A Novel Clustering Algorithm based on a Non-parametric "Anti-Bayesian" Paradigm

Hugo Lewi Hammer[1], Anis Yazidi[1], and B. John Oommen[2⋆]

[1] Department of Computer Science, Oslo and Akershus University College of Applied Sciences, Norway
[2] School of Computer Science, Carleton University, Ottawa, Canada

Abstract. The problem of clustering, or unsupervised classification, has been solved by a myriad of techniques, all of which depend, either directly or implicitly, on the Bayesian principle of optimal classification. To be more specific, within a Bayesian paradigm, if one is to compare the testing sample with only *a single* point in the feature space from each class, the *optimal* Bayesian strategy would be to achieve this based on the distance from the corresponding means or *central* points in the respective distributions. When this principle is applied in clustering, one would assign an unassigned sample into the cluster whose mean is the closest, and this can be done in either a bottom-up or a top-down manner. This paper pioneers a clustering achieved in an "Anti-Bayesian" manner, and is based on the breakthrough classification paradigm pioneered by Oommen *et al.* The latter relies on a radically different approach for classifying data points based on the non-central *quantiles* of the distributions. Surprisingly and counter-intuitively, this turns out to work equally or close-to-equally well to an optimal supervised Bayesian scheme, which thus begs the natural extension to the unexplored arena of clustering. Our algorithm can be seen as the Anti-Bayesian counter-part of the well-known k-means algorithm[3], where we assign points to clusters using quantiles rather than the clusters' centroids. Extensive experimentation[4] demonstrates that our Anti-Bayesian clustering converges fast and with precision results competitive to a k-means clustering.

1 Introduction

Clustering is a key task in data analysis [2] [3]. There exist a range of different clustering methods that vary in the understanding of what a cluster is. For

⋆ *Chancellor's Professor*; *Fellow: IEEE* and *Fellow: IAPR*. This author is also an *Adjunct Professor* with the University of Agder in Grimstad, Norway.

[3] The fundamental Anti-Bayesian paradigm need not just be used to the k-means principle. Rather, we hypothesize that it can be adapted to any of the scores of techniques that is *indirectly* based on the Bayesian paradigm.

[4] This paper contains the *prima facie* results of experiments done on one and two-dimensional data. The extensions to multi-dimensional data are not included in the interest of space, and would use the corresponding multi-dimensional Anti-Naïve-Bayes classification rules given in [1].

instance, density models, such as OPTICS [4] and DBSCAN [5], find the most dense regions in the space. As opposed to this, in hierarchical clustering [6] [7], the aim is to arrange the data points in an underlying hierarchy. A third group of clustering algorithms constitute the so-called "centroid" methods where each cluster is represented by a single point. The most prominent example of a scheme within this family is the k-means clustering algorithm where a centroid is represented by the mean value of the points in the cluster. The central strategy motivating *these* clustering schemes involves classifying data points to the different clusters based on the distances to the means (or centroids) of the clusters.

In this paper we introduce a novel alternative to the k-means clustering algorithm. We shall demonstrate that we can obtain excellent clustering performance by operating in a diametrically opposite way, i.e., a so-called Anti-Bayesian manner. Indeed, we shall show the completely counter-intuitive result that by working with a few points distant from the mean (centroid), one can obtain remarkable clustering performances. While the clustering algorithm in this paper follows the steps of a typical k-means clustering algorithm, it assigns the data points to the already-formed clusters using completely different criteria – by invoking the concepts of Anti-Bayesian Pattern Recognition (PR). Rather, unlike the k-means clustering strategies which rely on centroid-based criteria, our paradigm advocates the association of points to clusters based on *quantiles distant from the cluster means* [8] [1] [9], which is a concept that is unreported in the literature. Indeed, it is actually both un-intuitive and non-obvious.

We consider the non-parametric clustering problem where the distribution of each cluster is unknown. This is in contrast to the work in [8] [9] reported in the area of classification where the distributions are known, and more in the directions of [1], where a non-parametric case is considered. In [1] the quantiles are estimated by assuming that the data points are sampled from Gaussian distributions. In this paper we work with a totally distribution-free model which is more natural when the aim is to achieve flexible and robust clustering.

In this paper, in the interest of space and brevity, we merely propose the Anti-Bayesian clustering strategy for one and two-dimensional data. This is because the aim of the paper is to introduce the paradigm in a *prima facie* setting. Although the extensions to multi-dimensional data are still open, they would not be too complicated – they would use the corresponding multi-dimensional Anti-Naïve-Bayes classification rules given in [1].

2 Anti-Bayesian Clustering

Let $x_1, x_2, \ldots, x_n$ represent n points in $\mathbb{R}^p$. In this paper we present an algorithm, which is based on the Anti-Bayesian classification framework [8] [1] [9], to cluster these points into k clusters. The present algorithm follows the same steps as a k-means clustering algorithm. To motivate this discussion, we start this section with a short review of the classical k-means algorithm. Then, in Section 2.2 we briefly review the Anti-Bayesian classification framework which forms the basis for the Anti-Bayesian clustering algorithm presented in Section 2.3. Before we

proceed, we re-iterate our earlier comment: Since the fundamental Anti-Bayesian paradigm need not just be used to the k-means principle, our paradigm can be adapted to any of the scores of techniques that are *inherently* Bayesian.

2.1 k-means Clustering

Let $m_1, m_2, \ldots, m_k \in \mathbb{R}^p$ represent the respective centroids of each of the k clusters currently described, $C_1, \ldots, C_k$. The algorithm starts by associating an initial value to each centroid. A simple approach is to assign each of centroids to some of the points $x_1, x_2, \ldots, x_n$. The algorithm then consists of two steps:

1. **Assignment:** Assign each point x_i to the nearest[5] centroid.
2. **Update:** When all points are assigned to a cluster, update the centroid value of each cluster as the average of all the points in the cluster:

$$m_j = \frac{1}{|C_j|} \sum_{i=1}^{|C_j|} x_{j(i)}, \quad j = 1, \ldots, k,$$

   where $|C_j|$ is the number of points in cluster C_j and $j(1), j(2), \ldots, j(|C_j|)$ represent the points assigned to cluster C_j.
3. **Loop:** Repeat the above two steps until no points switch their clusters.

2.2 Anti-Bayesian Classification

The following classification method is based on the "Anti-Bayesian" methodology described and proven in [8] [1] [9]. To explain how it works, we start with the uni-dimensional case.

Let $x_1, x_2, \ldots, x_{n_1}$ and $y_1, y_2, \ldots, y_{n_2}$ be random samples from some unknown probability distributions $f_X(x)$ and $f_Y(y)$, respectively. Our task is to classify a new point z to see whether it is a sample from $f_X(x)$ or $f_Y(y)$. If we assume that the variances of $f_X(x)$ and $f_Y(y)$ are equal, the optimal classification strategy is to assign z to $f_X(x)$ or $f_Y(y)$ if z is closer to the means of X or Y respectively. This, in turn, assigns z to $f_X(x)$ if the average of the samples $x_1, x_2, \ldots, x_{n_1}$ is closer to z than the average of $y_1, y_2, \ldots, y_{n_2}$. It is otherwise assigned to $f_Y(y)$. The reader should observe this is precisely how points are assigned to clusters in the k-means paradigm.

In the Anti-Bayesian classification approach, classification is achieved based on quantile-based comparisons rather than comparisons with regard to the mean. To render this formal, we denote the quantiles as follows: $q_p^X = P(X > p)$. Although, in practice, the quantiles have to be, estimated (or learned), for ease of clarification, in the descriptions below, we assume that the quantiles are known. We also assume that $q_{1-p}^X < \text{Median}(X)$ so that q_p^X is always greater than q_{1-p}^X. In such a case, the Anti-Bayesian classification method operates as follows:

[5] It should be mentioned that the concept of "nearest" can be based on the specific metric being used, for example, a simple Euclidean metric.

1. Determine which of the distributions $f_X(x)$ or $f_Y(y)$ is to the left by using the quantiles of the distributions. We have three possible cases:
   Case 1: If $q_p^X < q_p^Y$ and $q_{1-p}^X < q_{1-p}^Y \implies f_X(x)$ is to the left of $f_Y(y)$.
   Case 2: If $q_p^X > q_p^Y$ and $q_{1-p}^X > q_{1-p}^Y \implies f_Y(y)$ is to the left of $f_X(x)$.
   Case 3: Else, we determine their relative positions by comparing the averages of the quantiles as follows:
      If $\frac{q_p^X + q_{1-p}^X}{2} < \frac{q_p^Y + q_{1-p}^Y}{2} \implies f_X(x)$ is to the left of $f_Y(y)$.
      Else[6] $f_Y(y)$ is to the left of $f_X(x)$.
   Figure 1 depicts the above three cases. We see that for Cases 1 and 2, $f_X(x)$ and $f_Y(y)$ is the distribution to the left, respectively. In the bottom figure (Case 3), the decision is not that obvious because the classes are highly overlapping.

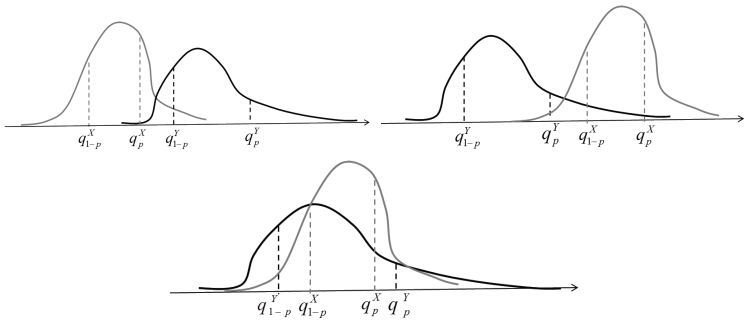

Fig. 1. This figure depicts Cases 1, 2 and 3 – arranged from left to right and from top to bottom respectively.

2. Once the relative positions of the distributions are determined, the classification rule must now be specified. For simplicity, we describe this just for Case 1 since the rules for the "mirrored" cases are analogous. The Anti-Bayesian rule classifies using the *right* quantile of the left distribution and the *left* quantile of the right distribution. If $B = \frac{q_p^X + q_{1-p}^Y}{2}$, we classify as follows:
   If $z < B$, classify z to $f_X(x)$.
   Else, classify z to $f_Y(y)$.
   This approach works even when the distributions overlap such that q_{1-p}^Y is to the left of q_p^X as shown in Figure 2.

The theoretical motivation for this algorithm is given in [8] and [9] and not repeated here.

We now go over to the case where the points are two dimensional. Let $x_{11}, \ldots, x_{n_1 1}$ and $x_{12}, \ldots, x_{n_1 2}$ be $2n_1$ independent samples from $f_X(x)$ and define the points $(x_{i1}, x_{i2}),\ i = 1, 2, \ldots n_1$. Similarly, define the points $(y_{i1}, y_{i2}),\ i =$

[6] This case occurs rarely in practice except when the classes are highly overlapping, in which case the PR problem is often meaningless.

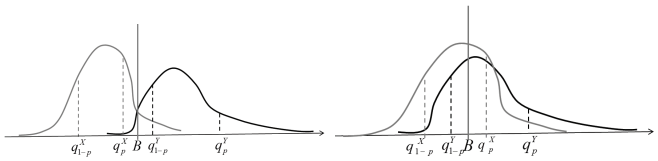

Fig. 2. The left panel shows the standard situation under Case 1, while the right panel shows a situation when q^Y_{1-p} is to the left of q^X_p.

$1, 2, \ldots n_2$ based on samples from $f_Y(y)$. Again we want to classify a point z to $f_X(x)$ or $f_Y(y)$. The classification is done as per the ideas in [1] and [9]. It is a natural generalization of the one dimensional case above and follows two steps.

1. Define the rectangle with corners (q^X_{1-p}, q^X_{1-p}), (q^X_{1-p}, q^X_p), (q^X_p, q^X_{1-p}) and (q^X_p, q^X_p) for $f_X(x)$, and the analogous rectangle corners for Y. Locate the corners in the two rectangles that are closest to each other.
2. If z is closer to the corner of the quantile rectangle of $f_X(x)$, classify z to $f_X(x)$. Else classify z to $f_Y(y)$.

Figure 3 shows the classification procedure for two typical cases.

Estimation of the Quantiles As described above, throughout this discussion, we have assumed that the distributions $f_X(x)$ and $f_Y(y)$ are unknown. Thus, in reality, the quantiles must be estimated from the samples. Let $x_{(1)}, x_{(2)}, \ldots, x_{(n_1)}$ be n_1 samples from $f_X(x)$ sorted by value. To achieve this estimation, we adapt the method recommended in [10] where the cumulative distribution function of $f_X(x)$ is estimated with a linear interpolation (spline) through the points $(x_{(i)}, p_i),\ i = 1, 2, \ldots, n_1$, and where p_i is defined as:

$$p_i = \frac{i - 1/3}{n_1 + 1/3},\ i = 1, 2, \ldots, n_1.$$

The estimate of q^X_p can then be easily read (or rather, inferred) from this curve.

2.3 Clustering based on Anti-Bayesian Classification

We now present a clustering algorithm that uses that Anti-Bayesian classification methodology presented in Section 2.2 combined with the assignment/update steps in the k-means clustering algorithm. Suppose that we have a set of points $x_1, x_2, \ldots, x_n$ that we want to group into k clusters denoted by $C_1, C_2, \ldots, C_k$. Let q^j_{1-p} and q^j_p represent the (1-p)-valued and p-valued quantiles of the data in cluster C_j. The algorithm starts by associating values to the quantiles of all the clusters $(q^j_{1-p},\ q^j_p),\ j = 1, 2, \ldots, k$. One way to achieve this is by randomly selecting $2k$ points and by associating the quantiles to the values of these points. To prevent an excessive initial overlapping of the clusters, it is natural to first sort the points before associating them to the quantiles. Similar to the k-means scheme, the present algorithm now consists of two steps:

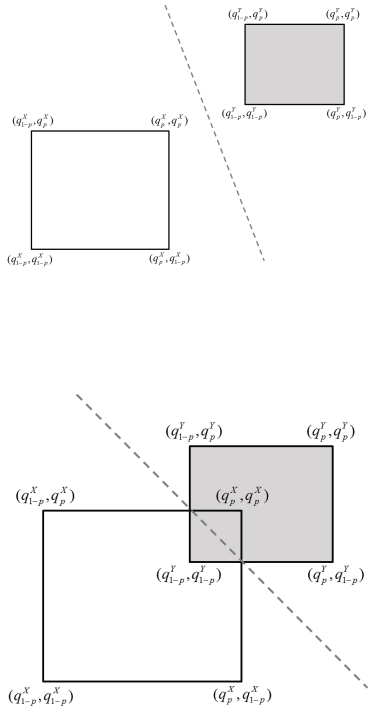

Fig. 3. Classification in the two dimensional scenario for two typical cases. The gray dashed line shows the border of the discriminant regions when z is classified to $f_X(x)$ or to $f_Y(y)$.

1. **Assignment:** Assign each point to a cluster. For each point, x_i, we do the following by repeated classifications using the methodology in Section 2.2. We start to determine if x_i is most likely to belong to C_1 or C_2. Assume that x_i is most likely to belong to C_2. We then say that C_2 is the current candidate cluster for x_i. Next we do an evaluation between C_2 and C_3 and repeat this for all the remaining clusters $C_4, \ldots, C_k$. In the last step we do an evaluation with the current candidate cluster from the previous evaluations, say C_a, and C_k. If x_i is more likely to belong to C_a, we assign x_i to this cluster, else we assign x_i to C_k.
2. **Update:** When all the points are assigned to clusters, we estimate the quantiles of all the clusters, i.e., $(q^j_{1-p}, q^j_p),\ j = 1, 2, \ldots, k$, using the estimator presented at the end of Section 2.2.

2.4 Evaluation of Clustering Performance

The question of how to measure the performance of a clustering algorithm is far from being obvious or trivial. Consider the following simple example. Suppose that six points $\{A_1, A_2, A_3, B_1, B_2, B_3\}$ are to be clustered into two clusters,

with the goal that the elements $\{A_i\}$ and $\{B_i\}$ are located in the same cluster. Consider now the case when the results of a specific clustering algorithm are:

Cluster C_1: $\{A_1, B_2, B_3\}$,

Cluster C_2: $\{A_2, A_3, B_1\}$.

If the requirements of the clustering problem required that all the elements $\{A_i\}$ are to be in cluster C_1, and that all the elements $\{B_i\}$ are to be in cluster C_2, we could immediately see that the number of errors incurred by the above clustering is 4. On the other hand, if the clustering problem merely stipulated that the $\{A_i\}$'s were to be in one cluster and that the $\{B_i\}$'s in another (irrespective of whether it is C_1 or C_2), the number of errors is 2. Since the latter is the more meaningful issue, in the experiments below, we evaluate the performance of a clustering algorithm by assuming that the cluster index is irrelevant as long as the elements that should belong together do, indeed, get clustered together.

3 Experiments

In this section we compare the performance of the k-means and Anti-Bayesian clustering algorithms when the data to be clustered was generated from a range of different distributions. For all the experiments we considered $k = 3$ clusters. The parameters of the distributions are shown in Table 1.

Distribution	Parameters Used	Shift: 1D Case	Shift 2D Case
Normal	$\mu = 0,\ \sigma = 1$	3, 6	2, 4
Uniform	$a = 0,\ b = 1$	9/10, 9/5	4/3, 8/3
Laplace	$\mu = 0,\ \lambda = 1/4$	4/5, 8/5	2/3, 4/3
Beta (symmetric)	$a = 2,\ b = 2$	5/8, 5/4	5/11, 10/11
Beta (asymmetric)	$a = 3,\ b = 1$	0.65, 1.29	0.42, 0.84
Gamma	$a = 3,\ s = 3$	2.13, 4.27	1.67, 3.33

Table 1. Distributions used in the experiments. The second column shows the parameters of the distributions of the leftmost cluster. The third and the fourth columns show how much the distribution is shifted to the right for cluster two and three for the one and two dimensional experiments.

A plot of all the distributions in the one dimensional experiment are shown in Figure 4. The distributions are the same for the two dimensional experiment, except that the distributions are less separated (shifted) from each other.

We considered two cases: In the first case, we generated $N = 10$ independent samples from each distribution, i.e. a total of 30 samples, and in the second where we generated $N = 1,000$ independent samples from each distribution. To evaluate the performance of the clustering algorithms, we first generated the synthetic data, performed the clustering and then measured the portion of samples that were classified to wrong cluster. This procedure was repeated a large amount of times to minimize the Monte Carlo error. In all the cases, we set $p = 1/3$ in the computation of the quantiles for the Anti-Bayesian clustering.

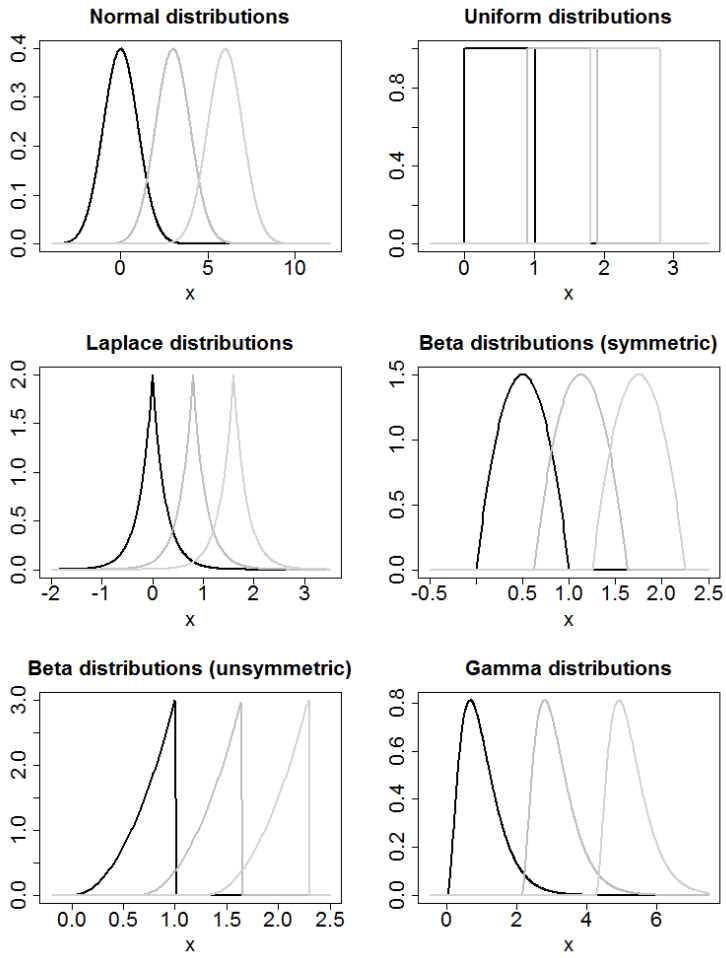

Fig. 4. Plot of all the distributions presented in Table 1 for the one dimensional case.

Tables 2 and 3 show the results for the one and two dimensional cases respectively. It is appropriate to mention that similar to the k-means algorithm, in some rare cases, the Anti-Bayesian algorithm did not converge, but cycled between a few different configurations. In such cases, we terminated it after 100 iterations and reported the final clusters. We also computed the 95% confidence intervals for the portions in Tables 2 and 3 and the width of these intervals were ≈ 0.002, implying that the Monte Carlo error was almost completely removed. These were thus not included in the tables.

There are many factors that distinguish the performances of the k-means and Anti-Bayesian clustering algorithms.

	$N = 10$		$N = 1000$	
Distribution	k-means	Anti-Bayes	k-means	Anti-Bayes
Normal	0.105	0.105	0.090	0.089
Uniform	0.116	0.104	0.106	0.157
Laplace	0.149	0.163	0.136	0.135
Beta (symmetric)	0.145	0.138	0.125	0.139
Beta (asymmetric)	0.081	0.087	0.074	0.098
Gamma	0.143	0.170	0.113	0.132

Table 2. Results from the one dimensional clustering experiment. The values show the portion of sample points that were classified to the wrong clusters. The values in parentheses represent the 95% confidence interval.

	$N = 10$		$N = 1000$	
Distribution	k-means	Anti-Bayes	k-means	Anti-Bayes
Normal	0.140	0.145	0.106	0.107
Uniform	0.102	0.102	0.074	0.078
Laplace	0.139	0.156	0.108	0.108
Beta (symmetric)	0.145	0.147	0.108	0.109
Beta (unsymmetric)	0.114	0.168	0.081	0.121
Gamma	0.141	0.202	0.102	0.124

Table 3. Results from the two dimensional clustering experiment. The values show the portion of sample points that were classified to the wrong clusters. The values in parentheses represent the 95% confidence interval.

1. The k-means scheme needs to estimate the means (centroids) of the clusters, while the Anti-Bayes scheme estimates the quantiles from the samples.
2. The performance of the Anti-Bayes clustering is remarkably accurate considering the fact that it operates from a completely counter-intuitive perspective, i.e., by comparing samples to elements of the clusters that are distant from the means.
3. Overall we see that both methods perform better when $N = 1000$ compared to $N = 10$. This is as expected since we are able to estimate centroids and quantiles with better precision. For $N = 10$, the k-means performs marginally better than Anti-Bayes for the symmetric distributions, but there are exceptions. For the one dimensional case, the Anti-Bayes performs the best for the Uniform and the Beta (symmetric) distributions.
4. When the distributions are asymmetric, it is known that the Anti-Bayes classification does not perform as well as the Bayes' bound [1] [9]. For these asymmetric distributions, the differences between the k-means and Anti-Bayes schemes is larger, but these difference are, really, quite small.
5. As mentioned in the introduction, in this paper, we have merely concentrated on the Anti-Bayesian clustering strategy for one and two-dimensional data. The extensions to multi-dimensional data are still open, but they would not be too complicated. Indeed, they would use the corresponding multi-dimensional *Anti-Naïve-Bayes* classification rules given in [1].

4 Conclusions

In this paper we have demonstrated how the "Anti-Bayesian" pattern classification framework formulated earlier in [8] [1] [9] can be used to build an efficient clustering algorithm competitive with the k-means clustering algorithm. The algorithm documents impressive clustering performance for a range of different distributions.

The k-means algorithm associates points to the clusters relying on the mean values (centroids) of the clusters which is known to be the optimal Bayesian bound. In contrast, the clustering algorithm presented in this paper associates points to clusters using quantiles distant from the cluster mean. Intuitively, one would thus expect a poor clustering performance and even serious convergence problems for the algorithm. But as demonstrated in [8] [1] [9], impressive classification precision can be achieved for the Anti-Bayesian approach which, thus, lays the foundation for the efficient clustering algorithm presented here.

The problems that are open are many. First of all, it would be interesting to demonstrate the power of such a strategy for multi-dimensional and real-life data. But more interestingly, we believe that the fundamental "Anti-Bayesian" paradigm can be applied to *any* clustering technique (apart from the k-means) that is inherently dependent on a Bayesian philosophy. We thus believe that this paper opens the doors to a host of unresolved problems.

References

1. Thomas, A., Oommen, B.J.: Order statistics-based parametric classification for multi-dimensional distributions. Pattern Recognition **46**(12) (2013) 3472–3482
2. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Dats. Prentice Hall, Englewood Cliffs, New Jersey, USA (1988)
3. Xu, R., Wunsch, II, D.: Survey of clustering algorithms. Trans. Neur. Netw. **16**(3) (2005) 645–678
4. Ankerst, M., Breunig, M.M., peter Kriegel, H., Sander, J.: Optics: Ordering points to identify the clustering structure, ACM Press (1999) 49–60
5. Ester, M., peter Kriegel, H., S, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise, AAAI Press (1996) 226–231
6. Murtagh, F., Contreras, P.: Methods of hierarchical clustering. CoRR **abs/1105.0121** (2011)
7. Sibson, R.: SLINK: An optimally efficient algorithm for the single-link cluster method. The Computer Journal **16**(1) (1973) 30–34
8. Thomas, A., Oommen, B.J.: The fundamental theory of optimal "anti-bayesian" parametric pattern classification using order statistics criteria. Pattern Recognition **46**(1) (2013) 376–388
9. Oommen, B.J., Thomas, A.: Anti–Bayesian parametric pattern classification using order statistics criteria for some members of the exponential family. Pattern Recognition **47**(1) (2014) 40–55
10. Hyndman, R.J., Fan, Y.: Sample quantiles in statistical packages. American Statistician **50** (1996) 361–365