INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATION & MANAGEMENT

IJRCM

IJRCM

A Monthly Double-Blind Peer Reviewed (Refereed/Juried) Open Access International e-Journal - Included in the International Serial Directories

Indexed & Listed at:

Ulrich's Periodicals Directory ©, ProQuest, U.S.A., EBSCO Publishing, U.S.A., Cabell's Directories of Publishing Opportunities, U.S.A.,

Open J-Gage, India [link of the same is duly available at Inflibnet of University Grants Commission (U.G.C.)],

Index Copernicus Publishers Panel, Poland with IC Value of 5.09 & number of libraries all around the world.

Circulated all over the world & Google has verified that scholars of more than 4255 Cities in 176 countries/territories are visiting our journal on regular basis.

Ground Floor, Building No. 1041-C-1, Devi Bhawan Bazar, JAGADHRI – 135 003, Yamunanagar, Haryana, INDIA

http://ijrcm.org.in/

CONTENTS

CALL FOR MANUSCRIPTS

We invite unpublished novel, original, empirical and high quality research work pertaining to recent developments & practices in the areas of Computer Science & Applications; Commerce; Business; Finance; Marketing; Human Resource Management; General Management; Banking; Economics; Tourism Administration & Management; Education; Law; Library & Information Science; Defence & Strategic Studies; Electronic Science; Corporate Governance; Industrial Relations; and emerging paradigms in allied subjects like Accounting; Accounting Information Systems; Accounting Theory & Practice; Auditing; Behavioral Accounting; Behavioral Economics; Corporate Finance; Cost Accounting; Econometrics; Economic Development; Economic History; Financial Institutions & Markets; Financial Services; Fiscal Policy; Government & Non Profit Accounting; Industrial Organization; International Economics & Trade; International Finance; Macro Economics; Micro Economics; Rural Economics; Co-operation; Demography: Development Planning; Development Studies; Applied Economics; Development Economics; Business Economics; Monetary Policy; Public Policy Economics; Real Estate; Regional Economics; Political Science; Continuing Education; Labour Welfare; Philosophy; Psychology; Sociology; Tax Accounting; Advertising & Promotion Management; Management Information Systems (MIS); Business Law; Public Responsibility & Ethics; Communication; Direct Marketing; E-Commerce; Global Business; Health Care Administration; Labour Relations & Human Resource Management; Marketing Research; Marketing Theory & Applications; Non-Profit Organizations; Office Administration/Management; Operations Research/Statistics; Organizational Behavior & Theory; Organizational Development; Production/Operations; International Relations; Human Rights & Duties; Public Administration; Population Studies; Purchasing/Materials Management; Retailing; Sales/Selling; Services; Small Business Entrepreneurship; Strategic Management Policy; Technology/Innovation; Tourism & Hospitality; Transportation Distribution; Algorithms; Artificial Intelligence; Compilers & Translation; Computer Aided Design (CAD); Computer Aided Manufacturing; Computer Graphics; Computer Organization & Architecture; Database Structures & Systems; Discrete Structures; Internet; Management Information Systems; Modeling & Simulation; Neural Systems/Neural Networks; Numerical Analysis/Scientific Computing; Object Oriented Programming; Operating Systems; Programming Languages; Robotics; Symbolic & Formal Logic; Web Design and emerging paradigms in allied subjects.

Anybody can submit the **soft copy** of unpublished novel; original; empirical and high quality **research work/manuscript** ***anytime*** in ***M.S. Word format*** after preparing the same as per our **GUIDELINES FOR SUBMISSION**; at our email address i.e. infoijrcm@gmail.com or online by clicking the link **online submission** as given on our website (***FOR ONLINE SUBMISSION, CLICK HERE***).

GUIDELINES FOR SUBMISSION OF MANUSCRIPT

1. **COVERING LETTER FOR SUBMISSION**:

DATED: _____________

THE EDITOR

IJRCM

Subject: **SUBMISSION OF MANUSCRIPT IN THE AREA OF _______________________.**

(e.g. Finance/Mkt./HRM/General Mgt./Engineering/Economics/Computer/IT/ Education/Psychology/Law/Math/other, please specify)

DEAR SIR/MADAM

Please find my submission of manuscript entitled '____' for possible publication in one of your journals.

I hereby affirm that the contents of this manuscript are original. Furthermore, it has neither been published elsewhere in any language fully or partly, nor is it under review for publication elsewhere.

I affirm that all the co-authors of this manuscript have seen the submitted version of the manuscript and have agreed to their inclusion of names as co-authors.

Also, if my/our manuscript is accepted, I agree to comply with the formalities as given on the website of the journal. The Journal has discretion to publish our contribution in any of its journals.

NAME OF CORRESPONDING AUTHOR :

Designation :

Institution/College/University with full address & Pin Code :

Residential address with Pin Code :

Mobile Number (s) with country ISD code :

Is WhatsApp or Viber active on your above noted Mobile Number (Yes/No) :

Landline Number (s) with country ISD code :

E-mail Address :

Alternate E-mail Address :

Nationality :

NOTES:

a) The whole manuscript has to be in ***ONE MS WORD FILE*** only, which will start from the covering letter, inside the manuscript. ***pdf. version is liable to be rejected without any consideration***.

b) The sender is required to mention the following in the **SUBJECT COLUMN of the mail**:

New Manuscript for Review in the area of (e.g. Finance/Marketing/HRM/General Mgt./Engineering/Economics/Computer/IT/ Education/Psychology/Law/Math/other, please specify)

c) There is no need to give any text in the body of mail, except the cases where the author wishes to give any **specific message** w.r.t. to the manuscript.

d) The total size of the file containing the manuscript is expected to be below **1000 KB**.

e) **Abstract alone will not be considered for review** and the author is required to submit the **complete manuscript** in the first instance.

f) ***The journal gives acknowledgement w.r.t. the receipt of every email within twenty four hours*** and in case of non-receipt of acknowledgment from the journal, w.r.t. the submission of manuscript, within two days of submission, the corresponding author is required to demand for the same by sending a separate mail to the journal.

g) The author (s) name or details should not appear anywhere on the body of the manuscript, except the covering letter and the cover page of the manuscript, in the manner as mentioned in the guidelines.

2. **MANUSCRIPT TITLE**: The title of the paper should be **bold typed**, **centered** and **fully capitalised**.

3. **AUTHOR NAME (S) & AFFILIATIONS**: Author (s) **name**, **designation**, **affiliation** (s), **address**, **mobile/landline number** (s), and **email/alternate email address** should be given underneath the title.

4. **ACKNOWLEDGMENTS**: Acknowledgements can be given to reviewers, guides, funding institutions, etc., if any.

5. **ABSTRACT**: Abstract should be in **fully italicized text**, ranging between **150** to **300 words**. The abstract must be informative and explain the background, aims, methods, results & conclusion in a **SINGLE PARA**. ***Abbreviations must be mentioned in full***.

6. **KEYWORDS**: Abstract must be followed by a list of keywords, subject to the maximum of **five**. These should be arranged in alphabetic order separated by commas and full stop at the end. All words of the keywords, including the first one should be in small letters, except special words e.g. name of the Countries, abbreviations.

7. **JEL CODE**: Provide the appropriate Journal of Economic Literature Classification System code (s). JEL codes are available at www.aeaweb.org/econlit/jelCodes.php, however, mentioning JEL Code is not mandatory.

8. **MANUSCRIPT**: Manuscript must be in ***BRITISH ENGLISH*** prepared on a standard A4 size ***PORTRAIT SETTING PAPER***. *It should be free from any errors i.e.* ***grammatical, spelling*** *or* ***punctuation.*** *It must be thoroughly edited at your end*.

9. **HEADINGS**: All the headings must be bold-faced, aligned left and fully capitalised. Leave a blank line before each heading.

10. **SUB-HEADINGS**: All the sub-headings must be bold-faced, aligned left and fully capitalised.

11. **MAIN TEXT**:

THE MAIN TEXT SHOULD FOLLOW THE FOLLOWING SEQUENCE:

INTRODUCTION
REVIEW OF LITERATURE
NEED/IMPORTANCE OF THE STUDY
STATEMENT OF THE PROBLEM
OBJECTIVES
HYPOTHESIS (ES)
RESEARCH METHODOLOGY
RESULTS & DISCUSSION
FINDINGS
RECOMMENDATIONS/SUGGESTIONS
CONCLUSIONS
LIMITATIONS
SCOPE FOR FURTHER RESEARCH
REFERENCES
APPENDIX/ANNEXURE

The manuscript should preferably range from ***2000 to 5000 WORDS***.

12. **FIGURES & TABLES**: These should be simple, crystal **CLEAR**, **centered**, **separately numbered** & self explained, and **titles must be above the table/figure**. **Sources of data should be mentioned below the table/figure**. *It should be ensured that the tables/figures are referred to from the main text.*

13. **EQUATIONS/FORMULAE**: These should be consecutively numbered in parenthesis, horizontally centered with equation/formulae number placed at the right. The equation editor provided with standard versions of Microsoft Word should be utilised. If any other equation editor is utilised, author must confirm that these equations may be viewed and edited in versions of Microsoft Office that does not have the editor.

14. **ACRONYMS**: These should not be used in the abstract. The use of acronyms is elsewhere is acceptable. Acronyms should be defined on its first use in each section: Reserve Bank of India (RBI). Acronyms should be redefined on first use in subsequent sections.

15. **REFERENCES**: The list of all references should be alphabetically arranged. ***The author (s) should mention only the actually utilised references in the preparation of manuscript*** and they are supposed to follow Harvard Style of Referencing. **Also check to make sure that everything that you are including in the reference section is duly cited in the paper**. The author (s) are supposed to follow the references as per the following:

- All works cited in the text (including sources for tables and figures) should be listed alphabetically.
- Use (**ed.**) for one editor, and (**ed.s)** for multiple editors.
- When listing two or more works by one author, use --- (20xx), such as after Kohl (1997), use --- (2001), etc, in chronologically ascending order.
- Indicate (opening and closing) page numbers for articles in journals and for chapters in books.
- The title of books and journals should be in italics. Double quotation marks are used for titles of journal articles, book chapters, dissertations, reports, working papers, unpublished material, etc.
- For titles in a language other than English, provide an English translation in parenthesis.
- ***Headers, footers, endnotes*** *and* ***footnotes*** *should* ***not be used*** *in the document.* However, **you can mention short notes to elucidate some specific point**, which may be placed in number orders after the references.

PLEASE USE THE FOLLOWING FOR STYLE AND PUNCTUATION IN REFERENCES:

BOOKS

- Bowersox, Donald J., Closs, David J., (1996), "Logistical Management." Tata McGraw, Hill, New Delhi.
- Hunker, H.L. and A.J. Wright (1963), "Factors of Industrial Location in Ohio" Ohio State University, Nigeria.

CONTRIBUTIONS TO BOOKS

- Sharma T., Kwatra, G. (2008) Effectiveness of Social Advertising: A Study of Selected Campaigns, Corporate Social Responsibility, Edited by David Crowther & Nicholas Capaldi, Ashgate Research Companion to Corporate Social Responsibility, Chapter 15, pp 287-303.

JOURNAL AND OTHER ARTICLES

- Schemenner, R.W., Huber, J.C. and Cook, R.L. (1987), "Geographic Differences and the Location of New Manufacturing Facilities," Journal of Urban Economics, Vol. 21, No. 1, pp. 83-104.

CONFERENCE PAPERS

- Garg, Sambhav (2011): "Business Ethics" Paper presented at the Annual International Conference for the All India Management Association, New Delhi, India, 19–23

UNPUBLISHED DISSERTATIONS

- Kumar S. (2011): "Customer Value: A Comparative Study of Rural and Urban Customers," Thesis, Kurukshetra University, Kurukshetra.

ONLINE RESOURCES

- Always indicate the date that the source was accessed, as online resources are frequently updated or removed.

WEBSITES

- Garg, Bhavet (2011): Towards a New Gas Policy, Political Weekly, Viewed on January 01, 2012 http://epw.in/user/viewabstract.jsp

KNOWLEDGE DISCOVERY IN DATABASES

ANANT KUMAR
LECTURER
NOBLES PG COLLEGE
RAMGARH

ABSTRACT

Knowledge discovery in databases((Data Mining) is a rapidly growing field, whose development is driven by strong research interests as well as urgent practical, social, and economical needs. While the last few years knowledge discovery tools have been used mainly in research environment. Sophisticated software products are now rapidly emerging. In this, we provide an overview of common knowledge discovery tasks and approaches to solve these tasks. We propose a feature classification scheme that can be used to study knowledge and data mining software. This scheme is based on the software's general characteristics, database connectivity and data mining characteristics. Finally, we specify features that we consider important for knowledge discovery software to possess in order to accommodate its users effectively. This research work, first of all, focuses on analyzing different processes of supervised learning. Secondly, it proposes a new improved process for developing knowledge extraction.

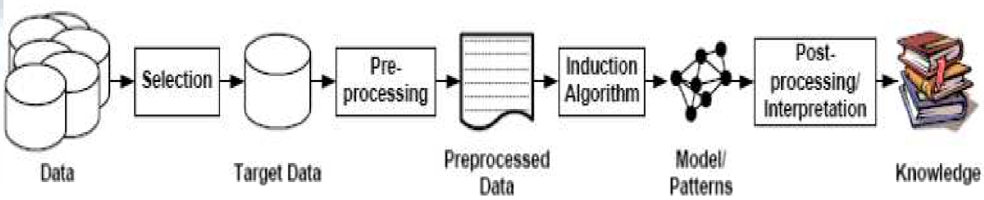
KEYWORDS
Data Mining, Knowledge Discovery in Databases, Supervised learning algorithms, Stacking, Classification, Regression.

INTRODUCTION

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction. Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association (patterns where one event is connected to another event, such as purchasing a pen and purchasing paper), sequence or path analysis (patterns where one event leads to another event, such as the birth of a child and purchasing diapers), classification (identification of new patterns, such as coincidences between duct tape purchases and plastic sheeting purchases), clustering (finding and visually documenting groups of previously unknown facts, such as geographic location and brand preferences), and forecasting (discovering patterns from which one can make reasonable predictions regarding future activities, such as the prediction that people who join an athletic club may take exercise classes).As an application, compared to other data analysis applications, such as structured queries (used in many commercial databases) or statistical analysis software, data mining represents a difference of kind rather than degree. Many simpler analytical tools utilize verification-based approach, where the user develops a hypothesis and then tests the data to prove or disprove the hypothesis. For example, a user might hypothesize that a customer who buys a hammer, will also buy a box of nails. The effectiveness of this approach can be limited by the creativity of the user to develop various hypotheses, as well as the structure of the software being used. In contrast, data mining utilizes a discovery approach, in which algorithms can be used to examine several multidimensional data relationships simultaneously, identifying those that are unique or frequently represented. For example, a hardware store may compare their customers' tool purchases with home ownership, type of automobile driven, age, occupation, income, and/or distance between residence and the store. As a result of its complex capabilities, two precursors are important for a successful data mining exercise; a clear formulation of the problem to be solved, and access to the relevant data. Data mining, popularly known as Knowledge Discovery in Databases (KDD), it is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. Though, data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

A lot of research efforts are performed around the world to find unknown facts hidden in millions of records, stored in huge databases. As databases are growing in size, there is a need to explore new patterns from existing database(s) and there is a need to develop new techniques that are efficient enough to find such patterns and should also be able to fulfill future needs of the companies demanding such solutions. There are different perspectives, why we search for such new patterns e.g. sales/marketing, customer retention, buyer behaviour, cost/utilization, quality control, inventory and fraud detection *etc.* Most important among them is marketing department(s) of the organizations, who are always looking for such patterns, which can furnish new prospects for marketing managers to think upon new trends, aspects and emerging scopes to take decisions at different managerial levels. Data mining techniques can be categorized in different ways. One of the most important categorization is based on kind of problems solved. Classification, Clustering and Association rules are most important kind of problems that are solved by data mining. Present work concentrates upon classification tasks. A large number of techniques and algorithms have been devised for extracting knowledge from databases. A large part of these techniques and algorithms concentrate upon classification problems. Classification task comprises of assigning objects to classes (groups) on the basis of measurements made on the objects. Classes of data are predefined, a set (the train set) of labeled objects are used to form a model through classifier for classification of future observations (or the test set). Such a processing is named as supervised learning. Supervised learning techniques are usually used for the solution of classification problems. Usually a general process is recommended for supervised learning as shown in Figure[1].

FIGURE 1: THE GENERAL KDD PROCESS



LITERATURE REVIEW

The amount of data continues to grow at an enormous rate even though the data stores are already vast. The primary challenge is how to make the database a competitive business advantage by converting seemingly meaningless data into useful information. How this challenge is met is critical because companies are increasingly relying on effective analysis of the information simply to remain competitive. A mixture of new techniques and technology is emerging to help sort through the data and find useful competitive data. By knowledge discovery in databases, interesting knowledge, regularities, or high-level information can be extracted from the relevant sets of data in databases. It is investigated from different angles, and large databases thereby serve as rich and reliable sources for knowledge generation and verification. Mining information and knowledge from large database has been recognized by many researchers as a key research topic in database systems and machine learning. Companies in many industries also take knowledge discovering as an important area with an opportunity of major revenue (Fayyad., 1996. Piatetsky-Shapiro.1991, Silberschatz.,1995). The discovered knowledge can be applied to information management, query processing, decision making, process control, and many other applications.

For example, Fayyad stated KDD process as follows (1996):

1. Learning the application domain 2.Creating a target dataset 3.Data cleaning and preprocessing 4.Data reduction and projection 5.Choosing the function of data mining 6.Choosing the data mining algorithm(s) 7.Data mining 8.Interpretation 9.Using the discovered knowledge

As the KDD process shows, data mining is the central of knowledge discovering, it requires complicated data preparation work. Data cleaning and preprocessing includes basic operations, such as removing noise or outliers, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time sequence information and known changes, as well as deciding DBMS issues, such as data types, schema, and mapping of missing and unknown values. Useful data are chosen from the formatted data to increase the effectiveness and focus on the task.

SUPERVISED LEARNING AND CLASSIFICATION

Different studies have been conducted from time to time for accounting the development of supervised learning techniques and algorithms. Data mining itself has emerged from other disciplines like Machine Learning, Artificial Intelligence and Statistics etc. Hence, it is obvious to get initial references related to comparisons of algorithms from its parent disciplines. Many researches were being performed before the time data mining was coined as a separate discipline for study. Fahrmeir et al., compared the results of a point awarding approach with the results obtained by the linear discriminate[4]. Gorman et al., reported that back-propagation outperformed nearest neighbour for classifying sonar targets[5], whereas Shadmehr et al., showed that some Bays algorithms were better on other tasks[118]. Kirkwood et al., developed a symbolic algorithm ID3, which performed better than discriminant analysis for classifying the gait cycle of artificial limbs[7].

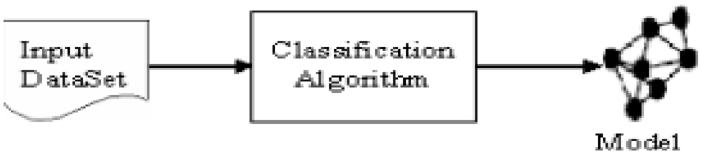
VARIOUS PROCESSES FOR SUPERVISED LEARNING

Supervised learning processes can vary from simple processing to very complex processing. Different techniques and algorithms are used to extract knowledge from data. These algorithms involve certain criteria to extract knowledge. A particular technique or algorithm may be suitable for a segment of problems only, but may not be suitable for others. There is no unique technique or algorithm that solves all types of problems. Supervised learning involves the use of the train set to train algorithms for the creation of a model of that technique or algorithm. This model is then applied on the test set to generate and compare results. Different supervised learning processes as discussed by Witten et al.[8], are as follows:

(i) SIMPLE SUPERVISED LEARNING

In its simplest form, input data is applied to classification algorithm and result is generated. It is also called Fixed Split experimentation as data is divided into the train and the test set. Such experimentation suffers with over fitting and under fitting model and results may not fulfill the reliability criteria. So there is need for pre-processing and post-processing of data.

FIGURE 2: SIMPLE SUPERVISED LEARNING PROCESS



(ii) PREPROCESSING OF THE DATA

A data set collected is not directly suitable for induction (knowledge acquisition); it comprises in most of cases noise, missing values, the data are not consistent, the data set is too large, and so on. Therefore, we need to minimize the noise in data, choose a strategy for handling missing (unknown) attribute values, use any suitable method for selecting and ordering attributes (features) according to their informative (so-called attribute mining), discretize/fuzzify numerical (continuous) attributes and eventually, process continuous classes.

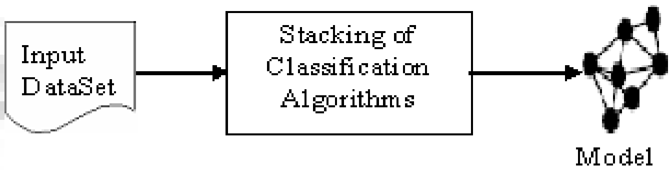
(iii) POST PROCESSING OF THE KNOWLEDGE DERIVED

The pieces of knowledge extracted in the previous step could be further processed. One option is to simplify the extracted knowledge. Also, we can evaluate the extracted knowledge, visualize it, or merely document it for the end user. There are various techniques to do this. Next, we may interpret the knowledge and incorporate it into an existing system, and check for potential conflicts with previously induced knowledge.

(iv) STACKING

Stacking combines the output of a number of classifiers. Stacked generalization, also known as Stacking in the literature. It is a method that combines multiple classifiers by learning the way that their output correlates with the true class on an independent set of instances. At a first step, N classifiers Ci, i = 1..N are induced from each of N data sets Di, i = 1..N. Then, for every instance ej , j = 1..L of an evaluation set E, independent of the Di data sets, the output of the classifiers Ci(ej) along with the true class of the instance class(ej) is used to form an instance mj , j = 1..L of a new data set M, which will then serve as the meta-level train set. Each instance will be of the form: C1(ej), C2(ej), . . . , CN(ej), class(ej). Finally, a global classifier GC is induced directly from M. If a new instance appears for classification, the output of all local models is first calculated and then propagated to the global model, which outputs the final result. Any algorithm suitable for classification problems can be used for learning the Ci and GC classifiers. Independence of the actual algorithm used for learning Ci, is actually one of the advantages of Stacking, as not every algorithm might be available for each data set and not the same algorithm performs best for every data set.

FIGURE 3: STACKED SUPERVISED LEARNING PROCESS.



COMPLEX PROCESSING

Different pre-processing, Post-processing and stacking of different algorithms may be combined to extract knowledge from databases. Such complex criteria may involve parallel processing of different algorithms as well.

OUTLINE OF THE METHOD

It is often observed that human beings deal with new and old situations or problems which they face, by first composing relevant queries from the information available about the situation or the problem at hand and its domain, and then posing these queries to the reserved knowledge in their minds. They then use their reasoning techniques to collect the relevant knowledge and try to make links among this knowledge in a way to help them dealing with the situation.

Reasoning is the process of drawing new conclusions from some premises which are known facts and/or assumed hypotheses. In general, a reasoning comprises a number of arguments or inferences. In any valid argument of human logical thinking, the premises must be relevant to the conclusion. Informally, we say a reasoning is ***relevant*** if and only if in every argument or inference of that reasoning the premises is relevant to the conclusion; a reasoning is ***irrelevant*** if and only if it **is** not relevant.

The previous observation inspires us to propose a new integration method to integrate the already exist domain knowledge bases into the KDD process. The key idea behind the proposed method, which is acquired from the previous observation, is that the relevancy among knowledge is dynamically changing and the most important factor which decides it is the situation or the problem at hand. Some knowledge in the knowledge base are relevant to each other in the current situation but are not relevant in another one. Therefore, in our proposed method the characteristics of relevant knowledge are identified from the application information and the database meta-knowledge (e.g.biases in data collection and data-base schema). Then, these characteristics are compiled automatically to a set of queries posed to the knowledge base. In order to solve the relevancy problem and assure the generality of the method, the queries and the knowledge base have to be rendered to a domain-independent form which both permits automated relevant reasoning and overcomes the language gab exist between the terminology of the knowledge base and of the queries' description. In the method we suggest that this form will be the relevant logic form and the reasoning engine will be the relevant reasoning based on strong relevant logic, which **is** proposed by Cheng recently [9]**.**

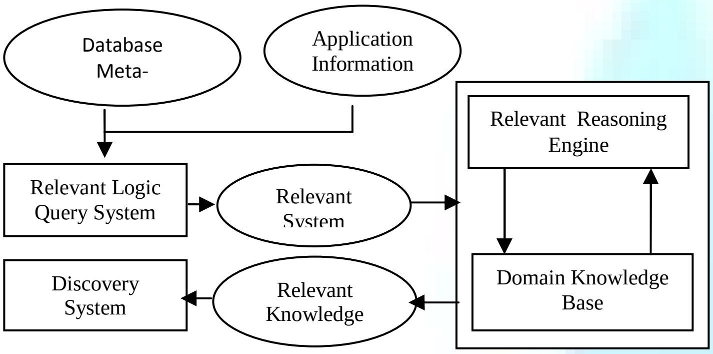
The method is shown in Figure, it consists exactly of three phases:

Phase 1: Specify the characteristics of the relevant knowledge by using the KDD task or application at hand and the database meta-knowledge.

Phase 2: Design relevant queries from the above characteristics in the form of relevant logic language.

Phase 3: Pose these queries to the domain knowledge base and use the relevant reasoning engine to collect the relevant knowledge.

FIGURE 4



Relevant reasoning based on Strong Relevant Logic(SRL) is preferred rather than any other reasoning techniques, e.g., those which are based on Classical Mathematical Logic (CML), because our aim is to get the relevant Knowledge automatically but the conclusions reasoned and/or deduced based on CML from the premises may be irrelevant at all, in the sense of meaning, to the premises. In the framework of the CML, there is no guarantee that the conclusion of a reasoning is necessarily relevant to its premises, even if the reasoning is valid in the sense of CML. As a result, for a conclusion reasoned and/or deduced based on the CML, we have to investigate whether it is relevant to its premises or not by ourselves. Obviously, the more information and/or knowledge we have, the more difficult investigation task we do. SRL gives us high level evidences that the reached conclusions are relevant to the given premises.

AN APPLICATION OF THE METHOD

Many of the existing KDD systems produce knowledge in the form 'If ... Then . . . ' represented as attribute-value rules[10] or relational rules[11] based on the similarities and differences between vectors of attributes. These attributes are static and exist in the initial description language of the examples in the database. In this section we present an algorithm designed under the framework of the proposed method. The goal of this algorithm is to use the prior domain knowledge bases where the knowledge is represented in the form of propositional rules to create new k relevant attributes, not already present in the attributes originally exist, for KDD systems to use. This algorithm works by using the attributes originally exist in the database schema to design relevant queries in the form of propositional rules and use the relevant reasoning engine to retrieve all their relevant rules which define new attributes. The new attributes are used to augment the description language a standard KDD system can operate with.

Algorithm (Task): Creating new k relevant attributes from the prior domain knowledge bases where the knowledge are represented in the form of propositional rules.

Input: An array of the old attributes A_l, ..., An,; and the required number k of new attributes.

Output: Array of the old and new attributes A_1, ..., A_n, $_{n+i}$....,A_{n+k}.

Method: The algorithms proceed in the following manner. For each encountered old attribute A, in the attributes array a relevant query ***SQ,*** in the form of propositional rule, is designed from it. In general,

suppose SQ = $A_i \rightarrow A_i$ is the most recently designed query. The relevant reasoning engine takes this query (rule) and finds all its relevant rules from the knowledge base. There are two forms of relevant rules important to the task, namely,

- A_i (=,< >,<,>)value → B(=,< >,<,>)value
- A_i (=,< >,<,>)value **^** H(=,< >,<,>)value
→ B(=,<>,<,>)value

where H is an attribute in the schema (new or old),
this can be tested by comparing H with the attributes
A_{i+1},...,A_f only, where A_f is the last constructed attribute. The previous rule's forms are produced from the reasoning engine as a result of applying the two inference rules Modus Ponen and Adjunction on the old and most recently constructed attributes. The algorithm is continued by selecting a rule from the list of ***SQ's*** relevant rules. If the rule has been previously selected, a new one is selected. When the new rule defines a new attribute, i.e., **(**B(=,<>,<,>)value) is new, this attribute is constructed, and a new SQ= **(**B(=,<>,<,>)value →B(=,<>,<,>)value) is designed from it. The algorithm starts again from the new SQ. After trying all the relevant rules of the new SQ, the algorithm return to the list of (A_i + A_i **)'s** relevant rules. This process of selecting a rule is continued until the list of these rules is exhausted. The overall process repeatedly works until the required attributes k are defined for most of the examples in the database or the last old attribute is reached in the array. The method is summarized as follows:

Step 1: Initialization: ***i*** := *1* (**i** is a counter on the old
attributes); l := 0 (l is a counter on the completely defined new attributes for most examples); ***f*** := *n+l;*

Step 2: REPEAT
SQ (Starting query):= $A_i \rightarrow A_l$
Relevance (SQ);
i := *i+l;*
UNTIL
l :=*k* OR *i:=n.*

The procedure Relevance is described as follows:

PROCEDURE Relevance (RULE: string);
begin
Find all the rules which are relevant to RULE;
FOR each rule R relevant to the task **DO**
CASE (R consequent variable) **OF**
New attribute and not constructed : {
CV := R antecedent;
A_f := R consequent variable;
f:= *f*+1
Assign R's consequent to all the examples
that satisfy CV;
SQ := R consequent → R consequent;
Relevance ***(SQ)*** }
New attribute and constructed before :{
CV := R antecedent;
Assign Rs consequent to all the examples
that satisfies CV;
IF (R consequent variable (=,<>,<,>)
value) is new value
THEN {
SQ := R consequent → R consequent;
Relevance ***(SQ)}}***
end CASE
end; FOR
IF most of the examples in the database is defined for the new attribute A_{n+1}
THEN ***l*** *:= l +1;*
end

Working Example: To illustrate how the algorithm works, consider the attributes list with their characteristics given in Table 1 and the rules given in Figure **.** These rules are constructed and given to the induction system CHARADE[12] by the knowledge shown in Table 2**.** Let these rules be exist in the domain knowledge base in a scattered form, Table **3** shows how the algorithm finds these rules and directly constructs new attributes from them. The column SQ's RELEVANT RULES in Table 3 contains all the relevant rules to the task. The assignment operations can be carried out on the examples given in [12].

TABLE 1: LIST OF ATTRIBUTES WITH THEIR CHARACTERISTICS

ATTRIBUTE	TYPE	DOMAINE
Humidity(H)	ordered set	low < high < very-high
Hygiene(h)	ordered set	very-bad < bad < average < good
Location(L)	string	NA
F_fruits/vegt.	Boolean	yes, no
Year(Y)	integer	NA
Temperature(T)	ordered set	severe-cold < cold < average
< hot < very-hot		
food-quantity (FQ)	ordered set	starvation<severe – restrictions < restrictions<ok
food-variety (FV)	ordered set	low < average < high
type-of-location (TL)	unordered set	land, sea
affection-severity(AS)	integer interval	[0,5]

IF (humidity = high)**THEN** (p e r s p i r a t i o n >= hard)

IF (hygiene = good) (humidity <= high) **THEN** (p e r s p i r a t i o n <= hard)

IF (humidity >= very--high) **THEN** (p e r s p i r a t i o n >= blocked)

IF (p e r s p i r a t i o n <= hard) **THEN** (f l u i d s <= healthy)

IF(fresh-fruits/vegetables = yes) **THEN** (f l u i d s <= healthy)

IF(fresh-fruits/vegetables <> yes) **THEN** (f l u i d s >= corrupted)

IF (hygiene <= average) (l o c a t i o n = sea) **THEN** (humidity >= very-high)

IF (hygiene = good) **THEN** (humidity <= high) **THEN** (humidity <= high)

TABLE 2: DEFINITIONS OF THE TWO NEW ATTRIBUTES

ATTRIBUTE	TYPE	DOMAINE
perspiration (P)	ordered set	normal < hard < blocked
fluids (P)	ordered set	healthy < corrupted

TABLE 3

ITERATION OF STEP 2	SQ	SQ' RELEVANT RULES	NEW ATTRIBUTES	ATTRIBUTES DOMAIN
1	H→H	H=high→ p>=hard	p	p>=hard
		H=very high p>=blocked		p>=blocked
		H<=high ∧ H>=good→p<=hard		p<=hard
	p>=hard→ p>=hard	no		
	p>=blocked→ p>=blocked	p>=blocked ∧F/V yes →+ F>=corrupted	F	F>=corrupt
	F>=corrupt→ F>=corrupt	no		
	P<=hard→p<=hard	p<=hard→F<=healthy		F<=healthy
2	h→h	h<=average ∧ L=sea →+ H>=very high h>=good -+ H<=high		
3	L→L	no		
4	F/V→F/V	F/V=yes →+ F<=healthy		

CONCLUSION

In this paper a new and general integration method for integrating the prior domain knowledge bases into the KDD process based on relevant reasoning to solve the relevancy problem and simulate the human way of thinking when one face a new or old situation. According to human being thinking, the principle factor which decides the relevancy among his knowledge is the situation or the problem at hand. The same idea is used by the integration method in order to increase the power and versatility of the KDD systems. This versatility is achieved by searching for and getting the relevant prior knowledge according to the application and KDD task at hand. This integration method can be considered as a step toward a full automatic discovery (including KDD) systems in the future. All discoveries made by current computer programs have been characterized as human/computer discoveries because the discovery process is far from being completely automated [13]. As we see one area where the human component has been vital is in guiding the KDD systems based on the relevant prior knowledge. Here an algorithm designed under the framework of the proposed method. This algorithm uses the prior domain knowledge bases where the knowledge are represented in the form of propositional rules to create new ***k*** relevant attributes, not already present in the attributes originally exist, for KDD systems to use. This algorithm can be integrated with any inductive learning methodology to increase its autonomy and improve learning.

REFERENCES

1. Fayyad U., Piatetsky-Shapiro G. and Smyth P.,(1996). “The KDD process for extracting useful knowledge from volumes of data”, *CACM 39 11*, 27-34.
2. Piatetsky-Shapiro Gregory, (1991). "Knowledge discovery in real databases", *IJCAI-89 Workshop , AI Magazine,* 11.
3. Silberschatz A., and Tuzhilin A., (1995). “On subjective measures of interest-ingness in knowledge discovery”.
4. Fahrmeir L., Haussler W. and Tutz G.,(1984). “Diskriminanz analyse”, *Multivariate statistische Verfahren, Verlag de Gruyter, Berlin*.
5. Gorman R. P. and Sejnowski T. J.,(1988). “Analysis of hidden units in a layered network trained to classify sonar targets”, *Neural networks,* 1(1), 75–89.
6. Rosenblatt,F.,(1958).“The perceptron:a probabilistic model for information storage and organization in the brain”, *Psychological Review*, 65, 386-408.
7. Kirkwood, C., Andrews B. and Mowforth P., (1989). “Automatic detection of gait events: a case study using inductive learning techniques”, *Journal of biomedical engineering*, 511–516.
8. Witten I. H. and Frank E.,(2005). “Data Mining –Practical Machine Learning Tools and Techniques”, *Elsevier Inc*.
9. J. Cheng, “Relevant Reasoning As The Logical Bases of Knowledge Engineering,” In Proc., 4th World Congress on Expert Systems. Vol.1, 1998 pp. 449-456.
10. J. Han, Y. Fu, “Attribute-Oriented Induction in Data Mining,” In U.M. Fayyad, G.P. Shapiro, P. Smyth, R. Uthurusamy, editors, *Advances in Knowledge discovery and Data Mining* AAAI/MIT Press 1996, pp. 399-421.
11. W.M. Shen, **K.** Ong, B. Mitbander, C. Zaniolo, “Metaqueries for Data Mining,” In U.M. Fayyad, G.P. Shapiro, P. Smyth, R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining* AAAI/MIT Press 1996, pp. 375-398.
12. V. Cormble, J.G. Ganascia, “Discovery of The Causes of Scurvy: Could Artificial Intelligence have saved a few enturies?,” In *Knowledge Discovery in Databases:* Papers from the 1993 *AAAI* Workshop, *AAAI* Press 1993, pp. 70-80.
13. R.Valdes-Perez, “Some Recent Human-Computer Discoveries in Science and What Accounts for Them,” ***AI*** *Magazine,* Vol.16. No.3, pp. 37-44.

REQUEST FOR FEEDBACK

Dear Readers

At the very outset, International Journal of Research in Computer Application & Management (IJRCM) acknowledges & appreciates your efforts in showing interest in our present issue under your kind perusal.

I would like to request you tosupply your critical comments and suggestions about the material published in this issue as well as on the journal as a whole, on our E-mail**infoijrcm@gmail.com** for further improvements in the interest of research.

If youhave any queries please feel free to contact us on our E-mail **infoijrcm@gmail.com**.

I am sure that your feedback and deliberations would make future issues better – a result of our joint effort.

Looking forward an appropriate consideration.

With sincere regards

Thanking you profoundly

Academically yours

Sd/-
Co-ordinator

DISCLAIMER

ABOUT THE JOURNAL

In this age of Commerce, Economics, Computer, I.T. & Management and cut throat competition, a group of intellectuals felt the need to have some platform, where young and budding managers and academicians could express their views and discuss the problems among their peers. This journal was conceived with this noble intention in view. This journal has been introduced to give an opportunity for expressing refined and innovative ideas in this field. It is our humble endeavour to provide a springboard to the upcoming specialists and give a chance to know about the latest in the sphere of research and knowledge. We have taken a small step and we hope that with the active co-operation of like-minded scholars, we shall be able to serve the society with our humble efforts.

Our Other Journals