Algorithms for solving discrete optimization and machine learning problems

I. I. Eremin*, E. Kh. Gimadi†, A. V. Kelmanov‡, and M. Yu. Khachay§

*Institute of Mathematics and Mechanics, UB of RAS, ermii@imm.uran.ru
†Sobolev Institute of Mathematics, SB of RAS, gimadi@math.nsc.ru
‡Sobolev Institute of Mathematics, SB of RAS, kelm@math.nsc.ru
§Institute of Mathematics and Mechanics, UB of RAS, mkhachay@imm.uran.ru

This survey presents a brief scientific report of the joint research collective uniting the collaborators from the Institute of Mathematics and Mechanics, Ural Branch of RAS and from the Sobolev Institute of Mathematics, Siberian Branch of RAS working together as a team of the projects "Development of a Common Theory for the Problems of Reconstruction, Inversion and Control" (under the guidence of academitians Yu.S. Osipov and I.I. Eremin, Program for basic research, UB of RAS) and "Development and Analysis of Discrete Optimization Problems in Operations Research and Pattern Recognition" (under the guidence of prof. E.Kh. Gimadi, Target program, SB of RAS). Main aims of the paper are the following.

1) to characterize the investigation scope of the project,

2) to present the most valuable recent results obtained by the joint research team.

Mathematical models of operations research — continuous and discrete optimization problems, techniques of machine learning, pattern recognition and data mining are intensively used for solving the broad range of applied problems. Herewith, induced instances of optimization problems usually possess a whole range of the additional features — infeasibility, unboundedness of the target function, integrality constraints, unformalized setting, etc. taking into account of which, on the one hand, makes the solution process of the appropriate problems more difficult and stimulates advances in theory and development of specialized algorithms, on the other.

In convex optimization, the most intractable case of a problem setting is called *an improper problem*. In this case, classical duality properties are violated for some or another reason.

Traditionally, two main approaches for correcting the improper problems are used, continuous, wich is followed to fundamental results by P.L. Chebyshev, and discrete, based on constructing of collective generalized solutions. This project is involved to development of the later, discrete, approach, on the basis of the cycles of immobility of Fejér iterational processes and committee generalized solutions of infeasible systems of equations and inequalities.

In the field of pattern recognition and data mining our research is mostly related to studying of special classes of combinatorial optimization problems, induced by optimal techniques of learning in terms of structural risk minimization and maximization of likelihood.

The main feature of this project is systematic investigation of discrete optimization problems, including

- making a classification of the problems, assigning them to one or another known complexity classes, proving of completeness and schemes of reduction, describing the polynomial time solvable subclasses, etc.;

- development of exact polynomial time and

pseudo-polynomial time algorithms on the basis of different implicit search schemes;

- studying the approximation capability of the problems in question, development of polynomial time approximation algorithms with proven approximation guarantees and polynomial time approximation schemes (PTAS), obtaining lower bounds of approximation thresholds;

- constructing the asymptotically exact and effective probabilistic algorithms for intractable problems;

- development and numerical analysis of heuristics and meta-heuristics, e.g., based on the local search scheme.

Over several decades, an active research in optimization, pattern recognition and data mining was conducted in our institutes. Priority results in almost all above mentioned fields were obtained.

The best known results, obtained by members of the team, belong to the field of optimal correction of improper linear and convex programs, exact penalty function method, discrete generalized committee solutions for infeasible systems of constraints and collective piecewise-linear training algorithms of pattern recognition, theory and methods of discrete optimization — generalizations of traveling salesman problem (TSP), scheduling, two-level optimization, multi-index assignment problem, facility location problem, etc.

Along with theoretical research, several software development projects, using the developed algorithms and techniques, have been successfully completed. In the Sobolev Institute of Mathematics, SB of RAS, the following software have been developed:

- a benchmark library "Discrete location problems", containing description of the fundamental models, survey of publications, demo-versions of algorithms, benchmark problems generators, and optimal solutions for generated instances;

- an updatable Web-resource "QPSLab System for Analysis and Recognition of quasi-periodic sequences" representing the most recent results related to combinatorial optimization problems induced by problems of robust off-line analysis of structured data and recognition of structured signals.

In the Institute of Mathematics and Mechanics, UB of RAS, a computational website "Quasar-offline", devoted to collective algorithms of pattern recognition has been developed.

Recent results

Optimal discrete correction of optimization problems. New construction schemes for iterative Fejér algorithms with easily interpretable attractors is introduced, new necessary and sufficient conditions of existence of committee solutions (with given properties), and effective algorithms for their construction is obtained for systems of linear inequalities.

Pattern recognition and data mining. The problem of structural risk minimization in the class of committee piecewise-linear decision rules is proven to be equivalent to the known NP-hard combinatorial optimization problem "Minimum affine separating committee" (MASC). Also, it is proven, that the later problem remains intractable in fixed dimension spaces within $n > 1$ under the additional *general position* constraint on subsets which should be separated. A nontrivial polynomial time solvable subclass is described. It is shown, that the MASC problem does not belong to APX class, unless $P = NP$, an appropriate efficient approximation threshold is obtained.

An approximation algorithm with $O(\log m)$ approximation ratio, which is effectively uses the general position condition, is developed.

On the other hand, it is shown, that this problem is MAX-SNP-hard, which implies an absence of PTAS for it, unless $P = NP$. Occasionally, the same result is proved for the well-known Minimum (planar) Point Covering (Min-PC) problem.

Exact polynomial time algorithms (with best known performance guarantee) for combinatorial optimization problems, induced by problems of robust estimation of vector alphabet, generating the quasi-periodic sequences and robust recognition of sequence (as a structure), including a repeating ordered fragments set taken from the given vector alphabet, are developed.

It is proved that several actual clusterization problems for finite subsets of Euclidean space are NP-complete in the strong sense. A new 2-approximation algorithm is proposed for the NP-hard problem, to which some partition into two clusters (by min-sum criterion of squared distances) problem for a finite subset of the Euclidean space can be reduced. For the same problem with fixed dimension and additional integrality constraint for separated vectors, a new exact pseudo-polynomial time algorithm is developed.

Asymptotic exactness conditions are proved for some randomized search algorithm for subset (of a Euclidean space) of a given cardinality with maximum norm of sum (of the elements).

Routing problems. A performance ratio for approximation solution of maximum-weight multiple traveling salesman problem (MAX m-TSP) in multidimensional Euclidean space, for which all of m found Hamiltonian circuits should be edge-disjoint, is proved. Conditions on number m of salesman routes, for which the algorithm with cubic time-complexity is asymptotically exact, are obtained.

Improved performance ratios of approximation algorithms for solving MAX and MIN 2-TSP on complete undirected graph are obtained in cases of the common or different edge-weight functions of Hamiltonian routes, where all (non-negative) weights of edges can be arbitrary, metric (satisfy the triangle inequality), should belong to the segment $[1, q]$, or valued by $\{1, 2\}$.

For vehicle routing problems with restricted number k of clients in each route (kVRP and multi-depot kVRP), performance ratios of approximation algorithms and conditions of their asymptotic exactness on random initial data, are obtained.

Asymptotic exactness conditions for an effective approximation algorithm for MAX TSP, formulated in finite dimension normed space, are proved. New approximation algorithms with improved performance guarantees are developed for the MAX TSP in spaces with a polyhedral norm.

Location problems. An exact polynomial time algorithm for the problem with common restriction on values of supply, for which all users and facilities are located on the chain-like network, is developed. For an arbitrary network and some special assumptions on values of users demand and amount of facilities, a new approximation algorithm is constructed and probabilistically tested. Conditions, for which the algorithm is effective and asymptotically exact, are obtained.

The Multi-stage Facility Location Problem (multi-FLP) is NP-hard in general case even for one-stage FLP. For the two-stage FLP on a tree-like network the exact algorithm with time complexity $O(nm^3)$ is constructed (n is the number of consumers and m is the total number of all facilities).

For the competitive facility location problem, formulated as two-level integer program, an equivalent representation in terms of pseudo-boolean function maximization problem is found. For this problem an overestimating algorithm (on some solution subset given by a partial boolean vector) is introduced. Also, an underestimating approximation algorithm is developed.

For the discrete (r, p)-centroid problem a new exact iterative algorithm, based on some ideas of local search, is proposed. Dependence of the problem complexity on parameters r and p is investigated. It is shown that the pricing problem with factory-chosen prices belongs to logAPX class, and related estimation problem is nontrivial one in complexity class $\sum_2^p$ of polynomial hierarchy. Exact and approximation algorithms for this problem, using the problem decomposition and genetic local search concept, are developed. Also, an approximation algorithm for max-min facility location problem for 2 facilities on rectangular area

with l_∞-metric is proposed.

Transportation problems. It is proved that decentralized transportation problem is NP-hard even in the particular case of the decentralized Semi-Assignment problem with identical demands. Nevertheless, for the last problem within random initial data polynomial-time asymptotic exact algorithm is presented.

Improved conditions of asymptotic exactness for some polynomial time approximation algorithm for solving planar m-layers 3-index assignment problem with random inputs are established.

Feasibility conditions for multi-index axial assignment problem on mono-cyclic permutations with index count $m \leq 8$ are obtained.

A sequential and parallel algorithms of dynamic programming for quadratic assignment problem on a tree are proposed.

Packing problems and scheduling. 2-dimensional rectangular bin packing problem with tabu areas, which is a generalization of well-known NP-hard bin packing problem. Solution coding schemes for guillotine and non-guillotine types of the problem, taking into account the specific of tabu areas, are developed. A new simulation annealing approximation algorithm, constructed on the basis of these schemes, is proposed.

Conditions of asymptotic exactness of some approximation algorithm for multi-project scheduling problem with one restricted resource and random inputs are examined.

A partial integral linear programming model for the scheduling problem with continuous resource and time is constructed.

A dynamic programming algorithm for the case of the problem with integer-valued time is developed. The algorithm has a polynomial time complexity when the machine count is over-bounded by any fixed constant.

Solution methods for integer linear programming. Application of the unimodular transforms for improving the inner structure of the problem in question (particularly, one-dimensional knapsack problem) and speeding-up the algorithms is examined. New L-class enumeration algorithms for boolean linear programming, based on enumeration of vertex and essential L-classes of the relaxation polyhedron of the problem, are developed.

In conclusion it should be noted that all theoretical results are used during the solution of applied problems.

ACKNOWLEDGMENTS. This work was done under partial support of Program for basic research of UB of RAS, proj. no. 09-C-1-1010, and Target program of SB of RAS, proj. no. 44.