Research Article

Radar Emission Sources Identification Based on Hierarchical Agglomerative Clustering for Large Data Sets

Janusz Dudczyk

R&D Department, WB Electronics S.A., Poznanska 129/133 Street, 05-850 Ozarow Mazowiecki, Poland

Correspondence should be addressed to Janusz Dudczyk; j.dudczyk@wb.com.pl

Received 11 January 2016; Revised 22 March 2016; Accepted 27 April 2016

Academic Editor: Fanli Meng

Copyright © 2016 Janusz Dudczyk. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

More advanced recognition methods, which may recognize particular copies of radars of the same type, are called identification. The identification process of radar devices is a more specialized task which requires methods based on the analysis of distinctive features. These features are distinguished from the signals coming from the identified devices. Such a process is called Specific Emitter Identification (SEI). The identification of radar emissions sources with the use of classic techniques based on the statistical analysis of basic measurable parameters of a signal such as Radio Frequency, Amplitude, Pulse Width, or Pulse Repetition Interval is not sufficient for SEI problems. This paper presents the method of hierarchical data clustering which is used in the process of radar identification. The Hierarchical Agglomerative Clustering Algorithm (HACA) based on Generalized Agglomerative Scheme (GAS) implemented and used in the research method is parameterized; therefore, it is possible to compare the results. The results of clustering are presented in dendrograms in this paper. The received results of grouping and identification based on HACA are compared with other SEI methods in order to assess the degree of their usefulness and effectiveness for systems of ESM/ELINT class.

1. Introduction

In the EW aspect, the way to increase the detail level of recognition is the SEI method [1–3]. It extracts distinctive features in the process of signal processing which comes from the emission source. The distinctive features above may be the result of transformations of received measurement data collections. As a result of these transformations, new collections should have distinctive features by which it is possible to identify precisely even a single copy of an emission source. Some example features which may appear as a result of specific identification are fractal features especially used in image transformation (Synthetic Aperture Radar, SAR) [4, 5], acoustic signal transformation, and radar signal analysis and transformation [6, 7]. By the extraction of distinctive features, received as a result of radar signal transformation, it is possible to identify radars of the same type. In work [7], there is a method of identification which is based on the extraction of features by which it is possible to have a correct identification with the probability 70% higher in comparison with classic methods (type identification). Also, measurement and analysis of out-of-band radiation are used in the radar identification process. The extraction of distinctive features coming from the out-of-band radiation analysis increases the precision of received results in the radar identification process from 50% to 70% [7]. Another way, which is used in SEI methods, is the analysis of inter-Pulse Repetition Interval modulation and intrapulse analysis of a radar signal. As it is presented in the work [8], the use of Linear Discriminant Analysis (LDA) and Karhunen-Loeve Transform (K-LT) makes it possible to identify specifically radars of the same type where the probability of correct identification equals 98% and Correct Identification Coefficient (CIC) value equals, for the new features, 0.98 and, for the old features, 0.47. The results above present that radar signal processing using intrapulse features and Karhunen-Loeve transform can be a useful tool for EW devices. In the work [9], there is a method of SEI which can be used in ESM systems. The main idea is to analyze the radar pulses and characterize those by extracting features that should be different for each radar. The applicability of the feature extraction procedure has been analyzed for different case studies.
to obtain a complete picture of the results achievable with the different radar signals. Also, the Fourier Transform has been widely used in radar signal and image processing. In the work [10], it is presented that Joint Time-Frequency (JTF) domain analysis is a useful tool for improving radar signal and image processing for time- and frequency-varying cases. Also, Vector Neural Network (VNN) with a supervised learning algorithm suitable for signal classification is very useful for emitter identification process, as shown in the work [11]. Also, the system for automatic recognizing of radar waveforms was introduced in the work [12], where the intercepted radar signal is classified to eight classes based on the pulse compression waveform: linear frequency modulation (LFM), discrete frequency codes (Costas codes), binary phase, and Frank, P1, P2, P3, and P4 polyphase codes. Simulation results show that the classification system achieves overall correct classification rate of 98% at signal-to-noise ratio (SNR) of 6 dB on data similar to the training data.

This paper deals with the problem of radar emission source identification with the use of the agglomerative method of hierarchical radar signal clustering. The problem of object clustering is connected with diversity of definitions resulting from no precise definition of a cluster. Thus, clustering is an issue which is not always solved explicitly. Clustering and classification are problems which are strictly connected with pattern recognition [13, 14]. Clustering concerns dividing a collection into groups (clusters). These clusters have the property owing to which elements in the same cluster are similar to each other while elements in different clusters are different from each other. Classification is assigning objects to classes once defined. As there are different criteria, there is also a different division of algorithms of clustering and classification [15]. These criteria are usually types of used measures, quality of solution, or the way of getting algorithm to the solution.

Methods of data clustering are also used in the process of radar recognition and identification. The term of radar source emission identification functions in radiotechnical identification in the majority of cases in two senses, that is, in a broad sense and in a narrow sense. The radar identification in a broad sense consists of a quite accurate definition of the place, the destination, and options of this signal on the basis of the results of detected parameters’ measurements and located signals from radar. The identification in a narrow sense is the classification of these signals. Depending on the number of details, the radar identification in a narrow sense might concern the classification of types and the identification of copies where the classification of the emission sources concerns the division of signal collections into groups corresponding to particular types of emission sources, while the identification of copies concerns the divisions of sets of signals into groups corresponding to particular copies of electromagnetic emission sources, which are of the same type.

It should be noticed that “measurement data” above are from radar devices which work physically on the battlefield, while basic measurement parameters of radar signal are received in the process of their analysis and initial processing. The second ones are described precisely in Section 3.2 (The Structure of Data Sets). Basic measurement parameters of radar signals are generated in a form of a signal sounding by a radar and are not enough to identify its emission source, that is, to define its particular copy for the same type of radar. This attitude to the process of radiotechnical emission source recognition is called Specific Emitter Identification. As a part of the advanced method of SEI analysis presented in this paper, it should be emphasized that there is a significant fact; namely, all measurement data in the form of recorded radar signals come from a dozen or so working radiolocation devices of the same type. Their structure, in the form of Pulse Description World (PDW), is described in Section 3.2. The structure of PDW is not the main subject of this work and its detailed description is presented in works [7, 16, 17]. A superheterodyne ELINT receiver is used in order to record data. The description of the procedure is in Section 3.2.

The aim of the authors is to emphasize the fact that the described emission source identification process is a really complex problem, where a number of aspects such as Database (DB) modelling [7, 16], the method of creating the pattern, the classification and identification process, used criteria, and methods estimating the CIC are currently a great challenge for researchers and have no optimal solutions. It is also good to point out the fact that their target is to be implemented to ESM/ELINT systems and to be used in EW in an optimal way, which causes no computational overload to such a recognition system.

2. Classic Model of Radar Emission Sources Recognition

The radar emission sources identification by the classification of signals (images), which come from them, can come down to the problem of object recognition through the recognition of objects’ images. Techniques of object recognition are currently developed fields of science; however, in many cases it is still not possible to formulate the optimal model of object recognition [1, 18]. Simultaneously, in the process of identification, the following problems appear:

(i) The side conducting the identification has no sufficient information to describe the classes in a way that would fit with the reality.

(ii) It is possible to define more or less accurate equivalents of particular classes in signal space through their models.

(iii) The decision about assigning a particular signal to a particular class is arbitrary.

(iv) The presence of random incidences causes the appearance of false classification.

(v) There is a lack of possibility of identifying copies of the same type without using the method of Specific Emitter Identification (SEI).

Disregarding the precise content of a recognized object (radar), what should be assumed is the fact that it can be defined with the use of a set of features. As a result of measure procedure of a radar signal, it is possible to present each of the analyzed signal features \( N \) in the form of a numerical value.
Thus, a formal description of a radar is a set of $N$ numbers $x = (x_1, x_2, \ldots, x_N)$, called the object image. In reality, the object image can be not only a set of numbers but also collections of logical expressions and collections describing its structure. The exact description of a source emission is significant as far as the proper construction register of a radar in DataBase is concerned. It is important as it eliminates redundancies from DataBase and designs a DataBase increasing the probability of a proper radar identification [7]. Methods based on unintentional emissions or distinctive features extraction with fractal properties [6] are often used in the process of radar recognition. These, also called Specific Emitter Identification (SEI), increase probability of source emission identification. As a result, identification of each of the copies is possible.

3. Hierarchical Clustering Dendrogram

Clustering is presented as a process of creating clusters of elements similar to each other. Clustering is based on the division of the finite set $V$ of elements on $L$ subsets $V_k$ of similar elements. This division can be written in the form of a family $Q(V)$ of subsets $V_k$ of the collection $V$, according to

$$Q(V) = \{V_k \subset V: k \in [1 : L]\}. \quad (1)$$

The set of all possible (or all acceptable) divisions $Q(V)$ of set $V$ on $L$ subsets is written as $Q^L(V)$. Division $Q$ will be defined as an optimal division $Q_{opt}$ if the criterion function $e(Q)$ from the criterion assumed is extremum for this division, as shown in

$$Q = Q_{opt} \iff e(Q) = \min_{Q \in Q^L} e(Q). \quad (2)$$

The definition of similarity measure between objects is necessary to start the clustering process. This definition will be also used to provide classes. A class in this case is considered to be a collection of objects whose similarity within the class is high whereas among classes is low. The definition of similarity plays a key role in such a method of clustering. There are three categories: agglomeration category, division category, and direct category in divisions based on the type of control used while building clusters. Agglomeration category is based on progressive merging of clusters, that is, iteration connection of separate small clusters in order to create bigger ones and make a final one. The process of such connection is usually presented in the form of a dendrogram. By using threshold which defines the minimum of similarity, the process of connection may be broken before the clusters become the final one. Thus, there will be few clusters and for each of them the value of similarity measure will be below the acceptable threshold. Division category is based on creating clusters by iteration division of the initial collection into smaller and smaller ones, finally, creating a situation in which every element will have its own cluster. The direct category is based on finding a specific number of clusters. What should be done here is finding such division of existing data that will be optimal in case of the measure defining the division into clusters. One of the first such methods is $k$-means MacQueen algorithm shown in work [19].

3.1. The Use of Threshold Criteria. Methods and criteria of elements clustering are significant in the process of clustering. During the completion of research procedure, Euclidean and Mahalanobis distances ($d_E$, $d_M$) were used according to (3) and (4). In this equation, $x$ and $y$ are the elements of the set in space $V$, $n$ is the number of elements in the set in space $V$, and $C$ is the covariance matrix according to (5), while $\overline{v}$ are the average elements of the set according to (6):

$$d^2_E (x, y) = d (x - y) (x - y)^T, \quad \forall \ x, y \in V^L \quad (3)$$

$$d^2_M (x, y) = d (x - y) C^{-1} (x - y)^T, \quad \forall \ x, y \in V^L \quad (4)$$

$$C = \frac{1}{n} \sum_{j=1}^{n} (v_j - \overline{v})^T (v_j - \overline{v}), \quad v_j \in V, \quad (5)$$

$$\overline{v} = \frac{1}{n} \sum_{j=1}^{n} v_j, \quad (6)$$

Two threshold criteria were used for the process of clustering. These were the furthest neighbour criterion (FNC) and the nearest neighbour criterion (NNC) according to (7) and (8), where $V_k$ are subsets of similar elements and $d$ is the distance function according to (9), in which $R^1$ is the real number system:

$$\forall \ V_k \in Q_{opt} \quad \forall \ x, y \in V_k \quad [d(x, y) < d_{max}], \quad (7)$$

$$\forall \ V_k \in Q_{opt} \quad \forall \ x, y \in V_k \quad [d(x, y) \leq d_{max}], \quad (8)$$

$$d : V \times V \rightarrow \{r \in R^1 : r \geq 0\}. \quad (9)$$

The advantage of these criteria is their simplicity and high agreement of optimal clustering results with those arbitrary realized by a human. The NNC highlights similarities of pairs. The FNC highlights the influence of the isolated elements of space $V$. This is some kind of disagreement as there is elements deviation of the space $V$ resulting from defining the place of elements in space $V$ on the basis of quantities defined experimentally.

3.2. The Structure of Data Sets and Measurement Procedure. Measurement data sets, which are received as a result of measurement procedure, are treated as homogeneous data, where interferences in PRI and RF measurement sequences are removed. The data selection and interferences removal methods are described precisely in the work [20]. All radar signals measurements are done during warfare. In total, there is a record of several thousand pulses coming from a dozen or so radar devices of the same type. In order to provide repetitiveness of the received measurement results, a measurement procedure was introduced, where radar signal measurement of every single radar is done in four measurement places with a constant distance from the radar emission source. The choice of three copies of the same type of radar to detailed verification is made in order to, by their dislocation,
provide comparable (in all cases) landform features in the area where the radar is dislocated. For measurement, there are constant values of detection thresholds and constant values of superheterodyne receiver sensitivity, which are set during measurement on the Radio Frequency of the analyzed emission source. The received measurement (recording) data set is described by the analyzed signal from the amplitude demodulator (AM Lin) output and frequency demodulator [21]. Recordings of radar signals in the forms of PDW were received during the measure procedure. The vector received is a formalized structure of a record type according to (10). Specific PDW fields contain frequency parameters and time parameters, where PRI\( (k) \) is the Pulse Repetition Interval, PW\( (k) \) is the Pulse Width, RF\( (k) \) is the Radio Frequency, \( n \) is the number of pulses in the recording which are qualified for analysis, and \( k \) is the number of the pulses in recording:

\[
\text{PDW} = \begin{bmatrix}
\text{PRI}(1) & \text{PW}(1) & \text{RF}(1) \\
\text{PRI}(2) & \text{PW}(2) & \text{RF}(2) \\
\vdots & \vdots & \vdots \\
\text{PRI}(k) & \text{PW}(k) & \text{RF}(k) \\
\vdots & \vdots & \vdots \\
\text{PRI}(n) & \text{PW}(n) & \text{RF}(n)
\end{bmatrix}. \quad (10)
\]

The holdout method \( (M_{H1}) \) was used in order to define the structure of measure vectors. This method divides the set of measurement data into two separate subsets, that is, the subset used to teach the classifier and the subset used to test the classifier. Usually the division is as follows: 2/3 available data is the teaching set and 1/3 is the testing set, [22, 23]. With the use of PDW vector, the vector of basic features \( V_p \) was defined according to (11) in which \( \text{RF}^{M_{H1}} \) is the average value of Radio Frequency, \( \text{PW}^{M_{H1}} \) is the average value of Pulse Width, and \( \text{PRI}^{M_{H1}} \) are average values of Pulse Repetition Interval:

\[
V_p = \left[ \text{RF}^{M_{H1}}, \text{PW}^{M_{H1}}, \text{PRI}^{M_{H1}} \right]. \quad (11)
\]

Individual component features of \( V_p \) vector were defined according to (12) \( \div (14) \), where \( n_w \) is the number of pulses in recording which were qualified for the analysis in order to use the holdout method:

\[
\text{RF}^{M_{H1}} = \frac{1}{n_w} \sum_{i=1}^{n_w} \text{RF}(i), \quad (12)
\]

\[
\text{PW}^{M_{H1}} = \frac{1}{n_w} \sum_{i=1}^{n_w} \text{PW}(i), \quad (13)
\]

\[
\text{PRI}^{M_{H1}} = \frac{1}{n_w} \sum_{i=1}^{n_w} \text{PRI}(i). \quad (14)
\]

4. The Implementation of GAS and Results of Analysis

The hierarchical clustering method was used in order to study particular features of \( V_p \) vector. This method uses hierarchical clustering and grouping algorithm which is based on the Generalized Agglomerative Scheme. It is a special type of hierarchical algorithms that are most appropriate for large data sets [24–26]. To divide a particular \( n \) cluster of elements into \( L \) groups, what needs to be done first is dividing it into \( n \) groups. This means that each of the groups consists of only one element. Secondly, two elements with the highest similarity should be connected into one group. As a result, there are \((n - 1)\) groups and later \((n - 2)\) groups until there are an acceptable number of clusters. The basic steps in the procedure above as are follows:

1. \( m = n \), whereas the group \( X_i = x[i], i = 1, 2, \ldots, n \).
2. If \( m \leq L \), stop.
3. Find the nearest pair of groups, for example, \( X_i \) and \( X_j \).
4. Combine \( X_i \) and \( X_j \), remove \( X_j \), and decrease \( m \) by one and jump to (2).

The procedure ends its activity after reaching the right number of groups \( m \leq L \). The hierarchical clustering algorithm which was implemented for research purposes was two-element parameterized. One of the parameters concerns the way of finding a similarity between features. It uses Euclidean distance \( (d_E) \) and Mahalanobis distance \( (d_M) \). The second parameter is connected with the clustering method by using the NNC and FNC criterion. The process of clustering in this algorithm may be divided into three stages, that is, measuring the distance, dendrogram building (on the basis of specific parameters), and dendrogram cutting on the specific level. The method above was used in a research experiment in order to group \( V_p \) vectors in the process of radar identification. \( V_p \) vectors are basic measure parameters of radar signals. What is important is that the measure vectors received, \( V_p \), are real radar signal recordings coming from few copies of radars of the same type. Thus, the process of identification is not easy. It is a special case of sources emissions identification. There were three radars chosen of the same type for the analysis. Their basic measure parameters merged most. The results of clustering the PRI were presented in the form of dendrograms shown in Figures 1–9. The labels on the axis in the Figures 1–9 present the initial value of these clusters and on the y-axis there is similarity between clusters. As a result of this presentation, it is possible to estimate the number of clusters and it is also possible to have a PRI outlier vector.

The process of cluster connection is carried out until the number of clusters received is sufficient. Graphically, shown in a dendrogram, the condition of an alloy (i.e., a specific number of clusters) presents the horizontal line which cuts the dendrogram. The value bracket \( \Delta d_x \) which is defined in each of the dendrograms presents the “safe area” for which there is a guaranteed correct result of data clustering. The received value \( \Delta d_x \), different for each of analyzed vectors \( V_p \), means that these are vectors coming from different radars.
Figure 1: The hierarchical clustering dendrogram of PRI (Euclidean distance, NNC) for copy of Radar Number 1.

Figure 2: The hierarchical clustering dendrogram of PRI (Euclidean distance, NNC) for copy of Radar Number 2.

Figure 3: The hierarchical clustering dendrogram of PRI (Euclidean distance, NNC) for copy of Radar Number 3.
copies of the same type. As a result of the use of agglomerative method (bottom-up), initially each PRI measurement vector is a separate cluster (class); in further iterations clusters are joined to bigger clusters until all PRI values belong to one cluster. In this way, clusterization is presented in the form of dendrograms.

Figures 1–3 present dendrograms for Euclidean distance and NNC, in which the distance between the first pair of clusters is quantified. The dendrograms which appeared here are cut at the level of the first pair. The function which cuts the structure of dendrogram restores the results as follows: 19.94, 19.9, and 19.82.

Figures 4–6 present dendrograms for Mahalanobis distances and NNC, in which the distance in the first pair of clusters is quantified and the rest of the dendrograms are cut at the level of the first pair. The results of clusterization $\Delta d_x$ are as follows: 0.104, 0.099, and 0.101.

Similarly, Figures 7–9 present dendrograms for Euclidean distances and FNC, where the results of clusterization $\Delta d_x$ are as follows: 20.08, 20.26, and 20.24. As a result of such an attitude, there is information about distinctive features of the vector of radar signal basic measurement parameters, $V_p$, in the aspect of PRI.

The use of clusterization result $\Delta d_x$ makes it possible to expand features of the vector $V_p$ into the received clusterization values. Thus, there is a measurement vector for every copy of a radar of the same type. The Hierarchical Agglomerative Clustering Algorithm used in the SEI process, based on GAS, makes it possible to receive hierarchical clustering for Pulse Repetition Intervals. In the process of clustering,
Euclidean and Mahalanobis distance measures are used as well as the nearest neighbour criterion (NNC) and the furthest neighbour criterion (FNC) for three different types of radar copies of the same type, that is, Number 1, Number 2, and Number 3.

5. Conclusion

The characteristic feature of the algorithm implemented is the possibility of presenting clustering structure in the form of a dendrogram. Such presentation of clustering results provides a wide range of options, for example, estimating the number of clusters (if it is not known before) and the possibility of analyzing appearing diverge vectors.

Using Euclidean distance and NNC criterion, the received values $\Delta d_x$ for particular radar copies are as follows: 19.94, 19.9, and 19.82. Using Mahalanobis distance and NNC criterion the received values $\Delta d_x$ for particular radar copies are as follows: 0.104, 0.099, and 0.101. Using Euclidean distance and FNC criterion the received values $\Delta d_x$ for particular radar copies are as follows: 20.08, 20.26, and 20.24. The determinant which influences the $\Delta d_x$ quantity is the type of measure distance used. Similarity between clusters was defined by quantities; thus, the dendrograms received will have particular proportions of similarity. As a result, it is possible to use the change of the distance to assess if the connection was natural or forced. This method makes it possible to differentiate particular radar copies of the same type on the basis of the dendrograms received.
The received measurement results have a significant influence on the radar emission sources specific identification of radar copies of the same type. Other methods mentioned in Section 1 of this paper such as the use of the out-of-band radiation [7], fractal features extraction [6], and methods based on the intrapulses analysis [8] increase the probability of identification to 50%–70%. In the work [6], there was an increase of the Correct Identification Coefficient level from the value $CIC = 0.169$ to the value $CIC = 0.916$, while in work [16] the value of decision function for the same radar types identification equals 63%. As it is presented in the work [8], radar signal processing using intrapulse features, Karhunen-Loeve Transform (K-LT), and Linear Discriminant Analysis can be a useful tool for Electronic Warfare devices. Both LDA and K-LT gave very similar results, and the received Correct Identification Coefficient (CIC) value equals 0.98 for the new features and 0.47 for the old features. The measurement results present that the new transformed features include about 90% of the recognized information needed to resolve the complicated problem of radar signal classification. The results of the speed and numerical stability of algorithms seem to be enough to put them into practice in the ESM devices. Simulation results presented in the work [12] show classification rate of 98% at signal-to-noise ratio (SNR) of 6 dB on data similar to the training data.

It must be admitted that it is an extremely high increase; however, the level of complexity of these methods and the used algorithms are complicated computationally, which causes the identification time to increase as well. The hierarchical PRI clustering method presented in this paper, based
on HACA, is realized on the basis of the use of MATLAB software and the received vectors $V_p$ are recorded in the dedicated DB for EW/ELINT system. Further works on the use of HACA in SEI process work out the matrix of mutual similarity, by which it is possible to estimate automatically the similarity among PRI vectors for different radars of the same type. Also, the automatic defining mechanism of $\Delta d_x$ value should be applied and an additional $\Delta d_y$ and a feature to $V_p$ measure vector should be added. The feature mentioned here is a good separation measure in the process of radars identification. This problem will be still examined in the presented SEI area.

Competing Interests

The author declares that he has no competing interests.

References
