Personalized Medicine Universe 5 (2016) 32–40

Contents lists available at ScienceDirect

Personalized Medicine Universe

journal homepage: www.elsevier.com/locate/pmu

Original article

Improvised prophecy using regularization method of machine learning algorithms on medical data

Vadamodula Prasad [a,*], T. Srinivasa Rao Dr [b], P.V.G.D. Prasad Reddy Prof [c]

[a] Dept of Computer Science & Engg, Raghu Institute of Technology, AP, India
[b] Dept of Computer Science & Engg, Gitam Institute of Technology, Gitam University, AP, India
[c] Dept of Computer Science & Systems Engg, AUCOE (A), Andhra University, AP, India

ARTICLE INFO

Article history:
Received 17 July 2015
Received in revised form
29 August 2015
Accepted 11 September 2015

Keywords:
Thyroid
Ridge regression
Machine learning
Regularization method

ABSTRACT

Patients with thyroid disease (TD) boast continuously increasing because of excessive growth of thyroid gland and its hormones. Automatic classification tools may reduce the burden on doctors. This paper evaluates the selected algorithms for predicting thyroid disease diagnoses (TDD). The algorithms considered here are regularization methods (RM) of machine learning algorithms (MLA). The analysis report generated by the proposed work suggests the best algorithm for predicting the exact levels of TDD. This work is a comparative study of MLA on UCI thyroid datasets (UCITD). The developed system deals with RM i.e., ridge regression algorithm (RRA) & least absolute shrinkage and selection operator algorithm (LASSO). The above algorithms personage produce at most 79% accuracy by RRA and 98.99% accuracy by LASSO. Thus, this paper shows the importance of LASSO, along with an example for parameter generation. The decisive factors (DF) also suggest the accuracy rate of LASSO is much better when compared with RRA.

1. Introduction

The objective of the proposed work is used for identification of the 7 categories of TD [17] i.e., hyperthyroid, hypothyroid, binding protein, replacement therapy, anti-thyroid therapy, non-thyroid illness and miscellaneous features by considering 28 major pristine attributes. The work is done by capturing the attention of MLA i.e., RM [14] and SMS [13]. The system developed is used for offline analysis [3] on TDD. Originally the system is developed by using a SMS, when the SMS fails to identify the relevant data in the knowledge, then automatically the system invokes to get the optimistic disease [1,2] by using two individual methods of RM which belongs to MLA namely (i) RRA [20] (ii) LASSO [5]. The method of RRA on TDD produced a result of maximum accuracy 79%, where as the LASSO on TDD produced 98.99%, but as per the comparisons made by DF [21], it is understood that LASSO [7] performance is best in prophecy, than RRA. Hence, the system is further enhanced with DF values namely meticulousness (*mT*), exactitude (eX), compassion (cP) and rigor (rI) in order to provide the disease name with its prevention and curing methods. Here, we obtain the diagnostic methods for different symptoms entered by the user dynamically. If the data entered by the user is sufficient and if it matches the knowledge, then the proposed system displays the actual disease with which the human is suffering, or else it displays the dialogue box stating that the knowledge is insufficient. For the purpose of calculating missing attribute values [10], RM is introduced, so that the nearby related diseases of TD can be determined based on the raised values [19], if the exact disease is unavailable in the knowledge of the expert, then the proposed system identically shows the probabilistic disease with which the human is suffering from. The LASSO method provides the actual disease by considering the values generated by the function

Abbreviations: TD, thyroid disease; TDD, thyroid disease diagnose; MLA, machine learning algorithms; RM, regularization methods; UCITD, UCI thyroid datasets; RRA, ridge regression algorithm; LASSO, least absolute shrinkage and selection operator algorithm; DF, decisive factors; SMS, string matching system; mT, meticulousness; eX, exactitude; cP, compassion; rI, rigor; EAS, expert advisory system; NTP, number of true positives; NFP, number of false positives; NTN, number of true negatives; NFN, number of false negatives; TP, true positives; FP, false positives; TN, true negatives; FN, false negatives; RS, rule-set.

* Corresponding author. Department of Computer Science & Engineering, Raghu Institute of Technology, Dakamarri, Visakhapatnam 531162, Andhra Pradesh, India. Tel.: +91 9440024661, +91 9177559425.

E-mail address: prasad_v542@yahoo.co.in (V. Prasad).

http://dx.doi.org/10.1016/j.pmu.2015.09.001
2186-4950/© 2015 International Society of Personalized Medicine. Published by Elsevier B.V. All rights reserved.

LASSO_K, if the raised values satisfies the RM then LASSO is responsible for providing the optimized disease which could be predicted.

2. Material & methods

2.1. Role of TDD, MLA & health professionals

2.1.1. Thyroid medical diagnosis (TDD)

Thyroid medical diagnosis (TDD) is known to be subjective and depends not only on the available data but also on the experience of the physician, intuition, biases and even on the psycho-physiological condition of the physician. Automatically derived diagnostic knowledge may assist physicians to make the diagnostic process more objective and more reliable. Database is constructed using 390 patient's data consisting 2096 rules obtained from the Intelligent system Laboratory of K.N.Toosi University of Technology from Imam Khomeini hospital [12,13,17].

2.1.2. Machine learning algorithms (MLA)

Machine learning algorithms (MLA) [1,15] has recompensed erudition, high parallelism and pace against noise. The readiness of reuse or reprocess feature in MLA made us to choose it as an option to develop this proposed work. Thus, we proposed a method of invoking MLA into this work of TDD. Nowadays, by developing machinery and information in therapeutic sciences, the computer science professionals are capable of providing expert advisory system EAS [2,9,12] to diagnose different kinds of diseases with high accuracy.

2.1.3. Health professionals

Health professionals are made to use these systems, due to some developed errors during general verdict process. Ailment diagnosis operations using EAS are performed based on set of disease symptoms. These systems are based on artificial intelligence [18] which helps the physician to minimize the costs and time in valuable diagnoses. Our previous proposed system [1] had succeeded in tracing out the missing attribute values for the identification of original TDD; the system solves all kind of problems raised by thyroid gland and the hormone produced by it.

2.2. Regularization method

RM [15] in fields of MLA and problems refers to a process of introducing additional information in order to solve an ill-posed problem in TDD. In this method, we provided some restrictions on data for smoothness or bounds in the data base.

The novelty which you can observe in this proposed RM algorithms are that, these are used to calculate the missing values, because it is already informed that the data of this thyroid is inconsistent and vague. As per the representation, the pseudo code which is implemented can show you the exact working style of RRA in RM. The continuous values which are obtained are used for prophecy of TDD. The method of obtaining the continuous numerical values is mentioned in the pseudo codes of Sections 2.2.1 & 2.2.2.

2.2.1. Ridge regression algorithm

RRA used for analyzing multiple regression data which is having multi co-linearity. When multi co-linearity [6,8,11] occurs, least squares estimates are unbiased (restrictions are avoided), but their variances are large so they are far from the true value. By adding a degree of bias (restriction) to the regression estimates, RR reduces the standard errors. It is hoped that the net effect will give estimates that are more reliable and useful for TDD.

Pseudo code of the RRA developed:

Step 1:

Input vector, **X = (S$_1$, S$_2$ … S$_p$)**, where ‘p’ value ranges from 1 to 28 considering the pristine attribute values, which is the count of symptoms. Let $\{S_j \mid j = 1, 2, \ldots, j\}$ be a set of J samples. The outcome of each sample S_J *is* denoted by the trait $\mathbf{t_J}$.

Step 2:

Output Y is real-valued. For the problem defined above, the I-dimensional binary vector ***C*** = {***b***$_1$, ***b2…..***, ***b***$_l$} where each bit indicates (‘1’ and ‘0’) based on the availability in the algorithm

Step 3:

Predict Y from X by **f (X)** so that the expected loss function **E (L(Y, f (X)))** is minimized. To define the objective function, we use ridge regression to feature subset selection. Through regression coefficients by forcing a penalty on the size of subset, ridge regression is able to smoothly approach the solution with less variance compared with forward and backward stepwise selection methods.

Give the trait ***t*** and ***C*** leading to a subset of selected attributes.

X = {**x$_{k1}$, x$_{k2}$,x$_{k3}$**, … …**x$_{kn}$**},we fit the following pair wise interaction model as Eq. (1)

From Eq. (1), we can obtain selected attributes as displayed in Eq (2), i.e.

$$t(C) = \beta_0 + \sum_{i=1}^{n} \beta_i x_{ki} \sum_{u=1}^{n-1} \sum_{v>u} \beta_{uv} x_{ku} x_{kv} \qquad \text{Eq. (2)}$$

Step 4:

Square loss: **L(Y, f (X)) = (Y − f(X))^2**. The optimal predictor **f*(X) = argmin$_{\mathbf{f\ (X)}}$E(Y − f(X))^2.** Then the ridge regression is to compute the coefficient set $\overset{\wedge}{\beta} = \{\beta_0, \beta_1, \beta_2, \ldots\ldots..\beta_n\}$ by minimizing the penalized residual sum of squares obtained from *trait **t*** from Eq. (2).

Therefore,

$$\overset{\wedge}{\beta} = \arg\min \beta \left\{ \sum_{j=1}^{J} \left(t_j - \beta_0 - n \sum_{i=1} \beta_i x_{ki}^{j} - \sum_{u=1}^{n-1} \sum_{v>u} \beta_{uv} x_{ku}^{j} \right)^2 + \lambda \left(\sum_{i=1}^{n} \beta_i^2 + \sum_{u=1}^{n-1} \sum_{v>u} \beta_{uv}^2 \right) \right\} \qquad \text{Eq. (3)}$$

Where complexity parameter is denoted as (λ) in Eq. (3)

Then, the objective function is defined as a multiple R^2 value, which is a decreasing function of the residual sum of squares from obtained Eq (3)

$$R^2(c) = 1 - \frac{\sum_{j=1}^{J} \left(t_j - \overset{\wedge}{\beta_0} - \sum_{i=1}^{n} \overset{\wedge}{\beta} x_{ki}^{j} - \sum_{u=1}^{n-1} \sum_{v>u} \overset{\wedge}{\beta_{uv}}\ x_{ku}^{j} x_{kv}^{j} \right)^2 + \varDelta}{\sum_{j=1}^{J} (t_j - t)^2} \qquad \text{Eq. (4)}$$

Where **f*(X) = E(Y |X)** and “**Δ**” values in Eq (4) is defined as

$$\varDelta = \lambda \left(\sum_{i=1}^{n} \beta_i^2 + \sum_{u=1}^{n-1} \sum_{v>u} \beta_{uv}^2 \right) \qquad \text{Eq. (5)}$$

$\bar{t} = 1/J \sum_{j=1}^{J} t_j$ is the mean outcome and $\sum_{j=1}^{J} (t_j - \bar{t})^2$ is the total outcome variation. The complexity parameter $\lambda \geq 0$ controls the eradication in accuracy levels. In the case of ordinary least squares regression $(\lambda = 0)$, *R2* lies between 0 and 1; $R^2 = 1$ indicates perfect model fit.