International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume: 3 | Issue: 4 | May-Jun 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 - 6470

Approaching Rules Induction: CN2 Algorithm in Categorizing of Biodiversity

Su Myo Swe[1], Khin Myo Sett[2]

[1]Lecturer, [2]Professor
[1]Department of Computer Studies, Dagon University, Yangon Region, Myanmar
[2]Department of Computer Studies, University of Mandalay, Mandalay Region, Myanmar

How to cite this paper: Su Myo Swe | Khin Myo Sett "Approaching Rules Induction: CN2 Algorithm in Categorizing of Biodiversity" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-4, June 2019, pp.1581-1584, URL: https://www.ijtsrd.com/papers/ijtsrd25153.pdf

IJTSRD25153

ABSTRACT

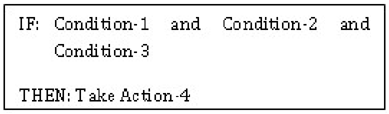
Rule induction is an area of machine learning in which formal rules are extracted from a set of observations. Machine learning is a field of artificial intelligence that uses statistical techniques to give computer systems the ability to "learn" from data, without being explicitly programmed. Machine learning applications are classification, regression, clustering, density estimation and dimensionality reduction. The CN2 algorithm is a classification technique designed for the efficient induction of simple, comprehensible rules of form “if cond then predict class”, even in domains where noise may be present. Biodiversity means biological diversity, the variety of life found in a place on Earth or, often, the total variety of life on Earth. This research used butterflies as biological dataset for categorizing biodiversity and passed it to CN2 Rule Induction. In this research, “The Fauna of British India, Ceylon and Burma. Butterflies. Vol. I and Vol. II” written by C.T Bingham are used as the required knowledge for resource and categorizing biodiversity of butterfly families by rules induction with CN2 algorithm system has developed. In this system, MS Visual Studio as a programming tool and MS SQL Server as for database development are used.

Keywords: *Machine Learning, Rule Induction, CN2 Algorithm, Biodiversity*

1. INTRODUCTION

In artificial intelligence, an expert system is a computer system that follows the decision-making ability of a human expert. Expert systems are designed to solve complex problems by reasoning through bodies of knowledge, represented mainly as if–then rules rather than through conventional procedural code. [11]

The most common form of architecture used in expert and other types of knowledge-based systems is the production system, also called therule-based system. In computer science, rule-based systems are used as a way to store and manipulate knowledge to interpret information in a useful way. They are often used in artificial intelligence applications and research. [5] This type of system use knowledge encoded in the form of production rules, that is, if…then rules. Rule have an antecedent or condition part, the left-hand side, and a conclusion of action pert, the right-hand side. [4]



```
IF: Condition-1  and  Condition-2  and
    Condition-3

THEN: Take Action-4
```

1.1. The Representative Architectures of Expert System

An expert system is divided into two subsystems: the knowledge base and the interface engine. The knowledge base represents facts and rules. The inference engine applies the rules to the known facts to deduce new facts. Inference engines can also include explanation and debugging abilities. [11] The architectures of Expert Systems today reflect knowledge engineers understanding of how to represent knowledge and how to perform intelligent decision-making tasks with the support of knowledge base system. The Expert System architecture is independent of specific computer hardware. Determinants for computer hardware selection would include the size of the knowledge database, the desire speed of the system’s responses and the level of sophistication for the user interface.

Figure (1.1) shows the architecture of a simple Expert System. The architecture of the simple Expert System could be extended. One common extension is to expand the knowledge base into a knowledge database and a domain database. These two databases could be managed by a database management system (DBMS). [9]

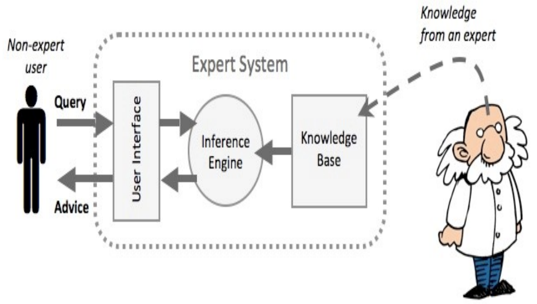


Figure (1.1) Architecture of a simple Expert System

1.2. Building A Knowledge Base

The procedure of building a knowledge base is called knowledge engineering. Investigation of a particular domain, determining what concepts are important in that the domain and creation of a formal representation of the objects and relations in the domain are the role of a knowledge engineer. Often, the knowledge engineer is trained in representation but is not an expert in the domain at hand, be it circuit design, space station mission scheduling or whatever. The knowledge will usually interview the real experts to become educated about the domain and to elicit the required knowledge, in a process called knowledge acquisition.

1.3. Architectures of a Rule-Based System

Each rule represents a small piece of knowledge relating to the given domain of expertise. A number of related rules collectively may correspond to a chain of inferences which lead from some initially known facts to some useful conclusions. The inference process is carried out in an interactive mode with the use providing input data needed to complete rule chaining process. Figure (1.2) shows the architecture of Rule-based system. [5]

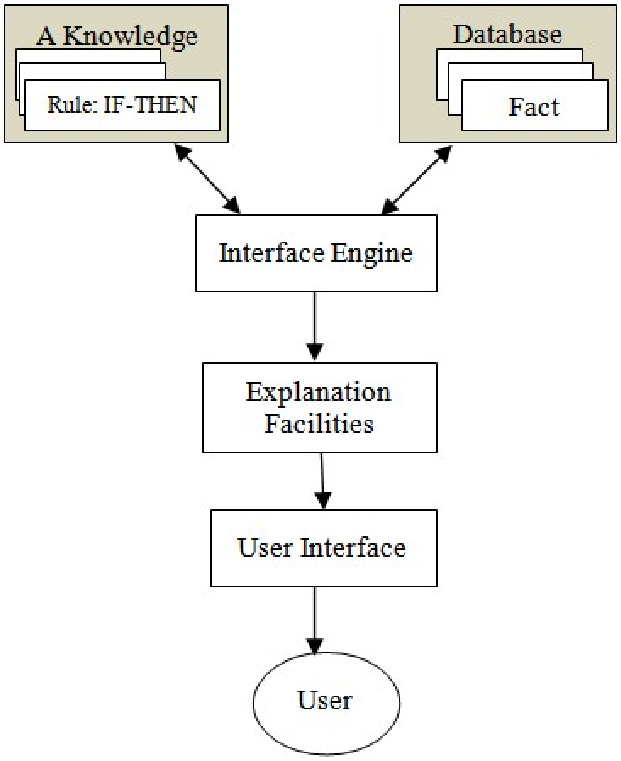


Figure (1.2) Architecture of Rule-Base System

A typical rule-based system consists of the following components are:

Knowledge base: contains the rules embodying expert knowledge about the problem domain

Database: contains the set of known facts about the problem currently being solved

Inference engine: carries out the reasoning process by linking the rules with the known facts to find a solution

Explanation facilities: provides information to user about the reasoning steps that are being followed

User interface: communication between the user and the system

In this research, rule induction with CN2 was used for classification of butterfly that is the simplest setting for classification.

1.4. Rule Induction

The rule induction system can create rule that fit the example cases. The rule can then be used to access other cases the outcome is not known. The heart of a rule induction system is an algorithm, which is used to induce the rules from the examples.

Induction methods use various algorithms to convert a knowledge matrix of attributes, values, and selections to rules. Some well-known rule learning algorithms are AQ, CN2, FOIL, RIPPER AND OPUS. [7]

1.5. Rule Induction with CN2

The **CN2 induction algorithm** is a learning algorithm for rule induction. It is designed to work even when the training data isimperfect. It is based on ideas from the AQ algorithm to produce rules and combine decision tree learning (such as C 4.5, ID3) to handle noise. As a consequence it creates a rule set like that created by AQ but is able to handle noisy data like ID3. [4]

2. Materials and Methods

In this research, family keys of butterflies are used as materials or biological data set and rule induction with CN2 algorithm is used as rule indention method.

2.1. Identifying Keys of the Butterfly

Butterflies are the most well-known of all insects. They are among the most beautiful creatures on Earth. They are popular among nature lovers as well as a subject for scientific study. Figure (2.1) shows the external features of the butterfly that are used in the scientific categorizing.

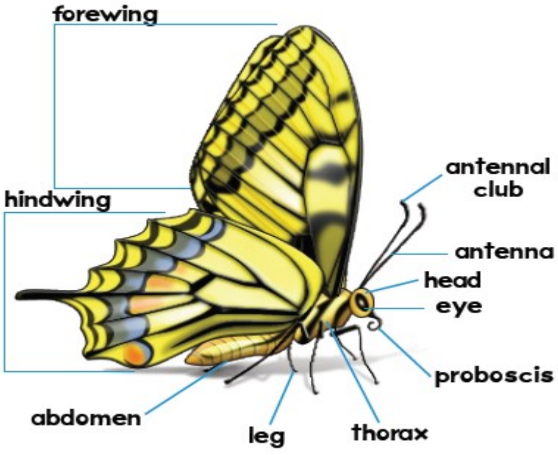


Figure (2.1) External Features of Butterfly

Butterflies and moths are insects that make up the order Lepidoptera, derived from the Greek words *lepidos* for scaly and *ptera* for wings. Four wings are present. The wings are membranous, with veins or nervures running longitudinal form base to the wing margins. The pattern formed by these veins (wing venation) is of primary importance in the classification ofLepidoptera. Wings of all the butterflies' families showed considerable variations in shapes and vein patterns reflecting their specific nature. [8] Figure (2.2) shows the parts of wings and venation pattern of butterflies that are also used in the scientific categorizing.

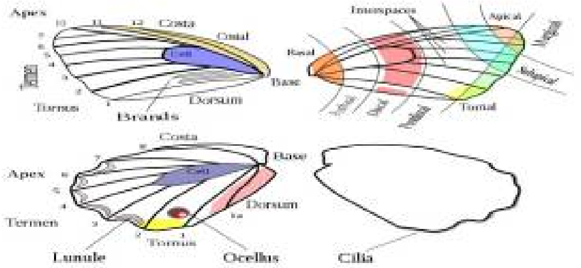


Figure (2.2) Parts of Wings and Venation Pattern of Butterfly

2.2. Exploring the Rule Induction with CN2 Algorithm

Rule induction is a popular method that automates the knowledge acquisition process when knowledge is expressed in terms of rules in classification-type problems. Rule induction examines historical cases and generates the rules that were used to arrive at certain recommendations. Rule induction can be used by a system engineer, an expert, or any other system builder.

CN2 based on the ID3 (Quinlan, 1983) and AQ (Michalski, 1969) algorithms. The ID3 algorithm provides itself to such modification by the nature of its general-to-specific search. The AQ algorithm's dependence on specific training examples during search makes it less easy to modify. CN2 was designed to modify the AQ algorithm itself in ways that removed this dependence on specific examples and increased the space of rules searched. CN2, a new induction algorithm combines the efficiency and ability to cope with noisy data of ID3 with the if-then rule form and flexible search strategy of the AQ family. Figure (2.3) shows the CN2 induction algorithm. [4]

```
Let E be a set of classified examples.
Let SELECTORS be the set of all possible
selectors.
            Procedure CN2 (E)
Let RULE-LIST be the empty list.
Repeat until BEST.CPX is nil or E is empty:
Let BEST.CPX be Find-Best.Complex(E).
        If BEST.CPX is not nil,
        Then let E' be the examples covered
            by BEST.CPX.
        Remove from E the examples E'
            covered by BEST.CPX.
        Let C be the most common class of
            examples in E'.
        Add the rule 'If BEST.CPX then the
            class is C' to the end of RULE-LIST.
        Return RULE-LIST.
```

Figure (2.3) The CN2 Induction Algorithm

3. Approaching Rules Induction: CN2 Algorithm in Butterfly Families

In this research, the proposed system contributed according with the following system flow diagram.

3.1. System Flow Diagram

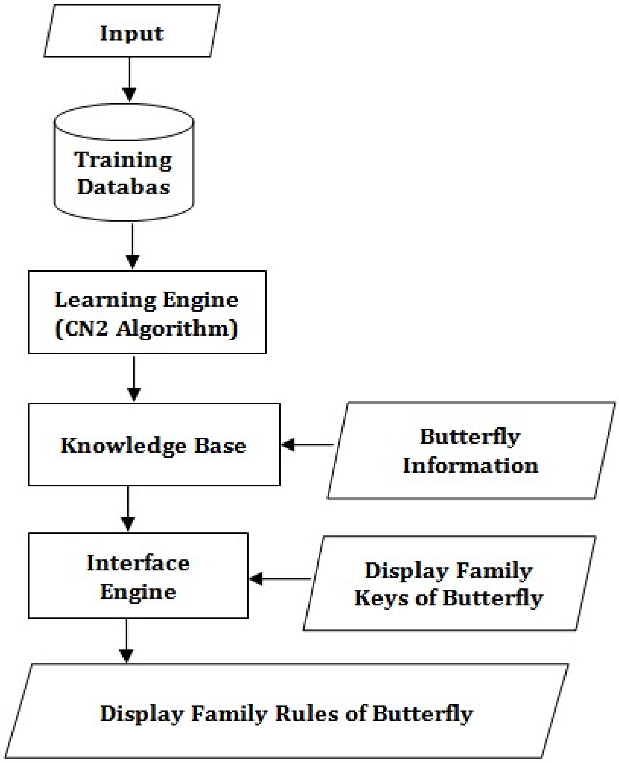

Input
Training Databas
Learning Engine (CN2 Algorithm)
Knowledge Base
Butterfly Information
Interface Engine
Display Family Keys of Butterfly
Display Family Rules of Butterfly

Figure (3.1) System Flow Diagram

3.2. Executing Family Keys for Categorizing

A dichotomous key is a tool that can be used to classify objects. Dichotomous means "divided in two parts" or "binary classification". Therefore, this key uses a series of yes or no questions to place objects into groups. [8] Butterflies are classified in the Kingdom: Animalia, Phylum: Arthropoda, and Order: Lepidoptera. [1] [2]

Kingdom Animalia	(animals)
Phylum Arthropoda	(arthropods, invertebrate animals with an exoskeleton, a segmented body, and jointed legs)
Class	**Insecta** (insects, rthropods with 6 legs, 2 antennae, and a 3-part body)
Order Lepidoptera	(butterflies and moths)

By using the rule induction with CN2 algorithm and identification key of family of butterfly, the following rules can be developed. Table 3.1 shows the family keys of butterfly.

Table (3.1) Family Keys of Butterflies

1	A. Antennae approximate at base, hind tibiae with only a terminal pair of spurs, one or more of the veins in the fore wing forked or coincident beyond the cell ……………………………………………………………………………**2** B. Antennae wide apart at base, hind tibiae generally with a medial as well as a terminal pair of spurs, all the veins in the fore wing from base or from cell, none forked or coincident beyond …………………**Hesperiidae**
2	a. Precostal nervure in hind wing present ……………………………………………………………………………**3** b. Precostal nervure in hind wing absent ……………………………………………………………**Lycaenidae**
3	a1. Front pair of legs imperfect in one or both sexes…………………………………………………………………**4** b1. Front pair of legs perfect in both sexes………………………………………………………………………**5**

4	a2. Front pair of legs perfect in both sexes……………………………………………**Nymphalidae** b2. Front pair of legs imperfect in male, perfect in female……………………………………………**Nemeobidae**
5	a2. Vein 1a in hind wing wanting, claws simple……………………………………………**Papilionidae** b2. Vein 1a in hind wing present, claws bifid……………………………………………**Pieridae**

4. Result and Discussion

After categorizing of the butterflies by using rule induction with CN2 algorithm, there have six rules for six families respectively as the result. Table (4.1) shows the results of the rules for butterfly family.

Table (4.1) Rule for Family of Butterfly

Rule No	Family Name	Rule for Family (Family Key)
1	Nymphalidae	A (a) (a1) (a2)
2	Nemeobidae	A (a) (a1) (b2)
3	Papilionidae	A (b) (a1) (a2)
4	Pieridae	A (b) (a1) (b2)
5	Lycaenidae	A (b)
6	Hesperidae	B

Rule induction can be used by a system engineer, an expert, or any other system builder. The categorization of butterfly family by using rule induction with CN2 was simple, clear and useful for the researcher. This research is starting for approach the co-operation of computer area and the biological area. That is showing the usefulness of the computer field of biological field for the categorization in biodiversity.

5. Conclusion

This research paper has introduced the idea of inducing rules from sets of examples in order to speed up the development of knowledge bases in expert systems. The rule induction with CN2 algorithm is the best technique for categorizing of biological dataset. The results give rules for categorization and the implemented system is easy to use and understand for biologist; who is willing to categorize the butterflies by computerization. A big advantage of the rule induction is that it enhances the thinking process of the expert. Butterflies are the best-known and best group of insects for examining pattern of terrestrial biotic diversity and distribution. For the children, schools are being encouraged to study the life and habit of butterflies and in this way come to appreciate the beauty of the natural environment and become involved in efforts for its conservation. So, this research will meet the aim of correct, easy, fast, simple and useful than traditional manual methods. As a future work, we will continue to do the research for the applied algorithm in biodiversity.

References

[1] Bingham. C.T, "The Fauna of British India, Ceylon and Burma. Butterflies. Vol. I Taylar and Francis, London", 1905

[2] Bingham. C.T, "The Fauna of British India, Ceylon and Burma. Butterflies. Vol. II Taylar and Francis, London", 1907

[3] Clark Peter and Boswell Robin. "Rule Induction with CN2: Some Recent Improvement" Machine Learning, Fifth European Conference, 1991

[4] Clark Peter and Niblett T. "The CN2 induction algorithm" Machine Learning Journal, 3(4); 261-283, 1989

[5] CSCU9T6 / ITNP60. "Reasoning Systems", University of Stirling, 2015

[6] Faruk Ertuğrul Ömer et al. "Identification of Butterfly Species by Similarity Indexes Based on Prototypes", International Journal of Computer Vision, Machine Learning and Data Mining, Volume 1, 2015

[7] Furnkranz Johannes and Kliegr Tomas. "A Brief Overview of Rule Learning"

[8] Omaha's Henry Doorly Zoo, "BUTTERFLIES. Taking Science to the Backyard", 2011

[9] Tan C. F. "The Application of Expert System: A Review of Research And Applications", ARPN Journal of Engineering and Applied Sciences, Vol. 11, No. 4, February 2016

[10] https://docs.orange.niolab.si/3/visual programming/widgets/model/cn2ruleinduction.html

[11] https://en.wikipedia.org/wiki