https://doi.org/10.15623/ijret.2018.0706018 Received: 13-04-2018, Accepted: 25-05-2018, Published: 15-06-2018

SIMULATION OF IMPROVED ASSOCIATION RULE DISCOVERY SYSTEM FOR DECISION SUPPORT

Macarthy Osuo-Genseleke[1], Asagba Prince O[2]

[1]PhD. Student, Department of Computer Science, IgnatiousAjuru University of Education, Rivers State, Nigeria
[2]Professor, Department of Computer Science, University of Port-Harcourt, Rivers State, Nigeria

Abstract

Recent computing transactions entails large sum of data which are retrieved, stored and used for operations. The data often contain association relationships which can be mined to aid management decision. We simulate a data mining system for the association of rule discovery using an improved C4.5 Algorithm. The system extracts data and its close relationships that will be used for decision making. The system analysis and its design was done using the Object - Oriented System Analysis and Design Methodology (OOADM). Weather data file was used to test run the system and results shows that mining associated rules in a large database is important in decision making. Algorithms and Designs in rule discovery associations that seem complex but very useful in making decisions need to be well implemented to be useful to users. This paper presents a simulation of an improved C4.5 Algorithm for association rule discovery system used in decision support. Java programming language is used for the implementation with Netbeans IDE. When the system was tested, users reported better usability, efficiency and clarity of results from the application.

Keywords: - *Rule Discovery, Data Mining, C4.5, Decision Tree,Algorithm, Association Rule.*

----------***----------

1. INTRODUCTION

Presently, information technology has experienced massive amount of data both in databases, data warehouses, etc. There has equally been a concordance difference in the applications of automated data processing and collection tools, causing the retrieval and use of complex amount of data transaction in various systems [1]. The findings of data items closely related with a large sum of data will definitely help in decision making.

The mining of large data with certain association rules is been largely researched. Applying data mining in this study to extract data items closely related in a database will aid in decision making [2]. A data mining pattern is established that will extract data closely related from a massive database used for transaction by extending the existing association rule used in data mining.

An Association relationships rule like {Milk,Tea}→{ Bread}, in a sales store represents a customer that bought milk and tea and might buy bread. This can be applied in decisions about marketing events, weather forecast, etc.

An Association rule W Z, is categorized as a set of items which implies that a transaction containing W items also contain Z items. Extracting data items closely related in databases are applied in organizations. In research, extraction of association rule data items from massive data, we assumed that the data contains:

- A set of transaction, Ti = { d1….di} where d is the set of all data items.
- The explanation of transaction, D, contains the description of each item in D, where di €D.

There is also a need to adopt an extended relational model in facilitating the managing of massive data that allows attribute values to be single or multiple. In this study, the discovery method developed can be applied to other data files, a relational table and in relational expressions.

Looking at A, as an item Ai or a set of conjunction items Ai A……A Ai where Ai…., Ai € 1. The support of pattern A in S, a (A/S) is the number of transactions in S containing A versus the total number of transactions found in S. The confidence A associated to B in set S, is the chance that the B pattern occurs in set S when pattern A also occurs in set S. Looking at the regularities of the occurrences of its patterns and the significance of the associated rule, an expert may be required to specify two thresholds: minimum or maximum support.

2. NEED OF DATA MINING

The need for data mining is to determine unknown and hidden patterns from data generated by organizations. Data mining is the answer to the problem organizations grapple with to get needed value from their stored data [3].

https://doi.org/10.15623/ijret.2018.0706018 Received: 13-04-2018, Accepted: 25-05-2018, Published: 15-06-2018

3. TECHNIQUES/FUNCTIONALITIES OF DATA MINING

Data mining have two vital goals which include: prediction and description. In prediction, existing variables are used to predict future values while description focuses on discovering key points that describe data [4]. Data mining techniques include: associations, classifications, heuristic clustering algorithm and clustering.

- Association Rules: Association rule is a class of items displayed as W Z, where W and Z are sets of items and W items is very likely to contain Z items. For example, a rule such as 85% of customers that buy milk and sugar also buy tea.
- Classification Rules: Classification is used in discovering rules that divide data into separate groups. Decision tree, artificial neural networks, genetic algorithms and some statistical models of mining data are classification discovery model.
- Clustering: Clustering are groupings of similar data and its aim is for these data to be used optimally.
- Heuristic Clustering Algorithm: Grouping of data into various groups. K-means, Kmenoids are clustering algorithms used in application developments. The particular clustering algorithm selected depends on the type of data and system requirement. Clustering algorithm in this process involves the separation of the hidden node's output values.

4. RULE EXTRACTION ALGORITHM

This algorithm can be applied to numerous data in the execution of precise responsibilities which include rule extraction, rule clustering and rule pruning [6].

- Rule Extraction (RE): It initializes the mined rule list by making it hollow. It arranges content in a certain frequency, it selects common sample as the base to produce and add the rule to the list of other extracted rules. Thereafter, all samples would be found and removed while the process is repeated until space is exhausted.
- Rule Clustering: This is a situation where rules that are of same category level are grouped and clustered together.
- Rule Pruning: Here, rules that are no longer in use, inactive and keeps repeating itself are removed as to get the best size of the rule list.

5. IMPROVED C4.5

This is the implementation a data mining system for the association of rule discovery using an improved C4.5 Algorithm. This model handles consistent, unequal, discrete and continuous data pattern. It handles complex data set. A sample of a complex Weather data was used for the implementation and data itself was mined first. The mined data and the mining rules are associated via mechanisms which associates items to data based on some predefined rules. The association is first discretized so they can easily associate one to the other in the process of mining the rules. When the discretization is complete the system generates rules that rebuild data and production of the decision tree that the system will deploy in the final storage of the decision tree system as a prolog knowledge base that will be necessary in the stage of producing the decision required by the system.

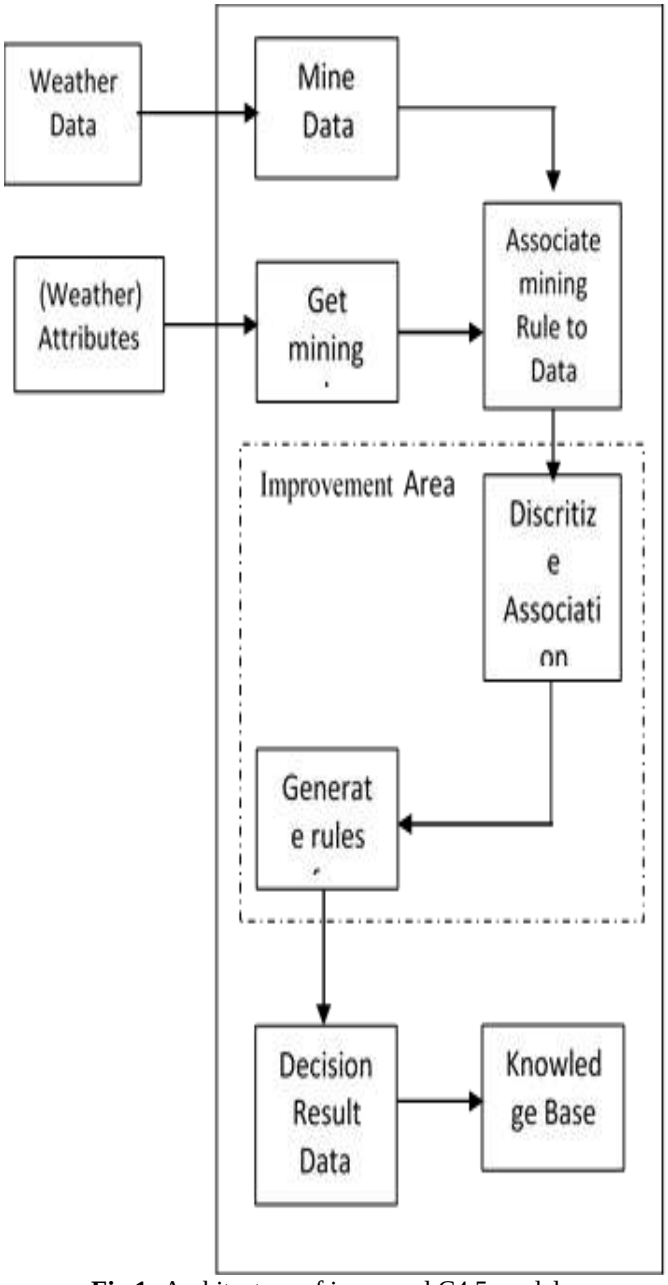

Fig 1: Architecture of improved C4.5 model

6. IMPROVED C4.5 ALGORITHM

- Select dataset as an input to the algorithm for processing.
- Create a given root node for the given tree.
- If all dataset are positive return the single-node tree root, with label = +
- If all dataset are negative return the single-node tree root, with label = -

https://doi.org/10.15623/ijret.2018.0706018 Received: 13-04-2018, Accepted: 25-05-2018, Published: 15-06-2018

- If attribute is empty, return the single-node tree Root, with label = most common value of Target attribute in dataset
- Select the classifiers
- A the attribute from Attributes that best classifies the dataset.
- The decision attributes for Root A
- Calculate entropy, information gain, gain ratio of attributes.
- For each given value vi,
- Insert a new branch of the tree below the Root that corresponds to the test A = vi
- Let dataset vi be subset of dataset that have value vi for A
- If dataset vi is empty
- Then below the new branch, add a leaf node with label = most common value of Target_ attribute in Dataset Else
- Below this new branch add the subtree
  Goto Step 2
- Tree generator generates the decision tree

7. IMPLEMENTATION OF THE IMPROVED SYSTEM

The implementation of the system is shaped by the user interface. Guideline for our interface design: (1) expose interaction opportunities in a straight forward and easily understood manner, and (2) provide immediate feedback in response to control actions. These both contribute to our overarching design goal, to support the progressive refinement of the models that contribute to improve rule discovery from given data set. As shown below, the system interface consists of eight major tabs: attributes, discretization, aggregation, casuality,C4.5 setting, Regression, Analysis and Recommend. Rule discovery users click a tab and the tab content displays activities that can be carried out on it.

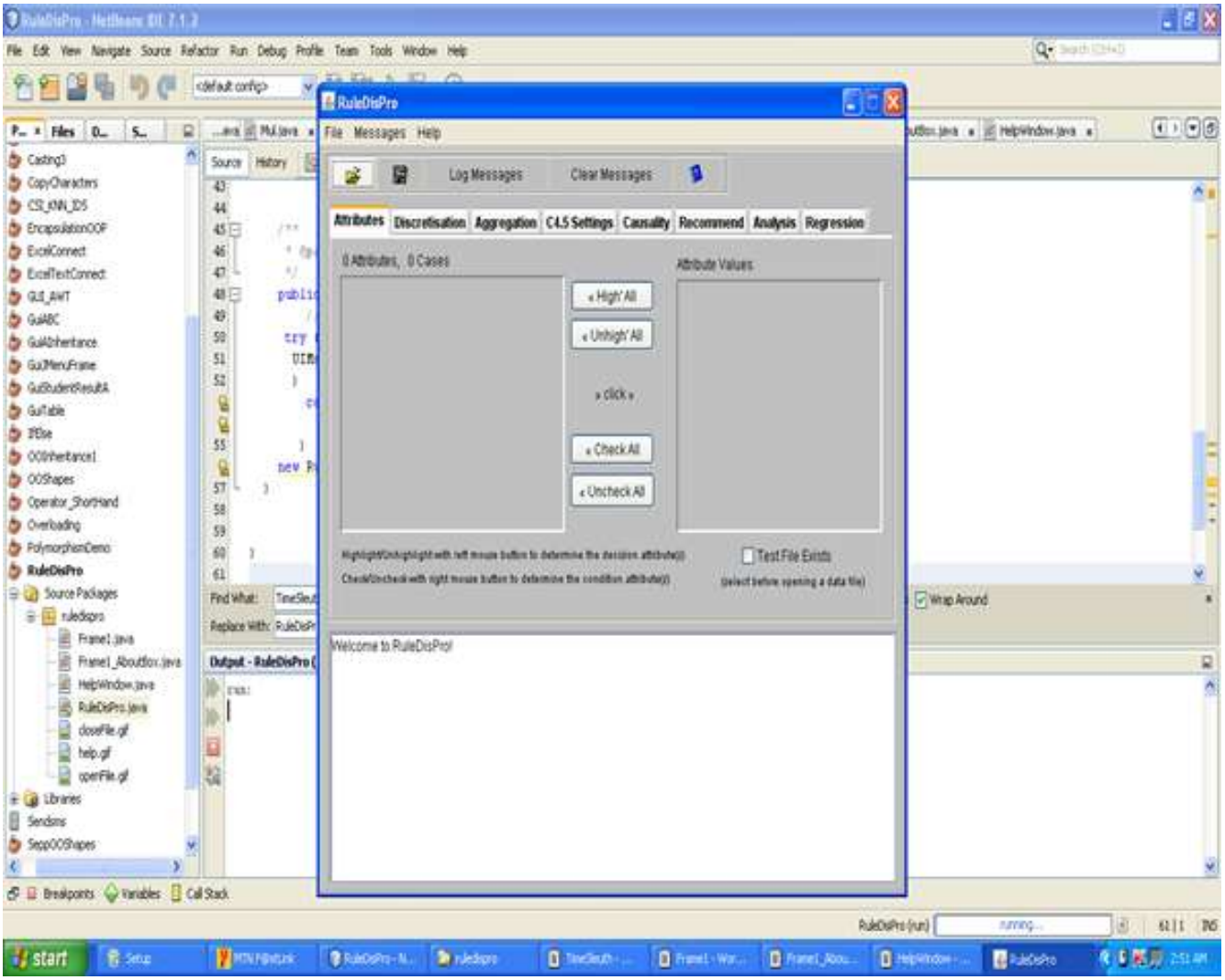

Fig 2: Output Screen showing attribute tab before data is loaded.

https://doi.org/10.15623/ijret.2018.0706018 Received: 13-04-2018, Accepted: 25-05-2018, Published: 15-06-2018

The windows show that attributes tab is the tab selected by default. Since no data is loaded, the attribute box is empty as shown in fig. 2.

The program Help menu output in fig. 3 illustrate the Rule Discovery Online documentation window. The window shows Introduction, Sample Session and GUI Guide which directs the user on how to use the graphical user interface. Other topic of the documentation also includes Investigating Causality, Contact Information and Input Files which directs the user to select and open an input data file. Once the input file is opened in the system the attribute window will display the data description file which contains the attribute variables on the window.

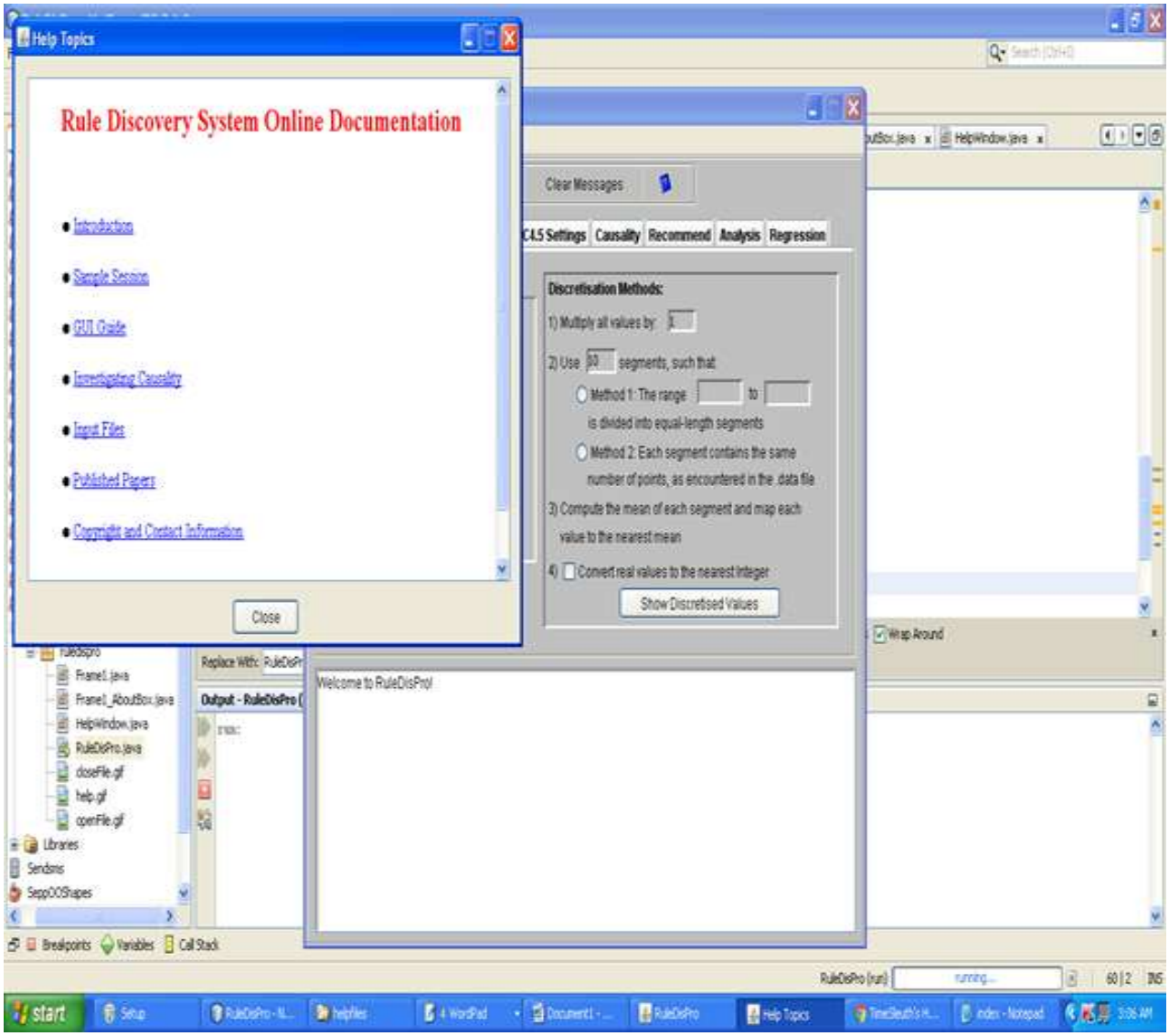

Fig 3: A Rule Discovery Online Documentation Window

The input file opened in the stage of executing the system is the Weather data file and the attributes are illustrated in figure 4. Some of the data attribute include 8 attributes- AirTemp, Rain, MaxWindSpeed, AvgWindSpeed, WindDirection, Humidity, SolarRad and originalDecision and 343 cases. The data reading process also shows the location where the data cases and the data attributes names are listed below the window. The date when the data file is upload is also displayed below the window.

https://doi.org/10.15623/ijret.2018.0706018 Received: 13-04-2018, Accepted: 25-05-2018, Published: 15-06-2018

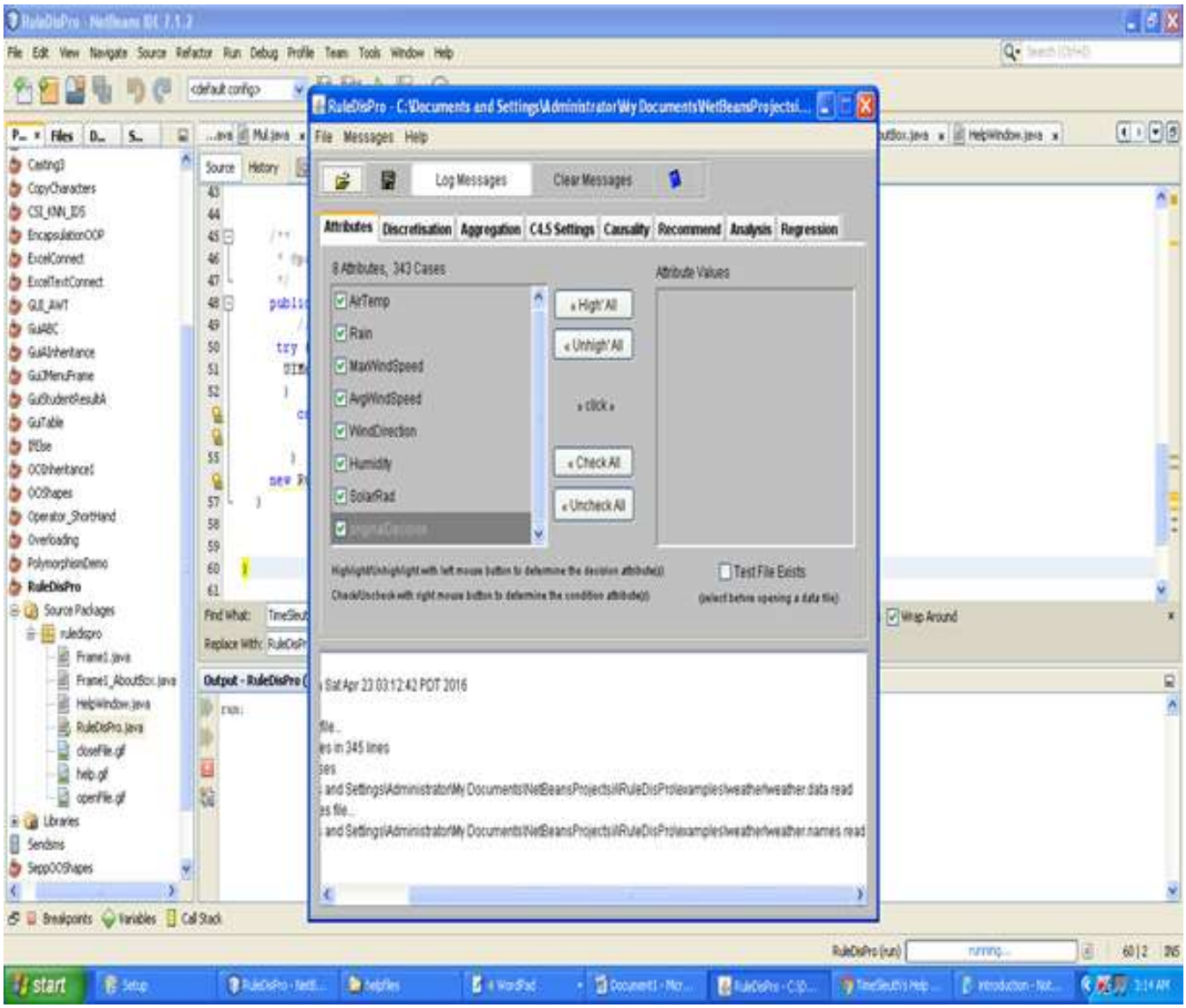

Fig 4: The Attribute Tab showing the Data Attributes on the Window

In fig. 5, the C4.5 Rule setting window shows different textboxes needed to be filled to set the arguments to handle the rule discovery system and call the algorithm that handle the decision process using the rule discovery process engine which is normally embedded in the C4.5 frame work. The window also provides options for output for the system in Prolog Output form or in No Prolog Output. The minimum confidence value has to be specified and the option of selecting the generation of analysis data and ignoring the C4.5 rules output can also checked.

https://doi.org/10.15623/ijret.2018.0706018 Received: 13-04-2018, Accepted: 25-05-2018, Published: 15-06-2018

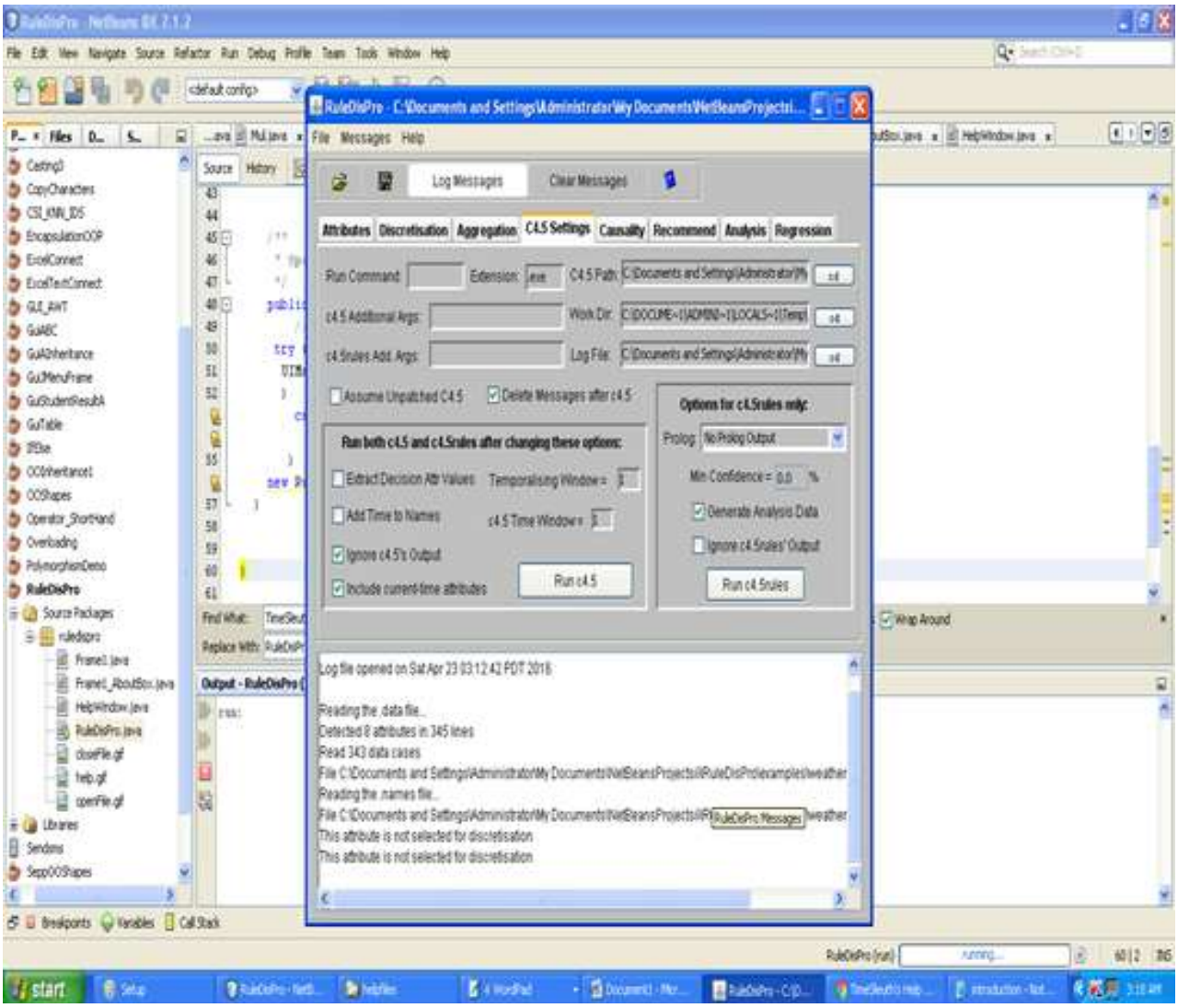

Fig 5: The C4.5 and C4.5 Rules Setting and Run Window

8. DOCUMENTATION OF IMPLEMENTED SYSTEM

Documentation involves writing down the manual of usage of the new system for reference purpose. The document can be saved in a file called System Specification and it contains:

- Data Input methods: This talks about the data required, data capture method, data checking and control procedures.
- Data Output methods: This addresses information produced from the system whether regular, exception or other reports. Ultimately, it is important for us to know the extent the systems assist real data miners to accomplish real tasks.

Fundamentally, there are ways we might try to learn this:

- To model the users' task in a way that permits automated evaluation. This is an approach adopted in this documentation, for example: the user generates data concerning the area to be modeled by the system. The data is stored in two file, one contains the raw data and the second contains the data attributes. When the system is in execution, the data is called up and the system uses it in the rule discovery process. Fully automated study designs with reusable evaluation resources was defined for summarization and machine translation. Separating components that are needed to evaluate rule discovery is possible. Modeling interaction is difficult, however, since the complex interplay between perception and cognition resists the type of binary (right/wrong, Yes/No, Event/No Event) characterization that has been so successfully employed in decision tree creation and automated evaluation frameworks for information data mining and rule discovery.
- At the other end of the spectrum, we can learn a lot by building real systems, giving them to real data rules, and watching them accomplish real tasks. Observational studies are most times employed when mature system designs and experienced data models are available. In

https://doi.org/10.15623/ijret.2018.0706018 Received: 13-04-2018, Accepted: 25-05-2018, Published: 15-06-2018

times like this, rich description and systematic qualitative analysis methods can offer insights into interactions between system, data, attributes and task that seems difficult to capture using more highly structured study designs as a result of the expensive nature of data collection and analysis also, controlled user studies in which two systems are compared quantitatively are often used when the research questions to be explored can be crafted in a way that is sufficiently narrowed. This is because both approaches bring strong, weak and mixed-method studies that wrap richer data collection around a quantitative controlled study design and can often affordably yield deeper insights than either approach to be tried in isolation. I therefore adopted a mixed-methods study design for the experiment. Weather data used from standard metrological stations are used for testing the system in this dissertation.

9. TESTING

A Java Run Time Environment (JRE) was Installed in the computer system and tested if the Java application will run as expected. Then a data file and the attribute file was created and named appropriately so it will be easy to call up in the application. The sample data file is shown below saved as weather.data. The attribute file illustrates the data set represented below, from Air Temperature to Solar Radian, the last value is sample range between 80 and 100 except the dummy value – 9999. Once the data is loaded in the system, the process of deploying the application in handling the decision results to the output. The implementation represents the presentation of the files including data gathering, loading and testing. The system application was designed to run on the Java platform and system need correct version of Run time environment for the application to function properly.

```
85,  0,  0,  0,  0,  25,  -9999,  -9999
82,  0,  8,  0,  101,  95,  -9999,  93
82,  0,  7,  1,  95,  99,  -9999,  92
80,  0,  0,  0,  113,  103,  0,  90
77,  0,  0,  0,  152,  106,  0,  89
77,  0,  0,  0,  123,  108,  0,  88
75,  0,  0,  0,  124,  108,  0,  87
75,  0,  0,  0,  259,  0,  0,  86
74,  0,  0,  0,  293,  0,  0,  85
74,  0,  0,  0,  131,  0,  0,  84
74,  0,  0,  0,  292,  0,  0,  84
74,  0,  0,  0,  60,  0,  0,  83
74,  0,  0,  0,  84,  0,  0,  83
75,  0,  0,  0,  106,  0,  1,  82
79,  0,  0,  0,  303,  0,  11,  82
83,  0,  8,  2,  40,  0,  25,  82
```

Fig 6: Part Sample of Weather Data [5]

10. THE ATTRIBUTE FILE FOR THE DATA

AirTemp: continuous, Rain: continuous, MaxWindSpeed: continuous, AvgWindSpeed: continuous, WindDirection:continuous, Humidity: continuous, SolarRad: continuous

Table 6: Results of C4.5 and Improved C4.5 showing parameters and the performance of the algorithms

Parameters	C4.5	IC4.5	Remark
Accuracy	94.7%	96.3%	Improved
False Value Rate	9.46%	6.45%	Reduced
Rule Selection Rate	97.36%	98.65%	Improved
Execution Time	34.6 (s/kb)	35.1(s/kb)	Increased

11. CONCLUSION

Finally, we have used data mining technique for association rule discovery to build a system that supports quick decision making. The improved system model was also implemented using C4.5 API and Java programming tools. This offers us the advantage of stepwise progress that includes scalability structures to systems when users' needs them to build a decision support system from data generated and made system mining process more exact, arranged and rule discovery route more reliable. The system developed also enable new innovation allowing various data stored in files to be processed via the system reusable components developed in the program. The challenge of discovering the necessary rules that will enable the developer to build a system for fast decision making was solved by using a C4.5 API package library.

REFERENCES

[1] MacarthyO. and Onyejegbu L. An improved design for association rule discovery system for decision support system. International Journal of ngineering Research and Technology. Vol 5(11), pp 236-241, Nov 2016

[2] Jiawei H. and YongjianF. Discovery of Multiple Level Association Rules from Large Databases, Proceedings of the 21st VLDB Conference Zurich, Swizerland.Sept 1995

[3] Yashpal S. and Alok S. C. Neural Networks In Data Mining, India, Journal of Theoretical and Applied Information Technology. July 2005.

[4] Sonalkadu, S. D. Effective Data Mining Through Neural Network, International Journal of Advanced Research in Computer Science and Software Engineering Vol 2 (3), Feb 2012.

[5] LACIS (2015) Weather Data Source, Louisiana AgriClimatic Information System's, http://typhoon.bae.lsu.edu/datatabl/current/sugcurrh.html

[6] Gaurab, T. Effective Data Mining For Proper Mining Classification Using Neural Networks,International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5 (.2),pp 112-125. 2015