Agnieszka DURAJ[1], Ewa KORZENIEWSKA[2], Andrzej KRAWCZYK[3,4]

Institute of Information Technology and Applied Mathematics, Lodz University of Technology (1),
Institute of Electrical Engineering Systems, Lodz University of Technology (2),
Electrical Faculty, Czestochowa University of Technology (3),
Military Institute of Medicine, Warsaw (4),

doi:10.15199/48.2015.12.19

Classification algorithms to identify changes in resistance.

Streszczenie. *W artykule została opisana podstawowa metoda pomiaru rezystancji elektrod medycznych wykonanych w oparciu o cienkie warstwy przewodzące powstałe w procesie PVD. Scharakteryzowano również krótko dwa algorytmy klasyfikacji danych: algorytm k najbliższych sąsiadów oraz klasyfikator bayowski, które zostały wykorzystane jako algorytmy identyfikacji zmian rezystancji elektrod.*

Abstract: *In the article the basic method for measuring the resistance of medical electrodes, made based on a thin conductive layer formed during the PVD process, is described. The authors also briefly characterized two algorithms for data classification: k-nearest neighbors and Bayes classifier, which were used as algorithms to detect changes in the electrode resistance.* ***(Zastosowanie algorytmów klasyfikacji danych do identyfikacji zmian rezystancji)***

Słowa kluczowe: k-NN, naiwny klasyfikator Bayesa, rezystancja warstw cienkich, wykrywanie wyjątków
Keywords: k-NN, naive Bayes classifier, resistance of thin layers, detect exceptions

Introduction

Every cognitive process associated with information processing is an innate ability of living organisms. Using all the senses (sight, hearing, taste and touch) allows not only the processing of data analyzed by the brain, but also enables more efficient inference and taking specific decisions. It is therefore essential to analyse large data sets using artificial intelligence techniques.

The authors of this study analyze data based on data classification algorithms. A set of measurement data was collected online while measuring the resistance of sensors made in the technology of vacuum application of PVD on flexible textile substrates. The surface resistance of the sensors is in the order of magnitude of Ω/square. While using the sensors, the important factors are: the ability to detect quickly local changes in the resistance over their entire surface and minimizing the levels of occurrence of errors and interference during signal transmission. This article presents the results of work related to the classification of surface resistance of the sensors as a method to identify changes in condition of the analyzed object.

The term classification can be understood as a method of data analysis the aim of which is to predict the value of a specific attribute based on a set of reference data. It covers not only the method of classifiers or functions describing the relationship between the properties of objects and their associated classifications, but also the classification models that are used to sort of new objects previously unsystematized. In the literature there are many different types of classification models such as k-nearest neighbor algorithm (k-NN) probabilistic classifiers (Bayesian classifier), decision trees, support vector method SVM, neural networks, metaheuristics (genetic algorithms), and the like.

In this paper, the authors described the use of two types of algorithms for data classification: k-nearest neighbor k-NN and naive Bayesian classifier. They have been used by the authors to identify changes in local surface resistance of the sensors as an element of diagnostic systems [1]. The study is a continuation of previous work aimed at detecting changes in the resistance thin film electrodes using the DBSCAN algorithm [2-7].

Identification of resistance – measurements

In the paper there are presented the results of measurements of the resistance of thin film electrodes produced in the vacuum deposition process. The selection of resistance measurement method was primarily due to the shape of the tested electrodes. Due to the small size of the created electrodes a method suggested in PN-EN 61340-2-3 describing the research methods used for determining the resistance and resistivity of solid planar materials - could not be used during this work [8]. Resistance measurements were made in the four-probe system utilizing the information processing system connected to the computer, which conducted the measurements at half-minute intervals.

The electrodes are connected with thin layer by electrically conductive glue. Before making the measurements, samples were conditioned for 24 hours in the following conditions: temperature 23°C and 55% relative humidity.

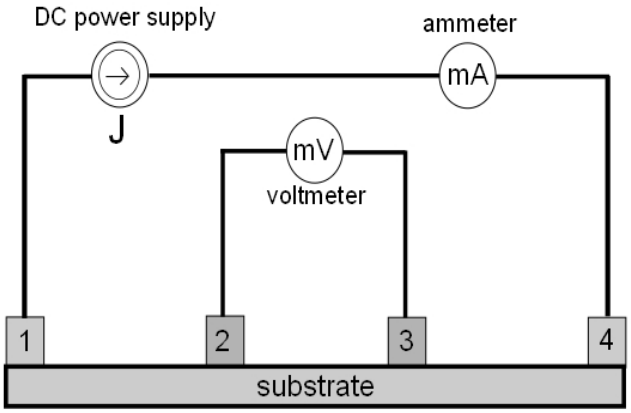

Fig.1. Four-probe system for resistance of the created electrodes measuring 1,4 – the current electrodes, 2,3 - the voltage electrodes

The resistance R which was measured in the system presented in figure 4 was calculated with formula (1):

$$R = \frac{1}{2}\left(\left|\frac{U_{2-3}}{I_{1-4}}\right| + \left|\frac{U_{2-3}}{I_{4-1}}\right|\right) \quad (1)$$

where R - measured resistance, U - measured voltage drop between the electrodes 2 and 3, I - current between the electrodes 1 and 4.

Proper substrate preparation for layer application is the key issue for the receipt of relevant electrical parameters of thin layers. [9] The resistance change may also be influenced by many external factors, eg. magnetic field

around the devices for magnetotherapy in the case of application of electrically conductive thin films as electrodes in medical systems [10,11].

The processes connected to the electromagnetic calculation and optimization of electronics equipment including obtaining respective values of resistance, are complex. In many of these processes for example parallel algorithms or multicore computers can be used [12-17]. In the case of data identification acquired from the electrodes to monitor vital functions, algorithms of exceptions detection can be filed and do not have to be complicated.

Classification of resistance changes

The study used two very popular algorithms for data classification, namely: the k - nearest neighbors algorithm and Bayesian classifier. The k-nearest neighbor algorithm is one of the simplest and yet the most frequently used data classifier. It is based on the following principle, "data pattern belongs to the class to which the highest number of its k-neighbouring patterns belong. The classification of the model is thus a determination of k samples from the training set that is closest in the context of applied metrics (eg. Euclidean, Manhattan, Hamming's). Then, we determine the class to which belongs the analyzed data sample. The methods of voting used in this case in are Majority Voting (on equal rights), or Inverse Distance Voting (method which takes distance into account.)

The classification model created in this way is an example of the so-called learning lazy, because it waits for the input parameters and does not generalize the adopted training data. It can therefore solve many problems simultaneously.

The rule of k nearest neighbors can be briefly presented as:

Let $k \in N$, $x_1, x_2, \ldots, x_n$ be objects in the analyzed a set classified into classes $C = \{C_1, C_2, \ldots, C_m\}$. Then the object x_{n+1} (observation) belongs to this class C_i to which belong the highest number of its k neighboring objects.

Therefore, the observation x is classified to the class to which most of the k observations from the training sample nearest to the point x belong. The more k points are compared, the greater the chance that the result will be accurate, but on the other hand, the number k should not be too large. The algorithm of the k - nearest neighbor in pseudo-code is shown in Listing 1.

Listing 1 Pseudo-code of the k nearest neighbor algorithm:

```
Measure the distance of the point which is to be categorized
to the points from the training set;
Sort the array of points by distance;
For k neighbors check one by one the training points,
        starting from the ones located closest to the point
        to the classification;
        Check to which groups the point belongs;
        If the group exists,
                Increase the number of occurrences of
                points in this category
        Otherwise:
                Add a new category and increase the
                number of occurrences;
Next, sort the list of categories according to occurrences;
Check if there are multiple categories having the most
occurrences with the same number of occurrences;
        If there are at least two such groups
        Display a prompt informing of the need to increase
        or decrease the amount of neighbors (k) by 1, in
        order to obtain an accurate result
    Otherwise:
        Classify the point to the group with the highest
        number of occurrences.
```

The most frequently used measures of distance are the Euclidean distance Euclidean distance, defined by the formula (2), Minkowski metric - formula (3), and Manhattan distance, also known as city block – formula (4).

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \tag{2}$$

$$d_{ij} = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^m \right)^{\frac{1}{m}} \tag{3}$$

$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}| \tag{4}$$

where: d_{ij}- the distance between the units numbered *i* and *j* respectively; x_{ik}- the value of the *k*-th static feature for the *i*-th unit in the set. *T* is the collection of all expressions.

For more information on distance measures see for example in [5-6].

The second classification algorithm used in the study is naive Bayes classifier. That naivete rests on the assumption of independence of variables in a given classification space. This simplifies the calculation of probabilities and the determination of belonging to classes. Most often this algorithm is used for discrete data, in case of continuous space data must be discretized.

An important element is the a priori information, based on which we obtain information. It should be noted that the Bayesian classifier, which depends on a priori information, does not guarantee getting accurate results if the information is inconsistent and inaccurate. When creating a priori knowledge it is important to avoid subjectivity by analyzing several models with different a priori distributions or use the largest possible dataset.

The basis of Bayesian classification in each case is, as the name suggests, Thomas Bayes' theorem.

Theorem 1 (Bayes' theorem)

Let *D* be a set of data, $h \in H$ be the set of possible hypotheses. Each hypothesis *h* can be assigned an unconditional probability (a priori), designated as *P(h)*. For any hypothesis $h \in H$ and set *D* there is equality called the conditional probability *P (h|D)* (also called Bayesian model).

$$P(h|D) = \frac{P(h) * P(D|h)}{P(D)} \tag{5}$$

During the classification process the event *C* is identified with a class *X* with the values of attributes describing the new object. Classification by the method of maximum a posteriori is optimal. However, dimensionality turns out to be a serious problem. Finding all collective conditional probabilities is costly. The large computational complexity was simplified by the assumption of independence of the explanatory variables in the form of naive Bayes classifier. It is linearly dependent on the number of variables m, thus simplifying the classification algorithm. On the other hand, the stronger the relationship between the variables, the lower the quality of classification. Worse classification results, therefore, will be obtained when the a priori set is dominated by instance of one class.

Research and experiments

The study was conducted on a set of 1200 records obtained by resistance measurement using method described in point 2.

The *k* nearest neighbors classifier obtained better classification results. Three distance measures were used, namely Euclidean distance, Minkowski, and Manhattan.

The experiments of the authors were aimed to verify the selected classification model. In papers [3,4] the influence of the selection of distance measures on the value of precision and entropy was demonstrated. The results obtained in [4] confirmed the significant advantage of

Cosine measure over the Euclidean measure. In the present study, however, we did not apply Cosine measure due to simple data set. The k-nearest neighbors algorithm was tested for two method of voting, namely Majority Voting and Inverse Distance Voting. The results of the classification errors are shown in Table 1, using the following designations: MV - Majority Voting, IDV - Inverse Distance Voting.

Table 1 The number of misclassifications of k-NN algorithm for different distance measures.

Distance measure	The number of misclassifications	
	MV	IDV
Euclid	2	1
Minkowski	5	7
Manhattan	3	3

By looking at table 1 it can be noted that most of the misclassifications were obtained for the k nearest neighbors algorithm using the Minkowski measurement. Regardless of the choice of voting method, the K-NN classification with Minkowski measure gave the highest number of classification errors. In addition, it was noted that the change in the method of voting did not affect the outcome of classification when Manhattan measure was used.

As far as the naive Bayesian classifier is concerned, the authors' presumptions were confirmed that the results will be worse due to domination of one class in the prepared set.

Summary

The article summarizes the basic research methods associated with resistance measurements. Two algorithms for data classification, namely k-nearest neighbors and Bayesian classifier were briefly characterized. The advantages and disadvantages of each algorithm were demonstrated. The results of research and experiments, which clearly confirm the authors' assumptions that k-nearest neighbors is the most efficient algorithm for classifying resistance, were presented.

In the subsequent works the authors will attempt to improve the performance of Bayesian classifier for classification of resistance changes.

REFERENCES

[1] Korzeniewska E., Duraj A., Koneczny C., Krawczyk A.; Thin film electrodes as elements of telemedicine systems; *Przegląd Elektrotechniczny* (Electrical Review) Vol. 2014, Nr. 12

[2] Korzeniewska E., Duraj A., Krawczyk A.; Detection of local changes in resistance by means of data mining algorithms" *Przegląd Elektrotechniczny* 2014, R. 90, 229-232

[3] Duraj A., Krawczyk A., Finding outliers for large medical datasets, *Przegląd Elektrotechniczny* (Electrical Review), R. 86 NR 12/2010, pp.188-191.

[4] Duraj, A. Krawczyk, Dobór miar odległości w hierarchicznych aglomeracyjnych metodach wykrywania wyjątków, *Przegląd Elektrotechniczny* (Electrical Review) Vol. 2011, Nr. 12b, ISSN 0033-2097, R. 87, str. 33-37.

[5] Wald, L., A European Proposal for Terms of References In Data Fusion, *In-ternational Archives of Photogrammetry and Remote Sensing*, Vol. 32, No. 7,1998, pp. 651–654

[6] Jain A. K., Dubes R., Algorithms for Clustering Data, *Prentice Hall*, New Jersey, 1988

[7] Duraj A., Krawczyk A.; Detekcja wyjątków sygnałów biomedycznych w systemach fuzji informacji, *Przegląd Elektrotechniczny* (Electrical Review) Vol. 2012, Nr. 12b.

[8] Korzeniewska E., Jakubas A.; Pomiar rezystancji powierzchniowej warstw cienkich o dowolnych kształtach wytworzonych na podłożach elastycznych; *Przegląd Elektrotechniczny* 2014, Nr. 12, str. 58-62

[9] Kubiak A., Rogowski J.,·Ruta Ł., Lisik Z., Influence of annealing temperature on electrical parameters and structure of Al contacts on p-type 4H-SiC substrate, *Proceedings of SPIE - The International Society for Optical Engineering* 07/2013; DOI:10.1117/12.2031031

[10] Cieśla A, Kraszewski W, Skowron M, Syrek P; Determination of safety zones in the context of the magnetic field impact on the surrounding during magnetic therapy; *Przegląd Elektrotechniczny* 87 (7), 2011, 79-82

[11] Cieśla A, Kraszewski W, Skowron M, Syrek P; Analiza rozkładu pola magnetycznego generowanego przez urządzenia do fizykoterapii; *Przegląd Elektrotechniczny* 91 (2), 2015, 162-165

[12] Kasprzyk L., Tomczewski A., Bednarek K., Efficiency and economic aspects in electromagnetic and optimization calculations of electrical systems, *Przegląd Elektrotechniczny*, No 12, 2010, p. 57-60, (in Polish)

[13] Kasprzyk L., Bednarek K., Speeding up of electromagnetic and optimization calculations by the use of the parallel algorithms, *Przegląd Elektrotechniczny*, No 12, 2009, p. 65-68

[14] Kasprzyk L., Tomczewski A., Bednarek K., The distribution of an electromagnetic and optimization computation of electrical systems by using multi-core processors, *Przegląd Elektrotechniczny*, No 12b, 2011, p. 82-85, (in Polish)

[15] Drzymała P., Welfle H., Szacowanie strat dodatkowych w uzwojeniach transformatorów z wykorzystaniem numerycznych metod polowych, *Przegląd Elektrotechniczny*, No 1, 2015, p. 133-135

[16] Drzymała P., Welfle H., DB2 pureXML - zaawansowanie składowanie danych w strukturach relacyjno-hierarchicznych, *Przegląd Elektrotechniczny*, No 3, 2014, p. 165-168

[17] Byczkowska-Lipińska L., Wosiak A., Multimedia NoSQL database solutions in the medical imaging data analysis, *Przegląd Elektrotechniczny*, No 12, 2013, p.234-237

***Authors**: PhD Ewa Korzeniewska, Lodz University of Technology, Faculty of Electrical Engineering, Electronics, Informatics and Automatic Control, Institute of Electrical Engineering Systems, Stefanowskiego 18/22 Street, 90-924 Łódź, Poland e-mail: ewa.korzeniewska@p.lodz.pl; PhD Agnieszka Duraj, Lodz University of Technology, Faculty of Technical Physics, Information Technology and Applied Mathematics, Institute of Information Technology, Wólczańska 215 Street, 90-924 Łódź, Poland e-mail: agnieszka.duraj@p.lodz.pl; Professor Andrzej Krawczyk, Czestochowa University of Technology, Electrical Faculty, Armii Krajowej 17 Street, 42-200 Częstochowa Poland, e-mail: ankra.new@gmail.com*