Comparative Study of Data Mining Tools and Analysis with Unified Data Mining Theory

Harshvardhan Solanki
Department of Computer Science Engineering
Gyan Vihar School of Engineering and Technology
Jaipur, Rajasthan, India

ABSTRACT
Today almost everyone has access to huge amount of data. Several wide spread organizations have their own large data repositories, data warehouses, which are still expanding with queries over data and the need for extraction of most beneficial data pattern and refined knowledge. This necessity is followed by the requirement of an apt data mining tool to help with decision making and query pre-processing. However, in this paper, a study will be presented of analysis of the selected tools to deal with the selection of the most apt tool for mining suitable for a particular data type. This research provides with the complete analysis of a tool regarding the features and functionalities offered by them. The tools are compared based on their specification and techniques and algorithms used in these tools along with the features it provides. This paper also introduces the Unification theory, one of the major issues with data mining and presents methodologies which can be used for formulation of this theory. At the end the shortcomings with these methodologies is also discussed.

Keywords
Data Mining tools, data classification, Unified data mining theory, Unified data mining process, Data mining with Multi Agent Systems, WEKA, Tanagra, KNIME.

1. INTRODUCTION
The advances in technology have pushed all the data from a local terminal to a remote terminal. Many of the organizations providing their services to the users are now managing a large data warehouse, information repositories, and several data marts for all similar data. With this large data resource, the manual analysis had become very arduous and led to the development of several automated multi-agent data miner or Data Mining Tools. Data Mining is the process of discovering interesting knowledge from large databases. Data mining is sometimes also referred as a part of knowledge discovery process (KDD). Data mining process involves several other tasks as well apart from just the extraction of information and analysis of data. These different tasks varies from data management aspects, and data pre-processing to generating new rules and measuring the interestingness, relativity, complexity considerations as well as signal and image processing, visualization and online updating of the database(Web Data Mining) [8]. Different tools use different algorithm base and techniques to carry out data mining tasks. Few of the popular data mining tools are Orange, WEKA (Waikato Environment for Knowledge Analysis), RapidMiner, KNIME. These tools provide a user friendly interface, visualization tools and modeling algorithms for the analysis of data and to aid with the ease of performing the task. All these tools are applicable to the mentioned processes and data analysis. Several of the functionalities provided by these tools include characterization and classification of data, patterns evaluation, associations and correlations, and prediction over the data. However, they do not facilitate for real time data challenges, that is, they do not provide for dynamic updating and real time pattern evolution. Whenever a new data is inserted to the database, each time the whole data is analyzed and pre-processed [1]. Although the requirement is for the real time pattern refinement, with these tools one has to manually schedule the process of data selection through information evaluation. And these tools also require some level of expertise for their use and so are not for the naïve.

These problems raise concern with following:

- Comparative study for selection of the most suitable tool for particular data domain.
- Selection of most appropriate algorithm for mining from particular data domain.
- A methodology to deal with manual implementation of mining algorithm with the data mining tools: A general framework.

The third of the above mentioned problem requires the implementation of unification theory of data mining. The present scenario of data mining involves the selection of a mining algorithm for each category of data or information. The selection of best suited algorithm of a particular data domain is vital to obtain most appropriate output of these algorithms. However, the selection of this best fit algorithm is trivial and also depends upon the knowledge of the user about that data domain, data type and algorithm required [5]. This is still a problem with many data mining tools as no unified theory has been adopted. Every time the task of classification or clustering is to be performed, the question of which algorithm is best suited arises. Present algorithms and techniques perform the consecutive tasks of classification separately from that of clustering. This bounds data mining tools to be compatible only with a particular application over the set of different data types for that application and to some problems defined exclusively for that data [5]. This has ultimately resulted in the requirement of a general framework that clearly defines the unification of different data mining tasks and hence would allow for development of a better data mining tool and a process which will be referred as unified data mining engine (UDME). This paper will discuss first the performance and functionalities of and issues with some popular data mining tools. The succeeding section will discuss formulation of unified theory and the extent of its applicability.

2. RELATED WORK

A considerable work has been done over this subject regarding the comparative study over few old tools and their versions. Several experts have proposed their work over different aspects of data mining tools. Various surveys and research had been conducted over the use of the most popular tool among the organizations. These studies were also conducted over different type of users of the tools and have revealed the flaws as well as some strong features with every tool.

The study conducted by Ralf Mikut represents some of the worth efforts taken for this categorization of tools based on several factors like target users, data organization and supported data and data structures, tools and services for data exploration, visualization and interface design and divided among nine different categories.. This categorization has also helped in the determination of the tool considered best for application of a particular data mining task [2].

Earlier suggestions by Carrier and Povel for tool classification were intended over the business objectives. A template for characterization was designed based on several dynamic sample databases of tools and other supporting attributes like services provided by system, business goals, and other features of processing data and user interface. Around 40 data mining tools were evaluated and a general schema was proposed for tool selection to achieve business goals [2].

The further research by King and Elder was done over by testing each selected tool with each one of the algorithm selected for the study. The study concluded that all of the tested data mining tools were similar to each other according to their performance but only affected by the type of dataset and methodology implemented for the classification and suggested the future work on clustering using different algorithms [3].

Quit a few attempts have been made to accomplish the actual implementation and functioning of Unified theory of data mining. Much of the work has been done regarding the core concepts of data mining to improve the preprocessing and to reduce the complexity measure of several data mining tools, and to increase efficiency of the tasks performed and enhance functionality of these tools [7]. These attempts made in order were to attain a general framework for the Data Mining process. Rather then using databases with only normal data, a data set of patterns of previous analysis was included in the conventional databases. This was followed by the observed reduction in the processing time of data as now there was no need to access data but was attained through querying patterns only [5]. This in turn simplified the whole knowledge mining process which has played a role to achieve functionality similar to that proposed by the Unifying theory of Data Mining.

3. THE COMPARATIVE STUDY

The comparative study of data mining tools is carried out by selecting some open source tools freely available on the internet and selecting of sample data sets from UCI machine learning repository. These data sets are then used with these tools in order to determine their performance and providing a complete analysis of their functionalities.

3.1 Data Mining Tools

Description of the data mining tools used in this comparative study.

3.1.1 Weka

Weka is the tool most commonly used due to its vast functionality and supported features. This java based data mining tool provides user with both GUI and simple CLI for performing and managing tasks to be performed. It supports all data mining tasks from preprocessing, classification, and clustering to visualization and feature selection. [9]

Figure 1: Weka GUI

3.1.2 Knime

KNIME (Konstanz Information Miner) is an open API workflow based data mining tool that provides easy accessibility to new nodes to be added into the workflow. It provides its user with the GUI which aid with the simplification of workflow generation by the user. It also provides with features to modify a particular node accordingly and execution of partial data flow [10].

3.1.3 Tanagra

This extension of SIPINA provides the users an easy to use interface for the analysis of either real or artificial data. It allows the researchers to easily add their own data mining research methodology or any newly identified data mining processing technique and also supports by providing them with architecture and a means to compare their methodology performances. It provides the beginners or naives with a platform where they can carry out their experimental procedures [6].

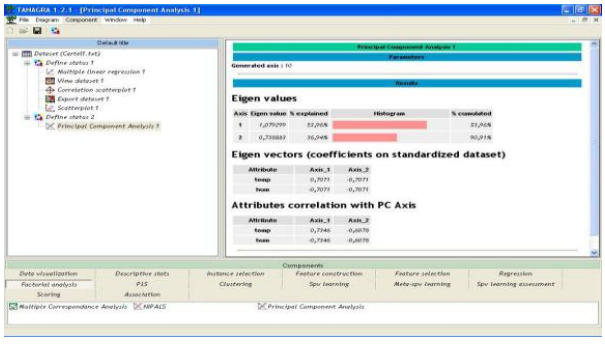

Figure 2: Tanagra explorer for result visualization

3.2 Experimental Analysis

The performance of these tools has been analyzed by first running them with several datasets available on UCI repository. Several different algorithms for classification and clustering were implemented in this analysis and performance of these algorithms was observed. In this section a sample of the experiment performed during the research is presented and conclusion of the results of different tools is discussed.

Data Set: Data set of (1)Audiology is used with data type multivariate, attribute type categorical, number of attributes 69, number of instances number of attributes, number of instances 226.

(2) Zoo: data type: multivariate, attribute type: categorical, integer, number of attributes: 18, number of instances: 101 [13].

Preliminaries: The classification was carried out of data set with percentage split methodology of 60% for training data and remaining 40% for the test data. The obtained measure of accuracy is used as the criterion for the performance analysis of the tools.

Experiment: The k nearest neighbor algorithm did not run on the data set, the possible reason being the incompatible data type as the data was of discrete values.

The accuracy measure of Weka with OneR is 42.85%, to 38.23%, for ZeroR it is 27.27% to 38.23%, for k-nearest neighbor it is 58.44% to 94.11%, for C4.5 it is 86.12% to 92.13%.

The KNIME does not provide any implementation for the ZeroR and OneR algorithms which proved to be a shortcoming in the accuracy of the tool. K-nearest neighbor is not compatible with the data type for Audiology but produced an accuracy of 41.54% for instances in Zoo data set. C4.5 achieved the accuracy between 62.55% and 89.36%.

Tanagra, similar to KNIME, does not provide with the implementation of ZeroR and OneR algorithm. K-nearest neighbor algorithm, like the case with other tools, does not conclude over the discrete values but runs well over the Zoo data set values. K nearest neighbor algorithm attains an accuracy percentage of 75.32% with percentage split. C4.5 obtains result between 74.38%and 80.51%.

Table 1: Accuracy for classification over audiology data set.

Accuracy% mapping	ZeroR	OneR	C4.5	KNN
WEKA	27.27%	42.85%	86.12%	58.44%
KNIME	NA	NA	62.55%	NA
TANAGRA	NA	NA	74.38%	NA

Table 2: Accuracy achieved over Zoo data set.

Accuracy% mapping	ZeroR	OneR	C4.5	KNN
WEKA	38.23%	38.23%	92.13%	94.11%
KNIME	NA	NA	89.36%	41.54%
TANAGRA	NA	NA	80.51%	75.32%

It has been observed that Weka has successfully run and implemented all the algorithms and produced appropriate results for the algorithms but with lowest accuracy that of ZeroR.

Though ZeroR and OneR did not provide result, KNIME produced an accuracy of 89.36% with C4.5 over the zoo data set which is very close then 92.13% of that confirmed by C4.5 in Weka.

Beside the non availability of ZeroR and OneR algorithm implementation, Tanagra's accuracy with C4.5 is satisfactory with result between 74.38% and 80.51% which is more stable then compared to 62.55% to 89.36% of what it is with KNIME.

3.3 Features and Functionalities:

3.3.1 Weka:

Database system support*:* Weka supports reading of files from several different data bases. It also provide feature to import the data over internet, from web pages or from a remotely located SQL database server by entering merely the URL of resource. This hence allows Weka to support variety of different data formats [9].

Graphical representation: Weka provides limited support to the visualization of concluded data. Although the Weka API provides with various data mining and processing methods, it lacks in the representation of the result of processing. The representation of the results, graphs and plots, lacks in detailed representation. But that really is not of much concern as Weka provides appreciable support for other functionalities and also the visualization provided is sufficient to represent the view of data on which the analyses has been performed and the results of the data analyses and preprocessing. Weka also allow for available add-on functions to be included in order to be able to interface with R statistical package with improved statistical analysis and representation of result [4].

Analysis and Processing Capabilities*:* It allows users to use R application as can directly interact with R package. Weka also provides with separate GUI for knowledge flow, for Experimenter, used to compare various results and explorer to analyze different data sets. It also provides for creating own filters for filtering out instances [4].

Issues: However, Weka is not better suitable option for the large data sets as they are roughly handled. The databases with large unstructured data are not suitable as it hinders the pre-processing and computing time of Weka. It can perform well and provide more accurate results with smaller databases. This tool has limited ability to partition dataset to training and test sets. It does facilitate to save parameters for future application. For cross validation or independent validation, Weka lacks to allow saving model in order to avoid rebuilding for other datasets [1].

3.3.2 Knime:

Database system support*:* KNIME provides with a great strength being able to establish database connections with any number of databases that provides JDBC. This tool also provides with a unique functionality of ports to different data sources and databases. With this there is no need to modify the SQL query. Users can use these ports in order to integrate from several databases and modify the dimension of the database accordingly [4].

Graphical representation: Workbench in KNIME provides user with an easy way to handle different functions and data flow of the process by merely dragging and dropping new nodes which can then be customized as required. It also provides with its up gradation to improve the workbench functionality. KNIME can also be implemented deploying further improvement with individual efforts over a particular function and over the data so as to be able to represent variety of data types and their properties [4]. Its applicability varies

from execution of basic processes to the integration with other software for enhanced visualization and performance. However, the manual effort required with KNIME makes it a less suitable option for large complex workflows.

Analysis and Processing Capabilities: KNIME provides all the supported operation for data mining with a graphical interface where the naïve user can import, filter and generate workflow without any need to formulate SQL queries and code [10]. On the other hand KNIME allows experts to write their own programming script for any new node or to download extensions from other users to attain more specific functionality.

Issues: KNIME also has limited partitioning ability like Weka but does provide the functionality to save parameters and validation model for cross validation and independent validation [1].

3.3.3 Tanagra:

Database system support: Importing of different data sources is not supported in Tanagra because, as it provides only tree representation, so it would disturb the tree nature of the stream diagram. Even it can also not read from any other data source format or integrated database directly [12].

Graphical representation: The visualization of various models is although not like other tools but it includes several statistical measures. Tanagra is based on stream diagram paradigm where a user defines data sources, data operation and a linking path through the sources and operations. This path describes the flow similar to workflow. But Tanagra allows the graph to be represented only as a tree so as to maintain the simplicity. However, this creates a bottleneck with the performance of Tanagra as now there can be only one data source to any operation [4]. Tanagra provides with different ways of visualization of data as in scatter plots, graphs and tables. However, it is difficult to develop graphs by specifying several parameter values as some may not be applied to a given data set but should be applied to the sub-node of the data indirectly. This leads to the missing values for some views. Instead the visualization of data result is more useful as high dimensional data can be represented in 2D. The representation of clustering and classification result is in text [12].

Analysis capabilities: Tanagra can import text files with whitespace delimited fields. It also provides with a conversion tool for files it cannot read from arff. This tool provides robust statistical analysis functionality by complementing with a range of uni- and multivariate parametric and nonparametric tests. It also includes correspondence analysis, principal component analysis, and the partial least squares methods. [11]

Issues: Tanagra cannot implement association rule mining for large datasets. It cannot integrate databases. It is required to integrate different datasets outside Tanagra and then import before performing any operations. Out of the available operator, not all are applicable to every node. Some of these often fail when applied to inappropriate location and that too without proper display of any error message. Tanagra has limited partitioning ability and does not provide to save parameter values. It also does not support for saving independent validation model [12].

4. EVALUATION OF COMPARATIVE STUDY

There are conclusions drawn from the study of these tools. The analysis of these tools has provided us with idea for the betterment of the whole data mining procedure.

Even though these tools have proved to be appropriate for the specific data domains and specific data mining tasks such as classification, clustering, etc, the shortcomings, flaws or specificity of these tools have acted as a pullback from the implementation of a general framework for data mining process. The major common drawback with theses tools is that their processing of data, classification, clustering, prediction and inferring of rules all is based on the selection of the algorithm for data mining over a particular type of data set. If the selection of algorithm is not appropriate regarding the domain of data then the produced patterns or predictions cannot be completely relied. For example, the classification over Audiology data set results with 84.41% correctly classified instances with SimpleLogistic while with ZeroR it results in only 27.27%. If an inappropriate algorithm is used for future data value prediction then it would produce incorrect results.

Another issue with these tools is that the current state of art does not provide an automated mining technique. All the tasks such as classification and clustering are performed consecutively and are specific to an application [5]. A theoretical framework is required for implementation of unified theory where these data mining tasks can be unified and overcome with the shortcomings of these tools.

5. UNIFIED DATA MINING THEORY

The unified theory suggest measures for development of a unification process for data mining softwares which can be used over a general set of databases, over a general set of data domains and which performs all the mining tasks, classification, clustering and visualization in a unified manner instead of individually performing each task [8].

This unified theory can be achieved by more then one means. Several research works is being carried on to come up with an implementation without any bugs. Here, two methodologies have been suggested for formulation of this theory.

5.1 UDMT By Means of MAS (Multi Agent Systems)

With arrival of newly extracted data from some new information being introduced every time consistently, it is required to integrate, and to analyze, clean, preprocess and incorporate the newly mined data into the database dynamically. Instead, presently the users themselves have to even set a time for the data mining process to be executed statistically and to mine again the newly added data in order to derive new interesting patterns.

These tools are not meant for a general user. They require some degree of expertise to be able to work with them, and to be just able to understand the functioning, and to analyze the outcome of these tools as a result of any data processing operation performed with them, where knowledge of data over which the tools will be used is also of concern. This technical complexity of implementing correct algorithm and technique also depends even on an individual expert's knowledge [1].

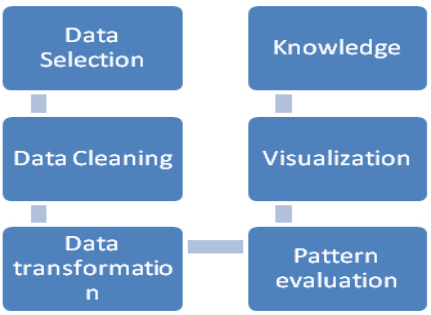

Figure 3: The conventional data mining process.

Multi agent systems are those systems which break a domain in components and each component performs its part of processing to achieve a common goal [5]. These agents are adoptive to new environment, autonomous, flexible, and easily modifiable and pertains some characteristics relating to artificial intelligence. These intelligent agents handle several independent tasks. Such multi agent system is modular and provides robustness to the system.

A MAS based data mining tool would provide with:

- Most suitable processing technique
- Preprocessing of new data and updating in the database.
- A data set of patterns obtained every time after any new data addition and suggest with possible knowledge that can be discovered.
- Reduced processing time for pattern recognition and representation of analyzed data.

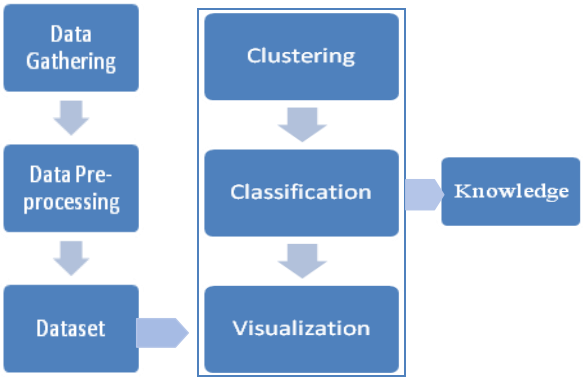

Figure 4: proposed data mining process.

In the above depicted process, the initial stages (data gathering, data preprocessing, data cleansing, and dataset preparation) are same as the data mining process. However, the next stage in this process, the mining tasks, classification, clustering and visualization are unified and the process is called Unified Data Mining Process UDMP [5]. In this technique of data mining, the only required input is a dataset in which the proceeding tasks are performed automatically and knowledge is mined. This knowledge is then analyzed by the user if is required by him according to his business rules [5]. There is no need to manually implement algorithms for classification and clustering in the whole process, however it is though required to specify them at the beginning of the process.

Mathematical formulation: This theory has been mathematically provided. Here algorithms are treated as functions namely; one function for classification algorithms; one function of clustering algorithms; one function of visualization algorithms.

The input and output to these are: data set and cluster set respectively for classification function; cluster set and set of rules as output for clustering function; set of rules as input and set of graphs as output of visualization function. The mapping of these functions onto the data set and its output represents a general workflow required for the unification process [5].

Issues: The implementation of MAS systems for unification has been successful but appropriate algorithm selection still remains a major issue in UDMT formulation with MAS. This issue has been arrived as for a function; it is needed to define the function properly before mapping it onto some input data. In case of data mining it means the definition of proper, appropriate algorithm.

There are several selection criterion defined, for eg; VC (Vapnik-Chervonenkis)-dimension, CV (Cross-validation), AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), (SRMVC) Structural Risk Minimize, but the accuracy and confidence over the accuracy of the algorithm is still an issue as not every tool provides support for different data types [5]. Which leaves us with the previously discussed problem of algorithm selection?

5.2 UDMT By Means of Inductive Database

The unified theory has also been proposed here with the idea of working over the core concepts of data mining (such as data, databases, queries) and coming up with adoption of some enhanced concepts of data mining.

Inductive databases and inductive queries: Inductive databases consist of mined patterns from previous processing of data in addition to the regular or local data [7]. In these IDB's these patterns are treated as general data and data mining tasks are performed as queries with consideration of these patterns. The relating queries (IQ's) are executed over these IDB's and they generate more of such patterns. Thus it results in a data mining process which is an extension of querying process where both data and patterns holding that data are queried at once.

The major benefit with this formulation is that previously only data was queried with a simple query which resulted in only data, now the IQ's over IDB's produce more refined patterns. There is no need to access the data, only querying the patterns within the IDB's can provide considerable results [7]. Several patterns and data can be processed over and manipulated by interlinked inductive queries replacing the traditional mining tasks, combined in order to discover hidden data and other patterns.

But there are some preliminary requirements for carrying over this formulation. The generation of such databases and querying language itself requires a measure for its development. The specification of the patterns to be included within a database requires satisfying some constraints defined over them. These constraints can be language constraint and evaluation constraints [7].

The language constraints deal with the pattern only while evaluation constraints conclude with the validity of that pattern within the considered database. Thus constraints play an important role and are helpful in more efficient pattern recognition. Similarly there are constraints defined for the IQ's over the databases. This is again then defined as

language based. (eg., association rule learning) and evaluation constraint. The generation of inductive databases for different types of patterns can be accomplished by an inductive query language and constraint based algorithms only. Main concern which arrives here is to deal with the definition of constraints which can be then used to generate inductive queries by applying them over data and various patterns. For each type of data and type of pattern a specific algorithm is defined [7].

This development of an inductive query language has helped with the formulation of unified theory. In this language the out put of one IQ can be given as an input to the other which is one of the requirement of the theory in order to achieve the sequenced execution of classification, clustering and visualization tasks supporting KDD process [7]. This IDB framework is more promising as it employs declarative queries instead of ad-hoc procedural queries.

Issues: There exist many theories regarding the approaches of inductive querying but there is no proper theoretical work defined to help with the practical approach of constraint based mining (IQ). Many of the approaches work in isolation and no measure defined for integration with DBs and DB tools [7].

5.3 General Framework Requirements

The general framework for data mining must support:

- Different types of data and data models with orthogonal representation so as to provide integration
- Mining of different types of structured data in a uniform fashion.
- Predictive modeling, clustering, mining frequent patterns in complex data, change and deviation detection,
- Different representations for the same data mining task [7].

These basic concepts can serve as the basis for the development of an IQ language and algorithms for multi-step KDD process.

6. CONCLUSION

The above study was conducted by using four algorithms over a data set: Zero Rule (ZeroR), One Rule (OneR), decision tree (C4.5), and k-nearest neighbor (KNN). Tools were run over the data set and results were observed for each algorithm. Accuracy percentage served as performance measure. Weka was indentified a better performer with the specified algorithms, followed by KNIME and Tanagra. This performance ranking based on the type of data set used and how the classifier is implemented within the tool, as task of classification is affected by so. But Weka still proved to be better as it provided with the implementation of ZeroR and OneR over data types where other tools did not.

The functionality offered by these tools, like API support and graphical presentation along with other features aid with the selection of tool best suitable according to the usage by different users. The methodologies discussed for UDMT formulation offer better approach towards data mining but still are left with some issues due to an incomplete theory for their correct formulation. The MAS suffer from the problem of selection of appropriate set of algorithms for classification, clustering and visualization. The correct application of an algorithm as a function is also an issue.

The Inductive Databases theory addresses the need for a language to design inductive queries and generation of databases including both data and patterns. All the defined patterns are then required to satisfy the constraints imposed upon. This results in need of better algorithms for constraint based data mining.

7. ACKNOWLEDGEMENT

The author sincerely thanks to his parents who motivated him throughout the duration of the work and Mr. Sandeep Bhargav (Department of Computer Science), Gyan Vihar University, Jaipur, without whose guidance this task would not have been accomplished.

8. REFERENCES

[1] "A comparative analysis of data mining tools in agent based systems", Sharon Christa, K. Laxmi Madhuri, V.Suma, research and industry incubation centre, Dhyanchand Sagar institutions.

[2] "A Comparison study between data mining tools with some classification methods", Abdullah H. Wahbeh, Qasem A. Al-Radaideh, Mohammed N. Al-Kabi, and Emad M. Al-Shawakfa.

[3] "Evaluation of Fourteen Desktop Data Mining Tools", Michel A. King and John F. Elder IV, Ph.D., Department of Systems Engineering- summary.

[4] "Suitability analysis of data mining tools and methods" Samuel Kováč, Masaryk university faculty of informatics.

[5] "Towards the Formulation of a Unified Data Mining Theory, Implemented by Means of Multiagent Systems (MASs)", Dost Muhammad Khan, Nawaz Mohamudally and D. K. R. Babajee.

[6] "A Study of Data Mining Tools in Knowledge Discovery Process", Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, A.S.K.Ratnam.

[7] "Towards a General Framework for Data Mining", Saˇso Dˇzeroski Joˇzef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

[8] Principles of Data Mining, by David Hand, Heikki Mannila and Padhraic Smyth

[9] WEKA, the University of Waikato, Available at: http://www.cs.waikato.ac.nz/ml/weka/downloading.html

[10] KNIME -Available at: http://www.knime.org/download-desktop

[11] "Open-Source Tools for Data Mining" Blaz Zupan, PhD, Janez Demsar, PhD.

[12] "Tanagra: An Evaluation" Jessica Enright Jonathan Klippenstein, November 5th, 2004

[13] UCI Machine learning repository. Available at: http://archive.ics.uci.edu/ml/

9. AUTHOR'S PROFILE

Harshvardhan Solanki, born on 9th Feruary, 1991 in Jaipur, Rajasthan. He completed his bachelor's degree (B. Tech.) in Computer Science from Suresh Gyan Vihar University, Jaipur in 2012. He is currently pursuing his master's (M. Tech.) in Software Engineering from Suresh Gyan Vihar University, Jaip