International Journal of Advanced Research in
IT and Engineering
ISSN: 2278-6244

APPLICATION OF NUMERICAL METHOD BASED ON INTERPOLATION FOR DETECTION OF BREAST CANCER

Navid Samimi*

Hedayat Bahadori*

Shabnam Azari*

Abstract: *Breast cancer is one of the leading causes of death among women all around the world. Therefore, true and early diagnosis of breast cancer is an important problem. The Lagrange interpolation was used in this study for the diagnosis of breast cancer. The Wisconsin Breast Cancer dataset (WBCD), derived from the University of California Irvine machine learning database, was used for the purpose of testing the proposed method was determined as 98%. In this method, we are looking for two equations by Lagrange interpolation because of two classes that are exist in breast cancer data set. We receive each sample and it will define rate of vicinity it's to each class (malignant or benign). So, the class of mentioned sample will be recognized. Moreover, the most appropriate attributes for the diagnosis of breast cancer were determined from the WBCD in this study. It is considered that the proposed method will be useful in similar medical practices.*

***Key words**: Breast cancer, Lagrange interpolation, extreme learning machine, expert system.*

---

*Sama Technical and Vocational Training College, Islamic Azad University, Omidiyeh Branch, Omidiyeh, Iran

1- INTRODUCTION

The breast is an appendage of the skin covering the external part of our body and it includes lactating glands. Breast cancer is defined as the existence of cells progressing abnormally within the tissue of the breast that cannot be controlled. A group of cells growing or changing abnormally is called a tumor. Any tumor may be benign (not dangerous) or malignant (having the potential for being dangerous)

Breast cancer is the most common cancer in women in many countries. Most breast cancers are detected as a lump/mass on the breast, or through self-examination/mammography [1]. Screening mammography is the best tool available for detecting cancerous lesions before clinical symptoms appear. Surgery through a biopsy or lumpectomy have been also been the most common methods of removal. Fine needle aspiration (FNA) of breast masses is a cost-effective, non-traumatic, and mostly invasive diagnostic test that obtains information needed to evaluate malignancy. Recently, a new less invasive technique, which uses super-cooled nitrogen to freeze and shrink a non-cancerous tumor and destroys the blood vessels feeding the growth of the tumor, has been developed [2] in the USA. Several Artificial Intelligence (AI) techniques including neural networks and fuzzy logic [3] are successfully applied to a wide variety of decision-making problems in the area of medical diagnosis. In this paper we examine the performance of Numerical method based on Lagrange interpolation for detection of breast cancer data.

The Wisconsin Breast Cancer datasets (WBCDs), derived from the University of California Irvine (UCI) machine learning database, were used for the purpose of testing the proposed model. As a result, we are of opinion that the proposed model will be a tool for assisting specialists in making decisions with respect to the patients at the final stage. This paper is organized as follows. Section 2 describes the characteristics of breast cancer data set. Section 3 describes prior studies. Section 4 provides details of the proposed method. The final section provides some conclusions relating to the performance of Lagrange Interpolation when applied to the breast cancer data.

2- DATASET

The WBCD, the dataset used in this study, was derived from the UCI machine learning database [4]. The dataset consists of 699 samples that were collected by Dr.Wolberg at the University of Wisconsin-Madison hospitals. A total of 16 instances were discarded from the dataset since they had missing observations and the RS + EML model was tested with the remaining 683 cases. The WBCD consists of 9 features and the values thereof range between 1 and 10. The target attribute was coded as benign (1 = benign) and malignant (0 = malignant). There are 444 benign cases and 239 malignant cases in the dataset. The attributes available in the dataset are detailed in Table 1.

Table 1. WBC data description of the attributes

feature	Code	Domain	Mean	Standard deviation
Clump Thickness	A1	1-10	4.44	2.83
Uniformity of cell size	A2	1-10	3.15	3.07
Uniformity of cell shape	A3	1-10	3.22	2.99
Marginal adhesion	A4	1-10	2.83	2.86
Single epithelial cell size	A5	1-10	2.23	2.22
Bare nucleoli	A6	1-10	3.54	3.64
Bland chromatin	A7	1-10	3.45	2.54
Normal nucleoli	A8	1-10	2.87	3.05
Mitoses	A9	1-10	1.60	1.73

3. STUDIES FOR THE DIAGNOSIS OF BREAST CANCER

When the performed studies are examined, it is observed that the machine learning studies carried out using the WBCD are widespread and that high success rates were achieved in all of these performed studies.

Ster and Dobnikar [5] achieved a classification success rate of 96.80% using linear discriminate analysis. Pena-Reyes and Sipper [6] achieved a classification success rate of 97.36% in the study that they performed using a hybrid model based on fuzzy logic and the genetic algorithm (GA).

The classification success rate achieved in the study by Setiono was 98.10%. Abonyi and Szeifert achieved a classification success rate of 95.57%.

Using the controlled fuzzy set method. A classification success rate of 96.66% was achieved in the study by Kim et al. using a fuzzy rule-based method. Sahan et al. [1] achieved a success rate of 99.14% using a hybrid model based on fuzzy artificial immunity and K-nearest neighbor in their studies. Polat and G�unes [2] achieved a success rate of 98.53% by least squares support vector machine (LS-SVM) in their studies.

4. LAGRANGE INTERPOLATION AND PRE-CALCULUS MATHEMATICS

Just as two points determine a line, three (non-collinear) points determine a unique quadratic function, four points that do not lie on a lower degree polynomial curve determine a cubic function and, in general, $n+1$ points uniquely determine a polynomial of degree *n*, presuming that they do not fall onto a polynomial of lower degree. The process of finding such a polynomial is called *interpolation* and the two most important approaches used are Newton's interpolating formula and Lagrange's interpolating formula. Each has its own advantages and disadvantages, as we will discuss. In this article, we show how both approaches can be introduced and developed at the precalculus level in the context of fitting polynomials to data.

The major drawback to Newton's interpolation formula is the fact that it requires uniform spacing for the *x*-values in a set of data. An alternative approach is Lagrange's interpolation formula, which does not require uniform spacing. But, it carries with it a cost – it is a more complicated formula that usually involves considerably more computational effort.

Suppose that we have three points (x_0, y_0), (x_1, y_1) and (x_2, y_2), where all of the x_i are different. The observation in the previous paragraph suggests that we may break up the quadratic function $f(x)$ into three quadratic functions $f_0(x)$, $f_1(x)$ and $f_2(x)$ by looking at the values of the function at x_0, x_1 and x_2 in the following way:

$$
\begin{array}{rcccccccl}
f(x_0) = & \boxed{\begin{array}{c} y_0 \\ 0 \\ 0 \end{array}} & + & \boxed{\begin{array}{c} 0 \\ y_1 \\ 0 \end{array}} & + & \boxed{\begin{array}{c} 0 \\ 0 \\ y_2 \end{array}} & & = y_0 \\
f(x_1) = & & & & & & & = y_1 \\
f(x_2) = & & & & & & & = y_2 \\
& \Downarrow & & \Downarrow & & \Downarrow & & \\
f(x) = & f_0(x) & + & f_1(x) & + & f_2(x) & &
\end{array}
$$

Look at the above display vertically. We pair each of the three numbers y_0, 0, and 0 in the first column with x_0, x_1, and x_2, respectively, to obtain the three points (x_0, y_0), $(x_1, 0)$, and $(x_2, 0)$. These points define the quadratic function $f_0(x) = y_0 \cdot \dfrac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}$, by formula (1). Similarly, the function $f_1(x)$ is defined by the three points $(x_0, 0)$, (x_1, y_1), and $(x_2, 0)$, while $f_2(x)$ is defined by $(x_0, 0)$, $(x_1, 0)$, and (x_2, y_2). The expressions for these two functions are

$$f_1(x) = y_1 \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} \text{ and } f_2(x) = y_2 \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}.$$

The sum of $f_0(x)$, $f_1(x)$ and $f_2(x)$ is the desired function that passes through the original three points (x_0, y_0), (x_1, y_1), and (x_2, y_2).

More generally, the quadratic Lagrange interpolating polynomial that passes through the points (x_0, y_0), (x_1, y_1) and (x_2, y_2) is

$$L_2(x) = y_0 \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} + y_1 \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} + y_2 \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}.$$

The ideas discussed here can be extended if there are more than three given points. For instance, the four points (x_0, y_0), (x_1, y_1), (x_2, y_2), and (x_3, y_3) determine a cubic polynomial (assuming that all the x_i are distinct and the points do not lie on a line or a parabola). A simple extension of the Lagrange interpolation formula used above gives a simple way to construct this cubic. The third degree Lagrange interpolating polynomial $L_3(x)$ is composed of four cubic components $f_0(x)$, $f_1(x)$ $f_2(x)$ and $f_3(x)$, each constructed in the comparable way. The result is

$$\begin{aligned} L_3(x) = {} & y_0 \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)} + y_1 \frac{(x - x_0)(x - x_2)(x - x_3)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)} \\ & + y_2 \frac{(x - x_0)(x - x_1)(x - x_3)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)} + y_3 \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)}. \end{aligned}$$

Notice that, at each of the four interpolating points, only one of the four cubic components is not automatically zero and so contributes precisely the associated value of *y* at each of those points. The other three cubic components must contribute zero at these points. For instance, at $x = x_0$, only the first component is non-zero and it contributes y_0 to the sum. That is, $L_3(x_0) = y_0$.

The equation of the parabola that passes through the three points $(1,2)$, $(3,8)$ and $(6,4)$ is therefore

$$f(x) = 2\frac{(x-3)(x-6)}{(1-3)(1-6)} + 8\frac{(x-1)(x-6)}{(3-1)(3-6)} + 4\frac{(x-1)(x-3)}{(6-1)(6-3)}$$

or

$$f(x) = \frac{1}{5}(x-3)(x-6) - \frac{4}{3}(x-1)(x-6) + \frac{4}{15}(x-1)(x-3) = -\frac{13}{15}x^2 + \frac{97}{15}x - \frac{18}{5}.$$

Figure 7 shows $f(x)$ along with its three component functions $f_0(x)$, $f_1(x)$ and $f_2(x)$. Let's focus on the three solid dots that represent the given points. First, consider the point $(1,2)$. Notice that only the graphs of $f(x)$ and $f_0(x)$ pass through this point while the other two curves go through the *x*-intercept at $x=1$. Similarly, only the graphs of $f(x)$ and $f_1(x)$ meet at $(3,8)$, while the other two curves have the *x*-intercept at $x=3$. Finally, only the graphs of $f(x)$ and $f_2(x)$ intersect at $(6,4)$, while the other two share the *x*-intercept at $x=6$.

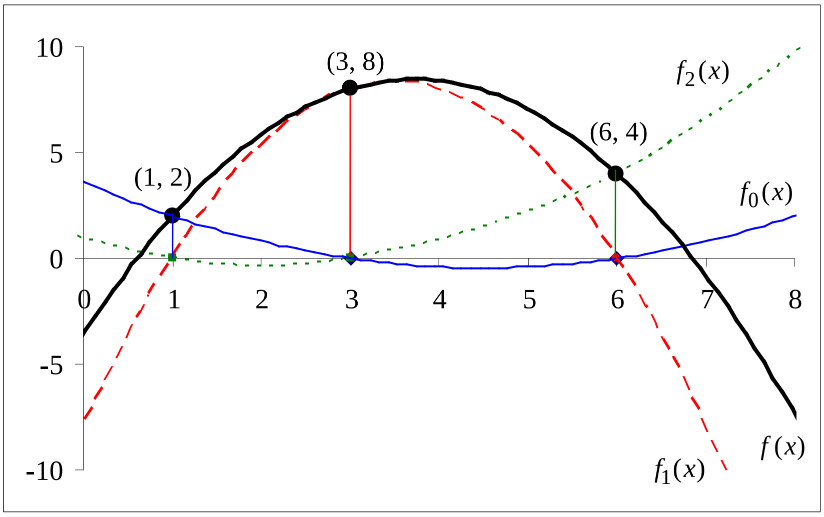

Figure 7: The three component quadratic functions of $f(x)$

$f_0(x) = \frac{1}{5}(x-3)(x-6)$

$f_1(x) = -\frac{4}{3}(x-1)(x-6)$

$f_2(x) = \frac{4}{15}(x-1)(x-3)$

$f(x) = f_0(x) + f_1(x) + f_2(x)$

This formula

$$f(x) = 2\frac{(x-3)(x-6)}{(1-3)(1-6)} + 8\frac{(x-1)(x-6)}{(3-1)(3-6)} + 4\frac{(x-1)(x-3)}{(6-1)(6-3)}$$

is an example of the quadratic *Lagrange interpolating formula* and is named after Joseph Louis Lagrange, a famous Italian-born French mathematician of the 18th century [1]. More generally, the quadratic Lagrange interpolating polynomial that passes through the points (x_0, y_0), (x_1, y_1), and (x_2, y_2) is

$$L_2(x) = y_0\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + y_1\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + y_2\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}.$$

The ideas discussed here can be extended if there are more than three given points. For instance, the four points (x_0, y_0), (x_1, y_1), (x_2, y_2) and (x_3, y_3) determine a cubic polynomial (assuming that all the x_i are distinct and the points do not lie on a line or a parabola). A simple extension of the Lagrange interpolation formula used above gives a simple way to construct this cubic. The third degree Lagrange interpolating polynomial, $L_3(x)$, is composed of four cubic components $f_0(x)$, $f_1(x)$, $f_2(x)$ and $f_3(x)$, each constructed in the comparable way. The result is

$$\begin{aligned} L_3(x) = y_0 \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + y_1 \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \\ + y_2 \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + y_3 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}. \end{aligned}$$

Notice that, at each of the four interpolating points, only one of the four cubic components is not automatically zero and so contributes precisely the associated value of *y* at each of those points. The other three cubic components must contribute zero at these points. For instance, at $x = x_0$, only the first component is non-zero and it contributes y_0 to the sum. That is, $L_3(x_0) = y_0$.

The authors also provide (URL here) an interactive spreadsheet to investigate graphically and numerically the way in which the linear, quadratic, and cubic Lagrange polynomials are constructed out of their component functions for user-defined sets of data [7].

5. CONCLUSION

In this paper, we examined the performance of Numerical method based on Lagrange interpolation with no time-consuming tuning procedures. We created two equations by Lagrange interpolation. We could recognize class of each sample by 10-fold crossing method. Rate of recognize was about 98%. This amount of rate is improvable by some preprocessing algorithms. Of course, we did not use these methods. We replaced value 1 instead of missing values.

REFERENCES

[1] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine. 2005 Jun; 34(2):113-27.

[2] A.Shukla,R.Tiwari and P.Kapur, "Knowledge based approach for diagnosis of breast cancer", Advance Computing Conference ,2009. IACC, 6 & 7March 2009. E-ISBN:978-1-4244-2928-8, published in IEEE xplore digital library.

[3] Santi Wulan Purnami, S.P. Rahayu and Abdullah Embong, “Feature selection and classification of breast cancer diagnosis based on support vector machine”, IEEE 2008.

[4] Farzaneh Keivanfard , Mohammad Teshnehlab , Mahdi Aliyari Shoorehdeli , “Feature Selection and Classification of Breast Cancer on Dynamic Magnetic Resonance Imaging by Using Artificial Neural Networks”, Proceedings of the 17th Iranian Conference of Biomedical Engineering (ICBME2010), 3-4 November 2010.

[5] D.S. O.L. Mangasarian, W.N. Street and W.H. Wolberg, Breast cancer diagnosis and prognosis via Linear programming, Operations Research, 43(4), pages 570-577, July-August 1995.

[6] Sariego, J., "Breast cancer in the young patient". The American surgeon 76 (12): 1397–1401, 2010.

[7] W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian, Computerized breast cancer Diagnosis and prognosis from fine needle aspirates. Archives of Surgery 1995; 130:511-516.