International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869 (O) 2454-4698 (P), Volume-6, Issue-1, September 2016

Performance assessment on spatial data by the development of Classification Techniques using P-Tree subspace method

D.V. Lalita Parameswari, Dr. M. Seetha, Dr. K.V.N. Sunitha

***Abstract*— Spatial data which has been collected through satellite imagery, sensor devices etc, is growing too fast to analyze. Spatial Image classification involves image interpretation, in which the image pixels are classified into various predefined land use/cover classes. It is difficult to classify such classes correctly using traditional classifiers and hence classification accuracy is low. To discover the hidden knowledge in spatial image data various classification techniques like naïve bayesian classification and decision trees applied on spatial data. Since the spatial data is huge to store and maintain, a new data compression technique i.e. Peano count tree (P-Tree) subspace method is proposed on spatial data. This paper emphasizes on the classification of LISS-III images using P-Tree subspace method, naïve bayesian and decision tree classification methods to improve the classification accuracy. The performance parameters like overall accuracy ,Kappa statistic and execution time were analyzed to identify the best classification method. It is ascertained that the classification accuracy has been improved by P-Tree subspace methods and decision tree classification with P-Tree subspace method is superior to other classification methods.**

***Index Terms*— Spatial data, spatial image classification, decision tree, naïve bayesian classification, P-Tree subspace method, classification accuracy.**

I. INTRODUCTION

Classification is used to extract significant data models, plays an important role in data mining techniques. Spatial data can be utilized in a number of applications like urban planning , evaluation of environmental damage, monitoring of land use, crop yield appraisal , growth regulation, radiation monitoring and soil assessment [8]. The major application of spatial data is to identify meaningful features or classes of land cover types in a scene [9]. Therefore, the principal product is a thematic map with themes like land use, vegetation types [6].

Spatial data classification is the process of assigning pixels of an spatial image to classes. These classes are formed by grouping identical pixels found in spatial data that match the categories of user interest by comparing pixels to one another and to those of known identity [7].

Some of the common classification methods used in data mining include decision tree classifiers, bayesian classifiers, k-nearest-neighbor classifiers, genetic algorithms, rough sets, and fuzzy logic techniques. Among these classification algorithms decision tree, naive bayesian classification techniques are commonly used because they are easy to understand and cheap to implement. ID3, C4.5 [1], [2] and CART [3] are some of the best known classifiers that use decision trees. Other decision tree classifiers include Interval Classifier [4] and SPRINT [4, 5 and 10] which concentrate on making it possible to mine databases that do not fit in main memory by only requiring sequential scans of the data. Decision tree classifier [12, 13 and 14] is a hierarchical structure where at each level a test is applied to one or more attribute values that may have one of two outcomes. The outcome may be a leaf, which allocates a class, or a decision node, which specifies a further test on the attribute values and forms a branch or sub-tree of the tree. Naive bayesian [17, 18] is a statistical classifier based on bayes theorem, which majorly depends on class conditional probabilities. The Peano count Tree (P-Tree) subspace method is a new invention to change the way spatial data is recorded, used, evaluated, and searched. It is basically a quadrant based and lossless image compression technique. It helps in building the classifier more efficiently and at a faster rate. It is ascertained that the proposed P-Tree subspace method outperforms the other classification techniques when evaluated with overall accuracy, kappa statistic and execution time.

A. Decision tree Classification

A decision tree classifier compactly stores the data in a simple form. It can perform automatic feature selection which can be used to efficiently classify the new data. Decision tree classifier exploits a hierarchical structure in which a test is applied to each level over one or more attribute values that may results one of two outcomes[10,11 and 16]. It may be a leaf or a decision node. The leaf denotes the class whereas decision node represents further test on attribute values to form a branch or sub tree of the tree.The leaf represents final classification. The rules are extracted from the decision tree classification process starting from the root node and ending at one of the leaf, to determine the label of the classified object. A decision has to be taken at every non-terminal node to determine the path for the next node. This process has to be repeated recursively until no remaining attributes may be further partitioned. The decision tree not only memorizes the training set but also generalizes the unseen data.

The algorithm for inducing a decision tree from the training set is as follows:

- Initially, the decision tree is a single node representing the entire training set.

D.V. Lalita Parameswari, Sr. Asst. Professor, Dept. of CSE, GNITS, Hyderabad-8, India
Dr. M. Seetha, Head and Professor, Dept. of CSE, GNITS, Hyderabad-8, India.
Dr. K.V.N. Sunitha, Principal, BVRITH for women, Hyderabad-8, India

- If all samples belong to the same class, this node becomes the leaf and is labelled with that class label.
- Otherwise, an entropy-based measure is used as a heuristic for selecting the attribute which best separates the samples into individual classes, the "decision" attribute. A branch is created for each value of the test attribute and samples are partitioned accordingly.
- The algorithm recursively advances to form the decision tree for the set at each partition. Once an attribute has been used, it is not considered in descendent nodes in the set.
- The algorithm stops if either all samples for a given node belong to the same class or when there are no attributes remaining in the set.

The attribute selected at each decision tree level is the one with the best measure for split which includes entropy, information gain and Gini index [3].

B. Naive Bayesian Classification

Naïve bayesian classification is a statistical classifier based on Bayes theorem. let X be a data sample whose class label is unknown. Let H be a hypothesis (ie, X belongs to class, C). P(H|X) is the posterior probability of H given X. P(H) is the prior probability [15,18] of H then P(H|X) = P(X|H)P(H) / P(X) where P(X|H) is the posterior probability of X given H and P(X) is the prior probability of X. Naïve bayesian classification uses this theorem in the following way. Each data sample is represented by a feature vector, $X=(X_1..,X_n)$ depicting the measurements made on the sample from $A_1,..A_n$. Given classes, $C_1,...C_m$, the naïve bayesian Classifier will predict the class label, C_j , that an unknown data sample, X (with no class label), belongs to the one having the highest posterior probability, conditioned on X $P(C_j|X) > P(C_i|X)$, where i is not equals to j. P(X) is constant for all classes so $P(X|C_j)P(C_j)$ is maximized. In naive bayesian the naive assumption 'class conditional independence of values' is made to reduce the computational complexity of calculating all $P(X|C_j)$'s. It assumes that the value of an attribute is independent of that of all others. Thus, $P(X|Ci) = P(X_k|C_i)*\ldots*P(X_n|C_i)$ .For categorical attributes, $P(X_k|C_i) = S_iX_k/S_i$ where S_i is the number of samples in class C_i and S_iX_k is the number of training samples of class C_i, having A_k the value X_k [19].

II. PEANO COUNT TREE SUBSPACE METHOD

Peano Count Tree subspace method (P-Tree subspace method) represents spatial data bit-by-bit in a recursive quadrant-by-quadrant arrangement. Each new component in a spatial data stream is converted in to P-Trees. A spatial image can be viewed as a 2-dimensional array of pixels. The idea of P-Tree subspace method is to recursively divide the entire spatial data, such as remotely sensed imagery data, into quadrants and records the count of 1-bits for each quadrant, thus forming a quadrant count tree. Using P-Tree subspace structure, all the count information can be calculated quickly. Associated with each pixel are various descriptive attributes called "bands".

Since each intensity value ranges from 0 to 255, which can be represented as a byte, each bit of the band can be split into a separate file, called a bSQ file. Each bSQ file can be reorganized into a quadrant-based tree (P-tree)[19]. For each band (assuming 8-bit data values), 8 basic P-trees can be obtained, one for each bit positions. The basic P-trees for band B_i are $P_{i,1}$, $P_{i,2}$, …, $P_{i,8}$, thus, $P_{i,j}$ 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.

Consider the input spatial image of size 8X8 with pixel values

241	146	227	213	128	229	120	51
177	211	234	193	200	90	110	85
209	146	227	213	128	229	120	51
145	211	234	193	228	173	158	85
241	146	227	213	142	229	248	179
177	211	234	193	230	173	158	221
209	146	227	157	138	229	248	183
17	211	234	169	236	173	158	213

Figure 1. 8 X 8 spatial image

The above sub image can be represented into binary format as follows.

11110001	10010010	11100011	11010101	10000000	11100101	01111000	00110011
10110001	11010011	11101010	11000001	11001000	01011010	01101110	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

Figure 2. Red band spatial data in a 64 pixel space (8 rows by 8 columns)

Initially all the first bit information of first column of each pixel is represented as first column in the P-tree subspace method. The second bit information from the second column, the process continues for all the eight columns. The resultant data represents 8X8 binary data for the Red band. The process is repeated to extract the pixel information for all the bands.

III. RESULTS AND DISCUSSIONS

The classification techniques like naïve bayesian, decision tree and naïve bayesian and decision tree with P-Tree subspace method were implemented on various input images. The study have been carried out by using sample images obtained from IRS 1D LISS III sensor. Total five classes were taken to analyze the classification process.

The classification process is analyzed by the accuracy assessment of the methods of naïve bayesian, decision tree, naïve bayesian and decision tree with P-Tree subspace method. The proposed P-Tree subspace method achieved better performance when compared with other classification techniques. The performance measures like accuracy, kappa statistic and execution time were used for the classification process.

A. *Overall Accuracy*

Once a classification has been sampled a contingency table (also referred to as an error matrix or confusion matrix) is developed. This table is used to properly analyze the validity of each class as well as the classification as a whole. In this way efficacy of the classification can be evaluated in more detail. Accuracy assessment can be performed by comparing two sources of information of classified 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 classified 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.

Overall accuracy can be calculated using the following formula:

$$\text{Overall Accuracy} = \frac{\text{Total number of correct classifications}}{\text{Total number of classifications}}$$

B. *Kappa statistic*

Kappa statistic is a measure of the proportional (or percentage) improvement by the classifier over a purely random assignment to classes. The Kappa statistic was derived to include measures of class accuracy within an overall measurement of classifier accuracy. It provides a better measure of the accuracy of a classifier than the overall accuracy, since it considers inter-class agreement.

Kappa statistic can be calculated in the following manner:

i. Construct an Error/Confusion matrix.
ii. For a confusion matrix with k rows, and k columns.
iii. Let A = the sum of k diagonal elements (Total number of correct classifications).
iv. Let B = sum of the k products (i.e. row total x column total)
v. N = Total number of pixels considered for classification.

Then Kappa statistic can be calculated by the following formula:

$$K = \frac{NA - B}{N^2 - B}$$

Interpreting Kappa statistic Measures:

- Poor agreement = Less than 0.20
- Fair agreement = 0.20 to 0.40
- Moderate agreement = 0.40 to 0.60
- Good agreement = 0.60 to 0.80
- Very good agreement = 0.80 to 1.00

C. *Execution Time*

Execution time is the time taken to run a particular algorithm. It can be defined as the difference in time between start and end execution (i.e. Time taken to display result). Basically it is system dependent. Here all the algorithms are executed on an Intel core i3 processor with 2GB RAM and 64 bit operating processor running windows 8.1.

The classification process is analyzed by the accuracy assessment of the methods of naïve bayesian, decision tree and P-Tree subspace method using naïve bayesian, decision tree . The table 1 , table 2 and table 3 shows the overall accuracy , kappa statistic and execution time of the above mentioned classification techniques for 10 different images.

Table 1: Performance evaluation of decision tree, naive bayesian classification with P-Tree subspace method based on accuracy.

Image no	Naive bayesian	Naive bayesian With P-Tree subspace method	Decision tree	Decision tree With P-Tree subspace method
Image1	39.05	53.47	47.28	52.11
Image2	42.67	69.34	54.31	60.99
Image3	40.88	61.60	56.53	58.85
Image4	46.42	60.73	62.61	69.31
Image5	51.11	70.90	63.41	71.21
Image6	40.23	54.40	45.53	52.54
Image7	56.34	62.87	65.23	70.73
Image8	58.42	63.86	65.76	68.32
Image9	40.36	54.06	45.43	62.14
Image10	51.73	66.43	52.93	63.62

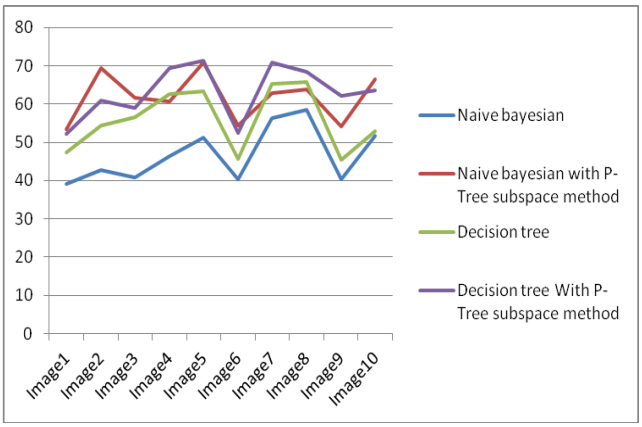

Figure 3: Performance evaluation of decision tree, naive bayesian classification with P-Tree subspace method based on accuracy.

Table 2: Performance evaluation of decision tree, naive bayesian classification with P-Tree subspace method based on Kappa statistic.

Image no	Naive bayesian	Naive bayesian With P-Tree subspace method	Decision tree	Decision tree With P-Tree subspace method
Image1	0.72	0.77	0.74	0.78
Image2	0.71	0.67	0.76	0.81
Image3	0.70	0.62	0.72	0.75
Image4	0.77	0.66	0.78	0.82
Image5	0.81	0.70	0.79	0.82
Image6	0.71	0.72	0.73	0.79
Image7	0.74	0.76	0.79	0.84
Image8	0.70	0.74	0.78	0.85
Image9	0.79	0.81	0.73	0.74
Image10	0.70	0.78	0.75	0.78

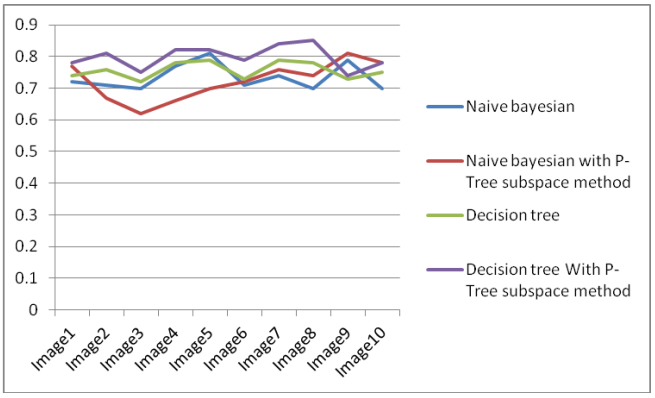

Figure 4: Performance evaluation of decision tree, naive bayesian classification with P-Tree subspace method based on Kappa statistic.

When naïve bayesian classification is compared with decision tree method because of the information gain achieved in the decision tree to select the split criteria at each attribute, it gives better result than the naïve bayesian. The proposed P-Tree subspace method is applied on these two classification techniques. Since naïve bayesian assumes independent assumption on each attribute the accuracy has been improved but the kappa statistic value for few images were less because kappa statistic measures the interrelations between the class information. The P-Tree subspace method classification provides better results in almost all cases. From fig.1, fig. 2 it is observed that for most of the images decision tree with P-Tree subspace method outperforms well. It has been observed that from Table 3 the execution time of P-Tree subspace methods reduced over conventional Naive bayesian and Decision tree classification techniques. Hence, it is ascertained that the accuracy is improved and faster execution time achieved for the P-Tree subspace method than its counter parts.

Table 3: Performance evaluation of decision tree, naive bayesian classification with P-Tree subspace method based on Execution time

Image no	**Naive Bayesian**	**Naive Bayesian with P-Tree subspace method**	**Decision tree**	**Decision tree with P-Tree Subspace method**
Image1	00:03.750	00:03.583	00:03.956	00:03.411
Image2	00:04.012	00:03.748	00:04.157	00:03.465
Image3	00:03.850	00:03.622	00:03.945	00:03.515
Image4	00:04.202	00:04.186	00:04.358	00:03.872
Image5	00:04.122	00:03.719	00:03.956	00:03.216
Image6	00:04.270	00:03.783	00:04.193	00:03.026
Image7	00:04.316	00:04.182	00:04.438	00:03.808
Image8	00:04.250	00:03.973	00:04.378	00:03.599
Image9	00:03.602	00:03.562	00:03.820	00:03.435
Image10	00:03.922	00:03.719	00:04.056	00:03.588

IV. CONCLUSIONS

The identification of class information is explored substantially by the spatial image classification. Various spatial classification techniques like naïve bayesian, decision tree were implemented in this paper. The main advantage of naïve bayesian classification technique is that it uses prior probability of class attribute to predict the unknown data. The naïve assumption achieved the class independent probability. Decision tree classification is simple to understand and resembles the human reasoning. The classification rules that are extracted from decision tree provides a knowledge base for further classification of new spatial image data. The P-Tree subspace method provides lossless and compressed data structure with its higher order bit representation of the original spatial image data. The information gain achieved through this facilitates fast processing and improves the classifier accuracy. The proposed P-Tree subspace method implemented on both naïve bayesian and decision tree classifiers. The decision tree with P-Tree subspace method achieved better accuracy and less execution time compared to naïve bayesian with P-Tree subspace method. It was ascertained that the accuracy of spatial image data classification has been improved by using P-Tree subspace method instead of normal naïve bayesian and decision tree classifiers. This work can be extended further by using association rules on classification techniques to improve the performance.

REFERENCES

[1] Amlendu Roy ,William Perrizo, Qin Ding, Qiang Ding, "Deriving High Confidence Rules from Spatial Data using Peano Count Trees", *Springer-Verlag*, LNCS 2118, July 2001.
[2] Buntine. W, Learning Classification Trees, *Statistics and Computing*, vol.2,Issue2,pp.63-73, june 1992.
[3] C. Apte, F. Damerau, and S. Weiss, " Automated Learning of Decision Rules for Text Categorization", *ACM transactions on Information Systems*, vol.12,Issue.3,pp.233-251, July 1994.
[4] C. Apte and S.J. Hong, "Predicting Equity Returns from Securities Data with Minimal Rule Generation", *Advances in Knowledge Discovery, AAAI Press / The MIT Press*, pp. 541-560, 1995.
[5] C.Z. Janikow, "Fuzzy Processing in Decision Trees", *In Proceedings of the Sixth International Symposium on Artificial Intelligence*, pp. 360-367, 1993.
[6] C. Palaniswami, A. K. Upadhyay and H. P.Maheswarappa, "Spectral mixture analysis for subpixel classification of coconut", *Current Science*, Vol. 91, No. 12, pp. 1706 -1711,December 2006.
[7] D. Lu, Q. Weng, "A survey of image classification methods and techniques for improving classification performance", *International Journal of Remote Sensing*, Vol. 28, No. 5, pp.823-870, January 2007.
[8] James A. Shine and Daniel B. Carr, "A Comparison of Classification Methods for Large Imagery Data Sets", *JSM 2002 Statistics in an ERA of Technological Change Statistical computing section*, pp.3205-3207, New York City, 11-15 August 2002.
[9] Jasinski, M. F., "Estimation of subpixel vegetation density of natural regions using satellite multispectral imagery", *IEEE Transactions on Geoscience Remote Sensing*, Vol. 34, pp. 804-813, 1996.
[10] Jehad Ali , Rehanullah Khan , Nasir Ahmad ,ImranMaqsood" Random Forests and Decision Trees" *International Journal of Computer Science Issues* (IJCSI), Vol. 9, Issue 5, No 3, pp:272-278, September 2012.
[11] K.Perumal ,R.Bhaskaran, "Supervised classification performance of multispectral images" *journal of computing*, vol 2, issue 2, ISSN: 2151-9617, Feb 2010.
[12] Marc Simard, Sasan S. Saatchi, and Gianfranco De Grandi,"The Use of Decision Tree and Multiscale Texture for Classification of JERS-1 SAR Data over Tropical Forest" *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 5, pp.2310-2321, 2000.
[13] Mahesh Paul and M.Mather, "Decision tree classification on remotely sensed data" *2nd Asian Conference on Remote Sensing*,5-9 November 2001,Singapore.
[14] Patel Brijain R,Kaushik K Rana,"A Survey on Decision tree algorithm for classification" *International Journal of Engineering Development and Research (IJEDR)*, Vol.2, Issue 1 , ISSN: 2321-9939, 2014.
[15] P.Langley ,W.Iba, and K. Thomas. "An analysis of Bayesian classifiers". *In Proceedings of the Tenth National Conferenceof Artificial Intelligence,* pages 223–228.AAAI Press, 1992.
[16] Raj Kumar, Dr. Rajesh varma, "Classification Algorithms for Data Mining: A Survey" *International Journal of Innovations in Engineering and Technology (IJIET)*,ISSN: 2319 – 1058, Vol. 1 Issue 2 August 2012.
[17] S. M. Weiss and C. A. Kulikowski, "Computer Systems that Learn: Classification and Prediction Methods from Satatistics, Neural Nets, Machine Learning, and Expert Systems", *Morgan Kaufman*, 1991.
[18] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", *Morgan Kaufmann*, 2001.
[19] William Perrizo, Anne Denton : "A Kernel-Based Semi-Naive Bayesian Classifier Using P-Trees", *Proceedings of the SIAM International Conference on Data Mining*, 2004