УДК 519.2: 004.6 DOI: https://doi.org/10.30837/IYF.IIS.2024.402

SUPPORT VECTORS MACHINE METHOD AND ITS SOFTWARE IMPLEMENTATION FOR DETECTING CANCER DISEASES

Gorishnia K. O.
Scientific Supervisor – Cand. Sc. (Techn.), assoc. prof. Kobziev V. G.
Kharkiv National University of Radio Electronics, dept. of Software Engineering, Kharkiv, Ukraine
e-mail: kateryna.horishnia@nure.ua

The application of the method of support vectors for the classification of cells on tomographic images of human organs with signs of oncological diseases is considered. The equation of the dividing hyperplane and the width of the border between two reference vectors, which allow classification of cells into three groups corresponding to different types of diagnosis, including in the early stages of the disease, are given. The Soft Margin SVM was chosen for the software implementation of the considered method.

Processing the accumulated results of medical research using Big Data methods allows to identify existing patterns in them, which are related to the course of diseases [1]. For many serious diseases, modern methods of tomographic studies are used, the result of which are sets of various images of organs with signs of diseases. Classification of such images based on formal rules provides medical personnel with a basis for establishing a diagnosis of a serious illness, as well as its stage.

We will focus on the analysis of the elements of tomographic images (their number, size, location, etc.) of human organs with various signs of oncological diseases, as one of the types of serious diseases. The relevance of statistical analysis of such information is explained by the following points.

1. Oncological diseases are one of the main causes of people's death in the world.

2. Data related to oncological diseases have a large volume of information, the use of methods for detecting regularities in which can help in identifying characteristic signs indicating the development of cancer.

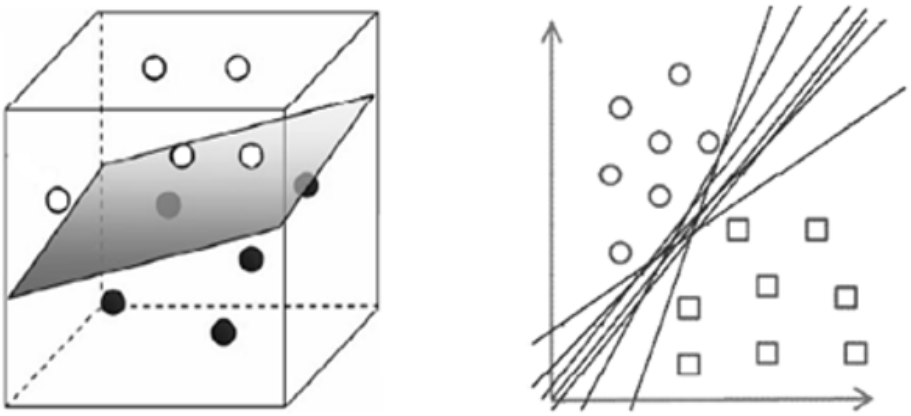
3. Identified regularities or anomalies help in delineating the characteristics of the patient, which determine an effective method of treatment.

Among the existing classification methods, it's need to choose a method that divides cells into healthy and diseased cells and most accurately takes into account the specified properties to identify hidden patterns of the location of diseased cells in the images in the early stages of the disease.

Solving these problems is based on fixing the location of specific elements of the image and classifying the situations corresponding to them. Cancer cells can be located inside the lining of certain organs, on the borders of such a lining or outside it.

The simplest option is observed when the shell in some area can be considered as a hyperplane. Cells of interest to the researcher are located in a

certain way around it. A conventional example of such a hyperplane (a plane in three-dimensional space) and its location in two-dimensional space is shown in Fig. 1. Cells are displayed by points with specific coordinates.



a) b)

Figure 1 – Separating hyperplane.
a) three-dimensional space, b) two-dimensional space

Among the huge set of existing ones, it is necessary to choose such a classification method, which, based on the training population, will choose the optimal option for the orientation of such a hyperplane in space (one of the possible lines in Fig. 1.b)), as well as predict a certain space between hyperplanes parallel to the chosen one, for a guaranteed distribution of cells on both sides of it.

The method of support vector machine (SVM) [2] best meets the stated conditions. This method for the training set, which is given by the set of vectors (x_1, x_2, ... , x_n) in the hyperspace R^k, allows you to determine the equation of the separating hyperplane H as follows:

$$\boldsymbol{w} \bullet \boldsymbol{x} + b = 0, \quad (1)$$

where $\boldsymbol{w}$ is a vector normal to the hyperplane, b is some constant.

In SVM the weights $\boldsymbol{w}$ and b are adjusted so that the class objects lie as far as possible from the separating hyperplane. This method maximizes the margin between the hyperplane and the class objects that are closest to it. Such objects are called support vectors.

The distance from the origin to the hyperplane is $|b|/||\boldsymbol{w}||$, where the denominator means the Euclidean norm (length) of the vector $\boldsymbol{w}$. The two hyperplanes H_1 and H_2 passing through support vectors are parallel to the hyperplane H. The width of the boundary (margin) separating the two populations (the distance between H_1 and H_2) will be equal to $2/||\boldsymbol{w}||$.

For a two-dimensional problem, the linear separation of points is presented in Fig. 2.

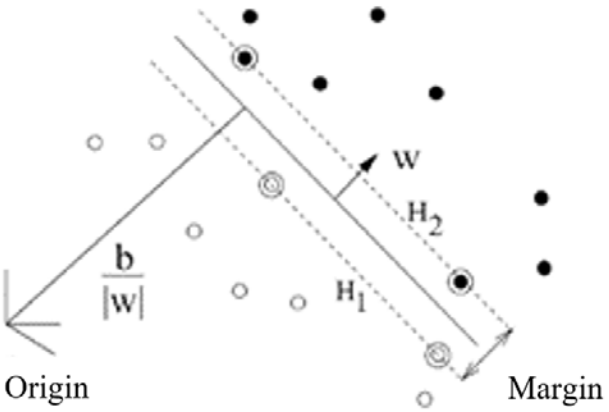


Figure 2 – Linear separation of points on the plane by SVM

At the stage of using the constructed classifier, the found boundaries will play important roles. The following options are possible for the cells of the test population (points on the test image): the location of points only inside the shell (from the origin of coordinates to H1), inside the shell and in margin (up to H2) and the location of a certain number of points outside the shell (hereinafter H2), which will mean a mild, a transitional and a severe disease state, respectively.

To reduce possible linear classification errors, a modification of the SVM associated with the so-called soft boundary is used. But considering the use of the support vector method for the classification of medical studies [3], it is important to focus on minimizing errors and matching the model with real data. This includes optimizing parameters, considering representativeness of data, adapting to changes in the population, and interpreting results for confidence and practicality in medical diagnostics.

Thus, the given classification method separates healthy and diseased cells into two groups with a certain margin in order to reveal hidden patterns in the early stages of the disease and distinguish three important diagnostic situations. Practical application of the described method is supported by program tool Soft Margin SVM.

References:

1. Choong, H.-J.Y., Lee, Choong Ho. "Medical big data: promise and challenges." Kidney Res. Clin. Pract., volume 36, № 1, pp. 3–11, April 2017. DOI: 10.23876/j.krcp.2017.36.1.3.

2. Steinwart, Ingo, i Christmann, Andreas. Support Vector Machines. New York: Springer-Verlag, 2008.

3. Gorishnia K., Kobziev V. The method of support vectors machine for detecting and classification of serious diseases in medical research. In: Computer science, information technologies and management systems. Proceedings of the International Scientific Young Scientists Conference, 2023, December, 21-22, Ivano-Frankivsk. Vasyl Stefanyk Precarpathian National University, 2023. pp. 137-139.