Preprocessor Agent Approach to Knowledge Discovery Using Zero-R Algorithm

Inamdar S. A
School of Computational Science
Swami Ramanand Teerth,
Marathwada University, Nanded

Narangale S.M.
School of Media Studies
Swami Ramanand Teerth,
Marathwada University, Nanded

G. N. Shinde*
Indira Gandhi College
CIDCO, Nanded-431603,
Maharashtra, India

***Abstract*— Data mining and multiagent approach has been used successfully in the development of large complex systems. Agents are used to perform some action or activity on behalf of a user of a computer system. The study proposes an agent based algorithm PrePZero-r using Zero-R algorithm in Weka. Algorithms are powerful technique for solution of various combinatorial or optimization problems. Zero-R is a simple and trivial classifier, but it gives a lower bound on the performance of a given dataset which should be significantly improved by more complex classifiers. The Proposed Algorithm called PrePZero-r has significantly reduced time taken to build the model than Zero-R algorithm by removing the Lower Bound Values 0 while preprocessing and comparing the result with class values. Also proposed study introduced new factor "Accuracy (1-e)" for each individual attribute.**

Keywords- Data mining; Zero-R algorithm; Lower Bound Value; Class values.

I. INTRODUCTION

Data mining has various techniques to extract useful information in large amounts of data. Data mining is defined as a technique of finding hidden information in a database [1]. It may be called as data driven discovery, explorative data analysis, deductive learning. Data mining in general falls in to the following categories: classification patterns, association patterns, sequential patterns, and spatial-temporal patterns. The important feature of Data Mining algorithms is running time of an algorithm must be predictable and acceptable in large database.

II. RELATED WORK

Research has shown that over the past few year data mining tools are heavily used in healthcare spectrum. Agent based approach has become an advanced trend in Knowledge discovery. The Classify Agent offers an alternative to achieving the data mining purpose of obtaining a good model with faster classification time from large database within reasonable timeframe. In Agent Based Meta Model Agents are basic modeling entities that maintain a set of properties and behaviors. By factoring agents, relationships, and behaviors into separate components, more modular and expressive models can be created. Research shows the knowledge discovery is using multi-agent approach for quicker and reliable information retrieval [2-11].

III. PRELIMINARIES

KDD (Knowledge Discovery from Databases) is the process of finding useful information and patterns in data. Data mining is the use of algorithms to extract the information and patterns derived by the KDD process. KDD is multistep process, the input to this process is the data and the output is the useful information desired by the users [1].

There are many techniques for classification such as neural networks, Bayesian, decision tree, instance based learning, genetic algorithm, rough set, and fuzzy logic [3].

A. Agents:

Agents are used to perform some action or activity on behalf of a user of a computer system. Agent refers to the entities which run in dynamic environment and have higher self-government capacity. Agent software is a type of computer program which simulates human intelligence behavior. Agent should be able to learn from experience and to act autonomously to the ever changing task.

B. MAS:

A group of agents can collectively and collaboratively form a Multi Agent System (MAS) to perform complex and lengthy tasks [1-7].

IV. WHY ZERO-R?

Rules can be extracted from the tree by search the tree path from root to the leaf. C4.5 is an algorithm used to generate a decision tree and an extension of Quinlan's earlier ID3 algorithm [8-12]. C4.5 is a technique to generalize rules associated with a tree that accumulate all the tests between the root node and the leaf node. This technique uses the training dataset to estimate the accuracy of each rule [2, 13-14]. The decision-formation and tree method like the nearest neighbors method, exploits clustering regularities for construction of decision-tree representation.

The decision tree learning method requires the data to be expressed in the form of classified examples. Genetic Algorithms are powerful technique for solution of various combinatorial or optimization problems. They are more an instrument for scientific research rather than a tool for generic practical data analysis. Rough classifier is an extension of logic and discrete mathematics from rough set theory [7]. Like decision tree, rough classifier is a nonparametric model which

suits for the exploratory knowledge discovery and without intervention from users.

In WEKA Zero-R is a simple classifier. Zero-R is a trivial classifier, but it gives a lower bound on the performance of a given dataset which should be significantly improved by more complex classifiers. As such it is a reasonable test on how well the class can be predicted without considering the other attributes. It can be used as a Lower Bound on Performance. Any learning algorithm in WEKA is derived from the abstract WEKA classifiers.

V. THE PROPOSED ZERO-R ALGORITHM

The Proposed Algorithm called PrePZero-R shows that we are removing lower bound values 0 and checking the results how it affects the class value. Following are the steps of PrePZero-R Algorithm shown in fig.1

In figure.1, in first step, we are selecting classifier Zero-R, simultaneously checking for its capabilities, if it does not satisfy its condition capabilities then directly Exit. If it is capable then build classifier. In next step getting Instance values, then set the Lower Boundary value for instance, remove Lower Boundary value form calculation, finally calculate Zero-R class value then Exit.

VI. EXPERIMENTAL SETUP

The experiment was conducted using the UCI Pima Indian data set [2]. We used Data mining library WEKA 3.6.5. In WEKA we introduce new term Accuracy (1-e) which gives better results in error.

VII. RESULT & DISCUSSION

The experiment was conducted using the UCI Pima Indian data set [5]. The dataset contains 768 instances of Pima Indian heritage females who were diagnosed for diabetes. The diagnostic result (diabetes negative or diabetes positive) in the data set. The five attributes are as follow: number of times pregnant, plasma glucose concentration, serum insulin, diabetes pedigree function and finally the test result.

The experiment is to measure time and accuracy of a classification on the UCI Pima Indian dataset [6]. Class variable value is mutually exclusive, either diabetes negative or diabetes positive. There are 4 standard methods for Data Mining: association, classification, clustering techniques and prediction.

In Table 1 the class value for each attribute is compared in Zero-R and PrePZero-R algorithm. The Accuracy is measured in Time (Nano-Seconds) in comparison with Zero-R and PrePZero-R algorithm and the difference is shown in the last column in Table 1.

For most medical applications the logical rules are not precise but vague and the uncertainty is present both in premise and in the decision. For this kind of application a good methodology is the rule representation from decision-tree method, which is easily understood by the user [4].

The experimental result shows that we are removing lower bound values 0 and checking the results how it affects the class value. As shown in table1 the proposed algorithm has significantly reduced the running time and Accuracy. This criterion is important in agent based data mining to obtain the good knowledge model from the complex and large database. This classifier simply predicts the majority class in the training data. it makes little sense to use this scheme for prediction, it can be useful for determining a baseline performance as a benchmark for other learning schemes. Zero-R tests how well the class can be predicted without considering other attributes. It can be used as a Lower Bound on Performance.

VIII. FUTURE SCOPE

Execution of this PrePZero-R algorithm experiment on parallel multiprocessor system will increase efficiency. Creation and implementation of fuzzy set algorithms tends to increase the accuracy of the output. This experiment assumes the dataset can be minimized for lower bound values while calculating class value.

If fuzzy min-max algorithm is implemented for removal of attribute values from testing parameters, the accuracy can be increased and the time taken to build the model will definitely be reduced. Thus preprocessor agent approach shall be used effectively.

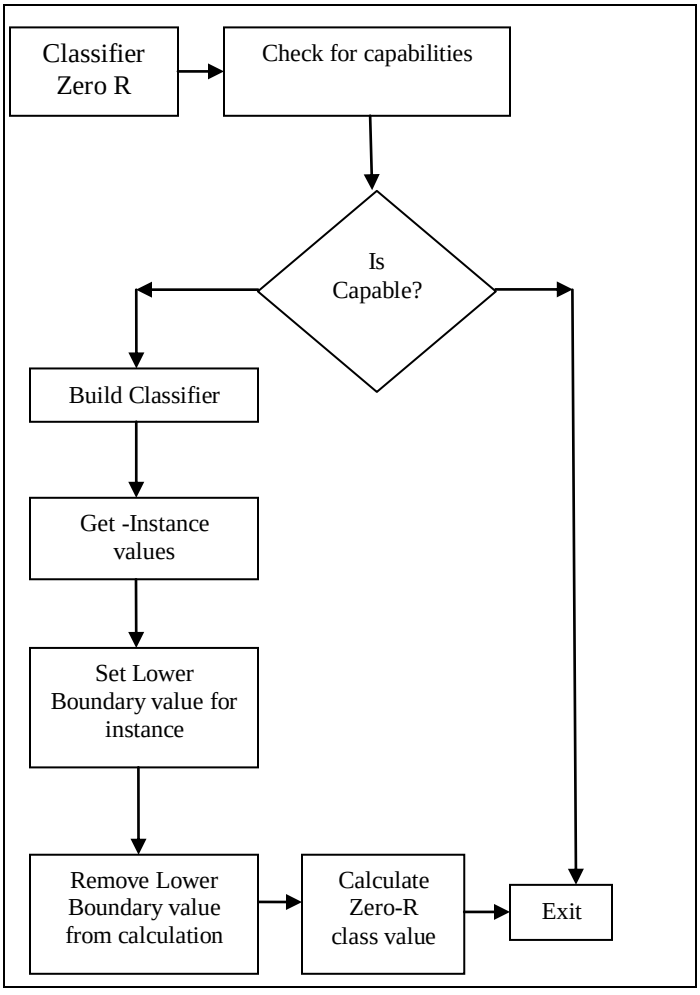

Figure 1. Preprocessing on individual attribute (Removal of lower boundary value from calculation)

TABLE1: COMPARISON OF RESULTS

Attribute	Class Value		Accuracy		Time (Nano Sec)		Difference
	ZeroR	PrePZeroR	ZeroR	PreP ZeroR	ZeroR	PreP ZeroR	
Number of times pregnant	3.8450520833333335	4.494672754946728	97.2255	97.0705	3445410	466400	2979010
Plasma glucose concentration	120.8945313	120.5388128	74.7986	74.8313	578705	462559	116146
Serum insulin	79.79947917	79.48249619	15.3368	15.4382	403892	437346	-33454
Diabetes pedigree function	0.471876302	0.463604262	99.7524	99.7542	523181	17540777	17017596
Test result	0.348958333	0.350076104	99.5452	99.5448	421213	3184273	-2763060

IX. CONCLUSION

To conclude the preprocessing approach for knowledge discovery shows significant increase in performance. The proposed PrePZero-R algorithm using WEKA has significantly reduced running time and increases accuracy. This algorithm does so by removing the Lower Bound Values and comparing the result with class values for each individual attribute. This criterion is important in Agent Based data mining to obtain the good knowledge from complex and large databases.

REFERENCES

[1] Margaret H. Dunham and Sridhar, Data Mining, Introduction and Advanced Topics, Prenticce Hall Publication, ISBN 81-7758-785-4

[2] M. Wooldridge, An Introduction to MultiAgent Systems. John Wiley & Sons Ltd, 2002. John Wiley & Sons, 2002.

[3] Abad-Grau, M. M., Arias-Aranda, D.: Operations Strategy and Flexibility: modeling with Bayesian Classifier. 106, 460--484 (2006).

[4] Plamena Andreeva, Maya Dimitrova, Petia Radeva, DATA MINING LEARNING MODELS AND ALGORITHMS FOR MEDICAL APPLICATIONS

[5] IDI.International diabetes institute - diabetes research, education and care. 2007(10/30/2007).

[6] U. P. Indian, "Pima Indians Diabetes Data Set." vol.2008,2008.

[7] Lenarcik, A., Piasta, Z.: Rough Classifier. In: Ziarko, W. (eds.) Rough Sets, Fuzzy Sets and Knowledge Discovery, pp. 298--316. Springer, London (1994).

[8] Ruggieri, S.: Efficient C4.5. In: IEEE Transactions on Knowledge and Data Engineering, pp. 438--444. IEEE Educational Activities Department, USA (2002).

[9] Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco (2001).

[10] Azuraliza Abu Bakar, Zulaiha Ali Othman, Abdul Razak Hamdan , Rozianiwati Yusof, Ruhaizan Ismail, 'Agent Based Data Classification Approach for Data Mining', 2008 IEEE

[11] Cuong Tong, Dharmendra Sharma and Fariba Shadabi, 'A Multi-Agents Approach to Knowledge Discovery', 2008 IEEE.

[12] Li Zhan, Liu Zhijing,' Web Mining Based On Multi-Agents ', COMPUTER SOCIETY,IEEE(2003)

[13] M. Wooldridge, An Introduction to MultiAgent Systems. John Wiley & Sons Ltd, 2002. John Wiley & Sons,2002.

[14] IDI. International diabetes institute - diabetes research, education and care.