BIG DATA clustering techniques

Vitaliy Vladimirovich Kurdyukov
L.N. Gumilyov Eurasian National University, 010008, Republic of Kazakhstan, Astana, Satpayev street, 2

Abstract
In the article methods of clustering are investigated and a model of big data is created. Developed Multi-extremal models are constructed in which it is necessary to determine the optimal values of more than one objective function, with or without constraint systems. The scope of obtained results is a wide range of applications of discrete programming block-symmetric tasks where an analysis of large volumes of data is required. Efficiency and accuracy of the proposed models and techniques are supported by comparative research, approbation and introduction.

Keywords: clustering, big data, method, data mining, discrete programming, data model.

1 Introduction
At the moment, in the modern information society in which we now live, the amount of information that requires analysis, processing, systematization, clustering, classification is constantly growing.

A new concept “Big Data” has appeared.

Big Data is a term meaning a data set, so volumetric and complicated that the use of traditional database control tools and applications for their processing becomes impossible. Collection, cleaning, storage, search, assess, transfer, analysis and visualization of such sets as a whole, not local fragments, make up a problem [1]. The big data is an important step of humanity in the constant striving to quantify and comprehend the world around us. What used to be impossible to measure, store, analyze and distribute finds its expression in the form of data. Use of huge assets and selection of quantity give a way to new capacities of the world. The term Big Data is borrowed by analysts from scientific publications. Big Data refers to terms with definite date of birth “September the third, year two thousand and eight”, when special issue of British magazine “Nature”, dedicated to the search of answer to the question “How technologies opening potential of work with big data sets can affect the future of science”. It includes not only a quantitative index, but also a set of techniques to work with huge data sets. Big data in information technologies is an aggregate of approaches, tools and techniques of processing of structured and unstructured data of huge volumes to obtain human-readable results, effective under conditions of non-stop growth and distribution by multiple units of a computer network. This series includes means of big data massively parallel processing, mainly, by NoSQL decisions, MapReduce algorithms, programming frameworks and Hadoop project libraries.

Three “V” are specified as critical characteristics for big data:
1. Volume (value of physical volume),
2. Velocity (velocity of both growth gain and necessity of processing with the highest speed and obtaining results).
3. Variety (variety as possibility of simultaneous processing of various types of structured and semi-structured data).

The predominant part of traditional data sources is structured. That means that traditional sources represent data in a distinctly predefined format. Sources of unstructured data, video and audio, can be hardly controlled. Image implies the format in which individual pixels are arranged in lines but their mutual arrangement determining what the viewer sees significantly differs in each individual case. Big volumes of information, even those stored (generated) by business production, exceed quantitative limits anticipating qualitative changes and enable to find new regularities, never before caught in small collected data volumes [2].

2 Methodology
Big data analysis and processing approaches and tools are based on proprietary methods, models and algorithms of tasks solution. They include tasks of data clustering, regression and time analysis, parallelising, merging of analytic environment with storage environment, and series of other tasks, specific for big data. High dimensionality of tasks solved at each design stage determines the necessity of research and development of new approaches, models, techniques and algorithms of big data processing system design. One of new directions of task setting and solution for effective design of big data processing system includes block-symmetric models and techniques to solve high dimensionality tasks [3]. Development of these techniques is a very actual problem.

Technique is a systematized aggregate of
steps/actions targeted to the solution of a definite task or achievement of a specific aim. Techniques develop in accordance with time, achievements of technical and scientific thought, and demands of the society. Development of techniques is a natural consequence of scientific thought development [10].

Clustering is partitioning of object sets into groups (clusters) based on properties of such objects [10]. Lack of any differences both among variables and records is characteristic for a clustering stage. Groups of the closest, similar records are searched. Advantage of the cluster analysis is that it enables to carry out grouping of objects not by one parameter but by the whole set of parameters.

Cluster analysis is a multivariate statistic procedure executing collection of data containing information of object selection and regulating objects in comparatively homogeneous groups [10]. Cluster is a group of objects with similar features. The purpose of clustering algorithms is creation of classes which are maximally connected inside but different from each other. This is why there two features of a cluster: intrinsic homogeneity - documents inside one cluster shall be maximally similar to each other; external isolation - documents of one cluster shall be minimally similar to documents from another cluster. Clustering is an example of the strategy “studying without a teacher”, i.e. adding a definite document to a particular cluster is automatic, without participation of a human expert [10]. Cluster content in clustering is determined only by distribution and structure of data. In clustering such parameters are not preset and aggregation is carried out automatically [8].

The task of discrete programming is the following [9]. The task of discrete programming (DP) is a task of searching extremum (max, min) of scalar function set on a discrete (disconnected) set, i.e. the task of mathematical programming (MP) with discreteness requirement imposed on all or part of variables determining feasibility region. The class of combinatorial models segregated in the process of discrete programming theory development [10]. Extremum of integral function set on the finite set of elements or elements of finite set providing extremum of objective function shall be determined.

Rearrangement minimizing the value of objective function shall be identified in a combinatorial set. Setting various combinatorial tasks can be often stated in the form a model with Boolean variables, taking only two values: 0 or 1. Large number of applied tasks is reduced to Boolean models that certifies the perspective of the model of this class. Methods of solutions of discrete programming tasks are often connected with their mathematical setting and peculiarities.

A typical example of combinatory technique is branch and bounds method [10]. This method includes directed search of admissible solutions on the basis of estimation. Researchers are especially interested in multi extremal models where optimal values of more than one function shall be determined with (without) constraint systems. As a rule, models of this class are complex in computation. In addition setting of the series of applied tasks reduces to the models of this class. Solution of the said tasks is very actual [13,14].

General formulation and solution of block-symmetric discrete programming tasks. Suppose that the set of objects $A = \{a_i: i = 1,1\}$. A set of objects $B = \{b_j: j = 1,1\}$ with elements of different types are given, as well as interrelations among elements of these sets identified by matrix $W = [w_{ij}], i = 1,1; j = 1,1$, elements of which are integral or Boolean.

Elements of set $A$ shall be distributed to n-group $A_n = 1,1$, and elements of subset $B$ in disjoint subsets $B_m = 1,1,1,1$ thus, to achieve extremum of the objective function $F(A,n)$. The following variables are introduced for formalized problem statement. Let, $X = [x_m], i = 1,1; n = 1,1,1$ be a Boolean matrix, where $x_m = 1$, if i-element is distributed to m-group, $x_m = 0$, otherwise. Similarly, $Y = [y_m], j = 1,1,1,1; m = 1,1,1,1,1$ be a Boolean matrix, where $y_m = 1$, if j-element is distributed to m-group, $y_m = 0$, otherwise. In general case matrixes of variables $X$ and $Y$ can be integral-valued. Function $F(X,Y)$ depending on distribution of $A$ and $B$ set elements into subsets $A_n$ and $B_m$ is identified on the set $A \times B$. Correspondingly, on the set $A$ – functions $\varphi(X), k = 1,1,1,1$ and on the set $B$ – functions $\psi(Y), s = 1,1,1,1,1$, identifying constraints on the sets $A,n$ and $B_m$.

Block-symmetric tasks of discrete programming are formulated the following way:

$F(X,Y) \rightarrow \text{extr}$

with constraints

$\varphi(X) \leq \varphi(X), k = 1,1,1,1 \quad (2)$

$\psi(Y) \leq \psi(Y), s = 1,1,1,1,1 \quad (3)$

In constraint sets (2) and (3) inequality signs may change into opposite depending on the task statement. In general case double-index matrixes of variables $X$ and $Y$ and given matrix $W$ can be integral-valued. Consider the task under the condition that the variables $X,Y$ and $W$ are Boolean matrices. Function $F(Z)$ is often used as function $F(X,Y)$ where $Z = XY$.

There is a formula (4), representing matrix composition of variables $X$ and $Y$ and the given matrix $W$ where objective function is identified. Unlike traditional task statements of discrete programming this formula has two types of variables $X$ and $Y$, and variables $X$ and $Y$ are symmetric to the given matrix $W$. In task (1) – (3) the set of constraints (2) depending on variable $X$ and set of constraints (3) depending on variables $Y$ can be identified.

Functional of type $F(X,Y)$ can be executed the following way:

$F(p(X),g(Y)) \rightarrow \text{extr}$

$p(X) \rightarrow \text{extr}$

$\varphi(X) \leq \varphi(X), k = 1,1,1,1 \quad (7)$

$g(Y) \rightarrow \text{extr}$

$\psi(Y) \leq \psi(Y), s = 1,1,1,1,1 \quad (9)$

In statement of the task (5) – (9) there is a unit of functions (6) – (7), depending only on variable $X$ and unit of functions (8) – (9), depending on the variable $Y$, integrated by common functional of the type (5). In a series of task settings constraint unit can be

$f(X,Y) \leq f(x,y), r = 1,1,1,1 \quad (10)$

which depends on variables $X$ and $Y$. In this case
functional unit of the type (5) and constraints of the type (10) can be determined. Thus, task (5) – (10) is called block-symmetric task of discrete programming. To solve the block-symmetric task of discrete programming providing \(X, Y\) and \(W\), are Boolean matrixes effective task solution system has been developed and proposed. Solution search pattern consists of the following stages: 

1. Boolean matrix \(W\) is established in submatrix \(Z = \|z_{mn}\|, n = 1, N; m = 1, M\) determined as task solution basis.

2. Matrix \(D = \|d_{ij}\|, i^l = \pi + 1; j = \lambda\); \(n = 1, N; j = \lambda\) is determined as direction of \(X\) solution search through logical addition of nonbasis lines of matrix \(W\) with basis lines and evaluation values are calculated by basis positions only. 3. Elements of set \(A\) are distribution by subsets \(A_n\). As a result solution \(X\) and intermediate matrix \(d_{ij}\) with \(\|\pi_{ij}\|, n = 1, N; j = \lambda\) are fixed.

4. Matrix \(D = \|d_{ij}\|, i^l = m + 1; j = 1, M\) of direction of \(Y\) solution search is determined though logical addition of nonbasis columns of intermediate matrix \(\Pi = \|\pi_{ij}\|\) to basis columns and evaluation values are calculated by matrix \(\Pi\) basis positions.

5. In accordance with the obtained values of matrix \(\Pi\) elements of set \(B\) are distributed by subsets \(B_n\). As a result solution \(Y\) and target matrix \(Z\), where value of objective function \(F(Z)\) is defined, are fixed.

It shall be noted that task solution search can be carried out both in forward direction according to the scheme \(DXDY\), and in reverse direction under the scheme \(DYDX\). Thus, basic model of a new class of block-symmetric tasks of discrete programming has been developed.

3 Results

Let us consider a numerical example of the solution of the task. The set of objects shall be divided by three classes and set of features by five classes in a way minimizing total number of interconnections among the obtained classes. Table 1 provides the initial matrix \(W\) with dedicated basis \(B\) in the upper left corner of initial matrix. Basis includes 1, 2, 3, 4, and 5 lines of matrix \(W\). Picture 1 demonstrates the process of \(X\) solution formation with the use of developed algorithm. Matrix \(D = \|d_{ij}\|\) is determined with the use of ratio.

\[
\begin{array}{ccccccccccccc}
\hline
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 \\
\hline
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
3 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
5 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
6 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
7 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
8 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
\hline
\end{array}
\]

**Table 1: Matrix \(W\) with the dedicated basis \(B\)**

**Picture 1. Process of \(X\) solution formation with the use of iterative display algorithm**

Display process is given in the table providing number of iteration \(k\); minimal element of \(D\) in accordance with which object number \(r\) is reflected in number of class \(\nu\). Picture 1 on the left provides matrixes \(X\) and \(\Pi\), content of which is determined by the basis of solution \(X\) search, and on the right - matrixes \(X\) and \(\Pi\), obtained with the use of iterative display algorithm. Matrix \(\Pi\) corresponds to a matrix of objective function \(Z\), reflecting interaction between the object and feature sets. Optimal value of objective function obtained at this basis and constraints is equal to \(\sigma(\Pi) = \sigma(Z) = 6\).
The developed models, techniques and algorithms are targeted to solution of big data tasks.

4 Discussion

Clustering task was first viewed in the 1930s. This problem has been studied in various aspects by many researchers, including: Lance U., William D., Hartigan D., McQueen D., Wong M. Kohonen T., Fritkze B., Kolmogorov A.N., Zagonuko N.G., Yuelkina V.N., Mkhitaryan V.S., Shchumskiy S.A. and others. Cluster analysis is a multidimensional statistical procedure executing collection of data containing information of object search and arranging objects in comparatively homogeneous groups [15,16,17,18]. Clustering task refers to statistic processing and a wider class of tasks for studying without a teacher. Clustering enables to identify definite regularities and trends in the whole aggregate of objects and their properties. Development of simple and fast clustering methods independent from parameters values of which can be rarely known a priori is very actual in solution of practical tasks in the field of social and economic applications when accuracy of obtained cluster decisions is of crucial importance. Types of clustering differ from application to application, but several common types can be identified.

5 Conclusion

The following results were obtained in the course of work:
1. Analysis of discrete programming methods as a design tool for big data processing systems was carried out.
2. The statement for design of modular flowcharts of data processing formalized as a block-symmetric task of discrete programming was developed and proposed.
3. Algorithm of task of automated design of modular flowcharts for big data processing was developed.
4. Computing experiments for assessment of task solution algorithm efficiency were carried out.
5. Visualization of graphs for computing experiments is performed and a model is created.

Any applied functional task or system can be represented in the form of the set of data processing procedures and information elements connected with them. In this case the task is to group data processing procedures by program modules and information elements by database sets with integration of connections between the modules and sets of the database. The obtained modular chart enables to encode the program of applied task solution by individual modules with further integration of individual programs of modules into one system. Big Data clustering is a result of technical revolution. The necessity of analytical work with big data significantly changed the face of IT industry and stimulates development of new software and hardware platforms. A new class of tasks, block-symmetrical tasks of discrete programming differing from traditional statements in properties: various types of variables, blocky structure and symmetry, has been developed and proposed. For the purpose of results visualization, tag clouds and the newest Clustergram, History Flow and Spatial Information Flow are used. Actual value of Big Data is in the use of big volumes of data for segmentation and clustering for effective building of large number of cluster models.

References
