Expert Systems With Applications xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Applying data mining techniques for spatial distribution analysis of plant species co-occurrences

Q1 Luís Alexandre Estevão da Silva[a,*], Marinez Ferreira de Siqueira[a], Flávia dos Santos Pinto[a], Felipe Sodré Barros[a], Geraldo Zimbrão da Silva[b], Jano Moreira de Souza[b]

[a] *Instituto de Pesquisas Jardim Botânico do Rio de Janeiro, Brazil*
[b] *COPPE/UFRJ, Universidade Federal do Rio de Janeiro, Brazil*

ARTICLE INFO

Keywords:
Data analysis
Data mining
Association rules
Knowledge management applications
Knowledge discovery

ABSTRACT

The continuous growth of biodiversity databases has led to a search for techniques that can assist researchers. This paper presents a method for the analysis of occurrences of pairs and groups of species that aims to identify patterns in co-occurrences through the application of association rules of data mining. We propose, implement and evaluate a tool to help ecologists formulate and validate hypotheses regarding co-occurrence between two or more species. To validate our approach, we analyzed the occurrence of species with a dataset from the 50-ha Forest Dynamics Project on Barro Colorado Island (BCI). Three case studies were developed based on this tropical forest to evaluate patterns of positive and negative correlation. Our tool can be used to point co-occurrence in a multi-scale form and for multi-species, simultaneously, accelerating the identification process for the Spatial Point Pattern Analysis. This paper demonstrates that data mining, which has been used successfully in applications such as business and consumer profile analysis, can be a useful resource in ecology.

1. Introduction

Amounts of available data have grown exponentially in recent years, causing discussions about the need for new methods to access, analyze and manage biological data (Howe et al., 2008), (Wu, Zhu, Wu, & Ding, 2014); one example is the Global Biodiversity Information Facility (Gbif, 2010). Databases on scientific collections have accompanied the increase in data, partly as a consequence of the need to obtain a shorter response time for research in biology and also due to the concern for biodiversity conservation. However, analysis of existing patterns in such large data sources has become a challenge, particularly for a few researchers who are needed for multiple studies in the area (Drew, 2011; Hopkins & Freckleton, 2002).

The difficulty in the extraction of knowledge from large databases has been perceived by many sectors of the economy, and much of the research has identified data mining as an efficient way to extract knowledge from these sources (Aggarwal, 2015; Han, Kamber, & Pei, 2011). It has been used successfully in areas such as customer relationship management, to analyze and build prediction models in the commercial area (Nettleton, 2014); education, predicting the success rate of students enrolled in courses (Natek & Zwilling, 2014; Peña-Ayala, 2014); text mining (Zhao, Cen, Javaheri, Sepehri, & Teimourpour, 2014); internet security (Abdelhamid, Ayesh, & Thabtah, 2014), among other areas. In ecology, its use has been increasing in applications, indicating potential uses of data mining (Hochachka et al., 2007); comparing and using classification algorithms to predicting the potential habitat of species (Dlamini, 2011; Lorena et al., 2011; Pino-Mejías et al., 2010); cluster analysis to identify regions with similar ecological conditions (Kumar, Mills, Hoffman, & Hargrove, 2011) or grouping species into disjunct sets with similar co-association values using *k-means* clustering algorithm (Flügge, Olhede, & Murrell, 2014); and for forest growing stock modeling with decision tree algorithm (Debeljak, Poljanec, & Ženko, 2014). But the adoption has been slower than the previously mentioned fields (Inman-Narahari, Giardina, Ostertag, Cordell, & Sack, 2010).

In addition to the large volumes of data, other difficult issues have demanded the work of ecologists, among which the study of plant communities (Baselga & Araújo, 2010) is notable for involving complex process analysis (Swenson, 2013) given that little is known of the processes governing the composition of plant communities (Uriarte, Condit, Canham, & Hubbell, 2004), particularly in tropical forests that are complex ecosystems where many species coexist (Johnson, Domínguez-García, Donetti, & Muñoz, 2014). In this context, one of

[*] Corresponding author. Tel.: +5502126671744.
E-mail addresses: estevao@jbrj.gov.br, luis.alexandre@ucp.br (L.A.E. da Silva), marinez@jbrj.gov.br (M.F. de Siqueira), flaviasantospinto@gmail.com (F.d.S. Pinto), felipe.b4rros@gmail.com (F.S. Barros), zimbrao@cos.ufrj.br (G.Z. da Silva), jano@cos.ufrj.br (J.M. de Souza).

http://dx.doi.org/10.1016/j.eswa.2015.08.031
0957-4174/© 2015 Published by Elsevier Ltd.

the most important ecological relationships between any species is their co-occurrence (Neeson & Mandelik, 2014). These interactions can be of positive (Monge & Gornish, 2014) or negative (Veech, 2014) type, such as facilitation and segregation, respectively. Quantifying this relationship between species allows several studies as selection of indicator species (Culmsee et al., 2014) and many other analyses in conservation ecology.

Spatial Points Process are defined as set of observations ($X^1, X^2, \ldots, X^3$) within study area 'A', where each point has at least a pair of coordinates (Lloyd, 2006). Others information can be associated, like species identification, elevation, collector, among other. This process is distance based and, analyze the spatial structure rather than its variation thru the space. Thus, it is possible to infer spatial association in a univariate (one points process. i.e: one species) or bivariate spatial point process (two different point process; two i.e: species). Among the methods, stands out Ripley's K-function (Ripley, 1977) that is commonly applied in plant ecology to detect the spatial distribution of individuals within communities and the underlying processes controlling the observed patterns (Haase, 1995; Zhang, Hu, Zhu, & Ni, 2012). The K-function estimates (Bivand, Pebesma, & Gómez-Rubio, 2008) the expected number of events found in a given distance (t) of each point or event and constructing ever-increasing concentric circles of radius t, as follows (Lancaster & Downes, 2004):

$$\mathrm{K}(t) = \mathrm{n}^{-2}\mathrm{A}\sum_{i}^{n}\sum_{j\neq i}^{n} w_{\mathrm{ij}} I_{\mathrm{t}}(d_{\mathrm{ij}})$$

As above, n is the total number of events; w_{ij} is a weighting factor to correct edge effects; A is the study plot area; I_t is a counter which is set to 1 if the distance d_{ij} between the ith and jth points (pairwise mode) is less than or equal to t, otherwise is equal 0. $K(t)$ is presented as the linearized L-function $\mathrm{L(t)} = [\mathrm{K(t)} / \pi]^{1/2}$, as proposed by Besag (Besag, 1977). Under Complete Spatial Randomness (CSR), $\mathrm{L(t)} = 0$; values of $\mathrm{L(t)} > 0$ indicate attraction between the two events; values less than 0 indicate repulsion.

However, the application of these techniques by bivariate form requires a full pairwise comparison, which is usually not practical in multivariate event sets (Perry, Miller, & Enright, 2006). This difficulty affects the type of application in analysis mainly in a megadiversity context. Another problem is the limited ability to separate scales (Detto & Muller-Landau, 2013), failing to demonstrating whether, for example, deviations from complete spatial randomness at small distances are due to causes acting at small scales or at larger scales (Loosmore & Ford, 2006). Another initiative is the kdot function (Baddeley & Turner, 2005) that investigates the relationship of co-occurrence of a point (species) to any point (species) in space. However, this approach does not allows identify which species have co-occurrence, which is another differential of data mining techniques presented in our approach.

Given the scenario presented and also considering that ecologists have long been researching effective methods (Veech, 2013) to understand the mechanisms of coexistence, competition and distribution of species (Wiegand et al., 2012) and, despite all the progress achieved in data mining area after studying the traditional methods used in ecology for the analysis of point patterns, we found no research using association rules with inventory plots data in ecological applications, which has motivated this study. Thus we offer a method for the analysis multi-scale form and for multi-species associated with the environment to assist ecologists in the assessment of patterns of occurrences of species in plant communities. This method is also based on the need for research with a larger number of variables to explain the occurrences in an environment as diverse as the tropical forest. Therefore, this study considers a new type of application of association rules in the investigation of patterns of occurrences of species' pairs and groups and emphasizes that the development of a specific method for extracting knowledge from biodiversity databases is necessary, given the large volumes of data available.

Table 1
Example with five transactions and six species.

Treeid	Items
1	sp_1, sp_2
2	sp_1, sp_3, sp_4, sp_5
3	sp_2, sp _3, sp _4, sp _6
4	sp _1, sp _2, sp _3, sp _4
5	sp _1, sp _2, sp _3, sp _6

2. Association rules

Notable among the categories in data mining is association rules (Agrawal, Imieliński, & Swami, 1993), which aims to discover frequently appearing items from a set of transactions, deriving rules from associations among the items involved in each transaction (Wu et al., 2008) without implying causality (Tan, 2007). A *transaction* corresponds to the set of items in an operation, such as products purchased by a particular customer for market basket analysis (Brin, Motwani, & Silverstein, 1997). The format of a *rule* can be exemplified as a logical statement between two items, *A* (antecedent) and *B* (consequent), as sp_A → sp_B, and can be comprehended as a pattern where sp_A and sp_B appear together. A pattern is interesting if it helps define a hypothesis, and an interesting pattern represents knowledge. This category of data mining is also used in such areas as business management (Cil, 2012), consumer profile analysis (Liao, Chen, & Deng, 2010), recommender systems (Adomavicius & Tuzhilin, 2005; Lazcorreta, Botella, & Fernández-Caballero, 2008), adverse drug reactions (Ji et al., 2013) and genetics (Lin, Huang, & Leu, 2011).

2.1. Notations and definitions

This study adapted the association rules to be used in data plots obtained from floristic inventories, or a method for assessment distribution of plant species in a local area, where the area is divided into plots of equal size. Several inventory protocols are used according to the purpose of the work (Gordon & Newton, 2006) and formalized as follows: (a) a set of items (itemset) is a subset of the set of specimens (individuals) from all species examined, and (b) each transaction is composed of all specimens within the specified distance (radius in meters) of a given specimen. To illustrate, Table 1 presents transactions involving six supposed species found near five trees identified by the attribute *Treeid* from 1 to 5.

Several metrics can be used to evaluate the quality of the rules generated by algorithms of association rules. We used the following set of measures: *support*, *confidence*, *lift* (Han et al., 2011); *chi-square* (Hahsler, Gruen, & Hornik, 2005) and *p-value* (Liu, Zhang, & Wong, 2011). The first two measures were used to define the species' pairs and groups, the third to evaluate the association type (positive or negative) and the last two to assess the degree of independence of the species. For example, considering two species sp_A and sp_ B, the *support* is the probability *P* of transactions with both species and is defined as *support* (sp_A → sp_B) = *P* (sp_A ∪ sp_B). The *confidence* is defined as the frequency with which items are found in the transaction sp_A containing sp_B and is defined as the conditional probability *conf* (sp_A → sp_B) = *P* (sp_A | sp_B). The *lift* is the measure of importance of a rule and can be defined by *P* (sp_A ∪ sp_B) / (*P* (sp_A) * *P* (sp_B)). Their values can be interpreted in the following ways. *A* value equal to one indicates independence between the antecedent and the consequent of the rule. *A* value greater than one means that the items have a positive correlation. In other words, the appearance of items with a

Please cite this article as: L.A.E. da Silva et al., Applying data mining techniques for spatial distribution analysis of plant species co-occurrences, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.08.031