Data Size versus Accuracy: Performance by different Data Mining Tools

D. Udhayakumarapandian and RM. Chandrasekaran

Abstract—**While increasing the data size the improvement in accuracy becomes better. This is true only up to a fixed size. After this point, the performance usually becomes stable. In the context of small data sets expecting better performance usually leads to failure. Data mining research community tries to address this problem case by case basis. In this paper we consider this study for diabetes and cancer datasets and establish the output showing the appropriate convergence of accuracy while increasing data size. This has been tested for different standard and familiar data mining tools. Comparative results are listed for the performance in classifying errors.**

Index Terms— **Weka, R package, KEEL, Knime, classifiers, Accuracy**

1. INTRODUCTION

Big data retrieved from databases and data warehouses is dealt practically with a data mining analytical tool. Frequent increases in cpu speed combined with frequent decreases in the cost of mass hard disk and other computer hardware have made it feasible to gather and maintain massive data storages. The software for data mining can be tool that can perform either fully undirected or do perform partially analysis. This leads to have lower cost analytical methods applied to larger data sets. Hence it demands the investigation for controlling the data size of various data sets.

"The volume of the data is probably not a very important difference: the number of variables or attributes often has a much more profound impact on the applicable analysis methods. For example, data mining has tackled width problems such as what to do in situations where the number of variables is so large that looking at all pairs of variables are computationally infeasible." Thus we find the inter dependence of the attributes determines the complexity rather than from the ability to process massive volumes of instances.

Decision tree based data mining methods are subject to over-fitting as the size of the data set increases, as stated in Domingos [1998] ,Oates and Jensen [1997]. As Oates and Jensen [1998] note, "Increasing the amount of data used to build a model often results in a linear increase in model size, even when that additional complexity results in no significant increase in model accuracy." Similarly vein Musick, Catlett, and Russell [1993] suggest that "often the economically rational decision is to use only a subset of the available data." While pruning in Decision trees helps to limit the proportion by which model complexity increases as the amount of data increases, its effectiveness can only be assessed by looking at the responsiveness of complexity and accuracy of the model to changes in data set size.

Prior studies of appropriate size of data sets in data mining have used public domain data modeling using relatively small data sets from the UCI repository. In this paper we describe the results of models generated from the systematic increasing of data from two medical datasets. Models are generated with each of four standard familiar data mining tools.

2. Related Works

The effectiveness of data mining models based on sampling from datasets has not been widely studied. However, there are a few studies that have addressed this topic which can be used as the starting point for this study.

Arithmetic progressive sampling is performed by John and Langley (e.g. samples of 100, 200, 300, 400, etc.) to 11 of the UCI repository datasets. Small data sets are used with replication to produce large data sets. It has been found that Naïve Bayesian Classifier is used to generate sample-based model for better accuracy.

The experiment is focused on determining the shape of the learning curve and made no attempt to determine an optimal data size. Provost, Jensen, and Oates [1999] modeled 3 of the larger UCI repository datasets using differing progressive sampling techniques. Then a larger sample is used to generate another model whose accuracy also is tested on the holdout set. The process is repeated for models based on progressively larger samples, until some standard accuracy criteria are met.

For example [6], increments of 100 (100, 200, 300, 400) or increments of 500 (500, 1,000, 1,500,2,000). Geometric progressive sampling uses equal proportional increments and an arbitrary initial size. For example, incremental doubling with an initial sample size of 100 would use samples of 100, 200, 400, 800. The dynamic progressive sampling technique used by Provost, Jensen, and Oates involved: (1) initially estimating and testing models based on samples of 100, 200, 300, 400, and 500, (2) estimating a power function based learning curve based on results for those models, and (3) selecting the next sample to be the size required to achieve the accuracy criteria according to the learning curve.

ISSN: 2278 – 1323

3. Methods and Materials

3.1 Identification of data mining tools

Tool A: Weka

'Waikato Environment for Knowledge Analysis' is known as Weka. Weka is a collection of machine learning algorithms for data mining tasks and implemented in java. It has GNU general public license. These algorithms can either be applied directly to a data set or can be called from your own Java code. The Weka (pronounced Weh-Kuh) workbench contains a collection of several tools for visualization and algorithms for analytics of data and predictive modeling, together with graphical user interfaces for easy access to this functionality. It has general feature including pre-processing on data, classification, clustering, and association rule extraction. It provides three graphical user interfaces i.e. the Explorer for exploratory data analysis to support preprocessing, attribute selection, learning, visualization, the Experimenter that provides experimental environment for testing and evaluating machine learning algorithms, and the Knowledge Flow for new process model inspired interface for visual design of KDD process. A simple Command-line explorer which is a simple interface for typing commands is also provided by Weka .
Advantages of Weka
This tool can accommodate various file types like ARFF, CSV and C4.5 binary. This can be integrated with other java packages seamlessly. However it lacks proper and adequate documentations.

Tool B: KEEL

'Knowledge Extraction based on Evolutionary Learning' is known as KEEL, an application package of machine learning software tools. Solution to Data mining problems and assessing evolutionary algorithms is provided by KEEL. Collection of libraries for preprocessing and post-processing techniques for data manipulating, soft-computing methods in knowledge of extracting and learning, and providing scientific and research methods is available in KEEL. It has been Licensed by GNU, general public license. It can run on any platform and supported by java language.

Tool C: R

R package is a free software programming language and software environment for statistical computing and graphics. The R package is widely used among statisticians and data miners for developing statistical software and data analysis. R has the flexibility and easiness for documenting mathematical symbols and formulae from the outputs generated. It has been licensed by GNU General Public License. It has very extensive statistical library. The GUI has with more readable graphics facilities.

Tool D: KNIME

'Konstanz Information Miner' is known as KNIME. It is an open source data analytics, reporting and integration platform. It has been implemented in the areas like pharmaceutical research, CRM customer data analysis, business intelligence and financial data analysis. It is supported in the Eclipse platform using modular API. It is easily extensible. Custom nodes and types can be implemented in KNIME within hours thus extending KNIME to comprehend and provide first-tier support for highly domain-specific data format. It has been licensed By GNU General Public License

4 Selected Method: ADT

The effectiveness of decision tree method is achieved by the method using alternating decision tree. The decision tree structure by the tree stumps yield a frame model for boosting. The base node is a prediction node with a numeric score. The second layers of node are called decision nodes essentially needed for tree stumps. The following layer varies between prediction nodes and decision nodes. The base search method is exhaustive search while the other methods are faster by heuristics approach. The saved instance data can be used for visualization.

4.1 Data description

The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care).The population lives near Phoenix, Arizona, USA. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. ADAP is an adaptive learning routine that generates and executes digital analogs of perceptron like devices. It is a unique algorithm; see the paper for details.

4.1.1 Dataset

The datasets for these experiments are from [12].The original data format has been slightly modified and extended in order to get relational format.

4.1.2 Dataset description:

The database of diabetes describes a set of eight attributes11 as shown in the below list 2.2. The class attribute has binary values 'tested negative' and 'tested positive'. The number of instances in this database is 768.

Table 1 Average Percentage of Correctly Classified by Tool for diabetes Dataset

Data set	Sample size	A	B	C	D
Diabetes	100	69.6615	65.1042	67.9688	65.1042
	200	69.9219	65.1042	68.099	65.1042
	300	70.7031	65.2344	68.4896	65.1042

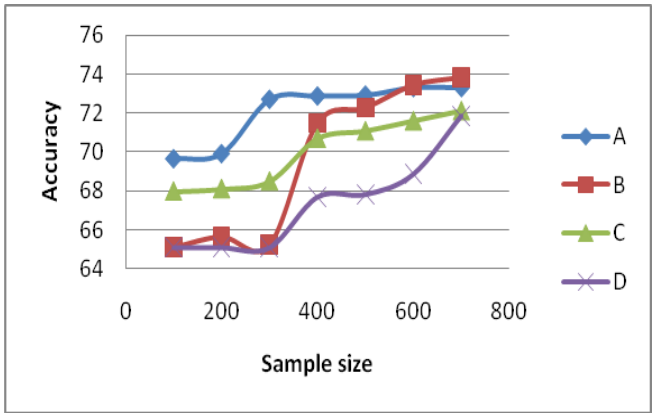
ISSN: 2278 – 1323

	400	71.875	71.4844	70.7031	67.7083
	500	72.9167	72.2656	71.0938	67.8385
	600	73.3073	73.4375	71.6146	68.8802
	700	73.3073	73.8281	72.1354	71.875

Table 2 Average Percentage of Correctly Classified by Tool for Cancer Dataset

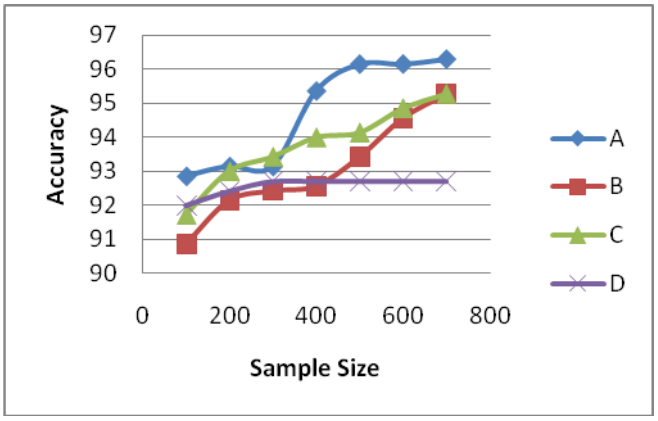
Data set	Sample size	A	B	C	D
Cancer	100	92.8469	90.8441	91.7024	91.9886
	200	93.133	92.1316	92.99	92.4177
	300	93.133	92.4177	93.4192	92.7039
	400	95.3453	92.5608	93.9914	92.7039
	500	96.1345	93.4192	94.1345	92.7039
	600	96.1359	94.5637	94.8498	92.7039
	700	96.279	95.279	95.279	92.7039

Figure 1 Model Accuracy vs Sample Size: Diabetes



From the figure1 we observe tool A performs better than other tools while incrementing data size with 100 instances as step size. Up to 400 instances, A and C dominates the B and D. Greater than 400 instances A and B dominates C and D.

Figure 2 Model Accuracy vs. Sample Size: Cancer



From the figure 2 we observe tool A performs better than other tools while incrementing data size with 100 instances as step size. Up to 300 instances, almost all tools perform closely and tool A dominates others for greater than 400 instances.

I. Conclusion

The experimental results obtained from ADTree of decision-tree models generated using systematic sets of progressive sample sizes for the data sets diabetes and cancer data sets. This experiment can be performed using each of 4 standard familiar data mining tools.

Acknowledgment

The authors would like to thank the management of Annamalai University for the support and encouragement for this research work.

REFERENCES

[1] V. Vapnik. Estimation of Dependences Based on Empirical Data [in Russian]. Nauka, Moscow, 1979.(English translation: Springer Verlag, New York, 1982).

[2] D.Udhayakumarapandian.,RM.Chandrasekaran., andA.Kumaravel "A Novel Subset Selection For Classification Of Diabetes Dataset By Iterative Methods" Int J Pharm Bio Sci ,5 (3) : (B) 1 – 8, July(2014)

[3] A.Kumaravel., Udhayakumarapandian.D.,Consruction Of Meta Classifiers For Apple Scab Infections , Int J Pharm Bio Sci, 4(4): (B) 1207 – 1213, Oct(2013)

[4] A.Kumaravel., Pradeepa.R., Efficient molecule reduction for drug design by intelligent search methods.Int J Pharm Bio Sci, 4(2): (B) 1023 – 1029,Apr (2013)

[5] https://www.waset.org/journals/waset/v68/v68-21.pdf world academy of science, engineering and technology, 2012.

[6] http://www.insight.nau.edu/downloads/Sample%20Size%20and%20Modeling%20Accuracy.pdf

[7] H.Dunham, Data Mining, Introductory and Advanced Topics, Prentice Hall, 2002

[8] Source about wekahttp://www.cs.waikato.ac.nz/ml/weka/ downloaded on 3rd august 2014

[9] L. Breiman, " RandomForests,"inMachine Learning, vol. 45, pp. 5-32, 2001.

[10] Steve R. Gunn., University Of Southampton,Support Vector Machines for Classification and Regression.

[11] Dietterich, T. G., Jain, A., Lathrop, R., Lozano-Perez, T. (1994). A comparison of dynamic reposing and tangent distance for drug activity prediction.Advances in Neural Information Processing Systems, 6. San Mateo, CA: Morgan Kaufmann. 216--223.

[12] A.Stensvand, T. Amundsen, L. Semb, D.M. Gadoury, and R.C. Seem. 1997. Ascospore release and infection of apple leaves by conidia and ascospores of Venturia inaequalis at low temperatures. Phytopathology 87:1046-1053.

[13] Website for attribute description http://archive.ics.uci.edu/ml/machine-learning databases/pima-indians-diabetes., accessed on 3rd august 2014

[14] Bal,Hp.2005.Bioinformatics-principles and applications.Tata McGraw-Hill Publishing company Ltd New Delhi.

[15] Bo.Th and Jonassen,I-2002 New feature subset selection procedures for classification of expression profiles.Genome Biology 3:research 00170.-0017.11

[16] Khalid AA Abakar & Chongwen Yua., Performance of SVM based on PUK kernel in comparison to SVM based on RBF

ISSN: 2278 – 1323

kernel in prediction of yarn tenacity, Indian Journal of Fibre & Textile Research, Vol. 39: (B) 55-59, March (2014).
[17] Steve R. Gunn., Support Vector Machines for Classification and Regression Technical Report., Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science .,10 May 1998
[18] F. Girosi., An equivalence between sparse approximation and Support Vector Machines.A.I. Memo 1606, MIT Artificial Intelligence Laboratory, 1997.
[19] N. Heckman., The theory and application of penalized least squares methods or reproducing kernel hilbert spaces made easy, 1997.
[20] G. Wahba. Spline Models for Observational Data. Series in Applied Mathematics,Vol. 59, SIAM, Philadelphia, 1990.
[21] Domingos, P., 1998, "Occam's Two Razors: the Sharp and the Blunt," Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, Menlo Park, CA: AAAI Press, pp. 37-43.
[22] Frey, L. and Fisher D., 1999, "Modeling Decision Tree Performance with the Power Law," Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics, San Francisco, CA: Morgan-Kaufmann, pp59-65.
[23] John, G. and Langley, P., 1996, "Static Versus Dynamic Sampling for Data Mining," Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, , AAAI Press, pp. 367-370.
[24] Lee, S., Cheung, D., and Kao, B., 1998, "Is Sampling Useful in Data Mining? A Case in the Maintenance of Discovered Association Rules," Data Mining and Knowledge Discovery, Vol. 2, Kluwer Academic Publishers, pp. 232-262.
[25] Mannila, H., 2000, "Theoretical Frameworks for Data Mining," SIGKDD Explorations, Vol. 1, No. 2, ACM SIGKDD, pp. 30-32.
[26] Musick, R., Catlett, J, and Russel, S., 1993, "Decision Theoretic Subsampling for Induction on Large Databases," Proceedings of the Tenth International Conference on Machine Learning, San Mateo, CA: Morgan Kaufmann, pp. 212-219.
[27] https://github.com/Dans-labs/recommender-systems/blob/.../datamining.r
[28] http://www.r-project.org/
[29] http://www.knime.org/
[30] http://rapidminer.com/

First Author D.Udhayakumarapandian received the MTech in Computer Science and Engineering and pursuing Phd Degree in Computer Science and Engineering from Annamalai University, TamilNadu.

He is currently working as an Assistant Professor at the Department of Computer Science and Engineering, Bharath University, Tamil Nadu, India. He has presented and published more than 8 papers in technical conferences and reputed Journals. His areas of research include Data Mining and its applications, Algorithms and Computer networks.

Second Author Dr. R. M. Chandrasekaran received the B.E Degree in Electrical and Electronics Engineering from Maduari Kamaraj University in 1982 and the MBA (Systems) in 1995 from Annamalai University, M.E in Computer Science and Engineering from Anna University and PhD Degree in Computer Science and Engineering from Annamalai University, Tamil Nadu, India in 1995,1998 and 2006 respectively.

He is currently working as a Professor as well the Controller of Examinations at the Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Tamil Nadu, India. From 1999 to 2001 he worked as a software consultant in Etiam, Inc, California, USA. He has conducted Workshops and Conferences in the Areas of Multimedia, Business Intelligence and Analysis of algorithms, Data Mining. He has presented and published more than 32 papers in conferences and journals and is the author of the book Numerical Methods with C++ Program (PHI, 2005). His Research interests include Data Mining, Algorithms, Networks, Software Engineering, Network Security, Text Mining. He is Life member of the Computer Society of India, Indian Society for Technical Education, Institute of Engineers, Indian Science Congress Association.

.

ISSN: 2278 – 1323