*International Journal Of Computer Science And Applications Vol. 1, No. 1, June 2008*
ISSN 0974-1003

# A Hybrid Model for Performance Analysis of Data Mining Tools

G.Indumathi,
Department of Computer Science, Pondicherry University, Puducherry, India
indumathigovindan@gmail.com

P.Dhavachelvan,
Department of Computer Science, Pondicherry University, Puducherry, India
pd_chelvoume@yahoo.co.in

## ABSTRACT
Seeking knowledge from massive data is a very essential and difficult; Data Mining tools facilitate the data analysts for mining information from any kind of sources. Mining tools can perform mining on data represented in quantitative, textual, or multimedia forms. The data mining tools available today use many new techniques and produce quality information indifferent aspects. In this work, a 3-tier architecture model is formed for analyzing the performance of the data mining tools. To evaluate the performance of the data mining tools, a complete survey is done and a set of quality attributes that are required for a data mining tools are defined. With the use of defined set, various data mining tools are evaluated. This performance analysis model consists of three primary components: a set of data mining tools, a set of performance metrics, and a set of application domains. These components form a 3-tier architecture for data mining performance assessments. The proposed also uses the statistical methods for interpreting the outcomes of the assessment model.

## Categories And Subject Descriptors
H.2.8 [Database Management]: Data Mining, H.3.3 [Information storage and retrieval]: Information Search and Retrieval, H.3.4 [Information storage and retrieval]: Performance Evaluation, G.3. [Probability and Statistics]: Statistical Computing.

## General Terms
Performance, Experimentation

## Keywords
Data mining tools, Performance Evaluation, Statistical Computing

## 1 INTRODUCTION
Nowadays, large quantities of data are being accumulated and seeking knowledge from massive data is a very difficult. Data could be large in two senses in terms of size, or in terms of dimensionality. Extracting knowledge and information facilitate by data mining, which is defined as a process consisting of data analysis and the use of algorithms to extract patterns in data or identify similar patterns in data. To get the knowledge from such immense data is one of the most desired attributes of Data Mining. Using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results [9]

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, machine learning methods, and mathematical algorithms. The mining tools should deliver the knowledge in user friendly way. And in addition knowledge is extracted in such a way that makes the highly complex tasks of data mining transparent to the user. Mining tools can perform mining on data represented in quantitative, textual, or multimedia forms.

The type of the data in the datasets vary depend on the domain of the database. Every mining tool is capable to mine certain type of data, so the mining tools are domain specific. The basic capabilities of the mining tools must be examined before choosing the tools. So choosing the correct mining tools from existing mining tools is not a simple task. The tools are also checked for the quality metrics that are expected for good mining tools.

These quality metrics can be quantitative and qualitative. In our mode we considered the metrics are accuracy, precision, specificity and sensitivity. Using these quality metrics the data mining tools are evaluated and the data analyst can choose the best among the available tools. These quality metrics that are quantitative are measured using sample databases on specific domain. Mining is done using all the tools and the values are compared and the tool having high quality values is chosen as the best tool.

## 2 PERFORMANCE ANALYSIS MODEL
A model is designed for analyzing the performance of the data mining tools and thereby facilitates the user to choose the best mining tool among the available tools. This model contains three major components; set of data mining tools, set of quality metrics, and set of application domains. The set of data mining tools specifies about the tools that are to be evaluated, the set of quality metrics defines about the quality metrics that the mining tools should have, and the set of application domain describes about the domains on which the mining tools are evaluated. These three major components are placed together to form hybrid model. This model is represented in the Figure1.

The purposed architecture is used for data mining performance assessment. The component set of data mining tools are depend on set of application domains and set of quality metrics


Published by Research Publications, Chikhli, India

components. The set of application domains implicitly specifies about the data types of the database. That is, if we specifics the domains as bank or finance, the data types will be textual and

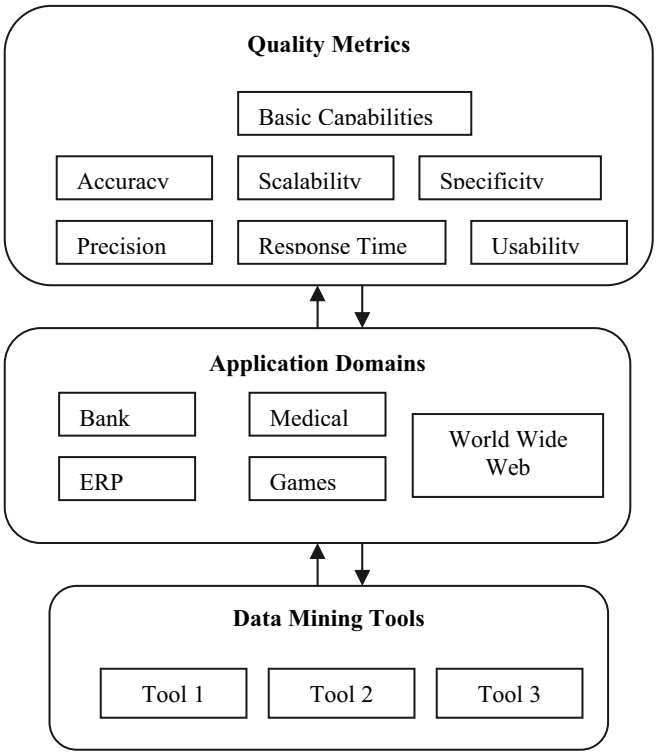


**Figure 1. 3 tier Architecture for Performance**

quantitative. In addition, the domain may have image data also.

## 3 EXPERIMENT IN AND RESULTS

We experiment the performance assessment for four different recent data mining tools; Tanagra, Weka 3.4.11, Rapid Miner and Alpha Miner. These tools are evaluated using two databases in the Medical and Statistical domain. Both these domains use textual and quantitative data. These two types of data are supported by the above mention four mining tools. The four tools are evaluated using four quality metrics accuracy, sensitivity, specificity and precision.

The tools are applied with the Medical database, and then knowledge is extracted from the database using the mining techniques. Here we used classification technique for mining the information. Then using the confusion matrix produced by the classification technique, the quality attributes are calculated. This procedure is done for the four tools. Then the same procedure is followed for the Statistical database also.

The four metrics are calculated using formula that are specified below[6][10]

**Metric 1:** Accuracy = (a + d)/ (a + b + c + d)

**Metric 2:** Precision = no. of True Positive / (no. of True Positive + no. of False positive)

i.e. Precision =a + (a + b)

**Metric 3:**Sensitivity=no. of True Positive / (no. of True Positive + no. of False Negative)

i.e. Sensitivity = a + (a + c)

**Metric 4:**Specificity=no. of True Negative / (no. of True Negative + no. of FalsePositive)

i.e. Specificity = d + (d + b)

The values of a, b, c and d are described in the confusion matrix as structured in the Figure 2.

| | Actual +ve | Actual -ve |
|---|---|---|
| Predicated +ve | a (TP) | b (FP) |
| Predicated -ve | c (TN) | d (FN) |

**Figure 2. Confusion Matrix**

Using the values in the confusion matrix, the values for the above specified metrics are calculated. The comparison table is show below. The figures 2, 3, 4 and 5 are the bar chart for each quality metric. The bar chart shows the metric values for each tool. The X axis is the list of the tools and y – axis specifies the metric values for each tools. The white bar is for medical database and the black bar is for statistical database.

| Tools | Metrics | Medical | Statistic |
|---|---|---|---|
| Tanagra | Accuracy | 0.8451 | 0.6496 |
| | Precision | 0.9918 | 0.7611 |
| | Specificity | 0.9000 | 0.7037 |
| | Sensitivity(Recall) | 0.8413 | 0.6144 |
| Weka | Accuracy | 0.9935 | 0.9562 |
| | Precision | 0.9918 | 0.9104 |
| | Specificity | 0.9696 | 0.9210 |
| | Sensitivity(Recall) | 1.0000 | 1.0000 |
| Rapid Miner | Accuracy | 0.6903 | 0.5985 |
| | Precision | 0.9365 | 0.5700 |
| | Specificity | 0.8750 | 0.3134 |
| | Sensitivity(Recall) | 0.4796 | 0.8714 |
| Alpha Miner | Accuracy | 0.9225 | 0.7445 |
| | Precision | 0.9837 | 0.6285 |
| | Specificity | 0.9167 | 0.6904 |

| | | | |
|---|---|---|---|
| | **Sensitivity(Recall)** | 0.9236 | 0.8301 |

**Table 1. Evaluation Table**

The tools are evaluated using two different domain specific databases and the metrics that values are calculated. For the medical domain, the above result shows that Weka is the best tool when compare to the other tools. The tool next to Weka is the Alpha Miner, followed by Tangara and Rapid Miner. For the Statistical domain the tools Weka shows good performance over the other tools. Then the tools Rapid Miner and Tangara are in the position. The preformed alike in this domain. From this experiment, among the four mining tools Weka given better performance for medical and Statistical domain.

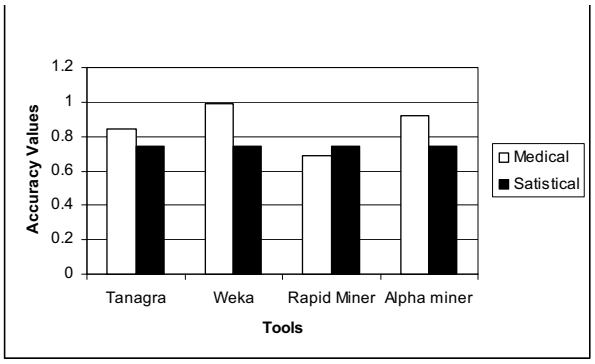


**Figure 3. Accuracy**

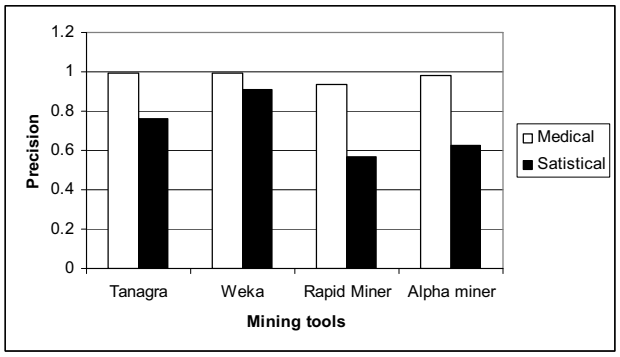


**Figure 4. Precision**

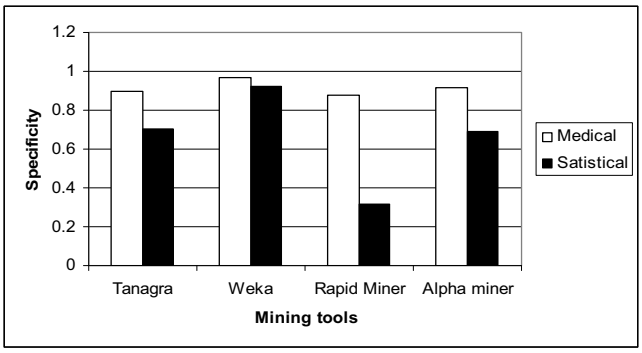


**Figure 5. Specificity**

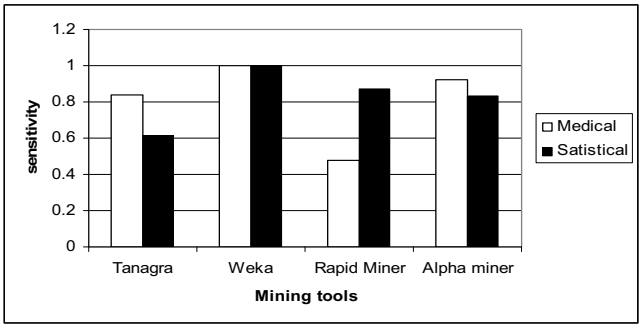


**Figure 6. Sensitivity**

# 4 CONCLUSION

The quality of mining depends on application domain. This is proved in the proposed work. Each tool has been evaluated using four different metrics in two different domains. It is observed that the performance of the tools is varied greatly and the variation is purely dependent on the application domain and type of data. This model facilitates the data analyst, business people, and researchers for selecting the exact tool as they need. And thereby, provides high quality information for using the appropriate tool.

# 5 REFERENCES

[1] Hanna M. Wallach, "Evaluation Metrics for Hard Classifiers", University of Cambridge, November 2004.

[2] Herna L. Viktor and Wayne M. Motha, "Creating Informative Data Warehouses: Exploring Data and Information Quality through Data Mining, Informing Science, InSITE - "Where Parallels Intersect", June 2002.

[3] "Introduction to Data Mining and Knowledge Discovery" Third Edition by Two Crows Corporation, 2005.

[4] Jeffrey W. Seifert, "Data Mining: An Overview", Analyst in Information Science and Technology Policy Resources, Science, and Industry Division , December 2004.

[5] Jesse Davis jdavis, Mark Goadrich,"The Relationship Between Precision-Recall and ROC Curves", International Conference on Machine Learning, 2006.

[6] José Ignacio Serrano, Marie Tomečková, Jana Zvárová, "Machine Learning Methods for Knowledge Discovery in

Medical Data on Atherosclerosis", European Journal for Biomedical Informatics.

[7] Michel A. King, John F. Elder, "Evaluation of Fourteen Desktop Data Mining Tools", 1999.

[8] "Oracle Data Mining Concepts", June 2005.

[9] Osmar R. Zaïane," Introduction to Data Mining" Principles of Knowledge Discovery in Databases, University of Alberta, Department of Computing Science, 1999.

[10] Vladimir Brusic And John Zeleznikow, "Knowledge Discovery And Data Mining In Biological Databases", The Knowledge Engineering Review, Vol. 14:3, 1999.