластера мінімально;  
- метод "k-средних" - вибираються довільні об'єкти, які вважаються центрами кластерів, потім усі аналізовані об'єкти послідовно перебираються і приєднуються до найближчого до них кластера. Після приєднання об'єкту розраховується новий центр кластера, який обчислюється як середнє значення атрибутів усіх об'єктів, що входять в кластер. Процедура повторюється до тих пір, поки змінюються центри кластерів.  
Будьякий з реалізованих в платформі методів кластеризації припускає явну вказівку кількості шуканих кластерів. Для атрибутів об'єктів можна вводити ваги, що дозволяє розставляти пріоритети між ними.

В результаті аналізу за допомогою кластеризації отримують наступні дані:  
- центри кластерів, що є сукупністю усереднених значень вхідних колонок в кожному кластері;  
- таблицю міжкластерних відстаней (відстані між центрами кластерів), що визначають міру відмінності між ними;  
- значення прогнозних колонок для кожного кластера;  
- рейтинг чинників і дерево умов, що визначили розподіл об'єктів на кластери.

**Литература**

1. Аджиев В. MiniSet-визуальный инструмент аналитика//Открытые системы, 1997.- № 3.- С. 73-77.
2. Киселев М., Соломатин Е. Средства добычи знаний в бизнесе и финансах. // Открытые системы, 1997. - №4. - С. 41-44.
3. Рузайкин Г.И. Орудие Data Mining: успех в анализе данных // Мир ПК, 1997. - № 1. - С. 102-103.
4. Шапот М. Интеллектуальный анализ данных в системах поддержки принятия решений // Открытые системы, 1998. - № 1. - С. 30-35.
5. Шапот М., Рощупкина В. Интеллектуальный анализ данных и управление процессами // Открытые системы, 1998. - № 4,5.
6. Губарев В.В. Информатика в рисунках и таблицах (фрагменты системного путеводителя по концептуальным основам) : Изд-во НГТУ, 1998. - 152 с.
7. Gubarev V.V. Experimental Data Analysis in the Systems Context // Proceedings The Third Russian-Korean International Symposium on Science and Technology. - Novosibirsk: NSTU, 1999. - Vol.1. - P.241-244; // Abstracts. - Vol.1. - P.190.
8. Головко М. Жизнь в мусорной куче, или нужны ли нам знания? // Computerwold Россия. - 17.08.1999. - С.41-43.
9. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. - Новосибирск: Изд-во ИМ СО РАН, 1999. - 270 с.
10. Туманов В. Data Warehouse: с чего начать? // PC WEEK, 1999. - № 29. - С. 15-16.

………………………………………………………………………….

Програма вивчення обов’язкової навчальної дисципліни «Методи та системи штучного інтелекту», складена відповідно до освітньо-професійної програми підготовки бакалавра.

**Предметом** вивчення навчальної дисципліни є поняття, які містяться в основі теорії вірогідності, математичної статистики, теорії інформації, теорії оптимального управління.

**Міждисциплінарні зв’язки**: нормативна навчальна дисципліна «Методи та системи штучного інтелекту», є складовою циклу професійної підготовки фахівців освітньо-кваліфікаційного рівня «бакалавр», є базовою для вивчення дисципліни «Прикладні системи штучного інтелекту».

Базується на курсі дисципліни «Вища математика», «Дискретна математика», «Теорія ймовірностей, ймовірнісні процеси та математична статистика».

**Програма навчальної дисципліни складається з таких змістових модулів:**

1. Інтелектуальний аналіз даних.

2. Регресійний аналіз даних.

3. Кластерний аналіз даних.

**1. Мета та завдання навчальної дисципліни**

1.1. Метою викладання навчальної дисципліни «Методи та системи штучного інтелекту», є формування у студентів знань, умінь, і навиків логічного і творчого мислення, необхідних для вирішення теоретичних і практичних завдань; оволодіння ними основних методів дослідження і дозволу практичних завдань управління складними об'єктами на основі інтелектуальних інформаційних технологій, необхідних для здобуття професійних знань.

1.2.Основними завданнями дисципліни «Методи та системи штучного інтелекту», є

* вивчення основних методів дослідження і розв’язку практичних завдань управління складними об'єктами на основі інтелектуальних інформаційних технологій, необхідних для здобуття професійних знань.

1.3. Згідно з вимогами освітньо-професійної програми студенти повинні:

***знати :***

* засади та методи складання математичних моделей фізичних процесів і складних систем на основі фундаментальних законів та застосовувати ці закони для характеристики стан процесу.
* способи витягання знань з масиву емпіричних даних в інтелектуальних системах.
* опис класів формалізованих евристик.
* виклад архітектури інтелектуальних систем, призначених для реалізації правдоподібних емпіричних міркувань технічними засобами;
* застосовування технології інтелектуальної обробки даних для вирішення практичних завдань класифікації, регресії, кластеризації при управлінні складними об'єктами, процесами, станами.
* вибір та обґрунтування засобів інтелектуалізації інформаційних систем.
* розробку та впровадження експертних систем.
* проектування нейрокомп’ютерів та прикладних нейросистем.
* розробку та впровадження баз знань в прикладних областях.
* інтелектуалізація інтерфейсу інформаційних систем.
* розробку та впровадження прикладних нейротехнологій.
* проектування інтелектуальних агентних систем.

***вміти*** :

* проводити збір, систематизацію та аналіз науково-технічної інформації з питань розробки та застосування систем штучного інтелекту та нейронних мереж.
* визначати методи та засоби для вирішення прикладних задач штучного інтелекту та розробляти конкретні вимоги до відповідних засобів.
* обґрунтовувати архітектуру та вимоги до компонент прикладних нейросистем та систем штучного інтелекту.
* розробляти методики формування масивів даних для навчання та тестування нейронних мереж та прикладних систем штучного інтелекту.
* розробляти та реалізовувати у вигляді комп’ютерних програм алгоритми штучного інтелекту, нейронних мереж та засобів пре- та постпроцесінгу даних.
* забезпечувати надійність та ефективність програм штучного інтелекту та нейросистем.

На вивчення навчальної дисципліни відводиться 120 годин / 4 кредитів ECTS.