Journal of Mathematical Modeling
Vol. 11, No. 3, 2023, pp. 441-450. Research Article

JMM

Strategies for disease diagnosis by machine learning techniques

Elham Hafezieh*, Ali Tavakoli, Mashallah Matinfar

University of Mazandaran, Babolsar, Iran
Email(s): hafeziehelham@gmail.com, a.tavakoli@umz.ac.ir, m.matinfar@umz.ac.ir

Abstract. Machine learning (ML) techniques have become a point of interest in medical research. To predict the existence of a specified disease, two methods K-Nearest Neighbors (KNN) and logistic regression can be used, which are based on distance and probability, respectively. These methods have their problems, which leads us to use the ideas of both methods to improve the prediction of disease outcomes. For this sake, first, the data is transformed into another space based on logistic regression. Next, the features are weighted according to their importance in this space. Then, we introduce a new distance function to predict disease outcomes based on the neighborhood radius. Lastly, to decrease the CPU time, we present a partitioning criterion for the data.

Keywords: Prediction, machine learning, classification, penalized logistic Ridge regression.
AMS Subject Classification 2010: 34A34, 65L05.

1 Introduction

Machine learning (ML) provides methods, techniques and tools to help clarify diagnostic and prognostic issues in medical domains. ML is being used for the analysis of the importance of clinical parameters and their combinations for prognosis, e.g., predicting disease progression, extracting medical knowledge for outcome research, therapy planning and support, and overall patient management [13]. Supervised ML algorithms are a type of ML technique that can be applied according to what was previously learned to get new data using labeled data and to predict future events or labels. Many studies have been conducted using supervised ML techniques in various medical domains which are summarized here.

In 2016, Graph-guided joint prediction of the class label and clinical scores for Alzheimer's disease have been proposed by Yu et al. [15]. Next in 2017, Zhu et al. in [18] have studied a novel relational regularization feature selection method for joint regression and classification in Alzheimer's diagnosis.

*Corresponding author
Received: 23 Januarry 2023/ Revised: 8 April 2023/ Accepted: 7 May 2023
DOI: 10.22124/JMM.2023.23678.2114

A logistic regression model was utilized to diagnose cardiac disease by Kumar et al. [11]. Anwar Zeb et al. developed a mathematical model to present the dynamical behavior of COVID-19 infection by incorporating isolation class [16]. Batista in [1] studied the logistic growth regression model, which is used for the estimation of the final size of the coronavirus epidemic. Li et al. suggested a ML-based approach such as K-nearest neighbors, support vector machines, and decision trees for classifying heart disease [12]. In 2022, Iwendi et al. used machine learning algorithms for COVID-19 health analysis and prediction [10]. Dritsas et al. [4], in 2023, predicted liver disease risk by using supervised machine learning models. In 2023, machine learning techniques were used for the prediction of mortality in lung cancer patients by Huang et al. [8].

Many researchers have used KNN [5, 9, 12] and logistic regression [5, 11] for predicting disease outcomes [11, 12]. These methods have their own problems; for example, in the KNN method, an appropriate K can be experimentally obtained for the entire samples in the dataset, which may not work for other datasets. Also, in the logistic regression model, the probability of belonging to some data is close to 1/2, which can cause incorrect predictions in data classification. Our methodology to overcome this problem is using the ideas of both probability-based and distance-based ML models. For this sake, by computing the parameters of logistic regression, we transform the data into another space which the features are weighted according to their importance. Then, we introduce a new distance function such that a query point is predicted inside of the neighborhood whose center is the query point and the radius is the new distance.

The organization of this paper is as follows. In Section 2, the logistic regression is explained briefly. Section 3 provides the methodology solution. Also, in Section 4, the idea of partitioning is added to the proposed method to reduce the CPU time. Finally, in Section 5 we evaluated the experimental research results.

2 Logistic Regression

One of the techniques for predicting disease outcomes is logistic regression that we explain briefly in the sequel. According to supervised learning, we have the correct response y for each input $\mathbf{x}$. Let $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be the training set where n is the number of training examples. Each training input $\mathbf{x}_i = (x_i^1, x_i^2, ..., x_i^p)^T$ is a p-dimensional vector of features. Let $Y = (y_1, y_2, ...y_n)$ denote the response vector where $y_i \in \{1, ..., c\}$ with c being the number of classes. To find the weight of each feature, we consider the Bernoulli distribution with logistic function as follows [2]

$$p(y_i|x_i;\theta) = h_\theta(x_i)^{y_i}(1-h_\theta(x_i))^{1-y_i}, \quad i = 1, \ldots, n,$$

where $h_\theta(x) = 1/(1+\exp(-\theta^T x))$ is referred to as logistic function with $\theta \in \mathbb{R}^p$ as the parameters. The overall likelihood function based on the ordinal logistic model can be expressed as

$$L(\theta) = \prod_{i=1}^{n} p(y_i|x_i;\theta) = \prod_{i=1}^{n} h_\theta(x_i)^{y_i}(1-h_\theta(x_i))^{1-y_i}.$$

The goal of maximum likelihood estimation is to quantify the parameters in order to maximize the likelihood function over the parameter space. The parameters θ can be estimated by minimizing the penalized negative log-likelihood, which is defined by

$$J(\theta) = \sum_{i=1}^{n} -y_i \log h_\theta(x_i) - (1-y_i)\log(1-h_\theta(x_i)) + \gamma\|\theta\|^2, \tag{1}$$

where $\|.\|$ denotes the L_2-penalty as in the ridge regression [7] and γ is a positive tuning parameter. The regularization term $\|\theta\|^2$ will be used to avoid uncontrolled growth of θ. The optimization problem can be solved by the gradient descent method [3] which is given in Algorithm 1.

Algorithm 1 Gradient descent algorithm with regularization

1: Input: θ , α
2: repeat until convergence
3: $\theta_0 = \theta_0 - \alpha \sum_{i=1}^{n} (h_\theta(x^{(i)}) - y^i) x_0^i$
4: $\theta_j = \theta_j(1 - \alpha\gamma) - \alpha \sum_{i=1}^{n} (h_\theta(x^{(i)}) - y^i) x_j^{(i)}$
5: output: θ

The parameters α and γ are chosen such that $0 < 1 - \alpha\gamma < 1$ (see [14]). As we mentioned before, here Step 2 of Algorithm 1 shrinks θ to tackle its uncontrolled growth. The logistic function takes the output between zero and one for each input; therefore, if the output is closer to one, the probability of belonging to class one is higher and vice versa. The class of each query point can be considered by

$$cl(x_i) = [h_\theta(x_i)], \tag{2}$$

where $cl(x)$ shows the class of x and $[h_\theta(x_i)]$ is defined by

$$[h_\theta(x_i)] = \begin{cases} 0, & 0 < h_\theta(x_i) \leq 0.5, \\ 1, & 0.5 < h_\theta(x_i) < 1. \end{cases}$$

The reality is that if $h_\theta(x_i)$ given by (2) is close to $1/2$, There may have been an incorrect classification of the classes.

3 Methodology solution

In this section, we explain our proposed method in the following. It is well known that in the diagnosis of disease, some features have more importance. To this end, for each training point x_i, we use the following transformation

$$X_i = \theta * x_i, \quad i = 1, \ldots, n,$$

where $\theta * x_i = (\theta_1 x_i^1, \ldots, \theta_p x_i^p)^T$, and θ_i's are the parameters of logistic regression. In addition to logistic regression, the method of KNN is one of the ML techniques that is used for predicting the disease outcomes. However, one of the difficulties of using KNN is that the optimum value of K is derived experimentally. Here, in our proposed method we use the distance-based technique explained as follows.

A neighborhood around a given query point $X_q = \theta * x_q \in \mathbb{R}^n$ with a radius of d is defined by

$$B_d(X_q) = \{X \in \mathbb{R}^n | \quad \|X - X_q\| < d\},$$

where $\|.\|$ is the Euclidean norm. An appropriate selection for the radius d is challenging. In the following, we present some suggestions.

Let $\Xi_G = D$ be the set of all data points and Ξ_R be a small randomly set of data in Ξ_G. One of the possible suggestions is

$$d = \max_{X_v \in \Xi_G \setminus \Xi_R} \min_{X_t \in \Xi_R} \|X_t - X_v\|. \tag{3}$$

This selection of d may gives a small value of d. However, changing the min and max in (3), i.e.,

$$d = \min_{X_v \in \Xi_G \setminus \Xi_R} \max_{X_t \in \Xi_R} \|X_t - X_v\|, \tag{4}$$

may leads to a big value of d. But, the main problem for the selection of d in (3) and (4) is that they do not depend on the query point. In other words, it had better to choose a floating d for each query point. For this sake, a suitable selection of d can be given based on the average distance in each class. In other words, let Υ_c and Υ_n be the given dataset in the classes 0 and 1, respectively. We define the radius d_q by

$$d_q := \max\{\frac{1}{\#\Upsilon_c} \sum_{X_{c_i} \in \Upsilon_c} \|X_{c_i} - X_q\|_\infty, \frac{1}{\#\Upsilon_n} \sum_{X_{n_i} \in \Upsilon_n} \|X_{n_i} - X_q\|_\infty\}, \tag{5}$$

where $\#A$ denotes the number of members of the set A. when data within range (5) are identified, split them based on their negative and positive results and let them into positive and negative sets.
Putting

$$\begin{aligned} c\ell_1 &= \#\{X_i \in \Upsilon_c : \|X_q - X_i\| < d_q\}, \\ c\ell_0 &= \#\{X_i \in \Upsilon_n : \|X_q - X_i\| < d_q\}, \end{aligned} \tag{6}$$

the class of each query point is introduced by

$$c\ell_q = \begin{cases} 1, & \text{if} \quad c\ell_1 > c\ell_0, \\ 0, & \text{if} \quad c\ell_1 \leq c\ell_0. \end{cases} \tag{7}$$

Now, we summarize the procedure of our proposed method in Algorithm 2.

Algorithm 2 Prediction

1: Input:Training dataset $\{(x_i, y_i)\}_{i=1}^n, y_i = \{0 \text{ or } 1\}$ and a query point x_q
2: Preprocessing datasets
3: Fit model (1) and find the parameters by Algorithm 1
4: Put $X_i = \theta * x_i$ and $X_q = \theta * x_q$
5: Compute d_q by (5) on $\{(X_i, y_i)\}_{i=1}^n$
6: Compute $c\ell_0$ and $c\ell_1$ by (6)
7: Output: Predict the class of X_q by (7)

Remark 1. *Taking the relation (5) into account, for each query point X_q, we should compute the distance of X_q from all data points. It is a reality that for standardized data points, the value of d_q for query points are close to each other, therefore, we can consider a fixed d_q for most of data points.*

4 Partitioning

When we are dealing with big data, computing d_q using (5) and the parameters of logistic regression take a notable CPU time. In this section, we present a partitioning criterion to reduce the CPU time in our proposed method, as expressed in the previous section. In fact, we applied our proposed method on a

part of the training dataset. In the sequel, we explain our strategy.
Taking

$$X_i \in [a_1,b_1] \times [a_2,b_2] \times \cdots \times [a_p,b_p], \qquad i = 1,\ldots,N,$$

we partition each $[a_j,b_j]$ into k_j parts so that

$$a_j = a_{j,0} < a_{j,1} < \cdots < a_{j,k_j} = b_j.$$

This creates $K_c = k_1 \times k_2 \times \cdots \times k_p$ cells. It is now easy to determine the cell of each query point. To predict the class of the query point X_q, we assume that this query point is in the cell C that includes n_c data points. Now, the class of the query point is identified based on our proposed method with n_c samples in cell C. In Figure 1, we observe that two features are partitioned into 25 cells in which the triangles and squares show the data in classes 0 and 1, respectively. Also, it shows that the query point lies in the red cell.

Remark 2. *We note that the cell of each datum is known before and so we do not take time on the cell location of data. However, the main disadvantage of the current method is that it takes fairly a lot of memory usage, especially when we are dealing with many features.*

Remark 3. *In our work, most of the features are binary. To determine the cells of data, just a few features are needed to be partitioned. On the other hand, we know that if an interval $[a,b]$ is partitioned into 2^r subintervals, then that of the real number in $[a,b]$ will be determined at most by $r+1$ searches. Therefore, fairly few searches are required to be determined the cell of a query point.*

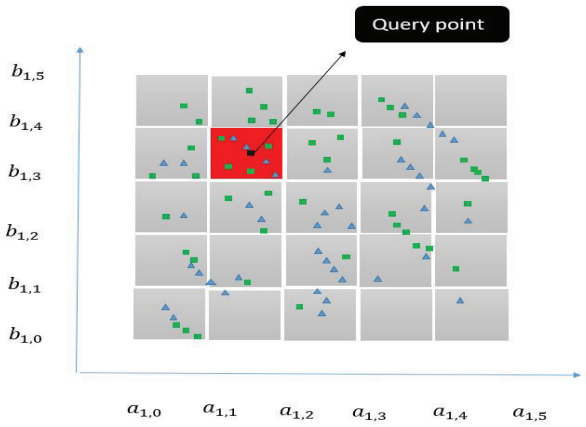

Figure 1: Schematic illustration partitioning two features.

5 Experimental results

In this research, we utilize the most commonly used metrics such as accuracy, precision, recall, F1-score, and AUC to evaluate the performance of the proposed method [6, 17]. The elements of these

metrics consist of true negative (TN), true positive (TP), false positive (FP) and false negative (FN). The F1 score is a weighted average of Precision and Recall calculated by

$$F1score = \frac{2 \times Precision \times Recall}{Precision + Recall},$$

where

$$Precision = \frac{TP}{TP+FP},$$

is the portion of retrieved instances that are relevant, and

$$Recall = \frac{TP}{TP+FN},$$

is the portion of relevant instances that are retrieved. The F1 score reaches its best value at 1 and worst value at 0. Accuracy summarizes the performance of the classification which is calculated by

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP}.$$

In order to verify the efficiency of our proposed method, we have conducted research on about 600 patients with suspected COVID-19, who were admitted to different hospitals in Iran at different time intervals. The symptoms which show up within 2 to 14 days of exposure to the virus, were assayed for each patient. These symptoms consist of Objective and Subjective. Objective symptoms are those evident to the doctor. In other words, these symptoms are measurable and called physical signs. Examples of such physical signs are temperature, Saturation of Peripheral Oxygen(SPO2), and C-Reactive-Protein(CRP). A subjective symptom is perceptible only to the patient such as fever, shortness of breath, and fatigue. The features of our dataset consist of 12 (COVID-19) symptoms such as sex, age, fever, sore throat, fatigue, loss of smell, dry cough, chest pain, shortness of breath, SPO2, temperature, and CRP. The results of the PCR test are thought of as binary (positive/negative) responses which denote the class label. The dataset's characteristics, and also the dataset's descriptions are shown in Figures 2 and 3.

A comparison between the proposed method and the KNN model is given in Tables 1 and 2. In addition, in Figure 4, we provide models evaluation based on AUC ROC curves. As is seen in Tables 1 and 2, and also in Figure 4, the proposed classification method outperforms compared to the KNN with an Accuracy of 97%, and an AUC equal to 1.00.

Table 1: Classification report for proposed method.

	precision	recall	F1 score
0	1.00	1.00	1.00
1	0.96	1.00	0.97
Accuracy	–	–	**0.97**
Weighted avg.	0.97	1.00	0.98

We have also verified the efficiency of the proposed method with partitioning on our data. Here, eight features are binary, and only four features SPO2, age, CRP, and temperature (TEM) are partitioned that

Feature	type	Description
Gender	nominal	This feature illustrates the patient's gender.
Age	numeric	The age range of the patients is 20–100 years.
Fever	nominal	Subjective symptom: 0 =no , 1= yes
Sore throat	nominal	Subjective symptom: 0 =no , 1= yes
Tiredness	nominal	Subjective symptom: 0 =no , 1= yes
Loss of smell	nominal	Subjective symptom: 0 =no , 1= yes
Dry cough	nominal	Subjective symptom: 0 =no , 1= yes
Chest pain	nominal	Subjective symptom: 0 =no , 1= yes
Shortness of breath	nominal	Subjective symptom: 0 =no , 1= yes
SPO2	numeric	Objective symptom: This feature captures patient's Saturation of Peripheral Oxygen
Temperature	numeric	Objective symptom: This feature captures patient's temperature
CRP	numeric	Objective symptom: This feature captures patient's C-Reactive-Protein

Figure 2: Dataset characteristics.

	count	mean	std	min	25%	50%	75%	max
Gender	611	0.53	0.49	0.00	0.00	1.00	1.00	1.00
Age	611	55.27	18.92	15.00	39.00	56.00	69.00	100.00
Fever	611	0.77	0.41	0.00	1.00	1.00	1.00	1.00
Sore throat	611	0.32	0.47	0.00	0.00	0.00	1.00	1.00
Tiredness	611	0.41	0.49	0.00	0.00	0.00	1.00	1.00
Loss of smell	611	0.09	0.29	0.00	0.00	0.00	0.00	1.00
Dry cough	611	0.48	0.50	0.00	0.00	0.00	1.00	1.00
Chest pain	611	0.29	0.45	0.00	0.00	0.00	1.00	1.00
Shortness of breath	611	0.50	0.50	0.00	0.00	1.00	1.00	1.00
SPO2	611	0.90	0.04	0.80	0.88	0.90	0.95	0.98
Temperature	611	38.41	0.58	36.40	38.10	38.40	38.70	39.80
CRP	611	1.77	1.11	0.00	1.00	2.00	3.00	3.00

Figure 3: Dataset description.

(see Figure 3). It is readily seen that there exist $3 \times 4 \times 4 \times 4 \times 2^8 = 3 \times 2^{14}$ cells that include the data points. We note that this information is saved on the system before.

In Table 4, it is seen that the proposed method with partitioning compared to the KNN method and proposed method without partitioning takes a little CPU time. All experiments run on a laptop with 8GB RAM, Intel Core i7-6500U, up to 3.1 GHz using python, version 3.8.5.

Table 2: Classification report for KNN.

	precision	recall	F1 score
0	0.80	0.94	0.86
1	0.98	0.91	0.94
Accuracy	–	–	**0.92**
Weighted avg	0.93	0.93	0.94

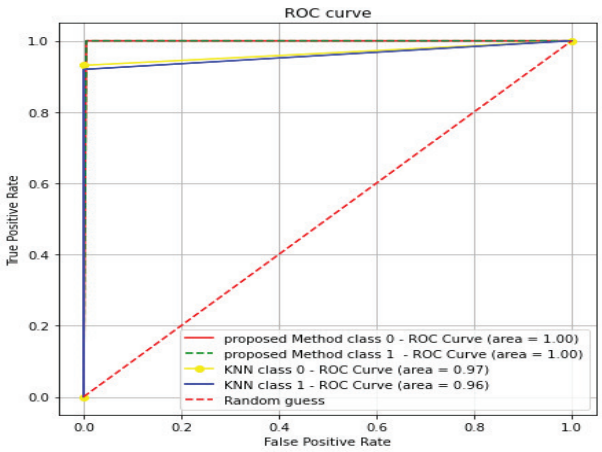

Figure 4: Models evaluation based on AUC ROC curves.

Table 3: Partitioning the features TEM, SPO2, PCR and AGE.

TEM	[37.3-37.9]	[38-38.9]	[39-39.9]	
SPO2	$< 85\%$	$86\% - 90\%$	$91\% - 94\%$	$> 95\%$
CRP	0	1	2	3
AGE	< 40	41-60	61-80	> 80

6 Conclusion

There are a lot of methods for prediction of disease outcomes. A combination of some methods can lead to a new method to improve the prediction results. In this paper, we used the idea of logistic regression and KNN methods that a better result was obtained. By applying the partitioning of features we decreased the total CPU time. The efficiency of our proposed method verified on 600 patients with suspected COVID-19.

Acknowledgements

The authors would like to thank the reviewers for their helpful comments and suggestions.

Table 4: CPU time & accuracy

Method	CPU time	Accuracy
KNN	2.89	0.92
Proposed method with partitioning	0.33	0.97
Proposed method without partitioning	2.62	0.97

References

[1] M. Batista, *Estimation of the final size of the coronavirus epidemic by the logistic model*, medRxiv, 2020.

[2] R. Bender, U. Grouven, *Ordinal logistic regression in medical research*, J. R. Coll. Physicians Lond. **31** (1997) 546.

[3] D.P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, Massachusetts, 1999.

[4] E. Dritsas, M.Trigka, *Supervised machine learning models for liver disease risk prediction*, Computers **12** (2023) 19.

[5] A. Govindu, S. Palwe, *Early detection of Parkinson's disease using machine learning*, Procedia Comput. Sci. **218** (2023) 249–261.

[6] G.S. Handelman, H.K. Kok, R.V. Chandra, A.H. Razavi, S. Huang, M. Brooks, H. Asadi, *Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods*, AJR Am J Roentgenol. **212** (2019) 38-43.

[7] A.E. Hoerl, R.W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics **12** (1970) 55–67.

[8] T. Huang, D. Le, L. Yuan, S. Xu, X. Peng, *Machine learning for prediction of in-hospital mortality in lung cancer patients admitted to intensive care unit*, PLOS one **18** (2023) e0280606.

[9] M.T. Huyut, *Automatic detection of severely and mildly infected COVID-19 patients with supervised machine learning models*, IRBM **44** (2023) 100725.

[10] C. Iwendi, C.G. Y. Huescas, C. Chakraborty, S. Mohan, *COVID-19 health analysis and prediction using machine learning algorithms for Mexico and Brazil patients*, J. Exp. Theor. Artif. Intell., 2022, https://doi.org/10.1080/0952813X.2022.2058097.

[11] P.M. Kumar, U.D. Gandhi, *A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases*, Comput. Electr. Eng. **65** (2018) 222-235.

[12] J.P. Li, A.U. Haq, S.U. Din, J. Khan, A. Khan, A. Saboor, *Heart disease identification method using machine learning classification in e-healthcare*, IEEE Access **8** (2020) 107562–107582.

[13] G. D. Magoulas, A. Prentza, *Machine Learning in Medical Applications*, In Advanced course on artificial intelligence, Springer, Berlin, Heidelberg, 1999.

[14] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Stat. Soc. B (Methodol.) **58** (1996) 267–288.

[15] G. Yu, Y. Liu, D. Shen, *Graph-guided joint prediction of class label and clinical scores for the Alzheimers disease*, Brain Struct. Funct. **221** (2016) 3787–3801.

[16] A. Zeb, E. Alzahrani, V.S. Erturk, G. Zaman, *Mathematical model for coronavirus disease 2019 (COVID-19) containing isolation class*, Biomed Res. Inter. **2020** (2020) 3452402.

[17] J. Zhou, A.H. Gandomi, F. Chen, A. Holzinger, *Evaluating the quality of machine learning explanations: A survey on methods and metrics*, Electronics **10** (2021) 593.

[18] X. Zhu, H.I. Suk, L. Wang, S.W. Lee, D. Shen, *Alzheimers Disease Neuroimaging Initiative. A novel relational regularization feature selection method for joint regression and classification in AD diagnosis*, Med. Image Anal. **38** (2017) 205–214.