




Parametric entropy based Cluster Centriod Initialization for k -means clustering of various Image datasets

Faheem Hussayn¹ and Shahid M Shah¹ 

Communication Control & Learning Lab (C^2L^2)
Department of Electronics & Communication Engineering,
National Institute of Technology Srinagar
{faheem_2023phacse002,shahidshah}@nitsri.ac.in
<https://sites.google.com/site/ccresearchlab/>

Abstract. One of the most employed yet simple algorithm for cluster analysis is the k -means algorithm. k -means has successfully witnessed its use in artificial intelligence, market segmentation, fraud detection, data mining, psychology, etc., only to name a few. The k -means algorithm, however, does not always yield the best quality results. Its performance heavily depends upon the number of clusters supplied and the proper initialization of the cluster centroids or seeds.

In this paper, we analyze the performance of k -means on image data by employing parametric entropies in an entropy-based centroid initialization method and propose the best-fitting entropy measures for general image datasets. We use several entropies like Taneja entropy, Kapur entropy, Aczel Daroczy entropy, and Sharma Mittal entropy. We observe that different entropies provide better results for different datasets than the conventional methods. We have applied our proposed algorithm on these datasets: Satellites, Toys, Fruits, Cars, Brain MRI, and COVID X-Ray.

Keywords: Entropy · Clustering · K -Means Algorithm · Unsupervised Learning

1 Introduction

A subset of artificial intelligence, machine learning, supplies machines with the ability to learn and make decisions without needing to be programmed explicitly. If learning is accomplished without supplying a machine with a labeled dataset, the machine needs to find the implicit data patterns without external support, this is known as unsupervised machine learning. Many real-life problems are actually modeled this way because, in real life, the pattern of data is difficult to know in advance [11].

Recent times have witnessed an avalanche of data. The use of data mining techniques has thus seen a tremendous increase and clustering has been one of the most used unsupervised techniques. Clustering, sometimes referred to as cluster

analysis, is an unsupervised machine learning technique wherein we tackle the problem of grouping or division of data points such that those data points that fall within the same group are more related to each other and less related to the data-points clustered into other groups [15]. It can also be simply defined as the collection or grouping of objects on the basis of similarity and dissimilarity between the objects. This technique finds its applications in Wireless networks, System diagnostics, Search engines, Fraud detection, Market Segmentation, Satellite imagery, pattern recognition, big data analytics, and so on.

The simplest unsupervised learning algorithm that is usually employed in solving clustering problems is the k -means algorithm. k -means, being one of the most famous algorithms employed for clustering, has also witnessed its use as part of other algorithms [19]. The k -means algorithm is iterative in nature and aims to assign every data-point of a data-set to one of the k clusters based on the features supplied. The k -means algorithm partitions ' n ' data points or observations into ' k ' groups or clusters. K -means algorithm has applications in various areas, like energy analytics [20], attack detection [2], [3], [21]. Very recently k -means algorithm has also been used in activity detection in smart grid-based systems [28]. However, the quality of the solution and convergence speed of the k -means algorithm largely depends on the number of clusters supplied and the position of the initial seed points or cluster centroids. The traditional k -means algorithm initializes cluster centroids randomly but this has obvious drawbacks. Various methods have been devised for alternate centroid initialization, which will be discussed in the next section. In this study, we will focus on an centroid initialization method based on the maximization of entropy measure.

2 Related Works

For the problem of proper centroid initialization, two groups of studies exist. The first group of studies has focused on improving the existing random initialization. When incorporating entropy, Steinbach et al. found that the "bisecting k -means" method generally outperformed the classical k -means and when not measuring entropy, it almost performed similarly [25]. However, they did not include any time-related metrics for comparison. k -means++ is another such modification of random centroid initialization. In this approach, Arthur and Vassilvitskii [4] initialized centroids from the data points at random while using the squared distance from the already initialized centroids to weigh potential centroids. The effect was that in contrast to the random initialization, this approach would ensure a maximal distance or "spread" of cluster centers or centroids.

The other group of studies has devised alternative methods to random centroid initialization for k -means and are consistent to the method we will discuss. A method for seed point selection that is recursive in nature was discussed by Duda and Hart [10]. A MaxMin algorithm was devised by Higgs et al. [14] and Snarey et al. [24] which was based on selecting a subset from original database to be used for initial centroids in order to create initial clusters. The bilinear

program was introduced by Bradley et al. [6] which determined initial points in such a way that the sum of distances of each data point should be minimized to the nearest centroid. Su & Dy [26] came up with a deterministic method for centroid initialization. The method is hierarchically divisive and is based on Principal Component Analysis (PCA). Cao et al developed an effective method to initialize clusters based on cohesion and coupling degree [7]. Bai et. al proposed an initialization method based on distance and density metrics [5]. Using the concept of Voronoi circles and their radii, Reddy et. al came up with an initialization technique [18]. Mahmud et al. came up with a method which is faster than traditional k-means. In this method, the selection of initial points is carried out using a weighted average score on the sorted data [17]. Gingles and Celebi developed a density-based approach [12] which was based on the hypothesis that centroids of clusters would naturally occur near the areas of high data-point density. Another density-based approach by Dalhatu et. al [9] has also been developed. One of the recent works for initialization based on entropy and for image segmentation (clustering of pixels) was carried out by Chowdhury et al. [8]. In their work, they employed the maximization of Shannon's Entropy to determine the optimal initial positions of the cluster centroids. This yielded lesser computation time and number of iterations concerning image datasets. The given method, however, was only tested on select images using Shannon's entropy. Despite all these methods, presently, there is no universally accepted method for centroid initialization of k -means algorithm which is the prime reason for the pursuit of this study.

3 Main Contribution

The traditional k -means algorithm initializes centroids randomly and as already discussed, the quality of clustering depends upon the location of the initial centroids. We have employed the entropy maximization algorithm devised by Chowdhury et al. [8] to initialize the centroids. However, in place of Shannon's Entropy, we test out different parametric entropy measures on contrasting image datasets with different parameters to yield the best-fitting entropy for centroid initialization. The entropy based initialization works on the entropy maximization principle. Exploiting the fact that for a multi spectral image, the intensity values for each color band of a particular pixel are mutually independent, we can easily calculate the probability of a pixel. Let N denote the total number of pixels in a given image and a , b , and c be the intensity values of the Red, Green, and Blue color bands respectively. Also, let n_a , n_b and n_c be the number of intensity values for a , b and c , respectively. Then, using the concept of independent random variables we can arrive at the equation:

$$\begin{aligned}
 &P(R = a, G = b, B = c) \\
 &P(R = a) * P(G = b) * P(B = c) \\
 &= \frac{n_a}{N} * \frac{n_b}{N} * \frac{n_c}{N}, \forall 0 \leq a, b, c \leq 255
 \end{aligned}$$

Using this probability, we can calculate the entropy measure for all the intensities in an image.

4 Methodology and Datasets

The algorithm for calculating the initial cluster centroids is given as Algorithm 1. For the entropy calculation step, we use the following entropy measures.

Shannon Entropy

Proposed by C.E Shannon [22], it is given as:

$$H_S(P) = - \sum_{i=1}^n p_i \log p_i \quad (1)$$

Kapur

Proposed by JN Kapur [16], this entropy is given by:

$$H_K(P) = \frac{1}{1-\alpha} \log \left(\frac{\sum_{i=1}^n p_i^{\alpha+\beta-1}}{\sum_{i=1}^n p_i^\beta} \right) \quad (2)$$

where $\alpha \neq 1, \alpha > 0, \beta \geq 1$

Aczél Daróczy

Proposed by J Aczél, Z Daróczy [1], it can be calculated using:

$$H_{AD}(P) = \frac{1}{\beta} \arctan \left(\frac{\sum_{i=1}^n p_i^\alpha \sin(\beta \log p_i)}{\sum_{i=1}^n p_i^\alpha \cos(\beta \log p_i)} \right) \quad (3)$$

where $\beta \neq 0, \alpha > 0$

Havrda and Charvát

Proposed by J Havrda and F Charvát [13], this entropy is given by:

$$H_{HC}(P) = \frac{1}{(2^{1-\alpha} - 1)} \left[\sum_{i=1}^n p_i^\alpha - 1 \right] \quad (4)$$

where $\alpha \neq 1, \alpha > 0$

Taneja

Proposed by I.J Taneja [27], this entropy is calculated using the formula:

$$H_T(P) = - \frac{2^{\alpha-1}}{\sin \beta} \sum_{i=1}^n p_i^\alpha \sin(\beta \log p_i) \quad (5)$$

where $\alpha \neq k\pi, k = 0, 1, 2, \dots, k > 0$

Sharma Mittal

Proposed by Sharma and Mittal [23], this entropy is given by:

$$H_{SM}(P) = \frac{1}{(2^{1-\alpha} - 1)} \left[\left(\sum_{i=1}^n p_i^\beta \right)^{\frac{\alpha-1}{\beta-1}} - 1 \right] \quad (6)$$

where $\alpha \neq 1, \alpha >, \beta \neq 1, \beta > 0$

In this research, instead of using a few images to test the initialization, we have used different real life data-sets containing similar images and averaged out the results to evaluate time-related metrics for each entropy measure. The data-sets were obtained from publicly available sources of Kaggle and some of the images were manually curated from Google. A summary of the datasets used is given in Table 1.

Since in this study we have focused on the initialization method of the cluster centroids, we need a way to determine the optimal number of clusters as it is also a factor on which the quality of k -means clustering depends as already discussed. We used the classic ‘elbow method’ to achieve this. The basic principle of the elbow method is that it plots the cost function (sum of square error values) for different values of k . Clearly, as the number of clusters increase, the SSE will reduce. A point will be reached where increasing the number of clusters will not have a drastic effect on the cost function. This we take as the optimal value of k . We can have the sum of squared distances of all data-points to the cluster as the cost function, where we call it ‘inertia’ or we can have the mean of squared distances of each data point to its nearest cluster, where we call it ‘dispersion’. Since there are multiple images in a particular dataset, using the fact that the distribution of pixel intensities will be similar, we employed the elbow method on any one of the images in a particular image dataset to determine the optimal value for the number of clusters k and conducted the analysis for all the images.

Table 1. Dataset Details

Dataset	Image Count	Attributes	Optimal k	Source
Satellite	25	3	3	Kaggle
Toys	50	3	4	Google
Brain MRI	30	2	3	Kaggle
X-Ray	25	2	2	Kaggle
Fruits	40	3	5	Google
Cars	50	3	3	Kaggle

An example of clustering using this approach is shown in Figure 1 and the corresponding comparison for the number of iterations utilized by k -means to converge is given in Figure 2. The th value for this image is set to 220.



Fig. 1. Original Image

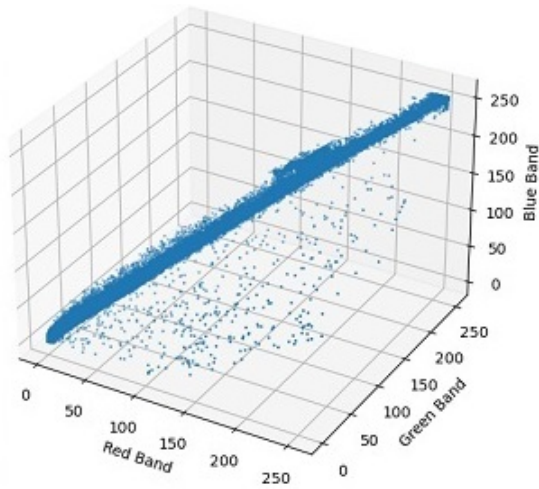


Fig. 2. Scatter Plot

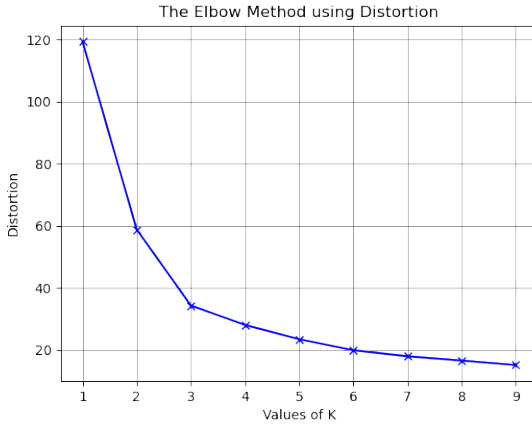


Fig. 3. Distortion based Elbow Method

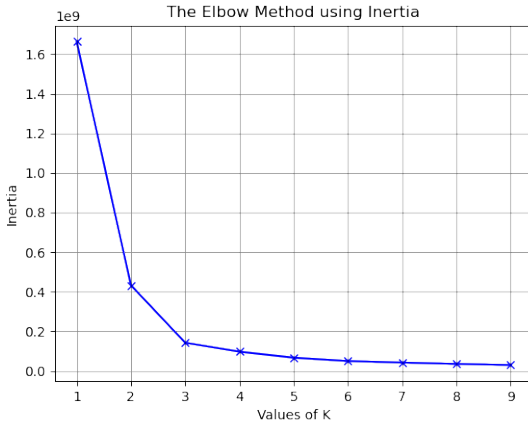


Fig. 4. Inertia based Elbow Method

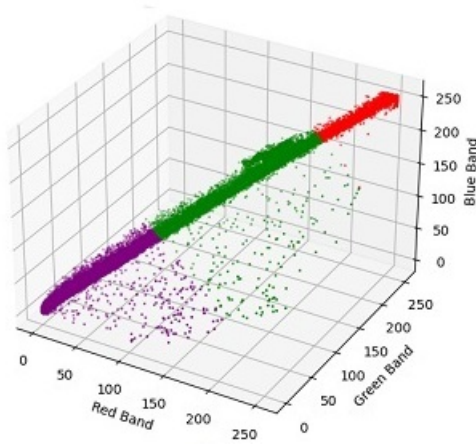


Fig. 5. Optimal Clustering (K=3)

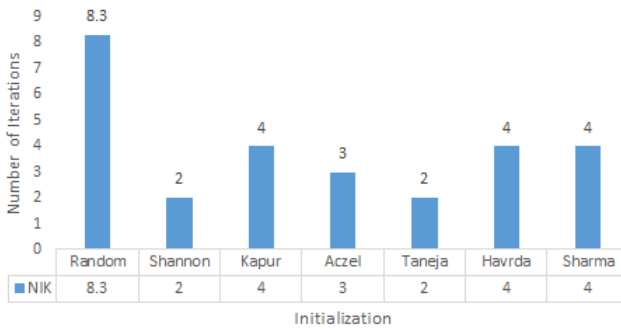


Fig. 6. Iteration Comparison of Initializations for Car Image

It is clear from the above graph that the most appropriate initialization for clustering this image using k -means algorithm is the Shannon and Taneja Entropy. The random initialization denotes the default random initialization of the k -means algorithm. The image shown in Figure 1: (a) is derived from the car dataset. We cannot possibly show the clustering for each image used in the experiment as we have employed image datasets and not single images.

The th or threshold value is a critical variable that essentially dictates the distance between the clusters. If it is not properly initialised, the distribution of centroids will not be appropriate. For example, if the value of th is too large, the computation cost will rise and so will be the time taken for convergence. On the contrary, a value that is too small can cause the centroids to be very near

Algorithm 1 Entropy Maximization Initialization

Input: Image (dataset) and number of clusters (K)

Output: Initial Centroid List

1. Input the number of cluster K and th (threshold for centroid spacing).
 2. Initialize the number of seeds $n_{cen} = 1$
 3. Calculate needed entropy for each pixel in the image.
 4. Sort the pixels in descending order of entropy values.
 5. Take the first pixel from the sorted list and include in centroid list.
 6. Take the next pixel from the list and calculate its euclidean distance with all the pixels in SE.
 7. if $min_{ED} > TH$, include this pixel in centroid list and perform $n_{cen} = n_{cen} + 1$.
Otherwise goto step 6.
 8. if $n_{cen} = K$, stop. Otherwise goto step 6.
-

causing the algorithm to converge prematurely. So, the choice of threshold value is important and should be decided by considering the "spread" of the data.

5 Results and Discussions

In our experiment, we used several images from contrasting image data-sets to analyse the performance of multiple entropy measures on the cluster initialization. The results were evaluated using the metrics: number of iterations of K-means (NIK), which essentially is the number centroid movements it takes to converge KMeans, Computation Time (C_T), which is the time to convergence, and Initialization Time (I_T). We club the Computation and Initialization time and call it Total Time. Since the k -means algorithm always converges and we are only modifying the initialization approach using different entropies, the sum of squared errors (SSE) metric may have similar results so our focus would be more on time-related metrics. The comparative description of the results is given in Table II. The initialization method using Shannon entropy is the original entropy maximization method devised by Chowdhury et al. [8]. [8]

From our study we concluded that there was no single entropy that was appropriate for the cluster initialization of every kind of image dataset. We get the insight that for certain datasets, certain entropy measures worked better. We summarize our results with the following insights:

- For the datasets with natural intensity levels and a higher dynamic range such as the images of cars, robots, toys, vegetables, fruits, etc. Taneja Entropy was the most appropriate.
- For the datasets with wide range of details like the satellite imagery, Shannon Entropy was the most appropriate for cluster initialization.
- For the datasets with similar saturation and less dynamic ranges, like the medical datasets of X-Ray and MRI Images, Kapur's Entropy was the most appropriate.

Table 2. Comparison of Initialization for Image Datasets

Dataset	Initialization	Avg. NIK	Total Time	SSE
Satellite	Random	4.76	2.148	3751.72
	Shannon	4.04	1.84	3751.5
Toys	Random	4.07	1.4335	1493.13
	Shannon	4.21	1.819	1492.3
	Taneja	3.11	1.079	1492.3
Fruits	Random	5.2	1.822	1564.1
	Shannon	6.3	1.938	1564.6
	Taneja	3.1	0.469	1563.6
Cars	Random	3.9	0.9242	1331.91
	Shannon	4.39	1.469	1332.02
	Taneja	2.01	0.401	1330.32
Brain MRI	Random	4.91	0.0203	1364.66
	Shannon	6.285	0.065	1364.5
	Kapur	4.285	0.036	1363.42
Covid X-Ray	Random	4.89	0.05	1379
	Shannon	3.69	0.079	1378.34
	Kapur	2.24	0.038	1377.43

6 Conclusion and Future Scope

In our study, we have extended the entropy based initialization method for image clustering using k -means by employing parametric entropy measures and demonstrated its effectiveness on contrasting image datasets. To generalize the results found in this study and establish facts based upon them would be irresponsible at this point due to the small size and lesser number of the datasets used. However, our findings do point out that further research should be pursued, and that the further exploration of the parametric entropies discussed in the previous sections would prove beneficial in other entropy based avenues of computing. In future, we would try to hypothesize or assert why certain entropy measures work better with certain kinds of data and include more entropy measures for testing. We would also study the effect of using the generalized entropy measures in place of Shannon's entropy with other research problems, such as cluster validation, metric evaluation, and so on. . 3rd-level heading).

References

1. János Aczél and Zoltan Daróczy. Über verallgemeinerte quasilineare mittelwerte, die mit gewichtsfunktionen gebildet sind. *Publ. Math. Debrecen*, 10:171–190, 1963.
2. Mir Shahnawaz Ahmad and Shahid Mehraj Shah. Mitigating malicious insider attacks in the internet of things using supervised machine learning techniques. *Scalable Computing: Practice and Experience*, 22(1):13–28, 2021.

3. Mir Shah Nawaz Ahmad and Shahid Mehraj Shah. Supervised machine learning approaches for attack detection in the IoT network. In *Internet of Things and Its Applications: Select Proceedings of ICIA 2020*, pages 247–260. Springer, 2022.
4. David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.
5. Liang Bai, Jiye Liang, Chuangyin Dang, and Fuyuan Cao. A cluster centers initialization method for clustering categorical data. *Expert Systems with Applications*, 39(9):8022–8029, 2012.
6. Paul S Bradley and Usama M Fayyad. Refining initial points for k-means clustering. In *ICML*, volume 98, pages 91–99. Citeseer, 1998.
7. Fuyuan Cao, Jiye Liang, and Guang Jiang. An initialization method for the k-means algorithm using neighborhood model. *Computers & Mathematics with Applications*, 58(3):474–483, 2009.
8. Kuntal Chowdhury, Debasis Chaudhuri, and Arup Kumar Pal. Seed point selection algorithm in clustering of image data. In *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, pages 119–126. Springer, 2018.
9. Kabiru Dalhatu and Alex Tie Hiang Sim. Density base k-mean’s cluster centroid initialization algorithm. *International Journal of Computer Applications*, 137(11), 2016.
10. Richard O Duda, Peter E Hart, et al. *Pattern classification and scene analysis*, volume 3. Wiley New York, 1973.
11. Martin Ford. *Architects of Intelligence: The truth about AI from the people building it*. Packt Publishing Ltd, 2018.
12. Caroline Gingles and M Emre Celebi. Histogram-based method for effective initialization of the k-means clustering algorithm. In *The Twenty-Seventh International Flairs Conference*, 2014.
13. Jan Havrda and František Charvát. Quantification method of classification processes. concept of structural α -entropy. *Kybernetika*, 3(1):30–35, 1967.
14. Richard E Higgs, Kerry G Bemis, Ian A Watson, and James H Wikel. Experimental designs for selecting molecules from large chemical databases. *Journal of chemical information and computer sciences*, 37(5):861–870, 1997.
15. Anil K Jain and Richard C Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.
16. JN Kapur. Generalized entropy of order α and type β . In *The Math. Seminar*, volume 4, pages 78–82, 1967.
17. Md Sohrab Mahmud, Md Mostafizer Rahman, and Md Nasim Akhtar. Improvement of k-means clustering algorithm with better initial centroids based on weighted average. In *2012 7th International Conference on Electrical and Computer Engineering*, pages 647–650. IEEE, 2012.
18. Damodar Reddy, Prasanta K Jana, and IEEE Senior Member. Initialization for k-means clustering using voronoi diagram. *Procedia Technology*, 4:395–400, 2012.
19. Amit Saxena, Mukesh Prasad, Akshansh Gupta, Neha Bharill, Om Prakash Patel, Aruna Tiwari, Meng Joo Er, Weiping Ding, and Chin-Teng Lin. A review of clustering techniques and developments. *Neurocomputing*, 267:664–681, 2017.
20. Shahid Mehraj Shah. Modelling energy consumption of domestic households via supervised and unsupervised learning: A case study. In *Machine Learning and Metaheuristic Algorithms, and Applications: Second Symposium, SoMMA 2020, Chennai, India, October 14–17, 2020, Revised Selected Papers 2*, pages 157–171. Springer, 2021.

21. Mir Shahnawaz Ahmad and Shahid Mehraj Shah. Unsupervised ensemble based deep learning approach for attack detection in iot network. *Concurrency and Computation: Practice and Experience*, 34(27):e7338, 2022.
22. Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.
23. Bhudev D Sharma and Dharam P Mittal. New non-additive measures of entropy for discrete probability distributions. *J. Math. Sci*, 10:28–40, 1975.
24. Michael Snarey, Nicholas K Terrett, Peter Willett, and David J Wilton. Comparison of algorithms for dissimilarity-based compound selection. *Journal of Molecular Graphics and Modelling*, 15(6):372–385, 1997.
25. Michael Steinbach, George Karypis, and Vipin Kumar. A comparison of document clustering techniques. Technical report, University of Minnesota Twin Cities, 2000.
26. Ting Su and Jennifer Dy. A deterministic method for initializing k-means clustering. In *16th IEEE international conference on tools with artificial intelligence*, pages 784–786. IEEE, 2004.
27. Inder Jeet Taneja. On generalized information measures and their applications. In *Advances in Electronics and Electron Physics*, volume 76, pages 327–413. Elsevier, 1989.
28. Nida Ul Islam and Shahid Mehraj Shah. A binary weight-based energy disaggregation framework for residential electricity consumption. In *2022 IEEE 10th Power India International Conference (PIICON)*, pages 1–6, 2022.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

