چهارمین سمینار آمار فضایی و کاربردهای آن

۵–۶ آبان ماه ۱۴۰۰
27-28 Oct. 2021

4th Seminar on Spatial
Statistics and Its Applications

Spatially Classification Decision Trees: Fundamentals and Some Extensions

Tahereh Alami[*], Mahdi Doostparast
Department of Statistics, Ferdowsi University of Mashhad, Mashhad, Iran.

Abstract:
In classical statistics, observations of a random variable are commonly assumed to be independent and identically distributed. Most statistical learning techniques such as Classification and Regression Trees (CART) assume independent samples to compute classification rules. But this assumption is often violated in spatial datasets and it may not be efficient for analyzing spatial data. The CART algorithm is adapted to the case of spatially dependent samples by three strategies; the first one is the weighting of the data according to their spatial pattern, the second is spatial Entropy used as the splitting criterion and in the third, we combine these two strategies to achieve more accuracy. This method is evaluated on a classical dataset to highlight its advantages and drawbacks.

Keywords: Classification, CART, Spatial data, Spatial Entropy, Kriging weight.
Mathematics Subject Classification (2010): 60G50, 60B34.

1 Introduction

Spatial Statistics is different from the classical Statistics because of the spatial correlation among observations. Most existing models are applicable for independent observations and may not be efficient for analyzing spatial datasets. The problem of prediction in a new location is of great importance and these models need to be adjusted for use by spatial data. Today, because of use of remote sensing (RS) devices, spatial data is widely used in the environmental and ecological researches. Examples of spatial data applications

[*] Speaker: tahereh.alami@mail.um.ac.ir

include Earth Science, Urban Informatics, Geosocial Media Analytics, and Public Health (Jiang , 2018). Recording the location of any data provides a lot of information, and so using this new information can make the results of the statistical learning algorithms more accurate. Recently, one of the methods that have attracted the attention of researchers is spatial decision trees. Several forms of the spatial decision tree were considered in the literature. In the following, we will discuss the process of formation of these methods.

The decision tree is a popular model in statistical learning algorithms because of its simplicity and comprehensibility. A tree-based algorithm uses a greedy top-down approach to create a tree through recursive binary splitting. There are various algorithms for making a decision tree such as ID3, C4.5, C5.0, all of which have been proposed by Quinlan, J. R. Furthermore, *Classification and Regression Tree* (CART) introduced by Breiman et al. (1984) is one of the most widely used tree-based methods. The CART algorithm is also capable of processing continuous and nominal attributes as targets and predictors.

There are several methods for the spatial decision tree that support spatial dataset. Li and Claramunt (2006) introduced a generalized ID3 algorithm as the spatial decision tree based on spatial diversity coefficients that measured spatial entropy for the georeferenced dataset. The spatial diversity is adapted to either discrete or continuous spaces and not limited to a bounded boundary as distances rather than neighborhoods are considered (Claramunt , 2005). An extension of the CART algorithm to the spatial scheme focuses on the applications of decision trees in environmental data. The CART algorithm applies to the case of spatially dependent samples by weighting the data according to their spatial pattern and also uses spatial estimates of class probabilities in algorithm structure and misclassification error (Bel et al. , 2005, 2009). Another spatial decision tree algorithm is suggested by Sitanggang et al. (2011). It is an extended version of ID3 that uses the spatial IG to select the best split layer and applies to discrete features shown as point, lines and polygonal. Jiang et al. (2014) proposed a focal-test-based spatial decision tree (FTDST) applied to the raster dataset.

In this paper, we provide a combination of the impurity function and the weighting process in the CART algorithm which takes into account their spatial dependencies of the observation. The proposed model uses both spatial Entropy as an impurity function and the weighted probabilities for the measure of the proportion of a node. The paper is organized as follows. In Section 2, the CART approach is briefly reviewed. Section 3 provides some existing spatial decision tree models and our proposed method. Section 4 shows results on a real spatial dataset and their efficiencies are compared. The conclusions will be presented at the end of the paper.

2 CART

Let (Ω, F, P) be a given triple probability space. Suppose that we have an independent random sample of the random vector $\mathbf{X} = (X_1, X_2, ..., X_p)$ as explanatory variables and Y be a categorical response variable. Consider a decision tree T with one of its nodes t. The CART algorithm splits a node t into two sub-nodes $t_L = \{\mathbf{X} : X_m \leq s\}$ and $t_R = \{\mathbf{X} : X_m > s\}$ according to a threshold s, called split point, on one of the explanatory variables or a subset of the labels of categorical explanatory variables. An optimal split selects the variable (Xm) and split point (s) such that the sub-nodes be purer than the parent node. For this aim, the reduction of an impurity criterion between a node t and the two sub-nodes t_L and t_R should be maximized, that is

$$\triangle \text{imp}(s,t) = \text{imp}(t) - [p_L \text{imp}(t_L) + p_R \text{imp}(t_R)], \tag{2.1}$$

where imp(t) is the impurity criterion of node t and p_L and p_R are the two corresponding proportions of the dataset falling into the sub-nodes given by N_{t_L}/N_t and N_{t_R}/N_t, respectively. imp(.) is a non-negative function of proportions and it is maximized when all classes in the node t are together with equal proportions and is minimized when only one class remains in the node t. Initially, CART considered the Gini index as the impurity criterion, but later Entropy was also added to CART impurity criteria (Wu , 2008). They are defined as follows, respectively

$$G(t) = \sum_{i \neq j} p(i|t)\ p(j|t) = 1 - \sum_j p(j|t)^2, \tag{2.2}$$

and

$$E(t) = -\sum_j p(j|t)\ \log\ p(j|t), \tag{2.3}$$

where $p(j|t) = \sum_{i \in t} I(Y_i = j)/N_t$ is the samples proportion of class j in a node t. If Entropy (2.3) is chosen as an impurity criterion, Equation (2.1) is called Information Gain (IG). The growth of the tree continues until all samples in a terminal node have the same class or their number are very small. Finally, the class with the maximum probability is assigned to each terminal node as the label. To avoid overfitting, the final tree T is pruned to reach an optimal size tree.

3 Spatially decision trees

Today, the performance of the non-spatial decision tree algorithms have been improved by involving spatial relationships among the spatial data. In this section, we focus on a well-known decision tree algorithm called CART and three frameworks are reviewed to adjust the CART algorithm with spatial observations: (1) CART with spatial entropy (Spatial CART), (2) CART with spatial weights (Weighted CART), and (3) CART with spatially entropy and weights (Spatial Weighted CART).

3.1 CART with spatial entropy (Spatial CART)

Spatial entropy can be availed when the values of the response variable have a highly spatial correlated with neighboring values. One of the well known spatial entropy criterion uses spatial diversity and is used for discrete and continuous domains (Claramunt , 2005). The diversity should increases as different entities are close to each other and decreases as similar entities are close to each other. According to the this rule, the spatial entropy must also take into account the spatial features of variables. Therefore, spatial diversity coefficients should increase when either the average distance between the entities belonging to a given category decreases or the average distance between the entities of a given category and the entities of all the other categories increases and vice versa. These quantities are, respectively, named *intra-distance* (${d_i}^{int}$) and *extra-distance* (${d_i}^{ext}$) and defined by

$$ {d_i}^{int} = \begin{cases} \frac{1}{|C_i| \times (|C_i|-1)} \sum_{j \in C_i} \sum_{k \in C_i; k \neq j} \mathrm{dist}(j,k) & |C_i| > 1 \\ \lambda & otherwise \end{cases} \tag{3.1} $$

and

$$ {d_i}^{ext} = \begin{cases} \frac{1}{|C_i| \times |C-C_i|} \sum_{j \in C_i} \sum_{k \in (C-C_i)} \mathrm{dist}(j,k) & C_i \neq C \\ \beta & otherwise \end{cases} \tag{3.2} $$

where C is the set of entities on a given dataset, C_i denotes the subset of entities of the ith category of the dataset and $\mathrm{dist}(j,k)$ is the distance between the entities j and k. Finally, to calculate spatial entropy of node t, the spatial diversity coefficients are used as follows:

$$ E_s(t) = -\sum_{i=1}^{n} \frac{{d_i}^{int}}{{d_i}^{ext}} \, p(i|t) \log_2 p(i|t). \tag{3.3} $$

Li and Claramunt (2006) introduced a spatial decision tree based on ID3 algorithm, which uses spatial entropy in the IG and is called spatial IG. In this paper, we apply spatial IG(IG$_t$) in the CART algorithm for the node t and its sub-nodes as follows:

$$ \mathrm{IG}_t(s,t) = E_s(t) - [p_L E_s(t_L) + p_R E_s(t_R)] \tag{3.4} $$

3.2 CART with spatial weights (Weighted CART)

Bel et al. (2009) is adapted the CART algorithm to the case of spatially dependent samples by weighting the data according to their spatial pattern. The idea of weighting the spatial data comes from the fact that in the spatial domain, the clustered data have similar behavior to each other than far data, and these data must give less weight, i.e. the data should be decluster. They considered two type of weights to determine the weight of data points: Voronoi and kriging weights. The weight in each of these methods is proportional to the inverse of the local density of the data and gives a lower weight to the clustered data. These methods are evaluated according to the locations of the data points $s_1, ..., s_n$. In weighted CART, the estimated proportions and misclassification are calculated by

$$\widehat{p}(j|t) = \frac{1}{\sum_{i\in t} w_i} \sum_{i\in t} w_i I(Y(\mathbf{s}_i) = j), \tag{3.5}$$

and

$$\widehat{R}(T) = \sum_{i=1}^{n} w_i I[T(X_1(\mathbf{s}_i), ..., X_p(\mathbf{s}_i)) \neq Y(\mathbf{s}_i)] \tag{3.6}$$

respectively, where w_i is weight corresponding to the ith sample of the dataset and $\sum_{i=1}^{n} w_i = 1$.

Voronoi weight

A Voronoï tessellation is generated by the sample locations, so that it is formed by the set of points of D closer to s than to any other sample location $s_1, ..., s_n$. Closer observations produce smaller cells while sparse data produce larger ones. An estimator of the local density of the sample design is the inverse of voronoi cells around each location of observations. Therefore, closer observations produce smaller cells and thus gives smaller weights which are proportional to the inverse of the local density of their location and vice versa.

Kriging weight

The kriging method is another way for deriving the weights, which are based on a covariance function modeling that takes into account spatial dependencies on the data under study. We first recall briefly some notions of geostatistics that will be used in this way. The value of a random filed $\{Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$ in specific location $\mathbf{s}_0$ is predicted based on the observed data $\{Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n)\}$. Consider the random function $Z(\mathbf{s})$ with a mean structure $\mu(\mathbf{s})$ and the variogram function $2\gamma(\mathbf{h})$, as follows:

$$(i)\ E(Z(\mathbf{s})) = \mu(\mathbf{s}), \qquad \forall \mathbf{s} \in D, \tag{3.7}$$

$$(ii)\ \mathrm{Var}(Z(\mathbf{s}_1) - Z(\mathbf{s}_2)) = 2\gamma(\mathbf{s}_1 - \mathbf{s}_2) = 2\gamma(\mathbf{h}), \qquad \forall \mathbf{s}_1, \mathbf{s}_2 \in D. \tag{3.8}$$

A random filed $Z(.)$ with constant mean and satisfying $\mathrm{cov}(Z(\mathbf{s}_1), Z(\mathbf{s}_2)) = C(\mathbf{s}_1 - \mathbf{s}_2) = C(\mathbf{h}), \forall \mathbf{s}_1, \mathbf{s}_2 \in D$ is called *second-order* or *weak stationary* where $C(.)$ is covariance structure of random filed $Z(.)$. Under the weak stationary property of the random filed $Z(.)$, the variogram and covariance function are related as follows:

$$2\gamma(\mathbf{h}) = 2(C(\mathbf{0}) - C(\mathbf{h})), \forall \mathbf{h} \geq 0. \tag{3.9}$$

In this situation the best linear unbiased predictor of random function $Z(\mathbf{s}_0)$ is obtained by minimizing the mean squared prediction error in the class of all linear unbiased predictors and denoted by $\widehat{Z}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i)$ where $\sum_{i=1}^{n} \lambda_i = 1$ and λ_i's are kriging weights. The constraint $\sum_{i=1}^{n} \lambda_i = 1$ guarantees unbiasedness, that is $E(\widehat{Z}(\mathbf{s}_0)) = \mu$. The vector $(\lambda_1, \lambda_2, ..., \lambda_n)$ is the solution of the system

$$\begin{pmatrix} \mathbf{C} & \mathbf{1} \\ \mathbf{1}^T & 1 \end{pmatrix} \begin{pmatrix} \Lambda \\ m \end{pmatrix} = \begin{pmatrix} \mathbf{C}_D \\ 1 \end{pmatrix} \tag{3.10}$$

where $\mathbf{1}$ is a vector of ones of length n, m is the Lagrange parameter, $\mathbf{C}$ is the matrix with elements of $C(\mathbf{s}_i, \mathbf{s}_j)$ and $\mathbf{C}_D$ is the vector with elements $C(\mathbf{s}_i, \mathbf{s}_0)$. For more details, see e.g., Cressie (1993).

In predicting the value of a random field in a new location, if two data points have a short distance from each other, they transmit almost identical information compared to separate data points, and thus less kriging weight is allocated to them. This property of kriging is known as de-clustering (Diggle and Ribeiro , 2007). The values of the kriging weights can be negative, but the algorithm needs positive weights to compute the probabilities. Therefore, a positiveness condition applies to the solution of Equation (3.10).

3.3 CART with spatially entropy and weights (Spatial Weighted CART)

Spatial Weighted CART uses both spatial impurity (especially spatial entropy) and spatial weights to build a decision tree. In the process of constructing a Spatial Weighted CART, we use weighted probabilities with two Voronoi and kriging weights to calculate the probabilities in each node and spatial entropy as the impurity criterion of the tree. The combination of weighted probabilities and spatial impurity criterion in tree construction significantly reduces misclassification error. In the next section, we compare the proposed method with others.

Table 1: Misclassification error of Swiss Jura data-set for different methods.

Model	**Weight**	**Impurity Criteria**		
		Gini index	Entropy	Spatial entropy
CART	-	0.43	0.39	0.41
Weighted CART	Voronoi	0.36	**0.353**	**0.315**
	Kriging	**0.33**	0.379	0.351

4 Numerical results

To evaluate the performance of the proposed model, we used a data set with numerical explanatory attributes and classification response that includes spatial information. The Swiss Jura data set, provided by Atteia et al. (1994), is on the concentration of seven heavy metals (cadmium, cobalt, chromium, copper, nickel, lead and zinc) in the topsoil at 359 locations. The area is made up of five types of Jurassic limestone types: Argovian, Kimmeridgian, Sequanian, Portlandian, Quaternary. Following Bel et al. (2009), we combine three classes Sequanian, Portlandian, and Quaternary into one class. Finally, three classes are coded as follow: 1 for Argovian, 2 for Kimmeridgian, and 3 for the others. The training sample contains 259 samples and the rest 100 samples are used as the test sample.

Table 1 provides misclassification errors values for different weights and impurity criteria. As shown in Table 1, the values of misclassification error in CART with two traditional impurity function, Gini index and Entropy, are 43% and 39% while the values of misclassification using the weighted probabilities are less than the traditional CART. In the weighted CART model with Gini index, the kriging weight gives lower misclassification which is equal to 33%, while the lower values of misclassification error of trees with Entropy and spatial entropy criterion are obtained by using Voronoi weight which are equal to 35.3% and 31.5%, respectively. The value of misclassification error for the spatial entropy criterion is 41% which is higher than standard CART, but then using weighted probabilities, it is observed that the misclassification decreases significantly. In this case, the lowest misclassification rate among all methods is obtained for Voronoi weight, which is equal to 31.5% and indicates the suitable performance of this method.

Conclusion

The notable conclusion of the paper is that the usage of weighted probabilities has a significant effect on reducing the amount of misclassification error when using the spatial entropy compared to the Gini and Entropy criteria, which indicates the efficiency of the proposed method. Also, the misclassification values of trees for each of the weights, in both Entropy and spatial entropy criteria, show that the use of spatial entropy criterion is more effective in the efficiency of weighting methods. More extensions and findings are given in Alami and Doostparast (2021).

References

Bel, L., Allard, D., Laurent, J.M., Cheddadi, R. and Bar-Hen, A. (2009). CART algorithm for spatial data: Application to environmental and ecological data, *Computational Statistics & Data Analysis,* **53**, 3082-3093.

Bel, L., Laurent, J. M., Bar-Hen, A., Allard, D., & Cheddadi, R. (2005), A spatial extension of CART: application to classification of ecological data, *In Geostatistics for environmental applications, Springer, Berlin, Heidelberg.*, 99-109.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*, CRC press.

Claramunt, C. (2005). A spatial form of diversity, *In International Conference on Spatial Information Theory, Springer, Berlin, Heidelberg,* 218-231.

Cressie, N. (1993). *Statistics for Spatial Data*, John Wiley, New York.

Diggle, P.J., Ribeiro, P.J. (2007). *Model-based Geostatistics*, Springer series in statistics.

Jiang, Z., Shekhar, S., Zhou, X., Knight, J. and Corcoran, J. (2014). Focal-test-based spatial decision tree learning, *IEEE Transactions on Knowledge and Data Engineering,* **27(6)**, 1547-1559.

Jiang, Z. (2018), A survey on spatial prediction methods, *IEEE Transactions on Knowledge and Data Engineering,* **31(9)**, 1645-1664.

Li, X. and Claramunt, C. (2006). A spatial entropybased decision tree for classification of geographical information, *Transactions in GIS,* **10(3)**, 451-467.

Sitanggang, I.S., Yaakob, R., Mustapha, N. and Nuruddin, A.A.B. (2011). An extended ID3 decision tree algorithm for spatial data, *In Proceedings 2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services,* 48-53, IEEE.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Philip, S.Y. and Zhou, Z.H. (2008). The top ten algorithms in data mining, *Knowledge and information systems,* **14***(1), pp.1-37.*