International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 3, March 2016

An Efficient Density Based Image Clustering Method with P-Trees

[1] D.V. Lalita Parameswari [2] Dr. M. Seetha, [3] Dr. K.V.N. Sunitha,
[1] Sr. Asst. Professor, Dept. of CSE, GNITS, Hyderabad-8, INDIA, [2] Professor, Dept. of CSE, GNITS, Hyderabad-8, INDIA, [3] Principal, BVRIT for women, Hyderabad., India,
[1] lalla_mks@yahoo.com [2] smaddala2000@yahoo.com [3] k.v.n.sunitha@gmail.com

***Abstract*:-- Image clustering analysis plays an important role in data mining applications which groups set of pixels. Traditional approaches of clustering are based on deviation of the Euclidean distance which leads to the clusters of spherical shapes and input parameters are to be specified which are hard to determine. To overcome this, density based clustering techniques like DBSCAN,OPTICS are used to cluster satellite images. Thus, only low-dimensional images can be processed with limited computer memory and computing speed. This paper emphasizes on the implementation of P-Tree which requires very less memory and is very efficient for lossless image representation and compression. Thus a new Peano count tree (P-Tree) method is proposed on DBSCAN and OPTICS clustering techniques on satellite images. The DBSCAN and OPTICS clustering techniques are implemented on satellite images by applying P-tree structure. These techniques are compared and analyzed with accuracy and kappa statistic performance measures. It is ascertained that the performance of DBSCAN and OPTICS clustering techniques with P-tree is efficient than the clustering without P-tree. Further the accuracy and kappa statistic are better for OPTICS method than DBSCAN.**

***Keywords*: Image clustering, remote sensing, density based methods, clustering accuracy Peano count tree.**

I. INTRODUCTION

Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. It can be used in the identification of areas of similar land usage in an earth observation or in merging regions with similar weather patterns etc. Spatial clustering algorithms have been proposed in many applications such as pattern recognition, data analysis, and image processing and so forth. Image segmentation is the process of partitioning a digital image into multiple segments. The segmentation can be categorized into edge detection, region growing, clustering [6, 13 and 14]. Edge detection is a well-developed field on its own within image processing. Region growing methods start usually from a pixel level and, using a homogeneity criterion, merge neighboring objects in a sequential order until the criterion exceeds a threshold defined by the user. The boundary-based methods will fail if the image is noisy. Both region boundary and edge detection based method often fail to produce accurate segmentation results. To overcome this density-based clustering methods like DBSCAN (density based spatial clustering of application along with noise) and OPTICS (Ordering points to identify the clustering structure) have been developed[2,3 and 7]. DBSCAN grows regions with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise [5,8 and 9]. It defines a cluster as a maximal set of density-connected points. DBSCAN develops clusters according to a density-based connectivity analysis. Although DBSCAN can cluster objects given input parameters such as ε and *MinPts*, it still leaves the user with the responsibility of selecting parameter values that will lead to the discovery of acceptable clusters. Actually, this is a problem associated with many clustering algorithms. To help overcome this difficulty, a cluster analysis method called OPTICS was proposed. OPTICS computes an augmented cluster ordering for automatic and interactive cluster analysis. This ordering represents the density-based clustering structure of the data. It contains information that is equivalent to density-based clustering obtained from a wide range of parameter settings[2,7]. Peano trees are a new contraption poised to alter the way spatial data is recorded, utilized, evaluated, and inspected. This novel technology efficiently plays a crucial step in making the technology mainstream. It has subdued many image compression methods. Primarily the individual evaluation of DBSCAN and OPTICS was done. Then a new method has been proposed for clustering the image by combining P-tree method with DBSCAN and OPTICS. The accuracy and kappa statistic of DBSCAN and OPTICS without and with P-Tree method are compared and analyzed. Hence this paper is organized as follows. Section 2 introduces the density based clustering techniques; Section 3 presents the P-Tree representation and construction. An efficient method of DBSCAN and OPTICS image clustering with P-tree is proposed in Section 4. The results and discussions are elucidated in Section 5 and conclusions are depicted in Section6.

II. SPATIAL CLUSTERING TECHNIQUES

A. DBSCAN ALGORITHM

The DBSCAN algorithm can discover clusters in huge spatial data sets by gazing the local density of database elements, with only one input parameter. The DBSCAN can furthermore decide which information has to be classified as noise or else outlier. DBSCAN can find clusters of an arbitrary shape [3,4]. If a point is found to be in a dense part of a cluster, then the ε-neighborhood of that point is also part of the same cluster. Thus, all points that are found in the ε-neighborhood are added. This method continues until the entire density-connected cluster is entirely found. Subsequently, a new unvisited point is retrieved and processed, leading to the finding of further cluster or noise.

B. OPTICS ALGORITHM

The OPTICS algorithm is similar to DBSCAN, but one of the major drawbacks in DBSCAN is the difficulty of identifying meaningful clusters in data of varying density [2, 7 and 13]. The objects of the database are ordered such that objects which are spatially closest become neighbors in their ordering. Furthermore, a special distance measure is calculated for each object that represents the density that needs to be accepted for a cluster in order to have both objects belong to the same cluster.

OPTICS requires two parameters: 'ε' (Epsilon) and MinPts (minimum points). It is sufficient to extract all density-based clustering with respect to any distance ε' which is smaller than the generating distance 'ε' from this order [5, 7] . This pixel's ε-neighborhood is retrieved, and if it contains sufficiently many pixels within or less than ε -neighborhood, a cluster is started. Otherwise, the pixel is labeled as noise. The reachability-distance of the current object is verified whether it is larger than the clustering-distance ε'. In this case, the pixel is not density-reachable with respect to ε' and MinPnts from any of the pixels which are located before the current pixel in the cluster-ordering. This is obvious, because if pixel had been density-reachable with respect to ε' and MinPnts from a preceding object in the order, it would have been assigned a reachability-distance of at most ε'. Therefore, if the reachability-distance is larger than ε', the core-distance of the pixel is considered to start a new cluster if pixel is a core object with respect to ε' and MinPnts; otherwise, pixel is assigned to NOISE.

III. PEANO COUNT TREES

The Peano Count Tree (P-tree) is a quadrant-based lossless tree representation of the original spatial data[1]. Unlike other methods, P-tree representation requires very less memory and is very efficient for lossless image representation and compression. Using P-tree structure, all the count information can be calculated quickly. This facilitates efficient ways for data mining. P-trees have been used in a variety of application including micro array data, stock market data, genetic information, and the original use in crop analysis. Unlike other methods, P-tree representation requires very less memory and is very efficient for lossless image representation and compression.

A. P-TREE IMAGE REPRESENTATION

Most of the spatial data collected, comes in a format called BSQ for Band Sequential [10,15] (or can be easily converted to BSQ).In BSQ format, values corresponding to each band is stored in a separate file. The Raster ordering of the data values [10,15] is followed internally in each file with respect to the spatial area represented in the dataset. For constructing P-Trees each BSQ band will be divided into several files, one for each bit position, called the 'bit Sequential' or bSQ. A simple transformation can be applied to convert image files to band sequential (BSQ) and then to bit sequential (bSQ) format. Each bSQ bit file is organized into Bij (the file constituting the jth bits of ith band), into a tree structure, called a Peano Count Tree (P-Tree) [11,12]. A P-Tree is a quadrant-based, lossless tree representation.

B. CONSTRUCTION OF P-TREES

In an Image each intensity value ranges from 0 to 255, which can be represented as a byte, each bit in one band can be represented as a separate file, called a bSQ file. Each bSQ file can be reorganized into a quadrant-based tree (P-tree). The Figures 1 & 2 shows 8 X 8 sub image and its bSQ representation.

241	146	227	213	254	229	120	51
177	211	234	105	98	80	110	116
152	136	106	107	136	85	98	102
137	112	87	82	121	88	106	100
135	108	89	87	109	83	104	115
94	101	101	91	111	117	128	138
98	86	113	120	109	106	114	138
94	68	110	145	140	108	94	137

Fig 1. 8x8 image

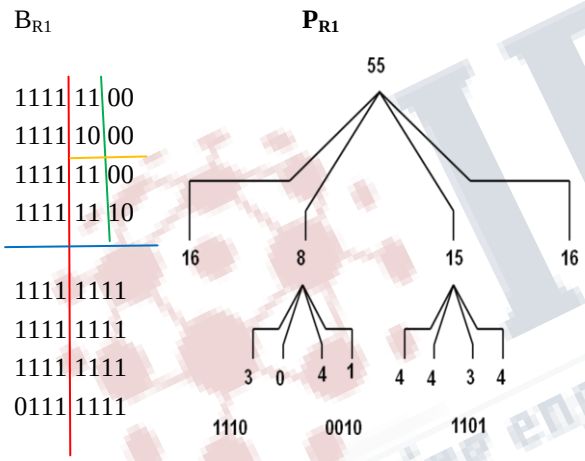
Red band data, B_R, in a 64 pixel space (8 rows by 8 columns)

11110001 10010010 11100011 11010101
10000000 11100101 01111000 00110011
10110001 11010011 11101010 11000001
11001000 01011010 0011110 01010101
11010001 10010010 11100011 11010101
10000000 11100101 01111000 00110011
10010001 11010011 11101010 11000001
11100100 10101101 10011110 01010101
11110001 10010010 11100011 11010101
10001110 11100101 11111000 10110011

10110001 11010011 11101010 11000001
11100110 10101101 10011110 11011101
11010001 10010010 11100011 10011101
10001010 11100101 11111000 10110111
00010001 11010011 11101010 10101001
11101100 10101101 10011110 11010101

Fig 2. bSQ representation

The root node of a Peano count Tree constitutes the count of 1's in the entire file (bit-band). The file is now divided into four equal quadrants and the next level (second level) of the P-Tree after the root constitutes the counts of 1's of the four quadrants into which the file is divided, in raster order. At the third level (next level), each quadrant is further partitioned into four sub-quadrants and their counts of 1's constitute the children of the quadrant node in raster order. This process of construction is continued recursively down along each tree path until the newly formed sub-quadrant is 'pure' i.e. either entirely 1-bits or entirely 0-bits, which need not be the leaf level. For example, the P-Tree for an 8-row-8-column bit-band is shown in the figure3.



In this example, 55 is the count of 1's in the entire sub image (called root count), the numbers at the next level, 16, 8, 15 and 16, are the 1-bit counts for the four major quadrants. Since the first and last quadrant is made up of entirely 1-bits (called pure-1 quadrants), we do not need sub-trees for them. Similarly, quadrants made up of entirely 0-bits are called pure-0 quadrant. This pattern is continued recursively. Recursive raster ordering is called Peano or Z-ordering in the literature. The process terminates at the leaf level (level-0) where each quadrant is a 1- row-1-column quadrant. If all sub-trees has to be expanded, including those pure quadrants, then the leaf sequence is just the Peano space-filling curve for the original raster image. For each band (assuming 8-bit data values), 8 basic P-trees can be generated, one for each bit positions. For band, B_i, the basic P-trees can be labeled as , P_{i1}, P_{i2}, …, P_{i8}, thus, P_{ij} is a lossless representation of the j^{th} bits of the values from the i^{th} band. However, P_{ij} provides more information and are structured to facilitate data mining processes. Finally, the basic P-trees can be generated quickly and it is only a one-time cost. So this structure can be viewed as a data mining ready and lossless format for storing spatial data.

IV. P-TREE WITH DBSCAN AND OPTICS

Peano Count Trees represent spatial data bit-by-bit in a recursive quadrant-by-quadrant arrangement. A spatial image can be viewed as a 2-dimensional array of pixels. Each image data can be divided into quadrants and records the count of 1-bits for each quadrant, thus forming a quadrant count tree. P-Tree requires less memory and is very efficient for lossless image representation and compression. To achieve better accuracy without any loss of information, a new P-Tree technique is proposed on DBSCAN and OPTICS clustering algorithms known as DBSCAN with P-Tree and OPTICS with P-Tree. Using P-tree structure, all the count information can be calculated quickly. Moreover the memory required for storage of the image is also reduced with bSQ format image representation as the number of bits required to store the image are less using P-tree representation.

V. RESULTS AND DISCUSSIONS

The spatial clustering techniques like DBSCAN and OPTICS are employed on 10 various images taken from IRS 1C LISS III satellite. To achieve better accuracy a novel P-Tree technique is implemented on DBSCAN and OPTICS clustering algorithms. These techniques are compared and analyzed by the performance measures like accuracy and kappa statistic. Accuracy assessment can be performed by comparing two sources of information of clustered data and reference test data. The relationship of these two sets is summarized in an error matrix where columns represent the reference data while rows represent the clustered data. An error matrix is a square array of numbers laid out in rows and columns that expresses the number of sample units assigns to a particular category relative to the actual category as verified in the field.

The table 1 and table 2 shows the accuracy and the kappa statistics for the spatial clustering techniques of DBSCAN, OPTICS and the proposed approach P-Tree on both DBSCAN and OPTICS for ten different images respectively. The OPTICS technique provides good improvement in accuracy than the DBSCAN due to an augmented cluster ordering computation. The accuracy of spatial clustering methods DBSCAN, OPTICS, DBSCAN with P-Tree OPTICS with P-Tree are as shown in Figure1 and Figure2. It is observed that the accuracy is improved

with P-tree method than the spatial clustering methods DBSCAN and OPTICS. The results show that for image6 DBSCAN with P-Tree gives more accuracy than OPTICS with P-Tree due to the abnormal distribution of intensity values on the image. The kappa statistic is also high for all the 10 images for the spatial clustering techniques of DBSCAN and OPTICS with P-Tree. It is also practical from Figure1 and Figure2 that the accuracy and kappa statistic are better for OPTICS with P-Tree method than DBSCAN with P-Tree method. Hence it is ascertained that the OPTICS with P-Tree method outperforms the DBSCAN with P-Tree method when evaluated with the performance measures like accuracy and kappa statistics.

Table 1: Performance evaluation of different clustering algorithms based on accuracy

IMAGES	DBSCAN	DBSCAN WITH P-TREE	OPTICS	OPTICS WITH P-TREES
Image1	42.5	72.5	49.78	75.2
Image2	51.23	68.58	52.55	79.40
Image3	63.64	73.00	63.72	78.33
Image4	69.88	63.03	71.00	80.42
Image5	58.00	60.89	68.70	75.18
Image6	61.43	70.23	62.21	68.93
Image7	66.36	68.76	69.01	72.234
Image8	62.56	66.42	63.70	70.23
Image9	48.55	69.08	55.29	72.351
Image10	53.80	59.85	55.48	65.44

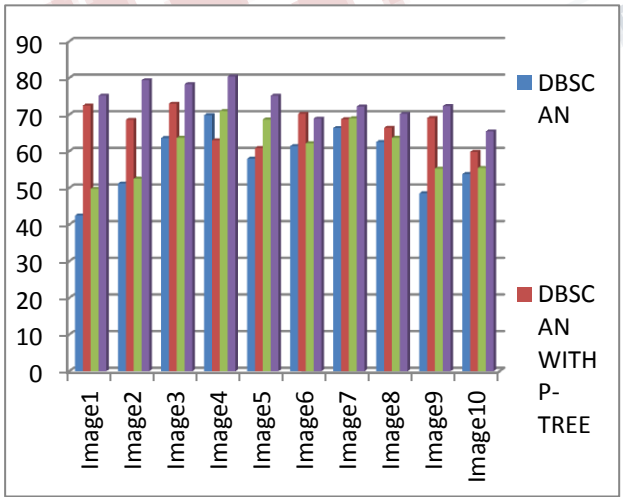


Figure 1: Performance evaluation of different clustering algorithms based on accuracy

Table 2: Performance evaluation of different clustering algorithms based on Kappastatisics

IMAGES	DBSCAN	DBSCAN WITH P-TREE	OPTICS	OPTICS WITH P-TREES
Image1	0.66	0.79	0.80	0.89
Image2	0.64	0.79	0.79	0.82
Image3	0.66	0.73	0.71	0.79
Image4	0.61	0.79	0.72	0.81
Image5	0.63	0.78	0.70	0.82
Image6	0.68	0.70	0.75	0.80
Image7	0.62	0.78	0.71	0.86
Image8	0.64	0.72	0.80	0.87
Image9	0.71	0.72	0.79	0.81
Image10	0.67	0.76	0.77	0.82

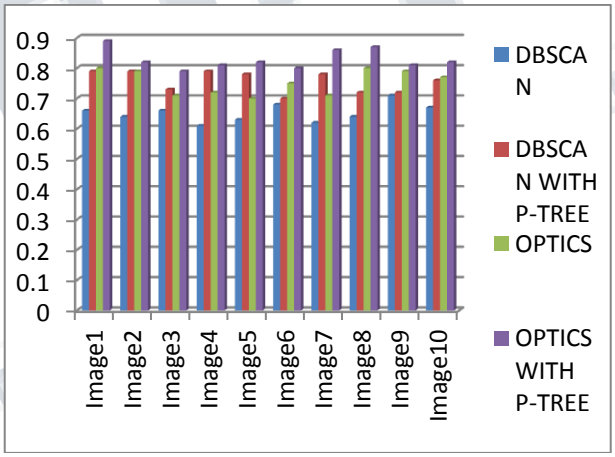


Figure 2: Performance evaluation of different clustering algorithms based on Kappastatisics

It has been ascertained that the classification accuracy improves remarkably upon the usage of Fuzzy based classification methods than the other classification methods.

VI. CONCLUSIONS

In the density based clustering, the OPTICS clustering algorithm gives good result compared to DBSCAN clustering algorithm for the same input parameters and it was done for different input parameters. OPTICS is a generalization of DBSCAN to multiple ranges, effectively replacing the ε parameter with a maximum search radius that mostly affects the performance.

REFERENCES

1) Amlendu Roy, William Perrizo, "Peano Count Tree Technology", "Deriving High Confidence Rules from Spatial Data using Peano Count Trees", LNCS 2118, July 2001.

2) D.V. Lalita Parameswari, Dr. M. Seetha, Dr. K.V.N. Sunitha **"** An improved grid based density methods for image clustering " published in International Journal of Computer Engineering and Applications (IJCEA) ISSN: 2321-3469, Volume IX, Special Issue

3) Ester M., Kriegel H.- P., Sander J., Xu X., A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD-96), Portland, OR, 1996, pp. 226-231, 1996.

4) Introduction to data mining, Pang-ning Tan, Vipin kumar, Michael Steinbach. Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2001.

5) J.O. Lapeyre and R. Strandh, "An efficient union-find algorithm for ex-tracting the connected components of a large-sized mage," Lab. Bordelais de Recherche en Inf., Bordeaux, France, Jan. 2004.

6) Tech. Rep. Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander."OPTICS: Ordering Points to Identify the Clustering Structure". Proc. ACM sigmod'99 Int. Conf. on Management of Data, Philadelphia PA, pp.49-60,1999

7) Martin Ester Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Published in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96) pp.226-231, 1996.

8) Matheus C.J.; Chan P.K.; and Piatetsky-Shapiro G. 1993. Systems for Knowledge Discovery in Databases, IEEE Transactions on Knowledge and Data Engineering 5(6): 903-913.

9) Maleq Khan, Qin Ding, and William Perrizo, k-Nearest Neighbour Classification on Spatial Data Streams, 2002.

10) Mohammad Hossain, Amal Shehan Perera and William Perrizo, Bayesian Classification on Spatial Data Streams Using P-Trees, Computer Science Department, North Dakota State University, Fargo, ND 58105, USA, 2002.

11) Qin Ding, Maleq Khan, Amalendu Roy and William Perrizo, "The P-tree Algebra", Computer Science Department, North Dakota State University Fargo, ND 58105-5164, USA, 2002.

12) Raymond T. Ng and Jiawei Han, CLARANS: A Method for Clustering Objects for Spatial Data Mining, IEEE transactions on Knowledge and Data engineering, Vol. 14, No. 5, September 2002.

13) Y. Tarabalka, J. A. Benediktsson, J. Chanussot, "Spectral-Spatial Classification of Hyperspectral Imagery Based on Partitional Clustering Techniques," IEEE Trans. Geosci. Remote Sensing, vol. 47, no. 8, pp. 2973-2987, 2009.

14) William Perrizo, Qin Ding, Qiang Ding, Amlendu Roy, "Deriving High Confidence Rules from Spatial Data using Peano Count Trees", Springer-Verlag, LNCS 2118, July 2001