International Journal of Computer Science and Informatics

Volume 1 | Issue 3 Article 11

January 2012

A System for Outlier Detection of High Dimensional Data

Bharat Gupta
Department of Electronics and Computer, Indian Institute of Technology, Roorkee, India,
bharatguptagg@gmail.com

Durga Toshniwal
Deptt. of Electronics and Computer, Indian Institute of Technology, Roorkee, India, durgafec@iitr.ernet.in

Recommended Citation
Gupta, Bharat and Toshniwal, Durga (2012) "A System for Outlier Detection of High Dimensional Data," *International Journal of Computer Science and Informatics*: Vol. 1 : Iss. 3 , Article 11.
DOI: 10.47893/IJCSI.2012.1037
Available at: https://www.interscience.in/ijcsi/vol1/iss3/11

Available online at www.interscience.in

A System for Outlier Detection of High Dimensional Data

Bharat Gupta & Durga Toshniwal
Department of Electronics and Computer, Indian Institute of Technology, Roorkee, India
E-mail : bharatguptagg@gmail.com, durgafec@iitr.ernet.in

Abstract - In high dimensional data large no of outliers are embedded in low dimensional subspaces known as projected outliers, but most of existing outlier detection techniques are unable to find these projected outliers, because these methods perform detection of abnormal patterns in full data space. So, outlier detection in high dimensional data becomes an important research problem. In this paper we are proposing an approach for outlier detection of high dimensional data. Here we are modifying the existing SPOT approach by adding three new concepts namely Adaption of Sparse Sub-Space Template (SST), Different combination of PCS parameters and set of non outlying cells for testing data set.

Keywords- *Data mining, Projected outlier, Stream projected outlier detector, Sparse subspace template (SST), Base cell summary (BCS), Projected cell summary (PCS).*

I. INTRODUCTION

Outlier detection is an important research field in data mining. This technology is to find a small group of data points which are different from the rest of large amount of data based on some measures. These data objects are considerably dissimilar, exceptional and inconsistent with respect to the majority of data in an input database [1]. Detecting outliers efficiently is an important issue in data mining, which has important applications in the field of fraud detection, network intrusion detection and monitoring criminal activities in electronic commerce etc. Because of the sparsity of high dimensional data, it is reasonable and meaningful to detect the outliers in suitable projected subspaces. Actually in high dimensional data outliers are actually embedded in some lower dimensional subspaces. Here, subspaces refer to as the data spaces consisting of a subset of attributes. We call such subspace and outliers in the subspace as anomaly subspace and projected outlier respectively. Projected outliers exist in high dimensional data because when the dimensionality of data goes up, data tend to become equally distant from each other and so the difference of data point's outlierness will become increasingly weak and thus undistinguishable. In this situation significant outlierness of data can be observed only in moderate or low dimensional subspaces, [2]. This phenomenon is commonly referred to as the curse of dimensionality. Most of existing outlier detection methods perform detection in the full data space or work in low dimensional data sets so it is difficult to find projected outliers by these methods [3]. So now it is required to develop an efficient method to find interesting and potentially useful abnormal patterns hidden in high dimensional data.

Rest of the paper is organized as section 2 gives an overview of SPOT, in section 3 we propose our approach, section 4 shows the result and we conclude in section 5.

II. AN OVERVIEW OF SPOT

Here in this paper, we are using existing technique SPOT (Stream projected outlier detector) to detect projected outliers [4]. In this technique first we find outlying subspaces corresponding to each data record (tuple in database) and then on the basis of those outlying subspaces we can find projected outliers along with their lower dimensional subspaces where the data points are showing outlierness. For this purpose, this technology uses the concept of data synopsis:

A. Data Synopsis

Like many other data mining tasks such as clustering and frequent item detection, one of central problems involved is the design of appropriate synopsis that is suitable for outlier detection purpose. For this technique, data synopsis uses base Cell Summary (BCS) and projected Cell Summary (PCS), two compact structures that are able to capture the major underlying characteristics of the data stream for detecting projected outliers.

Quantization of BCS and PCS provide an equi - width partition of the domain space, which partitions each attribute into a few disjoint subsets. Specifically, let D be a set of ϕ-dimensional streaming data objects. For each attribute, the space is partitioned into a set of non overlapping intervals. All the intervals in each such dimension are of identical length. The cells in hypercube can be classified into two categories, i.e., the base cells and projected cells.

A base cell is a cell in hypercube with the finest granularity. The dimensionality (i.e., number of attributes) of base cell is equal to ϕ.

A projected cell is a cell that exists in a particular subspace s, which is a projection of a number of base cells from the full data space into s. The dimensionality of a projected cell is equal to |s| and $|s| < \phi$, where |s| denotes the number of attributes in s.

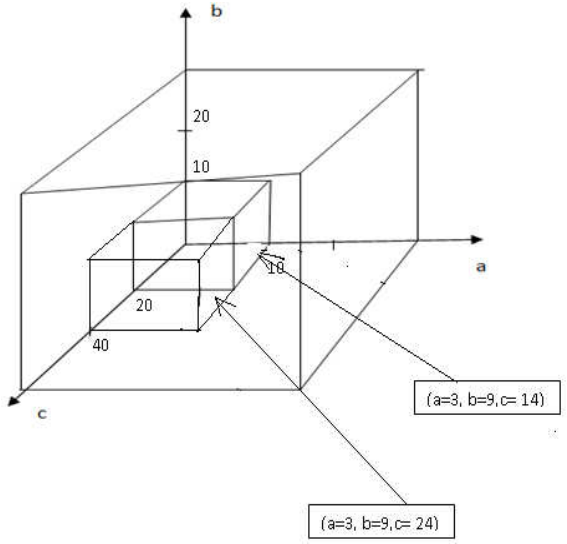

Fig. 1 : Allocation of data objects to base cell.

B. BCS:

BCS of a base cell c in the hypercube is defined as BCS(c) = {Dc, LSc, SSc), where Dc, LSc and SSc denote the number of points, the sum and squared sum of data values in each dimension of points in c respectively, LSc= vector sum of(p_i) and SSc= vector sum of (p_i^2), for pi located in c, $1 \leq i \leq \phi$. Dc is a scalar while both LSc and SSc are ϕ-dimensional vectors.

C. PCS:

Then with the help of already calculated BCS values for each base cell we will find out the Projected Cell Summary for each projected cell in all subspaces, where PCS value contain three parameters relative density and inverse relative standard deviation. The PCS of a cell c in a subspace s is defined as PCS(c, s) = (RD, IRSD, IkRD), where RD , IRSD and IkRD are respectively relative density, inverse relative standard deviation and inverse k relative distance of data points in c of s, respectively.

D. Outlying Cell:

A cell c in subspace s is called the outlying cell of a point p in s if one or more components of PCS(c, s) (i.e., RD, IRSD, IkRD) exceed their corresponding thresholds, where p is located in c.

E. Outlying Subspace:

An outlying subspace (s) of a point (p) is a subspace in which (p) is an outlier in a particular cell of (s). In other words, an outlying subspace (s) of (p) is a subspace that contains the outlying cell of (p) [5] [6].

F. Relative Density:

Relative density of a cell (c) in subspace (s) measures the relative density of c w.r.t. the expected level of density of non empty cells in (s). If the density of (c) is significantly lower than the average level of cell density in the same subspace, then the data can be labeled as outlier.

G. Inverse Relative Standard Deviation:

Inverse relative standard deviation of a cell (c) in subspace (s) is defined as inverse of the ratio of standard deviation of (c) in (s) against the expected level of standard deviation of non empty cells in (s). Under a fixed density, if the data in a cell features a remarkably high standard deviation, then the data are generally distributed more sparsely in the cell and the overall outlierness of data in this cell is high.

H. Projected Outliers:

A data point p is considered as a projected outlier if there exists at least one outlying subspace of p. Projected outliers are important entities as outliers that exist in high dimensional spaces. They carry abnormal behavior or patterns in a subset of attributes of the data. Aim of SPOT is to screen out these projected outliers from high dimensional database.

III. PROPOSED APPROACH

A. Adaptation of SST:

The proposed work is to implement adaptation of SST by using concept of outlier repository, where we will keep the record of all outlier points detected, along with there outlying subspaces. For adaptation of SST periodically remove those subspaces from SST where most of outliers (outliers in outlier repository) do not show there outlierness. Now add those subspaces in SST where most of the outlier points are detected as outliers. To improve SST find all outlying subspaces for outliers

in outlier repository and also add those subspaces in SST where outlier points show higher outlierness.

Thus it reduces time complexity due to the use of only sparse sub spaces instead of all sub spaces. Also it reduces false positive rate.

B. Usage of different combination of PCS parameters:

Earlier in outlier definitions we only used minimum threshold of PCS parameters. But here in this approach we will also add the concept of "Max. Threshold", because some times one PCS parameter may be less than minimum threshold value but its other 2 parameter may be very large. So in this case it should not be detected as outlier because here it has more probability to be as false result.

The benefits of this approach are thus by doing this we can reduce the false positive rate as the maximum threshold value will stop the false positive data to be detected as outlier.

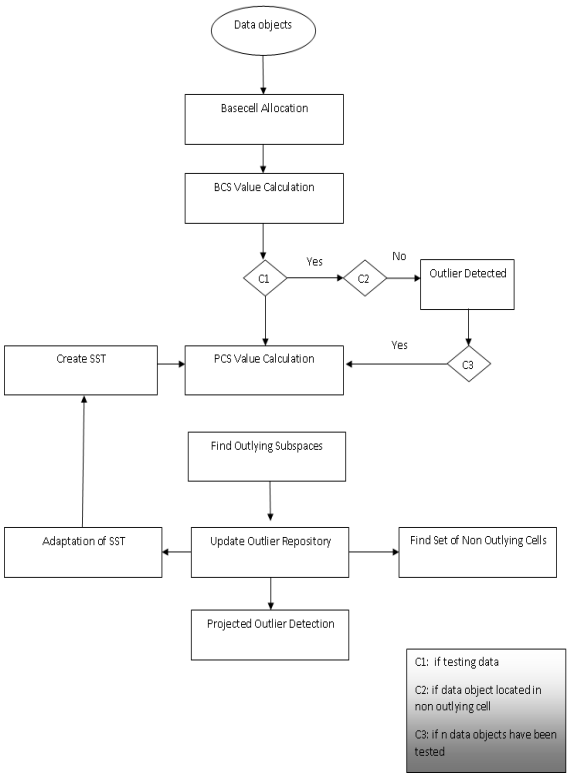

Fig. 2 : Diagram for Outlier Detection.

C. Checking in Non-Outlying cells:

With the help of training data set we will create set of non outlying cell. Then each data object of testing data set will be located, if it is not lying in non outlying cell then it will be outlier. If it will be located in non outlying cell then it will only increase PCS values rather our aim for finding outliers is to minimize the PCS values. Then for each data object we will update BCS value of its cell using incremental update property shown by BCS. As few data objects can't affect the PCS parameters of any cell, so we will update this set of non outlying cell after some fixed no of data object (periodically). And as we don't have to find BCS value each cell again so it will be very easy to update set of non outlying cell.

The benefits of this approach is therefore it reduces time complexity of program as do not need to find BCS values of all cells again because BCS show incremental update property. Also it will be easy to find outliers because we have to just check is it lying in non outlying cell or not.

Figure 2 shows the working of the proposed approach to find outliers.

IV. RESULTS

Here we are taking data set of breast cancer analysis from UCI machine learning repository. It is numerical data set with 10 attributes. In the data set only 6.7% person have cancer which we are treating as outliers with respect to data of all person in the data set. We trained and tested the above dataset under 5 different conditions. These conditions are listed in Table 1.

We performing the training and testing on the dataset with these varying conditions. Results obtained are shown in Table 2.

The graphs in Fig. 3 and Fig. 4 show that condition 1 has the least false positive rates.

TABLE – I : DESCRIPTION OF CONDITIONS FOR TRAINING AND TESTING

Condn	Included SST adaption	Included All PCS Parameter	Concept of Maximum Throughput	Extra Conditions
1	Yes	Yes	Yes	-
2	No	Yes	Yes	-
3	Yes	Yes	No	-
4	Yes	Yes	Yes	Check in Non outlying cell
5	No	No (Only RD)	No	-

TABLE II : RESULTS OF VARIOUS CONDITIONS

	Cond^n 1	Cond^n 2	Cond^n 3	Cond^n 4	Cond^n 5
True +ve	.062	.0625	.064	.043	.037
True -ve	.921	.919	.884	.943	.89
False +ve	.0112	.0139	.0494	.0114	.037
False -ve	. 0049	. 0044	.0274	. 0056	. 029

A. Result Analysis

In 1st condition we used both adaptation of SST and added concept of maximum threshold where both reduces false positive rate and adaptation of SST also reduces time complexity.

In 2nd condition we only added concept of maximum threshold which reduces false +ve rate effectively but did not reduce time complexity.

In 3rd condition we included adaptation of SST which slightly reduces false +ve rate but help to reduce time complexity.

In 4th condition we applied adaptation of SST and also added concept of maximum threshold to find set of non outlying cell then for testing data object we directly check those objects lying in non outlying cell or not.

In condition 5th we only used RD parameter of PCS and also did not included adaptation of SST so it gives large false +ve rate.

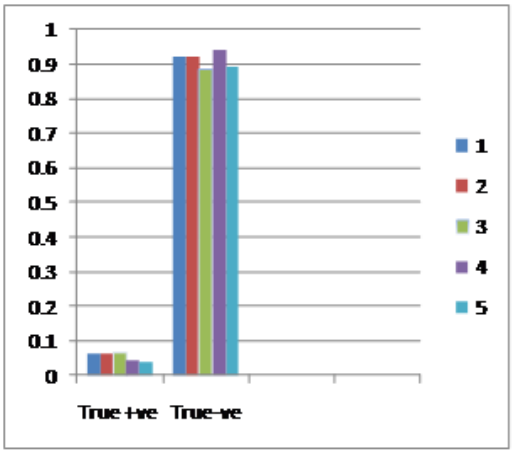

Fig. 3 : Graph showing the true +ve and True –ve of various conditions.

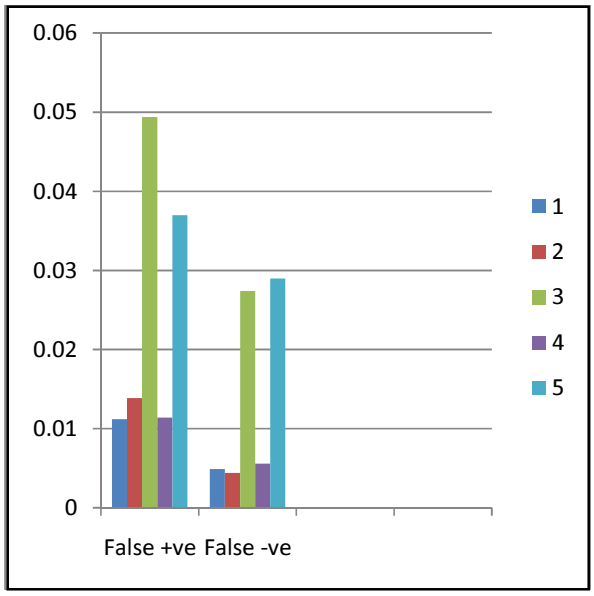

Fig. 4 : Graph showing the false +ve and false –ve of various conditions.

Here in 1st condition we applied our approach which is useful because it included adaption of SST which provide only sparse subspaces where it is more probability for data object to be detected as outlier, as it reduces no of subspaces to be checked it also affect the time complexity and false +ve rate. When we added concept of maximum threshold in outlier definition it reduces the chance of false +ve rate because even if, for non outlier data object its one PCS parameter has the value less then minimum threshold its other parameter will show higher value then maximum threshold.

V. CONCLUSIONS

SPOT uses the concept of compact data synopsis, which includes BCS and PCS to capture necessary data statistical information for outlier detection and because of efficient computation and maintenance of these two, it is possible for SPOT to meet the one pass constrained and time criticality posed by applications. If we use adaptation of SST it will reduce time complexity effectively and it will also help to reduce false positive rates. If we include different combinations of PCS parameters and add some concept of maximum threshold with already defined minimum threshold concept, it will also help to reduce the false positive rate. Specifically for testing data if we will check it is lying in set of non outlying cell, it will help to avoid use of any window based model and it will reduce complexity of method.

REFERENCES

[1] J. Han and M Kamber. “Data Mining: Concepts and Techniques”, Morgan Kauf - man Publishers, 2009.

[2] C. C. Aggarwal and P.S. Yu. “Finding generalized projected clusters in high dimensional spaces” ACM SIGMOD International Conference on Management of Data (SIGMOD’00), April 2000, pp 70-81.

[3] C. C. Aggarwal and P. S. Yu. “Outlier Detection in High Dimensional Data”, SIGMOD International Conference on Management of Data (SIGMOD’01), April 2001, pp 84-95.

[4] J. Zhang, Q. Gao, and H. Wang, "SPOT: A System for Detecting Projected Outliers From High-dimensional Data", Data Engineering, IEEE 24th International Conference, 7-12 April 2008, pp.1628-1631.

[5] J. Zhang and H. Wang. “Detecting Outlying Subspaces for High dimensional Data: The New Task, Algorithms and Performance. Knowledge and Information Systems (KAIS)”, Data Engineering, IEEE 22th International Conference, Jan 2007, pp. 333-355.

[6] R. Agrawal, J. Gehrke, D. Gunopulos and P. Raghavan. “Automatic subspace clustering of high dimensional data for data mining applications”,ACM SIGMOD International Conference on Management of Data (SIGMOD’98), May 1998, pp 94-105.

❑❑❑