УДК 303.722.3:631.1.517.11

Ю.В. Кернасюк, канд. екон. наук
Кіровоградська державна сільськогосподарська дослідна станція НААН України

Інтелектуальний аналіз даних для навчання і практичного застосування в економіці

В статті розкрито основні аспекти та перспективи розвитку нового напрямку інформаційно-аналітичного забезпечення комплексного оцінювання результатів економічної діяльності сільськогосподарських підприємств з використанням засобів інтелектуального аналізу даних. На прикладі сільськогосподарських підприємств Кіровоградської області доведено перспективність застосування кластерного аналізу для оцінки економічної ефективності результатів їх діяльності. Встановлено раціональний напрям спеціалізації виробництва, що забезпечує максимальний прибуток.
інтелектуальний аналіз даних, ефективність, кластер, сільськогосподарські підприємства

Ю.В. Кернасюк, канд. экон. наук
Кировоградская государственная сельскохозяйственная опытная станция НААН Украины
Интеллектуальный анализ данных для обучения и практического применения в экономике

В статье раскрыты основные аспекты и перспективы развития нового направления информационно-аналитического обеспечения комплексной оценки результатов экономической деятельности сельскохозяйственных предприятий с использованием средств интеллектуального анализа данных. На примере сельскохозяйственных предприятий Кировоградской области доказана перспективность применения кластерного анализа для оценки экономической эффективности результатов их деятельности. Установлено рациональное направление специализации производства, обеспечивающее максимальную прибыль.
интеллектуальный анализ данных, эффективность, кластер, сельскохозяйственные предприятия

Постановка проблеми та її актуальність. Ринкове реформування економіки, постійне загострення конкуренції товаровиробників та необхідність запобігання загроз фінансових ризиків вимагають від фахівців сучасних професійних якостей, підвищують відповідальність керівників за результативність і наслідки рішень. В цих умовах стратегічної складової бізнесу стає стійка тенденція розвитку інформатизації процесів управління, а ефективність діяльності організації визначається застосуванням інноваційних методів і програмно-технічних засобів їх підтримки [3, с. 3].

Зростання інформаційних потоків в різних видах діяльності зумовлює потребу застосування новітніх інструментів аналізу даних, що дозволяють ефективно використовувати обмежений час для їх опрацювання та швидко надавати аналітичні звіти у зручній для користувача формі.

Вирішення цих проблем безпосередньо залежить від широкого сприйняття в навчальному процесі та практичній діяльності фахівців економічної сфери інноваційних напрямів обробки інформації, створення ефективної системи її аналізу та інтеграції з уже існуючими способами та засобами.

Серед останніх досягнень у галузі новітніх інформаційних технологій, які знайшли широке практичне застосування в наукових дослідженнях та бізнесі, окремо виділяють інтелектуальний аналіз даних, що найбільш поширений в англомовній термінології під назвою «Data Mining».

Аналіз останніх досліджень і публікацій. Особливості становлення і розвитку інтелектуального аналізу даних у різних сферах діяльності вивчали А.І. Петренко, А.А. Барсегян, В.А. Дюк, А.П. Самойленко, І.О. Чубукова, В.Ф. Ситник, М.Т. Краснюк, К.В. Ілляшенко та ін.

Дослідженням практичних аспектів використання інформаційних технологій Data Mining займалися В.П. Боровіков, П.С. Большаков, А.А. Халафян та ін.

Разом з тим, у більшості з представлених публікацій і окремих наукових роботах ще недостатньо уваги приділяється питанням використання інтелектуального аналізу даних для навчання і практичної діяльності в сфері оцінювання підприємств агробізнесу на предмет їх ефективності, що пов'язано як з порівняно незначним проміжком часу від його виникнення та поширення, так і певною складністю застосування та інтерпретації одержаних результатів.

Усі вище перелічені моменти зумовили актуальність продовження досліджень у даному напрямку та необхідність подальших наукових розробок.

Постановка завдання. Метою даної публікації є оцінка сучасного стану розвитку інтелектуального аналізу даних і огляд можливостей його використання в навчальному процесі та практичній діяльності на прикладі аграрної сфери економіки.

Виклад основного матеріалу. Розвиток технології інтелектуального аналізу даних (термін походить від англійського «Data Mining» і дослівно перекладається, як розробка, виявлення прихованих закономірностей або взаємозв'язків між змінними у великих масивах необроблених даних) зумовлений рядом об'єктивних факторів. Основні з них, на які слід звернути увагу: поява великої кількості даних в різних сферах людської діяльності та прискорення накопичення інформації, що набагато перевищує швидкість її обробки та можливість адекватного сприйняття.

Інтелектуальний аналіз даних (скорочено ІАД) – це сукупність сучасних методів добування знань. Вибір методу проведення аналізу, як правило, залежить від типу наявних даних і від того, яку інформацію потрібно знайти чи дістати зі сховища їх зберігання.

До найпоширеніших методів ІАД можна віднести такі [8, с. 67]:

– об'єднання (association; іноді вживають термін affinity, що означає подібність, структурну близькість) – виокремлення структур, що повторюються в часовій послідовності. Цей метод визначає правила, за якими можна встановити, що один набір елементів корелює з іншим. Користуючись ним, аналізують ринковий кошик пакетів продуктів, розробляють каталоги, здійснюють перехресний маркетинг тощо;

– аналіз часових рядів (sequence-based analysis, або sequential association) дає змогу відшукувати часові закономірності між даними (трансакціями). Наприклад, можна відповісти на запитання: купівля яких товарів передує купівлі даного виду продукції ? Метод застосовується, коли йдеться про аналіз цільових ринків, керування гнучкістю цін або циклом роботи із замовником (Customer Lifecycle Management);

– кластеризація (clustering) – групування записів, що мають однакові характеристики, наприклад за близькістю значень полів у базі даних. Використовується для сегментування ринку та замовників. Можуть залучатися статистичні методи або нейромережі. Кластеризація часто розглядається як перший необхідний крок для подальшого аналізу даних;

– класифікація (classification) – віднесення запису до одного із заздалегідь визначених класів, наприклад під час оцінювання ризиків, пов'язаних із видачею кредиту;

– оцінювання (estimation);

– нечітка логіка (fuzzy logic);

– статистичні методи, що дають змогу знаходити криву, найближче розміщену до набору точок даних;

– генетичні алгоритми (genetic algorithms);

– фрактальні перетворення (fractal-based transforms);

– нейронні мережі (neural networks) – дані пропускаються через шари вузлів, «навчених» розпізнавати ті чи інші структури – використовуються для аналізу переваг і цільових ринків, а також для приваблювання замовників.

На думку А.І. Петренка основна особливість «Data Mining» – це поєднання широкого математичного інструментарію (від класичного статистичного аналізу до нових кібернетичних методів) і останніх досягнень у сфері інформаційних технологій. У технології «Data Mining» гармонійно об'єдналися строго формалізовані методи і методи неформального аналізу, тобто кількісний і якісний аналізи даних [6].

Як вважають В.А. Дюк і А.П. Самойленко, інтелектуальний аналіз даних є міждисциплінарною областю, що виникла й розвивається на базі досягнень прикладної статистики, розпізнавання образів, методів штучного інтелекту, теорії баз даних та ін.[2, с. 20].

На рис. 1. наведено схему поєднання найпоширеніших методів проведення ІАД.

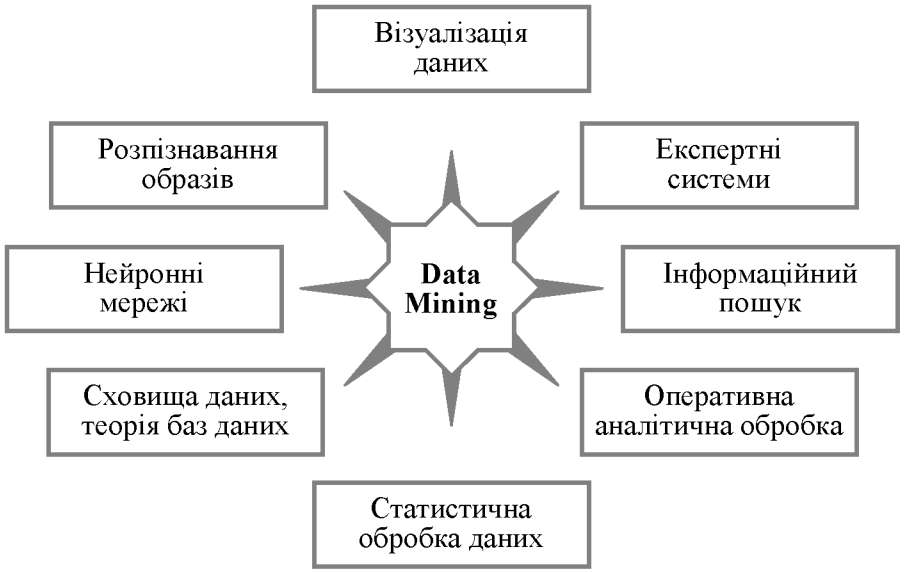

Рисунок 1 – «Data Mining» як міждисциплінарна область для обробки значних масивів інформації та пошуку нових знань

Джерело: за даними [8, с. 69]

Для ознайомлення з практичними можливостями ІАД, доцільно привести приклад його застосування при опрацюванні масиву статистичної інформації основних економічних показників діяльності сільськогосподарських підприємств 21 району Кіровоградської області в середньому за 2009-2013 рр.

Необхідно дослідити зведені статистичні дані в розрізі кожного із 21 районів за напрямами спеціалізації, які сформувалися в регіоні впродовж останніх років.

З метою групування районів за критерієм спеціалізації застосовано процедуру прямої класифікації з використанням кластерного аналізу. Суть її полягає у формуванні кластерів на основі утворення об'єктів, які мають однорідні показники спеціалізації. Аналіз багатовимірних даних проводився з використанням методу ЕМ – алгоритму

(термін походить від англійського «expectation-maximization» і дослівно перекладається, як очікування–максимізація) за допомогою спеціальної функції «Generalized cluster analysis» в середовищі модулю «Data Mining» випробувальної версії пакету прикладних програм STATISTICA.

Алгоритм «expectation-maximization» заснований на методиці ітеративного обчислення оцінок максимальної правдоподібності, яку запропонували в 1977 р. A. P. Demster, N. M. Laird, D. B. Rubin [4].

«Означений прийом є розширенням методів, доступних в модулі кластерного аналізу STATISTICA. Даний модуль спеціально розроблений для обробки великих масивів даних. Він дозволяє розбивати на кластери неперервні або категоріальні змінні, реалізує функціональність повного навчання без учителя для розпізнавання образів (кластеризації), включає можливість впровадження моделей прогнозуючої кластеризації. У модулі реалізована розширена кластеризація методом EM. Її іноді називають кластеризацією на основі ймовірності або статистичної кластеризації. За допомогою програми проводиться кластеризація спостережень на основі категоріальних і неперервних змінних, припускаючи різний розподіл аналізованих змінних (заданих користувачем). Модуль дозволяє будувати підсумкові результати і графіки (наприклад, графік розподілу для кластеризації EM), обчислювати описові статистики класифікації для кожного спостереження» [5].

За визначенням П.С. Большакова, «Generalized EM & k-Means Cluster Analysis» є узагальненим методом максимуму середнього і кластеризації методом k-means. Даний модуль – це розширення методів кластерного аналізу, він призначений для обробки великих масивів даних і дозволяє кластеризувати як неперервні, так і категоріальні змінні; забезпечує всі необхідні функціональні можливості для розпізнавання образів [1].

На практиці перспективність його застосування підтверджено шляхом проведеного кластерного аналізу сільськогосподарських підприємств Кіровоградської області. Первинна вихідна інформація наведена в табл. 1.

Таблиця 1 – Основні економічні показники діяльності сільськогосподарських підприємств Кіровоградської області, середнє за 2009-2013 рр.

Назва районів	Рівень рентабельності діяльності господарств, %	Прибуток на 1 га, грн	У середньому на одне підприємство		Питома вага у структурі усієї товарної продукції і послуг, %				
			розмір сільськогосподарських угідь, га	чисельність працівників, осіб	зернові культури	олійні культури	продукція тваринництва	інші продукція	послуги в сільському господарстві
1	2	3	4	5	6	7	8	9	10
Область	**30,5**	**635**	**2001**	**41**	**44,0**	**42,7**	**7,1**	**3,3**	**2,9**
Олександрійський	12,4	903	3153	93	53,7	29,4	7,2	1,1	8,6
Олександрівський	23,2	859	1789	36	57,6	34,8	0,6	4,4	2,6
Бобринецький	29,2	905	1902	31	32,9	64,4	1,9	0,1	0,6
Гайворонський	27,3	665	1249	32	50,3	37,6	9,1	1,1	1,8
Голованівський	16,3	1465	2905	65	51,1	47,0	0,4	0,2	1,2
Добровеличківський	40,9	768	2236	42	44,2	51,8	3,2	0,2	0,6
Долинський	31,5	1609	1465	21	34,6	61,3	2,1	0,6	1,3
Знам'янський	33,6	2043	2748	56	53,8	33,8	4,0	5,6	2,8

Продовження таблицы 1

1	2	3	4	5	6	7	8	9	10
Кіровоградський	40,8	1075	1825	46	31,0	34,1	26,3	4,8	3,8
Компаніївський	36,2	1116	1539	31	39,4	53,1	5,0	0,4	2,0
Маловисківський	37,7	1416	1817	32	46,9	49,6	1,6	0,6	1,3
Новгородківський	39,0	874	1765	34	34,2	41,8	9,6	12,8	1,7
Новоархангельський	28,5	1701	1857	38	43,4	51,4	2,7	0,6	1,9
Новомиргородський	48,4	965	1725	42	32,6	51,1	4,1	10,5	1,8
Новоукраїнський	25,5	1017	2609	51	48,3	41,0	5,4	3,2	2,1
Вільшанський	37,6	566	1630	39	55,0	37,0	3,7	0,7	3,6
Онуфріївський	24,2	780	1440	21	45,6	47,9	4,4	0,3	1,8
Петрівський	22,2	1519	2717	54	37,4	51,1	6,7	0,4	4,4
Світловодський	32,5	830	1523	22	68,0	20,2	10,0	0,7	1,1
Ульяновський	26,7	1298	2082	40	41,9	33,4	17,1	6,5	1,2
Устинівський	44,1	635	2482	44	31,4	61,6	2,9	0,2	4,0

Джерело: розроблено автором за даними Головного управління статистики в Кіровоградській області (форма 50 – с.г.).

У якості категоріальної змінної було використано показник середнього рівня рентабельності діяльності господарств в районі, % (РР), а неперервними є: середній розмір сільськогосподарських угідь на одне господарство, га (ПСГ); середньооблікова чисельність працівників на одному підприємстві, осіб (СЧП); частка зернових культур в структурі усієї товарної продукції і послуг (ЗК), %; частка олійних культур в структурі усієї товарної продукції і послуг, % (ОК); частка продукції тваринництва в структурі усієї товарної продукції і послуг, % (ТП); частка іншої продукції в структурі усієї товарної продукції і послуг, % (ІП); частка послуг сільського господарства в структурі усієї товарної продукції і послуг, % (П); прибуток з розрахунку на 1 га, грн. (ПР).

Після створення робочої таблиці і експорту її в середовище STATISTICA за допомогою модулю «Generalized cluster analysis» було відібрано категоріальну і неперервні змінні. При цьому зроблено припущення, що райони доцільно згрупувати за ознакою ефективності діяльності сільськогосподарських підприємств:

- вище від середньої ефективності;
- нижче від середньої ефективності;
- середня ефективність.

З огляду на означене було вибрано 3 кластери. Після проведення розрахунку одержано звіт у табличній та графічній формі.

За результати проведеного кластерного аналізу райони області розподілилися наступним чином:

1 кластер (6 районів) – Олександрійський, Голованівський, Добровеличківський, Новоукраїнський, Петрівський і Устинівський;

2 кластер (10 районів) – Олександрівський, Бобринецький, Гайворонський, Долинський, Компаніївський, Маловисківський, Новоархангельський, Вільшанський, Онуфріївський і Світловодський;

3 кластер (5 районів) – Знам'янський, Кіровоградський, Новгородківський, Новомиргородський і Ульяновський.

На рис. 2 для порівняльного аналізу і оцінки взаємозв'язку зображено зміну центрів тяжіння з назвами об'єктів, які віднесені до визначених кластерів в середньому за 5 років.

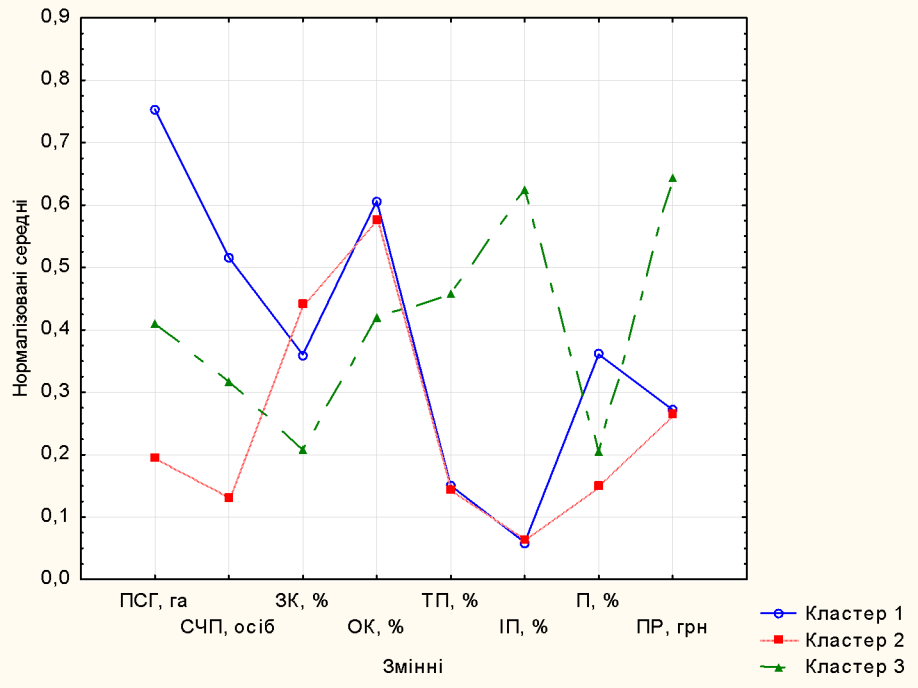

Рисунок 2 – Графік середніх для неперервних змінних кожного кластера

Джерело: розроблено автором.

Так, для 1 і 2 кластерів досліджуваних районів простежується закономірність в частині схожості вибору напрямку спеціалізації господарств і відмінність в розмірах землекористування, чисельності працівників та показниках ефективності їх економічної діяльності. Водночас сільськогосподарські підприємства тих районів, які увійшли до 3 кластеру, нічого спільного не мали в порівняні з групою у 1 і 2 кластерах.

Результати дисперсійного аналізу значимості впливу окремих чинників на рентабельність діяльності сільськогосподарських підприємств наведено в табл. 2.

Таблиця 2 – Дисперсійний аналіз результатів досліджень

ANOVA for continuous variables (БД спеціалізація райони) Number of clusters: 3 Total number of training cases: 21						
	Between - SS	df	Within - SS	df	F	p value
ПСГ, га	**4231174**	**2**	**1653406**	**18**	**23,03158**	**0,000011**
СЧП, осіб	**2901**	**2**	**2445**	**18**	**10,67893**	**0,000875**
ЗК, %	251	2	1770	18	1,27674	0,303029
ОК, %	214	2	2448	18	0,78722	0,470162
ТП, %	**247**	**2**	**488**	**18**	**4,56156**	**0,024969**
ІП, %	**191**	**2**	**68**	**18**	**25,12822**	**0,000006**
П, %	11	2	54	18	1,82242	0,190224
ПР, грн	**1180328**	**2**	**1963945**	**18**	**5,40899**	**0,014470**

Джерело: складено автором.

Жирним шрифтом виділено змінні (у робочих звітах програми STATISTICA вони відображаються червоним кольором – прим. авт.) з високою статистичною значущістю $p<0,05$, тобто які підтверджують вплив розміру землекористування, чисельності працівників, частки продукції тваринництва та інших видів в її загальній структурі, а також суми одержаного прибутку з 1 га на рівень рентабельності діяльності сільськогосподарських підприємств.

«Завдання кластеризації полягає в поділі досліджуваної кількості об'єктів на групи "подібних" об'єктів, які називають кластерами» [7, с. 159-160].

Узагальнення результатів кластерного аналізу для порівняльного співставлення між існуючими напрямами спеціалізації сільськогосподарських підприємств регіону, показниками економічної ефективності та розмірами за площею сільськогосподарських угідь і середньообліковою чисельністю працівників наведено у табл. 3.

Таблиця 3 – Порівняльна оцінка результатів кластерного аналізу

Показники	Кластери		
	1	2	3
Кількість районів в кластері	6	10	5
У середньому в кластері			
- прибуток на 1 га, грн	968,0	960,0	1519,6
- площа сільськогосподарських угідь на одне господарство, га	2684	1621	2029
- чисельність працівників на підприємстві, осіб	58	30	44
- частка зернових культур в структурі виручки, %	44,3	47,4	38,7
- частка олійних культур в структурі виручки, %	47,0	45,6	38,8
- частка продукції тваринництва в структурі виручки, %	4,3	4,1	12,2
- частка іншої продукції в структурі виручки, %	0,9	1,0	8,0
- частка послуг сільського господарства в структурі виручки, %	3,5	1,8	2,3

Джерело: розраховано автором

З урахуванням визначених координат кластерних центрів було проведено їхнє профілювання, тобто подана комплексна оцінка кожному з них за їх особливостями.

Встановлено, що в Кіровоградській області впродовж досліджуваного періоду найвища прибутковість аграрного виробництва спостерігалася в групі районів 3 кластеру, де сільськогосподарські підприємства мали зерно-олійний напрям спеціалізації з розвинутим виробництвом тваринницької продукції. Це, зокрема, такі райони: Знам'янський, Кіровоградський, Новгородківський, Новомиргородський і Ульяновський. Водночас, найнижчий рівень прибутку з 1 га одержували господарства 2 та 3 кластерів, які знаходилися, відповідно на території Олександрівського, Бобринецького, Гайворонського, Долинського, Компаніївського, Маловисківського, Новоархангельського, Вільшанського, Онуфріївського, Світловодського районів і Олександрійського, Голованівського, Добровеличківського, Новоукраїнського, Петрівського та Устинівського районів. У цих районах сформувався зерно-олійний і олійно-зерновий напрям спеціалізації, галузь тваринництва займала лише 4,3 та 4,1 % від усієї товарної продукції сільського господарства.

Порівняльний аналіз за розміром землекористування засвідчив, що в 3 кластері він складав 2029 га проти 2684 і 1621 га, відповідно у 1 та 2 кластерах.

Таким чином, на основі даних системного аналізу економічно обґрунтовано, що для регіону раціональним є саме зерно-олійний напрям спеціалізації з розвинутим виробництвом тваринницької продукції. У структурі всієї товарної продукції і послуг сільського господарства питома вага зернових і олійних культур повинна складати, відповідно порівну (39 і 39 %), а частка тваринницької продукції не менше 12 %.

Висновки та перспективи подальших досліджень. Отже, в умовах постійного росту інформаційних потоків і знань застосування новітніх інструментів аналізу даних дозволяє ефективно використовувати обмежений час для їх опрацювання за рахунок потужних інструментів інтелектуального аналізу даних. У аграрній сфері означений напрям аналізу ще не знайшов широкого поширення, в той час як у багатьох інших галузях бізнесу, освіти і медицини його використання дозволило значно підвищити результативність роботи. На прикладі проведеного системного аналізу економічної

ефективності діяльності сільськогосподарських підприємств 21 району Кіровоградської області встановлено раціональний напрям і співвідношення основних галузей аграрного виробництва, що забезпечує максимальний прибуток з 1 га використання сільськогосподарських угідь. Для товаровиробників регіону рекомендуються раціональні параметри спеціалізації: у структурі всієї товарної продукції питома вага зернових і олійних культур повинна складати, відповідно рівну частку близько 39 і 39 %, а тваринницька – не менше 12 %. Перспективи подальших досліджень пов'язані, безпосередньо, з необхідністю поглиблення вивчення інших методів інтелектуального аналізу та практичного їх застосування в аграрній сфері, а також поширення знань для навчання фахівців економічних спеціальностей.

Список літератури

1. Большаков П.С. Возможности Statistica Data Miner [Текст] / П.С. Большаков // Exponenta Pro. Математика в приложениях. – 2003. – № 1 (1). – С. 13–16.
2. Дюк В.А. Data Mining: учебный курс (+CD) [Текст] / В.А. Дюк, А.П. Самойленко – СПб.: Питер, 2001. – 368 с.
3. Информационные системы в экономике [Текст]: [учебник для студентов вузов / под ред.: Г. Титоренко]. – М. : ЮНИТИ-ДАНА, 2008. – 463 с.
4. Maximum Likelihood from Incomplete Data via the EM Algorithm [Електронний ресурс] / A.P. Dempster, N.M. Laird and D.B. Rubin // Journal of the Royal Statistical Society. Series B (Methodological) – 1977. – Vol. 39, No. 1. – pp. 1–38. – Режим доступу до журн. : http://links.jstor.org/sici?sici=0035-9246%281977%2939%3A1%3C1%3AMLFIDV%3E2.0.CO%3B2-Z.
5. Модули Data Mining [Електронний ресурс]. – Режим доступу : http://www.statsoft.ru/products/STATISTICA_Data_Miner/modules-statistica-data-miner.php#Обобщенный_EM.html
6. Петренко А.І. Grid та інтелектуальна обробка даних Data Mining [Текст] / А.І. Петренко // Системні дослідження та інформаційні технології. – 2008. – № 3. – С. 97–110.
7. Технологии анализа данных. Data Mining, Visual Mining, Text Mining, OLAP [Текст] / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко та ін. – СПб. : БХВ- Петербург, 2007. – 384 с.
8. Шарапов О.Д. Економічна кібернетика: Навч. посібник. [Текст] / Шарапов О.Д., Дербенцев В.Д., Семьонов Д.Є – К.: КНЕУ, 2004. – 231 с.

Yuriy Kernasyuk
Kirovohrad State Agricultural Experimental Station of the National Academy of Agrarian Sciences of Ukraine, Kirovohrad, Ukraine

Data Mining for Learning and Practical Application in the Economy

The market economic reforms, the increasing of producer's competition and the need to prevent threats of financial risk require advanced professional features from professionals; increase the responsibility of managers for performance and consequences of decisions. In these conditions, the strategic component of business development is a steady trend of information management processes and efficiency of the organization is determined by using innovative methods and software and hardware for supporting. The purpose of the article is to assess the current state of data mining and an overview of its use in teaching and practice on the example of the agrarian economy.

The growth of information flow requires the use of advanced data analysis tools. Data mining – the collection of modern methods of information processing. The ways of using data mining to study the economic efficiency of agricultural production were grounded. The information source is the average economic performance of agricultural enterprises for 2009-2013 years. Data mining had been performed by using EM – algorithm and direct classification procedures with using cluster analysis. The three clusters of areas Kirovohrad region were determined.

The systematic analysis of the economic efficiency of farms in 21 districts of Kirovohrad region was conducted and it was established a rational direction and value of basic industries of agriculture, which provides maximum profit from 1 hectare of agricultural land. The analysis showed (for 21 districts) the rational direction of specialization of agricultural production, ensuring maximum profit from 1 hectare of agricultural land: the proportion of grains and oilseeds - 39 and 39%, and livestock production - not less than 12%.

data mining, efficiency, the cluster, agricultural enterprises

Одержано 11.11.14