Міністерство освіти і науки України

Національний університет водного господарства та природокористування

Навчально-науковий інститут автоматики, кібернетики та обчислювальної техніки

Кафедра прикладної математики

ЗАТВЕРДЖУЮ

Проректор з науково-педагогічної,
методичної та виховної роботи
_____ О.А. Лагоднюк

“_____”_____ 20___ р.

04-01-18

РОБОЧА ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

Program of the Discipline

Спеціальні методи інтелектуального аналізу даних

Special methods of data mining

спеціальність	122 "Комп’ютерні науки "
specialty	122 "Computer science "

Рівне – 2017

Робоча програма дисципліни “ Спеціальні методи інтелектуального аналізу даних ” для аспірантів, які навчаються за спеціальністю 122 "Комп’ютерні науки та інформаційні технології", 15 с.

Розробник: Турбал Юрій Васильович, д.т.н., професор кафедри прикладної математики.

Робочу програму схвалено на засіданні кафедри прикладної математики

Протокол від «06» грудня 2016 року № 4

Завідувач кафедри _____ П.М. Мартинюк

Схвалено науково-методичною комісією за спеціальністю 122 “Комп’ютерні науки та інформаційні технології”

Протокол від «12» грудня 2016 року №3

Голова науково-методичної комісії _____П.М. Мартинюк

ВСТУП

Програма навчальної дисципліни “ Спеціальні методи інтелектуального аналізу даних” складена відповідно до освітньо-наукових програм підготовки здобувачів третього (освітньо-наукового) рівня спеціальності 122 "Комп’ютерні науки ”.

Предметом вивчення навчальної дисципліни є перелік питань, розгляд яких становить передумови для успішної роботи над науковим дослідженням та які стосуються,зокрема, методів екстраполяції та прогнозування. Результатом вивчення дисципліни є готовність здобувача до аналізу отриманих ним результатів спостережень, що описуються вимірюваними параметрами та готовність здійснювати прогнозні оцінки.

Анотація

Метою програми є вивчення методів сучасної обробки даних – інтелектуального аналізу даних (Data Mining, Knowledge Discovery in Data), аналітичного дослідження великих масивів інформації з метою виявлення нових раніше невідомих, практично корисних знань і закономірностей, необхідних для прийняття рішень; огляд методів, програмних продуктів і різних інструментальних засобів, які використовуються Data Mining; розгляд практичних прикладів застосування Data Mining; підготовка здобувачів до самостійної роботи з вирішення задач засобами Data Mining і розробки інтелектуальних систем.

В межах програми розглядаються п’ять змістових модулів: “Методологічні основи інтелектуального аналізу даних”, “Алгоритми Data Mining: класифікація, регресія, прогнозування”,” Інтелектуальний аналіз часових рядів”, “Алгоритми Data Mining: кластеризація, пошук асоціативних правил”, “Сховища даних та оперативний аналіз даних (OLAP)”.

Особливістю курсу є вивчення підходів до інтелектуального аналізу даних за умов малих виборок та сучасні підходи до прогнозування.

Ключові слова: аналіз даних, інтелектуальні системи, Data Mining, Knowledge Discovery in Data, часові ряди, регресійні моделі, кластеризація, класифікація, асоціативні правила .

Abstract

The purpose of the program is to study methods of modern data processing - data mining (Data Mining, Knowledge Discovery in Data), analytical research of large amounts of information in order to identify new previously unknown, practically useful knowledge and patterns necessary for decision making; an

overview of the methods, software and various tools used by Data Mining; Consideration of practical examples of Data Mining application; preparation of applicants for independent work on solving problems by means of Data Mining and development of intelligent systems.

The program covers five content modules: "Methodological Foundations of Data Mining", "Data Mining Algorithms: Classification, Regression, Prediction", "Time Series Intelligence", "Data Mining Algorithms: Clustering, Finding Associative Rules", " Data Warehouses and Operational Data Analysis (OLAP) ".

A feature of the course is the study of approaches to data mining in the context of small samples and modern approaches to forecasting.

Keywords: data analysis, intellectual systems, Data Mining, Knowledge Discovery in Data, time series, regression models, clustering, classification, associative rules.

1. Опис навчальної дисципліни

<table>
<tr><td rowspan="2">Найменування показників</td><td rowspan="2">Галузь знань, спеціальність, спеціалізація, рівень вищої освіти</td><td colspan="2">Характеристика навчальної дисципліни</td></tr>
<tr><td>денна форма навчання</td><td>заочна форма навчання</td></tr>
<tr><td rowspan="2">Кількість кредитів – 4</td><td>Галузь знань
12 Інформаційні технології</td><td colspan="2" rowspan="2">Вибіркова</td></tr>
<tr><td>Спеціальність
122 "Комп’ютерні науки та інформаційні технології”.</td></tr>
<tr><td></td><td rowspan="5">Спеціалізація
-</td><td colspan="2">Рік підготовки</td></tr>
<tr><td>Змістових модулів – 3</td><td>2-й</td><td>2-й</td></tr>
<tr><td></td><td colspan="2">Семестр</td></tr>
<tr><td rowspan="2">Загальна кількість годин – 120</td><td>4-й</td><td>4-й</td></tr>
<tr><td colspan="2">Лекції</td></tr>
<tr><td rowspan="10">Тижневих годин для денної форми навчання:
аудиторних – 4
самостійної роботи – 7</td><td rowspan="10">Рівень вищої світи:
PhD</td><td>20 год.</td><td>4 год.</td></tr>
<tr><td colspan="2">Практичні, семінарські</td></tr>
<tr><td>-</td><td>-</td></tr>
<tr><td colspan="2">Лабораторні</td></tr>
<tr><td>20 год.</td><td>8 год.</td></tr>
<tr><td colspan="2">Самостійна робота</td></tr>
<tr><td>80 год.</td><td>108 год.</td></tr>
<tr><td colspan="2">Індивідуальні завдання:</td></tr>
<tr><td colspan="2">-</td></tr>
</table>

		Форма контролю:	
		зал.	зал.

Примітка.

Співвідношення кількості годин аудиторних занять до самостійної та індивідуальної роботи становить (%):

для денної форми навчання – 33% до 67%;

для заочної форми навчання – 10% до 90%.

2.Мета та завдання навчальної дисципліни

Мета: вивчення методів сучасної обробки даних – інтелектуального аналізу даних (Data Mining, Knowledge Discovery in Data), аналітичного дослідження великих масивів інформації з метою виявлення нових раніше невідомих, практично корисних знань і закономірностей, необхідних для прийняття рішень; огляд методів, програмних продуктів і різних інструментальних засобів, які використовуються Data Mining; розгляд практичних прикладів застосування Data Mining; підготовка здобувачів до самостійної роботи з вирішення задач засобами Data Mining і розробки інтелектуальних систем.

Завдання:

У результаті вивчення навчальної дисципліни здобувач повинен

з н а т и :

- основні поняття, задачі та стадії інтелектуального аналізу даних;
- підходи до збереження, представлення та обробки інформації в сучасних інформаційних системах;
- методи побудови моделей та аналізу залежностей у великих та малих масивах даних;
- сучасні програмні засоби для проектування і розробки систем інтелектуального аналізу даних;
- концепції сховищ даних, їх оперативної аналітичної обробки;

в м і т и :

- обґрунтовувати вибір конкретного типу моделі та методу інтелектуального аналізу даних при вирішенні поставленої практичної задачі; –
- проводити необхідну попередню обробку даних, визначати тип задачі аналізу, вирішувати її адекватно обраним методом з оптимально визначеними параметрами, оцінювати результати, робити змістовні висновки та інтерпретацію;
- використовувати сучасні програмні засоби для проектування та дослідження систем інтелектуального аналізу даних;
- застосовувати технології роботи зі сховищами даних, здійснювати їх аналітичну обробку та інтелектуальний аналіз для забезпечення надійної роботи інформаційних систем;
- проектувати інформаційне забезпечення (логічну та фізичну структури баз даних) інформаційних систем.

3. Програма навчальної дисципліни

Змістовий модуль 1. Методологічні основи інтелектуального аналізу даних.

Тема 1. Основи інтелектуального аналізу даних. Визначення Data Mining і область застосування. Задачі, моделі та методи Data Mining. Методи, стадії, задачі Data Mining. Поняття Business Intelligence.

Тема 2. Процес виявлення знань. Цикл одержання, попередньої обробки, аналізу даних, інтерпретації результатів та їхнього використання. Етапи процесу Data Mining, пов'язані з побудовою, перевіркою, оцінкою, вибором и корекцією моделей. Методи первісної обробки даних. Інструментальні засоби Data Mining. Методи дослідження структури даних: візуалізація даних.

Змістовий модуль 2. Алгоритми Data Mining: класифікація ,регресія, прогнозування

Тема 3. Задачі класифікації . Постановка задачі класифікації та представлення результатів. Методи побудови правил класифікації. Методи побудови дерев рішень. Методи побудови математичних функцій. Методи опорних векторів, «найближчого сусіда», Байеса. Аналіз багатомірних угрупувань. Класифікація об'єктів у випадку невідомих розподілень даних. Методи оцінювання помилок класифікації.

Тема 4. Задачі регресії Регресія та її види. Інтерпретація параметрів регресії. Методи вирішення задач регресії. Регресія та прогнозування.Системи функцій Чебишева . Проблема моментів та задачі аналізу даних. Степенева та загальна проблема моментів, алгоритми розв'язання.

Тема 5. Прогнозування за умов недостатньої кількості даних.

Специфіка побудови прогнозів за умов недостатньої кількості даних, проблема детермінованості та недетермінованості моделей. Обмеження методів, переваги та недоліки стохастичних та детермінованих підходів.Застосування пірамідального підходу в умовах малих виборок у поєднанні з методами згладжування.

Змістовий модуль 3. Інтелектуальний аналіз часових рядів

Тема 6. Методи аналізу часових рядів Часовий ряд. Визначення й типологія часових рядів. Компоненти часових рядів. Основні показники часового ряду. Прогнозування на основі часового ряду. Тренд, циклічні коливання, сезонні коливання, нерегулярна компонента. Адитивна й

мультиплікативна моделі часового ряду. Основні характеристики часових рядів. Перевірка стаціонарності часових рядів. Критерій Стьюдента, критерій Фішера. Метод поворотних точок. Поняття нечітких часових рядів. Методи моделювання часових рядів. Методи аналізу та прогнозування поведінки часових рядів.

Змістовий модуль 4. Алгоритми Data Mining: кластеризація, пошук асоціативних правил

Тема 7. Задачі кластеризації Постановка задачі кластеризації та представлення результатів. Види кластерів. Міри близькості, засновані на відстанях. Базові алгоритми кластеризации. Адаптивні методи кластеризации.

Тема 8. Вирішення задачі пошуку асоціативних правил Постановка задачі пошуку асоціативних правил та представлення результатів. Секвенціальний аналіз. Різновиди задач пошуку асоціативних правил. Методи подання результатів. Алгоритми пошуку асоціативних правил. Методи пошуку асоціативних правил: метод Apriori, побудова FP-дерев пошуку шаблонів даних. Min-max асоціації у базах даних. Побудова hash-дерев.

Змістовий модуль 5. Сховища даних та оперативний аналіз даних (OLAP)

Тема 9. Сховища даних. Визначення сховища даних, порівняння з базами даних, використання.. Архітектура сховища даних. ETL-процеси (добування, перетворення й завантаження даних). Вітрини даних, куби даних, багатомірна модель даних.

Тема 10. Оперативний аналіз даних Розгортання OLAP-кубів. Операції над OLAP-кубами (зріз, обертання, консолідація, деталізація). Архітектура OLAP-систем: MOLAP, ROLAP, HOLAP.

Структура навчальної дисципліни

Назви змістових модулів і тем	Кількість годин											
	денна форма						заочна форма					
	усього	у тому числі					усього	у тому числі				
		л	п	лаб.	інд.	с. р.		л	п	лаб.	інд.	с. р.

1	2	3	4	5	6	7	8	9	10	11	12	13
Змістовий модуль 1. Методологічні основи прогнозування.												
Тема 1. Основи інтелектуального аналізу даних.	12	2		2		8	12	1		1		10
Тема 2. Процес виявлення знань.	12	2		2		8	12	1		1		10
Разом за змістовим модулем 1	24	2		2		16	24	2		2		20
Змістовий модуль 2. Алгоритми Data Mining: класифікація, регресія, прогнозування												
Тема 3. Задачі класифікації	12	2		2		8	12					12
Тема 4. Задачі регресії	12	2		2		8	12					12
Тема 5. Прогнозування за умов недостатньої кількості даних.	12	2		2		8	12					12
Разом за змістовим модулем 2	36	6		6		24	36					36
Змістовий модуль 3. Інтелектуальний аналіз часових рядів												
Тема 6. Методи аналізу часових рядів	12	2		2		8	12					12
Разом за змістовим модулем 3	12	2		2		8	12					12
Змістовий модуль 4. Алгоритми Data Mining: кластеризація, пошук асоціативних правил												
Тема 7. Задачі кластеризації	12	2		2		8	12			1		11
Тема 8. Вирішення задачі пошуку асоціативних правил	12	2		2		8	12			1		11
Разом за змістовим модулем 4	24	4		4		16	24			2		22
Змістовий модуль 5. Сховища даних та оперативний аналіз даних (OLAP)												
Тема 9. Сховища даних.	12	2		2		8	12	1		1		10
Тема 10. Оперативний аналіз даних	12	2		2		8	12	1		1		10
Разом за змістовим модулем 5	24	4		4		16	24	2		2		20
Усього годин	120	20		20		80	120	4		8		108

5. Теми лабораторних занять

№ з/п	Назва теми	Кількість годин	
		денна форма	заочна форма
1	Тема 1. Знайомство з програмою інтелектуального аналізу даних WEKA та підготовка даних	2	1
2	Тема 2 Задача класифікації	2	
3	Тема 3 Прогнозування, задача регресії	2	1
4	Тема 4 Задача кластеризації	2	1
5	Тема 5 Пошук асоціативних правил	2	1
6	Тема 6. Прогнозування за умов недостатньої кількості даних.	2	
7	Тема 7. Методи прогнозування багатомірних процесів .	2	
8	Тема 8. “Пірамідальний” метод екстраполяції та його особливості.	2	
9	Тема 9. Сховища даних.	2	2
10	Тема 10. Оперативний аналіз даних	2	2
	Разом	20	8

6. Самостійна робота

Розподіл годин самостійної роботи для здобувачів денної форми навчання:

Підготовка до аудиторних занять – 0,5 год/1 год. занять.

Підготовка до контрольних заходів – 6 год. на 1 кредит ЄКТС.

Опрацювання окремих тем програми або їх частин, які не викладаються на лекціях.

6.1. Теми для самостійної роботи

№ з/п	Назва теми	Кількість годин	
		денна форма	заочна форма
1	Тема 1. Особливості обробки даних. 5. Модель Map Reduce		
2	Тема 2. Візуальний аналіз даних. Методи візуалізації		
3	Тема 3. Стандарти Data Mining: CWM, CRISP, PMML		
4	Тема 4. Методи класифікації. Алгоритм побудови елементарних правил (1- rule), алгоритм Naive Bayes.	8	10
5	Тема 5. Точність класицікації. Оцінка рівня помилок.	8	10
6	Тема 6. Застосування нейронних мереж для задач класифікації	8	12
7	Тема 7. Дерева прийняття рішень. Алгоритм найближчого сусіда	8	12
8	Тема 8. Постановка задачі пошуку асоціативних правил, її різновиди. Представлення результатів. Алгоритм Apriori та його різновиди.	8	10
9	Тема 6. Ієрархічні алгоритми кластеризації. Алгоритм k-Means.	8	10
10	Тема 10. Адаптивні методи кластерізації.	8	12
	Разом	80	108

7. Методи навчання

1) Лекції проводяться з використанням технічних засобів навчання і супроводжуються демонстрацією за допомогою відеопроектора лекційного матеріалу.

2) Лабораторні роботи проводяться в комп'ютерному класі з використанням роздаткового матеріалу, методичних вказівок.

8. Методи контролю

Для визначення рівня засвоєння здобувачами навчального матеріалу використовуються такі методи оцінювання:

1) поточний контроль проводиться на лабораторних заняттях шляхом усного опитування і перевірки виконаних лабораторних робіт та домашніх завдань;

2) виконання додаткових індивідуальних завдань під час лабораторних робіт і консультацій;

3) поточне тестування після вивчення кожного змістового модуля;

Введена кредитно-трансферна система організації навчального процесу зі 100-бальною шкалою оцінювання знань здобувачів.

Усі форми контролю включені до 100-бальної шкали оцінювання.

Оцінювання здобувачів проводиться відповідно до вимог ECTS.

9. Розподіл балів, які отримують здобувачі

Поточне тестування та самостійна робота										Сума
Змістовий модуль 1								Змістовий модуль 2		
Т1	Т2	Т3	Т4	Т5	Т6	Т7	Т8	Т9	Т10	100
10	10	10	10	10	10	10	10	10	10	

Т1, Т2 ... Т20 – теми змістових модулів.

Шкала оцінювання

Сума балів за всі види навчальної діяльності	Оцінка за національною шкалою
	для заліку
90-100	зараховано
82-89	
74-81	

64-73	
60-63	
35-59	не зараховано з можливістю повторного складання
0-34	не зараховано з обов'язковим повторним вивченням дисципліни

10. Рекомендована література

Базова

1. Марченко О. О., Россада Т.В. Актуальні проблеми Data Mining: навчальний посібник для студентів факультету комп'ютерних наук та кібернетики. — Київ. — 2017. — 150 с.
1. Барсегян А. А. Анализ данных и процессов: учеб. пособие / А.А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. – 3-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2009. – 512 с.
2. Паклин Н.Б. Бизнес-аналитика: от данных к знаниям / Н.Б. Паклин, В.И. Орешков. – СПб.: Питер, 2009. – 624 с.
3. Сегаран Т. Программируем коллективный разум / Т. Сегаран. – СПб.: Символ-Плюс, 2008. – 368 с.
4. Чубукова И.А. Data Mining: учебное пособие / И.А. Чубукова. – М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. – 382 с.
5. Дубровин В.И. Интеллектуальные средства диагностики и прогнозирования надежности авиадвигателей: Монография / В.И. Дубровин, С.А. Субботин, А.В. Богуслаев, В.К. Яценко. – Запорожье: ОАО "Мотор-Сич", 2003. – 279 с.
6. Leskovec J. Mining of Massive Datasets / Jure Leskovec Anand Rajaraman, Jeffrey David Ullman // Stanford Univ. – 2010.

Допоміжна

1. Han J. Data Mining: Concepts and Techniques (Second Edition) / J. Han, M. Kamber – Morgan Kaufmann Publishers, 2006. – 800 p.
2. Witten, I. H. Data mining : practical machine learning tools and techniques. / Ian H. Witten, Frank Eibe, Mark A. Hall. – 3rd ed. – Morgan Kaufmann Publishers, 2011. – 630 p.

3. Макленнен Д. Microsoft SQL Server 2008: Data Mining – интеллектуальный анализ данных / Д. Макленнен, Ч. Танг, Б. Криват. – СПб.: БХВ-Петербург, 2009. – 720 с.
4. Дюк В. Data Mining : учебный курс / В. Дюк, А. Самойленко. – СПб.: Питер, 2001. – 368 с. 15.
5. Барсегян и др. Методы и модели анализа данных: OLAP и Data Mining. – СПб., 2004
6. Berry, Michael J. A. “Data mining techniques: for marketing, sales, and customer relationship management “/ Michael J.A. Berry, Gordon Linoff. – 2nd ed.

12. Інформаційні ресурси

1. Національна бібліотека ім. В.І. Вернадського / [Електронний ресурс]. – Режим доступу: http://www.nbuv.gov.ua/
2. Рівненська обласна універсальна наукова бібліотека (м. Рівне, майдан Короленка, 6) / [Електронний ресурс]. – Режим доступу : http://www.libr.rv.ua/
3. Рівненська централізована бібліотечна система (м. Рівне, вул. Київська, 44) / [Електронний ресурс]. – Режим доступу: http://cbs.rv.ua/
4. Цифровий репозиторій НУВГП / [Електронний ресурс]. – Режим доступу: http:// http://ep3.nuwm.edu.ua/
5. Наукова бібліотека НУВГП (м. Рівне, вул. Олекси Новака, 75) / [Електронний ресурс]. – Режим доступу: http://nuwm.edu.ua/naukova-biblioteka
6. Weka 3: Data Mining Software in Java [Електронний ресурс] – Режим доступу: http://www.cs.waikato.ac.nz/ml/weka/
7. Weka 3 Wiki documentation [Електронний ресурс] – Режим доступу: http://weka.wikispaces.com/
8. Курс лекций Николая Анохина / [Електронний ресурс]. – Режим доступу: https://www.youtube.com/playlist? list=PLrCZzMib1e9pyyrqknouMZbIPf4l3CwUP
9. Data is the New Oil By Michael Palmer / [Електронний ресурс]. – Режим доступу: http://ana.blogs.com/maestros/2006/11/data_is_the_new.html
10. Анализ данных как область знания / [Електронний ресурс]. – Режим доступу: http://postnauka.ru/video/34960 4. Материалы на тему анализа данных http://www.basegroup.ru/library/methodology/data_mining/
11. Наивный Байесовский классификатор в 25 строк кода / [Електронний ресурс]. – Режим доступу: http://habrahabr.ru/post/120194/
12. Фильтрация смс спама с помощью наивного байесовского классификатора/ [Електронний ресурс]. – Режим доступу: http://habrahabr.ru/post/184574/
13. Лекции курса «Машинное обучение» от yandex / [Електронний ресурс]. – Режим доступу:https://yadi.sk/d/V9p7E6uAFjHcD
14. Воронцов К. В. Лекции по алгоритмам кластеризации и многомерного шкалирования / [Електронний ресурс]. – Режим доступу: http://www.ccas.ru/voron/download/Clustering.pdf
15. Котов А., Красильников Н. Кластеризация данных. 2006 / [Електронний ресурс]. – Режим доступу:http://logic.pdmi.ras.ru/~yura/internet/02ia-seminar-note.pdf

16. Информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных / [Електронний ресурс]. – Режим доступу:www.machinelearning.ru/
17. Н.Ю. Золотых Как обучаются машины? научно-популярная лекція http://www.uic.unn.ru/~zny/ml/Pop/ml_pop.pdf
18. .Главы из книги на тему машинного обучения и презентации уроков Сергея Николенко/ [Електронний ресурс]. – Режим доступу: http://logic.pdmi.ras.ru/~sergey/teaching/ml/ Бази данних http://vincentarelbundock.github.io/Rdatasets/datasets.html