Journal of Engineering and Applied Sciences 13 (15): 6146-6150, 2018
ISSN: 1816-949X

Modification of a Cluster Analysis Problem for Build the Structure of a Distributed Information-Measuring System

S.P. Vorobyev, D.V. Shaykhutdinov, E.V. Kirievskiy and V.V. Grechikhin
Department of Information and Measurement Systems and Technologies,
Platov South Russia State Polytechnic University (NPI), Novocherkassk, Russia

Abstract: The study shows the necessity and features of the design of distributed information-measuring systems of precision multifunctional measuring complexes. The use of the ethernet family standard is the basic technical solution for building the information level. The formulation of a mathematical model of the structure optimization of a distributed information-measuring system which is reduced to a modified clustering problem is presented. It is proposed to use an approach based on a genetic algorithm to solve the problem. The research of the algorithm is performed by analyzing the modifications of the classical algorithm based on the variable mutation, under which the mutation probability varies depending on the needs of the algorithm. Results of numerical experiments showing the advantage of the proposed formulation of the problem for constructing the architecture of a distributed information-measuring system are presented.

Key words: Mathematical model, information-measuring system, cluster analysis, genetic algorithm, architecture, probability

INTRODUCTION

The need to design distributed information-measuring systems of precision multifunctional measuring complexes is caused by the complication of technological processes in production, monitoring of parameters under uncertainty, under the influence of destabilizing factors, the requirements for a high metrological control level. Currently, information-measurement systems are used in production and scientific research, however do not fully meet all the necessary technical characteristics: efficiency, a wide range of parameters under study, the ability to control a set of parameters to be determined, the required metrological level of measurement results. The problems of increasing resistance to destabilizing factors, functioning in conditions of uncertainty, the possibility of choosing the method of control and changing the structure of the information-measuring system in the process of monitoring and fabricating materials are also not solved in existing information-measuring systems.

The design of distributed information systems of precision multifunctional measuring complexes usually involves the implementation of a complex interconnected complex of software and hardware solutions of the information level and the measuring subsystem. The information part provides automated collection, transmission, storage, processing, visualization of information necessary to optimize management decisions over measurement and control processes in accordance with the accepted criterion, the development of the corresponding necessary control actions on the measurement object and measuring equipment, harmonious interaction of the operator and technical systems of different complexity level or transfer of this information for further processing to the following levels, as well as auxiliary functions that provide the solution of intrasystem and service problem. The measuring subsystem provides the implementation of direct, indirect, joint or aggregate measurements of physical quantities, transformation of measurement information with a specified accuracy, direct control of the measurement process and direct impact on the measurement object, evaluation and presentation of the residual uncertainty characteristics of the measured values, representation to the operator of measurement results and implementation of processes in the required form.

The use of the Ethernet family standard is the basic technical solution for building the information level. Currently, Ethernet offers a wide range of data transmission facilities, both using copper and optical fiber with data rates up to 10 GB/sec and above. This technology is very important not only for the creation of modern LAN but also for industrial automation. Along

Corresponding Author: S.P. Vorobyev, Department of Information and Measurement Systems and Technologies, Platov South Russia State Polytechnic University (NPI), Novocherkassk, Russia

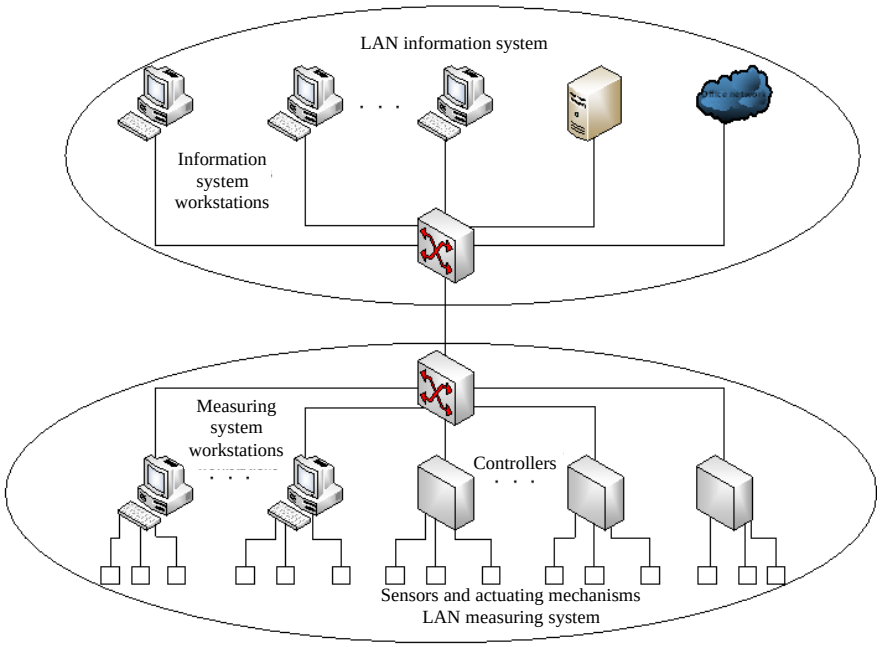

Fig. 1: The structure of the distributed information-measuring system

with the benefits of standardized communication, ethernet provides a seamless network architecture that connects both office equipment and production sites into a single unit. Given the above, ethernet is the most convenient environment for building a distributed precision multifunctional information-measurement system (Fig. 1).

MATERIALS AND METHODS

Mathematical statement of the optimization problem: In formalization, the problem of constructing the optimal structure of a distributed precision multifunctional information-measuring system can be reduced to the problem of cluster analysis in which it is necessary to divide a set of network objects (workstations and controllers) into subsets of information and measurement parts in accordance with the selected criterion.

The problem of non-uniform cluster analysis of the structure of a distributed information-measuring system can be most clearly visualized as a problem of integer programming. The source data is a set of network nodes $NU = \{NU_1, NU_2, ..., Nu_j, ..., NU_n\}$. Each object $NU_j \in NU$ is described by the characteristic s_1. In the general case, each cluster GL_v contains n_v objects in compliance with the general requirement:

$$n = \sum_{v=1} n_v$$

The cost of including NU_j in the cluster GL_v is functional. F_v (NU_j, GL_v, node type, cluster type, cluster characteristics, information flow characteristics) which depends on the type of cluster and which nodes make up the cluster. Variables are introduced that describe the inclusion of an object Nu_j in the cluster Gl_v x_{jv}: $x_{jv} = 1/x_{jv} = 0$ and define the leading element of the cluster y_j: $y_j = 1$ /$y_j = 0$. The criterion assumes minimization of the total amount of costs for all clusters:

$$\sum_{j=1}^{n}\sum_{v=1}^{m} F_v\begin{pmatrix} NU_j, GL_v, \text{node type, cluster characteristics,} \\ \text{information flow characteristics)} x_{jv} \rightarrow \min \end{pmatrix}$$

when constraints are executed: each object is included in only one cluster:

$$\sum_{v=1}^{m} x_{jv} = 1$$

and the number of clusters is fixed and equal to m.

$$\sum_{j=1}^{n} y_j = m$$

In addition, a functional is introduced that describes the distance function for different nodes:

$$d(NU_j, NU_r) = FR_{jr}(\text{node type, information flow characteristics})$$

and restrictions on the reaction time when traffic is transmitted within a cluster for a single node:

$$FT_v(NU_j, CL_v, x_{jv}, \text{node type, type of traffic}) \le VK_v,\ v = \overline{1, m}$$

The following distance functions can be used to solve the problem:

- Distance according to l_1
- Distance according to the supremum-norm
- Distance according to l_p-norm
- Mahalanobis distance
- The square of the Euclidean distance
- Distance between city blocks (Manhattan distance)
- Chebyshev distance
- Exponential distance
- Percentage of disagreement

It is also necessary to analyze the application of the following rules for combining or linking:

- Single connection (nearest neighbor method)
- Full connection (the method of the most remote neighbors)
- Unweighted pairwise means
- Weighted pairwise mean
- Unweighted centroid method
- Weighted centroid method (median)
- Ward method
- Two-way coupling
- K-means method (the main idea is to minimize the difference between cluster elements and maximize the distance between clusters)
- K-median method

Solution algorithm: Given the sufficiently large dimension of certain information-measuring systems, the process of their design requires the solution of NP-complete problem and therefore, it is advisable to use heuristic algorithms in particular from the family of genetic algorithms. The advantage of genetic algorithms is that they work with codes that represent a formalized form of a set of parameters that are arguments of the objective function. When implementing the search procedure, the genetic algorithm processes several points of the search space at the same time which allows overcoming the danger of falling into the local extremum of the polymodal

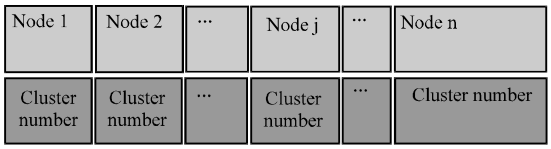

Fig. 2: Chromosome structure

objective function. Additional information is not used in the process of work. In the genetic algorithm are use both probabilistic and deterministic rules for generating new points of the search space simultaneously which gives a much greater effect (Back *et al*., 1997; Goodman and Kovalenko, 1966; Kureichik, 1988, 1998, 2002; Goldberg, 1989).

The main idea of the genetic algorithm is to create a population of individuals, each of which is represented as a chromosome. Any chromosome is a possible solution to the optimization problem under consideration. Only the value of the objective function (or fitness function) is used to find the best solutions. The value of the fitness function of the individual shows how well the individual described by this chromosome is suitable for solving the problem. The chromosome consists of a finite number of genes, representing the genotype of the object that is the totality of its hereditary traits. The process of evolutionary search is conducted only at the level of the genotype. The basic biological operators are applied to the population: crosses, mutations, inversions, etc. The population is constantly updated by generating new individuals and destroying old ones and each new population becomes better and depends only on the previous one.

The use of a genetic algorithm requires the development of a method for coding a solution. The structure of the chromosome (Fig. 2) which represents a string consisting of cluster numbers for each node is proposed to implement the algorithm. The chromosome is an array of pairs (node, cluster). The length of such an array will always be equal to the number of nodes that make up the computer network of the information-measuring system.

Each chromosome is evaluated by a measure of its "Fitness" (fitness-function). The most adapted individuals have a great opportunity to participate in the reproduction of offspring. Proportional selection is that each i-th chromosome has a probability P(i), equal to the ratio of its fitness to the total fitness of the population. The multipoint crossover in this case has a break point which is the boundary between adjacent elements of the array (that is the node is selected at random). The number of them will be <1 the number of genes in the chromosome or the number of clustered nodes. Preference is given to

the most adapted representative of the population when choosing from which parent the offspring will take the next gene. The node numbers for which cluster values change are selected randomly.

The genetic algorithm consists of the following steps. Determination of the initial population of n chromosomes, each of which represents a solution variant that is the points are randomly distributed across clusters.

Recalculation of the cluster center for each population, using the best chromosome from the current population. After receiving new centers, the algorithm continues to work until the stopping criterion is reached.

The chromosome (a possible solution) is obtained as a result of the application of genetic operators. During the investigation of the genetic algorithm, a number of approaches and modifications of the classical algorithm were used to solve the problem of choosing the optimal variant in distributed information-measuring systems which implied a variable mutation in which the mutation probability varies depending on the needs of the algorithm.

RESULTS AND DISCUSSION

The numerical experiment was performed for different values of the number of network nodes nu and the number of clusters cl. Experimental research of solving the problem of non-uniform clustering (Fig. 3) showed the advantage of the proposed formulation of the problem to build the architecture of a distributed information-measuring system. The cost of including a single node in a cluster is a function of the type of cluster and its composition.

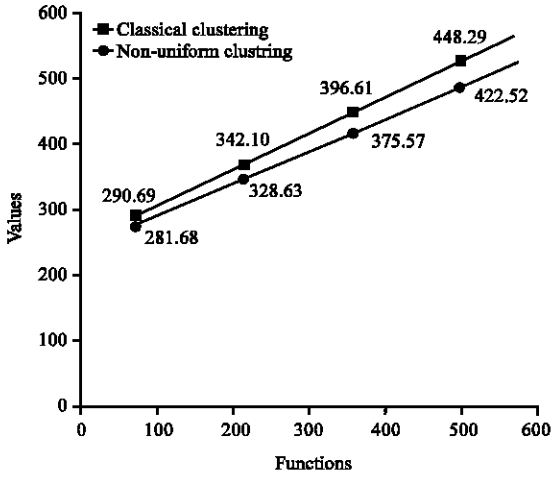

Fig. 3: Graph of the dependence value optimized function on the number of nodes and clusters

The modification of the cluster analysis problem by means of a genetic algorithm allows one to find the best feasible or optimal (from the point of view of the constraints in the problem) solution than the usual classical clustering problem. This advantage is shown as a result of the experiments performed on the test examples.

CONCLUSION

The distributed information system of the precision multifunctional measuring complex is a seamless connection of two LAN the information part and the measuring part, constructed on the basis of fast ethernet or gigabit ethernet and industrial ethernet, respectively.

Optimization of the architecture of the measuring complex can be performed within the framework of a multi-level topological structure of the computer network (Vorobyev, 2009, 2016). The difference from the classical problem of cluster analysis is that the cost of including a single node in the cluster depends on the cluster type, the composition of the cluster and additional conditions that determine the structure of information flows in the cluster.

ACKNOWLEDGEMENTS

The study results are obtained with the support of the project #2.7193.2017/8.9 "Development of scientific bases of design, identification and diagnosis systems for highly accurate positioning with application of the methodology of inverse problems of electrical engineering", carried out within the framework of the base part of state job.

REFERENCES

Back, T. D.B. Fogel and Z. Mi-chalewuz, 1997. Handbook of Evolutionary Computation. IOP Publishing, Bristol, England, UK., Pages: 1130.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Pub. Co., New York, USA., ISBN: 9780201157673, Pages: 412.

Goodman E.D. and A.P. Kovalenko, 1966. Evolutionary calculations and genetic algorithms. Surv. Appl. Ind. Math., 3: 760-760.

Kureichik, V.M., 1988. Genetic Algorithms. TRTU Publishing, Taganrog, Russia, Pages: 242.

Kureichik, V.M., 1998. Methods of Genetic Search: A Tutorial. TRTU Publishing, Taganrog, Russia, Pages: 118.

Kureichik, V.M., 2002. Genetic Algorithms and their Application. 2nd Edn., TRTU Publishing House, Taganrog, Pages: 242.

Vorobyev, S.P., 2009. Possible directions of using the concept of multilevel topology and optimization of distributed corporate systems. Issues Mod. Sci. Pract., 8: 131-143.

Vorobyev, S.P., 2016. The mathematical model of building a multi-level topology of computer network for distributed corporate system based on the inverse problem. J. Eng. Appl. Sci., 11: 1243-1247.