Study of the Influence of Data Volume on the Quality of Regression to Restore the Distribution of Temperatures inside Tissue during Hyperthermia

Evgeny Kostyuchenko * and Elena Amletova

Posted Date: 29 January 2024

doi: 10.20944/preprints202401.1978.v1

Keywords: Hyperthermia, Regression, Data reduction, Decision Tree, Random Forest

Article

Study of the Influence of Data Volume on the Quality of Regression to Restore the Distribution of Temperatures Inside Tissue during Hyperthermia

Elena Amletova and Evgeny Kostyuhenko *

Tomsk State University of Control Systems and Radioelectronics, Tomsk, Russia

* Correspondence: key@fb.tusur.ru

Abstract: The use of hyperthermia is one of the effective methods of treating cancer. In this case, the problem arises of constructing and predicting the distribution of the thermal field inside tissues depending on their type. This paper discusses the use of linear and polynomial regression methods, as well as regression algorithms based on decision tree, random forest and K nearest neighbors methods to recover missing temperature values. The influence of the size of the training sample when performing regression on the quality of the reconstructed values is investigated. Experimental recommendations on sample size are given for the selected decision tree and random forest methods. The possibility of a significant (10 times or more) reduction in the initial sample size, without leading to a decrease in the R^2 metric by more than 0.05 for various tissues, has been shown.

Keywords: hyperthermia; regression; data reduction; decision tree; random forest

1. Introduction

It has been clinically proven that hyperthermia, an increase in tissue temperature to 39-44°C, increases the effectiveness of radiation therapy and chemotherapy in the fight against cancer [1]. For the use of hyperthermia in clinical settings, research into the use of this type of treatment has been carried out for 30 years. To determine the relevance of the use of magnetic hyperthermia, an analysis of scientific publications was carried out.

The authors [2] believe that the use of magnetic hyperthermia, known for more than 75 years, should be discussed for clinical use, since incorrectly selected parameters for the procedure can cause devastating harm to surrounding healthy tissues. A scientific article [3] states that there are cases with positive results from the use of hyperthermia, which can increase the survival rate of patients with glioma. Temperature is a key parameter for optimal functioning and growth of cells; changes in temperature can lead to cell death, so that healthy human cells do not suffer during treatment; the authors [4] propose their own method for measuring intracellular temperature, since they believe that treatment with hyperthermia is applicable in the field cancer therapy.

Regression analysis is used in many areas of human activity, and the medical field is no exception. Statistical data processing methods are key tools for analyzing theoretical, experimental and clinical observations. Medical statistics is aimed at developing specialized methods for studying general processes, identifying significant patterns and trends in the health status of patients. Regression analysis is one of the methods used to achieve these goals. This analysis belongs to the known methods of prediction, which in turn is a pressing issue in modern medicine. Forecasting affects both the results of targeted activities and the characteristics of the course of diseases.

Regression analysis is used in many areas of human activity, and the medical field is no exception. Statistical data processing methods are key tools for analyzing theoretical, experimental and clinical observations. Medical statistics is aimed at developing specialized methods for studying general processes, identifying significant patterns and trends in the health status of patients. Regression analysis is one of the methods used to achieve these goals. This analysis belongs to the

 doi:10.20944/preprints202401.1978.v1

known methods of prediction, which in turn is a pressing issue in modern medicine. Forecasting affects both the results of targeted activities and the characteristics of the course of diseases.

To select regression analysis models, scientific articles were reviewed where this method was used in the context of medical research. The authors of the study [5] conducted a binary logistic regression analysis to predict severe influenza in pregnant women. This forecast allows you to quickly organize treatment in order to reduce the level of complications. In the article [6], work was carried out to predict the likelihood of complications with tonsillitis. A logistic regression model using 10 clinical features to determine the likelihood of developing abscesses was studied and developed.

In medicine, each sign is considered significant, since indicators such as temperature, hemoglobin level, creatinine, and platelet count represent characteristics of a single organism. A change in any of these signs can lead to disruption of the functioning of a particular organ, which, in turn, can affect the functioning of other organs. The authors of the work [7] conducted a study to improve the effectiveness of treatment for patients diagnosed with acute pancreatitis. In their study, they used a ridge regression model, which is effectively used when processing large amounts of data and in the presence of multicollinearity of features.

Scientific work [8] examined the number of Covid-19 cases in Turkey using machine learning methods: Support Decision Regression (SVR), Linear Regression (LR), Bagged Tree Regression (BG), Decision Tree Regression Fine Tree (FT) . The best result was shown by the linear regression model; the coefficient of determination R was equal to 0.991, the average absolute error MAE was equal to 0.014, and the root of the root mean square error RMSE was equal to 0.017. The authors of the scientific work [9] developed a model based on the support vector regression (SVR) method to predict the values of forced vital capacity.

Because obtaining large volumes of individual temperature distribution values is difficult, then the question arises of the dependence of the accuracy of the obtained values on the number of bottom values used in constructing the regression model.

The purpose of this work is to select a regression model and assess its quality to solve the problem of restoring the temperature distribution for bone, fat and muscle tissue, depending on the amount of data used to build the regression model.

2. Materials and Methods

2.1. Dataset used

To carry out the work, a data set was selected from a scientific article entitled “Dataset from “In silico assessment of collateral eddy current heating in biocompatible implants subject to magnetic hyperthermia treatments” [10].

The authors of this paper conducted a study to evaluate the risk of thermal damage caused by heating of two types of prostheses in the treatment of three different diseases: colorectal cancer, prostate cancer, and head and neck cancer. When treating these diseases with magnetic hyperthermia, the jaw and pelvic area are affected, where hip implants may be located. For each case, two implant alloys were considered: Ti6Al4V and CoCrMo. At the same time, to determine safe conditions for the procedure, in addition to temperature, the specific absorption rate of electromagnetic energy (SAR) was calculated. Parameters such as exposure time to the tumor (5 and 30 minutes) and the maximum permissible temperature threshold (1 or 5 ºC) were also taken into account. The selected maximum permissible threshold of 5 ºC is based on research that shows that tissue heating of more than 5 ºC leads to adverse health effects. The 1ºC threshold is based on a more conservative approach, which takes into account the potential for adverse effects from even small changes in temperature. A set of values for the magnetic induction of the coil used for the experiment was considered: 5 and 15 mT. For all cases, the magnetic field frequency was 300 kHz, a value within the range intended for hyperthermia treatment. The heating of three types of tissues is analyzed: bone, fat and muscle.

 doi:10.20944/preprints202401.1978.v1

This work does not consider the possibility of the presence of implants that significantly reduce the quality of the resulting regression models; for this reason, a fragment related to "pure" human tissues is used.

2.2. Classifiers and quality metrics used

The study compared regression models based on linear and polynomial regression and K nearest neighbors, decision tree and random forest methods [11].

The quality of the constructed model was assessed using the following metrics [11]:

1. Determination coefficient R^2:

$$R^2 = 1 - \frac{\frac{1}{n}\Sigma_i^n (y - y_{pred})^2}{\frac{1}{n}\Sigma_i^n (y - \bar{y})^2}, \quad (1)$$

2. Mean absolute error (MAE):

$$MAE = \frac{1}{n}\Sigma_i^n \left| y - y_{pred} \right|, \quad (2)$$

3. Mean square error (MSE):

$$MSE = \frac{1}{n}\Sigma_i^n (y - y_{pred})^2, \quad (3)$$

4. Root of the mean square error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\Sigma_i^n (y - y_{pred})^2}, \quad (4)$$

where $\bar{y}$ is the average value of the target variable;

y – real values of the target variable;

y_{pred} – predicted values;

n – number of observations.

When building the models, 10-fold cross-validation was used. To ensure the significance of the data, experiments were carried out 20 times, and the results obtained were averaged. This approach allows us to ensure the reliability of the results.

3. Results

3.1. Subsection

At the first stage of research, models were built for the complete data set based on linear (LR) and polynomial regression (PR) and the K nearest neighbors (KNN), decision tree (DT) and random forest (RF) methods.

The obtained values for induction of 5 mT and 5 minutes of exposure for bone, fat and muscle tissue are given in Tables 1–3.

Table 1. Quality metrics for regression models, bone tissue, induction 5 mT, exposure 5 minutes.

Model	MAE	MSE	RMSE	R2
LR	0.006	0.0001	0.01	0.71
PR	0.004	0.00007	0.008	0.84
DT	0.003	0.00006	0.008	0.86

 doi:10.20944/preprints202401.1978.v1

KNN	0.006	0.00001	0.01	0.66
RF	0.004	0.00006	0.008	0.87

Table 2. Quality metrics for regression models, fat tissue, induction 5 mT, exposure 5 minutes.

Model	MAE	MSE	RMSE	R2
LR	0,01	0,0004	0,02	0,7
PR	0,007	0,0002	0,01	0,86
DT	0,006	0,0001	0,01	0,9
KNN	0,008	0,0002	0,01	0,87
RF	0,005	0,0001	0,01	0,92

Table 3. Quality metrics for regression models, muscle tissue, induction 5 mT, exposure 5 minutes.

Model	MAE	MSE	RMSE	R2
LR	0,003	0,00005	0,007	0,89
PR	0,002	0,00004	0,006	0,903
DT	0,0026	0,000039	0,0062	0,88
KNN	0,0064	0,00012	0,011	0,64
RF	0,0023	0,000034	0,0058	0,904

Since building a regression model takes time, and when conducting research on the dependence of the quality of the model on sample sizes, it is necessary to repeatedly build classifiers for each set of values, and also due to the significant scatter of quality metrics, it was decided to reduce the number of classifiers considered in the future to the two best.

For each of the tables, pairs of the worst (-) and best (+) classifiers were selected according to the R2 metric. This choice is justified by the fact that of the given metrics, only the coefficient of determination is normalized and has a limitation in the form of a maximum value equal to 1.

Table 4. Selecting two classifiers for further research.

Tissue	Bone	Fat	muscle
-	KNN, LR	LR, PR	KNN, DT
+	RF, DT	RF, DT	RF, PR

Based on the results of this comparison, for further study of the influence of the size of the training sample on the quality of regression, the Random Forest models (as consistently showing the best results) and the Decision Tree model were selected as twice included in the list of the best and only once in the list of the worst.

A similar analysis was carried out for other values of induction and heating time; the results confirmed the above choice of a pair of classifiers for further analysis.

3.2. The influence of the size of the training sample on the quality of regression

In the experiment, the models were trained on data sets for three types of tissues with varying data volumes. The quality of the model was assessed using ten-fold cross-validation, that is, the model was first trained on 9 parts of the data set, and then tested on the remaining part, which allows us to evaluate the accuracy of the model on independent data. This approach was repeated 20 times to reduce the degree of randomness in the obtained values. Tables 5–7 show the results of the models

 doi:10.20944/preprints202401.1978.v1

when changing the volume of the data set for bone, fat and muscle tissue, induction 5 mT, exposure time 5 minutes.

Table 5. Influence of data size on model performance, bone tissue, induction 5 mT, exposure 5 minutes.

Data set size	Decision Tree			Random Forest		
	MAE	RMSE	R2	MAE	RMSE	R2
10000	0,0038	0,0077	0,873	0,0035	0,0073	0,886
5000	0,0039	0,0079	0,861	0,0036	0,0075	0,874
1000	0,0043	0,0089	0,823	0,0040	0,0082	0,849
500	0,0043	0,0087	0,810	0,0039	0,0077	0,842
100	0,0050	0,0086	0,684	0,0044	0,0075	0,758
50	0,0066	0,0105	0,679	0,0051	0,0079	0,693
40	0,0047	0,0072	0,591	0,0051	0,0074	0,501
30	0,0068	0,0406	0,267	0,0051	0,0073	0,475

Table 6. Influence of data size on model performance, fat tissue, induction 5 mT, exposure 5 minutes.

Data set size	Decision Tree			Random Forest		
	MAE	RMSE	R2	MAE	RMSE	R2
10000	0,0062	0,0122	0,916	0,0054	0,0109	0,933
5000	0,0062	0,0121	0,915	0,0055	0,0111	0,929
1000	0,0063	0,0122	0,907	0,0056	0,0109	0,924
500	0,0068	0,0134	0,866	0,0060	0,0117	0,900
100	0,0095	0,0164	0,648	0,0076	0,0136	0,772
50	0,0091	0,0146	0,607	0,0077	0,0122	0,668
40	0,0109	0,0167	0,466	0,0109	0,0162	0,586
30	0,0097	0,0145	0,387	0,0074	0,0109	0,554

Table 7. Influence of data size on model performance, muscle tissue, induction 5 mT, exposure 5 minutes.

Data set size	Decision Tree			Random Forest		
	MAE	RMSE	R2	MAE	RMSE	R2
10000	0,0022	0,0055	0,913	0,0019	0,0052	0,922
5000	0,0023	0,0060	0,887	0,0020	0,0055	0,907
1000	0,0026	0,0065	0,870	0,0024	0,0060	0,891
500	0,0030	0,0070	0,840	0,0025	0,0061	0,883
100	0,0038	0,0068	0,719	0,0031	0,0059	0,818
50	0,0034	0,0062	0,704	0,0032	0,0057	0,688
40	0,0031	0,0046	0,689	0,0028	0,0044	0,766
30	0,0062	0,0095	0,356	0,0029	0,0041	0,131

 doi:10.20944/preprints202401.1978.v1

4. Discussion and Conclusions

By analyzing the obtained values, the following conclusions can be drawn.

Reducing the number of temperature measurement points leads to a decrease in the quality of the models. This conclusion is quite obvious and indicates that there are no contradictions between the experimental results and logic.

In experiments without implants, it is possible to reduce the volume of data by a factor of 10 for bone and fat tissue, and in some cases for muscle tissue by a factor of 20. This change in data size allows you to maintain the quality of the model, since the value of the average determination coefficient decreases by no more than 0.05 compared to the results of training and testing models on an initial sample of 10,000 examples.

As a result of the study, regression models were built, trained on different data sets, representing the results of measuring the temperature of three types of human tissue (bone, fat, muscle) during the treatment of colorectal cancer tumors with magnetic hyperthermia without the use of implants.

In experiments without implants for all types of tissues, it is better to use the Random Forest model, which shows excellent results when reducing the number of temperature measurements by about 10-20 times, depending on the type of tissue. The quality of the model does not deteriorate compared to estimates obtained on the initial sample size of 10,000 examples, and the value of the coefficient of determination decreases by no more than 0.05.

It should be noted that preliminary testing of the same models for fragments of the data set containing values for tissues with installed implants gave unsatisfactory results in terms of regression accuracy; the construction of adequate models and the study of their behavior for such cases remains a topic for further research.

Author Contributions: Examples of such statement(s) are shown below:

Funding: This research was funded by the Ministry of Science and Higher Education of the Russian Federation within the framework of scientific projects carried out by teams of research laboratories of educational institutions of higher education subordinate to the Ministry of Science and Higher Education of the Russian Federation, project number FEWM-2020–0042.

Data Availability: The analyzed data was taken from a data set that is publicly available at https://figshare.com/articles/by-resource-doi/10.1080/02656736.2021.1909758 and described in [10].

Declarations

- The authors have no relevant financial or non-financial interests to disclose.
- The authors have no competing interests to declare that are relevant to the content of this article.
- All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.
- The authors have no financial or proprietary interests in any material discussed in this article.

Free text: All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by [full name], [full name] and [full name]. The first draft of the manuscript was written by [full name] and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Example: CRediT taxonomy:

Conceptualization: [Evgeny Kostyuchenko]; Methodology: [Evgeny Kostyuchenko]; Formal analysis and investigation: [Elena Amletova]; Writing - original draft preparation: [Elena Amletova]; Writing - review and editing: [Evgeny Kostyuchenko]; Funding acquisition: [Evgeny Kostyuchenko]; Resources: [Evgeny Kostyuchenko]; Supervision: [Evgeny Kostyuchenko].

References

1. Van der Zee, J. (2002). Heating the patient: a promising approach?. Annals of oncology, 13(8), 1173-1184.

 doi:10.20944/preprints202401.1978.v1

2. Chang, D., Lim, M., Goos, J. A., Qiao, R., Ng, Y. Y., Mansfeld, F. M., ... & Kavallaris, M. (2018). Biologically targeted magnetic hyperthermia: Potential and limitations. Frontiers in pharmacology, 9, 831.
3. Rivera, D., Schupper, A. J., Bouras, A., Anastasiadou, M., Kleinberg, L., Kraitchman, D. L., ... & Hadjipanayis, C. G. (2023). Neurosurgical Applications of Magnetic Hyperthermia Therapy. Neurosurgery Clinics, 34(2), 269-283.
4. Silva, P. L., Savchuk, O. A., Gallo, J., García-Hevia, L., Bañobre-López, M., & Nieder, J. B. (2020). Mapping intracellular thermal response of cancer cells to magnetic hyperthermia treatment. Nanoscale, 12(42), 21647-21656.
5. Tarbaeva, D. A., Belokrinitskaya, T. E., & Serkin, D. M. (2019). Binary logistic regression in predicting severe forms of influenza in pregnant women. Siberian Medical Review, (4 (118)), 113-116.
6. Yastremsky, A. P., Izvin, A. I., Sannikov, A. G., & Zakharov, S. D. (2019). Prediction of the probability of developing peritonsillar abscess based on the logistic regression method. Russian otorhinolaryngology, 18(2 (99)), 95-102.
7. Cherdantsev, D. V., Stroev, A. V., Mangalova, E. S., Kononova, N. V., & Chubarova, O. V. (2019). Using ridge regression to assess the severity of acute pancreatitis. Bulletin of Siberian Medicine, 18(3), 107-115.
8. Atik, I. (2022). Performance comparison of regression learning methods: Covid-19 case prediction for turkey. Int. J. Mech. Eng, 7, 6297-6308.
9. Wang, C., Chen, X., Zhao, R., He, Z., Zhao, Z., Zhan, Q., ... & Fang, Z. (2019). Predicting forced vital capacity (FVC) using support vector regression (SVR). Physiological measurement, 40(2), 025010.
10. Rubia-Rodríguez, I., Zilberti, L., Arduino, A., Bottauscio, O., Chiampi, M., & Ortega, D. (2021). In silico assessment of collateral eddy current heating in biocompatible implants subjected to magnetic hyperthermia treatments. International Journal of Hyperthermia, 38(1), 846-861.
11. Scikit-learn Machine learning in Python Retrieved September 15, 2023, from https://scikit-learn.ru