Статистичні методи моделювання з використанням технології data-mining

Виконав студент 4 курсу групи КА-34

Саркісов Степан Юрійович

Керівник: д.т.н., професор кафедри ММСА

Бідюк Петро Іванович

Актуальність дослідження

- Людський розум сам по собі не пристосований для сприйняття величезних масивів різнорідної інформації. В середньому людина, за винятком деяких індивідуумів, не здатна уловлювати більше двох-трьох взаємозв'язків навіть у невеликих вибірках. Тому методи математичної статистики виявляються корисними головним чином для перевірки заздалегідь сформульованих гіпотез, тоді як визначення гіпотези іноді буває досить складною і трудомісткою задачею.
- Сучасні технології Data Mining переробляють інформацію з метою автоматичного пошуку шаблонів (патернів), характерних для будь-яких фрагментів неоднорідних багатовимірних даних.

Мета, предмет і об’єкт дослідження

Мета: Побудова математичних моделей вибраних процесів на основі масивів даних та розробка статистичних критеріїв для перевірки адекватності побудованої моделі використовуючи технологію data mining.

Об’єкт: Статистичні дані стосовно розвитку досліджуваних процесів.

Предмет: Методи статистичного аналізу масивів даних з метою побудови адекватних моделей досліджуваних процесів.

Постановка задачі

1. Класифікація і аналіз вибраних типів моделей для опису даних.
2. Визначення критеріїв для перевірки побудованих моделей.
3. Формулювання власних критеріїв та перевірка їх по відношенню до існуючих.
4. Аналіз критеріїв, які більш точно характеризують побудовані моделі.

Деякі системи для статистичної обробки даних

- **EViews**

+ Компактність

+ Доступність

- **SAS**

+ Потужність

- **SPSS**

+ Гнучкість

+ Допоміжні матеріали

- **STATISTICA**

+Візуалізація

+Продуктивність

Недоліки: Висока вартість та складність адаптації

Критерії адекватності моделей

Коефіцієнт множинної детермінації R^2

$$R^2 = \frac{\mathrm{var}(\hat{y})}{\mathrm{var}(y)} = 1 - \frac{SSE}{SST}$$

Сума квадратів похибок для вибраної моделі повинна бути мінімальною, тобто,

$$\sum_{k=1}^{N} e^2(k) = \sum_{k=1}^{N} [\hat{y}(k) - y(k)]^2 \to \min_{\hat{\theta}}$$

Інформаційний критерій Акайке: $AIC = N \cdot ln\left(\sum_{k=1}^{N} e^2(k)\right) + 2n$

Статистики Дарбіна-Уотсона (DW): $DW = 2 - 2\rho$,

де ρ: $\rho = E[e(k)e(k-1)]/\sigma_e^2$

Критерії якості оцінок прогнозів

- Середньоквадратична похибка:

$$\text{СКП} = \frac{1}{N}\sum_{k=1}^{N}[y(k) - \hat{y}(k)]^2 = \frac{1}{N}\sum_{k=1}^{N} e^2(k)$$

- Середня похибка:

$$\text{СП} = \frac{1}{N}\sum_{k=1}^{N}[y(k) - \hat{y}(k)] = \frac{1}{N}\sum_{k=1}^{N} e(k)$$

- Середнє абсолютне значення похибки у процентах:

$$\text{САПП} = \frac{1}{N}\sum_{k=1}^{N}\frac{[y(k) - \hat{y}(k)]}{y(k)} \times 100\%$$

Проаналізовані моделі дослідження

У ході виконання роботи розглянуті **4 типи моделей**:

1. Моделі, побудовані на основі регресійного аналізу
2. Моделі у формі нейромереж
3. Моделі на основі генетичних алгоритмів

Для використання обрано моделі на основі регресійного аналізу у зв’язку з простотою їх реалізації і практичного використання, у тому числі для оцінювання прогнозів.

Моделі, використані в СППР

Авторегресія: рівняння авторегресії описує пам'ять процесу, тобто вплив значень попередніх станів на його поточний стан:

$$y(k) = a_0 + a_1 y(k-1) + \ldots + a_p y(k-p) = a_0 + \sum_{i=1}^{p} a_i y(k-i) + \varepsilon(k)$$

Авторегресія з ковзним середнім порядку (p,q) :

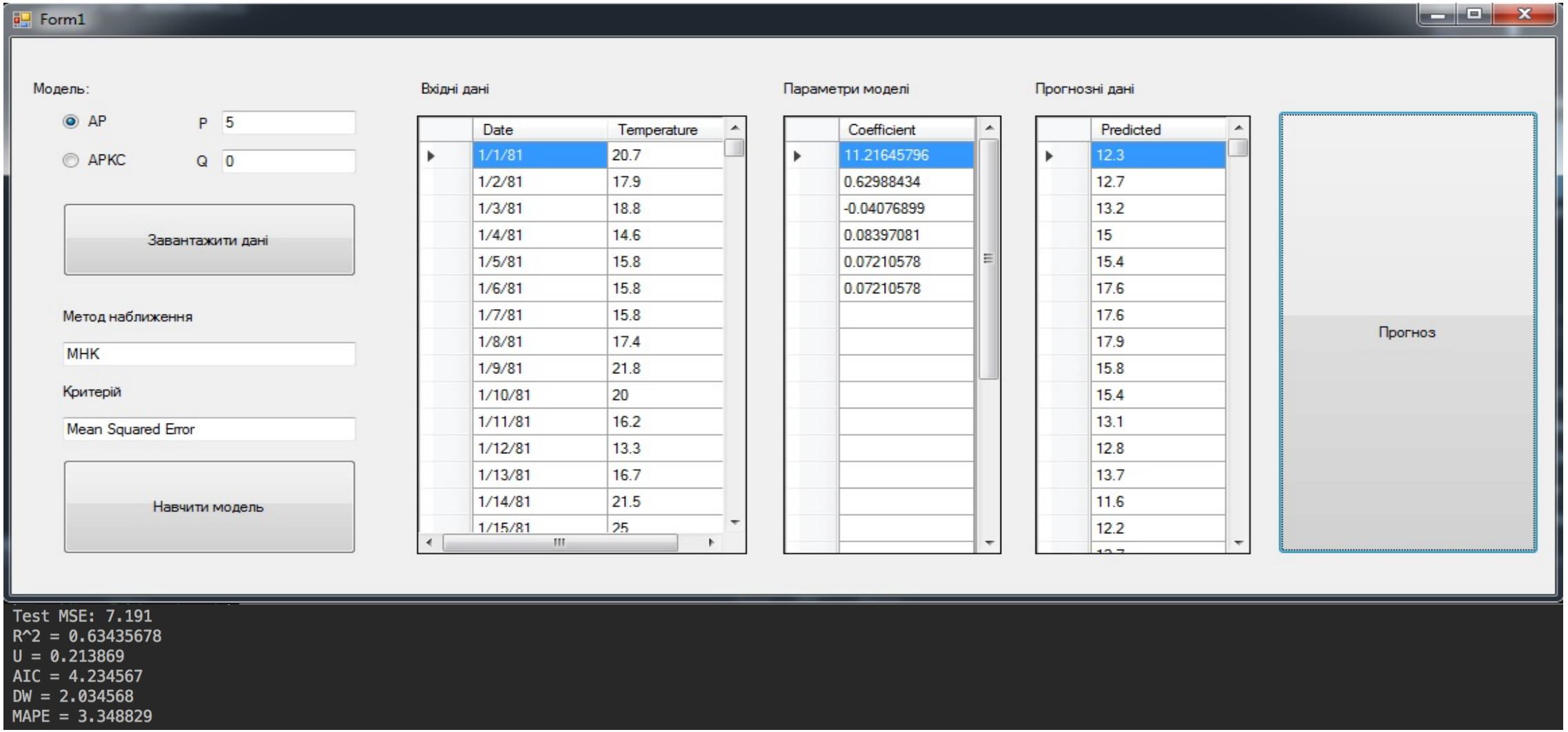
$$y(k) = a_0 + \sum_{i=1}^{p} a_i y(k-i) + \sum_{j=1}^{q} b_j \varepsilon(k-j) + \varepsilon(k)$$

Інтегральні критерії якості

$$I_1 = |1 - R^2| + U$$

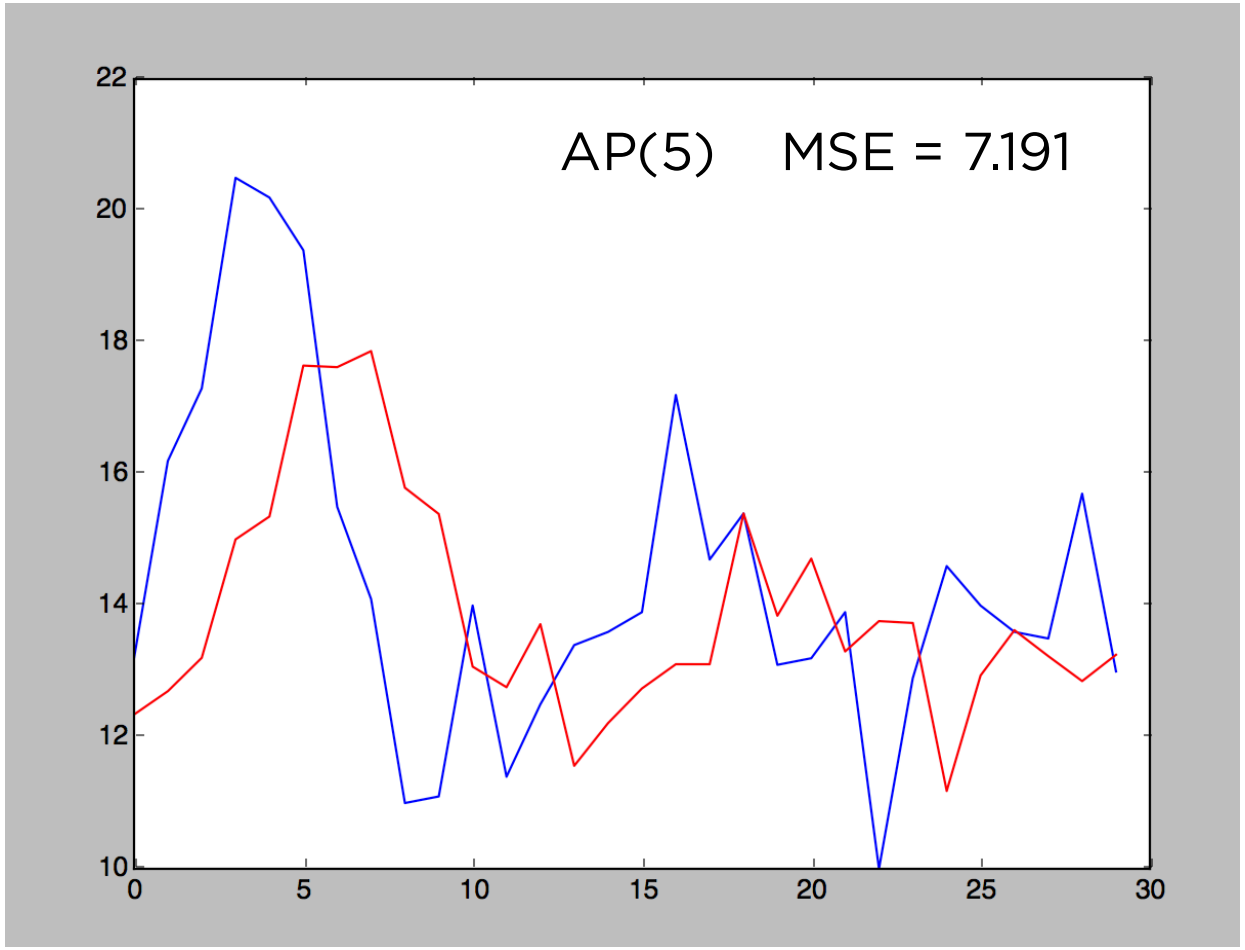
$$I_2 = |1 - R^2| + |2 - DW| + U.$$

Приклади роботи програми

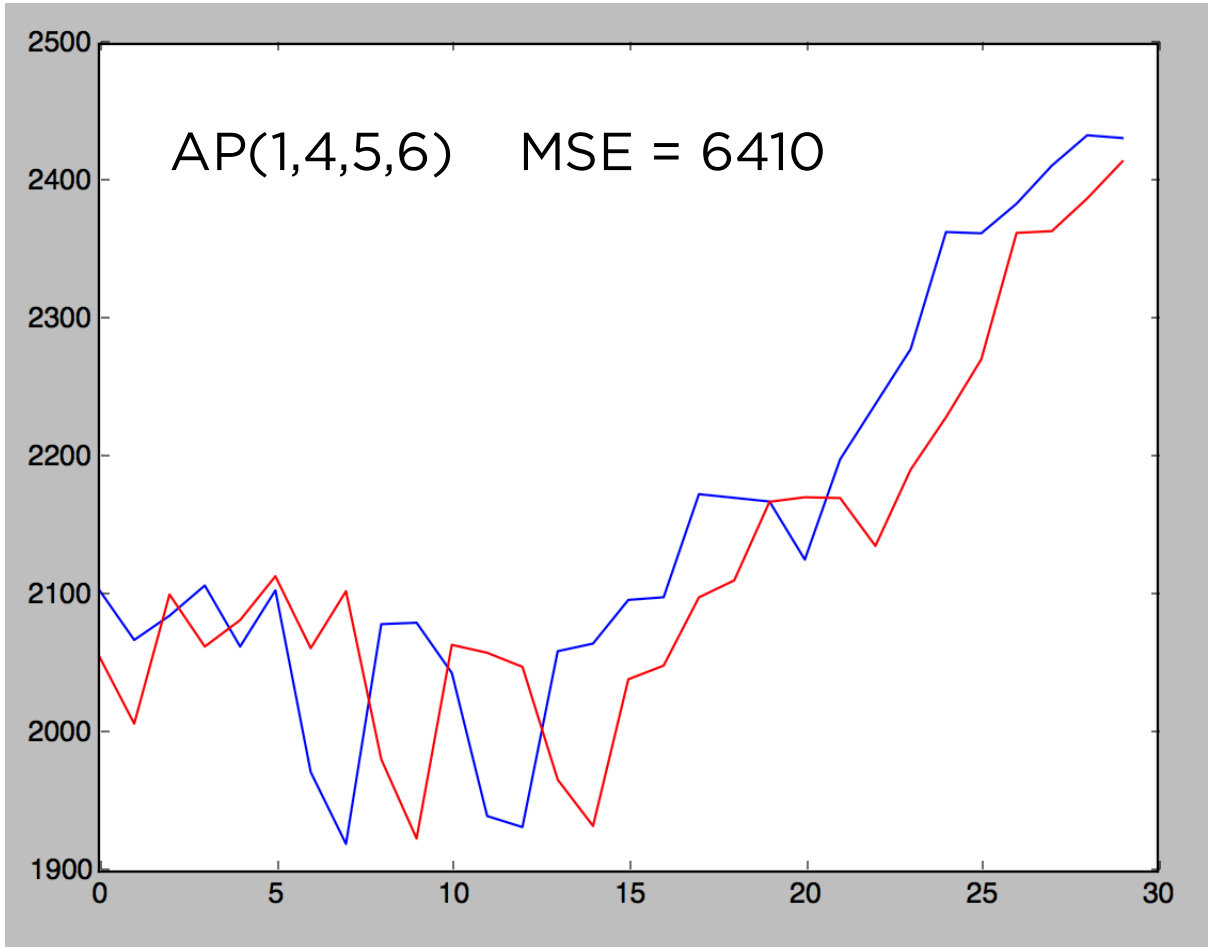


Прогнози

Мінімальна температура повітря в Мельбурні



S&P 500



Реальні Прогнозні

Аналіз результатів

Ряд	Тип моделі	Адекватність моделі			Характеристика прогноза			Интегральний критерій	
		MSE	R^2	DW	MAPE	U	AIC	I_1	I_2
Мінімальна температура повітря в Мельбурні	AP(1)	6.475	0.588	2.14	4.51	0.175	4.75	0.587	0.727
	AP(4)	7.15	0.62	2.04	3.63	0.175	4.68	0.555	0.595
	AP(5)	7.19	0.63	2.03	3.34	0.21	4.67	0.58	0.61
	AP(7)	6.99	0.63	2.01	3.21	0.17	4.64	0.54	0.55

Ряд	Тип моделі	Адекватність моделі			Характеристика прогноза			Интегральний критерій	
		MSE	R^2	DW	MAPE	U	AIC	I_1	I_2
S&P 500	AP(1)	7352	0.96	2.38	3.57	0.02	10.8	0.06	0.44
	AP(1,4,5,6)	6410	0.95	2.13	3.53	0.02	10.8	0.07	0.2
	AP(1,5,6)	6862	0.95	2.18	3.48	0.02	10.8	0.07	0.25
	AP(1)+trend	6806	0.98	2.19	3.58	0.02	10.8	0.04	0.23

Висновки

- Виконано огляд моделей, які можуть бути використані для формального опису лінійних та нелінійних процесів в економіці та інших сферах діяльності.

- Розроблена та програмно реалізована СППР для обробки статистичних даних, побудови моделей ФЕП та оцінювання прогнозів.

- Для вибраних ФЕП побудовані регресійні моделі, які характеризуються високим ступенем адекватності і забезпечують можливість обчислення високоякісних оцінок прогнозів.

Подальше дослідження

- Розширення номенклатури типів математичних моделей для опису даних.
- Використання методів інтелектуального аналізу даних.
- Застосування методів комбінування оцінок прогнозів, обчислених за допомогою альтернативних методів.
- Розробка та реалізація комерційної СППР.

Дякую за увагу