Структурна Ідентифікація Класифікаторів в Системах Кризового Моніторингу

Голуб С.В.
кафедра інтелектуальних систем прийняття рішень
Черкаський національний університет
імені Богдана Хмельницького
Черкаси, Україна
fpkpk@ukr.net

Авраменко А.С.
кафедра інтелектуальних систем прийняття рішень
Черкаський національний університет
імені Богдана Хмельницького
Черкаси, Україна
RedStar929@gmail.com

Structural Identification of Classificators in Crisis Monitoring Systems

Golub S.V.
Department of intelligent decision support systems
Bogdan Khmelnitsky Cherkassy National University
Cherkassy, Ukraine
fpkpk@ukr.net

Avramenko A.S.
Department of intelligent decision support systems
Bogdan Khmelnitsky Cherkassy National University
Cherkassy, Ukraine
RedStar929@gmail.com

***Анотація*—Запропоновано вдосконалити метод оптимізації структури інформаційної системи кризового моніторингу. Збільшення точності та якості класифікаторів досягається шляхом кластеризації масивів вхідних даних. Кожен кластер відноситься до свого алгоритму синтезу моделей. Експериментально підтверджено ефективність покращеного методу. Час перебудови структури зменшується на 60%. Похибка моделювання значимо не погіршується.**

***Abstract*—Modernization of a method for optimizing the structure of the crisis monitoring information systems is proposed. Improving of accuracy and quality in classification models achieved by clustering the input data arrays. Best algorithm of model synthesis is selected for each cluster in the input data arrays. Effectiveness of the improved method was experimentally confirmed. Time for restructuring of models reduced by 60%. Errors of modelling is not significantly worse.**

Ключові слова—кризовий моніторинг, багаторівневе моделювання, кластеризація, час перебудови системи, похибка моделювання.

Keywords—crisis monitoring, multilevel modeling,clasterisation, system restructuring time,modelling error

I. Вступ

Основним завданням технологій моніторингу є не лише збір інформації о характеристиках об’єкту моніторингу але і забезпечення інформацією процесу прийняття рішень. Інформація отримується в результаті моделювання властивостей об’єкта моніторингу на основі даних, отриманих в процесі вимірювання чисельних характеристик цього об’єкта.

Прийняття рішень в умовах надзвичайних ситуацій накладає ряд обмежень на технології, що забезпечують інформацією ці процеси. Так як розвиток надзвичайних ситуацій зазвичай є ланцюговим лавиноподібним динамічним процесом, що полягає в різкому погіршенні стану деякого об'єкта, як правило, представляє собою сукупність території і розташованих на ній об'єктів економіки і житлових комплексів, що призводить до катастрофічних для цього об'єкта і його оточення наслідків. То рішення в даних ситуаціях повинні отримуватися як найшвидше. Також незадовільна прогнозованість та динамічність надзвичайних ситуацій генерує велику кількість параметрів для моделювання, частина з яких може виникнути вперше, що означає велику можливість помилки в попередньо отриманих моделях

Головним завданням моделювання в моніторингових системах є забезпечення інформацією процесу прийняття рішень. Ця інформація здобувається за результатами моделювання властивостей об’єкта моніторингу на основі даних, отриманих в процесі вимірювання чисельних характеристик цього об’єкта.

Саме зараз перспективними системами моніторингу є системи засновані на технології багаторівневого перетворення даних, яка реалізована у вигляді інформаційної системи з ієрархічним поєднанням багатопараметричних моделей [1].

Такі моделі синтезуються за допомогою індуктивних алгоритмів, нейронних мереж, генетичних алгоритмів та інших. В даній технології сценарій вибору алгоритму синтезу багатопараметричних моделей (АСМ) реалізовано шляхом послідовного їх випробування та вибору кращого [1]. Далі з синтезованих моделей формується ієрархія(рис. 1).

Моделі на кожному рівні ієрархії розв'язують локальні задачі з перетворення даних. В таких ієрархічних структурах можливе поєднання великої кількості моделей, а саме, від п'ятдесяти і більше.

В процесі моніторингу кризових ситуацій властивості масиву вхідних даних постійно змінюються. Тому існує висока ймовірність того, що одна або декілька моделей можуть почати видавати не адекватні результати. Для виправлення таких «пошкоджень» проводиться заміна цих моделей з їх повторним синтезом та синтезом усіх моделей, які з ними пов'язані.

Процес синтезу усіх «пошкоджених» моделей та моделей, які з ними пов'язані, займає досить тривалий час. Тривалість процесу синтезу залежить від кількості моделей в структурі.

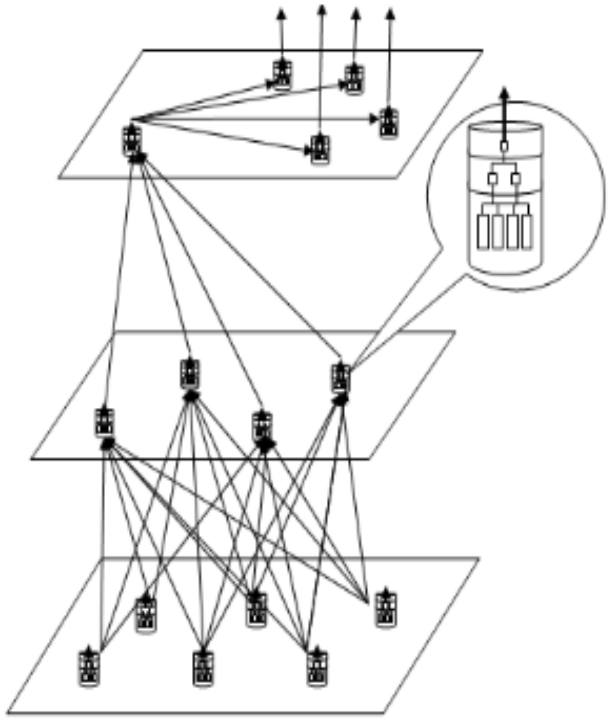

Рис. 1. Структура системи перетворення інформації

Зважаючи на те, що в умовах кризового моніторингу на обґрунтування рішень виділяється не більше 2-3 хвилин, а властивості МВД змінюються динамічно, є необхідність в зменшенні часу перенавчання ієрархічної системи моделей. Ми цього досягаємо шляхом удосконалення процесу вибору кращого АСМ із існуючого в синтезаторі їх переліку. Процес послідовного випробування АСМ замінюється на процес розпізнавання його кращого варіанта серед присутніх в синтезаторі моделей шляхом застосування вирішуючого правила [3]. Це правило має вигляд поліноміальної моделі, що синтезована одним із алгоритмів МГУА [1]. Дана модель класифікує МВД до відповідного АСМ.

Вибір найбільш придатного АСМ проводиться на основі таких інформативних параметрів таблиць первинного опису [3]:

- кількість спостережень;
- кількість незалежних змінних;
- кількість параметрів максимально суміщених з функцією мети;
- кількість не суміщених параметрів;
- середній коефіцієнт кореляції незалежних змінних;
- середній коефіцієнт кореляції незалежних змінних та функції мети;
- середній коефіцієнт детермінації незалежних змінних;
- середній коефіцієнт детермінації незалежних змінних та функції мети;
- визначник нормованої таблиці первинного опису;
- визначник нормованої матриці значень незалежних змінних;
- власне число нормованої таблиці первинного опису;
- власне число матриці значень незалежних змінних.
- максимальне сингулярне число нормованої таблиці первинного опису;
- максимальне сингулярне число матриці значень незалежних змінних.

Коефіцієнт кореляції обраховується за формулою (**Ошибка! Источник ссылки не найден.**), а коефіцієнт детермінації – за формулою (2).

$$r_{xy} = \frac{\operatorname{cov}(x,y)}{\sqrt{s_x^2 \cdot s_y^2}} = \frac{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2 \cdot \sum_{i=1}^{m}\left(y_i - \bar{y}\right)^2}} \qquad (1)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}\left(y_i - x_i\right)}{\sum_{i=1}^{m}\left(y_i - \bar{x}\right)} \qquad (2)$$

Для розпізнавання найкращого алгоритму синтезу моделей потрібно використовувати модель, побудовану одним із індуктивних алгоритмів – наприклад, методом групового урахування аргументів[2]. Використання

класифікатора на основі моделі, отриманої з допомогою МГУА, складається з двох етапів:знаходження оптимальної структури моделі, та навчання моделі на таблиці, що задає функцію класифікації;використання створеної моделі для розпізнавання найкращого алгоритму синтезу моделей.

Даний метод був запропонований та апробований у [4]. При використанні класифікації швидкість перенавчання системи зростає, але при цьому зростає і похибка моделювання.

Також в результаті дослідження запропонованого в [4] методу було виявлено ряд недоліків. А саме, проблематичність в забезпеченні алгоритму навчання класифікаторів якісними даними. Користувачу потрібно знайти МВД, які добре навчаються представленими в нашому набору АСМ алгоритмами. Проте кожен такий МВД, що може складатися з сотень тисяч спостережень, формує лише одне спостереження при навчанні класифікаторів, а для якісного навчання може бути необхідно сотні або тисячі таких спостережень. Тому це вимагає від користувача великої кількості зусиль для забезпечення достатньої кількості МВД та отриманих з них спостережень.

Для усунення цієї проблеми нами запропоновано подальше покращення методу класифікації АСМ шляхом кластеризації даних, що використовуються для навчання класифікатора.

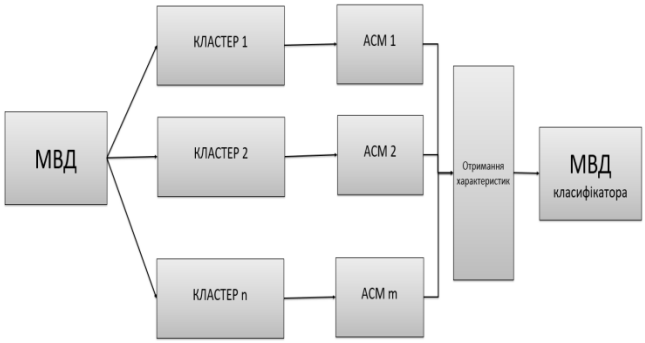

Рис. 2. Схема використання кластерів

Кожен кластер буде мати свій перелік характеристик та свій найкращий АСМ до якого його потрібно буде класифікувати. Отже з кожного МВД буде отримано не одне спостереження, а декілька в залежності від результатів кластеризації. Даний підхід повинен спростити роботу по забезпеченню даними процесу навчання.

II. Мета роботи

Метою цієї роботи є зменшення часу створення структури ієрархії моделей без значної втрати точності результатів моделювання на виході системи.

Для досягнення поставленої мети була сформована гіпотеза про те, що спостереження в МВД можливо кластеризувати та кожен кластер можна навчати окремо. Кожен кластер буде мати свій перелік характеристик та свій найкращий АСМ до якого його буде класифіковано. Це забезпечить вирішення проблеми з кількістю та якістю даних для навчання класифікаторів.

III. Результати досліджень

Сформульована гіпотеза була перевірена експериментально. Вирішуюче правило створювалось за допомогою багаторядного алгоритму МГУА [2]. В якості класифікаційних ознак МВД використовувався набір характеристик, запропонований в [3].

Для проведення дослідження в інструменті було реалізовано метод кластеризації К-середніх, мінімального дерева [5]. Відстані між об’єктами визначаються за допомогою відстаней Евкліда, Чебишева та L-норми [5].

Алгоритм починається з отримання ПО та формування МВД з нього. Наступним кроком є використання обраного методу для кластеризації МВД. Далі кожен кластер обробляється окремо для нього проводиться навчання усіма запропонованими АСМ після чого обирається найбільш придатний АСМ. Далі з кластера отримуються його характеристики [3]. Отриманий набір характеристик поміщається у МВД класифікатора до відповідного АСМ.

Для синтезу моделей даного дослідження використані результати моніторингу захворюваності населення Черкаської області впродовж 2000-2014 років [1].

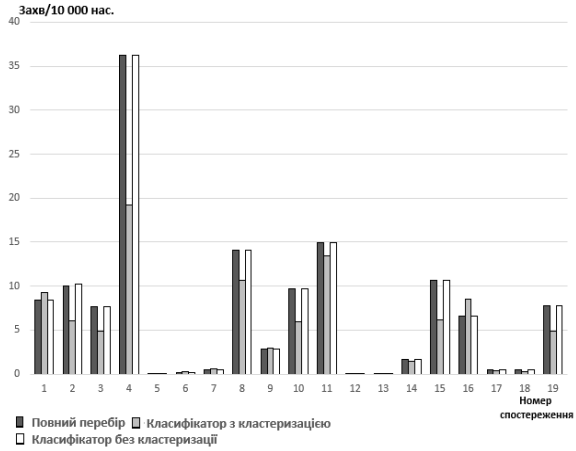

Рис. 3. Порівняння СКО

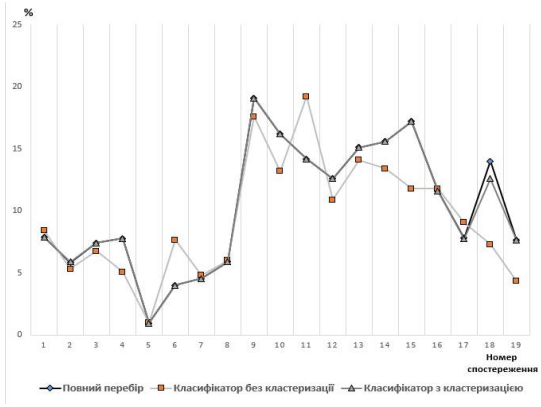

Рис. 4. Порівняння абсолютної похибки

Дослідження абсолютного (рис. 4) та середньо квадратичного (рис. 3) відхилення показало, що застосування процесів кластеризації даних дозволяє знизити похибку моделювання в середньому до10%.

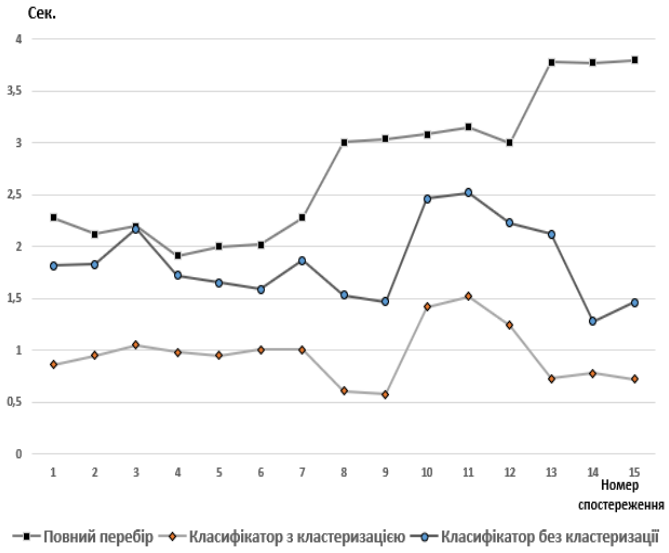

Рис. 5. Порівняння часу навчання/ перенавчання

При дослідженні часу навчання моделей (рис. 5) з використанням кластеризації було виявлено незначне погіршення в порівнянні з стандартним методом класифікації в обсязі до 30%.

Проте даний метод на 60% швидший в порівнянні з повним перебором.

В процесі дослідження було виявлено, що більша точність моделей класифікатора призводить до випадків фальшивого спрацювання. А це в свою чергу призводить до вповільнення алгоритму. Також найкращим методом кластеризації в даних умовах виявився метод мінімального дерева, а найкращою метрикою відстані L-норма.

Наступні дослідження будуть присвячені подальшому процесу параметричної оптимізації описаного методу. Також буде досліджено можливі способи впізнання класифікаторами не тільки придатності але і якості обраних АСМ та їх ранжування.

Висновки

Зростання похибки моделювання в методі [4] є «платою» за скорочення часу синтезу моделей. Зважаючи на те, що в структурі інформаційної системи багаторівневого перетворення даних міститься від 50 моделей і більше, вдається досягнути значного скорочення часу адаптації структури системи до зміни властивостей МВД. В умовах кризового моніторингу такі результати дають надію на можливість застосування моніторингових інформаційних систем із технологіями багаторівневого перетворення даних для підтримки прийняття рішень із локалізації наслідків надзвичайних ситуацій.

Проте в деяких випадках дана «плата» не допустима і алгоритм повинен враховувати це. Тому запропонована в даній роботі модернізація ціною деякого сповільнення дозволяє наблизити похибку моделювання до результатів повного перебору. В середньому в порівнянні з повним перебором похибка не перевищує 10%. Проте, в порівнянні з попереднім методом, швидкість впала в середньому на 30%.

Таким чином запропонований вдосконалений метод класифікації масиву вхідних даних в інформаційних системах багаторівневого моніторингу.

В даному методі зменшення часу перебудови структури моніторингової інформаційної системи для розв’язку нових задач в умовах надзвичайних ситуацій як і раніше досягається шляхом розв’язку задачі розпізнавання кращого алгоритму синтезу моделей за правилом, що створене за багаторядним алгоритмом МГУА. Проте покращення результатів моделювання досягається завдяки попередній кластеризації даних для навчанні класифікаторів, що призводить до збільшення їх якості та кількості, при використанні тих самих наборів даних що і попередній метод.

Література References

[1] Багаторівневе моделювання в технологіях моніторингу оточуючого середовища / С.В.Голуб ; Черкас. нац. ун-т ім. Б.Хмельницького. - Черкаси : ЧНУ, 2007. - 218 с.

[2] Ивахненко А.Г. Индуктивный метод самоорганизации моделей сложных систем. Киев: Наук. Думка, 1981. 296 с.

[3] Колос П.О. Визначення множини інформативних параметрів таблиці первинного опису об’єкта моделювання./ Вісник Черкаського університету, випуск 173. – Черкаси: Вид. ЧНУ, 2009. – С. 121-128.

[4] Avramenko A.: Classification models in information systems for social and environmental crisis monitoring / Avramenko A., Golub S. // Engineer of XXI Century – We design the future: Wydawnictwo Naukowe Akademii Techniczno-Humanistycznej w Bielsku-Bialej - Bielsku-Biala : ATH, 2016. – 928 p.

[5] J. Kogan, C. Nicholas, M. Teboulle – «Clustering Large and High Dimensional data» (http://www.csee.umbc.edu/ nicholas/clustering/tutorial.pdf)