ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

Fuzzy Bayesian Classification for Spatial Data Streams with p-trees

D.V.Lalita Parameswari[1], Dr. M.Seetha[2], K. Ragha Deepika[3]

Sr.Asst. Professor,Dept.of CSE,GNITS,Hyderabad,India[1]

Professor,Dept.of CSE,GNITS,Hyderabad,India[2]

M.Tech Student, Dept.of CSE,GNITS,Hyderabad,India[3]

Abstract: Enormous amount of geographic data have been and continuously being collected with the advent of modern data acquisition systems like remote sensing, Global positioning system etc. To efficiently process this data, there is a great need to extract the hidden knowledge from these spatial data streams which are unpredictably large in size, complexity and dimensionality. To address these challenges, spatial data mining and geographic knowledge discovery has emerged as an active research field, focusing on the development of theory, methodology, and practice for the extraction of knowledge and useful information from massive and highly complex spatial data streams.

This paper emphasizes on Bayesian classification for spatial data. Bayesian is combined with a new data structure called peano count tree for compressing the spatial data, enhancing the scalability and reducing the classifier build time. A new technique called Fuzzy Bayesian is introduced which dramatically increased the performance of the Bayesian classifier. It is ascertained that the accuracy has been improved by Fuzzy Bayesian method with P-trees.

Keywords: Spatial data, Raster format, Bayesian classification, Fuzzy Bayesian classification, Peano Count Tree, bit sequential format (bSQ), Band sequential format(BSQ).

I. INTRODUCTION

One of the important areas of data mining is Classification [4, 6, 7, 8]. The major aim of classification is to predict the target class label for each object in the data set accurately. Basically in classification there are two phases, a training phase and a validation phase. In the training phase, a training set will be identified for building a classifier. Every record in the training set will have several attributes, one of which is the class label, indicating the class to which that record belongs. The classifier, once built and tested, can be used to predict the class label of new records that do not yet have a class label attribute value. In the validation phase, a test set will be used to test the correctness of the classifier. The classifier, once tested and certified, can be used to predict the class label of unclassified data in the future.

Spatial data is an emerging area for classification. It refers to the information related to a particular location on the earth's surface, and it allows users to look at an area in relation to other areas. This Spatial data can be satellite imagery, aerial photograph of earth’s surface, Global positioning system’s data, Remote sensed or Radar collected data etc. Spatial images were considered for classification. Here the image pixels are classified into various classes (say Greenery, Water, Urban area etc.).

Spatial data classification offers promising ways to uncover enormous amount of knowledge from the data that has been collected from satellite imagery and remote sensing applications. This data can be help full in various applications say for example to describe or explain locations of human settlements, to prepare land-use maps from satellite imagery, to predict habitat suitable for endangered species, to detect unusual warming of ocean affecting weather in nearby locations, prediction of natural calamities etc. Different techniques for classification have been proposed such as Decision trees [2, 3, 5], Bayesian, neural networks, K-Nearest Neighbours [13, 14, 16], Bayesian belief networks, fuzzy sets, and generic models, etc. Among these models, decision trees, Bayesian and K Nearest Neighbours classifiers are widely used for classification. We focus on Bayesian classification, Bayesian with P-Trees and a new technique called Fuzzy Bayesian is introduced.

Bayesian classification is a statistical classifier based on Bayes theorem, which is based on class conditional probabilities. The Peano count Trees [1, 11] are 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.

Fuzzy logic [5, 9, 19] is an emerging theory. The major advantage of this technique is that it provides natural description of the problem in simple linguistic terms rather than complex relationships among the features. This advantage of dealing with the complicated systems in a very simple way is the major reason why fuzzy logic is widely applied in various techniques. It is also possible to use fuzzy logic in various classification algorithms to classify the spatial or remotely sensed image, such that certain land cover classes are represented clearly in the resulting output. In this, a priori knowledge about spectral information for certain land cover classes is used in order to classify image in fuzzy logic classification procedure. The ability to accommodate ambiguity in the training and test data necessitates a fuzzy approach for managing the classifier training and test data.

II. Data Smoothing and Attribute Relevance

At first the data is made ready to be classified at the data preparation stage called Pre-processing. Data Pre-processing involves data cleaning which in turn involves noise reduction by applying various missing value management techniques and smoothing techniques.

III. Bayesian Classification

Given a relation, R(k, A_1, …, A_n, C), where k is the key of the relation R and A_1, …, A_n, different attributes C is the class label attribute. Given an unclassified data sample (which has values for all the attributes except for the class label C).a classification technique will predict the class label (C) value for the given sample and thus determine its class.

Here as spatial data is considered, the key, k in R, represents some pixel value (location) over a space and each A_i is a descriptive attribute of the pixels or the locations. A typical example of such spatial data is an image of the earth's surface, collected as a satellite image or an aerial photograph. The attributes can be different reflectance bands such as red, green, blue, near infra -red, infra-red, thermal infra-red, etc. The attribute values may also include the synchronized ground measurements such as soil type, yield, a weather attribute, zoning category, etc. A classifier may predict the yield (It can be considered as class label attribute) value from different reflectance band values extracted from a satellite image.

The Naïve Bayesian classification is as follows [18]:

Consider X to be a data sample with unknown class label i.e. X is a newly arrived object in the data set. Let H be the hypothesis that, X belongs to class, C. P(H|X) is the posterior probability [7, 18] of H given X. P(H) is the prior probability [7, 19] of H then according to Bayes theorem [7, 18],

$$P(H|X) = P(X|H)P(H) / P(X).$$

The 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 Bayesian Classifier will predict the class label, C_j , for a labelled data sample, X in such a way that X is labeled to that class C_j which has 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 the classes so $P(X|C_j)P(C_j)$ is to be maximized.

A. *Conditional Independence assumption*

The naive assumption 'class conditional independence of values' (The presence or absence of an attribute is independent of the presence or absence of any other attribute in the feature space) [7] is made to reduce the computational complexity. Thus,

$$P(X|C_i) = P(X_k|C_i)*\ldots*P(X_n|C_i)\text{-------}(1)$$

For categorical attributes,

$$P(X_k|C_i) = S_iX_k/S_i\text{--------}(2)$$

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 [12].

IV. Peano Count Trees (P-Trees)

Most of the spatial data collected, comes in a format called BSQ for Band Sequential [10] (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] 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) [1, 11]. A P-Tree [11] is a quadrant-based, lossless tree representation.

A. *Construction of a P-Tree*[11]

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 following figure 3.

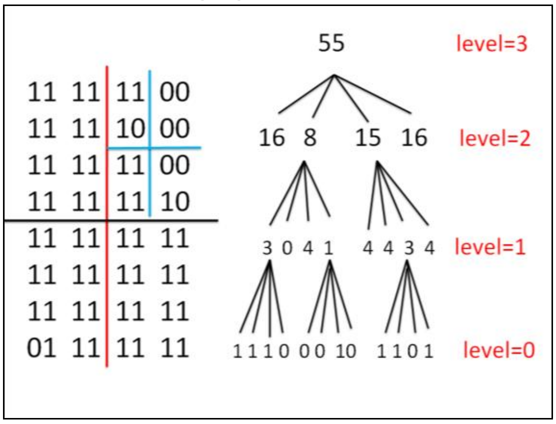


Fig 3: 8x8 Image and its P-Tree

B. *The P-Tree Algebra*

Certain operations like AND, OR, NOT [11] are performed on the basic P-Trees to combine multiple P-trees (P-Trees for the original values at any level of bit precision).

Let $P_{b,v}$ denote the P-Tree for band 'b', and value 'v', here v can be expressed in n-bit precision. For example, consider 8-bit precision for 'v', having the value 11010011. $P_{b,v}$ can be written as $P_{b,11010011}$. This can be constructed from the P-Trees as follows: $P_{b,11010011} = P_{b1}$ AND P_{b2} AND P_{b3}' AND P_{b4} AND P_{b5}' AND P_{b6}' AND P_{b7} AND P_{b8}. Where' indicates NOT operation.

The AND operation is the bit wise AND of the pixels. Similarly, any data set (Here, features extracted from the spatial image) can be represented as P-Trees.

V. Bayesian Classification with P-Trees

Most of the spatial data comes in a format called Band sequential (BSQ) [10] (bits of each band in raster order will be stored in a separate file), for ease of classification it is converted into a format called bit sequential (bSQ) [19] (each bit position of each band will be stored in a separate file). For example if a spatial image has 3 bands say Red, Green and Blue, and each pixel has 8 bits for each of the bands (R1 to R8 for red, G1 to G8 for Green and B1 to B8 for Blue).In bSQ format each bit position will have a separate file, so for the above example we get 24 files i.e. 8 for Red, 8 for Green and 8 for Blue bands respectively. These files will be converted to P-tree structure as mentioned above. Also we construct P-Tree for each of the features identified.

A. *Probability calculation*

In order to classify a tuple $X=(X_1..,X_n)$ we need to find the value of $P(X_k |C_i)$. To find the values of S_iX_k and S_i we need value P-Trees, P_k,X_k (Value P-Tree of band k, value X_k) and P_c,C_i (value P-Tree of label C, value C_i). S_iX_k= RootCount [(P_k,X_k) AND (P_c,C_i)] and
S_i= RootCount[P_c,C_i].

In this way we find the value of all probabilities for all the features extracted and substitute in (2).

VI. Fuzzy Bayesian Classification

To improve the performance of the Bayesian classifier, fuzzy logic is applied on it.

A. *Fuzzy Logic* [5,9,19]

Fuzzy rules are nothing but simple conditional IF THEN rules of the form

IF x is A
THEN y is B.

Where x and y - linguistic variables, A and B -linguistic values, here the linguistic values are determined according to the universe of discourse i.e. x and y in this case.

In general algorithms the rules can have precise values to define IF THEN rules but in fuzzy logic we may consider certain missing values in between the precise rules being defined, by formulating the fuzzy rules accordingly.

For example, consider a simple temperature regulator in an air conditioner that regulates temperature in a room according to the room temperature.

The fuzzy rules to operate the temperature regulator in an air conditioner can be as follows:

IF room temperature IS very cold THEN stop Air conditioner.

IF Room temperature IS cold THEN increase the air conditioner temperature to the prefixed value.

IF Room temperature IS normal THEN maintain level

IF Room temperature IS hot THEN reduce the air conditioners temperature's value.

There is no "ELSE" statement here because all of the rules need to be evaluated, because the temperature can be "normal" and "cold" at the same time at different degrees. There can also be adverbs such as "very", or "somewhat", which allows more number of situations to be considered

Certain operators of Boolean logic such as AND, OR, and NOT also exist in fuzzy logic.

Combining fuzzy logic with Bayesian classification facilitates increase in the probabilities of each of the classes by considering many of the pixels that were left un-classified or incorrectly classified by applying Bayesian classification alone.

VII. Fuzzy Bayesian Classification with P-Trees

The Fuzzy logic and the Peano count tree data structure explored in the previous sections have been combined and applied over Bayesian classification for spatial data.

This new technique not only increases the accuracy of classification drastically but also helps in building the classifier at a faster rate.

VIII. RESULTS & DISCUSSIONS

The Classification process is analyzed by the accuracy assessment of the methods of Bayesian, Bayesian with P-trees ,Fuzzy Bayesian and Fuzzy Bayesian with P-trees techniques. Accuracy assessment can be performed by comparing two sources of information. 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. Accuracy assessment was done using independent sample points. The table 1 shows the accuracy of the Bayesian Classification for 5 different images. The accuracy was calculated by kappa statistic measure. Bayesian with P-Trees improves the accuracy than the Bayesian because P-Trees stores the Higher order bit information. When HOB information is combined with prior probability this technique gives better than Bayesian Technique. For all 5 images the Fuzzy Bayesian with P-trees method gives more accuracy compared to other Techniques. The Table 2 represents the sample images and their classified images.

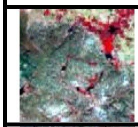
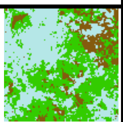
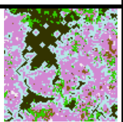
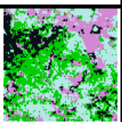
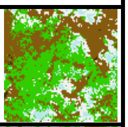
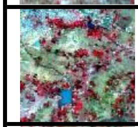
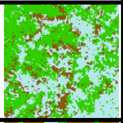
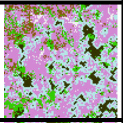
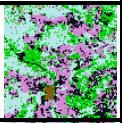
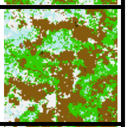
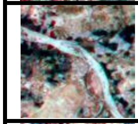
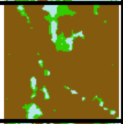
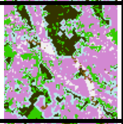
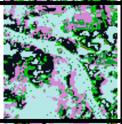
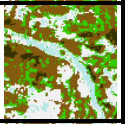
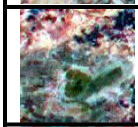
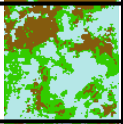
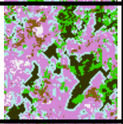
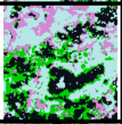
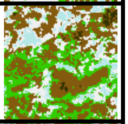
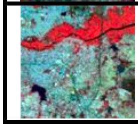
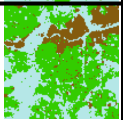
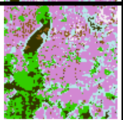
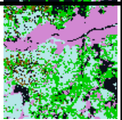
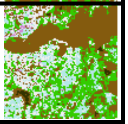
ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

Table 1: Accuracy Analysis for Bayesian Classification

SNo	Bayesian		Bayesian with P-Trees		Fuzzy Bayesian		Fuzzy Bayesian with P-Trees	
	A	**K**	**A**	**K**	**A**	**K**	**A**	**K**
1	39.05	0.72	53.47	0.77	58.60	0.73	87.83	0.98
2	42.67	0.71	69.34	0.67	66.87	0.72	82.86	0.99
3	40.88	0.70	61.60	0.62	61.86	0.71	83.71	0.96
4	46.42	0.77	60.73	0.66	64.06	0.78	87.63	0.98
5	51.11	0.81	70.90	0.70	76.23	0.82	86.36	0.96

A-Overall Accuracy, K-Cohen's Kappa coefficient.

Table 2: Classified Images

Input Image	Output Image			
	Bayesian	Bayesian with P-Trees	Fuzzy Bayesian	Fuzzy Bayesian with P-Trees
				
				
				
				
				

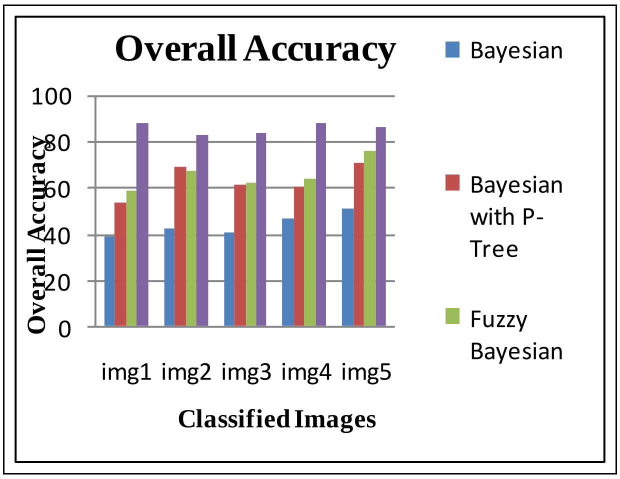


Fig 4: Overall accuracy for classified images

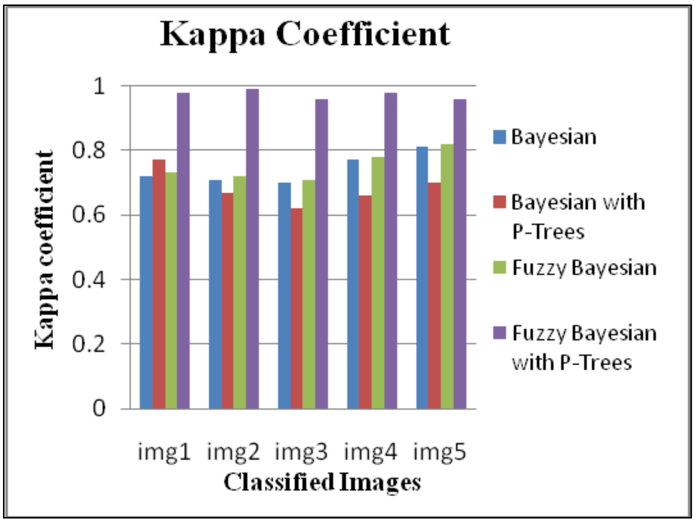


Fig 5: Kappa coefficient for classified images

For classification, IRS 1D satellite's LISS 3 sensor images were used. Fig 4 shows the Overall accuracy for classified images, Fig 5 represents the Kappa coefficient for classified images. It was observed that the Classification accuracy improves remarkably upon the usage of the P-tree and Fuzzy based classification methods than the normal Bayesian classification. It has been ascertained that P-tree based fuzzy Bayesian gives better results.

IX. CONCLUSIONS

Bayesian classification was employed effectively for Image Classification. The main advantage of this technique is that it uses prior probability of class attribute to predict the unknown data. Bayesian classification when combined with a new data structure called Peano count tree and an emerging technique called fuzzy logic gives better performance than in its pure state. P-trees are highly distributed, fault tolerant and a quadrant based tree.

P-Trees provide a lossless and compressed representation of image data. Fuzzy logic proved to be an excellent choice for image classification applications since it mimics human control logic. Fuzzy rules are very helpful in interpreting good training data for the Bayesian classifier there by proving to be a better classifier of spatial data than the normal Bayesian classification Technique. It was ascertained that the accuracy of satellite images classification can be improved by using Peano count tree or Fuzzy based classification methods than the normal Bayesian classifier.The study can be extended further for taking more classes to identify that are misclassified.

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] Buntine. W, "Learning Classification Trees, Statistics and Computing", 2:63-73, 1992.

[3] C. Apte, F. Damerau, and S. Weiss, "Automated Learning of Decision Rules for Text Categorization, ACM Transactions on Information Systems", 12(3):233-251, July 1994.

[4] C. Apte and S.J. Hong, "Predicting Equity Returns from Securities Data with Minimal Rule Generation, In Advances in Knowledge Discovery", AAAI Press / The MIT Press, pages 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] 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.

[7] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2001.

[8] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, "Machine Learning, Neural and Statistical Classification", Ellis Horwood, 1994.

[9] C.Z. Janikow, "A Genetic Algorithm Method for Optimizing the Fuzzy Component of a Fuzzy Decision Tree", In GA for Pattern Recognition, S. Pal & P. Wang (eds.), CRC Press, pp. 253-282,1995.

[10] Joseph Collins,"Understanding Rasters", UNRUH, 2006.

[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] Md Abdul Maleq Khan, Fast Distance Metric Based Data Mining Techniques: k-Nearest-Neighbour Classification and k-Clustering. A Thesis Submitted to the Graduate Faculty Of the North Dakota State University Of Agriculture and Applied Science, 2002.

[13] Maleq Khan, Qin Ding, and William Perrizo, k-Nearest Neighbour Classification on Spatial Data Streams, 2002.

[14] George Y. Lu, David W. Wong, An adaptive inverse-distance weighting spatial interpolation technique, 2007.

[15] W. K. Pratt. Digital Image Processing, 2007.

[16] http://en.wikipedia.org/wiki/Inverse_distance_weighting.

[17] Pang-ning Tan, Vipin kumar, Michael Steinbach. "Introduction to data mining",2006 .

[18] 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.

[19] http://en.wikipedia.org/wiki/Fuzzy_logic

[20] Pablo Ruiz, Javier Mateos, Gustavo Camps, Rafael Molina and Aggelos K. Katsaggelos. Bayesian Active Remote Sensing Image Classification. IEEE 2013.