Expert Systems With Applications 151 (2020) 113317

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A comparison of clustering algorithms for automatic modulation classification

Jacques P. Mouton, Melvin Ferreira*, Albertus S.J. Helberg

School of Electrical, Electronic and Computer Engineering, North-West University, Potchefstroom 2531, South Africa

ARTICLE INFO

Article history:
Received 11 October 2019
Revised 14 February 2020
Accepted 15 February 2020
Available online 16 February 2020

Keywords:
Automatic modulation classification
Constellation diagram
I/Q plane
Clustering algorithm
Centroid estimation

ABSTRACT

In this paper, the k-means, k-medoids, fuzzy c-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Ordering Points To Identify the Clustering Structure (OPTICS), and hierarchical clustering algorithms (with the addition of the elbow method) are examined for the purpose of Automatic Modulation Classification (AMC). This study compares these algorithms in terms of classification accuracy and execution time for either estimating the modulation order, determining centroid locations, or both. The best performing algorithms are combined to provide a simple AMC method which is then evaluated in an Additive White Gaussian Noise (AWGN) channel with M-Quadrature Amplitude Modulation (QAM) and M-Phase Shift Keying (PSK). Such an AMC method does not rely on any thresholds to be set by a human or machine learning algorithm, resulting in a highly flexible system. The proposed method can be configured to not give false positives, making it suitable for applications such as spectrum monitoring and regulatory enforcement.

1. Introduction

Automatic modulation classification (AMC) refers to a process that identifies modulation schemes used to modulate an unknown signal without human input. AMC has several military and civilian applications. In civilian applications, spectrum users are assigned specific bands in the electromagnetic spectrum which depend on the regulatory domain, user application, radio technology used, and the transmitter. For example, 802.11 Wi-Fi commonly operates at 2.4 GHz in unlicensed spectrum, while most cellular providers have dedicated bands assigned in licensed spectrum operating at 900 MHz and 1800 MHz. AMC can be used in cognitive radios that can sense the spectrum in order to automatically switch to the optimal modulation scheme. In 5G networks it will also be necessary to continuously monitor the electromagnetic spectrum to ensure that only authorised transmissions occupy allocated frequency slots in space and time. In military applications, a signal could be intercepted and demodulated if the modulation scheme can be determined. Alternatively, signal jamming can be made more effective if the modulation scheme is known (Zhu & Nandi (2015)). AMC methods typically fall into two categories, likelihood-based and feature-based. Likelihood-based methods typically outperform feature-based methods when it comes to classification accuracy but require accurate channel parameter estimation. Feature-based methods are typically used in situations where not all channel parameters can be determined (Zhu & Nandi (2015)).

Except for a limited number of legacy broadcast transmissions that use analogue modulation, most modern communications systems use digital modulation techniques. Digital modulation schemes, such as Pulse-amplitude Modulation (PAM), QAM, and PSK, allow both higher data rates and spectral efficiency. These modulation schemes can be represented in a two-dimensional Cartesian plane known as the In-phase and Quadrature (I/Q) plane. Each sample is represented by a single point. These I/Q samples can be clustered together to estimate the unknown modulation scheme's order and symbol level. This feature is, however, rarely used and merits further investigation (Ali, Yangyu, & Liu (2017)).

In this paper, the constellation representation of the received modulated samples in the I/Q plane is used as a candidate feature. Digitally modulated signals are received through an AWGN channel. Various clustering algorithms are compared for their ability to estimate the order and level of the received symbols in the I/Q plane. These symbol levels are then compared with a pre-defined pool of modulation schemes' known symbol levels, to identify the modulation scheme.

The rest of this paper is ordered as follows: Section 2 provides a background on the clustering algorithms used in the comparison. Related work in AMC is given in Section 3. An overview of the three stages of the proposed AMC method is given in Section 4. Section 5 discusses the results and findings from the evaluations

* Corresponding author.
E-mail addresses: 24911658@student.nwu.ac.za (J.P. Mouton), melvin.ferreira@nwu.ac.za (M. Ferreira), albert.helberg@nwu.ac.za (A.S.J. Helberg).

https://doi.org/10.1016/j.eswa.2020.113317
0957-4174/

and comparisons that were made in this paper. More specifically, Section 5.2 evaluates applicable algorithms for modulation order estimation, Section 5.3 evaluates the abilities of different clustering algorithms to determine symbol levels, and Section 5.4 uses the best performing clustering algorithms from the prior evaluations to compose the proposed AMC method. A comparative evaluation with popular AMC methods from related work is shown in Section 5.5. Finally, advantages and disadvantages are discussed in Section 6.

2. Clustering algorithms

For the purpose of this paper, clustering algorithms can be functionally split into two categories which will be defined as: *known-order*, where the number of clusters is used as an input parameter, and *unknown-order*, where this is not required. In known-order algorithms, data is partitioned under a specified number of clusters according to a set of rules. Unknown-order algorithms do not require the modulation order as input, instead relying on factors such as the density of received symbols in the I/Q plane to determine the number of clusters.

Several of the mentioned clustering algorithms have been investigated with the purpose of improving computational tractability. Techniques such as parallel computing, distributed processing and algorithmic improvements are employed to improve scalability and execution time. (Cordova and Moh (2015); Corizzo, Pio, Ceci, and Malerba (2019); He et al. (2011); Mao, Xu, Li, and Ping (2015)). We will focus only on single-threaded performance in order to compare the execution time of these clustering algorithms.

2.1. Unknown-order algorithms

2.1.1. Density-based spatial clustering of applications with noise (DBSCAN)

DBSCAN is a density-based clustering algorithm that does not require the number of clusters as input as proposed by Ester, Kriegel, Sander, and Xu (1996) and implemented by Yarpiz (2015). Instead, the algorithm requires two input parameters, epsilon (ϵ) and MinPts. DBSCAN makes a distinction between core points and border points. Points that are within ϵ distance of each other are said to be core points if the number of these points is greater than or equal to MinPts. Border points are points that can be reached by a core point, but cannot be used to reach another point. Border and core points together form clusters with all other points being classified as noise. MinPts, therefore controls the minimum size of the clusters, while ϵ controls how dense those clusters are. DBSCAN has previously been used for modulation order estimation by Zhu and Fujii (2016).

In our evaluation of DBSCAN, the optimal values of ϵ and MinPts were found to vary significantly between modulation schemes, modulation order as well as the amount of noise present in the system. Fig. 1(a) and (b) shows a sensitivity analysis conducted for 8-PSK and 16-QAM respectively, by sweeping MinPts and ϵ over a range of values and recording the resulting number of clusters. The value of MinPts is shown on the y-axis and ϵ on the x-axis, with the number of clusters shown by the intersection of these two values. These two figures demonstrate that in noisy environments, only a narrow region between these two parameters will produce the correct number of clusters, implying that the value of ϵ and MinPts will have to be uniquely determined for each set of parameters under evaluation.

2.1.2. Ordering points to identify the clustering structure (OPTICS)

OPTICS is a modified version of DBSCAN proposed by Ankerst, Breunig, Kriegel, and Sander (1999) and implemented by Daszykowski (2002). DBSCAN was used to estimate modulation order in Jajoo, Kumar, Yadav, Adhikari, and Kumar (2017). OPTICS does not necessarily cluster the data in the traditional sense—instead it provides an ordering of the data based on density. The initial purpose of OPTICS was to improve on DBSCAN by reducing its sensitivity to changing input parameters by only requiring MinPts. The reachability distance (ϵ) can then be chosen in a post-processing step. This can be seen in Fig. 2, with each peak representing a cluster and the points succeeding each peak belonging to that specific cluster. Peaks can be determined by identifying regions in which the reachability distance rises and drops again below a certain threshold. The threshold is determined as a fraction of the largest reachability distance. The maximum value between those regions can then be considered a peak. As proposed by Jajoo et al. (2017), in our evaluation, this threshold is set as 60% of the average between the two highest peaks, marked with a green star, disregarding the large initial peak. A sensitivity analysis also verified this threshold value as suitable.

2.1.3. Hierarchical clustering

In hierarchical clustering, each data point starts either as its own cluster that is iteratively merged with its closest neighbour until only one cluster is present (agglomerative), or a single cluster comprising all data points which is then divided until each data point is its own cluster (divisive) (Xu & Tian (2015)). The distance between merging clusters is then represented in a dendrogram, as shown in Fig. 3. The number of clusters can be determined by using a maximum distance as a stop-criterion. Alternatively, the distance between the merging of clusters can be examined in a post-processing step by using a technique like the elbow method (Antunes, Gomes, & Aguiar (2018)). Hierarchical clustering has not previously been used for AMC in literature.

2.2. Known-order algorithms

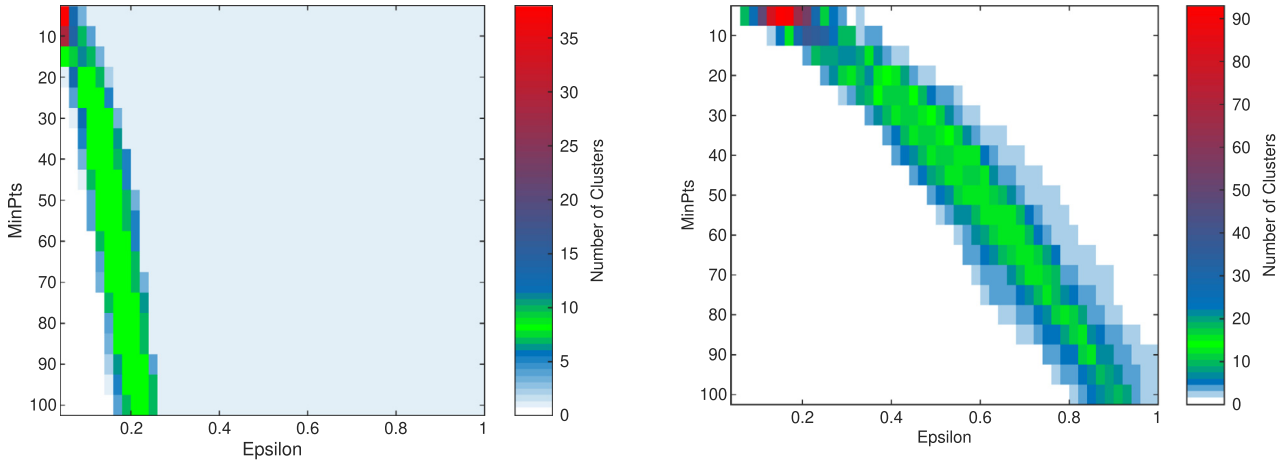
2.2.1. K-means

The k-means algorithm (also known as Lloyd's algorithm) is an iterative clustering algorithm for minimising the total distance between data points and their assigned cluster's centroid (Lloyd (1982)). The only required inputs are the data points, the number of clusters required, and a stopping condition. In its original form initial clusters are distributed randomly, however Arthur and Vassilvitskii (2007) developed the k-means++ algorithm, which improves on the random initial clusters and significantly accelerates the k-means algorithm's convergence rate.

The algorithm functions by determining which cluster's centroid is closest to each point. Each point is then assigned to its nearest cluster. Each cluster's centroid is recalculated as the arithmetic mean or geometric center of all its assigned points' locations. This process is repeated either until no points can be assigned to a closer cluster or until the stopping condition has been reached. K-means is typically used in AMC literature to estimate the symbol levels (Azarmanesh and Bilén (2013); Jajoo et al. (2017)), but not to estimate the modulation order.

2.2.2. K-medoids

The k-medoids algorithm is similar to the k-means algorithm in the sense that data is iteratively partitioned to minimise the total distance error. The main difference between the k-means algorithm and the k-medoids algorithm is that while k-means uses the arithmetic mean of a cluster's data points as the cluster's centroid, k-medoids uses the median data point, or medoid, belonging to the cluster. This allows k-medoids to be more robust to outliers and therefore more representative of the data (Xu & Tian (2015)). Variations of the k-medoids algorithm have been used in literature to



(a) 8-PSK (b) 16-QAM

Fig. 1. Sensitivity analysis of the number of clusters for the DBSCAN algorithm for an input signal with 12 dB SNR in an AWGN channel.

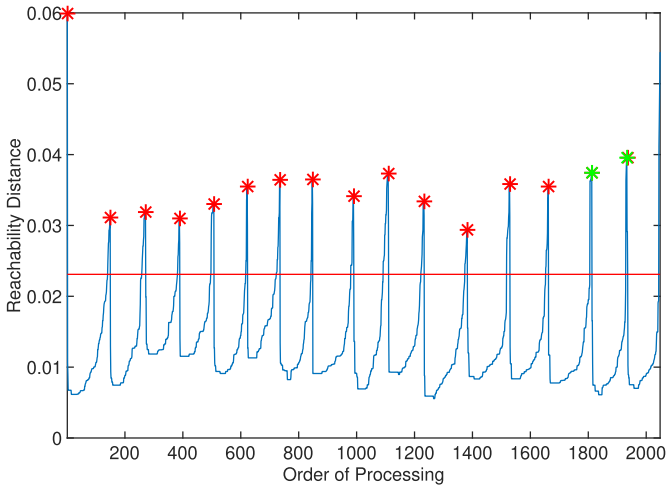


Fig. 2. Extracting information from the OPTICS clustering algorithm. The input signal has 16 symbol levels at 15 dB SNR in an AWGN channel. Identified peaks (clusters) shown in red with the two highest peaks indicated in green.

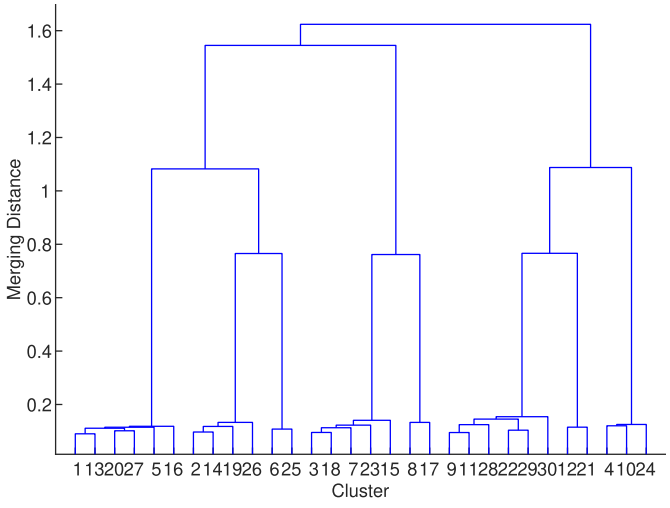


Fig. 3. Dendrogram for eight symbol levels. Horisontal lines represent the linking of two data points or clusters. Vertical lines represent the distance between the two merging clusters.

determine the symbol levels in the I/Q plane but not to estimate the modulation order.

2.2.3. Fuzzy c-means

The fuzzy c-means algorithm was first proposed in Bezdek, Ehrlich, and Full (1984) and used by Ahmadi (2010) and Mobasseri (2000) to estimate the number of symbol levels. The fuzzy c-means algorithm also aims to minimise the distance error. However the output of the fuzzy c-means algorithm is a degree of membership rather than a cluster number, allowing each point to belong to more than one cluster. The cluster with the highest degree of membership is generally accepted as the cluster that the point belongs to. The number of symbol levels, i.e. the modulation order, can be estimated by first clustering data with the highest number of symbol levels in the pool and iteratively merging clusters whose points show a large degree of membership overlap. This process is repeated until points show majority membership to only one cluster. Fuzzy c-means has been used in literature for AMC (Ahmadi (2010); Mobasseri (2000)).

3. Related work in AMC

Clustering techniques can also be used in statistical approaches for pattern recognition that does not directly rely on the location of symbols in the I/Q plane. This study will limit its focus to clustering I/Q data for estimating the modulation order and symbol levels of an unknown digital modulation scheme.

The statistical approach envolves clustering features extracted from a signal in a feature space where each cluster represents a modulation scheme type. An example of these features is the standard deviation of the direct phase component; and the signal spectrum symmetry around the carrier. These two features were used by Guldemr and Sengur (2006) to classify analogue modulation schemes with several clustering algorithms. Currently, the use of analogue modulation is limited to specific applications, such as terrestrial broadcast radio and land mobile radio services. Terrestrial broadcast television in analogue remain in some countries where the digital migration process, as mandated by the International Telecommunications Union (ITU), has not yet been concluded (ITU (2006)).

The k-means, fuzzy c-means, DBSCAN and OPTICS algorithms have previously been used for AMC in literature in some capacity. However, these clustering algorithms have never been directly compared in the context of AMC.

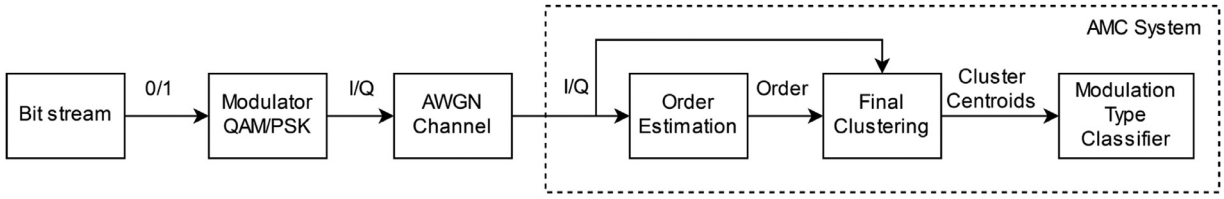


Fig. 4. Stages of the proposed AMC method.

Jajoo et al. (2017) were the first to use the OPTICS algorithm for AMC. First an error value (e), is calculated and compared with a threshold value to detect Amplitude Shift Keying (ASK). Incoming I/Q symbols are then clustered with OPTICS to estimate the number of symbol levels. The I/Q data is then reclustered, with the k-means algorithm, using the estimated number of symbol levels. Next, a distance error is calculated and compared with a threshold to distinguish M-QAM from M-PSK.

Zhu and Nandi (2014) investigated the use of clustering algorithms for the purpose of AMC. In their approach, I/Q symbols are clustered together with the different number of clusters as an input parameter. For each number of clusters the distance error between each symbol and its cluster centroid is calculated. This information is then used to develop a "minimum distance centroid estimator", incorporating a Non-Parametric Likelihood Function (NPLF) to blindly classify modulation schemes.

Azarmanesh and Bilén (2013) used k-means, k-means++, and k-centre algorithms to develop a method for classifying single carrier signals. In this work, specific attention is given to identifying Orthogonal Frequency-Division Multiplexing (OFDM) signals and the separation of sub-carriers into single carrier signals that can then be identified. After a single carrier signal is identified, the greedy k-centre algorithm is used rather than k-means++ for initialisation to improve accuracy. The symbols are then clustered together with k-means, and a cumulative deviation error is calculated and compared with a threshold to distinguish between modulation types such as M-QAM and M-PSK.

The fuzzy c-means clustering algorithm was used by Ahmadi (2010) to determine the modulation order. With this algorithm the I/Q symbols are clustered together with a varying number of samples. The distance error between the known symbol levels and estimated symbol levels is then used to determine the highest possible modulation order. The Two-Threshold Sequential Algorithm Scheme (TTSAS) is then used as it does not require the precise number of symbol levels, but does require two thresholds to be trained.

Mobasseri (2000) argues that the shape of a modulation scheme's constellation representation is a robust feature in a noisy environment. The fuzzy c-means algorithm is used to cluster I/Q data into different number of clusters, starting with the highest modulation in a pool and lowering the number of clusters until low membership overlap is shown. The modulation type is then identified using a Maximum Likelihood (ML) rule based on known modulation scheme models.

4. Proposed method

The proposed AMC method in this study is composed of several stages as shown in Fig. 4 namely Order Estimation, Final Clustering, and Modulation Type Classification. The Order Estimation stage accepts arbitrary I/Q samples, applies the relevant clustering algorithm and outputs the estimated number of clusters, which, represents the estimated modulation order. The Final Clustering stage clusters the I/Q samples into the number of clusters determined in the previous stage, with the relevant clustering algorithm, to output the cluster centroids to the next stage. The centroids of these clusters are then compared to the constellation representation of known modulation schemes in the Modulation Type Classification stage in order to determine the modulation type.

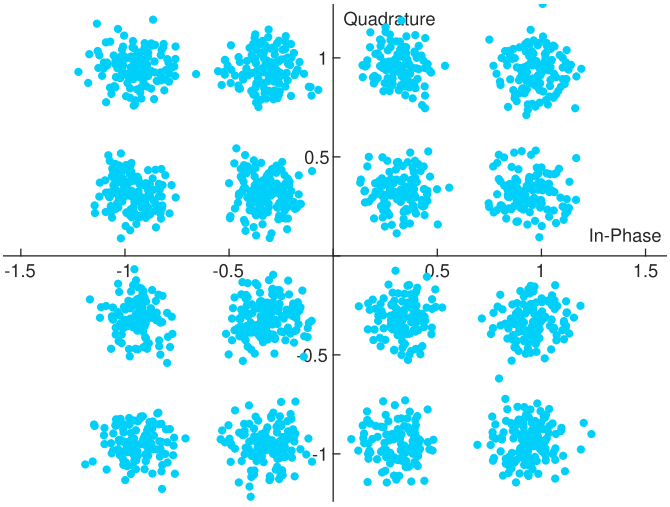


Fig. 5. Received I/Q data for 16-QAM with 18 dB SNR in an AWGN channel.

The implementation specifics and trade-offs for the individual stages will be discussed in the sections that follow. Note that all the clustering algorithms mentioned in Section 2 will be implemented to be evaluated for suitability in our proposed AMC method, as detailed later in Section 5.

4.1. Order estimation stage

The first stage in this study's proposed method, requires the modulation order to be estimated. The modulation order is used as an input for the next stage that will re-cluster the data to estimate the symbol levels. Fig. 5 shows the I/Q plane representation of the received samples for a 16-QAM modulation scheme transmitted through an AWGN channel. Examining these data points, two observations can be made: Firstly, the received symbols are scattered around certain centroids, easily identifiable due to the presence of sufficient Signal to Noise Ratio (SNR) in the example. The number of centroids correspond with the modulation scheme's order, also referred to as the number of symbol levels. The algorithms presented in Section 2 are used to determine how many of these reference points are present by clustering the received symbols. Unknown-order clustering algorithms will be capable of determining the modulation order directly, based on density. Known-order clustering algorithms will, however, require a post-processing step.

4.1.1. The elbow method, for known-order algorithms

The elbow method can be used to determine the optimal numerical parameter in a system by examining the change in variance as this parameter changes (Ketchen & Shook (1996)). Therefore, this method can be used to determine the number of symbol levels present by clustering the received symbols into each possible

number of symbol levels in a pool with a known-order clustering algorithm such as k-means, k-medoids, or fuzzy c-means. Where the total distance error is calculated and plotted as a function of the number of symbol levels present in the pool, a sharp elbow will form that indicates the probable number of symbol levels in the pool.

Several methods exist to identify the elbow, such as the L-method where two straight lines are fit to the data or evaluating the curvature (Antunes et al. (2018)). These methods can, however, be computationally complex. We adopt a simpler method that is similar to the approach used by Handaka, Wijaya, and Muljono (2018), exploiting the property that modulation orders come in 2^n steps. This property allows for easier identification of the elbow due to a large separation between the possible number of clusters. We calculate what percentage the distance error changes for each number of clusters relative to the preceding one. The largest percentage difference indicates the position of the elbow.

4.1.2. Fuzzy c-means with fuzzy overlap

The degree of membership returned by the fuzzy c-means algorithm indicates how well the cluster centroids fit the data for a certain number of clusters. Therefore, the data is clustered into the highest modulation order in the pool to determine the most likely number of clusters. The degree of membership is then summed for all data points belonging to each cluster to determine each cluster's average degree of membership. Clusters are then merged where a large degree of membership is shown to more than one cluster. This is repeated until all points show predominant membership to only one cluster.

4.1.3. DBSCAN

Fig. 1(a) and (b) demonstrated how only certain combinations of ϵ and MinPts deliver the correct number of clusters, as discussed in Section 2. To determine this region of correct MinPts and ϵ, MinPts is set to slightly less than the number of samples divided by the largest modulation order in the pool. This allows MinPts to be slightly less than the number of samples in each cluster in the case where the largest modulation order in the pool is the correct one. DBSCAN is then applied with a constant MinPts, and ϵ is incremented from 0 at a set rate until a number of clusters greater than 0 is achieved- this value for ϵ represents the minimum ϵ value. The same process is repeated from an overestimated ϵ, where ϵ is decremented to find the maximum value for ϵ. If one of these points cannot be determined, MinPts is lowered and the process is repeated. The maximum number of clusters between these minimum and maximum values for ϵ is then accepted as the estimate for the number of clusters.

4.1.4. OPTICS

The only parameters required for OPTICS is the data to be clustered, and the minimum number of points to be considered as a cluster. The order is determined by counting the number of peaks larger than the threshold, as was discussed in Section 2.1.2. The minimum number of points is calculated in the same way as with DBSCAN.

4.1.5. Hierarchical clustering

The number of clusters can be determined by examining the merging distances represented by vertical lines in a dendrogram, as shown in Fig. 3. The merging distance is plotted as a function of the number of clusters, after which the elbow method is applied.

4.2. Final clustering stage

The I/Q symbols can be clustered together in order to determine the used modulation scheme's symbol levels in the I/Q plane with the estimated parameters from the Order Estimation stage (Section 4.1). These estimated symbol levels can then be used with a classifier in the final stage in order to determine the modulation scheme.

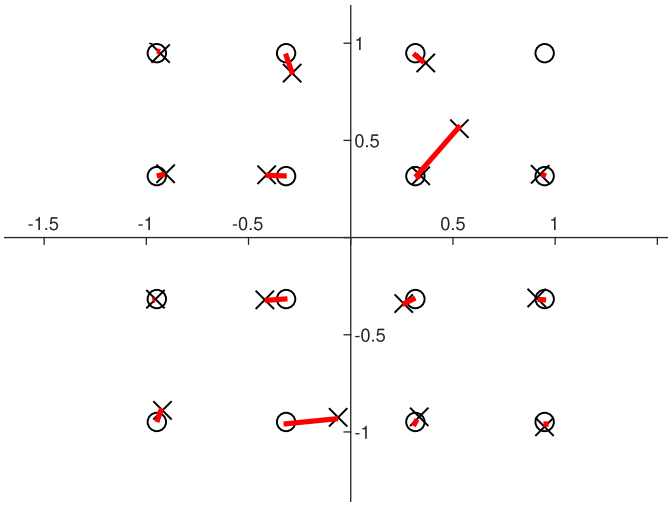


Fig. 6. Minimum distance classifier. Estimated locations are marked with an X while actual locations are marked with an O.

Density-based clustering algorithms typically do not provide the centroids of clusters as an output. The proposed method requires cluster centroids in the Final Clustering stage to estimate the symbol levels. Therefore, we calculate the centroid of each cluster for these density-based algorithms as the mean location of all points associated with each cluster.

The output of the Final Clustering stage is the estimated symbol levels. This ouput is fed to the Modulation Type Classification stage, discussed next. Alternative classification algorithms can also be used by considering the output of the Final Clustering stage as an input feature (Ali et al. (2017)).

4.3. Modulation type classification stage

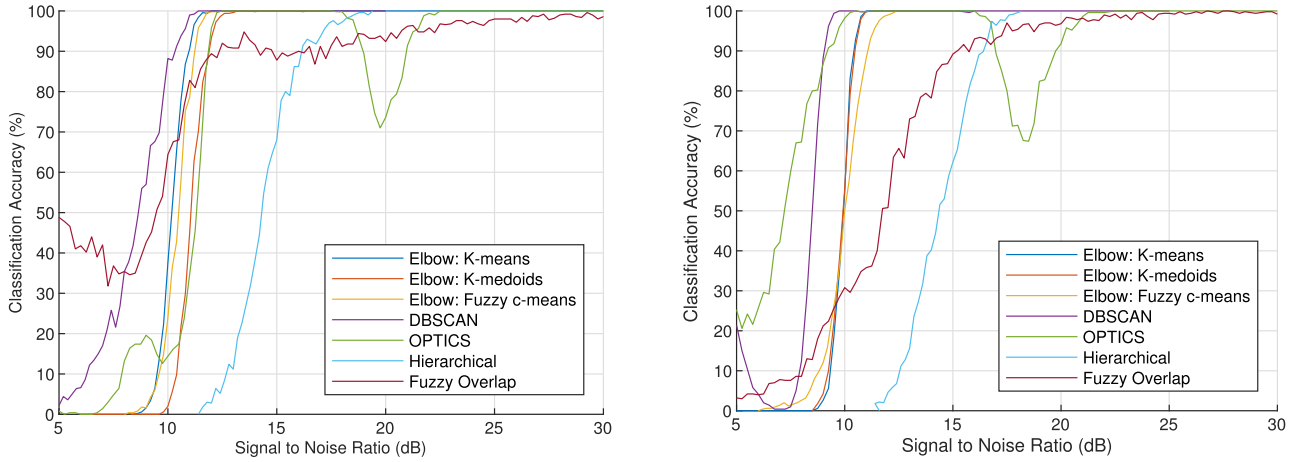
The Modulation Type Classification stage compares the estimated symbol levels from the Final Clustering stage with symbol levels of known modulation schemes in a pool. We will use a simple, deterministic, minimum distance classifier to keep complexity and execution time low. As shown in Fig. 6, the distance between each estimated symbol level (X) and its closest reference symbol level (O) can be calculated and summed together for a total distance error for each modulation scheme with the same number of symbols. The modulation scheme with the lowest score is then accepted as the most likely used modulation scheme.

5. Evaluation

5.1. Method and parameters

The aforementioned clustering algorithms and simulation model was implemented in the Matlab™ 2018a simulation environment, with Linux Mint 18.3 as host operating system, and run on a computer with an i7 6500u clocked at 2.6 GHz.

Refer to Fig 4. Randomly-generated bit streams are respectively modulated with 8-PSK or 16-QAM modulations in an AWGN channel. An SNR step size of 0.2dB was used with each SNR level repeated 500 times. The number of samples was varied between 1024, 2048, and 4096 samples. The possible pool of modulation schemes were M-QAM and M-PSK, with $M \in \{2, 4, 8, 16, 32, 64, 128, 256\}$ for all simulations.



(a) 16-QAM. (b) 8-PSK.

Fig. 7. Order estimation accuracy for 2048 samples in an AWGN channel.

Table 1
SNR at which 90% order estimation accuracy is achieved (dB).

	8-PSK			16-QAM		
Algorithm	1024 Samples	2048 Samples	4096 Samples	1024 Samples	2048 Samples	4096 Samples
Elbow: K-means	11.0	10.6	10.4	11.6	11.0	10.6
Elbow: K-medoids	11.0	10.8	10.4	12.2	11.8	11.6
Elbow: Fuzzy c-means	11.2	11.2	15.0	11.8	11.2	13.6
DBSCAN	9.5	9.2	9.5	9.6	10.4	10.5
OPTICS	12.0	9.2	8.2	12.0	11.8	12.0
Hierarchical	16.0	16.4	16.6	16.4	16.2	16.4
Fuzzy overlap	17.0	15.3	15.0	30.0	12.5	12.5

The I/Q data stream from the AWGN channel is first fed to the various stages of the proposed AMC system to evaluate the suitability of each clustering technique for the relevant stage. Section 5.2 details the comparison of the clustering algorithms for the Order Estimation stage and Section 5.3 evaluates the algorithms for the Final Clustering stage.

With suitable clustering algorithms identified for the relevant stages of the proposed AMC system, the performance of the proposed AMC system is evaluated in Section 5.4. In Section 5.5 the proposed AMC system is compared with popular AMC methods found in related work. The I/Q data stream from the AWGN channel is fed to the proposed AMC system as shown in Fig. 4 for the system-level analyses.

Small deviations from the method or parameters are applied in some of the evaluations. Where applicable this will be discussed together with the relevant evaluation.

5.2. Modulation order estimation accuracy and execution time evaluation

We evaluate the clustering algorithms in respect of accuracy as a function of SNR in an AWGN channel and the number of samples in this section.

Fig. 7(a), (b), and Table 1 illustrate that DBSCAN achieved the highest accuracy for modulation order estimation at low SNRs. Table 2 shows that hierarchical clustering has the lowest execution time for a various number of samples when all other variables are kept constant. Comparing each algorithm's complexity with its execution time reveals that there is little correlation between low complexity and execution time, mainly as a result of parameter estimation and iteration requirements. Table 1 shows the SNR in dB where 90% order estimation accuracy for 8-PSK and 16-PSK is achieved for various sample lengths.

The required minimum number of samples depends on the highest modulation order in the pool and the desired accuracy. More samples are required for higher-order modulation schemes to allow a sufficient number of samples per symbol level. Even though DBSCAN demonstrates the best accuracy at low SNR levels, Table 2 shows that it does not scale well when increasing the number of samples. K-means together with the elbow method is, however, only slightly less accurate and scales more linearly when keeping the number of allowed iterations constant, which is confirmed by its Big O complexity.

OPTICS and DBSCAN exhibit interesting behaviour due to their ability to identify noise, which can be seen in Fig. 7(a) and (b), where OPTICS undergoes a drop in accuracy after achieving 100% classification accuracy. However, for both DBSCAN and OPTICS, this behaviour dissipates with a larger number of samples.

5.3. Final clustering accuracy and execution time evaluation

The true number of clusters is used for *known-order* clustering algorithms without estimation for evaluation purposes while the estimated parameters from the order estimation stage (see Fig. 4) are used for *unknown-order* clustering algorithms. OPTICS allows the value of ϵ to be chosen after the data has been clustered. This allows for either the use of a value for ϵ that has been determined in the previous stage or for choosing one that results in a set order. The unknown order clustering algorithms are only tested until the data cannot be clustered into the correct number of clusters, as the estimated symbol levels at that point no longer represent the data.

The clustering algorithms were evaluated with the parameters and method as described in Section 5.1. For brevity, only the 16-

Table 2
Order Estimation complexity and execution time comparison with 16-QAM.

Algorithm	Complexity[a]	Run Time (s)		
		1024 Samples	2048 Samples	4096 Samples
Elbow: K-means	$O(knt)$	0.998	2.346	5.737
Elbow: K-medoids	$O(k(n-k)^2)$	17.610	82.984	340.963
Elbow: Fuzzy c-means	$O(n)$	3.923	11.236	24.985
DBSCAN	$O(n*logn)$	0.968	3.666	14.785
OPTICS	$O(n*logn)$	0.172	0.536	1.915
Hierarchical	$O(n)$	0.014	0.0775	0.420
Fuzzy overlap	$O(n)$	9.029	25.128	54.950

[a] *Source:* Xu and Tian (2015).

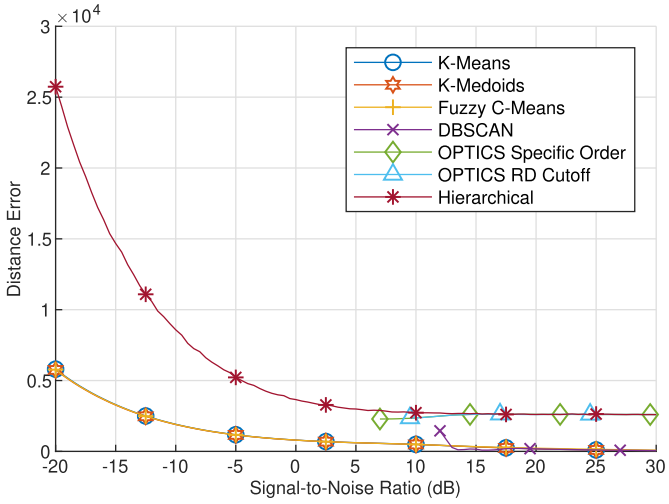


Fig. 8. Final Clustering accuracy for 16-QAM.

QAM results are discussed here, as analysis of the 8-PSK results yielded similar findings. Evaluation metrics are the algorithm execution time and the clustering accuracy, expressed in terms of the distance error.

The distance error (e_d) is defined as the sum of the Euclidean distances between all I/Q samples and their assigned cluster's centroid as given by:

$$e_d = \sum_{k=1}^{N} \sqrt{(C_{di} - D_{ki})^2 + (C_{dq} - D_{kq})^2}, \quad (1)$$

where N is the number of samples, D_k is an individual I/Q sample and C_d is the sample k's assigned cluster's centroid. In each instance the subscripts i and q denote the In-Phase and Quadrature values, respectively. The distance error is calculated for each sample point evaluated and averaged over the 500 repetitions.

As seen in Fig. 8, all known-order clustering algorithms are comparable in accuracy when determining symbol levels. The unknown-order clustering algorithms, other than DBSCAN, are also equally accurate, with the only distinction being the lowest SNR where the algorithm can be used. Table 3 shows that, for the Final Clustering stage, hierarchical clustering has the lowest execution time and that k-means scales the best with the number of samples when keeping all other variables constant. This corresponds with the trend observed in Section 5.2.

5.4. Performance evaluation of the proposed AMC method

The best performing algorithms from the comparative evaluations in Sections 5.2 and5.3 are utilised to complete the proposed AMC method as shown in Fig. 4.

Table 3
Final Clustering execution time comparison with 16-QAM (s).

Algorithm	Number of samples				
	512	1024	2048	4096	8192
K-means[a]	0.115	0.179	0.269	0.691	2.408
K-medoids[a]	0.148	0.730	2.712	11.256	59.312
Fuzzy c-means[a]	0.053	0.129	0.300	0.650	1.592
DBSCAN[b]	0.016	0.671	2.599	12.844	53.757
OPTICS[b](Specific Order)	0.072	0.240	0.779	3.536	23.701
OPTICS[b](RD Cutoff)	0.070	0.236	0.777	3.437	24.768
Hierarchical[b]	0.004	0.010	0.043	0.282	2.131

[a] Known-order: Actual number of clusters used as input.
[b] Unknown-order: Estimated number of clusters used as input.

The elbow method with k-means is proposed for the Order Estimation stage and k-means for the Final Clustering stage. The k-means clustering algorithm was chosen for its accuracy and scaling with a larger number of samples, as revealed in the comparative evaluations. The Modulation Type Classifier stage is implemented as described in Section 4.3.

The proposed AMC method was evaluated with the parameters and method as described in Section 5.1, and, the number of samples was varied between 512, 1024, 2048, 4096, and 8192 samples. Fig. 9(a) and (b) show the classification accuracy for 16-QAM and 8-PSK respectively in an AWGN channel with a varying sample length.

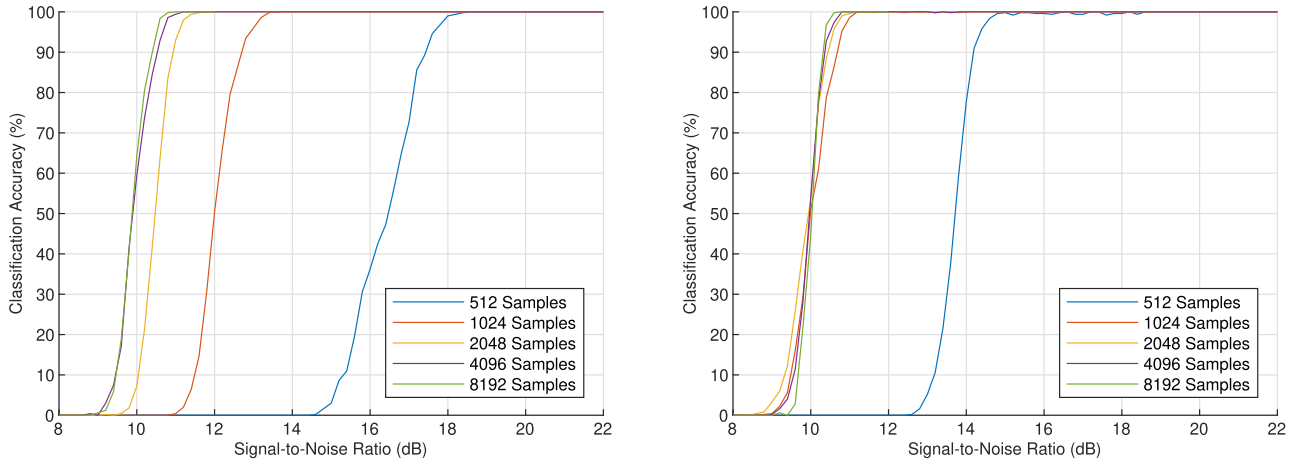
Fig. 9(a), (b), and Table 1 illustrate that the required minimum number of samples depends on the highest possible order in the pool of possible modulation schemes. Therefore, a certain minimum number of samples for each symbol level is required for clustering algorithms to be viable for AMC. However, the accuracy gain does not scale linearly when increasing the number of samples, leading to an optimal number of samples that depends on the highest order modulation scheme in the pool.

5.5. Comparison with popular AMC methods

In 2015, Zhu and Nandi (2015) published "Automatic Modulation Classification Principles, Algorithms and Applications" in which the most popular AMC methods were simulated on the same test platform to provide an equal comparison between the following methods:

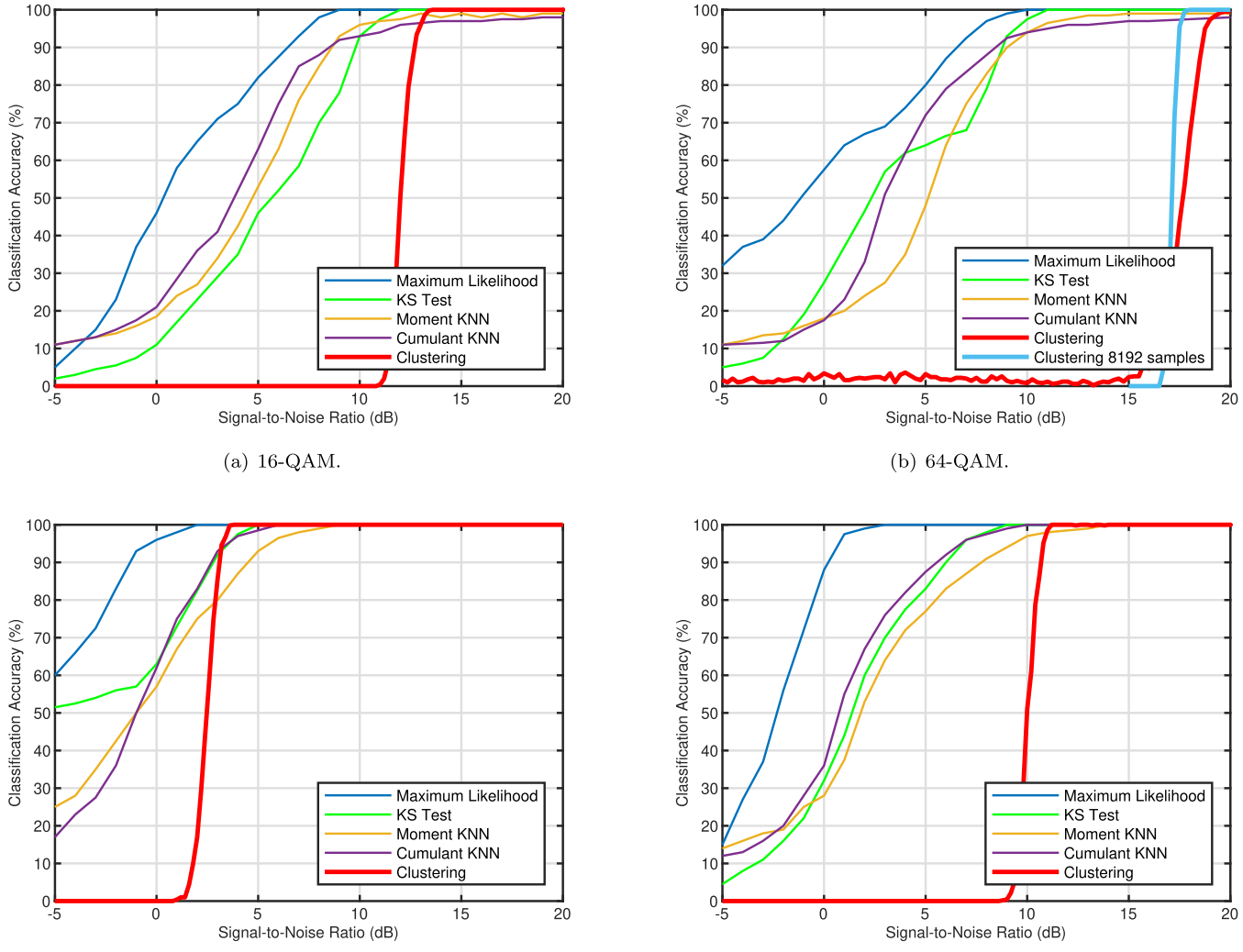
- The Maximum Likelihood Ratio Test (MLRT)
- The Kolmogorov Smirnov (KS) Test
- High order cumulants with a K-Nearest Neighbour (KNN) classifier
- High order moments with a KNN classifier

Zhu and Nandi (2015) did not examine clustering techniques as these methods are relatively unexplored in literature. In Zhu's comparison between the above mentioned AMC methods, 1024 samples were used for all methods and modulation orders. Accu-



(a) 16-QAM. (b) 8-PSK.

Fig. 9. Classification accuracy of proposed AMC method at various sample sizes.



(a) 16-QAM. (b) 64-QAM.

(c) Q-PSK. (d) 8-PSK.

Fig. 10. Classification accuracy of proposed AMC method (show in red) vs. popular AMC methods.

rate channel information was also used where required in order to represent ideal situations for each AMC method. The pool of possible modulation schemes evaluated were 2-PAM, 4-PAM, 8-PAM, BPSK, QPSK, 8-PSK, 4-QAM, 16-QAM, and 64-QAM. Execution times were not stated in the related work.

The proposed AMC method was evaluated with the parameters and method as described in Section 5.1, however, the sample length and modulation pool were modified to correspond with the parameters as described above.

Fig. 10(a)–(d) show the classification accuracy for 16-QAM, 64-QAM, Q-PSK, and 8-PSK respectively for the proposed method (shown in red) in comparison with the four AMC methods described above.

The proposed clustering-based method's classification accuracy drops off at a faster rate in low SNR situations when compared with the other methods. The proposed method is, however, capable of 100% classification accuracy at a similar SNR for lower order modulation schemes.

A larger number of samples is recommended for higher-order modulation schemes like 64-QAM as each cluster will contain fewer samples. For example, if 1024 samples are used, 8-PSK will have 128 samples for each symbol level while 64-QAM will only have 16, assuming each symbol level appears equally often. The effect of increasing the number of samples for the proposed method is shown in Fig. 10(b) for 8192 samples (shown in cyan). Increasing the sample size improves the classification accuracy.

The nature of the operation of the elbow method allows it to also give a measurement of how intact the underlying geometry of the constellation remained during transmission depending on how pronounced the elbow is. This behaviour can allow the proposed method to fail in such a manner, that when the SNR is too low, a *could not classify* state is given rather than a false positive. This will allow the proposed method to classify modulation schemes with a higher degree of confidence in sub-optimal environments than related work.

6. Conclusion

Clustering algorithms for the use of AMC in the I/Q plane are relatively unexplored in literature. The evaluation conducted in Section 5 show that clustering algorithms can be used in various ways that deliver benefits that other AMC methods do not.

In Sections 5.2 and 5.3, the order estimation accuracy and classification accuracy were evaluated as functions of SNR and the number of samples. The best performing clustering algorithms from each section were then incorporated into the proposed AMC method, which was evaluated for classification accuracy in Section 5.4. The results show that increasing the number of samples allows for accurate classification at lower SNR levels.

In Section 5.5, the proposed method was compared with other popular non-clustering feature- and likelihood-based AMC methods in literature. Fig. 10(a)–(d) show that clustering-based techniques have a sharper drop-off in accuracy compared to MLRT, the KS test, and moments and cumulants with a KNN classifier. These figures also show that the proposed method achieves 100% classification accuracy at a similar SNR compared to other methods, if a sufficient number of samples is used, but also shows a sharp drop-off at low SNR.

In summary, the proposed method does not rely on machine learning, nor does it require thresholds to be set, resulting in a highly adaptable AMC method. Furthermore, the proposed method relies on the k-means algorithm that is well known for its relatively low algorithmic complexity and linear scalability. This allows the proposed method to be adapted for situations where execution time is preferred over classification accuracy or vice versa by changing the number of samples. The proposed method is also made deterministic in its number of operations by limiting the number of allowed iterations of the k-means algorithm. A large pool of modulation schemes can be used without significantly increasing the execution time of the proposed method. Future work will focus on exploring the proposed method's ability to detect false positives and evaluating the performance under the European Telecommunications Standards Institute (ETSI) multipath channel models.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Credit authorship contribution statement

Jacques P. Mouton: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Melvin Ferreira:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing - review & editing. **Albertus S.J. Helberg:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing - review & editing.

Acknowledgements

The authors gratefully acknowledge the financial support of this study by the Telkom CoE at the NWU, GEW Technologies and the National Research Foundation under grant number TP14081892668.

References

Ahmadi, N. (2010). Using fuzzy clustering and TTSAS algorithm for modulation classification based on constellation diagram. *Engineering Applications of Artificial Intelligence, 23*(3), 357–370. doi:10.1016/j.engappai.2009.05.006.

Ali, A., Yangyu, F., & Liu, S. (2017). Automatic modulation classification of digital modulation signals with stacked autoencoders. *Digital Signal Processing, 71*. doi:10.1016/j.dsp.2017.09.005.

Ankerst, M., Breunig, M. M., Kriegel, H., & Sander, J. (1999). OPTICS: ordering points to identify the clustering structure. In *Proceedings of the ACM SIGMOD international conference on management of data* (pp. 49–60). ACM. doi:10.1145/304182.304187.

Antunes, M., Gomes, D., & Aguiar, R. L. (2018). Knee/elbow estimation based on first derivative threshold. In *Proceedings of the IEEE fourth international conference on big data computing service and applications (bigdataservice)* (pp. 237–240). IEEE. doi:10.1109/BigDataService.2018.00042.

Arthur, D., & Vassilvitskii, S. (2007). K-Means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms* (pp. 1025–1027). ACM.

Azarmanesh, O., & Bilén, S. G. (2013). I-Q diagram utilization in a novel modulation classification technique for cognitive radio applications. *EURASIP Journal on Wireless Communications and Networking*, (289). doi:10.1186/1687-1499-2013-289.

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences, 10*, 191–203. doi:10.1016/0098-3004(84)90020-7.

Cordova, I., & Moh, T. (2015). Dbscan on resilient distributed datasets. In *Proceedings of the international conference on high performance computing simulation (hpcs)* (pp. 531–540). doi:10.1109/HPCSim.2015.7237086.

Corizzo, R., Pio, G., Ceci, M., & Malerba, D. (2019). Dencast: Distributed density-based clustering for multi-target regression. *Journal of Big Data, 6*(1), 43.

Daszykowski, M. (2002). Matlab code for OPTICS. 10.13140/RG.2.1.3998.3843

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the second international conference on knowledge discovery and data mining, KDD* (pp. 226–231). AAAI Press.

Guldemr, H., & Sengur, A. (2006). Comparison of clustering algorithms for analog modulation classification. *Expert Systems with Applications, 30*(4), 642–649. doi:10.1016/j.eswa.2005.07.014.

Handaka, S., Wijaya, E., & Muljono (2018). The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In

Proceedings of the international seminar on application for technology of information and communication (ISEMANTIC), (pp. 533–538). doi:10.1109/ISEMANTIC.2018.8549751.

He, Y., Tan, H., Luo, W., Mao, H., Ma, D., Feng, S., & Fan, J. (2011). Mr-dbscan: An efficient parallel density-based clustering algorithm using mapreduce. In *Proceedings of the IEEE 17th international conference on parallel and distributed systems* (pp. 473–480). doi:10.1109/ICPADS.2011.83.

ITU (2006). Final acts of the regional radiocommunication conference for planning of the digital terrestrial broadcasting service in parts of regions 1 and 3, in the frequency bands 174-230 mhz and 470-862 mhz. In *Proceedings of the RRC.*

Jajoo, G., Kumar, Y., Yadav, S., Adhikari, B., & Kumar, A. (2017). Blind signal modulation recognition through clustering analysis of constellation signature. *Expert Systems with Applications, 90*, 13–22. doi:10.1016/j.eswa.2017.07.053.

Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal, 17*(6), 441–458. doi:10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129–137. doi:10.1109/TIT.1982.1056489.

Mao, Y., Xu, Z., Li, X., & Ping, P. (2015). An optimal distributed k-means clustering algorithm based on cloudstack. In *Proceedings of the IEEE international conference on information and automation* (pp. 3149–3156). doi:10.1109/ICInfA.2015.7279830.

Mobasseri, B. G. (2000). Digital modulation classification using constellation shape. *Signal Processing, 80*(2), 251–277. doi:10.1016/S0165-1684(99)00127-9.

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science, 2*(2), 165–193. doi:10.1007/s40745-015-0040-1.

Yarpiz (2015). Implementation of DBSCAN clustering in Matlab. https://yarpiz.com/255/ypml110-dbscan-clustering.

Zhu, X., & Fujii, T. (2016). A novel modulation classification method in cognitive radios based on features clustering of time-frequency. In *Proceedings of the IEEE radio and wireless symposium (RWS)* (pp. 45–47). IEEE. doi:10.1109/RWS.2016.7444364.

Zhu, Z., & Nandi, A. K. (2014). Blind digital modulation classification using minimum distance centroid estimator and non-parametric likelihood function. *IEEE Transactions on Wireless Communications, 13*(8), 4483–4494. doi:10.1109/TWC.2014.2320724.

Zhu, Z., & Nandi, A. K. (2015). *Automatic modulation classification principles, algorithms and applications* (1st). Wiley.