Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 10 (2013) 781 – 787

Procedia Technology

International Conference on Computational Intelligence: Modeling, Techniques and Applications (CIMTA) 2013

Approximate Data Mining using Sketches for Massive Data

Parul Gupta*, Swati Agnihotri, Suman Saha

Jaypee University of Information Technology, Solan, India

Abstract

With the popularity of the Web and Internet, massive data is generated.However, this enormous datasets present the challenge to apply data mining techniques in order to extract useful information. Dimensionality reduction can be used to improve both efficiency and effectiveness while extracting information from data. In this paper we have proposed an algorithm to reduce the dimensionality of the datasets such that after applying data mining techniques on reduced datasets we get almost same results as with the original datasets. Random Sketch is used to reduce the dimensions of the dataset.

Selection and peer-review under responsibility of the University of Kalyani, Department of Computer Science & Engineering

Keywords: Dimension Reduction, Random Sketch.

1. Introduction

Data mining is to extract useful information from extremely large datasets such as to find recurrently occurring patters or finding similar items or clustering of data [1].Modern Internet applications have created a need to manage enormous amount of data promptly. The data is too massive that it does not fit into the main memory, therefore it is difficult to apply data mining techniques on the data. So, we need to devise techniques to reduce the datasets such that we get similar results, as with the original datasets, if we apply data mining techniques on reduced dataset.

The techniques used to process large datasets are (1) Parallel processing: Algorithms like BFR that processes data in parallel, in order to apply data mining on large datasets [2].(2)Dimension reduction: Done using Singular value decomposition, Random Projection or Sampling [1]. Dimensionality reduction techniques are well explored in databases [3,4].

In this paper we have proposed to reduce the dimensionality of the dataset such that data fits into the main memory and data mining techniques could be applied on the datasets in order to mine useful information from the data. We have used Random Sketch in order to reduce the dimensionality of the dataset. The size of the data, after applying Random Sketch has reduced considerably and then clustering is applied on the reduced datasets.

* Corresponding author
E-mail address: parulgupta39@gmail.com

Selection and peer-review under responsibility of the University of Kalyani, Department of Computer Science & Engineering
doi:10.1016/j.protcy.2013.12.422

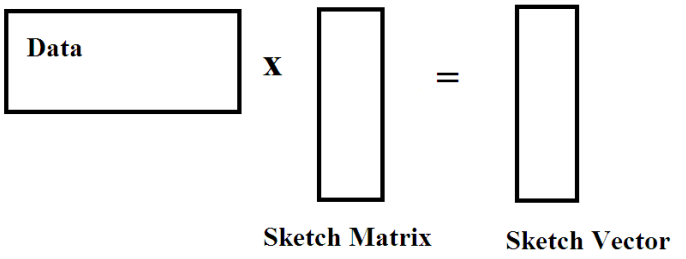

Fig. 1. Schematic view of random sketch.

1.1. Related Work

Problems of clustering as well as classification have been studied in data stream. Some methods for clustering in data stream are proposed in [5–7]. Partition based approach in [7] uses adaptation of k-means to create clusters over the entire data stream. Count-Min Sketch based approach for clustering massive domain data stream is discussed in [8] .Various techniques on classification for data stream have been studied in [9,10].Mining of massive domain data streams using sketch is discussed in [11]

2. Data Sketching

Our main focus in this paper is on massive data, that is, data is too large that it cannot be incorporated in the primary memory and also a lot of time is consumed while accessing data from the disk [12].Disk cannot transfer data to primary memory at more than a hundred million bytes per second. That is not a problem when the dataset is in megabyte. But in massive dataset (data in gigabytes or terabyte)it presents problems just accessing it. Therefore some techniques must be designed so that various data mining techniques could be applied on the massive datasets.

In this paper we are mainly focused on dimension reduction. It is used to map set S to S in space of much smaller dimensionality while preserving important properties of set S [1] .Hence, we have used Random Sketch, based on Johnson-Lindenstrauss lemma [13].This lemma states that a set of points in high dimensional space can be mapped to much lower dimension such that pairwise distance of the points in the higher dimensional space are almost preserved. The cardinality of the lower dimension space depends on the number of input points and degree to which the pairwise distance need to be preserved. Formally it can be written as follow:
Lemma: Johnson-Lindenstrauss Lemma For any such that $1/2 > \epsilon > 0$, and any set of points $S \epsilon R^d$ *with* $|S| = n$ upon projection to a uniform random k-dimensional subspace where k = O(logn), the following property holds with probability atleast 1/2 for every pair u, v ϵ S, $(1-\epsilon)\|u-v\|^2 \leq \|f(u)-f(v)\|^2 \leq (1+\epsilon)\|u-v\|^2$, where f(u), f(v) are projection of u, v

Random sketch is used to create compact synopsis or the summary of the data which is smaller than the original data. Sketches capture salient properties while occupying little memory. [14]

2.1. Random Sketch

In random sketch, at first relation is modelled as defining a vector or matrix, and then the sketch is formed by multiplying the data by a vector. Figure 1 is demonstrating it in which a fixed sketch matrix multiplies the data to generate the sketch (vector) [14]. Sketch vector forms the synopsis of the entire data, which is much smaller than the original data. Data mining algorithms can now be applied on the sketch vector.

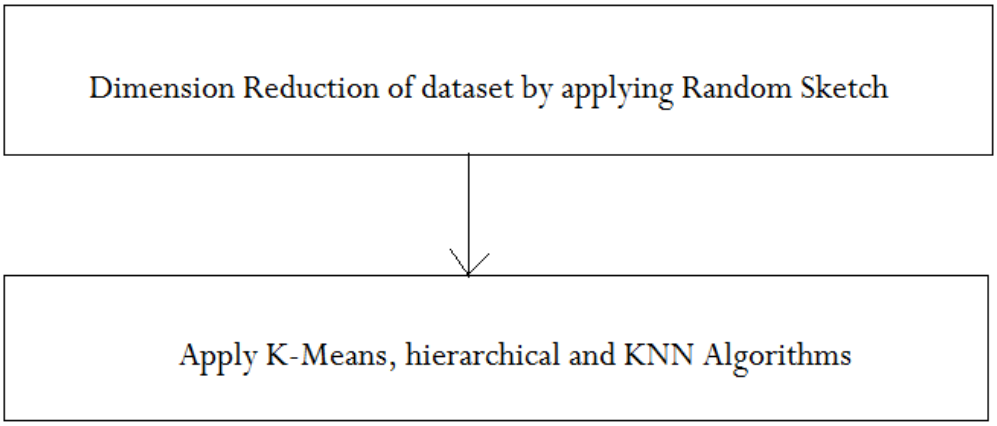

Fig. 2. Flowchart of Proposed algorithm.

3. Mining Sketches

Enormous data poses problems in mining, as the data does not fit into the main memory, therefore various data mining algorithms cannot be applied on the data. So, in this paper we are proposing an algorithm which reduces the dimensionality of the dataset and thereafter applying distance based clustering and classification algorithms on the reduced dataset. Figure 2 shows the flowchart of the proposed algorithm. In this the random sketch is used to reduce the dimensionality of the dataset and then k-Means, hierarchical clustering and K-nearest neighbour algorithms are applied on the reduced dataset.

In this algorithm we have used random sketch to reduce the dimensionality of the dataset. Step 1 counts the number of attributes in the dataset and in step 2 sketch matrix is generated which have rows equal to the number attributed in the dataset. Sketch matrix have 3 sets of values (1) between 0 and 1 (2) 0 and 1 (3)between maximum and minimum value of the dataset. Step 3 multiplies each row of the dataset with the sketch matrix in order to generate sketch vector. Once the Sketch vector is generated K-Means, Hierarchical clustering and K-Nearest Neighbour algorithms are applied on the Sketch Matrix.

3.1. Algorithmic steps

The proposed algorithm have the following steps:
Step 1: Count the number of attributes in the input dataset.
Step 2: Generate a random vector equal to number of attributes in the input file(Sketch matrix).
Step 3: FOR each row R in dataset DO
Multiply R with random vector and store the result in a file (so a random sketch of the initial dataset is created)
END;
Step 4: Apply clustering and classification algorithms on the reduced dataset.

4. Experimental Results:

In this section, we have evaluated the effectiveness and performance of our approach. We reduced the dataset using Java and then applied K-Means, Hierarchical clustering and KNN using Weka Tool[15]. The results presented in this paper are obtained from experiments conducted on Windows machine with 1.5GHz CPU and 2 GB of memory.

Random values selected	K-Means	Hierarchical Clustering	K NN
0 and 1	0 1 0 44 45 | y 1 21 20 | n	0 1 0 46 43 | y 1 20 21 | n	a b <-- classified as a 57 32 | a = y b 30 11 | b = n
Between 0 and 1	0 1 0 46 43 | y 1 22 19 | n	0 1 0 46 43 | y 1 20 21 | n	a b <-- classified as a 58 31 | a = y b 30 11 | b = n
Between maximum and minimum values in the dataset	0 1 0 46 43 | y 1 22 19 | n	0 1 0 35 54 | y 1 17 24 | n	a b <-- classified as a 57 32 | a = y b 29 12 | b =n

Fig. 3. Confusion matrix for different random values of sketch matrix.

4.1. Dataset

In our experiment we have used datasets URL Reputation. Dataset is taken from UCI repository [16]. The dataset is used to identifying Malicious URLs. It has 3231961 attributed to identify whether the URL is malicious or not. First attribute is the class (-1 or 1) -1 indicates that the URL is malicious and 1 indicates the data is not malicious.

4.2. Evaluation of Clusters

Weka tool evaluates the clustering using Classes to clusters evaluation. In this mode Weka first ignores the class attribute and generates the clustering. Then during the test phase it assigns classes to the clusters, based on the majority value of the class attribute within each cluster. After that it computes the classification error, based on this assignment and also shows the corresponding confusion matrix. An example of this for k-means is shown below.

```
   0   1 — assigned to cluster
0  5   4 — yes
1  3   2 — no
```

Above is the confusion matrix which indicated that 5 items of cluster 0 was assigned correctly and 4 items were assigned incorrectly. Similarly row 2 represents that 3 out of 5 items were assigned correctly and 2 assigned incorrectly.

4.3. Results

The proposed algorithm is also applied on URL Reputation dataset. In this the size of original file is 1.58GB and the size of resultant file which we get after applying our proposed algorithm is1.47KB.Time taken to perform k-Means, Hierarchical Clustering was .02 sec and for KNN .01 sec . Figure3 is demontrating confusion matrix which is obtained for different random values of sketch matrix when K-Means, Hierarchical clustering K-Nearest Neighbour algorithms are applied on sketch vector.

Figure 4 shows the percentage of data correctly assigned to the clusters when K-Means is applied on the reduced dataset i.e. the sketch vector. X-axis represents the different random values taken of sketch matrix. Pre-eminent

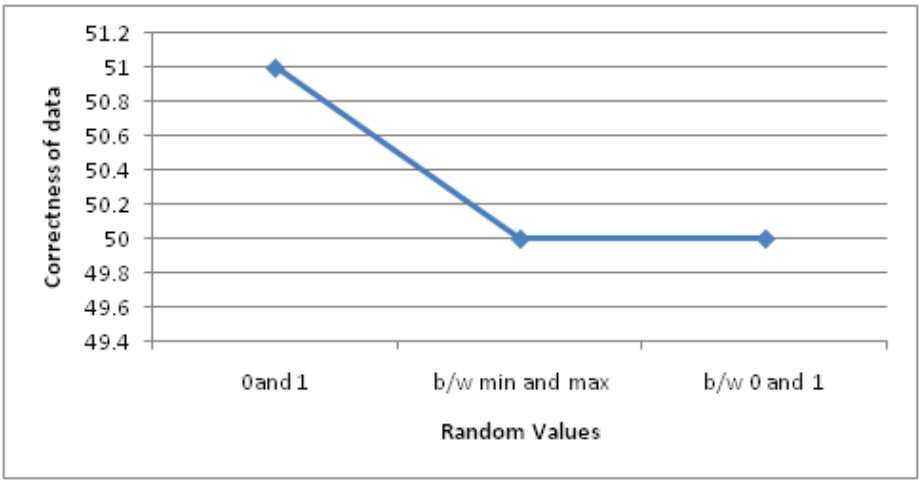

Fig. 4. K-Means applied on sketch vector for different values of sketch vs correctness of data.

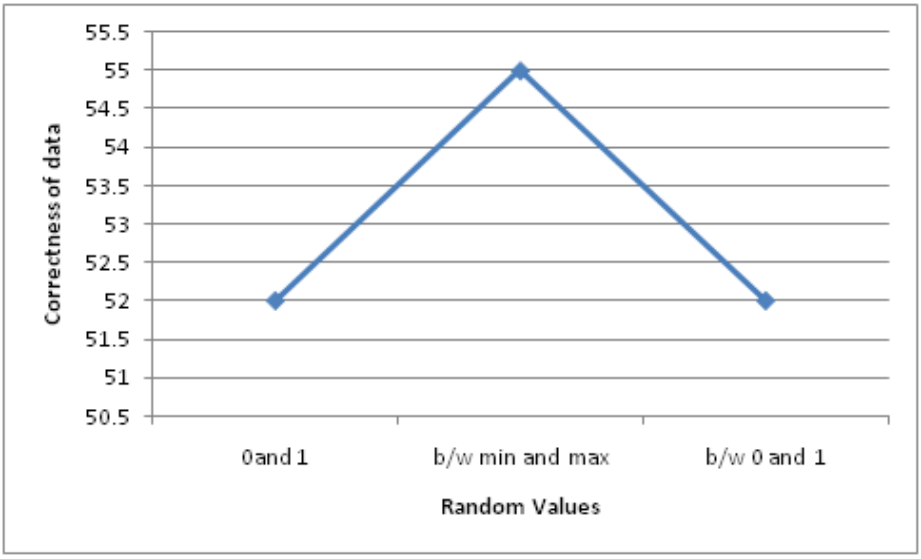

Fig. 5. Hierarchical Clustering applied on sketch vector for different values of sketch matrix vs correctness of data.

results were obtained when sketch matrix has values between 0 and 1.

Figure 5 shows the percentage of data correctly assigned to the clusters when Hierarchical Clustering is applied on the reduced dataset i.e. the sketch vector. X-axis represents the different random values taken for sketch matrix. Pre-eminent results were obtained when sketch matrix has values between maximum and minimum values of the dataset.

Figure 6 demonstrates the percentage of correctly classified data after applying K-Nearest Neighbour algorithm on sketch vector obtained. X-axis represents the different random valued taken and Y-axis percentage of correctly classified data. Pre-eminent results were obtained when random values are between minimum and maximum values of dataset and also for if the random values are between 0 and 1.

In k-means ,hierarchical clustering as well as in K-Nearest Neighbour algorithm if the values were taken between minimum and maximum values of the dataset correctness of data is best in all cases. Therefore if we take random

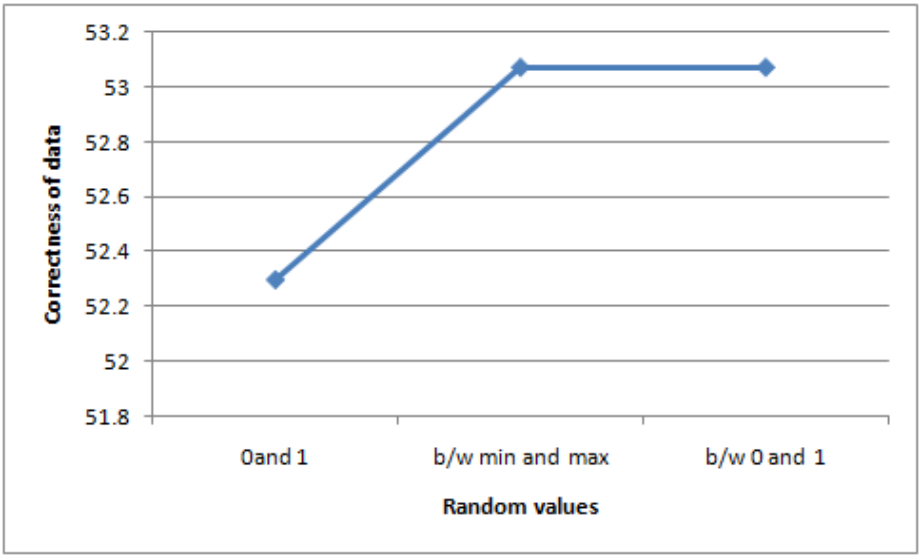

Fig. 6. K-Nearest Neighbour Classification algorithm is applied on the sketch vector for different values of sketch matrix vs correctness of data.

values in sketch matrix to be in the range of maximum and minimum value of the dataset and thereafter applying distance based clustering and classification algorithms on sketch vector we can get good enough results.

5. Conclusion

We have presented an algorithm for dimensionality reduction of massive datasets. Data that does not fit into the main memory is trimmed down using Random sketch and then distance based clustering and classification algorithms are applied on the reduced dataset which could not be done if we have the original data as it could not be incorporated into the primary memory .We have also presented results which shows that if the random value chosen are between maximum and minimum values of the dataset for sketch matrix, good results are obtained when K-Means, Hierarchical clustering and K-Nearest Neighbour algorithms were applied on sketch vector obtained after reduction of dataset.

References

[1] Afrati, F.N.. Efficient approximation and online algorithms. chap. On approximation algorithms for data mining applications. Berlin, Heidelberg: Springer-Verlag; 2006, p. 1–29.
[2] Bradley, P., Fayyad, U., Reina, C.. Scaling clustering algorithms to large databases. AAAI Press; 1998, p. 9–15.
[3] Faloutsos, C.. Indexing and mining streams. In: Proceedings of the 2004 ACM SIGMOD international conference on Management of data. SIGMOD '04; New York, NY, USA: ACM; 2004, p. 969–969.
[4] Garofalakis, M., Gehrke, J., Rastogi, R.. Querying and mining data streams: you only get one look a tutorial. In: Proceedings of the 2002 ACM SIGMOD international conference on Management of data. SIGMOD '02; New York, NY, USA: ACM; 2002, p. 635–635.
[5] Aggarwal, C.C., Han, J., Wang, J., Yu, P.S.. A framework for clustering evolving data streams. In: Proceedings of the 29th international conference on Very large data bases - Volume 29. VLDB '03; VLDB Endowment; 2003, p. 81–92.
[6] Guha, S., Meyerson, A., Mishra, N., Motwani, R., O'Callaghan, L.. Clustering data streams: Theory and practice. IEEE Trans on Knowl and Data Eng 2003;15(3):515–528.
[7] O'Callaghan, L., Mishra, N., Meyerson, A., Guha, S., Motwani, R.. Streaming-data algorithms for high-quality clustering. In: Proceedings of IEEE International Conference on Data Engineering. 2001, p. 685.
[8] Aggarwal, C.C.. A framework for clustering massive-domain data streams. In: ICDE. 2009, p. 102–113.
[9] Aggarwal, C.C., Han, J., Wang, J., Yu, P.S.. A framework for on-demand classification of evolving data streams. IEEE Trans Knowl Data Eng 2006;18(5):577–589.
[10] Domingos, P., Hulten, G.. Mining high-speed data streams. In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. KDD '00; New York, NY, USA: ACM; 2000, p. 71–80.

[11] Aggarwal, C.C., Yu, P.S.. On classification of high-cardinality data streams. In: SDM. 2010, p. 802–813.
[12] Rajaraman, A., Ullman, J.D.. Mining of Massive Datasets. New York, NY, USA: Cambridge University Press; 2011.
[13] Johnson, W., Lindenstrauss, J.. Extensions of Lipschitz mappings into a Hilbert space. In: Conference in modern analysis and probability (New Haven, Conn., 1982); vol. 26 of *Contemporary Mathematics*. American Mathematical Society; 1984, p. 189–206.
[14] Cormode, G., Garofalakis, M., Haas, P.J., Jermaine, C.. Synopses for massive data: Samples, histograms, wavelets, sketches. Found Trends databases 2012;4:1–294.
[15] Witten, I.H., Frank, E.. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2005.
[16] A. Asuncion, D.N.. UCI machine learning repository. 2007.