

Rough Set Theory: Approach for Similarity Measure in Cluster Analysis

Shuchita Upadhyaya

Deptt. of Computer Science and Application, Kurukshetra University, Kurukshetra, INDIA

Alka Arora

Indian Agricultural Statistics Research Institute, New Delhi-110012, INDIA

Rajni Jain

National Center for Agricultural Economics and Policy Research, New Delhi-110012, INDIA

Abstract - Clustering of data is an important data mining application. One of the problems with traditional partitioning clustering methods is that they partition the data into hard bound number of clusters. Rough set based Indiscernibility relation combined with indiscernibility graph, leads to knowledge discovery in an elegant way. Indiscernibility relation has a strong appeal to be applied in clustering as it creates natural clusters in data. Indiscernibility relation is used for measuring the similarity among the data items based on which clustering is performed. In the proposed approach the strict notion of indiscernibility is relaxed and classes are formed on the basis that objects are similar rather than identical. Indiscernibility relation creates indiscernible classes and representation of these classes with indiscernibility graph aids in better representation of clusters.

Keywords: Clustering, Rough Set, Similarity Measure, Indiscernibility Graph, Indiscernibility.

1. Introduction

A cluster is a collection of data objects that are similar to one another within the same class/group and are dissimilar to the objects in other class/group. The process of grouping the set of objects into classes of similar objects is called clustering. Clustering of data is important component of Data Mining. As a data mining task, data clustering identifies clusters, or densely populated regions, according to some distance measure [2]. Typical steps involved in the clustering procedure are Feature Selection/Extraction from data (this is optional), Inter pattern proximity/similarity measure appropriate to data domain and Clustering or grouping of data [3]. Since similarity is fundamental to the definition of a cluster, a measure of the similarity between two patterns drawn from the same feature space is essential to most clustering procedures. Pattern proximity is usually measured by a distance function defined on pairs of objects. A variety of distance measures are available depending upon the data type [3] and [6]. In this paper, we explore the feasibility of Rough Set Theory (RST) for similarity measure. Our motivation for this work comes from the notion of indiscernibility in RST that we consider very attractive, since each indiscernible relation is also a sort of cluster. In the proposed approach indiscernibility is used as a measure of similarity without any distance function for clustering the objects. In addition very less work has been done in applying RST to clustering.

The Paper is organized as follows. Section 2 discusses some important concepts of rough set theory. Section 3 focuses on approach used in this study. Section 4 presents application of proposed approach followed by conclusion and future work.

2. Rough Set Concepts

Rough Sets Theory (RST) is a mathematical approach, proposed by Z. Pawlak in the early eighties and since has come into focus as an alternative to the more widely used method of machine learning and statistical data analysis [4], [5] and [7]. Data is represented in the Rough Set (RS) framework in the form of information system or

table. Each row of the table represents an object and every column represents an attribute that can be measured for each object. Rough set theory is derived from the set theory therefore usual assumptions of traditional quantitative research techniques do not apply.

Formally an information system is a pair $A = (U, A)$ where U is a non-empty, finite set of objects called the universe and A is a non-empty, finite set of attributes on U . With every attribute $a \in A$ we associate a set V_a such that $a: U \rightarrow V_a$. The set V_a is called the domain or value set of attribute a . Indiscernibility is core concept of RST and is defined as equivalence between objects. Objects in the information system about which we have the same knowledge form an equivalence relation. (The notion of equivalence is recalled first, a binary relation $R \subseteq X * X$ which is reflexive (i.e. an object is in relation with itself xRx), symmetric (if xRy then yRx) and transitive (if xRy and yRz then xRz) is called an equivalence relation.)

Formally any set $B \subseteq A$ there is associated an equivalence relation called B-Indiscernibility relation defined as follows:

$$IND_A(B) = \{(x, x') \in U^2 \mid \forall a \in B a(x) = a(x')\}$$

If $(x, x') \in IND_A(B)$, then objects x and x' are indiscernible from each other by attributes from B . Equivalence relations lead to the universe being divided into equivalence class partition and union of these sets make the universal set. The indiscernibility class of an object $x \in U$ is defined as $R_A(x)$, and consist of those objects that stand in relation to object x by R_A .

$$R_A(x) = \{x' \in U \mid x R_A x'\}.$$

An alternative way to represent R_A is an Indiscernibility Definition Graph (IDG). An IDG for an attribute a is a graph with the elements of V_a as nodes or vertices, and a set of edges $E_a \subseteq V_a^2$ such that:

$$IDG_a = (V_a, E_a).$$

An edge $(v_1, v_2) \in E_a$ is to be interpreted as that an object with value v_1 is indiscernible with object with value v_2 . In IDG, the spatial layout of each vertex is really irrelevant; this is no. of edges between two vertex that determines the distance between them [8] and [9].

3. Proposed Approach

In the proposed approach, we have used indiscernibility relation of rough set theory for similarity measure. Similarity between any two objects to be clustered is used in the decision on whether to put them into same cluster or disjoint cluster. In RST objects are considered indiscernible if they share the same values for all the attributes. In the proposed approach this strict requirement of indiscernibility defined in canonical rough set theory is relaxed. Clustering is done on the basis that, objects are similar rather than identical. Means even if the attributes exist that discern between two objects, we may still deem the objects indiscernible if the number of such attributes is low. To illustrate consider the data set with n attributes. Objects are considered indiscernible if they share the same values for $n-r$ attributes where $r=1,2,3\dots$ and so on depending upon domain knowledge and size of the dataset. Experiment need to be carried out for different values of r for better quality of clusters.

Representation of indiscernible classes obtained through this approach by IDG aids in visualization of the clusters formed. Combining these approaches together provides better representation of data clusters than partitioning clustering approach that divides the data into hard bound number of clusters.

4. Example of Application: The Animal Taxonomy Problem

In this section proposed approach has been demonstrated on animal taxonomy data sets. For brevity small dataset has been taken up for demonstration as shown in Table 1. This table contains information on eight animals that we know, a priori, to belong to the classes of amphibians and mammals. The classes, however, are not informed. It is expected that some near-real grouping come up from the process [1].

To illustrate, Universal set (U) contains 8 objects and the attribute set A consist of 6 attributes (Name, Metabolism, Cover, Dentition, Reproduction and No. of feet) as shown in table 1. Attribute Name is unique and is used to identify the objects hence not used for clustering objects. Remaining 5 attributes are used to describe these objects.

Table 1: Animal Taxonomy Example

| Name | Metabolism | Cover | Dentition | Reproduction | No. of feet |
|----------|-------------|---------|--------------|--------------|-------------|
| Frog | ectothermic | wetskin | superior | oviparous | 4 |
| toad | ectothermic | wetskin | no | oviparous | 4 |
| elephant | endothermic | hair | complete | viviparous | 4 |
| dog | endothermic | hair | complete | viviparous | 4 |
| cat | endothermic | hair | complete | viviparous | 4 |
| rabbit | endothermic | hair | complete | viviparous | 4 |
| jaguar | endothermic | hair | complete | viviparous | 4 |
| whale | endothermic | hair | the youngest | viviparous | 0 |

Taking strict notion of indiscernibility and value of $r = 0$ (zero), only objects elephant, dog, cat, rabbit and jaguar form an equivalence class as they share same values for all the attributes (Metabolism, Cover, Dentition, Reproduction and No. of feet). When the condition of indiscernibility is relaxed and value of $r=1$ (one) is taken, objects frog and toad too becomes indiscernible and form another equivalence class as shown in Fig 1.

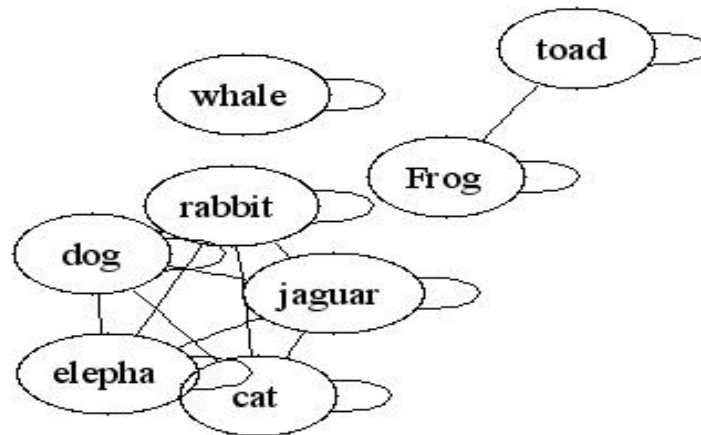


Fig 1: Indiscernibility based Clustering of objects with $r=1$

In order to refine the cluster further experiment was carried out with value of $r = 2$ then whale becomes indiscernible with objects elephant, dog, cat, rabbit, jaguar and form one group of mammals and frog & toad in another group of amphibians as shown in Fig 2. A highly significant result has been obtained with the proposed approach and data was clustered into two different groups of amphibians and mammals representing the true picture of data.

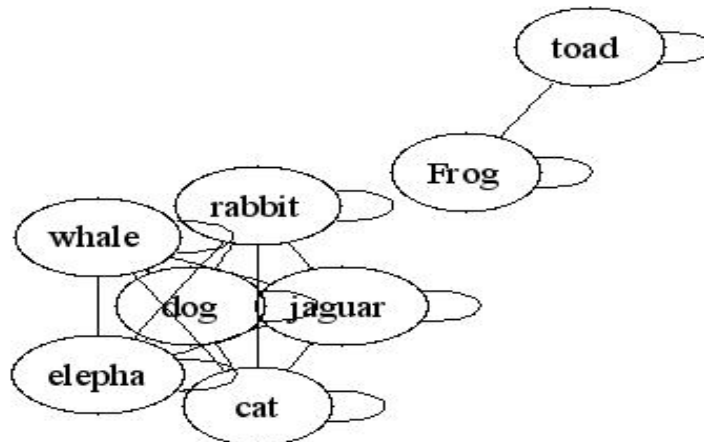


Fig 2: Indiscernibility based Clustering of objects with $r=2$

This approach was carried out for better understanding on Zoo data set from the Machine learning data repository at UC Irvine [10]. Zoo dataset consist of 101 instances of animals divided into seven groups and 17 attributes describing those 101 objects. Applying indiscernibility based rough set approach with value of $r = 2$ gave significant results and data was grouped into five distinct clusters and two mixed clusters giving real grouping of the data.

5. Conclusion and Future Work

In this study, we presented results on small and medium size dataset on exploring the indiscernibility based approach of RS in clustering. Encouraging results are obtained for these two data sets. In future more research is required to apply the same approach for large data set and on data from the real life situation. This study has been carried out on data sets with binary values therefore more work need to be carried out on quantitative data as that is important part of any dataset.

6. References

- [1] Do Prado, H.A., Engel, P.M. & Filho, H.C. 2002, 'Rough clustering: An alternative to find meaningful cluster by using the reducts from a dataset', in *Rough Sets and Current Trends in Computing, Third International Conference, RSCTC 2002. Lecture Notes in Computer Science, Vol. 2475*, eds. J.J. Alpigini, J.F. Peters, A.Skowron & N.Zhong, Springer Verlag, Berlin, pp. 234-238.
- [2] Han, J. and Kamber, M. *Data Mining Concepts and Techniques*, Morgan Kaufmann.
- [3] Jain, A.K, Murty, M.N., and Flynn, P.J. Data Clustering : A review, *ACM Computing Surveys*, 31 ,3, 264-323.
- [4] Komorowski, J., Pawlak, Z., Polkowski, Skowron, A. Rough sets: A tutorial. In: S. K. Pal, A. Skowron (Ed.). *Rough Fuzzy Hybridization: A new Trend in Decision-Making*. Berlin: Springer-Verlag, 1999, 3–98 [KPP 99].
- [5] Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*, Boston, MA, Kluwer Academic Publishers, 1991.
- [6] Pavel Berkhin, Survey of Clustering Data Mining Techniques from Internet.
- [7] Yao, Y.Y., Wong, S. K. M. and Lin, T.Y. A Review of Rough Set Models in: LYN, T.Y; Cercone. N. (Eds.). *Rough Sets and Data Mining: Analysis of Imprecise Data*, Kluwer Academic Pub. 1997.
- [8] The GraphViz homepage. [<http://www.research.att.com/sw/tools/graphviz/>].
- [9] The Rossata homepage. [<http://www.rossata.com>].
- [10] <http://www.ics.uci.edu/mllearn/MLRepository.html>.