Analysis of Indian Estuarine Data of Flora & Fauna

Moumita Ghosh, Anirban Roy, and Kartick Chandra Mondal

Abstract Estuaries represent the transitional ecosystem between freshwater and marine environment. Being dominated by both kinds of aquatic realms, it offers one of the most diverse ecosystems. However, Indian estuaries need a more exhaustive survey for the proper management of the wetlands as the estuarine ecological niche of flora and fauna is at risk. Mainly anthropogenic movements including trading, industrial as well as recreational activities, are the underlying reasons behind the deteriorating estuarine ecosystem and biodiversity. Comprehending the importance of the estuarine ecosystem, this article is concentrating on knowledge discovery from Indian estuarine data of flora & fauna. Here, we show the efficient use of the combining approach for bi-clustering and association rule mining on a manually curated real dataset. We come up with a set of rules, presentable to the ecologists as it can summarize closely occurred member lists, predicted list of sites for member expansion, etc. Hence, our study would assist in reinforcing the estuarine diversity that could pioneer region-based further studies.

Keywords Indian estuaries · Data mining · Knowledge discovery

1 Introduction

Context and the motivation:

Different physicochemical parameters, for example, temperature, pH, salinity, electrical conductivity, total dissolved solids, total suspended solids, turbidity, etc., along with various biological parameters are the main influencing factors for the unique supporting ecosystem in estuaries. Though estuaries are dynamic systems favoring the proliferation of diverse biota, estuary health is losing its dignity [2, 8]. Pollution,

M. Ghosh · K. C. Mondal (✉)
Department of Information Technology, Jadavpur University, Kolkata, India

A. Roy
West Bengal Biodiversity Board, Governmet of West Bengal, Kolkata, India

M. Saraswat et al. (eds.), *Proceedings of International Conference on Data Science and Applications*, Lecture Notes in Networks and Systems 287,
https://doi.org/10.1007/978-981-16-5348-3_31

overfishing, and nutrient run-off are the effect of different human activities, which are the major identified reasons for this perpetuating loss of the estuarine ecosystem.

To prevent the depletion of this self-sustaining habitat, global concern, as well as the implementation of policies and mechanisms, is obvious. Focused study for exploring useful data and revealing knowledge can help in generating eco-awareness, thus preserving biodiversity at estuaries. A study has shown that National Centers For Coastal Ocean Science (NCCOS) has completed a research project [12] aiming at developing a database, namely Estuarine Living Marine Resources (ELMR) program database. This publicly available repository contains information of five regions in total: West Coast, Gulf of Mexico, Southeast, Mid-Atlantic, and North Atlantic. For each class, the database includes habitat type (tidal/freshwater/mixing zone), class presence (monthly distribution), and relative abundance (e.g., not present to highly abundant). The dataset is updated regularly and is approachable for an analytical job.

Leading by such kind of practice, here, we would like to investigate Indian estuarine biodiversity database containing the information related to diversity in presence along with the frequency of presence data. Government of India Web site for The Environment and Information System (ENVIS) Center on Wildlife and Protected Areas has summarized multiple resources for Indian flora, fauna, and data regarding multiple estuaries at the State level. We feel the necessity for gathering and summing up data in compliance with the research level. Currently, no such study describing Indian estuary database in terms of flora–fauna class diversity is available, to the best of our knowledge.

This drives us to introduce an estuarine floral–faunal diversity database. We have collected the data from a book published by the Zoological Survey of India [3]. This book contains information related to integrated flora and fauna diversity, along with 20 major estuaries of India. We explore the database by exploiting data mining methodology. This kind of knowledge exploration would help the ecologists in taking suitable measures for estuarine ecosystem preservation.

Contribution:

Recapitulating our studies, the main contributions are as follows:

- Introduce a real dataset of presence/absence status for the flora and fauna of Indian estuary.
- Showing up the detailed methodology on data preprocessing for making the dataset usable for employing data mining methodology.
- Discussion on the domain-specific significance of frequent closed itemsets and rule mining in investigating ecological data.

Organization of the paper:

The whole paper is divided into multiple sections. Section 2 describes the dataset and its prepossessing, mainly discretization for further employment. Next, Sect. 3 highlights a few existing methodologies, the motivation behind our approach, and briefly explains our study. The information that could be gathered is shown in Sect. 4. The conclusion is drawn in Sect. 5.

2 Preparation of the Dataset

India has a long coastal area along the east and west. Our database contains data on 20 major estuaries from both coastal regions. Following 15 estuaries are from the east coast, Hooghly-Matla, Subarnarekha, Baitarani-Brahmani, Mahanadi, Rushikulya, Bahuda, Vamsadhara, Nagavali, Godavari, Krishna, Penner, Ennore, Adyar, Veller, and Cauveri (i.e., E1 to E15 in Table 1). Cochin, Zuari, Mandovi, Tapi, and Narmada are situated at the west coast of India (i.e., E16 to E20 in Table 1). All of these estuaries are pointed on Indian Map on page 2, Fig. 1 of the book "Indian Estuarine Biodiversity" [3]. Twenty-three rows for faunal groups and three rows for floral groups are manually curated from the book, and it stores the estuary-wise number of classes found for each group of flora and fauna. This actual dataset is shown in Table 1. Our main aim, here, is to discover useful information from the estuarine dataset that could assist to preserve the diversity of the estuarine ecosystem. For this purpose, we would use data mining approach that demands a preprocessing step on this numerical dataset for a better understanding of the resultset.

Data preprocessing task consists of five elementary sub-tasks, namely data cleaning, integration, transformation, reduction, and discretization. Here, the database contains no missing, noisy, or inconsistent data. Also, we are not performing operations on multiple databases. Our dataset contains count data for different classes. Again, data volume is small enough, and no redundancy is occurring here. Thus, data cleaning, integrating, transformation, and reduction are not relevant in this case. We have performed only data discretization that allows us to categorize the column values into a set of discrete levels. Binning is one such discretization technique that is simple and widely used. It works in two ways, as mentioned below:

- Equal-width partitioning:
  Partition the range into N equal interval.
  According to the data values in the dataset, the highest and lowest data values are 377 and 1, respectively. For $N = 7$, the width is 54. A high value of N guarantees less information loss, and a low value of N gives low categorization of data that in turn makes result interpretation to difficult. 76% of the data fall in the category of very very low which takes the data range of 1–54. It is almost straightforward, but we can see that it has skewed the data toward the left (belong to very very low). Next, we have tried for equal-depth partitioning.

Table 1 Floral and faunal data at Indian estuaries

Taxonomic group	Major estuaries of India																			
	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16	E17	E18	E19	E20
Floral groups																				
Phytoplankton	45	40			162				61	68	12	53		41	39		67		53	47
Mangroves	23	9	26	10			8		16	13	12	3			8	11	15	15	3	
Other Flora			19				9		19	13		1			3		10	10		
Faunal groups																				
Protozoa					25	26					20	3		23					4	3
Foraminifera					5					47	11					73		14		
Porifera	1								2								2	2		
Cnidaria	24	12		11	20	5			13	3	10					34	3	3		
Ctenophora	1			1	2						1	1				1				
Rotofera	5								14		2	16		13						
Nematoda	2				11												20			
Acanthocephala									1											
Sipuncula	1	1																		
Mollusca	83	49	19	152	47		28	43	73	103	82	10		11	51	26	40	41	30	32
Annelida	91	37	11	34	19		13	4	70	45				24	48	47	70	3		
Arthopoda	377	53	88	45	159	99	24	17	88	118	125	56	58	35	55	167	72	21	25	60
Bryozoa	4																			
Brachiopoda	1	2								1										
Chaetognatha	4			3	6	2				1	2	3				4	6	6		3
Echinodermata	22	6	1						7	2	1									
Hemichordata	1																			
Urochordata				3	6	4					3					1				
Class Pisces	314	146	157	177	45	91	64	71	307	268	63	17	135	82	135	126	73	44	64	49
Class Amphibia	13	3	14							4		3				27				
Class Reptilia	57	5	45	6	2	1	1	1	1	10		1				4	7	7		
Class Aves	156	108	269	46	1			52	75	17						45	43	150	23	23
Class Mammalia	41	2	27		4				1	11		8			2	5				
Subtotal	1198	424	631	478	352	228	130	188	652	630	320	118	193	188	291	560	336	291	146	170
Total	1266	473	676	488	514	228	147	188	748	724	344	175	193	229	341	571	428	316	198	217

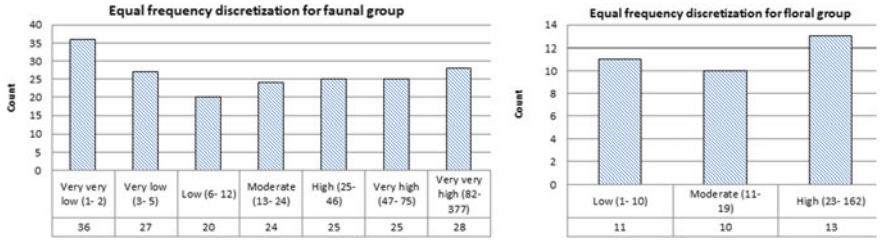

Fig. 1 Histogram distribution of flora and fauna for data discretization

- Equal-depth partitioning:
  - Partition the data range into N intervals containing approximately. Here, $N = 7$ is taken for the fauna, and $N = 3$ is for flora datasets. Dataset of flora contains 185 numbers of samples. If we divide it into seven intervals, as in the previous case, it can be justified in a much better way. Histogram distribution of the data samples for both flora and fauna is shown in Fig. 1.
  - It gives an almost homogeneous distribution of data at each bin.

We rename the data attribute values according to the above binning techniques as shown in Fig. 1. Discretized datasets of fauna and flora are shown below in Table 2 and Table 3, respectively. Corresponding to the raw dataset of fauna (Table 1), their presence-only data is shown in Table 4.

A statistical representation of the dataset can be depicted below in Fig. 4. Fauna presence data based on their frequency of class occurrence and a total number of estuaries having a similar kind of frequency of class are identified and shown in graphical format.

It is evident from Fig. 4 that the presence of Pisces and Arthropoda is moderate to very very high, and they cover all 20 estuaries. Pisces, Arthropoda, Aves, and Mollusca are also found with very very high frequency with 11, 8, 4, and 4 numbers of estuaries, respectively. Most of the classes are not reported from many of the estuaries. The tool will efficiently be able to extract information related to these without any manual intervention. Thus, the information regarding the chances of class occurrence at the estuaries where they have not found could establish a proper measure to build the diversity richer. Section 4 would deal with this issue.

3 Background Study and the Proposed Methodology

The advantageous usage of data mining in biodiversity data analysis is not unfamiliar to the research community. Previously, a few studies have identified the beneficial use of it in ecological data analysis [4]. Clustering, classification, rule mining, etc., are the major tasks that can be performed using data mining algorithms. Clustering identifies a similar group of data items from a huge number of data based on the predefined

Table 2 Discretized faunal data at Indian estuaries(E1 to E20): VVH→ very very high, VVL→ very very low, VH→ very high, VL→ very low, H→ high, L→ low, M→ moderate, and ? refers to the unavailability of the information

Taxonomic group	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16	E17	E18	E19	E20
Protozoa	?	?	?	?	H	H	?	?	?	?	M	VL	?	M	?	?	?	?	VL	VL
Foraminifera	?	?	?	?	VL	?	?	?	?	VH	L	?	?	?	?	VH	?	M	?	?
Porifera	VVL	?	?	?	?	?	?	?	VVL	?	?	?	?	?	?	?	VVL	VVL	?	?
Cnidaria	M	L	?	L	M	VL	?	?	M	VL	L	?	?	?	?	H	VL	VL	?	?
Ctenophora	VVL	?	?	VVL	VVL	?	?	?	?	?	VVL	VVL	?	?	?	VVL	?	?	?	?
Rotofera	VL	?	?	?	?	?	?	?	M	?	VVL	M	?	M	?	?	?	?	?	?
Nematoda	VVL	?	?	?	L	?	?	?	?	?	?	?	?	?	?	?	M	?	?	?
Acanthocephala	?	?	?	?	?	?	?	?	VVL	?	?	?	?	?	?	?	?	?	?	?
Sipuncula	VVL	VVL	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Mollusca	VVH	VH	M	VVH	VH	?	H	H	VH	VVH	VVH	L	?	L	VH	H	H	H	H	H
Annelida	VVH	H	L	H	M	?	M	VL	VH	H	?	?	?	M	VH	VH	VH	VL	?	?
Arthopoda	VVH	VH	VVH	H	VVH	VVH	M	M	VVH	VVH	VVH	VH	VH	H	VH	VVH	VH	M	H	VH
Bryozoa	VL	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Brachiopoda	VVL	VVL	?	?	?	?	?	?	?	VVL	?	?	?	?	?	?	?	?	?	?
Chaetognatha	VL	?	?	VL	L	VVL	?	?	?	VVL	VVL	VL	?	?	?	VL	L	L	?	VL
Echinodermata	M	L	VVL	?	?	?	?	?	L	VVL	VVL	?	?	?	?	?	?	?	?	?
Hemichordata	VVL	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Urochordata	?	?	?	VL	L	VL	?	?	?	?	VL	?	?	?	?	VVL	?	?	?	?
Class Pisces	VVH	VVH	VVH	VVH	H	VVH	VH	VH	VVH	VVH	VH	M	VVH	VVH	VVH	VVH	VH	H	VH	VH
Class Amphibia	M	VL	M	?	?	?	?	?	?	VL	?	VL	?	?	?	H	?	?	?	?
Class Reptilia	VH	VL	H	L	VVL	VVL	VVL	VVL	VVL	L	?	VVL	?	?	?	VL	L	L	?	?
Class Aves	VVH	VVH	VVH	H	VVL	?	?	VH	VH	M	?	?	?	?	?	H	H	VVH	M	M
Class Mammalia	H	VVL	H	?	VL	?	?	?	VVL	L	?	L	?	?	VVL	VL	?	?	?	?

Table 3 Discretized floral data of Indian estuaries: VVH→ very very high, VVL→ very very low, VH→ very high, VL→ very low, H→ high, L→ low, M→ moderate, and ? refers to the unavailability of the information

Taxonomic group	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16	E17	E18	E19	E20
Phytoplankton	H	H	?	?	VVH	?	?	?	VH	VH	L	VH	?	H	H	?	VH	?	VH	VH
Mangroves	M	L	H	L	?	?	L	?	M	M	L	VL	?	?	L	L	M	M	VL	?
Other Flora	?	?	M	?	?	?	L	?	M	M	?	VVL	?	?	VL	?	L	L	?	?

Table 4 Presence-only dataset corresponding to table 1 where 1 represents the presence, and ? refers to the unavailability of the information

Taxonomic group	Major estuaries of India																			
	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16	E17	E18	E19	E20
Floral groups																				
Phytoplankton	1	1	?	?	1	?	?	?	1	1	1	1	?	1	1	?	1	?	1	1
Mangroves	1	1	1	1	?	?	1	?	1	1	1	1	?	?	1	1	1	1	1	?
OtherFlora	?	?	1	?	?	?	1	?	1	1	?	1	?	?	1	?	1	1	?	?
Faunal groups																				
Protozoa	?	?	?	?	1	1	?	?	?	?	1	1	?	1	?	?	?	?	1	1
Foraminifera	?	?	?	?	1	?	?	?	?	1	1	?	?	?	?	1	?	1	?	?
Porifera	1	?	?	?	?	?	?	?	1	?	?	?	?	?	?	?	1	1	?	?
Cnidaria	1	1	?	1	1	1	?	?	1	1	1	?	?	?	?	1	1	1	?	?
Ctenophora	1	?	?	1	1	?	?	?	?	?	1	1	?	?	?	1	?	?	?	?
Rotofera	1	?	?	?	?	?	?	?	1	?	1	1	?	1	?	?	?	?	?	?
Nematoda	1	?	?	?	1	?	?	?	?	?	?	?	?	?	?	?	1	?	?	?
Acanthocephala	?	?	?	?	?	?	?	?	1	?	?	?	?	?	?	?	?	?	?	?
Sipuncula	1	1	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Mollusca	1	1	1	1	1	?	1	1	1	1	1	1	?	1	1	1	1	1	1	1
Annelida	1	1	1	1	1	?	1	1	1	1	?	?	?	1	1	1	1	1	?	?
Arthopoda	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Bryozoa	1	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Brachiopoda	1	1	?	?	?	?	?	?	?	1	?	?	?	?	?	?	?	?	?	?
Chaetognatha	1	?	?	1	1	1	?	?	?	1	1	1	?	?	?	1	1	1	?	1
Echinodermata	1	1	1	?	?	?	?	?	1	1	1	?	?	?	?	?	?	?	?	?
Hemichordata	1	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Urochordata	?	?	?	1	1	1	?	?	?	?	1	?	?	?	?	1	?	?	?	?
Class Pisces	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Class Amphibia	1	1	1	?	?	?	?	?	?	1	?	1	?	?	?	1	?	?	?	?
Class Reptilia	1	1	1	1	1	1	1	1	1	1	?	1	?	?	?	1	1	1	?	?
Class Aves	1	1	1	1	1	?	?	1	1	1	?	?	?	?	?	1	1	1	1	1
Class Mammalia	1	1	1	?	1	?	?	?	1	1	?	1	?	?	1	1	?	?	?	?

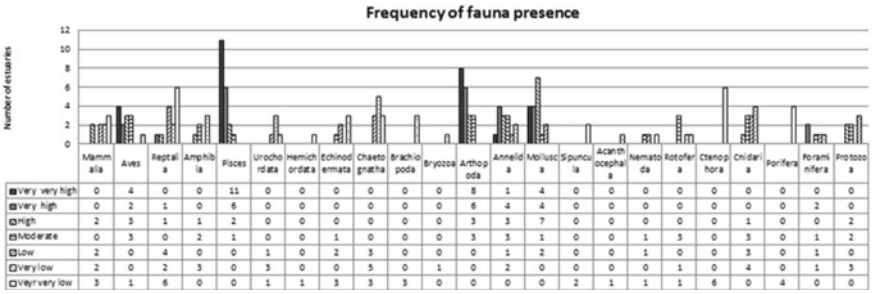

Fig. 2 Frequency of fauna presence at Indian estuaries

similarity threshold. It can be in one or more dimensions. Two-dimensional clustering is mainly getting popularity in bioinformatics for gene data analysis. Exceptionally, [1] has used it in assessing migratory bird population data. Classification is another major task of data mining, and it has used for decades for species classification in multiple research works [9, 13, 15]. Classification of huge data where data is available in image, video, or audio format rather than simple text is a challenging task, and deep learning-based approaches [6, 7] are proven to be useful here. Rule mining for future prediction is followed by few authors where biodiversity domain is taken into account [5, 14]. Relation among data items and their dependency can be found using rule-based approaches.

Background study directs that association rule mining and bi-clustering are the two promising approaches for analyzing ecological data where the dataset can be clustered down first in both row and column, then based on frequently occurred data, useful rules can be mined. We would like to use a tool proposed in [10, 11] where the computational approach of both the above-mentioned task is found in one algorithm. The whole approach has appeared in Fig. 2.

The algorithm [11] extracts the itemsets that are occurring frequently in the dataset. These are identified as closed frequent itemsets. Then, it generates the underlying facts in the form of antecedent and consequent. These facts are known as association rules. Along with the generated rules, it shows the object lists satisfying the facts which will make the algorithm more reliable. In addition to this, the data structure (suffix-tree) used behind the tool makes it better both in terms of time and space complexity [10].

4 Result and Discussion

We have followed the operations as shown in Fig. 2. The information retrieval task is performed on the discretized datasets shown in Tables 2 and 3 and presence-only dataset as shown in Table 4.

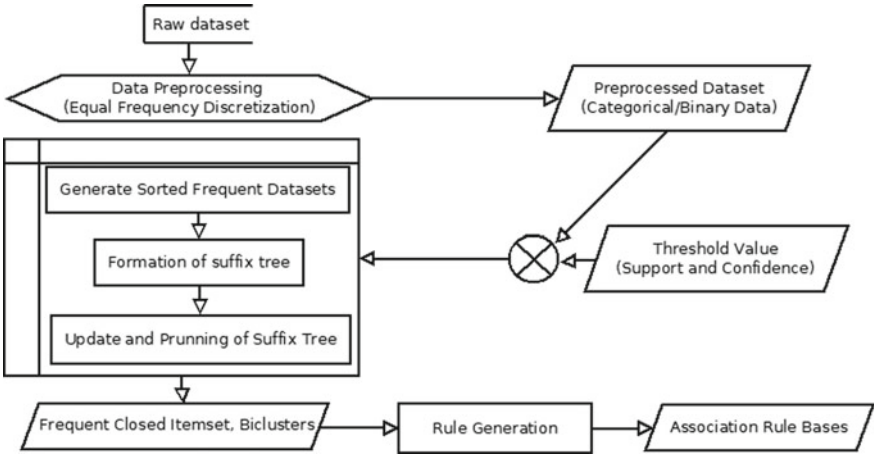

Fig. 3 Block diagram of the used approach

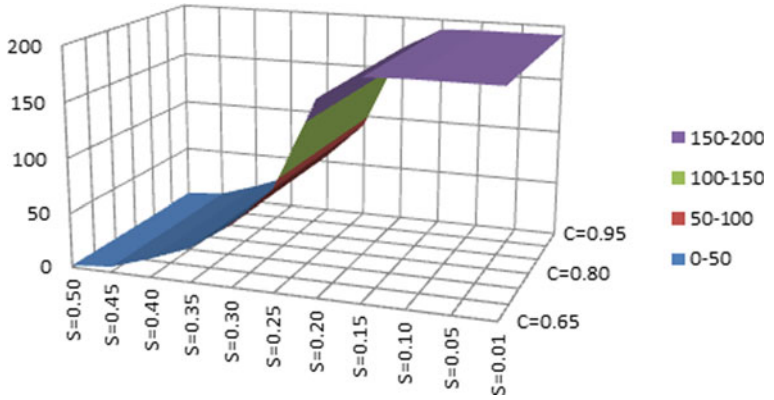

Fig. 4 Number of rules generated by applying tool: *x* axis: min-support; *y* axis: min confidence; *z* axis: rule count

A huge set of rules can be generated by lowering the constraints (support and confidence). By changing the values of minimum support and minimum confidence (step size 5%), we have listed down the statistics of rules generated with the presence-only dataset of fauna for all combinations of min-support and min-confidence value and shown the variation in the number of rules generated in Fig. 3. It can be seen that the number of rules is larger for the lower support values as the constraints are relaxing (Fig. 4).

Below, we are going to highlight our observations from the generated set of rules and show the way how the tool generated resultant file can easily be interpreted and can be sorted and searched according to the need of the user.

4.1 Result Obtained from the discretized dataset of fauna(Table 2)

- Consider the example of class Pisces. Extracted rows from the result file (Table 5, the rule set 1) show that the frequency of occurrence of species count for this class is very high to very very high. Class Pisces belongs to one of the most diverse classes. Mollusca, Arthropoda, and Aves form a closed group along with Pisces for class frequency level of very very high (shown in Table 5 of the rule set 2.)
- This states that co-occurrence among these classes is quite familiar, suggesting the probability of more availability at not-found estuaries. Like, both Mollusca and Pisces are found in very very high frequency at estuary E1, E4, and E10. As Pisces is very very high at 11 estuaries in total (E1, E2, E3, E4, E6, E9, E10, E13, E14, E15, E16), remaining eight estuaries are probable sites for Mollusca with very very high frequency.

- Table 6 says, class numbered C10 (Mollusca), C11 (Annelida), and C22 (Class Aves), in 60% cases, can be found with very high frequency at estuary 9 when at estuary 1 they occur with very very high frequency. Hence, it can be inferred that for the remaining 40% cases, estuary 9 may have a very high frequency of occurrence for all the classes having very very high frequency at estuary 1. We obtain [E1 = VVH] for classes 10, 11, 12, 19, 22 from frequent closed itemsets. So, it can be concluded that C12 and C19 have the probability of occurring with very high frequency at E9 and confidence of this inference is 60%.
- Again, as the antecedent for all the rules is the same, $R1 - R2 = 11, 22$ indicates that C11 (Annelida) and C22 (Class Aves) should occur at estuary 10 with 60% confidence. Similarly, $R1 - R3 = 22$ indicates that C22 (Class Aves) should occur at estuary 15 with very high frequency.

Table 5 Sample rule set 1 (upper) and rule set 2 (lower) for the class Pisces extracted from discretized fauna dataset

Closed set	Support	Estuary list
[Class Pisces = VVH]	11	1, 2, 3, 4, 6, 9, 10, 13, 14, 15, 16
[Class Pisces = VH]	6	7, 8, 11, 17, 19, 20
[Arthopoda = VVH, Class Pisces = VVH]	6	1, 3, 6, 9, 10, 16
[Mollusca = VVH, Class Pisces = VVH]	3	1, 4, 10
[Mollusca = VVH, Arthopoda = VVH]	3	1, 10, 11
[Class Aves = VVH, Class Pisces = VVH]	3	1, 2, 3

Table 6 Generated rules having same antecedent extracted from discretized fauna dataset

Rule	Antecedent	Consequent	Support	Confidence	Class list
R1	[E1 = VVH]	[E9 = VH]	3	0.6	[10, 11, 22]
R2	[E1 = VVH]	[E10 = VVH]	3	0.6	[10, 12, 19]
R3	[E1 = VVH]	[E15 = VH]	3	0.6	[10, 11, 12]
R4	[E1 = VVH]	[E4 = H]	3	0.6	[11, 12, 22]
R5	[E1 = VVH]	[E17 = VH]	3	0.6	[11, 12, 19]
R6	[E1 =VVH]	[E3 = VVH]	3	0.6	[12, 19, 22]

4.2 Result obtained from the presence-only dataset of fauna (Table 4)

- We get the following frequent closed set without thoroughly investigating the dataset which specifies that Arthopoda and Pisces are found in all the estuaries $[Arthopoda = 1, Class Pisces = 1]$ Support = 20 Estuary list = 1, 2, 3, 4, 5, 6, 7,8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20

- Mollusca is the third highest found class after Arthopoda and Pisces. The following frequent closed itemset highlights that Mollusca is found at 18 different sites listed in the second position.
  $[Mollusca = 1, Arthopoda = 1, Classpisces = 1]$ Support = 18 Estuary list = 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 14, 15, 16, 17, 18, 19, 20

- Hooghly-Matla is the most diversified estuary. Followed by, Rushikulya, Krishna, and Cochin. 19 different types of faunal classes are found in Hooghly-Matla estuary as shown in Table 7.
- From the generated rule set, we obtain the rule shown in Table 8. It can be interpreted that a class will occur at E16 with 83.33% probability when that class has found at both the estuaries E10 and E1. Alternatively, it can be explained as 83.33% of the classes found at E10 and E1, are also found at E16. For remaining classes, they have the probability of occurrence with 83.33% confidence.
- Future probable habitat for a class can be identified from the closed set; alternatively, a future probable class for an estuary can be identified from the closed set.

  - Here, Hooghly-Matla, Subarnarekha, and Krishna estuary are forming closed set as 11 types of classes shown in Table 9 are common for all of them. $[E2 = 1, E10 = 1, E1 = 1]$ Support= 11 Class= 4, 10, 11, 12, 14, 16, 19, 20, 21, 22, 23

It could be concluded that C9 and C15 class has the probability of occurrence at estuary E10, E2, respectively.

Table 7 Top four diversity-rich estuaries extracted from presence-only fauna dataset

Closed item	Support	Classes list
[E1=1]	19	3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 21, 22, 23
[E5=1]	14	1, 2, 4, 5, 7, 10, 11, 12, 15, 18, 19, 21, 22, 23
[E10=1]	13	2, 4, 10, 11, 12, 14, 15, 16, 19, 20, 21, 22, 23
[E16=1]	13	2, 4, 5, 10, 11, 12, 15, 18, 19, 20, 21, 22, 23

Table 8 Rule generated from the presence-only dataset of fauna

Antecedent	Consequent	Support	Confidence	Class list
E10=1, E1=1	E16=1	10	0.8333333	4, 10, 11, 12, 15, 19, 20, 21, 22, 23

- A few closed sets are summarized below in Table 10. Using transition law on them, we can say that E1, E2, E5, E10, E16 are forming a closed group. Thus, these estuaries should have homogeneous presence data for classes.
  Presence data of these estuaries is as follows in Table 11 where C4, C10, C11, C12, C19, C21, C22, and C23 have a presence in all five estuaries. For other classes, probable estuarine habitat can be predicted.

  - Rules generated from the fauna dataset in Table 12 have a similar object list that consists of C6, C10, C12, and C19. Also, E14, E12, E11, E9, and E1 are forming a closed group, and the occurrence of any class is associated with all others belonging to the same group with a minimum of 67%,

4.3 Result obtained from the discretized dataset of flora (Table 3)

- Table 13 shows that E7, E9, and E10 are forming the closed group based on mangroves and other flora.

- Mangroves with medium dense are associated with phytoplankton with very high density and are found at three estuaries E9, E10, and E17 (observed from frequent closed itemset). However, Table 14 extracts a rule showing medium dense mangrove and very high dense phytoplankton accompanying other flora with medium dense with 66.6% confidence at estuary E9 and E10. This directs that E17 may be expected to have other flora with medium or higher density.

Table 9 Presence data for all fauna classes (C1 to C23) at estuaries E1, E2, and E10 where ? refers to the unavailability of the information

Classes	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23
E1	?	?	1	1	1	1	1	?	1	1	1	1	1	1	1	1	1	?	1	1	1	1	1
E2	?	?	?	1	?	?	?	?	1	1	1	1	?	1	?	1	?	?	1	1	**1**	1	1
E10	?	1	?	1	?	?	?	?	?	1	1	1	?	**1**	1	**1**	?	?	**1**	**1**	1	1	1

Table 10 Closed itemsets extracted from the presence dataset of fauna

Closed item	Support	Estuary list
[E2=1, E1=1]	12	4, 9, 10, 11, 12, 14, 16, 19, 20, 21, 22, 23
[E10=1, E1=1]	12	4, 10, 11, 12, 14, 15, 16, 19, 20, 21, 22, 23
[E16=1, E5=1]	12	2, 4, 5, 10, 11, 12, 15, 18, 19, 21, 22, 23
[E16=1, E1=1]	11	4, 5, 10, 11, 12, 15, 19, 20, 21, 22, 23
[E5=1, E1=1]	11	4, 5, 7, 10, 11, 12, 15, 19, 21, 22, 23

4.4 Result obtained from the presence-only dataset of flora (Table 4)

Table 15 highlights a rule with a support value of 9 and a confidence value of 0.75. Table 16 says that phytoplankton appears in 12 estuaries which are E1, E2, E5, E9, E10, E11, E12, E14, E15, E17, E19, and E20 (generated from frequent closed itemset file) out of which consequent (for the rule in Table 15) has occurred in nine estuaries (E1, E2, E9, E10, E11, E12, E15, E17, and E19). Therefore, it can be thought of as reasonably high to predict that mangrove class may also be present in estuaries E5, E14, and E20 with 75% confidence.

5 Conclusions

The aim of this paper is to assist the ecologists in taking suitable steps for ecosystem conservation via an algorithmic approach. We gather Indian estuarine data and illustrate the way of knowledge extraction using data mining-oriented methodology. Our followed approach internally utilizes suffix-trees. This data structure enables the efficient storage of data and computation of relevant patterns. It requires only a unique scan of the dataset to extract all valid patterns. It can discover the minute details for each class regarding its level of appearances, co-occurrences, and for each estuary regarding its homogeneity in diversity. It is also able to find out the future probable occurrence of a particular class for a particular estuary. Undoubtedly, this kind of information could help the ecologists in managing estuarine biodiversity by establishing policy and suitable measures.

In the future, we want to prepare a digitized and downloadable estuarine floral–faunal diversity database for accelerating the research design and implementation. Also, we would like to derive useful knowledge by employing the addressed methodology in the study of dark diversity that is the study on the species that can potentially colonize in a particular ecological condition with the local community but are absent. Another useful task could be proposing a methodology for uncertain frequent itemset mining on species occurrence dataset.

Table 11 Presence data for all fauna class (C1 to C23) at estuaries E1, E2, E5, E10, and E16, where ? refers to the unavailability of the information

Classes	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23
E1	?	?	1	1	1	1	1	?	1	1	1	1	1	1	1	1	1	?	1	1	1	1	1
E2	?	?	?	1	?	?	?	?	1	1	1	1	?	1	?	1	?	?	1	1	1	1	1
E5	1	1	?	1	1	?	1	?	?	1	1	1	?	?	1	?	?	1	1	?	1	1	1
E10	?	1	?	1	?	?	?	?	?	1	1	1	?	1	1	1	?	?	1	1	1	1	1
E16	?	1	?	1	1	?	?	?	?	1	1	1	?	?	1	?	?	1	1	1	1	1	1

Table 12 Rules generated from presence-only fauna dataset

Anticedent	Consequent	Support	Confidence	Class list
E14=1	E12=1, E11=1, E9=1, E1=1	4	0.6666667	6, 10, 12, 19
E14=1, E12=1, E11=1	E9=1, E1=1	4	0.8	6, 10, 12, 19
E14=1, E9=1, E1=1	E12=1, E11=1	4	0.8	6, 10, 12, 19
E12=1, E11=1, E1=1	E14=1, E9=1	4	0.6666667	6, 10, 12, 19
E12=1, E9=1, E1=1	E14=1, E11=1	4	0.6666667	6, 10, 12, 19
E11=1, E9=1, E1=1	E14=1, E12=1	4	0.6666667	6, 10, 12, 19

Table 13 Rules generated from the discretized flora dataset

Antecedent	Consequent	Classes list
[E7= Low]	[E9= Moderate, E10= Moderate]	2, 3
[E9= Moderate]	[E7= Low, E10= Moderate]	2, 3
[E10= Moderate]	[E7= Low, E9= Moderate]	2, 3

Table 14 Rules generated from the flora dataset

Antecedent	Consequent	Support	Confidence	Estuary list
[Mangroves=M, Phytoplankton=VH]	[Other Flora=M]	2	0.6666667	[9, 10]

Table 15 Rules generated from the presence-only dataset of flora

Antecedent	Consequent	Support	Confidence	Estuary list
Phytoplankton=1	Mangroves=1	9	0.75	[1, 2, 9, 10, 11, 12, 15, 17, 19]

Table 16 Closed itemsets extracted from the presence-only dataset of flora

Closed item	Support	Estuary list
Phytoplankton=1	12	1, 2, 5, 9, 10, 11, 12, 14, 15, 17, 19, 20

Acknowledgements The authors are grateful to the Department of Science & Technology, Government of India, New Delhi, for financial assistance under the scheme of WOS-A (File no SR/WOS-A/ET-112/2017(G)) to carry out this Ph.D. research project.

References

1. Adams, E.M.: Using migration monitoring data to assess bird population status and behavior in a changing environment. Electron Theses Dissertations **2238** (2014)
2. Alfred, J.R.B., Das, A.K , Sanyal, A.K.: Ecosystems of India 1–410 . ENVl5-Zool .Surv. India, KoIkata (2001)
3. Chandra, K., Raghunathan, C., Dash, S.: Current status of estuarine biodiversity in India: 1–575. the Director, Zool. Surv. India, Kolkata (2018)
4. Dzeroski, S.: Environmental Applications of Data Mining. University of Trento, Lecture Notes of Knowledge Technologies (2003)
5. Gougeon, F.A., et al.: Automatic individual tree crown delineation using a valley-following algorithm and rule-based system. In: Proceedings of International Forum on Automated Interpretation of High Spatial Resolution Digital Imagery for Forestry, Victoria, British Columbia, Canada, pp. 11–23 (1998)
6. Guirado, E., Tabik, S., Alcaraz-Segura, D., Cabello, J., Herrera, F.: Deep-learning convolutional neural networks for scattered shrub detection with google earth imagery. arXiv preprint arXiv:1706.00917 (2017)
7. Heredia, I.: Large-scale plant classification with deep neural networks. In: Proceedings of the Computing Frontiers Conference, pp. 259–262 . ACM (2017)
8. Jha, B.C., Nath, D., Srivastava, N.P., Satpathy, B.B.: Estuarine fisheries management options and strategies. CIFRI Policy Pap. **3**, 1–23 (2008)
9. Liu, K., Li, X., Shi, X., Wang, S.: Monitoring mangrove forest changes using remote sensing and GIS data with decision-tree learning. Wetlands **28**(2), 336–346 (2008)
10. Mondal, K.C.: Algorithms for Data Mining and Bio-informatics. Ph.D. Thesis, Nice Sophia-Antipolis (2013)
11. Mondal, K.C., Pasquier, N., Mukhopadhyay, A., Maulik, U., Bandhopadyay, S.: A new approach for association rule mining and bi-clustering using formal concept analysis. In: International Workshop on Machine Learning and Data Mining in Pattern Recognition, pp. 86–101. Springer, Berlin (2012)
12. Nelson, D.M., Monaco, M.E.: National overview and evolution of NOAA's estuarine living marine resources (ELMR) program (2000)
13. Sesnie, S.E., Finegan, B., Gessler, P.E., Thessler, S., Ramos Bendana, Smith, A.M. : The multispectral separability of costa rican rainforest types with support vector machines and random forest decision trees. Int. J .Remote Sens. **31**(11):2885–2909 (2010)
14. Silva, M., Trevisan, D.Q., Prata, D.N., Marques, E.E., Lisboa, M., Prata, M.: Exploring an ichthyoplankton database from a freshwater reservoir in legal amazon. In: International Conference on Advanced Data Mining and Applications, pp. 384–395. Springer, Berlin (2013)
15. Sitanggang, I.S., Yaakob, R., Mustapha, N., AN, A.: Application of classification algorithms in data mining for hotspots occurrence prediction in Riau province, Indonesia. J. Theor. Appl. Inf. Technol. **43**(2) (2012)