Theoretically optimal distributed anomaly detection

Aleksandar Lazarevic*, Nisheeth Srivastava†, Ashutosh Tiwari*, Josh Isom*
Nikunj C Oza‡ and Jaideep Srivastava†
* United Technologies Corporation
Hartford, CT
† Dept of Computer Science
University of Minnesota, MN
‡ NASA Ames Research Center
Moffett Field, CA

***Abstract*—A novel general framework for distributed anomaly detection with theoretical performance guarantees is proposed. Our algorithmic approach combines existing anomaly detection procedures with a novel method for computing global statistics using local sufficient statistics. Under a Gaussian assumption, our distributed algorithm is guaranteed to perform as well as its centralized counterpart, a condition we call 'zero information loss'. We further report experimental results on synthetic as well as real-world data to demonstrate the viability of our approach.**

***Index Terms*—anomaly detection; distributed; data mining**

I. Introduction

In real-life situations where access to data in multiple locations can present a strategic advantage, a metaphorical prisoner's dilemma is often seen to unfold. For example, there are numerous cases where data cannot be localized at a central location (even by a trusted third party) for proprietary reasons, notwithstanding the advantages that may arise from such cooperation. In other cases, particularly concerning medical data, statutory requirements with respect to privacy rights render such cooperation infeasible even though, as before, such cooperation might prove to be advantageous to all parties concerned. In some situations, the logistics of transferring data to various locations are too imposing, e.g. when real time predictions are needed for mobile platforms such as an armada of ships. Finally, transferring large amounts of data to and fro would impose significant bandwidth costs and might prove to be a computational bottle-neck in situations where the amount of data at each location is large. Thus, opportunities for consolidating information from multiple sources, when weighed against the risks involved, create what could be called a 'distributed data mining prisoner's dilemma'. In this paper, we propose a novel solution to this problem in the specific data mining domain of anomaly detection.

Current research in anomaly detection using advanced data mining techniques has so far focused on detecting different types of anomalies from individual data sources [6], [9], [12]. However, in several real-life situations, the data containing anomalies may be dispersed across multiple locations. Therefore, there is a need for distributed algorithms that work well with limited exchange of relevant information from distributed sites in order to achieve anomaly detection on a global scale. If such distributed algorithms can guarantee detection of the same set of anomalies as the centralized approach, they may be seen to possess significant practical advantages over existing centralized methods.

In pursuit of this goal, we propose a novel framework for anomaly detection from distributed data sources that guarantees the same prediction performance as the centralized algorithm and exchanges very limited information across distributed sites. The rest of this paper is organized as follows: in Section 2, we briefly review existing work in this area. In Section 3.A, we present an efficient anomaly detection method based on T^2 and Q statistics methods [1] and show a novel distributed extension based on exchanging only local mean vectors and covariance matrices across distributed sites. Our experimental results performed on synthetic and several publicly available real data sets (documented in Section 4) indicate that the proposed distributed anomaly detection method is very effective in detecting anomalies. Finally, we conclude with some observations on possible lines of future research in Section 5.

II. Related Work

To solve the problem of detecting anomalies from distributed data sources, researchers have proposed several approaches: (i) modifications of simple distance based anomaly detection algorithms [2], (ii) tracking the changes in principal components [3], [4], and (iii) adapting standard ensemble schemes [5]–[12]. Recently proposed distributed version of distance based anomaly detection algorithm [2] is event based and requires exchange of set of data records across distributed sites. The number of data points that has to be sent to other distributed sites is still relatively large thus increasing time complexity of this distributed algorithm. Two methods for monitoring principle components [3], [4] have been used in the distributed eigen monitoring algorithm to detect changes in distributed and dynamic astronomical data streams. These methods are modifications of a very simple approach for anomaly detection based on the fact that the top few principal components capture the bulk of variability in a given data set, and the smallest principal components result in constant values. Therefore, any data point that violates this structure for the smallest components corresponds to an anomaly.

Using ensemble methods for distributed anomaly detection has also gained a lot of attention among researchers recently. However, there have been only a limited number of proposed techniques for distributed unsupervised learning. To detect anomalies from very large and distributed databases, some researchers have proposed modifications of standard distributed data mining framework [5]–[11]. In this framework, instead of merging all data at a central location, local models are built at each local data site, and global anomaly detection is performed by exchanging these local models. These local models are represented using different forms such as data boundary descriptions (minimal bounding rectangles, convex hulls [8], etc), or using specific machine learning models (neural networks [7], association rules [5], clustering [8], [11], Principal Component Analysis [6], genetic programming [8], etc). However, the main problem with exchanging local models thus far has been inability to guarantee the same prediction performance as the centralized method. In the absence of such guarantees, the case for using distributed computing for anomaly detection has to be weighed in the balance against the risk of losing performance by failing to detect anomalies. Since anomaly detection is generally sought in scenarios where false negatives have severe negative consequences, the lack of theoretical performance guarantees in distributed approaches is an important handicap in anomaly detection, even more so than in less false-negative sensitive data mining applications.

It is felt that our proposed approach represents an advance in that it successfully addresses this problem with limited data exchange among distributed sites. Not only does it guarantee the same performance as centralized anomaly detection, it also, in effect, parallelizes the existing algorithm by improving computational efficiency by the order of the number of sites. Thus, our approach may be used not only to perform distributed computing in relevant domains, but also to parallelize anomaly detection in scenarios where the data is originally centralized.

III. Methodology

In this section, we first present an efficient anomaly detection method from a single dataset based on T^2 and Q statistics algorithms, followed by a distributed version of this anomaly detection method.

A. Statistical anomaly detection

T^2 statistic based anomaly detection. T^2 statistics have been frequently used in statistical process control applications to detect various faults in multivariate datasets [1]. T^2 statistics can be computed directly by generating a PCA (principal component analysis) model of the multivariate data. PCA is a well-known multivariate technique that is used primarily for dimensionality reduction. Assume that $\mathbf{X}$ is a given data set of n data records which have m features. The PCA model is calculated using the singular value decomposition (SVD) on the autoscaled $\mathbf{X}$:

$$\frac{1}{\sqrt{n-1}} \cdot \mathbf{X} = U \cdot \Sigma \cdot V^T, \tag{1}$$

where U is a $n \times m$ orthogonal matrix, V is a $m \times m$ orthogonal matrix and Σ is a $n \times n$ diagonal matrix that contains positive real singular values of decreasing magnitude along its main diagonal $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{min(m,n)}$. Factorization in equation (1) is equivalent to solving an eigenvalue decomposition of the sample covariance matrix Σ.

$$\Sigma = \frac{1}{\sqrt{n-1}} \mathbf{X}^T \cdot \mathbf{X} = V \cdot \Lambda \cdot V^T, \tag{2}$$

where the diagonal matrix $\Lambda = \Sigma^T \cdot \Sigma \in \mathbb{R}^{m \times m}$ contains the positive real eigenvalues of decreasing magnitude and the i^{th} eigenvalue equals the square of the i^{th} singular value (i.e., $\lambda_i = \sigma_i^2$). In order to optimally capture the variations of the data while minimizing the effect of random noise corrupting the PCA representation, the loading vectors corresponding to a largest singular values are typically retained. These vectors are then stacked into a $m \times a$ loading matrix P forming the orthogonal basis for a-dimensional space. Thereafter, the T^2 statistic can be calculated using the following formula [16]:

$$T^2 = x^T \cdot P \cdot \Sigma_a^{-2} \cdot P^T \cdot x, \tag{3}$$

where x is an $m \times 1$ observation vector and Λ_a contains the the first a rows and columns of Λ. The threshold for the T^2 statistic [17] is given as:

$$T_\alpha^2 = \frac{a(n-1)(n+1)}{n(n-a)} \cdot F_\alpha(a, n-a), \tag{4}$$

where $F_\alpha(a, n-a)$ is the upper $100\alpha\%$ critical point of the F-distribution with a and n - a degrees of freedom.

Q statistics based anomaly detection. While T^2 statistics is typically focused on the a largest eigenvalues, Q statistic [16] monitors the portion of the observation space called residual space that corresponds to the (m-a) smallest eigenvalues. In other words, T^2 statistics captures the variations within the PCA model and Q statistics is the measure of amount of variation not captured by the PCA model:

$$Q = e^T \cdot e, \tag{5}$$

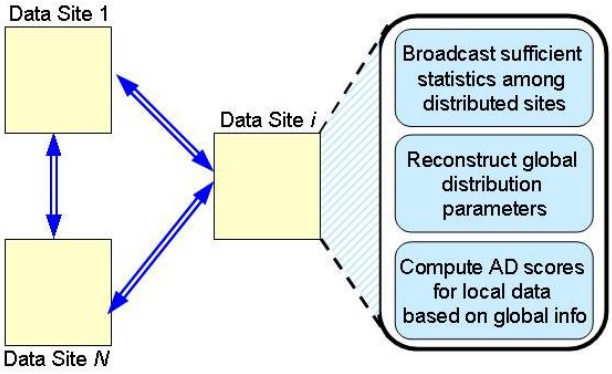
where $e = (\mathbb{I} - P \cdot P^T) \cdot x$ is the residual vector and $\mathbb{I}$ is the $m \times m$ identity matrix.

B. Distributed Anomaly Detection

The general framework for distributed anomaly detection based on T^2 and Q statistics based methods is shown in Figure 1. Assume there are N distributed sites, where site j contains data set S_j with n_j examples, $j = 1, 2, \cdots, N$. Data sets contain the same set of m features, but they do not necessarily have the same number of data records and cannot be completely merged for the purpose of anomaly detection. The distributed version of anomaly detection algorithm has to achieve exactly the same prediction performance as the centralized one when all distributed data sets are merged together. Therefore, T^2 and/or Q statistics anomaly detection scores computed from all distributed sites have to be exactly the same as in the scenario when all data sets are merged at a centralized site. It can be observed from equations 3

and 5 that both T^2 and Q statistics for the centralized data can be achieved simply by computing the covariance matrix Σ and loading matrix P for the centralized data. Here, we present a simple approach of determining the global covariance matrix without merging all distributed data sets but by simply exchanging the local covariance matrices and mean vectors. In the scenario of N disjoint sets $S_j, j = 1, 2, \cdots, N$, centralized data set S would be computed as the union of all S_j sites $(S = S_1 \cup S_2 \cup \cdots \cup S_N)$.

Fig. 1. Schematic of a general framework for distributed anomaly detection.



Each distributed set is considered as an independent site with limited communication capability with other sites. Simple covariance matrix Σ for the data set S would be computed as:

$$\Sigma = \frac{1}{n-1}(S-\mu)^T \cdot (S-\mu) = \begin{bmatrix} \sigma_{1,1} & \cdots & \sigma_{1,m} \\ \cdots & \cdots & \cdots \\ \sigma_{m,1} & \cdots & \sigma_{m,m} \end{bmatrix},$$

where,

$$\begin{aligned} \sigma_{i,j} &= \frac{1}{n-1}\sum_{k=1}^{n}(s_i(k)-\mu_i)\cdot(s_j(k)-\mu_j), \\ &= \frac{1}{n-1}\sum_{k=1}^{n} s_i(k)\cdot s_j(k) - \frac{n}{n-1}\mu_i \cdot \mu_j, \end{aligned} \tag{6}$$

n is the size of the dataset S, μ is the mean vector of S for all m features, i.e. $\mu = [\mu_1, \mu_2, \cdots, \mu_m]$. Considering the distributed scenario, however, sample covariance matrix Σ_l computed from the data present at site l may be calculated as:

$$\Sigma_l = \frac{1}{n_l-1}(S_l-\mu_l)^T \cdot (S_l-\mu_l) = \begin{bmatrix} \sigma^l_{1,1} & \cdots & \sigma^l_{1,m} \\ \cdots & \cdots & \cdots \\ \sigma^l_{m,1} & \cdots & \sigma^l_{m,m} \end{bmatrix},$$

where

$$\begin{aligned} \sigma^l_{i,j} &= \frac{1}{n_l-1}\sum_{k=1}^{n}(s^l_i(k)-\mu^l_i)\cdot(s^l_j(k)-\mu^l_j), \\ &= \frac{1}{n_l-1}\sum_{k=1}^{n} s^l_i(k)\cdot s^l_j(k) - \frac{n_l}{n_l-1}\mu^l_i \cdot \mu^l_j. \end{aligned} \tag{7}$$

The key question here is how to compute the covariance matrix Σ from N covariance matrices Σ_l and mean vectors $\mu^l, (l = 1, 2, \cdots, N)$ computed locally at N distributed sites.

The $(i,j)^{th}$ element of global covariance matrix defined in equation (6) can be rewritten as:

$$\begin{aligned} \sigma_{i,j} = \frac{1}{n-1}\sum_{l=1}^{N}\left(\sum_{k=1}^{n_l} s^l_i(k)s^l_j(k)\right) \\ - \frac{1}{n(n-1)}\sum_{l=1}^{N}\sum_{l'=1}^{N} n_l n_{l'} \mu^l_i \mu^{l'}_j \end{aligned} \tag{8}$$

Note that equation (8) is a result of splitting the first term in equation (6) among N distributed sites. Furthermore, the global means are written as a linear combination of local means, i.e.,

$$\mu_i = \frac{1}{n}\sum_{l=1}^{N} n_l \mu^l_i.$$

The sum $(\sum_{k=1}^{n_l} s^l_i(k)s^l_j(k))$ in equation (8) can be substituted using equation (7), thus resulting in the following expression for the $(i,j)^{th}$ element of the global covariance matrix:

$$\begin{aligned} \sigma_{i,j} = \frac{1}{n-1}\sum_{l=1}^{N}\left((n_l-1)\sigma^l_{i,j} + n_l \mu^l_i \mu^l_j\right) \\ - \frac{1}{n(n-1)}\sum_{l=1}^{N}\sum_{l'=1}^{N} n_l n_{l'} \mu^l_i \mu^{l'}_j. \end{aligned} \tag{9}$$

In matrix notation, this equation may be written as:

$$\begin{aligned} \Sigma = \frac{1}{n-1}\sum_{l=1}^{N}\left((n_l-1)\Sigma^l + n_l \mu^{l,T}\cdot\mu^l\right) \\ - \frac{1}{n(n-1)}\sum_{l=1}^{N}\sum_{l'=1}^{N} n_l n_{l'} \mu^{l,T}\cdot\mu^{l'}. \end{aligned} \tag{10}$$

Here, μ^l is the row mean vector of the dataset corresponding to the l^{th} distributed site. Equation (10) provides a method to exactly determine the global covariance matrix from the local covariance matrices and local mean vectors, thus allowing exact computation of global T^2/Q statistics as in the centralized scenario when all data sets are merged together.

Space constraints require us to be extremely brief in our description of the theoretical properties of our algorithm. If the underlying data may be assumed to be drawn from a Gaussian distribution, our distributed algorithm is guaranteed to perform as well as a centralized one, since the statistics shared between data sites (mean and covariance) are sufficient to reconstruct the density function representing the underlying data distribution. Our algorithm, thus, displays 'zero information loss'.

IV. Experimental results

In this section, we report results on on synthetic data as well as on several real life data sets summarized in Table 1. Synthetic data corresponds to two scenarios where (i) data at multiple sites may be assumed to be drawn from the same underlying distribution (referred to henceforth as the homogeneous case) and (ii) when data at multiple sites are drawn from different heterogeneous distributions (the heterogeneous case).

A. Experiments on Synthetic Data

In anomaly detection, anomalous data records are typically detected as deviations from normal data modeled as unimodal data distributions. Multi-modal data distributions are, of course, possible but are less frequently encountered. In practice, it is often found that a Gaussian distribution serves as an adequate fit for most two-sided unimodal data distributions, which accounts for its overwhelming prevalence of use (apart from considerations regarding simplicity of analysis). Section 3.4 shows how our algorithm can be applied in cases where the Gaussian assumption is inapplicable. Extending this generalization to account for multi-variate families of non-Gaussian distributions, however, is non-trivial. Therefore, for the purpose of our experiments, we have exclusively studied cases where the underlying data distributions can be modeled using Gaussians.

In the homogeneous case, approximately 20,000 data samples are drawn from a 30-dimensional Gaussian distribution at each of ten independent sites. A random number of anomalies are seeded within these ten data sites. The total number of anomalous data samples over the entire dataset is about 200 (0.1%), a percentage that corresponds to realistic domain applications. In the heterogeneous case, 20,000 data samples are again drawn at each of ten independent sites. However, in this case, data samples at different sites are drawn from different 30-dimensional Gaussian distributions. Anomalies are seeded in the same manner as in the homogeneous case, with a total number again approximately 200.

All results are described in terms of ROC curves, plotting the true positive rate (sensitivity), against the false positive rate (1 - specificity). T^2 algorithm had a perfect score on the homogeneous data, while the Q statistics did not perform as well on the same data. The relative inferior performance of the Q statistics can be explained by the complementarity of the T^2 and Q statistics. Since the T^2 statistics is a measure of variations within a PCA model, anomalous records that lie closer to the direction of first few eigenvectors would be easily discriminated from normal data records by the T^2 statistics. On the other hand, Q statistics is more useful when the anomalies are more apparent in the residual subspace i.e. when the anomalous records lie closer to the eigenvectors that are not considered in the PCA model. In our homogenous case, anomalies were generated closer to the primary eigenvectors, hence, the T^2 statistics turned out to be better suited for detecting anomalies, as opposed to the Q statistics. However, for the heterogeneous data the Q statistics performed better than the T^2 as the anomalies were seeded such that they were more visible in the residual subspace. Therefore, for the purpose of multivariate anomaly detection, it is recommended to monitor T^2 and Q statistics, simultaneously, as they will pick anomalies lying in different but complementary subspaces.

To conclude, both in cases where data at multiple sites is generated from the same normal distribution, where good performance is theoretically assured, and in cases where data is drawn from multiple normal distributions, where theoretical guarantees need not apply, our algorithmic approach performs well. In the latter case, note that while the generative distributions are different in terms of their moments, in several practical scenarios, they are not very far apart[1], which causes the merged distribution to continue to look relatively unimodal (though not as symmetric). Generally speaking, if the shared data distribution is not perturbed too far away from a relatively Gaussian form by this merging of local sufficient statistics, our anomaly detection algorithms should still be able to perform well on the associated data set. Note that the zero information loss guarantee will continue to hold in this case, since the sufficient statistics of local distributions are being combined. It is another matter that the combination of these statistics might create a merged distribution that is non-Gaussian and hence not very tractable for anomaly detection. Thus, performance will suffer, but no more than if the data samples had been combined in the first place. Hence, there will be a performance guarantee, even if the asymptote of the performance (the centralized case) performs extremely poorly.

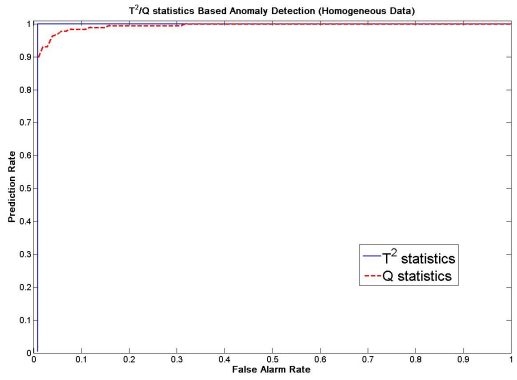
B. Experiments on Real Life Data

All real life datasets used in our experiments have been widely used by other researchers for anomaly detection. Table 1 gives a summary of those data sets. KDD CUP 1999 dataset includes a set of 41 features derived from each network connection and a label that specifies the status of a network connection record which is either normal or presents a specific attack type. Attacks fall into four main categories: DoS (Denial of Service), R2L (Remote to Local), U2R (User to Root) and Probe. We selected U2R, which covers only 246 instances, to detect the smallest intrusion class. Since the anomalies are detected as deviations from the normal behavior, we modified the original dataset (311029 records), and collected only normal class (60593 records) and U2R attack records for the experiment.

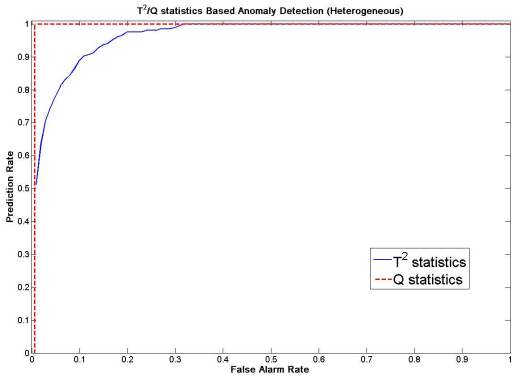
For Satimage dataset we chose the smallest class to represent anomalies and collapsed the remaining classes into one class. This procedure gave us a skewed 2-class dataset, with 5809 majority class examples and 626 minority class examples (anomalies). When performing experiments on mammography and rooftop datasets, we did not change any of the class distributions. All real life data sets have been split into ten subsets with approximately same number of data records per site.

For our experiments performed on real life data sets from Table 1, the computed ROC curves for both T^2 and Q statistics based anomaly detection approaches are presented in Figure 2. Analyzing Figure 2 bears out our earlier observation on the complementarity of our two anomaly detection algorithms. For the mammography and KDDCup99 data set, Q statistics had better detection performance while for rooftop and satimage data sets T^2 statistics was more successful. Thus, we reiterate our earlier observation that a sequential application of both

[1] e.g. functionally identical devices transmitting maintenance statistics from multiple locations will presumably have expected parameter values within a narrow range, though operational variance might be high.



(a) T^2 statistics based distributed anomaly detection for homogeneous data



(b) Q statistics based distributed anomaly detection for homogeneous data

Fig. 2. ROC curves for synthetic data. Data at multiple sites is drawn from the same multi-variate Gaussian distribution. Homogeneous data assumes all sites draw samples from the same multi-variate Gaussian distribution. Heterogeneous data assumes multiple sites drawn samples from different distributions.

TABLE I
SUMMARY OF DATA SETS USED IN EXPERIMENTS

Dataset	Modification made in the data set	Size of Dataset	Numbers of Continuous features	Numbers of Discrete features	Number of anomalies	Percentage of anomalies
KDDCUP 1999	U2R vs. normal	60839	34	7	246	0.4%
Mammography	-	11183	6	0	260	2.32%
Rooftop	-	17829	9	0	781	4.38%
Satimage	Smallest class vs. rest	16435	36	0	626	9.73%

anomaly detection methods should likely result in good results in at least one of the cases.

In all cases, performance of the distributed algorithm was identical to the performance obtained from data consolidation and centralized computation. This phenomenon occurs not because local sufficient statistics are transmitted accurately, but because the anomaly detection schemes use only the global covariance, which is obtained analytically from the sample means and covariance matrices as shown in Section 3. Since the actual distribution is not Gaussian, the statistical anomaly detection algorithms will not perform optimally well. Thus, in this case, the performances are identical not because of zero information loss, but because the error in transmission is pushed through to the anomaly detection stage as an artifact of the anomaly detection algorithm chosen. In practical terms, the high area under curve (AUC) for at least one of our two anomaly detection algorithms in all four cases presents significant evidence of the robustness of our algorithms for computation with real-life data drawn from arbitrary distributions, which in turn suggests that our algorithm may be useful in real world applications.

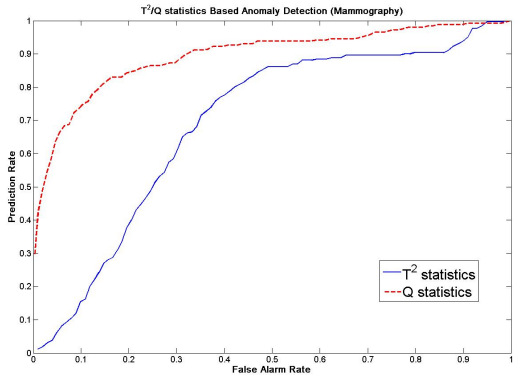
C. Computational efficiency

The time complexity for centralized T^2 and Q statistics based anomaly detection is $O(nm^2)$. In the distributed version of anomaly detection algorithm, each site computes PCA, computes local mean vector and covariance matrix, broadcasts them to all other sites, computes the global mean and covariance matrix and finally computes the anomaly detection score. The time complexity for all computations is $O(n_l m^2)$, while the communication overhead is minimal since the local mean is m-dimensional, and the local covariance matrix is of size $m \times m$. Thereby, computational savings of $O(\frac{n}{n_l})$ are achievable.

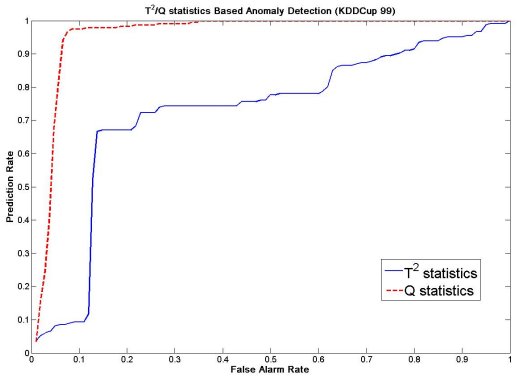
V. CONCLUSIONS

In this paper, we have presented an algorithm for distributed anomaly detection with a theoretical guarantee of detection performance equivalent to detection performance of a centralized algorithm, assuming that the parametric assumption we make on the distribution underlying the data is justified. Experimental results on real data demonstrate that our algorithm performs well even in cases where the theoretical guarantee does not hold explicitly.

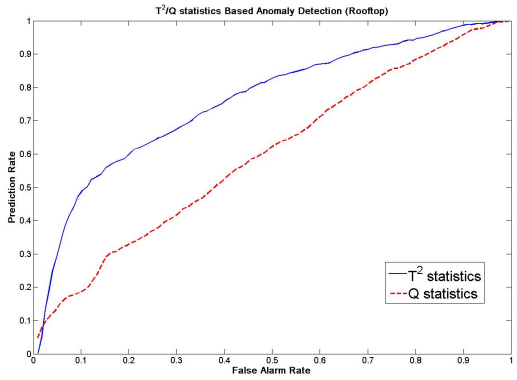
We foresee two avenues of promising work emerging from our proposed approach. The first avenue involves deeper investigation of our method of distributing sufficient statistics among data sites to anomaly detection methods (or indeed any other data mining algorithms) for more general distributions. We feel that since, for well-behaved (in our case, exponential family) distributions, sufficient statistics of the relevant distributions can be made available at all sites, it should be relatively straightforward to construct a viable scheme for zero information loss distributed anomaly detection (data mining) for any exponential family distribution. Second, since real world data is never exactly modeled by a Gaussian or any



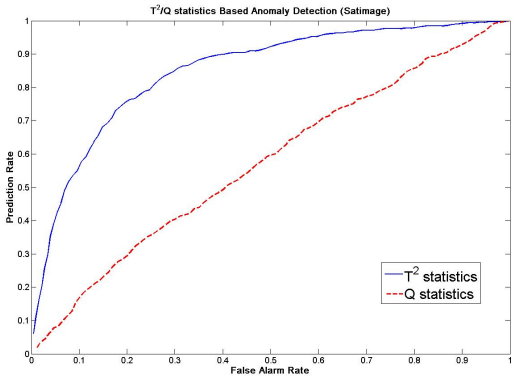
(a) Mammography data



(b) KDD data



(c) Rooftop data



(d) sat image data

Fig. 3. ROC curves comparing performance of T^2 and Q statistics based distributed anomaly detection algorithms on real data.

other standard distribution, it might be fruitful to design an information divergence based loss measure to compute the information loss with respect to the centralized case. This may then be further used as a measure of confidence in the distributed algorithm's predictions.

REFERENCES

[1] L.H. Chiang, E.L. Russell, R.D. Braatz, Fault Detection and Diagnosis in Industrial Systems, Springer; 1st edition, January 2001.

[2] J. Branch, B. Szymanski, R. Wolff, C. Gianella, H. Kargupta. (2006). In-Network Outlier Detection in Wireless Sensor Networks. Proceedings of the 26th International Conference on Distributed Computing Systems (ICDCS), 2006.

[3] H. Dutta, C. Giannella, K. Borne and H. Kargupta. (2007). Distributed Top-K Outlier Detection from Astronomy Catalogs using the DEMAC System. Proceedings of the SIAM International Conference on Data Mining, Minneapolis, USA, April 2007.

[4] K. Das, K. Bhaduri, S. Arora, W. Griffin, K. Borne, C. Giannella, H. Kargupta. (2009). Scalable Distributed Change Detection from Astronomy Data Streams using Local, Asynchronous Eigen Monitoring Algorithms. SIAM International Conference on Data Mining, Nevada. 2009.

[5] G. Deshmeh and M. Rahmati Distributed anomaly detection, using cooperative learners and association rule analysis,Intelligent Data Analysis,12(2008),pp. 339-357.

[6] V. Chatzigiannakis, S. Papavassiliou, M. Grammatikou and B. Maglaris Hierarchical Anomaly Detection in Distributed Large-Scale Sensor Networks,ISCC,2006, pp. 761-767.

[7] N. Srinivasan and V. Vaidehi Anomaly Detection in a Distributed Environment using Neural Networks on a Cluster, Proceeding Communication, Network, and Information Security,2005.

[8] S. Rajasegarar, C. Leckie, M. Palaniswami and J. C. Bezdek Distributed Anomaly Detection in Wireless Sensor Networks, Communication systems, ICCS, 2006 pp. 1-5.

[9] J. B.D. Cabrera, C. Gutirreza and R. K. Mehraa, Ensemble methods for anomaly detection and distributed intrusion detection in Mobile Ad-Hoc Networks Information Fusion, 9(2008), pp. 96-119.

[10] G. Folino, C. Pizzuti and G. Spezzano, 2005. GP Ensemble for Distributed Intrusion Detection Systems, 3683, pp. 54-62.

[11] Y.-F. Zhang, Z.-Y. Xiong and X.-Q. Wang, 2005. Distributed intrusion detection based on clustering, In Proc Int'l Conf. on Machine Learning and Cybernetics, 4, pp. 2379-2383.

[12] Otey, M. E., Ghoting, A., and Parthasarathy, S. 2006. Fast distributed outlier detection in mixed-attribute data sets. Data Min. Knowledge Discovery 12, 2-3, 203-228.

[13] P. Chan and S. Stolfo,On the Accuracy of Meta-learning for Scalable Data Mining, Journal of Intelligent Integration of Information, 1998.

[14] W. Fan, S. Stolfo and J. Zhang, The Application of AdaBoost for Distributed, Scalable and On-line Learning, In Proc. SIGKDD 1999, pp. 362-366.

[15] A. Lazarevic and Z. Obradovic, The Distributed Boosting Algorithm, In Proc. SIGKDD 2001.

[16] Jackson, J.E., 1959. Quality control of methods for several related variables. Technometrics 1, pp. 359377.

[17] MacGregor, J.F. and Kourti, T., 1995. Statistical process control of multivariate processes. Control Engineering Practice 3 3, pp. 403414.

[18] Y. Rosenberg and M. Werman, A general filter for measurements with any probability distribution, In Proc CVPR 1997, 654659