Запорізький національний університет
Міністерство освіти і науки України

Кваліфікаційна наукова праця
на правах рукопису

**СЄЛЮТІН ЄВГЕН КИРИЛОВИЧ**

**УДК**

**ДИСЕРТАЦІЯ**

# ФРАГМЕНТАРНІ МОДЕЛІ В ЗАДАЧАХ ОПТИМАЛЬНОЇ КЛАСИФІКАЦІЇ

113 прикладна математика
11 фізико-математичні науки

Подається на здобуття наукового ступеня доктора філософії. Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

Є.К. Сєлютін

Науковий керівник
**Козін Ігор Вікторович**
доктор фізико-математичних наук, професор

Запоріжжя 2021

# АНОТАЦІЯ

*Сєлютін Є.К.* Фрагментарні моделі в задачах оптимальної класифікації. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 113 – Прикладна математика. – Запорізький національний університет, Запоріжжя, 2021.

Дисертаційна робота присвячена узагальненню і розробці теоретичних основ математичного апарату для побудови та дослідження фрагментарних моделей та метаевристичних методів для розв'язку задачі класифікації. Розглянуті в дисертації моделі та методи можуть використовуватися під час пошуку розв'язків у багатьох наукових та практичних задачах, в тому числі, задачі розміщення виробництва.

У **вступі** обґрунтовано актуальність теми дисертаційної роботи, сформульовано мету та основні задачі дослідження, показано їх зв'язок з науковими програмами. Визначено методи дослідження, наукову новизну та практичне значення отриманих результатів.

У **розділі 1** проведено огляд результатів за тематикою дисертаційної роботи. Обґрунтовано вибір напрямків подальших досліджень, пов'язаних з розв'язком задачі оптимальної класифікації. Розглянуто та проаналізовано існуючі задачі оптимальної класифікації та методів пошуку їх розв'язків. Досліджено обчислювальну складність існуючих задач класифікації.

Виявлено, що постановка задачі пошуку оптимальної класифікації дозволяє віднести цю задачу до багатокритеріальних. Майже всі математичні методи оптимізації призначено для відшукання оптимального рішення однієї функції – одного критерію. Тому одним з варіантів вирішення багатокрітеріальних завдань є її приведення до однокрітеріальною з одним узагальненим критерієм.

Проаналізовано, що для вирішення завдання класифікації досить тривалий час використовувалися методи найближчого сусіда (Nearest Neighbor), K-найближчого сусіда (k-Nearest Neighbor); байєсовські мережі (Bayesian Networks); індукція дерев рішень; нейронні мережі (neural networks), метод опорних векторів; статистичні методи, зокрема, лінійна регресія; класифікація cbr – методом або за допомогою генетичних алгоритмів. Але більш ефективними себе показали метаевристичні методи.

Виявлено, що більшість задач оптимальної класифікації є важкорозв’язуваними (складними в обчислювальному сенсі), оскільки до них поліноміально зводиться хоча б одна NP-повна задача. Для таких задач на сьогодні невідомі алгоритми пошуку точного розв'язку простіших, ніж повний перебір всіх допустимих розв'язків задачі. Тому є сенс шукати прості наближені алгоритми, які хоч і не дають точного розв'язку, але мають високу швидкодію. Серед таких алгоритмів виділяється клас «жадібних» алгоритмів.

Виділено невирішені задачі класифікації, зокрема, розпізнавання спаму електронної пошти, зображень та мовлення в якості цифрового паролю, опис мікроматриці ДНК тощо.

У **розділі 2** наведено ряд основних понять та результатів, які використовуються в роботі та стосуються комбінаторного об’єкту «фрагментарна структура». Вивчено особливості та властивості фрагментарних структур, встановлено зв’язок з категорією «опуклість» та задачею покриття графів. Досліджено поняття метаевристики та проаналізовано метаевристичні методи пошуку оптимальних рішень у задачі класифікації. Встановлено зв’язок між задачею оптимальної перестановки та класифікації.

Визначено, що задачі оптимальної класифікації на дискретних множинах мають велику кількість переваг. Зокрема, вони адекватно відображають нелінійні залежності, неподільність об'єктів, враховують логічні та технологічні обмеження, а також «якісні» вимоги. Але у прикладних задачах,

як правило, є багато обмежень, які ускладнюють застосовність відомих алгоритмів. Тому для таких задач є виправданим застосування метаевристик.

Виявлено, що експериментальні дослідження розподілу локальних оптимумів свідчать про високу концентрацію їх в безпосередній близькості від глобального оптимуму (гіпотеза про існування «великої долини» для задач на мінімум або «центрального гірського масиву» для задач на максимум).

Показано, що проблема пошуку початкових популяцій для реалізації еволюційно-фрагментарної моделі зводиться до *задачі покриття (або розбиття) простору перестановок опуклими множинами.*

У **розділі 3** розглянуто фрагментарний алгоритм покриття графу зірками. Проаналізовано найпростіші варіанти задачі розміщення виробництва з точки зору задачі класифікації, а також наведено метаевристичні алгоритми пошуку оптимальних рішень складних задач розміщення виробництва.

Сформульована задача покриття: заданий граф $G = (V, E)$, ребра якого $E$ зважені функцією $p: E \to R_+^1$. Знайти підмножину мінімальної ваги з неперетинних по вершинах зірок у цьому графі, об'єднання яких містить всі вершини графу $G$.

Побудовано фрагментарну модель задачі. Множиною елементарних фрагментів є множина $E$ всіх ребер графу, елементи якого нумеруються числами 1,2,...,$m$.

Визначено умову приєднання елементарного фрагменту: або ребро $e_i \in E$ має рівно одну спільну вершину з однією з зірок фрагмента і ця вершина – центр зірки, або це ребро не має загальних вершин з уже обраними ребрами. Розглядаються довільні впорядкування ребер графу, кожне з яких описується перестановкою з групи перестановок $S_m$.

Розглянуто задачі розміщення виробництва – одні з найбільш поширених та актуальних в наш час, що мають масовий характер.

Визначено, що найкращі результати при розв’язанні задачі про розподіл економічного навантаження з урахуванням впливу на навколишнє

середовище демонструє метод рою часток із підбором коефіцієнтів соціалізації та персоналізації на основі генетичного алгоритму. В тому числі, ці результати є кращими за результати генетичного алгоритму, який є досить відомим при розв’язанні даної задачі. Більше того, результати генетичного алгоритму є найгіршими у порівнянні із усіма розглянутими алгоритмами. Визначена доцільність використання методу рою часток та фрагментарної моделі, а також їх модифікацій.

У **розділі 4** було запропоновано програмні реалізації генерації випадкових графів та розв’язку задачі класифікації за допомогою метаевристичних алгоритмів. Було проведено порівняння ефективності роботи алгоритмів.

Наведено алгоритм генерації графів з 50 вершинами у розрідженому (50 ребер) та насиченому (500 ребер) варіантах. Розріджений граф не є зв'язковим, оскільки кожна його вершина з'єднана тільки з невеликою кількістю інших вершин; насичений граф, безсумнівно, є зв'язковим, тому що кожна його вершина пов'язана в середньому з 20 іншими вершинами.

Розглянуто реалізацію створення випадкових ребер. Для заданої кількості вершин *V* генеруються довільні ребра, тобто пари випадкових чисел від *0* до *V-1*. Результатом, швидше за все, буде довільний мультиграф з петлями. Будь-яка пара може містити два однакових числа (тобто можливі петлі); і будь-яка пара може повторитися кілька разів (тобто можливі паралельні ребра). Програма генерує ребра до тих пір, поки не набереться *E* ребер; рішення про видалення паралельних ребер залишається за реалізацією. Якщо видаляти паралельні ребра, то в насичених графах кількість генеруються ребер буде значно більше, ніж кількість використаних ребер (*E*); тому даний метод зазвичай використовується для розріджених графів.

Проведено оцінку якості метаевристичних алгоритмів розв’язку задачі класифікації на основі генерації випадкових графів за допомоги порівняння результатів роботи цього алгоритму з іншими алгоритмами на досить великий серії задач.

Для оцінки ефективності метаевристик на фрагментарних структурах було розроблено програму оцінки ефективності для різних модельних задач, в якій для багатьох дискретних задач, що допускають фрагментарну модель, були реалізовані універсальні алгоритми ряду метаевристик. Зокрема, реалізовані метод випадкового пошуку, метод ітеративного локального пошуку, метод імітації відпалу, еволюційно-фрагментарний алгоритм, метод перемішаних стрибаючих жаб. У програмі реалізовані генератор випадкових завдань з різними обмеженнями, база даних завдань і програма порівняння ефективності алгоритмів.

Для задачі покриття графа зірками проведено чисельний експеримент на базі 53 випадково згенерованих завдань. Розглядалися зв'язкові графи з числом вершин від 20 до 50 і з щільністю ребер 0,5-0,8. Для кожної з задач було побудовано фрагментарну модель і застосовувалася група алгоритмів (випадковий пошук, еволюційно-фрагментарний алгоритм, метод імітації відпалу тощо). Параметри алгоритмів підбиралися таким чином, щоб трудомісткість обчислень була приблизно однаковою.

Результати порівняння різних алгоритмів показують, що жоден з них не володіє явною перевагою перед іншими. Це побічно підтверджує відому теорему «про відсутність безкоштовних обідів» для метаевристик. Таким чином, в умовах реальної експлуатації розумно застосовувати не одну, а кілька метаевристик і вибирати найкращий результат. Розглянутий в дисертаційній роботі метод використання фрагментарних моделей для пошуку субоптимальних рішень задач класифікації, дозволяє порівняно просто побудувати універсальну комп'ютерну систему для таких завдань. Причому універсальними будуть програми реалізації метаевристик, а індивідуальними методи побудови фрагментарних моделей і алгоритми розрахунку значень критеріїв.

**Ключові слова:** фрагментарна модель, метаевристичні методи, генетичний алгоритм, метод стрибаючих жаб, задача оптимальної класифікації.

# ABSTRACT

Selyutin E.K. Fragmentary models in optimal classification problems. – Qualifying scientific work on the rights of the manuscript.

The dissertation on competition of a scientific degree of the doctor of philosophy on a specialty 113 Applied Mathematics. – Zaporizhzhya National University, Zaporizhzhya, 2021.

The dissertation is devoted to the generalization and development of the theoretical basis for the mathematical apparatus for constructing and researching fragmentary models and metaheuristics methods for solving a classification problem. The models and methods considered in the thesis can be used for resolving many scientific and practical problems, including the problem of manufactory location.

The introduction substantiates the relevance of the study, formulates major goals and objectives of the research, shows their connection with scientific programs. Defined research methods, scientific novelty and practical significance of the results.

**The Chapter 1** shows the overview of the results on the subject of the dissertation work. The choice of directions of further research related to the solution of the optimal classification problem is substantiated. The existing problems of optimal classification and methods of searching for their solutions are considered and analyzed. The computational complexity of existing problems of classification is investigated.

It was found that setting the task of finding the optimal classification allows you to attribute this task to multicritical. Almost all mathematical optimization methods are designed to find the optimal solution of one function – one copy. Therefore, one of the solutions to multi-copy problems is to bring it to a single-key with one generalized criterion.

It was analyzed that the methods of the nearest neighbor (Nearest Neighbor), K-Nearest Neighbor (k-Nearest Neighbor) were used to solve the classification problem for quite a long time; Bayesian Networks; induction of decision trees; neural networks, method of reference vectors; statistical methods, in particular, linear regression; classification CBR – by method or by means of genetic algorithms. But metaheuristic methods have shown themselves to be more effective.

It was found that most optimal classification tasks are difficult to solve (complex in the computational sense), since at least one NP-complete problem is polynomially reduced to them. For such problems, today unknown algorithms for finding an exact solution are easier than a complete selection of all permissible solutions to the problem. Therefore, it makes sense to look for simple approximate algorithms that, although they do not give an accurate solution, but have high performance. Among such algorithms stands out class of "greedy" algorithms.

Unsolved classification tasks such as e-mail spam recognition, images and speech as a digital password, description of DNA microarray are highlighted.

**The Chapter 2** provides a number of basic concepts and results that are used in the work and relate to the "fragmentary structure" combinatoric object. The peculiarities and properties of fragmentary structures were studied, the connection with the category "bulge" and the task of covering graphs was established. The concept of meta-heuristics is studied and metaheuristic methods of searching for optimal solutions in the problem of classification are analyzed. The connection between the task of optimal permutation and classification has been established.

It is determined that the problems of optimal classification on discrete sets have a large number of advantages. In particular, they adequately reflect nonlinear dependencies, the indivisibility of objects, take into account logical and technological restrictions, as well as "high-quality" requirements. But in applied tasks, as a rule, many limitations complicate the applicability of known algorithms. Therefore, for such tasks, the use of meta-heuristics is justified.

It was found that experimental studies of the distribution of local optimums indicate a high concentration of them near of the global optimum (the hypothesis about the existence of a "great valley" for minimum tasks or "central mountain range" for maximum tasks).

It is shown that the problem of finding initial populations for the implementation of an evolutionary-fragmentary model is reduced to the task of covering (or breaking) the permutation space with convex sets.

**The Chapter 3** discusses the fragmentary algorithm for covering graph stars. The simplest options for the task of placing production in terms of the classification task are analyzed, as well as metaheuristic algorithms finding optimal solutions to complex production placement tasks are given.

The coverage problem is formulated: a given graph $G = (V, E)$ whose edges E are weighted by a function $p: E \to R_+^1$. Find a subset of the minimum weight of non-intersecting vertices of stars in this graph, the union of which contains all the vertices of the G`s graph.

A fragmentary model of the problem is constructed. The set of elementary fragments is the E set of all edges of the graph, the elements of which are numbered 1,2, ..., m.

The condition of joining an elementary fragment is determined: either the edge $e_i \in E$ has exactly one common vertex with one of the stars of the fragment and this vertex is the center of the star, or this edge has no common vertices with the already selected edges. Arbitrary orderings of the edges of the graph are considered, each of which is described by a permutation from the group of permutations $S_m$.

The problems of production location, the most common and relevant in current time and which have a mass character, are considered.

It is determined that the best results in solving the problem of distribution of economic burden considering the impact on the environment demonstrate the method of swarming of particles with the selection of coefficients of socialization and personalization based on a genetic algorithm. In particular, these results are

better than the results of the genetic algorithm, which is well known for solving this problem. Moreover, the results of the genetic algorithm are the worst in comparison with all the considered algorithms. The expediency of using the method of particle swarm and fragmentary model, as well as their modifications is determined.

**The Chapter 4** proposed software implementations for generating random graphs and solving the classification problem using metaheuristic algorithms. A comparison of the efficiency of the algorithms was performed.

An algorithm for generating graphs with 50 vertices in sparse (50 edges) and saturated (500 edges) variants is given. A sparse graph is not connected, since each of its vertices is connected to only a small number of other vertices; a saturated graph is undoubtedly connected, because each of its vertices is associated with an average of 20 other vertices.

The implementation of creating random edges is considered. For a given number of vertices V, arbitrary edges are generated, which means pairs of random numbers from 0 to V-1. The result is likely to be an arbitrary multigraph with loops. Any pair can contain two identical numbers (possible loops), and any pair can be repeated several times (parallel edges are possible). The program generates edges until E edges are typed; the decision to remove the parallel edges remains to be implemented. If you delete the parallel edges, then in saturated graphs the number of generated edges will be much greater than the number of used edges (E); therefore, this method is usually used for sparse graphs.

The quality of metaheuristic algorithms for solving the classification problem based on the generation of random graphs is evaluated by comparing the results of this algorithm with other algorithms on a fairly large series of problems.

To estimate the effectiveness of metaheuristics on fragmentary structures, a program for evaluating the effectiveness of various model problems was developed, where for many discrete problems that allow a fragmentary model, universal algorithms of a number of metaheuristics have been implemented. In particular, the method of random search, the method of iterative local search, the

method of simulation of annealing, the evolutionary-fragmentary algorithm, the method of mixed jumping frogs are implemented.

For the problem of covering the graph with stars, a numerical experiment was performed based on 53 randomly generated problems. Connected graphs with the number of vertices from 20 to 50 and with edges` density of 0.5-0.8 were considered. A fragmentary model was built for each of the problems and a group of algorithms was used (random search, evolutionary-fragmentary algorithm, annealing simulation method, etc.). The parameters of the algorithms were selected so that the complexity of the calculations was approximately the same.

The results of comparing different algorithms show that none of them has a clear advantage over the others. This indirectly confirms the well-known "No free lunch" theorem for metaheuristics. Thus, in the conditions of real operation, it is reasonable to apply not one, but several metaheuristics and choose the best result. The method of using fragmentary models for finding suboptimal solutions to classification problems considered in the dissertation allows building a relatively universal computer system for such problems. Moreover, the programs of realization of metaheuristics will be universal, and the methods of construction of fragmentary models and algorithms of calculation of values of criteria will be individual.

**Keywords:** fragmentary model, metaheuristic methods, genetic algorithm, jumping frog method, optimal classification problem.

**Список опублікованих праць за темою дисертації**

*Статті у наукових фахових виданнях України:*

1. Козін, І. В., **Селютін, Є. К.** Особливості пошуку оптимальних класифікацій: еволюційні алгоритми / Вісник Запорізького національного університету. Фізико-математичні науки, (2), 2020. – С. 62 – 68.

2. **Selyutin Y.,** Kozin I. Comparative effectiveness of metaheuristic methods / Науковий вісник Ужгородського університету : серія Математика і Інформатика / редкол. : М. М. Маляр, Г. І. Сливка-Тилищак та ін. – Ужгород : Говерла, 2020. – Вип. 1 (36). – С. 105–111.

3. Козин И.В., **Селютин Е.К.** Метаэвристики для поиска оптимальных классификаций / Питання прикладної математики і математичного моделювання [Текст]: зб. наук. пр. / редкол.: О.М. Кісельова (відп. ред.) [та ін.]. – Дніпро, 2020. – Вип. 20. – С. 93 – 101.

4. Козін І. В., Максишко Н.К., **Селютін Є.К.** Використання еволюційних алгоритмів пошуку оптимальний класифікацій / Запорізький національний університет [Електронний ресурс]. – Режим доступу: http://visnykznu.org/issues/2019/2019-econ-2/15.pdf

*Статті у наукових періодичних виданнях Європейського Союзу з наукового напряму, з якого підготовлено дисертацію:*

5. Kozin I.V., **Selyutin E.K.,** Polyuga S.I. Jumping frog method for optimal classifications / International Academy Journal Web of Scholar. 2(52). doi: 10.31435/rsglobal_wos/30042021/7519

*Наукові праці, які засвідчують апробацію матеріалів дисертації:*

6. **Selyutin Ye.,** Kozin I. Features of metaheuristic methods / Журнал «Молодий вчений». – Київ, 2021. – № 2. – С. 109 – 113.

7. **Селютин Е.К.** Применение метода прыгающих лягушек для задачи размещения производства / Комбінаторні конфігурації та їхні застосування: Матеріали XXII Міжнародного науково-практичного семінару імені А.Я. Петренюка (Запоріжжя - Кропивницький, 15-16 травня 2020 року) / за ред. Г.П. Донця – Кропивницький: ПП «Ексклюзив-Систем», 2020. – С. 133 – 137.

8. **Selyutin Y.,** Kozin I. The Metaheuristic Application in Classification Problems / Інформаційні технології: теорія і практика: Тези доповідей III-ї Всеукраїнської науково-практичної інтернет-конференції здобувачів вищої освіти і молодих учених, 2020 р., м. Харків) [Електронний ресурс] / Редкол. : М. В. Новожилова, І.О.Яковлева, Г. Л. Козіна, Г.В. Бакурова, Т.А. Желдак. Електрон. дані. – Харків : ХНУМГ імені О.М.Бекетова, 2020. – С. 22 – 24.

9. **Селютин Е.К.** Применение метода прыгающих лягушек для поиска оптимальных классификаций / Математичне та програмне забезпечення інтелектуальних систем (МПЗІС-2020): Тези доповідей XVIII Міжнародної науково-практичної конференції, Дніпро, 18-20 листопада 2020 р. / Під загальною редакцією О.М. Кісельової. – Дніпро: ДНУ, 2020. – С. 227 – 229.