ISSN 2079-0570, Advances in Gerontology, 2014, Vol. 4, No. 4, pp. 296–301. © Pleiades Publishing, Ltd., 2014.
Original Russian Text © A.I. Michalski, 2014, published in Uspekhi Gerontologii, 2014, Vol. 27, No. 2, pp. 321–327.

Aspects for Implementation of Data Mining in Gerontology and Geriatrics

A. I. Michalski

Institute of Control Sciences, Russian Academy of Sciences, ul. Profsoyuznaya 65, Moscow, 117997 Russia
e-mail: ipuran@yandex.ru

Abstract—Current challenges facing the theory and practice in aging sciences require the use and development of new methods of investigation of observational and experimental data. This is associated with both the extensive development of the measuring and experimental basis of biological research and the progress in information support of studies in aging. As a result, large databases containing information on the state of health of vast groups of people who survived to an advanced age have been created. The combination of achievements in these directions make it possible to apply data mining methods that are successfully used for solving intricate tasks in economics, medical diagnostics, organization of the Internet, and other fields of science and technology for solving tasks in gerontology and geriatrics. This review provides some examples of the use of data mining methods in gerontology.

Keywords: gerontology, data mining, classification, dependence reconstruction, heterogeneity, factors and genetic determinants of longevity, health condition at advanced age

DOI: 10.1134/S207905701404016X

Data analysis, often referred to in English literature as *data mining*, is an actively developing field of application of methods used in machine learning, pattern recognition, dependence reconstruction, and artificial intelligence. These fields strongly overlap and complement each other in solving specific practical problems. The combination of data mining and modern methods of efficient automatic data processing ensured the penetration of purely theoretical and algorithmic methods into wide practice. These methods are used in the design of airplanes and rockets, in the development of "smart" energy systems that automatically respond to failures and overloads to maintain stability and performance, in the development and mining of minerals, and in medicine and biology [7]. The most striking example of using data mining methods in medicine is computed tomography, which makes it possible to restore the spatial structure of an object by its flat projections.

Currently, the development of the technological basis of biological experiments, the creation of extensive databases of experimental observations of model organisms, and the results of large-scale studies of human health have created prerequisites for the application of data mining methods in gerontological research. The mathematical modeling methods used in gerontology [4] will be enhanced by the data mining methods. With the aid of data mining methods, information about the aging process of different organisms can be reflected at different levels of organization, including the molecular one, which, in turn, will help integrate heterogeneous experimental data on aging and the lifespan of different organisms.

This article considers examples of using modern data mining methods to study the survival of nematodes *C. elegans* in the presence of heterogeneity as well as the uniformity of their aging. The assessment of the age of *Drosophila* embryo development by gene expression and the search for genetic determinants of longevity in yeast *Saccharomyces cerevisiae* is described. Examples of assessment of the effects of adverse external factors on the lifespan of *C. elegans* and *Drosophila*, genetic factors on the lifespan of mice, and genetic polymorphism on the state of the human body in old age are given. In addition, examples of predicting the lifespan of mice by indices measured in middle age and prediction of the life expectancy of centenarians on the basis of data on the health condition and mental and social wellbeing are considered. An example of searching for relationships between different diseases and cancer incidence in humans is described. All these examples illustrate the effectiveness of the introduction of data mining methods into the studies of patterns of aging, longevity, and active aging.

EXAMPLES OF APPLICATION OF DATA MINING METHODS IN GERONTOLOGY

Analysis of Survival in the Presence of Heterogeneity

An example of effective use of the idea of taking into account the effect of unobservable factors on sur-

vival is the analysis of survival of nematodes *C. elegans* after thermal exposure [13]. Nematodes of strain TJ1060 (*spe*-9; *fer*-15) were grown for three days on a rigid substrate at a temperature of 25.5°C. On day 4 of life, the worms were divided into ten groups, which were kept for different times at a temperature of 35°C. One group of worms, which were not heated, served as the control. Each group was comprised of 100 to 200 worms. Starting from day 5 of life, the number of surviving and dead worms in each group was determined.

To explain the experimentally observed survival curves, the hypothesis was that the worms represented a heterogenous population in terms of stress tolerance; during development, individual stress tolerance is maintained at a high level until the end of the reproductive period and then drops sharply. It was shown that heating in the early period of life led to the redistribution of the organisms in the population between the groups "weak," "normal," and "stable" [20] and that these differences are manifested at moderate heating. From the biological standpoint, this phenomenon can be explained by the hypothesis that stress exposure leads to activation of various antistress systems (production of heat shock proteins, SOD, catalase, and DNA repair enzymes), the effectiveness of which differs at the individual level. The balance between the agents that damage the cellular structure as a result of thermal stress and the defense system products is expressed in reduced mortality and redistribution between the groups of organisms in the population. The steady-state point that is responsible for this balance shifts with age towards smaller effectiveness of antistress protection, which entails a sharp increase in mortality in the postreproductive period [2].

Assessment of the Uniformity of the Aging Process

An unexpected application of data mining methods for solving gerontological problems is described in [16]. In gerontology, some scientists believe that the aging process is associated with random accumulation of damages in the body as a result of generation of reactive oxygen species (ROS), whereas others point to a natural, programmed course of the aging process [1]. The prospects and strategy of influence and management of the aging process depend on what point of view is closer to the truth.

To obtain additional quantitative evidence in favor of one point of view or another, the authors investigated changes in the morphological structure of the muscle tissue of the nematode *C. elegans* with age. For this purpose, in the age range from 0 to 12 days, every two days a section of muscle tissue was photographed and then analyzed by the formal mathematical methods for image structure assessment. At each age, 50 animals were examined. The calculations were performed using Haralik's cooccurrence matrix and Tamura's directionality methods. The entropy of the cooccurrence matrix characterizes the amount of information contained in the distribution of pairs of pixels in the image, located at different distances and having set levels of gray, and the directionality characteristics quantitatively indicate the degree of orientation of the image. The results of analysis of the muscle tissue of nematodes of different ages showed that its structure changed with age from structured to chaotic and that these changes were nonmonotonic in time. The greatest changes were observed on days 2–4 and 8–10 of life. This applies to both the magnitude of the cooccurrence matrix entropy and the degree of image orientation. The obtained data support the fact that the aging of the nematode *C. elegans* is not the result of accumulation of stochastic defects but obeys the development program. This conclusion is also confirmed by the results of analysis of gene expression in aging nematodes [9].

Assessment of the Age of Drosophila Embryo Development

The modern method of restoring regression dependences, which is based on the support vector regression (SVR), was used to assess the age of *Drosophila* embryos by gene expression [14]. Using the data from a training set, which consisted of 103 measurements of gene expression and measurements of embryo age by the degree of membrane invagination, the regression dependence was built using polynomial kernels of the first and second degree. The quality of the dependence built was tested by cross-validation. The method consists in sequential removal of one observation from the training, construction of regression dependence on the basis of the remaining data, and calculation of the error made by the constructed dependences by remote observation. The criterion characterizing the accuracy of the regression construction method is the sum of errors made by the dependences built from remote observations. The same criterion was used for adjusting additional parameters of the model.

The calculation results showed that the support vector regression makes it possible to use gene expression data to predict the age of a *Drosophila* embryo with an accuracy of 2 min, although the duration of embryo development is 20–60 min. This result is improved if factors (i.e., linear combinations of expressions determined by factor analysis methods) are used instead of gene expression. For this purpose, sets of genes with poorly correlated expression are selected, which make the major contribution to the reduction of the prediction error of embryo developmental age. The authors showed that, in that study, the number of such very significant factors was only three.

Search for Genetic Determinants of Longevity

The search for the genes affecting longevity with the use of data mining methods is described in [12]. In

addition to the elucidation of the molecular mechanisms of aging, the identification of such genes indicates potential targets of therapeutic treatment of diseases of the elderly, such as cancer, cardiovascular and neurodegenerative diseases, and diabetes. This study was performed with the yeast *Saccharomyces cerevisiae*.

In that paper, the algorithm for constructing the shortest path network using protein interaction data was used. A network consisting of 171 genes was constructed, including 33 of 40 genes known from the published data as the genes associated with longevity. The network also included 45 genes essential for sustaining life. To confirm the results of the study, the potential of replicative senescence of 88 yeast strains with deletion of one of the genes that were components of the constructed network was determined. These strains contained significantly more mutations that affected the replicative lifespan than 564 strains with the deletion of a randomly selected gene. Additionally, genes whose association with the lifespan was not known earlier were identified, which, in the framework of the study, affected the lifespan in response to nutrition deficiency.

The author of [17], to study the effect of genetic manipulation on the lifespan of mice, used an accelerated ~~test~~ model, which is widely used in calculations of the reliability of technical systems. Data on the lifespan of mice subjected to genetic manipulation from 16 published studies were used. Genetic manipulation included knockout mutations and transgenic models with increased expression of specific genes. Each study included experimental and control cohorts of mice. The number of animals in the groups varied from 20 to 190; in some experiments discriminating by gender was not performed.

Weibull, Gompertz, log-logistic, and log-normal models were considered as basic survival models. A particular model was selected using the Akaike information criterion, which is popular in data mining [5]. By this criterion, it was found that, in 15 cases, the most appropriate model was Weibull's model, and in only one study was it the log-normal model. The Akaike information criterion was also used to select covariates, such as gender, date of birth, etc., included in the model in each study.

The study showed that the majority of genetic manipulations had a multiplicative effect on the survival rate of mice, which was independent of age and was described well by the deceleration factor in the accelerated test model. In addition, the change of scale is a more natural, intuitive measure of the influence on the lifespan than the ratio of mortalities in the proportional hazard model. The accelerated test model was more stable than the proportional hazard model in terms of deviations from the theoretical assumptions about the ~~basal~~ model.

The results of calculations showed that, among the considered genetic manipulations, mutations in the *Prop1* gene had the greatest effect on the lifespan of mice. The product of this gene is involved in pituitary gland stimulation. These mutations, compared to the control cohort, caused deceleration equal to 1.48 ~~at~~ a 95% confidence interval (1.34, 1.63).

Assessment of the Influence of Factors and Exposures on Lifespan

The effects of exposure to various adverse factors on the lifespan are described in extensive literature. The effect of reduced mortality of the nematode *C. elegans* in the postreproductive period as a result of short-term heating at the beginning of life is described in [13]. In the same study, a mathematical model explaining the effects of exposure at the beginning of life on the mortality at the end of life is proposed. The authors of [15] described the delayed effect of an increased lifespan of *Drosophila melanogaster* as a result of nonlethal heating in early life. To elucidate the genetic basis of this delayed effect, changes in gene expression in the experimental and control groups 10–51 days after the last exposure were studied. To assess the effect of heating, the methods of constructing linear models, Bayesian methods of analysis, and hierarchical clustering, which are used in data mining, were applied. It was found that the expression of the *hsp70* gene in *Drosophila* of the experimental group changed by a factor of 1.7–2. The results of clustering showed that the flies subjected to thermal stress were grouped into clusters according to age, which confirms the effect of events in the early period of life on the subsequent life of the organism at the genetic level.

Identification of Genes that Affect the Organism Condition in Old Age

The condition of the human body in old age, except the increased morbidity and mortality compared to the middle age, is also characterized by a decrease in a number of physical, functional, and cognitive abilities [3]. According to many authors, one of the main causes underlying these changes at the molecular level is the imbalance between the damages caused by ROS and the ability of the body to defend against this impact and eliminate the damages. The authors of [10] studied the association of 38 genes involved in the proantioxidant pathways with physical and cognitive abilities of people older than 90 years. The assessed parameters were the hand grip force, the ability to perform actions required for everyday life, the speed of movement on feet, and cognitive abilities, which were assessed by the results of implementation of a number of tests and exercises. Additionally, the influence of genes involved in the proantioxidant pathways on the life expectancy was studied.

The correlation between genes and the human condition in old age was studied using linear and logistic regression models, which are conventionally used in

data mining, as well as a new approach of constructing a robust procedure for selecting significant genes [11]. The correlation between the selected genes and the life expectancy was assessed using the Cox regression, which is widely used in survival studies.

As a result of analysis of data obtained for 311 men and 769 women older than 90 years, 14 genes involved in the proantioxidant pathways, associated with at least one of the studied parameters of the human condition in old age, were identified. The probability of the absence of correlation in this case was limited to 0.05. The hand grip strength was associated with the *UCP3* and *NDUFS1* genes, and the speed of movement on legs was associated with the *GCLC, UCP2, UCP3*, and *NDUFS1* genes. In the cohort studied, a positive statistical correlation between the cognitive abilities and the *NDUFV1, MT1AI*, and *GSTP1* genes was found. This correlation was less strong in the case of the *UQCRFS1* gene. The most statistically significant correlation with the ability to perform activities required for daily life was found for the *TXNRD* gene. None of the genes listed showed a statistically significant correlation with survival. However, the analysis revealed three genes statistically associated with longevity in the studied cohort of humans; the *LOX* gene had a positive effect on longevity, whereas the effect of the *SOD2* and *UQCRFS1* genes was negative.

Life Expectancy Prediction

An important direction in the application of modern data mining methods in gerontology is the prediction of life expectancy and the search for its determinants. The authors of [19] described the results of the study of life expectancy of centenary inhabitants of Rome by using neural networks. The study involved 110 people. On the basis of a questionnaire that formed 100 indices reflecting the state of health, as well as mental and social wellbeing, using the conventional geriatric scales, a three-level neural network without feedback was built. The aim of the study was to determine the factors that best predicted the chances of survival to the next year. The most accurate results were obtained by using 23 variables that reflected the comorbidity structure, risk factors for cardiovascular diseases, cognitive state, mood, functional status, and social relations. Among the selected factors, the presence of social relations was most important for maintaining longevity at older ages, even with a high degree of disability.

The authors of [18] compared different classification methods for predicting the lifespan of mice. The aim of the study was to demonstrate the possibility of predicting an individual lifespan on the basis of indices measured in middle age and to compare the effectiveness of solving this problem with the aid of 22 different machine learning algorithms.

The experiment was performed with a genetically heterogeneous mouse population obtained by crossing CB6F1 females and C3D2F1 males. A total of 1188 mice were studied, of which 403 were virgin males, 457 were virgin females, and 299 were nonvirgin females. After the removal of those who did not die of old age or did not survive the age of 2 years, 741 mice remained alive. After the experiment, each mouse was classified into one of four classes by the lifetime quartile. This parameter was selected because many of the considered machine learning algorithms were developed to solve classification problems and, as a result of their operation, formed the trait of class. Based on the results of previous studies and published data, blood T-cell parameters (measured at the age of 8 and 18 months), the level of serum hormones in blood (measured at the age of 4 and 15 months), body weight (measured at the age of 8 and 18 months), and cataract parameters (measured at the age of 8 and 24 months) were chosen as lifespan-associated traits.

In the study, 22 advanced machine learning algorithms belonging to a broad range of methods from the linear regression, naive Bayes method, and neural networks to methods of random trees and support vector machine were used. As a result, it was shown that the body weight of mice and the set of characteristics of Tcells at an age before 2 years allow predicting the quartile into which an individual mouse lifespan falls with an accuracy of 35.3% (±0.10%). This study shows that the combination of data mining methods and the biological approach helps improve the accuracy of mouse lifespan prediction and should contribute to progress in the development of multidisciplinary approaches in biogerontology and study of mammalian aging.

Comorbidity Study

The statistical correlation between various diseases and cancer was considered in [6]. Nonpersonal data on diseases affecting humans at the end of life and the causes of death, collected by the National Center for Health Statistics in the United States in 1980 among people older than 64 years, were analyzed. The methodology consisted in identifying the diseases in which the distribution of persons who died from cancer of a specific localization differed most significantly from the distribution of people who died from other causes. For this purpose, the methods of assessing the uniform deviation of the Kullback–Leibler divergence, which characterizes the divergence of the distribution of random variables from its empirical analogue calculated with the use of histograms, was used.

The analysis of the three most common forms of cancer (cancer of the digestive organs and peritoneum, respiratory organs and chest, and urogenital system) has shown that cardiovascular diseases are significantly associated with the group of subjects who died from diseases other than cancer. This is associated

with a high rate of mortality from cardiovascular diseases in older age, and they act as a competing risk of death in relation to cancer. As a result of analysis, diseases serving as the leading risk factors for cancer have been distinguished. For cancer of the digestive organs and peritoneum, these factors are represented by the class "other diseases of the digestive system;" for respiratory system cancer, by the class "respiratory diseases;" for urogenital system cancer, by the class "diseases of genitourinary organs." These findings point to the importance of early treatment of these diseases for the prevention of oncology.

CONCLUSIONS

The examples described above demonstrate the practical value of the use of data mining methods for solving problems of gerontology and geriatrics. These methods make it possible to compile statistical data, formalize the procedures for constructing meaningful conclusions, and verify the statistical significance of findings. The entire range of modern data mining methods—from classification methods to nonparametric predictive methods and dependence reconstruction—is applicable in the studies of consistent patterns of development and aging of different organisms as well as in description and identification of factors affecting the state of the organism at the end of life and on morbidity and mortality. The possibility of interpreting the state of the organism at the molecular level, which is provided by the data mining methods, makes it possible to synthesize data sets combining and complementing the observations of yeast, worms, flies, mice, and humans in searching for determinants of health and longevity.

Data mining methods play an auxiliary role because they are intended to identify dependences hidden in empirical data and indicate directions for further research. All conclusions derived from mathematical processing can be considered final only after experimental verification in real biological objects.

Problems of biology and gerontology, in turn, create new challenges for data mining. An example is the study of gene expression on microarrays. The development of this technology made it possible to automatically obtain data on expression of tens of thousands of genes using a single tissue sample. To perform reliable statistical analysis of expression of all these genes and determine the influence of genes on the phenotypic characteristics of an organism and to search for possible targets for new drugs, the number of tissues should be an order of magnitude greater than the number of genes. All these samples should have the same properties and characteristics and differ only in statistically independent perturbations. Apparently, it is almost impossible to collect such a vast amount of material due to both organizational difficulties and natural heterogeneity of populations of living organisms. Therefore, there is the problem of the search for genes and genetic associations (genetic networks) that are statistically most important in the formation of the investigated phenotype and are associated with the risk of serious diseases [8]. An additional task is the problem of identification of statistical models whose dimensions (e.g., the number of parameters) are comparable to or greater than the number of samples used.

Due to the necessity to build complex statistical models, the problem of verification of results is exacerbated. The conventional criteria used for testing the statistical significance of the results of calculations can be either too optimistic, if the complexity of the model is not taken into account, or too pessimistic when it is taken into account roughly. Modern trends in the development of data mining methods, such as nonparametric estimation with complexity control and methods of screening and building assessments adapted to data, can overcome these difficulties.

REFERENCES

1. Anisimov, V.N., *Molekulyarnye i fiziologicheskie mekhanizmy stareniya* (Molecular and Physiological Mechanisms of Aging), St. Petersburg: Nauka, 2008.
2. Michalski, A.I. and Yashin, A.I., Biological control and lifespan, *Probl. Uprav.,* 2003, vol. 3, pp. 61–65.
3. Michalski, A.I., Rodionov, Yu.A., Manton, K.G., et al., Frequency of disability among elderly man and women, *Usp. Gerontol.*, 2009, vol. 22, no. 4, pp. 569–587.
4. Novosel'tsev, V.N. and Michalski, A.I., Mathematical modeling and aging: a research program, *Usp. Gerontol.*, 2009, vol. 22, no. 1, pp. 117–128.
5. Nosko, V.P., *Ekonometrika. Elementarnye metody i vvedenie v regressionnyi analiz vremennykh ryadov* (Econometrics: Elementary Models and Introduction to Regression Analysis of Time Series), Moscow: Inst. Ekon. Perekh. Perioda, 2004.
6. Tsurko, V.V. and Michalski, A.I., Statistical analysis of relationship cancer and associated diseases, *Usp. Gerontol.*, 2013, vol. 26, no. 4, pp. 766–774.
7. Azzalini, A. and Scarpa, B., *Data Analysis and Data Mining: An Introduction*, Oxford Univ. Press, 2012.
8. Barabasi, A.L., Gulbahce, N., and Loscalzo, J., Network medicine: a network-based approach to human disease, *Nat. Rev. Genet.*, 2011, vol. 12, pp. 56–68.
9. Budovskaya, Y.V., Wu, K., Southworth, L.K., et al., An elt-3/elt-5/elt-6 GATA transcription circuit guides aging in *C. elegans*, *Cell*, 2008, vol. 134, pp. 291–303.
10. Dato, S., Soerensen, M., Lagani, V., et al., Contribution of genetic polymorphisms on functional status at very old age: a gene-based analysis of 38 genes (311 SNPs) in the oxidative stress pathway, *Exp. Gerontol.*, 2014, vol. 52, pp. 23–29.
11. Li, Q., Yu, K., Li, Z., et al., MAX-rank: a simple and robust genome-wide scan for case-control association studies, *Hum. Genet.*, 2008, vol. 123, pp. 617–623.
12. Managbanag, J.R., Witten, T.M., Bonchev, D., et al., Shortest-path network analysis is a useful approach toward identifying genetic determinants of longevity, *PLoS One,* 2008, vol. 3, p. e3802.

13. Michalski, A., Johnson, T., Cypser, J., et al., Heating stress patterns in *Caenorhabditis elegans* longevity and survivorship, *Biogerontology*, 2001, vol. 2, pp. 35–44.

14. Myasnikova, E., Samsonova, A., Samsonova, M., et al., Support vector regression applied to the determination of the developmental age of a Drosophila embryo from its segmentation gene expression patterns, *Bioinformatics*, 2002, vol. 18, suppl. 1, pp. 87–95.

15. Sarup, P., Sorensen, P., and Loeschcke, V., The long-term effects of a life-prolonging heat treatment on the *Drosophila melanogaster* transcriptome suggest that heat shock proteins extend lifespan, *Exp. Gerontol.*, 2014, vol. 50, pp. 34–39.

16. Shamir, L., Wolkow, C.A., and Goldberg, I.G., Quantitative measurement of aging using image texture entropy, *Bioinformatics*, 2009, vol. 25, pp. 3060–3063.

17. Swindell, W.R., Accelerated failure time models provide a useful statistical framework for aging research, *Exp. Gerontol.*, 2009, vol. 44, pp. 190–200.

18. Swindell, W.R., Harper, J.M., and Miller, R.A., How long ~~shall~~ my mouse live? Machine learning approaches for prediction of mouse lifespan, *J. Gerontol., Ser. A*, 2008, vol. 63, pp. 895–906.

19. Tafaro, L., Cicconetti, P., Piccirillo, G., et al., Is it possible to predict one-year survival in centenarians? A neural network study, *Gerontology*, 2005, vol. 51, pp. 199–205.

20. Yashin, A., Cypser, J., Johnson, T., et al., Heat shock changes the heterogeneity distribution in populations of *Caenorhabditis elegans*: Does it tell us anything about the biological mechanism of stress response? *J. Gerontol., Ser. B*, 2002, vol. 57, pp. 83–92.

Translated by M. Batrukova

SPELL: 1. antistress, 2. catalase