Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 1; January-March, 2015 pp. 39-42
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/acsit.html

Data Quality and the Performance of the Data Mining Tools

Mrs. Rekha Arun[1] and J. Jebamalar Tamilselvi[2]

[1]*Research Scholar, Sathyabama University, Chennai*
[2]*Research Supervisor, Sathyabama University, Chennai*

Abstract—*This investigation focuses on the impact of data quality on the performance of data mining tools used at national research institutes of northern India. Performance criteria namely: Computational Performance, Functionality, Usability and Ancillary Task Support were considered for the study. Regression models were developed from the data collected with the help of a suitable and tested questionnaire. The analysis revealed that 'Computational Performance' is significantly affected by the completeness of data, while 'Functionality' is mainly affected by the consistency of data. Validity of data has an affect on the 'Usability' of the tool. Consistency of data and completeness of data have significant impact on 'Ancillary task support of the tool' with consistency having greater influence than completeness. Thus it is concluded that data quality has vital impact on the performance of the data mining tools.*
Keywords: Data mining, Data quality, Computational performance, Functionality, Usability, Ancillary task support

1. INTRODUCTION

Majority of research and business organizations are moving towards data mining and data warehousing now a days. This technology switching requires integration of data collected over long periods of time and through multiple generations of database technology. Researchers typically utilize diverse information from multiple database support planning of experiments or analysis and interpretation of results. This assimilation of data from diverse schemas and data sources may cause low quality data. Four sources of error are observed in bioinformatics databases.

1. Attribute level–incorrect values of individual field, the cause may be errors in original data submitted or from automated systems for record processing.
2. Record Level–conflicts between or misplacement of fields within a record
3. Single Source Database Level–Conflicting or duplicate entry.
4. Multi Source Database Level–imperfect data integration and source synchronization.

High quality data or clean data are essential to almost any information system that requires accurate analysis of large amount of real world data. To improve the quality of data, four essential tasks are suggested by DOD (Department of Defence) as

a) Define scope problem, identify objectives, identify and review Documentation, Develop quality metrics.
b) Measure Apply organization metrics, Flag suspect data.
c) Analyze Indentify conformance issues, Provide recommendations, Prioritize conformance issues, Validate conformance issues.
d) Improve Select improvement opportunities, Implement improvements, Document improved quality, Update organizations standards.

The data mining tools bring together techniques from machine learning, pattern recognition, statistics databases, and visualization to address the issue of information extraction from large data bases.

2. IMPACT OF DATA QUALITY ON PERFORMANCE OF DATA MINING TOOL

Impact of data quality on the performance of data mining tools can be analyzed by investigating four categories of criteria namely Computational Performance, Functionality, Usability, and Ancillary task support as suggested by Collier et al. (1999). The data quality attributes are adopted from the guidelines provided by the Department of Defence which comprise ' Validity', 'Timeliness', 'Consistency', 'Completeness', 'Uniqueness', and 'Accuracy'. Regression models were developed to study the impact of data quality on data mining tool for the four performance criteria.

Computational Performance of Data Mining Tool

Computational performance is the tool's ability to handle a variety of data sources in an efficient manner i.e. to easily handle data under a variety of circumstances rather than on performance variables that are driven by hardware configurations and/or inherent algorithmic characteristics. The regression model for the same is given table 2

Impact of Data Quality on Computational Performance of Data Mining Tool

Model	R	R square	Adjusted R Square	Std. Error of the Estimate
1	.490a	.240	.222	.503

ANOVA b

Model	Sum of squares	DF	Mean Square	F	Sig.
Regression	3.437	1	3.437	13.589	.001a
Residual	10.874	43	.253		
Total	14.311	44			

Coefficients a

Unstandardized coefficients			Standardized Coefficients		
	B	Std. error	Beta		
(constant)	.827	.162		5.109	.000
Completeness	.253	.069	.490	3.686	.001

Excluded Variables b

Model	Beta Ins	T	Sig.	Partial Correlation	Collinearity Statistics Tolerance
Accuracy	.007a	.043	.966	.007	.715
Consistency	-.054a	-.367	.715	-.057	.848
Timeliness	.183a	1.223	.228	.185	.781
Uniqueness	.008a	.049	.961	.008	.710
Validity	.010a	.066	.948	.101	.866

a) Predictors, completeness
b) Dependents variable : Computational Performance

This table represents the regression model, it can be analyzed that the impact of the data quality on computational performance of a data mining tool is 24 percent. Rest of the performance is affected by other factors. The F value is 13.589 significant at 0.01 level. From the coefficients table it can be analyzed that data quality attribute significant at 0.01 level and t-value of COMP is 3.686 significant at .010 level. Excluded variables table represents the variables having significant value greater than 0.05 and are thus excluded from the regression equation and do not majorly affect the computational performance of data mining tool.

Functionality of Data Mining Tool:

Software functionality helps access how well the tool will adapt to different data mining problem domains. It is the inclusion of a variety of capabilities, techniques and methodologies for data mining. Results of the multiple regression analysis are made available in table 3 the value of R square is .417 which is significant at level 0.01. This indicates that the quality of data slightly affects the functionality of a data mining tool. Data quality attribute 'Consistency of Data' explains the 41.70 rest of the variance is affected by some other factors

Impact of Data Quality on Computational Performance of Data Mining Tool

Model	R	R square	Adjusted R Square	Std. Error of Estimate
1	.646a	.417	.404	.811

ANOVA b

Model	Sum of squares	DF	Mean Square	F	Sig.
Regression	20.277	1	20.277	30.809	.000a
Residual	28.301	43	.658		
Total	48.578	44			

Coefficients a

Unstandardized coefficients			Standardized Coefficients		
	B	Std. error	Beta		
(constant)	-1.546	.796		-1.942	.059
Consistency	.978	.176	.646	5.551	.000

Excluded Variables b

Model	Beta Ins	T	Sig.	Partial Correlation	Collinearity Statistics Tolerance
Accuracy	.113a	.644	.523	.099	.447
Completeness	.159a	1.263	.213	.191	.848
Timeliness	-.106a	-.816	.419	-.125	.813
Uniqueness	.114a	.890	.379	.136	.836

a) Predictors in the model : (constant), Consistency
b) Dependents variable : Functionality

Usability of Data Mining Tool

Usability refers to the quality, how easy a tool is to learn and use i.e. accommodation with different levels and types of users without loss of functionality or usefulness. Multiple regression is used to analyze the impact of data quality on 'Usability' of data mining tool and is illustrated in table 4

Impact of Data Quality on Computational Performance of Data Mining Tool

Model	R	R square	Adjusted R Square	Std. Error of the Estimate
1	.431a	.186	.167	.742

ANOVA b

Model	Sum of squares	DF	Mean Square	F	Sig.
Regression	5.418	1	5.418	9.834	.003a
Residual	23.693	43	.551		
Total	29.111	44			

Coefficients a

Unstandardized coefficients			Standardized Coefficients		
	B	Std. error	Beta		
1 (constant)	.957	.836		1.145	.259
Validity	.554	.177	.431	3.136	.003

Excluded Variables b

Model	Beta Ins	T	Sig.	Partial Correlation	Collinearity Statistics Tolerance
Accuracy	.142a	.842	.405	.129	.671
Completeness	.074a	.499	.621	.077	.866
Consistency	.066a	.338	.737	.052	.510
Timeliness	-.001a	-.005	.996	-.001	.862
Uniqueness	-.009a	-.057	.955	-.009	.863

The result of multiple regression exhibited in table 4 indicate that the impact of data quality attributes on the 'usability' of the data mining tool is 18.6 percent. The F-value is 9.834 which is significant at 0.01 levels. A variable 'validity of date' is significance affecting the usability and having t-value 3.136 at 0.01 level of significance. Excluded variable table list the variable with significant greater than 0.05.

Ancillary task support of data mining tool

Ancillary task support allows the user to perform the variety of data cleansing, manipulation, transformation, visualization and other tasks that support data mining. Theses task include data selection, cleansing, enrichment, value substation, data filtering, binning of continuous data, generating derived variables, randomization, deleting records, etc.

Impact of data quality on Ancillary task support of data mining tool

Model	R	R square	Adjusted R square	Std. Error of the estimate
1	.481a	.231	.231	.624
2	.555b	.308	.275	.599

ANOVA a

Model	Sum of square	Df	Mean square	F	Sig.
1regression	5.037	1	5.037	12.937	.001a
Residual	16.741	43	.389		
Total	21.778	44			
2regrssion	6.704	2	3.352	9.339	.000b
Residual	15.074	42	.359		

Coefficients a

	B	Std. error	Beta	T	sig
1(constant)	2.045	.612		3.339	.002
Consistency	.487	.136	.481	3.597	.001
2(constant)	2.175	.591		3.608	.001
Consistency	.369	.141	.364	2.611	.012
completeness	.191	.089	.300	2.155	.037

Excluded variables b

Model	Beta In	t	Sig.	Partial correlation	Collinearity statistics Tolerance
Accuracy	.203a	1.015	.316	.155	.447
Completeness	.300a	2.155	.037	.316	.848
Timliness	-.009a	-.058	.954	-.009	.813
uniqueness	.046b	.310	.758	.048	.836
validity	.151a	.806	.425	.123	.510
2accuracy	.046b	.218	.829	.034	.376
timeliness	-.115b	-.898	.374	-.139	.707
uniqueness	-.115	-.724	.473	-.112	.665
validity	.099b	.539	.593	.84	.500

a. predictors in the model(constant),consistency
b. predictors in the model:(constant),consistency, completeness
c. dependent variable: ancillary task support

From the regression model given in the bove table 5 the value of R square is .308 which is significant at level 0.01 this indicates that the quality of data significantly affects the ancillary task support of a data mining tool data quality attributes 'consistency of data" and 'completeness of data' together explain the 30.8percent of the variance (R square) in the ancillary task support of the data mining software rest of variance is affected by some other factors consistency of data has and completeness of data are significant at 0.01level.the beta value indicates the relative influence of the entered variables, that is, 'completeness of data'(beta=0.300).this can be analysis that the rest of the factors do not significantly contribute and are kept in excluded variable list

3. CONCLUSION

From the above analysis it can be observed the data quality has vital impact on the performance of data mining tool at research list organizations

a) Computational performance of the data mining tool is significantly affected by the 'completeness of data'.
b) Data quality attribute 'consistency of data' majorly affects the functionality of the data mining software .reset the variance is affected by some other factors
c) Validity of data has been affect on the usability of data mining tool
d) Two data quality attributes 'consistency of data' and 'completeness of data' have impact and ancillary task support of data mining tool, where consistency has greater influence than completeness.

Consequently, data for the research must be collected are measured keeping in mind the six attributes of data quality since they have great impact on the performance of data

mining tools. The study has been carried for the research institutes related to health in Rajasthan and Gujarat including the two institutes.

REFERENCES

[1] Yi-Ping Phoebe Chen(Ed.) 2005. Bio informatics Technologies, Springer–Verlag Berlin Heidelberg, Chapter 3: Dtata ware housing in Bioinformatics, Judice L Y Koh and Valdimir Brusic.

[2] C.Sumithradevi, M.Punithavalli 2009, "Detecting Redundancy in Biological Databases–An efficient Approach", Global Journal of Computer Science and Technology, Page 141-45 Vol 9, No.4

[3] DOD Data Administration Guidelines 2003, DOD guidelines on data quality management.

[4] V.Ganti, J.Gherke, and R.Ramakrishnan 2001. "DEMON: Mining and Monitoring Evolving Data", IEEE Transactions on Knowledge Management and Data Engineering, Vol.13, No.1, pp.50-62.

[5] J.Hipp, U.Guntzer, and U.Grimmer 2001. "Data Quality Mining", Workshop on Research Issues in Data Mining and Knowledge Discovery.

[6] D.Luebbers, U.Grimmer, and M.Jarke 2003. "Systematic Development of Data Mining–Based Data Quality Tools", Proceedings of the 29th VLDB Conference, Berlin, Germany.

[7] M.Ge, and M.Helfert 2007. "A review of information quality research–develop a research agenda", in The International Conference of Information Quality, Cambridge, Massachusetts, USA.