

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
“ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ”
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Кваліфікаційна наукова праця
на правах рукопису

Ямковий Клим Сергійович

УДК 004.891.2

ДИСЕРТАЦІЯ
ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ПОБУДОВИ
КОМПОЗИТНИХ ІНДИКАТОРІВ НА ОСНОВІ
МЕТОДІВ МАШИННОГО НАВЧАННЯ

122 – Комп’ютерні науки
12 – Інформаційні технології

Подається на здобуття наукового ступеня доктор філософії

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

 К. С. Ямковий

Науковий керівник:
Любчик Леонід Михайлович,
доктор технічних наук, професор

Харків – 2023

АНОТАЦІЯ

Ямковий К. С. Інформаційні технології побудови композитних індикаторів на основі методів машинного навчання. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії (PhD) за спеціальністю 122 – Комп’ютерні науки. — Національний технічний університет “Харківський політехнічний інститут”, Харків, 2023.

У дисертаційній роботі вирішено науково-практичне завдання розробки методів та інформаційних технологій побудови композитних індикаторів на основі ядерних методів машинного навчання та оптимального узгодження експертної та статистичної інформації.

Об’єкт дослідження – процеси побудови композитних індикаторів у задачах ранжування і багатокритеріального оцінювання та вибору.

Предмет дослідження – методи та інформаційні технології побудови композитних індикаторів з на основі оптимального узгодження експертної і статистичної інформації та агрегації даних.

Мета і задачі дослідження – розробка методів та інформаційних технологій побудови композитних індикаторів основі ядерних методів машинного навчання та оптимального узгодження експертної та статистичної інформації з метою підвищення точності отриманих моделей та обмеження їх складності.

У вступі обґрунтовано актуальність теми дисертаційної роботи, зазначено зв’язок роботи з науковими темами, сформульовано мету, задачі та цілі дослідження, визначено об’єкт, предмет та методи дослідження, показано наукову новизну та практичне значення отриманих результатів, наведено інформацію про практичне використання, особистий внесок здобувача, апробацію результатів дослідження та їх висвітлення у публікаціях. Приводяться відомості щодо структури та обсягу дисертаційної роботи.

У першому розділі проведено аналіз задачі побудови композитних індикаторів та огляд різних підходів до їх побудови, зокрема, методів машинного навчання. Наведено приклади використання композитних індикаторів

у багатьох сферах для побудови різних узагальнених показників: людського розвитку, екологічної ефективності, інвестиційного портфеля тощо. Наведені приклади використання композитних індикаторів в задачах ранжування: рекомендаційні системи інтернет-магазинів, рекомендаційні системи медіаконтенту, пошукові системи (Google). Проаналізовані різні підходи та алгоритми у навчанні ранжуванню: точковий, парний та списковий. Розглянуто метрики оцінки якості роботи алгоритмів ранжування, які використовуються для оцінювання запропонованих алгоритмів. Визначено особливості збору та використання експертно-статистичної інформації для побудови композитних індикаторів. Обґрунтовано вибір мети і задач роботи.

У другому розділі задача побудови композитного індикатора сформульована в термінах машинного навчання, отримано розв'язання задачі побудови нелінійної моделі композитного індикатору на основі ядерної рідж-регресії. Проаналізовано методи узгодження різнорідної експертної інформації, що дозволяють знайти компроміс між експертними оцінками композитних індикаторів та статистичними оцінками часткових показників. Обґрунтовано запропонований метод оптимального узгодження експертної і статистичної інформації за допомогою регуляризації ядерної регресії з використанням апіорної інформації щодо важливості часткових показників, що суттєво підвищує точність отриманих моделей.

У третьому розділі наведено принципи концепції великих даних та описані проблеми, які виникають при збільшенні обсягів інформації, що використовується для побудови композитних індикаторів. Запропоновано використовувати методи агрегування даних для зменшення складності ядерної моделі. Розглянуто методи групування та кластеризації для агрегації даних. Для підвищення точності та ефективності кластеризації запропоновано використовувати регуляризацию за допомогою цільової змінної на етапі розрахунку відстані між точками в просторі ознак, викладено запропонований метод регуляризованої кластеризації. Визначено проблему недостатньої розмітності даних, яка особливо часто виникає при збільшенні кількості даних. Для розв'язання цієї проблеми запропоновано використовувати методи напівкерованого навчання на основі графової регуляризації та ядерного трюку під час оптимізації нелінійної функції переваг. Для розв'язання зазначених проблем розроблено двоетапний алгоритм агрегації даних, який використо-

вує як глобальні, так і локальні патерни у структурі набору під час агрегації. Такий підхід дозволяє значно скоротити розмір вибірки при збереженні всіх властивостей та шаблонів.

У четвертому розділі наведено опис запропонованої інформаційної технології для побудови композитних індикаторів методами машинного навчання, яка реалізує розроблені у роботі методи та алгоритми. Розроблена інформаційна технологія реалізована у вигляді бібліотеки мовою програмування Python з відкритим похідним кодом та наслідуванням інтерфейсів бібліотеки scikit-learn та відповідає усім вимогам методологій розробки проєктів у сфері машинного навчання та аналізу даних, а саме KDD та CRISP-DM. Перевірено працездатність розробленої інформаційної технології, точність запропонованих алгоритмів та проаналізовано отримані результати дослідження. Для цього використано декілька багатомірних наборів даних, які представляють різні домені області. Результати показали працездатність та ефективність запропонованих у роботі методів та алгоритмів.

У висновках наведено основні результати дисертаційної роботи щодо вирішення поставлених наукових задач дослідження. За результатами дослідження отримано такі наукові результати:

1. Вперше розроблено метод побудови композитних індикаторів з використанням нелінійної функції переваг, що відрізняється поєднанням ядерних методів машинного навчання з рідж-регресією та забезпечує можливість побудови нелінійних моделей;
2. Отримали подальший розвиток методи узгодження експертної і статистичної інформації, вперше запропоновано метод оптимального узгодження з використанням регуляризуючого функціоналу, що містить експертні оцінки, що забезпечує підвищення точності отриманих моделей;
3. Вперше розроблено алгоритм двоетапної агрегації даних, який використовує ідею пошуку глобальних і локальних патернів у вибірці та регуляризовану кластеризацію, що дозволяє значно зменшувати розмір набору даних при збереженні його властивостей та характеристик та будувати редуковані моделі індикаторів;
4. Вперше запропоновано метод регуляризації при побудові нелінійних моделей індикаторів, який використовує немарковані експертами спосте-

реження з використанням напівконтрольованого графового підходу у поєднанні з ядерними методами, що забезпечує можливість використання даних з обмеженою експертною інформацією;

5. Вдосконалена інформаційна технологія для побудови композитних індикаторів методами машинного навчання, що відрізняється комбінацією підходів, а саме, використанням ядерних методів машинного навчання, оптимального узгодження експертної та статистичної інформації, агрегації даних та напівкерованого навчання.

Результати роботи впроваджено у науково-дослідних роботах та навчальному процесі кафедри комп'ютерної математики і аналізу даних Національного технічного університету "Харківський політехнічний інститут". Запропоновані методи та інформаційна технологія можуть бути повною мірою впроваджені в організаціях всіх галузей економіки для підвищення якості вибору альтернатив та автоматизації процесів побудови композитних індикаторів і ранжування.

Ключові слова: машинне навчання, багатокритеріальна оптимізація, ранжування, композитний індикатор, експертна оцінка, узгодження, агрегація, кластеризація, функція переваг, ядерна функція, регуляризація, порядкова регресія, напівконтрольоване навчання.

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА

Публікації здобувача за темою дисертації, в яких опубліковані основні наукові результати:

1. K. Yamkovyi, “Adaptation of LambdaMART model to semi-supervised learning,” *Вісник Національного технічного університету «ХПІ»*. Серія: Системний аналіз, управління та інформаційні технології, № 1 (9), с. 76—81, 2023
2. L. Lyubchuk та K. Yamkovyi, “Comparative Analysis of Modified Semi-Supervised Learning Algorithms on a Small Amount of Labelled Data,” *System Research & Information Technologies*, № 4, с. 34—43, 2022
3. О. Ахієзер, Г. Грінберг, Л. Любчик та К. Ямковий, “Failure rate regression model building from aggregated data using kernel-based machine learning methods,” *Вісник Національного технічного університету «ХПІ»*. Серія: Системний аналіз, управління та інформаційні технології, № 2 (8), с. 51—56, 2022
4. K. Yamkovyi, “Development and comparative analysis of semi-supervised learning algorithms on a small amount of labeled data,” *Вісник Національного технічного університету «ХПІ»*. Серія: Системний аналіз, управління та інформаційні технології, № 1 (5), с. 98—103, 2021
5. L. Lyubchuk, G. Grinberg, Z. Konokhova та K. Yamkovyi, “Composite Indicators Building Based on Concordant of Expert-Statistical Information Using Biased Ridge Kernel Regression,” в *13th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, IEEE, т. 1, Дортмунд, Німеччина, 2023, с. 617—620
6. L. Lyubchuk, О. Akhiezer, G. Grinberg та K. Yamkovyi, “Machine Learning-Based Failure Rate Identification for Predictive Maintenance in Industry 4.0,” в *2022 12th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, IEEE, Афіни, Греція, 2022, с. 1—5. DOI: 10.1109/DESSERT58054.2022.10018614

7. L. Lyubchuk, G. Grinberg та K. Yamkovyi, “Integral Indicator for Complex System Building Based on Semi-Supervised Learning,” в *2018 IEEE First International Conference on System Analysis & Intelligent Computing (SAIC)*, IEEE, Київ, Україна, 2018, с. 1—5

Публікації здобувача за темою дисертації, які засвідчують апробацію матеріалів дисертації:

8. L. Lyubchuk, G. Grinberg та K. Yamkovyi, “Learning to rank on aggregated data in real time,” *XXXVI International Conference Problems of Decision Making Under Uncertainties (PDMU-2021)*, с. 62, 2021
9. К. Ямковий та Л. Любчик, “Експертно-статистичне оцінювання інтегральних індикаторів методами машинного навчання,” в *XIII Міжнародна науково-практична конференція магістрантів та аспірантів*, Харків, Україна, 2019, с. 170—171

ABSTRACT

Yamkovyi K. S. Information technologies for building composite indicators based on a machine learning approach. – Manuscript.

The thesis is submitted to obtain a scientific degree of Doctor of Philosophy in the specialty 122 – Computer Sciences. – National Technical University “Kharkiv Polytechnic Institute”, Kharkiv, 2023.

The scientific and practical task of developing methods and information technologies for building composite indicators based on kernel methods of machine learning and optimal concordance of expert and statistical information is solved in the dissertation work.

Research object – the processes of building composite indicators in the tasks of ranking and multi-criteria evaluation and selection.

Research subject – methods and information technologies for building composite indicators based on optimal concordance of expert and statistical information and data aggregation.

The purpose and objectives of the research – the development of methods and information technologies for building composite indicators based on kernel methods of machine learning and optimal concordance of expert and statistical information, to increase the accuracy of the obtained models and limit their complexity.

The introduction substantiates the relevance of the topic of the dissertation, indicates the connection of the work with scientific topics, formulates the goal, tasks and objectives of the research, defines the object, subject and methods of the research, shows the scientific novelty and practical significance of the obtained results, provides information about practical use, personal contribution of the recipient, the approbation of research results and their coverage in publications. Information on the structure and scope of the dissertation work is given.

In the first chapter, an analysis of the task of constructing composite indicators and an overview of various approaches to their construction, in particular, machine learning methods, was carried out. Examples of the use of composite indicators in many areas for the construction of various generalized indicators

are given: human development, environmental efficiency, investment portfolio, etc. Examples of the use of composite indicators in ranking tasks are given: recommender systems of online stores, recommender systems of media content, and search engines (Google). Different approaches and algorithms in learning to rank are analyzed: pointwise, pairwise, and listwise. Metrics for evaluating the quality of ranking algorithms, which are used to evaluate the proposed algorithms, are considered. Defined the specifics of the collection and use of expert statistical information for the construction of composite indicators. The choice of the goal and tasks of the work is substantiated.

The second chapter formulates the task of constructing a composite indicator in terms of machine learning, and a solution to the task of constructing a nonlinear model of a composite indicator based on kernel ridge regression is obtained. The methods of concordance of disparate expert information, which allow finding a compromise between expert assessments of composite indicators and statistical assessments of partial indicators, are analyzed. The proposed method of optimal concordance of expert and statistical information using kernel regression regularization with the use of a priori information on the importance of partial indicators is justified, which significantly increases the accuracy of the obtained models.

The third chapter provides the principles of the concept of big data and describes the problems that arise when the amount of information used to construct composite indicators increases. It is proposed to use data aggregation methods to reduce the complexity of the kernel model. Methods of grouping and clustering for data aggregation are considered. To increase the accuracy and efficiency of clustering, it is proposed to use regularization with the help of a target variable at the stage of calculating the distance between points in the feature space, and the proposed method of regularized clustering is outlined. The problem of insufficient data marking, which especially often arises when the amount of data increases, is identified. To solve this problem, it is proposed to use semi-supervised learning methods based on graph regularization and kernel trick during the optimization of the nonlinear preference function. To solve these problems, a two-stage data aggregation algorithm was developed, which uses both global and local patterns in the set structure during aggregation. This approach allows to significant reduce the size of the sample while preserving all properties

and patterns.

The fourth chapter describes the proposed information technology for building composite indicators using machine learning methods, which implement the methods and algorithms developed in the work. The developed information technology is implemented in the form of a library in the Python programming language with open-source source code and inherits scikit-learn library interfaces and meets all the requirements of project development methodologies in the field of machine learning and data analysis, namely KDD and CRISP-DM. The functionality of the developed information technology, the accuracy of the proposed algorithms, and the obtained research results were analyzed. For this, several multidimensional data sets representing different domains of the area were used. The results showed the efficiency and effectiveness of the methods and algorithms proposed in the work.

In the conclusions, the main results of the dissertation work on the solution of the set scientific research problems are presented. According to the results of the research, the following scientific results were obtained:

1. For the first time, a method of constructing composite indicators using a non-linear preference function was developed, which is characterized by a combination of kernel machine learning methods with ridge regression model and provides the possibility of constructing non-linear models;
2. The methods of concordance of expert and statistical information have received further development, for the first time the method of optimal matching using a regularizing functional containing expert evaluations, which ensures an increase in the accuracy of the obtained models, has been proposed;
3. For the first time, a two-stage data aggregation algorithm was developed, which uses the idea of finding global and local patterns in the sample and regularized clustering, which allows you to significantly reduce the size of the data set while preserving its properties and characteristics and build reduced indicator models;
4. For the first time, a regularization method was proposed for the construction of nonlinear models of indicators, which uses unlabeled observations by experts using a semi-supervised graph approach combined with kernel methods, which provides the possibility of using data with limited expert

information;

5. Advanced information technology for building composite indicators using machine learning methods, characterized by a combination of approaches, namely, the use of nuclear machine learning methods, optimal matching of expert and statistical information, data aggregation and semi-supervised learning.

The work results are implemented in the research works and educational process of the Department of Computer Mathematics and Data Analysis of the National Technical University “Kharkiv Polytechnic Institute”. The proposed methods and information technology can be fully implemented in organizations of all sectors of the economy to improve the quality of the choice of alternatives and automate the composite indicators building and ranking processes.

Keywords: machine learning, multicriteria decision analysis, ranking, composite indicator, expert evaluation, concordance, aggregation, clustering, preference function, kernel function, regularization, ordinal regression, semi-supervised learning.

LIST OF APPLICANT PUBLICATIONS

Publications of the applicant on the topic of the dissertation, in which the main scientific results are published:

1. K. Yamkovyi, “Adaptation of lambdamart model to semi-supervised learning”, *Bulletin of National Technical University «KhPI». Series: System Analysis, Control and Information Technologies*, no. 1 (9), pp. 76–81, 2023
2. L. Lyubchyk and K. Yamkovyi, “Comparative analysis of modified semi-supervised learning algorithms on a small amount of labelled data”, *System Research & Information Technologies*, no. 4, pp. 34–43, 2022
3. O. Akhiezer, G. Grinberg, L. Lyubchyk, and K. Yamkovyi, “Failure rate regression model building from aggregated data using kernel-based machine learning methods”, *Bulletin of National Technical University «KhPI». Series: System Analysis, Control and Information Technologies*, no. 2 (8), pp. 51–56, 2022
4. K. Yamkovyi, “Development and comparative analysis of semi-supervised learning algorithms on a small amount of labeled data”, *Bulletin of National Technical University «KhPI». Series: System Analysis, Control and Information Technologies*, no. 1 (5), pp. 98–103, 2021
5. L. Lyubchyk, G. Grinberg, Z. Konokhova, and K. Yamkovyi, “Composite indicators building based on concordant of expert-statistical information using biased ridge kernel regression”, in *13th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, IEEE, vol. 1, Dortmund, Germany, 2023, pp. 617–620
6. L. Lyubchyk, O. Akhiezer, G. Grinberg, and K. Yamkovyi, “Machine learning-based failure rate identification for predictive maintenance in industry 4.0”, in *2022 12th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, IEEE, Athens, Greece, 2022, pp. 1–5. DOI: 10.1109/DESSERT58054.2022.10018614
7. L. Lyubchyk, G. Grinberg, and K. Yamkovyi, “Integral indicator for complex system building based on semi-supervised learning”, in *2018 IEEE First*

International Conference on System Analysis & Intelligent Computing (SAIC), IEEE, Kyiv, Ukraine, 2018, pp. 1–5

Publications of the applicant on the topic of the dissertation, which certify the approbation of the dissertation materials:

8. L. Lyubchyk, G. Grinberg, and K. Yamkovyi, “Learning to rank on aggregated data in real time”, *XXXVI International Conference Problems of Decision Making Under Uncertainties (PDMU-2021)*, p. 62, 2021
9. K. Yamkovyi and L. Lyubchyk, “Expert-statistical evaluation of integral indicators using machine learning methods”, in *XIII International Scientific and Practical Conference of Master’s and Postgraduate Students*, Kharkiv, Ukraine, 2019, pp. 170–171

ЗМІСТ

Перелік позначень і скорочень	4
Вступ	6
1 Завдання та методи побудови композитних індикаторів	11
1.1 Поняття композитного індикатора, приклади, сфери застосування	11
1.2 Методи ранжування на основі композитних індикаторів	16
1.3 Метрики оцінки якості ранжування	19
1.4 Експертна і статистична інформація для побудови композитних індикаторів	23
1.5 Формулювання цілей та завдань роботи	27
1.6 Висновки до розділу 1	28
2 Побудова композитних індикаторів методами машинного навчання з узгодженням інформації	31
2.1 Постановка задачі побудови композитних індикаторів з використанням експертної і статистичної інформації	31
2.2 Побудова композитних індикаторів за допомогою методу порядкової регресії	32
2.3 Побудова нелінійних ядерних моделей композитних індикаторів	34
2.4 Оптимальне узгодження експертної і статистичної інформації .	36
2.5 Висновки до розділу 2	41
3 Побудова композитних індикаторів на основі агрегованих даних	42
3.1 Проблеми побудови композитних індикаторів на великих даних	42
3.2 Побудова композитних індикаторів на агрегованих даних	44
3.3 Побудова критеріїв кластеризації на основі показників близькості у просторі показників та близькості рангів на кластерах .	51

3.4	Побудова функції переваг за допомогою методів напівконтрольованого навчання	53
3.5	Висновки до розділу 3	58
4	Інформаційні технології вирішення практичних завдань побудови композитних індикаторів	60
4.1	Розробка інформаційної технології побудови композитних індикаторів	60
4.1.1	Функціональні можливості інформаційної технології . . .	60
4.1.2	Архітектура інформаційної технології	64
4.1.3	Опис інтерфейсу	66
4.2	Вихідна інформація для проведення контрольних розрахунків .	68
4.2.1	Синтетичні дані	68
4.2.2	Набір даних “Laptop Ranked Specifications”	68
4.3	Побудова базової моделі композитних індикаторів	71
4.4	Застосування методів узгодження на прикладі набору даних “Laptop Ranked Specifications”	73
4.5	Побудова моделей композитних індикаторів на агрегованих даних	79
4.5.1	Агрегація синтетичних даних	79
4.5.2	Агрегація набору даних “Laptop Ranked Specifications” . .	81
4.6	Побудова моделей з використанням напів-контрольованого навчання	83
4.7	Висновки до розділу 4	84
	Висновки	86
	Список джерел інформації	89
	Додаток А	103
	Додаток Б	105
	Додаток В	108