SCIENTIFIC PUBLICATIONS OF THE STATE UNIVERSITY OF NOVI PAZAR
SER. A: APPL. MATH. INFORM. AND MECH. vol. 9, 1 (2017), 91-101.

Optimization for Classifying the Patients Using the Logic Measures for Missing Data

Natasa Glisović, Miodrag Rašković

Abstract: The aim of this paper is to show the system for making a decision in the diagnostics of the patients with autoimmune diseases. In this paper we consider the patients with systemic lupus erythematosus (SLE), Sy Sjogren and systemic sclerosis. In medicine, unlike many other fields, much of the relatively homogeneous and well systematized medical knowledge is implicitly given in the histories of patients. This opens the possibility that medical experts formulate a framework for their knowledge, and those systems based on the effective strategies judgment, produce the useful knowledge from such formulated framework and the adequate history of the illness of the patients. In this research we proposed new distances for the missing values in the cluster optimization. The support system in the diagnostics uses these proposed distances implemented in the program language c#.

Keywords: Systemic diseases, missing data, decision support, systemic lupus erythematosus, Sy Sjogren, systemic sclerosis.

1 Introduction

It is often required to establish how the data are connected, how certain data differ or do not go together with each other and what the measure of their comparison is. An important part in detecting the similarities and grouping the data into clusters has the choice of metrics and the accuracy of the cluster algorithm operation. For clustering, we use the machine learning. The machine learning can be observed as determining the dependence on the available data [4]. The methods of the classification machine learning according to the examples do the estimation of copying the unknown dependence between the input (of the data) and the system output (of the classification) according to the available examples of the right classification.

Manuscript received March 12, 2016; accepted Janyary 23, 2017.

Natasa Glisović is with the State University of Novi Pazar, Department of Mathematical Sciences, Novi Pazar, Serbia; Miodrag Rašković is with the Mathematical Institute of the Serbian Academy of Sciences and Arts (SANU), Beograd, Serbia

In our paper the method uses the proposed distances which have an aim that according to the patients base, which consists of patients ill with three systemic autoimmune diseases (SLE, Sy Sjogren and Sclerosis systemica), decides for a new patient to which group of diseases they belong [9][12].

The paper is divided into several sections. The mathematics problem, which was used for the support decision in section 2 is described. In section 3 the database will be described which was obtained by applying the model described in section 2. The conclusion of research will be given in section 4.

2 Mathematical model

The Metric learning has become a popular issue in many learning tasks and can be applied in a wide variety of settings, since many learning problems involve a definite notion of distance or similarity [1]. A metric or distance function is a function which defines a distance between the elements of a set [7][15]. A set with a metric is called a metric space. In many data retrieval and data mining applications, such as clustering, measuring the similarity between the objects has become an important part. Normally, the task is to define a function $sim(X,Y)$, where X and Y are two objects or sets of a certain class, and the value of the function represents the degree of "similarity" between the two [8][17]. Formally, a distance is a function D with nonnegative real values, defined on the Cartesian product $X \times X$ of a set X. It is called a metric on X if for every $x,y,z \in X$:

- $D(x,y) = 0$ if $x = y$ (the identity axiom);
- $D(x,y) + D(y,z) \geq D(x,z)$ (the triangle inequality);
- $D(x,y) = D(y,x)$ (the symmetry axiom).

A set X provided with a metric is called a metric space.

The missing data have challenged researchers since the beginning of the field research. What is the reason for the data missing? When people talk about the mechanisms of the missing, three terms come up: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Although statisticians prefer not to use the word cause, they do often use the words due to or depends on in this context. With MAR, the missing (i.e., whether the data are missing or not) may depend on the observed data, but not on unobserved data [14]. MCAR is a special case of MAR in which the missing does not depend on the observed data either [14]. With MNAR, the missing does depend on the unobserved data.

To calculate the distance between patients we need all the information about tests (analyses). The method based on the distance are very popular in the literature because they can be used for every type of data, while the corresponding function of the distance is suitable for that type of the data. Therefore, the problem of the data classification can be reduced to the problem of finding the distance function for that kind of the data. It results from this that the finding of the distance function has become an important research areas for the data processing and their accuracy.

The aim of the research is to suggest a new distance by which we can measure the distance of the object is which the data are missing.

The problem of the missing data is of great importance. When the mechanisms of the missing data are talked about, first of all, the reason for the missing data should be established. The data can miss out of many reasons. Some of them are: the data are not available, the mistakes which are made while working with the equipment, the nonconsistency with the other data so they are deleted, they are not filled in because of the look of understanding; they were not considered as important at the moment of input etc. The decision is important: what should be done with the missing data?

- Deleted the parameters where the missing data appear-which is not advisable, especially with the classification, if the missing values vary from one parameter to the other.
- Manual filling in the missing data is tiring and often impracticable.
- Automatic filling in: by a general constant, the mean value of the parameter for all who belong to the same class.
- The most probable value: the conclusion which is given on the basis of Bayes's formula or the decision trunk.

In the base with it was worked in this research, there were the missing data because for some of the patients there was no need for the analyses to be done, before them the disease had been found out, and some of the data are missing because of the loss of that datum at the moment of the input into the database. Also, the data were coded, so there was on idea of using Hammings distance and the proposal of a new distance which uses Hammings.

2.1 The description of the proposed distance

Let F be the final set with q elements.

Definition 2.1. *Hamming's distance $d\left(\underline{x},\underline{y}\right)$ between the two vectors $\underline{x},\underline{y} \in F^{(n)}$ is the number of places on which these two vectors differ.*

The proposed distance uses Hamming's distance and the stated formula. The usage of the stated formulas in the distance definition is because of the generality of the distance.

Let φ and ψ be two sets of the stated formulas to which the formulas which represent the literal conjunction belong. The proposed distance between these two sets of the stated formulas is defined with:

$$D(\varphi,\psi) = \frac{\max\limits_{A\in\varphi}\min\limits_{B\in\psi} d(A,B) + \max\limits_{B\in\psi}\min\limits_{A\in\varphi} d(A,B)}{2} \tag{1}$$

where $d(A,B)$ is Hamming's distance.

Example 2.2. *Let the two patients p_1 and p_2 be given with the values of the three analyses a, b and c so we have the following data about the patients:*

patient p_1 has $a \wedge b$ (from some reason there is a lack of the datum about the values of the finding c)

patient p_2 has $a \wedge \neg b \wedge c$

If we apply the proposed distance, the formula φ which describes the first patient is $a \wedge b \wedge (c \vee \neg c)$ if we change only with the conjunction we will have two formulas $a \wedge b \wedge c$ and $a \wedge b \wedge \neg c$ (those two formulas belong to the set of formulas φ). While the set of the formulas of the second patient that is the set ψ is the formula $a \wedge \neg b \wedge c$. If we apply the proposed distance to these two patients, we will get the distance between them:

$$D(\varphi,\psi) = \frac{\max\limits_{A\in\varphi}\min\limits_{B\in\psi} d(A,B) + \max\limits_{B\in\psi}\min\limits_{A\in\varphi} d(A,B)}{2} =$$

$$= \frac{\max\left\{\min\left\{d(a\wedge b\wedge c, a\wedge\neg b\wedge c)\right\}, \min\left\{d(a\wedge b\wedge\neg c, a\wedge\neg b\wedge c)\right\}\right\}}{2} +$$

$$+ \frac{\max\left\{\min\left\{d(a\wedge\neg b\wedge c, a\wedge b\wedge c), d(a\wedge\neg b\wedge c, a\wedge b\wedge\neg c))\right\}\right\}}{2} =$$

$$= \frac{\max\left\{\min\left\{1\right\}, \min\left\{2\right\}\right\} + \max\left\{\min\left\{1,2\right\}\right\}}{2} = \frac{2+1}{2} = \frac{3}{2}.$$

Remark 2.3. *The proposed distance does not fill in the missing data. It serves to determine the distance (similarity) of the patient with the other using only the information which value the missing data can have.*

2.2 The formulation of the optimization by clustering and the algorithm description

The proposed system (see the algorithm) uses the technique of clustering [10]. In order to carry out the analysis of clustering at all, it is necessary to define the measures of closeness of the two objects on the basis of their characteristics. The concept of similarity is determined according to the data themselves. The proposed distance (1), between the data can serve as a measure of that difference. The methods based on the distances are often desirable because of their simplicity of the application in different scenarios.

Let X be a set of m objects described with the help of n characteristic. The cluster algorithm for the weight characteristics, which groups the set X, k cluster, is based on the minimization of the aim function [3]:

$$F(W,Z,\Lambda) = \sum_{l=1}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n} w_{l,j}\lambda_{l,i}^{\beta} d(z_{l,i},x_{j,i}) \tag{2}$$

Subject to

$$w_{l,j} \in \{0,1\}, 1 \le l \le k, 1 \le j \le m \tag{3}$$

$$\sum_{l=1}^{k} w_{l,j} = 1,\ 1 \le j \le m \tag{4}$$

$$0 < \sum_{i=1}^{m} w_{l,j} < m,\ 1 \le l \le k \tag{5}$$

$$0 \le \lambda_{l,i},\ 1 \le l \le k,\ 1 \le i \le n \tag{6}$$

$$\sum_{i=1}^{n} \lambda_{l,i} = 1,\ 1 \le l \le k \tag{7}$$

where $k(\le m)$, the number of clusters, $\beta < 1$, $W = [W_{i,j}]$ the matrix of binary numbers of the order $k \times m$, $Z = [z_1, z_2, \ldots, z_k] \in \mathbb{R}^{n \times k}$ contains the centers of clusters $\Lambda = [\lambda_{l,i}]$ the matrix of real numbers of the order $k \times n$ and $0 \le d(z_{l,i}, x_{j,i})$ the proposed measure of the difference (1) between the center z_i and the object x_j for ith characteristic. The main idea of the optimization problem is a minimization of the difference measure between the centers of the clusters and the objects.

Algorithm

Step 1. Choose an initial matrix $Z^{(1)} \in \mathbb{R}^{n\times k}$ and set $\Lambda^{(1)}$ be a $k \times m$ matrix with all the entries being equal to $\frac{1}{n}$. Set $t = 1$.

Step 2. Determine $W^{(t+1)}$ so that $F(W^{(t+1)}, Z^{(t)}, \Lambda^{(t)})$ is minimized. If

$$F(W^{(t+1)}, Z^{(t)}, \Lambda^{(t)}) = F(W^{(t)}, Z^{(t)}, \Lambda^{(t)})$$

then stop; otherwise go to step 3.

Step 3. Determine $Z^{(t+1)}$ so that $F(W^{(t+1)}, Z^{(t+1)}, \Lambda^{(t)})$ is minimized.
If $F(W^{(t+1)}, Z^{(t+1)}, \Lambda^{(t)}) = F(W^{(t+1)}, Z^{(t)}, \Lambda^{(t)})$ stop; otherwise go to step 4.

Step 4. Determine $\Lambda^{(t+1)}$ so that $F(W^{(t+1)}, Z^{(t+1)}, \Lambda^{(t+1)})$ is minimized. If

$$F(W^{(t+1)}, Z^{(t+1)}, \Lambda^{(t+1)}) = F(W^{(t+1)}, Z^{(t+1)}, \Lambda^{(t)})$$

stop; otherwise go back to step 2.

Theorem 2.4. *[3] Let $\tilde{Z}$ and $\tilde{\Lambda}$ be fixed. The minimizer of the matrix $\tilde{W}$ for the optimization problem $\min_W F(W, \tilde{Z}, \tilde{\Lambda})$ where (3)-(7) is given by:*

$$\hat{w}_{l,j} = \begin{cases} 1, & \sum_{i=1}^{n} \tilde{\lambda}_{l,i}^{\beta} d(\tilde{z}_{l,i}, x_{j,i}) \leq \sum_{i=1}^{n} \tilde{\lambda}_{h,i}^{\beta} d(\tilde{z}_{h,i}, x_{j,i}),\ 1 \leq h \leq k \\ 0, & \textit{otherwise} \end{cases}.$$

Theorem 2.5. *[3] Let $\tilde{W}$ and $\tilde{\Lambda}$ be fixed. The minimizer $\tilde{Z}$ of the optimization problem $\min_Z F(\tilde{W}, Z, \tilde{\Lambda})$ is given with*

$$\hat{z}_{l,i} = \frac{\sum_{j=1}^{m} \tilde{w}_{l,j} x_{i,j}}{\sum_{j=1}^{m} \tilde{w}_{l,j}}$$

where $1 \leq l \leq k$.

When ith is numerical, or $\hat{z}_{l,i} = d_i^{(r)} \in DOM(D_i)$, where

$$|\{\tilde{w}_{l,j} |\ x_{i,j} = d_i^{(r)},\ \tilde{w}_{i,j} = 1\}| \leq |\{\tilde{w}_{l,j} |\ x_{i,j} = d_i^{(t)},\ \tilde{w}_{i,j} = 1\}|, \forall t \in DOM(D_i),$$

When ith is categorical. With $|Y|$ is marked the cardinality of the set Y.

Theorem 2.6. *[3] Let $\tilde{W}$ and $\tilde{Z}$ be fixed. The minimizer of the matrix $\tilde{\Lambda}$ for the optimization problem* $\min_{\Lambda} F(\tilde{W},\tilde{Z},\Lambda)$*, for which (3)-(7) and is given by*

$$\hat{\lambda}_{l,i}=\begin{cases}\frac{1}{n_i}, & \sum_{j=1}^{m}\tilde{w}_{l,j}d(\tilde{z}_{l,i},x_{j,i})=0, n_i=|\{t:\sum_{j=1}^{m}\tilde{w}_{l,j}d(\tilde{z}_{l,t},x_{j,t})=0\}|,\\ 0, & \sum_{j=1}^{m}\tilde{w}_{l,j}d(\tilde{z}_{l,i},x_{j,i})\neq 0,(\exists t)\sum_{j=1}^{m}\tilde{w}_{l,j}d(\tilde{z}_{l,t},x_{j,t})=0\\ \frac{1}{\sum_{t=1}^{n}\left[\frac{\sum_{j=1}^{m}\tilde{w}_{l,j}d(\tilde{z}_{l,i},x_{j,i})}{\sum_{j=1}^{m}\tilde{w}_{l,t}d(\tilde{z}_{l,t},x_{j,t})}\right]^{\frac{1}{(\beta-1)}}}, & \sum_{j=1}^{m}\tilde{w}_{l,j}d(\tilde{z}_{l,t},x_{j,t})\neq 0,\forall t,\ 1\le t\le n\end{cases}$$

.

The algorithm represents the cluster algorithm for the weight characteristics where the matrix W is determined in accordance with Theorem2.4, the centers of the cluster Z in every interaction are in accordance with the Theorem2.5, and the matrices of the weights for the characteristics in accordance with Theorem2.6.

Theorem 2.7. *[3] The optimization cluster algorithm converges into the final number of iteration.*

On the basis of the Theorem2.7, this algorithm for the pondered characteristics converges. However, it stops at the local minimum [2]. The time complexity of the algorithm is $O(l\times m\times n\times k)$ where l is the total number of iterations, k the number of clusters, n the number of characteristics and m the number of characteristics in the observed set. So, the proposed algorithm is adapted for the work a great number of set data.

3 The description of the database and the results of the research

In our paper we used the database of the patients with one of the mentioned three systemic autoimmune diseases: systemic lupus erythematosus (SLE), Sy Sjogren and systemic sclerosis (SScl).

Every patient was diagnosed on the basis of appropriate criteria. The selection of variables is established by the expert committee but is also under the strict influence of the statistical analyses. The classification criteria for SLE were established by the American College of Rheumatology from 1982. Later, in 1997 the criteria were revised [6]. Systemic Lupus International Collaborating Clinics (SLICC) set up the new classification criteria in 2012 [11]. The diagnostic criteria for Sy Sjogren [13][16] and the systic sclerosis [5], are given in a similar way.

The data set consists of 45, at random, chosen patients with the systemic autoimmune diseases, among them, 15 of the subjects were diagnosed as SLE, 19 as Sjogren's syndrome, 11

as progressive systemic sclerosis and 3 had both SLE [18] and Sjogren's syndrome. These patients are diagnosed and treated at the Clinic for Allergology and Immunology Clinical Centre Belgrade during the period of 2012-2014. We used a set of 87 variables, formed on the basis of 'availability' and 'prices' of the diagnostic methods. The first group of 33 variables consisted of the data obtained by anamneza and clinical check-up of patients, that is 'the most available group of data'. The second group of variables of 37-70 consisted of the data obtained by the labaratory treatment. The third group of 71-87 consisted of the variables obtained by the clinical check-ups which included invasive diagnostic procedures such as a biopsy of salivary glands, a biopsy of kidneys. In this research, the system for making a decision about the diagnosis of the new patient is implemented according to the database in which we already have the patients with diagnoses.

The implemented system The System for the Decision Support and Optimization for Classifying the Patients (SDSCP) using the measures for missing data is the desktop application. It has been developed in .NET technology and c# language. NET is a software platform which is developed by Microsoft and which is used with Windows operating system. The application is designed for the Windows system, but there is a possibility of using the same on Linux and Android operating systems through Mono platform.

The database is local and information is stored in JSON (Java Script Object Notation) format. JSON is textually based, an open standard, designed for the understandable data exchange among the people. It is from Javascript language and serves for the representation of the data structures. It is independent of the programme language in which it is used.

The used developing surroundings is Microsoft Visual Studio 2003 Express. It is one of the most popular surroundings for the development, especially for NET platform and c# language. It contains the editor of the code, debugger, the designer tools for graphical applications. It enables an easy organization of the project. As a result it gives quick, effective and comfortable work.

For following the project version GIT was used, a free distributed system for the control of the versions.

The success of the classification, thus defined distance implemented through the described algorithm, has been shown through the comparison of the results with the other methods of filling in the missing data such as the middle value and linear regression. The results of the comparison are that the middle value method was the worst in clustering then the linear regression which was much better than the middle value. The proposed distance gave the best results (Table 1.).

Method of filling in missing data	Performance as a percentage
Middle value	80%
Linear regression	93%
The proposed distance	98%

Table 1. The performance results of different methods. Optimization problem is given the best results when applying the proposed distance

4 Conclusion

The proposed system showed better results than the existing methods. The percentages of the successful completion show that on the basis of the proposed distance, the system gave a wrong conclusion with one patient, while at the other method that number was considerably bigger, so with three (linear regression) and nine (middle value) patients there was a wrong decision.

In the medicals opinion, who had tested the proposed system in practice, this system is useful because the results were very reliable and enabled giving the suggestions of an adequate therapy. For this class of disease there is an additional therapy which can be, on the basis of the proposed system, determined.

The proposed distances can be used in other areas as well as applied to the other problems where there is a need for finding out the most similar or the most different attributes.

In future research is planned to improve in the direction of finding an adequate account of the weight of each of the analysis.

Acknowledgement We would like to thank Prof. Dr Sanvili Raskovic, Prof. Dr Aleksandri Popadic Peric and Prof. Dr Vojislavu Djuricu doctors at the Clinic for Allergology and Immunology Clinical Centre Belgrade for the suggestions, the data and the testing which contributed to the improvement of this research.
The work presented here was supported by the Serbian Ministry of Education and Science (project III44006).

References

[1] AGRAWAL R., FALOUTSOS C., SWAMI A.,*Efficient similarity search in sequence databases*, Proc. 4th Int. Conf. On Foundations of Data Organizations and Algorithms, 1993. – Chicago. pp. 69-84.

[2] BEZDEK JC., *A convergence theorem for the fuzzy ISODATA clustering algorithms*, IEEE Trans Pattern Anal Mach Intell. 1980; 2:1-8.

[3] CHAN EZ, CHING WK, NG MK, HUANG JZ., *An optimization algorithm for clustering using weighted dissimilarity measures*, Pattern Recognit. 2004;37:943-952.

[4] CHERKASSKY V., MULIER F. M., *Learning from Data: Concepts, Theory, and Methods,* 2^{nd} Ed, John Wiley-IEEE Press, 2007.

[5] FRANK VAN DEN HOOGEN, DINESH KHANNA, JAAP FRANSEN,SINDHU R JOHNSON, MURRAY BARON, ALAN TYNDALL, 2013 classification criteria for systemic sclerosis: an American college of rheumatology/European league against rheumatism collaborative initiative Ann Rheum Dis 2013;72:1747-1755.

[6] GLADMAN DD, UROWITZ MB., *Prognosis, mortality and morbidity in systemic lupus erythematosus*. In: Wallace DJ, Hahn BH. Dubois' lupus erythematosus. 7th ed. Philadelphia: Lippincott Williams & Wilkins; 2007:1333-53.

[7] GORONZY JJ, WEYAND CM., The innate and adaptive immune systems. In: Goldman L, Ausiello D, eds. *Cecil Medicine* . 23rd ed. Philadelphia, Pa: Saunders Elsevier;2007: chap 42.

[8] LI M., CHEN X., MA B., VITANYI P., *The similarity metric*, IEEE Transactions on Information Theory, 2004, vol.50, No. 12, pp.3250-3264.

[9] NADASHKEVICH O, DAVIS P, FRITZLER MJ., *A proposal of criteria for the classification of systemic sclerosis*, Med. Sci. Monit. 2004 Nov;10(11):CR615-21. Epub 2004 Oct 26.

[10] PANG-NING TAN, ET AL., *Introduction to data mining*. Pearson Education India, 2007.

[11] PETRI M, ORBAI AM, ALARCÓN GS, GORDON C, MERRILL JT, FORTIN PR,, Derivation and validation of the Systemic Lupus International Collaborating Clinics classification criteria for systemic lupus erythematosus. Arthritis Rheum. 2012 Aug;64(8):2677-86.

[12] RAHMAN A, ISENBERG DA., SYSTEMIC LUPUS ERYTHEMATOSUS, N Engl J Med. 2008 Feb 28;358(9):929-39.

[13] SC SHIBOSKI, CH SHIBOSKI, LA CRISWELL, AN BAER, S CHALLACOMBE, H LANFRANCHI, American College of Rheumatology Classification Criteria for Sjögren's Syndrome: A Data-Driven, Expert Consensus Approach in the SICCA Cohort. Arthritis Care Res (Hoboken). Apr 2012; 64(4): 475–487.

[14] SCHAFER JL, GRAHAM JW, 2002. Missing data: our view of the state of the art. Psychol. Methods 7:147–77

[15] SIEGEL RM, LIPSKY PE, Autoimmunity. In: Firestein GS, Budd RC, Harris Ed, et al, eds. *Kelley's Textbook of Rheumatology* . 8th ed. Philadelphia, Pa: Saunders Elsevier; 2009:chap 15.

[16] VITALI C, BOMBARDIERI S, JONSSON R, MOUTSOPOULOS HM, ALEXANDER EL, CARSONS SE, DANIELS TE, FOX PC, FOX RI, KASSAN SS, PILLEMER SR, TALAL N, WEISMAN MH, European Study Group on Classification Criteria for Sjögren's Syndrome. Classification criteria for Sjögren's syndrome: a revised version of the European criteria proposed by the American-European Consensus Group. Ann Rheum Dis. 2002 Jun;61(6):554-8.

[17] VITANYI P., Universal similarity, ITW2005, Rotorua, New Zealand, 2005.

[18] WATANABE N, TAKABAYASHI K., Recent investigations on the basis of pathogenesis of SLE and new therapeutic approaches. Nihon Rinsho. 2009 Mar;67(3):500-5.