Hybrid Data Reduction Technique for Classification of Transaction Data

Ifiok James Udo, Babajide Afolabi

Hybrid Data Reduction Technique for Classification of Transaction Data

Ifiok J. Udo and Babajide S. Afolabi

Abstract— Data classification problems during the process of mining transaction data requires robust and efficient data reduction technique to guard against loss of essential level information. In this paper, we have addressed the concepts of data reduction in transaction processing systems. The tradeoffs of data reduction techniques are being presented and a hybrid technique for data reduction suitable for addressing classification problems of transaction data is proposed.

Index Terms— Database, hybrid data reduction, classification and transaction data.

1 Introduction

Data classification problems during the mining process of transaction data requires robust and efficient data reduction technique. This is to ensure accuracy in the resulting data as well as guarantee the retention of essential level information for the proper mining process. The extraction of important features embedded in the contents of data is also made possible due to the technique of data reduction technique that is being adopted. Nevertheless, transaction data is made up of quantity of instances and number of features and there exists one-to-many association constraint as a result of data normalization. However, concurrency which is of utmost importance in transaction processing systems is fully optimized with data normalization which render some data to be of little importance or redundant nature to such system.

Furthermore, organizations apart from incurring costs in terms of buying machineries (i.e. required hardware) and hiring required numbers of personnel to classify huge databases, may also fail woefully in business competitiveness which require just-in-time and accurate information made possible by being able to extract significant features from the contents of the available data. This is because the general view of data is inhibited to decision makers at the expense of informed decisions, thus making accurate information difficult to be obtained in overloaded databases.

Data reduction as an essential element of data preparation [1], seeks to reduce large databases to manageable sizes and also help decision makers to know the true dimensionality of its business database(s). In transaction processing system, where transaction data are contained, daily business operations are often being supported according to [2] and there may also abound some feature vectors and associations which are of little or no relevance to the system during data classification and related tasks. These many feature vectors and associations often result when the level of data normalization increases in transaction processing systems to check the level of data redundancy. Normalization which also brings multiple instance problems to the fore is indispensable for achieving high level of concurrency and reduced redundancy in transaction processing systems.

The justification for data reduction apart from reducing the cost of data management and aiding informed decision making, reduction of data are advantageous in many other fields of Computer Science [3] such as data mining, information retrieval, web intelligence, machine learning and knowledge discovery in databases to name but just a few. Data reduction is also aimed at diminishing the quantity of irrelevant data in databases to enhance the quality of data for further analysis.

Research approaches to addressing data reduction problems have concentrated efforts in developing task-specific techniques which are often times purpose-built (i.e. to suit only a particular computing task); thus tending to neglect a multi-purpose approach that tends to be more scalable and robust in nature.

In this paper, we have addressed the concepts of data reduction in transaction processing systems, the tradeoffs of the existing reduction approaches have also been reviewed and the proposed hybrid data reduction technique suitable for addressing classification problems in relational data mining and also capable of performing the twin functions of reducing a transaction data in both features and sizes concurrently is also presented.

The remaining part of this paper is structured as follows: section 2, focuses on the concepts of data reduction including its merits and section 3 addresses related works and some drawbacks. Transaction processing systems and multi-relational setting as a store for transaction data is discussed in section 4. The proposed hybrid data reduction approach is subsumed in section 5 and lastly,

- *I.J. Udo is with the Information Storage and Retrieval Group, Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria.*
- *B.S. Afolabi is with the Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria.*

conclusion is drawn in section 6.

2 Data Reduction

The concepts of data reduction came into existence in a bit to make databases reflect the true dimensionality of business enterprises in which they represents, apart from rendering data for easy to manage. Data reduction otherwise known as data editing, filtering, thinning and condensing, depending on the objective of the tasks to be achieved is also considered to be an essential element of data pre-processing [4] aimed at preparing data for mining and/ or further machining to help improve the quality of the data obtained as a consequent of reduction operations. Data reduction also finds its usefulness in data classification and document retrieval among others. In transaction processing systems where record often exhibits one-to-many association constraints due to high level of data normalization, data reduction is important because a single record described by many associations with feature vectors may but only few are relevant for the observed classification and other related tasks performed on such records. An ideal reduction of data in transaction processing systems allows decision makers to know the real dimensionality of its enterprise database(s), thus leading to an informed decision making. The need for concurrency and data consistency that can be made possible by reduction of irrevlevant data in transaction processing systems cannot be over- emphasized.

2.1 Importance of Data Reduction

Although data reduction is not without a disadvantage (otherwise known as the curse of dimensionality), many advantages accrue to scientists, decision makers and analysts as a result of making data size(s) reduced.These benefits that far outweighs the disadvantages are among others include:

1. The reflection of the true dimensionality of business database(s).
2. Optimal usage of minimal or limited memory size.
3. Efficient and fast retrieval of data.
4. Increased efficiency and performance in subsequent data analysis and data mining.
5. Visualization of high dimensional data for explorative data analysis.
6. Low bandwidths consumption during data transmission as a result of data compression.
7. Cost reduction in terms of required numbers of personnel and storage requirements.

3 Related Work

Many data reduction approaches in both transaction systems and data warehouses in the past abounds. According to [5], there are two major approaches of data reduction. These approaches are feature selection and data size reduction. The significant drawback of these reduction approaches are being majorly tending to task-specifics (such as data mining, pattern recognition, information retrieval, web intelligence and machine learning) rather than a general purpose approach. Other drawbacks of these approaches besides being limited to task-specific approach are that they also do not combine the twin functions of reducing the data size (i.e. number of samples reduction) as well as its dimensionality (i.e. number of features reduction or feature selection) reduction concurrently. Hence this work seeks to combine the functionalities of feature selection and data size reduction in a single algorithm to achieve the twin functions of reducing transaction data in sizes and features concurrently.

3.1 Feature Selection

This method is also referred as dimensionality reduction or reducing the number of features and has been implemented with dimensionality reduction techniques. It allows for compact representation of data by mapping each point to a lower dimensional continuous vector. It may be supervised, semi-supervised or unsupervised. According to [1], feature extraction and feature subset selection are the two main paradigms involved in dimensionality reduction techniques. Dimensionality reduction can also be achieved by method of aggregation [6], [7] and Graph embedding and extension [8] as well as Discernibility [9].

3.2 Data Size Reduction

Reducing the number of samples have been implemented with several methods such as sampling procedures (e.g. simple random sampling and stratified or cluster sampling). These methods are based on statistical sampling which view data as expensive resources and assumes that it is practically impossible to collect population data. This approach does not suit data reduction in databases where population data is assumed to be known. Other methods are Adaptive sampling [1] and adaptive sampling with Genetic Algorithm [10] and Discernibility [9]. Moreso, other approaches such as Wavelets [11] and Clustering [12] for reducing the quantity of instances have been used. The adaptive sampling approach which employs chi-square criterion in our view is simple and adaptive in nature. It segments data into categories to ease computation; but it is intractable with very large and high-dimensional data. The approaches of [1] and [10] are only based on dimensionality reduction.

4 Transaction Processing Systems

According to [2], transaction processing systems are databases that support daily business operations. It usually involves large number of users who simultaneously perform transactions to change real time data. Concurrency and atomicity are the major characteristics of transaction processing systems. In a bit to enhance concurrency, data consistency and reduced redundancy, data normalization is often used in these systems. Moreso, one-to-many association constraints which pose multiple-instance problem [7] is brought to the fore due to high level of data normalization. Therefore, transaction processing environment depicts a multi-relational setting as described in fig. 1.

The representation in fig. 1 depicts part of a higher

education electronic portal schema in Nigeria for instance, whereby a student (i.e. a single object) may offer more than one course (i.e. a multiple-instance) and each course (i.e. a single object) may also be assigned to more than a lecturer (i.e. a multiple-instance) in the relation Student-Course-Lecturer schema. The notation "RN" on students' table stands for the student registration number, "CC" on course table stands for course code, and LID stands for lecturer's identity number. The two levels of one-to-many association comprising student, course and lecturer relationship is as shown in fig. 1. This illustration presents a student one-to-many relationship with course relation, through the association with Lecturer through the association of the Title (i.e. Course title) and LID (i.e. Lecturer Identity Number). The objects in fig. 1 can be represented in a vector space model [13]; therefore making it possible to be manipulated algebraically.

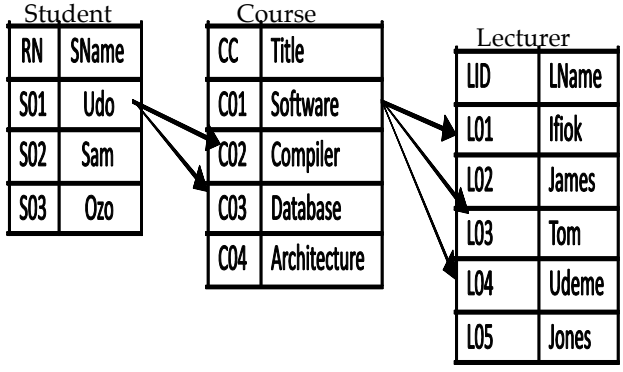

Fig. 1. Transaction data schema showing one-to-many association constraints.

2.1 Data Model in a Multi-Relational Setting

According to [14], relational data model (X) comprises number of features (J) and quantity of instances(N); the basis of which relational databases evolved. The transaction data model is therefore presented mathematically in equation (1). The equation (1) simply shows that tansaction data is a function of quantity of instances (i.e. number of records) and number of features (i.e. number of attributes) it is made up of.

$$X = (N\ x\ J) \tag{1}$$

Due to increasing volumes of data in transaction processing systems, database(s) grows in both features and sizes, thereby making data normalization an appropriate means of checking data redundancy, hence optimizing concurrency and data consistency. This often leads to multiple-instance problems.

In the work of [15], records exhibiting one-to-many association characteristics in multi-relational setting have been encoded into target and non-target tables which can also be represented in a vector space model as presented in equation (2). This vector space model presentation will allow similar instances to be aggregated and also permit data size reduction to be effectively carried out on objects. The relational object model in multi-relational setting is given as shown in equation (2).

$$O_i = \left(rf_1.log\left(\frac{n}{of_1}\right), rf_2.log\left(\frac{n}{of_2}\right), \ldots, rf_m.log\left(\frac{n}{of_m}\right)\right) \tag{2}$$

Where rf_m is the frequency of mth representation in ith object, of_m is the frequency of the mth representation of in each object, or the number of objects containing feature "m", n is the total number of objects and $O_i\ \varepsilon\ DB$ and $j = 1\ to\ m, i = 1\ to\ n$. Also, DB denotes a database.

5 Hybrid Data Reduction Technique

Our objective of developing a hybrid data reduction technique is to allow reduction of transaction data in both the number of features and quantity of instances at the same time. During a hybrid data reduction we consider the need for a compact representation of the observed data features to ensure retention of essential level information. In the other hand, the reduction of data size is also achieved with the Adaptive sampling procedure so that the overall reduction process is simplified and is made interactive. These different approaches of data reduction are integrated into a single algorithm. The scalability and efficiency that is achieved in our proposed algorithm is quite significant, in terms of the quality of the resulting data, and when compared with existing reduction approaches. This measure can be achieved with the closeness-of-fit measure.

In our approach, a full dataset is considered as a known object which reduction operations are performed on it. The two aspects of data reduction are performed simultaneously starting from the feature selection and then followed by the data size reduction. The flowchart of the proposed multi-purpose data reduction approach is as shown in fig. 2. The feature selection phase in our approach adopts a feature subset selection paradigm [5] and it is performed by calculating the average distance to nearest neighbour object(s) using Euclidean distance measure [16]. The object(s) are then grouped depending on the computed distance (i.e. objects with the same distance measure are grouped together to form nodes). The maximum number of nodes to be formed (Q) is determined and calculated by equation (3) as shown.

$$Q = p_1 * J \tag{3}$$

where p_1 is the proportion of feature selection to be performed and J is the total number of data attributes.

To further carry out data reduction process, node centre(s) which uniquely identifies each node or clusters are determined from the respective nodes. This phenomenon is otherwise called as feature subset selection. These node centres are computed as the average of the connected features to each node or cluster as shown in equation (4).

$$z_q = \frac{1}{N_i} \sum_{x_{j \in B_q}} x_j , q = 1, 2, \dots, Q \quad (4)$$

Where z_q is the qth cluster or node centre, N_i is the ith number of interconnected features in node B_q (q= 1, …, Q) and x_j is the jth objects feature contained in the node.

The output of the feature selection phase is a full database with aggregated features and respective nodes centres is/ are therefore taken as input to the data size reduction problem. The data size reduction problem is formulated as a modified chi-square criterion that minimizes the frequency between the original and the reduced dataset. By subjecting data size reduction to binary quadratic equation presented in equation (5), our aim is to minimize the closeness-of-fit between the expected and the observed distributions while selecting the true data subset to ensure accuracy. However, we incorporate a pattern search technique to the algorithm to select a true data subset representative from the observed dataset. The data reduction process continues till a specified number of iterations and trials in iteration (i.e. stopping criteria) that the algorithm obeys to fathom feasibility are met. Our data size reduction model which is in line with the work of [1] is given by equation (4).

$$min\ X_m^2 = \sum_{q=1}^{Q} \sum_{z=1}^{Z_q} \frac{(n_{qz} - pN_{qz})^2}{pN_{qz}} \quad (5)$$

subject to:

$$\sum_{s=1}^{n} W_{sqz} = n_{qz} \quad (6)$$

$$W_{sqz} = \begin{cases} 1; & if\ W(s,q)\ is\ contained\ in\ X(NxQ) \\ 0, & otherwise. \end{cases} \quad (7)$$

where $W_{sqz}(q = 1, \dots, Q; z = 1, \dots, Z_q)$ is a row which must correspond to a row in $X\ (N\ x\ Q)$. The row vector in equation (7) indicates a Boolean vector, the entries of the vectors are either 1 or 0, depending on whether or not the selected data sample presents the result that matches the entry in the observed dataset. If the entry is found in the observed dataset the vector model assumed a value of 1 and 0 otherwise. X_m^2 denotes the modified chi-square criterion, n_{qz} and N_{qz} are the zth category of the qth attribute's node centre in the reduced and true dataset respectively, $p = n/N$ is the sampling proportion of the data size reduction, $Z_q(q = 1, 2, 3, \dots, Q)$ is the qth frequency of features' node centres of the total attributes (J) and Q is the total number of nodes formed.

5.1 Flowchart of a Hybrid Data Reduction Technique

This is the step-by-step graphical description of data reduction algorithm based on multi-purpose approach. These steps are as as discussed below:

1. Initialization of dataset: The dataset to be used for the reduction process is defined and initialized. The preferred dataset should be large in quantity of instances and also high-dimensional.
2. The input and target values for the computation are obtained. These are parameters such as proportion of reduction in quantity (p_1) and number of features to be obtained after the execution of the algorithm (Q), number of iterations and number of trials within each iteration.
3. Feature selection is performed as discussed in subsection 3.1. The result is tested with the set reduction proportion of number of features until the feature selection phase is completed.
4. The result obtained from the feature selection phase is taken as the input to data size reduction phase. These results contain parameters such as the total number of nodes computed (Q), with respective node centres and the number of features that is/ are contained in each of the nodes to its centres. These parameters obtained and the quantity of instance in the observed dataset is used for the evaluation of the objective function as presented in equation (3). This is achieved by selecting a random sample (n_{qz}) of size same as the observed dataset (n_{qz}) and continuously swap it based on the criteria presented in equations (5), (6) and (7).
5. Step 5: The result of the reduced dataset is presented. This result can be tested to ascertain the level of its accuracy and error rate.
6. The algorithm terminates.

CONCLUSION

Modern days business organizations dependency on databases or data warehouses calls for adequate attention in addressing the issues of increasing volumes of data in such systems, which render the true dimensionality of business enterprise difficult to understand. However, this inherent problem can be minimized in organizational databases to assist decision makers in informed decision making. In transaction processing systems, where concurrency, atomicity and data consistency must be ensured, these systems functions are difficult to be optimized with large volumes of irrelevant data or data with redundant nature. As a sequel to the aforementioned problems, transaction data need to be fully reduced to reflect the real dimensionality of its enterprise. This can be achieved by adopting the hybrid technique to reduce transaction data by extracting significant and relevant features from the database while diminishing the quantity of irrelevant information.

While enhancing concurrency with data normalization, many associations and feature vectors that are irrelevant and/ or are of redundant nature to data classification and other related tasks can be sufficiently reduced with a hybrid data reduction technique. The proposed approach apart from being able to reduce transaction data in both features and sizes for all round usage can improve the performance of resulting databases with minimum relevant data, thus making transaction processing systems more efficient and scalable. The twin functions of reducing the number of features and quantity of instances are combined in a hybrid data reduction technique, thus minimizing the

overall time of reducing a transaction data.

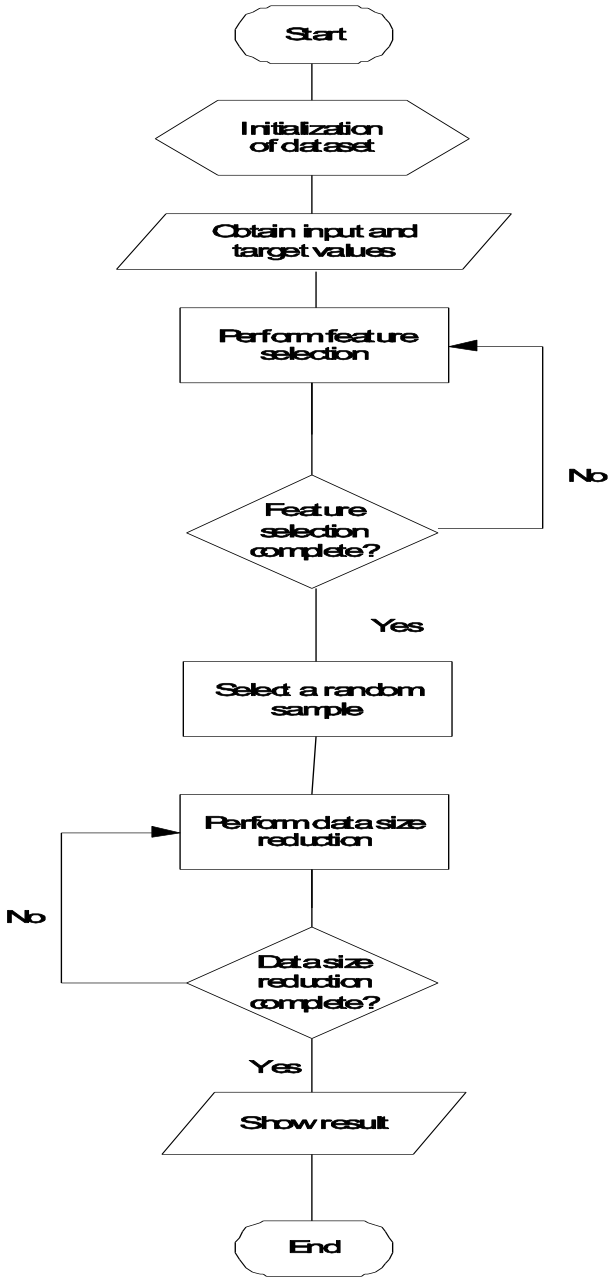

Fig. 2. A flowchart of a hybrid data reduction technique.

REFERENCES

[1] X. Li, "Data Reduction via Adaptive Sampling". Communication in Information and Systems, vol 2, no. 1. pp 53-68, 2002.

[2] E. Malinowski, and E. Zimanyi, "Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Application". Springer-Verlag, Available at http://www.springer.com/978-3-540-74404-7. Visited: November, 2008.

[3] Z. Zhang,, Zhang, C.. and Zhang, S. "An Agent-based Hybrid Framework for Database Mining" Applied Artificial Intelligence, vol.17 no.(5–6). pp 383–398, 2003.

[4] J.R. Cano, H. Francisco and L. Manuel, "Using Evolutionary Algorithms as Instance Selection for Data Reduction in KDD: An Experimental Study". *IEEE Transactions on Evolutionary Computation*, vol.7. no. 6, pp. 561-575, 2003.

[5] Q. Hu, Y. Daren and X. Zongxia. Information-preserving Hybrid Data Reduction based on Fuzzy-rough Techniques. *Pattern Recognition Letters*, vol. 27, no. 2006, pp.414-423, 2005.

[6] J. Skyt, C.S. Jensen and T.B. Pedersen, "Specification-based Data Reduction in Dimensional Data Warehouse". *Information Systems*, vol.33, no 1. pp 36-63, 2007.

[7] R Alfred and K. Dimitar. "Aggregating Multiple Instances in Relational Database Using Semi-Supervised Genetic Algorithm-based Clustering Technique". *In Local Proceedings of ADBIS, Varna*. pp. 136 -147, 2007.

[8] S. Yan, X. Dong, Z. Benyu, Z. Hong-Jiang, Y. Qiang. and L. Stephen. "Graph Embedding and Extensions: A General Framework for Dimensionality Reduction". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29 no. 1, pp.40-51, 2007.

[9] Z. Voulgaris and G.D Magoulos. "Dimensionality Reduction for Feature and Pattern Selection in Classification Problems". *ICCGI Proceedings of the Third International Multi-Conference on Computing in the Global Information Technology*. pp 60-65, 2008.

[10] X. Li, and V.S. Jacob "Adaptive Data Reduction for Large-Scale Transaction Data". *European Journal of Operational Research*, vol. 188. no. 3. pp 910-924, 2008.

[11] S. Russsell and V. Yoon "Application of Wavelets Data Reduction in a Recommender System". *Expert Systems* vol. 34. No. 4. pp 2316-2325, 2008.

[12] O. Okun and H. Priisalu, "Unsupervised Data Reduction. *In Signal Processing*, vol. 87, no. 9. pp. 2260-2267, 2007.

[13] G. Salton, A. Wong, C.S. Yang, "A vector space model for automatic indexing". *Communications of the ACM*, Issue.18, pp.613-620. 1975,

[14] E.F. Codd. A Relational Model of Data for Large Shared Databanks. *Communications of the ACM*, vol.13, no.6. pp 337-387. 1970.

[15] R. Alfred and K. Dimitar. "A Clustering Approach to Generalized Pattern Identification Based on Multi-instanced Objects with DARA". *In Local Proceedings of ADBIS,Varna*. pp. 38 – 49, 2007

[16] J.C. Gower. "Eucidean Distance Geometry". *Math. Scientist* vol. 7, pp. 1-14, 1982.

Ifiok J. Udo holds a B. Sc degree in Computer Science and he is at present an M.Sc student of the Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria and a member of Information Storage and Retrieval Group in the same department. His research work is on the Development of a Hybrid Data Reduction Tecnique for OLTP Environments. He has published articles in reputable Journals and Conference.

Babajide S. Afolabi holds a Ph. D in Information and Communication Sciences from Universite Nancy 2, Nanacy, France. He is the head of Information Storage and Retieval Group. He is a member of Nigerian Computer Society (NCS) and Computer Professional Registration Council of Nigeria (CPN). He is a senior lecturer in the Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria. He has published articles in reputable Journals and Conference.