

Detection of Anomalies in Large Datasets Using an Active Learning Scheme Based on Dirichlet Distributions

Karim Pichara, Alvaro Soto, and Anita Araneda

Pontificia Universidad Católica de Chile

kpb@ing.puc.cl, asoto@ing.puc.cl, aaraneda@mat.puc.cl

Abstract. Today, the detection of anomalous records is a highly valuable application in the analysis of current huge datasets. In this paper we propose a new algorithm that, with the help of a human expert, efficiently explores a dataset with the goal of detecting relevant anomalous records. Under this scheme the computer selectively asks the expert for data labeling, looking for relevant semantic feedback in order to improve its knowledge about what characterizes a relevant anomaly. Our rationale is that while computers can process huge amounts of low level data, an expert has high level semantic knowledge to efficiently lead the search. We build upon our previous work based on Bayesian networks that provides an initial set of potential anomalies. In this paper, we augment this approach with an active learning scheme based on the clustering properties of Dirichlet distributions. We test the performance of our algorithm using synthetic and real datasets. Our results indicate that, under noisy data and anomalies presenting regular patterns, our approach significantly reduces the rate of false positives, while decreasing the time to reach the relevant anomalies.

1 Introduction

In this paper, we propose a new algorithm for the detection of anomalous records in large datasets. Depending of the domain, these anomalies may correspond to fraudulent transactions in a financial database, new phenomena in scientific information, or records of faulty products in a manufacturing database [6]. Our approach is based on the active learning paradigm. Under this scheme, our algorithm selectively asks a human expert for feedback searching for informative data points which, if labeled, would improve the performance of the overall process.

We build upon our previous work [4] [15] that allows us to efficiently find a Bayesian network (BN) [11] [10] to model the joint probability density function (pdf) of the attributes of records in a large database. This joint pdf provides a straight forward method to rank the records according to their oddness. In effect, while highly common records, well explained by the BN receive a high likelihood, strange records, poorly explained by the BN, receive a low likelihood.

Although our previous approach has shown to be effective in the detection of strange records, in practical applications the real relevance of an unusual record

is highly dependent of the domain under consideration. For example, in a fraud detection application, an unusual business transaction might not correspond to a fraud but it can be just a legal and irrelevant operation. As suggested by this example, our experience indicates that the raw unsupervised BN, constructed only from the low level features stored in a database, usually provides a great number of false positives.

In this paper, we augment our previous approach with an active learning scheme that helps us to bridge the gap between the blind unsupervised results provided by the BN and the domain knowledge provided by an expert. Our rationale is that while computers can process huge amounts of low level data, an expert can provide high level semantic knowledge to efficiently lead the search. In this way, starting from an initial set of candidate anomalies provided by a BN, our active learning algorithm selectively asks the expert for data labeling, looking for relevant semantic feedback to improve its knowledge about what characterizes a truly relevant anomaly.

The basic ideas behind our active learning approach are based on two main observations:

1) Our first observation is that, usually, the anomalies present in large databases are not isolated points but they exhibit certain regularities or patterns that arise in selective subspaces. In effect, in many domains, it is possible to find “types” of anomalies that form microclusters characterized by specific subsets of attributes of the database. The main goal of our active learning approach is to use the feedback from the expert to rapidly discover these microclusters.

2) Our second observation is rooted in a key feature of our probabilistic model, that is, the factorization of the joint pdf provided by the BN. From a clustering point of view, this factorization can be understood as model fitting in selective dimensions or subspaces. In effect, each factor in the joint pdf is given by a local conditional pdf over a subset of variables. These subsets of variables correspond to relevant subspaces of the feature space. As we explain in Section 2, our active learning approach makes use of the relevant factors provided by the BN and the clustering properties of Dirichlet distributions as the guiding tools to use the feedback provided by the user to find the micro clusters with relevant anomalies.

This paper is organized as follows. Section 2 discusses the details of our approach. Section 3 shows the results of applying our methodology to synthetic databases. Section 4 briefly reviews relevant previous work. Finally, Section 5 presents the main conclusions of this work.

2 Our Approach

This section describes the main steps of our active learning approach. As mentioned before, this algorithm actively asks for feedback from the expert to efficiently explore an initial set of candidate anomalies provided by a BN. In this exploration, the algorithm uses the factorization provided by the BN to identify key subspaces to detect the anomalies. Within the most prominent subspaces, the

algorithm identifies relevant microclusters that contain the anomalies. Our algorithm is based on three main steps: 1) Identification of initial set of candidate anomalies, 2) Selection of relevant subspaces using the factorization provided by a BN, and 3) Use of active learning to identify relevant microclusters. In the rest of this section, we refer to the details of these 3 steps.

In what follows, we use lowercase boldface letters, such as \mathbf{x} , to denote sets of single random variables, such as x_i . We use lowercase letters, such as x , to denote an instance of \mathbf{x} . We assume that the input database contains unlabeled observations and that there is no missing data.

2.1 Identification of Initial set of candidate anomalies

As a first step of our algorithm, we fit a BN to the records in the database (see [4] for details). If the training of the BN is successful, most anomalies appear as low probability objects. Therefore, we use the likelihood values provided by the BN as an indicator of the degree of rareness of each record in the database. This helps us to filter the dataset by identifying as candidate anomalous records only the first τ records with lowest likelihood. Deciding the correct value of τ depends directly on the capacity of the BN to fit the data. In our experience, the anomalous records usually fall between the 5 to 10% of the records with lowest probability under the BN model.

The factorization of the joint pdf provided by the BN allows us to efficiently estimate the likelihood of a record x in the database, as

$$P(x) = \prod_i^n p(x_i | Pa^G(x_i)) \quad (1)$$

where G is the acyclic directed graph that defines the BN, $Pa^G(x_i)$ is the set of direct parents of x_i in G , and n is the total number of attributes in the database.

2.2 Selection of relevant subspaces

Our next step is to find the relevant subspaces, or sets of attributes, that we use to identify microclusters containing anomalies. Our intuition is that the target subspaces are closely related to the factors provided by the BN, as these factors model the most relevant relations or patterns arising from the data.

In most situations, as it is in our experiments, the anomalies in the database correspond to a very small fraction of the total number of records, so we do not expect that the initial set of factors found by the BN highlights the target subspaces. Therefore, we fit a second BN to the reduced set of candidate anomalies, with the goal of obtaining a set of factors with a closer relation to the subspaces that determine the patterns of the anomalies.

For a given record in the candidate set of anomalies, we define its relevant factors as the set of factors that contribute the most to its likelihood obtained from the new BN. In particular, let \bar{p}_i be the mean value of factor $p(x_i | Pa^G(x_i))$

over the candidate anomalies k , $k = 1, \dots, \tau$. Also, let S be the set of all records with values of the i -th factor greater than \bar{p}_i . Then, for a given record we define the factor $p(x_i|Pa^G(x_i))$ as relevant, if:

$$p(x_i^k|Pa^G(x_i)^k) > \bar{p}_i + \delta, \quad (2)$$

$$\text{where } \delta = \frac{1}{|S|} \sum_{x^k \in S} (p(x_i^k|Pa^G(x_i)^k) - \bar{p}_i).$$

After obtaining the relevant factors, we relate them to the respective records. We visualize this as a bucket filling process. We represent each factor $p(x_i|Pa^G(x_i))$ of the joint pdf by a bucket i , $i = 1, \dots, n$. A record x from the set of candidate anomalies is included in that bucket, if the corresponding factor is relevant for that record. In this way, each record can be assigned to several buckets.

Given that our goal is to find microclusters with potential anomalies, we perform a clustering process within each bucket, using the elements inside the bucket. We use Gaussian Mixture Models (GMMs) to find the microclusters. Each GMM is trained with an accelerated version of the Expectation Maximization (EM) algorithm [15] that incorporates a model selection step to estimate a suitable number of Gaussians. When training the mixture in a given bucket, we use only the dimensions or attributes that identify that bucket. In this way, in bucket i the mixture is trained in the subspace generated by x_i and those attributes contained in $Pa^G(x_i)$.

After we find the microclusters, we assign to each datapoint a weight that is inverse to the initial joint probability value under the initial BN model. This ensures higher likelihood values to less probable elements according to the BN used to model all the data. As a result, the strangest records have a higher probability of being sampled in the next step the algorithm.

2.3 Use of active learning to identify relevant microclusters

In this step of our algorithm, we implement the active learning scheme that uses the feedback from the expert to guide the search for anomalous records. The selection of candidate anomalous records shown to the expert is based on the buckets and microclusters found in the previous step. The selection is performed by a 3-step sampling process: 1) First, we select one of the buckets. 2) Then, from this bucket, we select a microcluster and, 3) Finally, from this microcluster, we select the candidate anomalous record that is shown to the expert. According to the classification assigned by the expert to the selected record, we refine our model, increasing or decreasing the likelihood of retrieving again a relevant anomaly from the same bucket and microcluster.

We perform the 3-step sampling process and model refinement using a probabilistic approach, where probabilities represent the uncertainty about whether or not a bucket or microcluster contains anomalous records. We model the problem using Dirichlet distributions [5], and take advantage of their clustering properties. In particular, our microcluster selection process corresponds to an instance of a Polya Urn model [3].

In the process of sampling buckets, we use a Multinomial distribution to model the probability of selecting each bucket. Initially, each bucket has the same probability p_i of being selected. Later, after receiving feedback from the expert, we update the parameters of the Multinomial distribution according to the equations of a Polya Urn process [5]. In this way, a successful bucket increases its own probability of been selected while an unsuccessful bucket decreases this probability.

Once a given bucket is selected, we sample an observation from the GMM used to model the microclusters inside that bucket. Given that we use Gaussian functions, it is possible that the sample from the Gaussian mixture does not correspond to the position of a real record inside the bucket. To solve this problem, we use Euclidean distance to select the record in the bucket that is closest to the sampled observation. As we do in the bucket selection process, we also use the feedback from the user and the equations of a Polya urn process to update the probabilities of selecting a given microcluster from the relevant GMM. Here, the Dirichlet distribution controls the parameters of a Multinomial distribution over the set of mixture weights of the GMM. In this way, a positive feedback from the user increases the probability of selecting again a record from the given microcluster.

3 Results

We test our algorithm under different conditions using synthetic and real datasets. Synthetic datasets correspond to samples from GMMs, where we artificially add anomalous records as datapoints in microcluster of low density areas. In the real case, we use a dataset from the UCI repository [2], corresponding to a pen-based recognition of handwritten digits.

In the experiments, we use as a baseline method an scheme that we called BN-detection. This scheme consists on showing sequentially to the experts, the records sorted in ascending order according to the likelihood values provided by the BN applied to the complete dataset. In this way, we can observe the advantages of adding the active learning scheme.

3.1 Anomaly detection in synthetic datasets

To analyze the performance of our algorithm, we use synthetic datasets to conduct 3 main experiments: 1)Evaluation of anomaly detection, 2)Evaluation of capabilities to learn the relevant subspaces for the anomalous records, 3) Evaluation of sensibility under presence of noisy records.

The synthetic datasets contains 90.000 records and 10 attributes. We add to this dataset 2000 anomalous records, simulated on five different subspaces with 400 anomalies contained in each one. To build the BN that models the candidate set of anomalies, we use 10% of the records with lowest likelihood, i.e., $\tau = 9200$.

Figure 1 shows the number of anomalies detected by the algorithm in each of the 5 subspaces with anomalies versus the number of records shown to the expert.

We can see that the algorithm is able to detect around 90% of the anomalies when the expert has analyzed only 2% of the database.

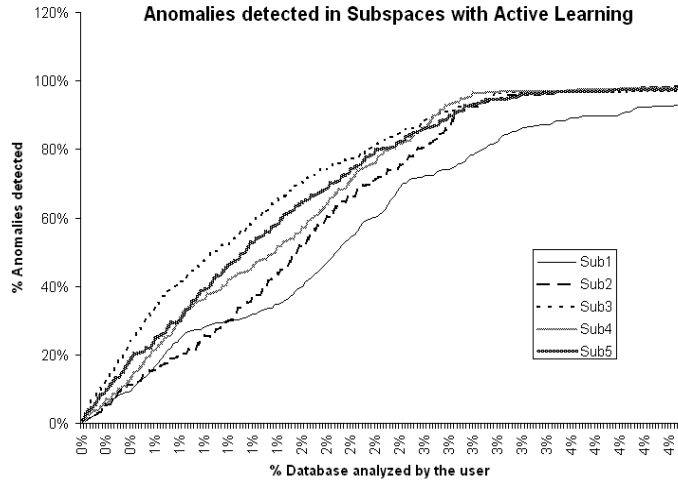


Fig. 1. Detection of anomalies in a database containing 90.000 records, 10 attributes, and 2000 anomalous records.

In terms of the impact of our active learning scheme to speed up the detection of the anomalies, our tests indicate that the proposed approach speed up in 25% the anomaly detection rate with respect to the baseline method.

3.2 Detection of relevant subspaces for anomalous records

To test the effectiveness of our algorithm to detect relevant subspaces where anomalies are generated, we conduct the following experiment. We simulate a database containing 11.000 records and 8 attributes. We add to this dataset 249 anomalous records, simulated on five different subspaces with 83 anomalies contained on each one. The anomalies were simulated on subspaces $S_1 = \langle x_1, x_3, x_5 \rangle$, $S_2 = \langle x_2, x_4, x_6 \rangle$, and $S_3 = \langle x_8, x_7 \rangle$. To build the BN that models the candidate set of anomalies we used $\tau = 2000$.

Figure 2 shows the percentage of the anomalous records detected in the different buckets provided by the BN factorization. In the figure, most of the anomalies generated in subspace $S_1 = \langle x_1, x_3, x_5 \rangle$ are detected in the bucket related to subspace $\langle x_3, x_5 \rangle$. A similar situation occurs with anomalies generated in the subspace $S_2 = \langle x_2, x_4, x_6 \rangle$, which are mostly detected in the bucket related to subspace $\langle x_4, x_6 \rangle$. This result shows that indeed, there is a close relation

between the factorization provided by the BN and the subspaces that are relevant for the anomalies.

In the case of subspace $S_3 = \langle x_8, x_7 \rangle$, there is an indication of some indirect relations occurring between variables in subspaces and buckets, where some variables are related with others subspaces through the BN structure. In effect, the bucket related to the subspace $\langle x_2, x_4 \rangle$ detects almost 20% of the anomalies from subspace $S_3 = \langle x_8, x_7 \rangle$, whose variables are not included in the set of variables that represent the bucket. Following the BN structure, however, we see that x_4 is a parent of x_8 and x_2 is a direct descendant of x_7 .

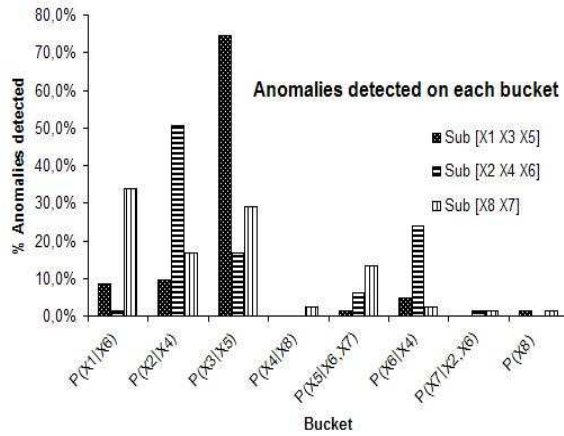


Fig. 2. Anomalies detected inside each bucket. Higher levels of detection are inside the buckets representing variables related to the subspaces where the anomalies are generated.

3.3 Sensibility under presence of noise

As this experiment we compare the performance of the BN-detection versus our algorithm under databases containing different levels of noisy records. Figure 3 shows the number of questions required to detect 100% of the total number of anomalies for a synthetic database with 15.000 records and 10 attributes, where 400 frauds are simulated on five different subspaces under different level of noisy records. The noisy records are generated using samples from uniform distributions in the range of values of each attribute.

Figure 3 shows that the performance of the BN-detection schemes does not scale well with the level of noisy records. This result was expected because most of the noisy records appear as low likelihood points, then they are shown to the expert. In contrast, the clustering properties of our active learning scheme

provides a better scaling with the level of noise because the search concentrates in the buckets and microclusters with relevant anomalies.

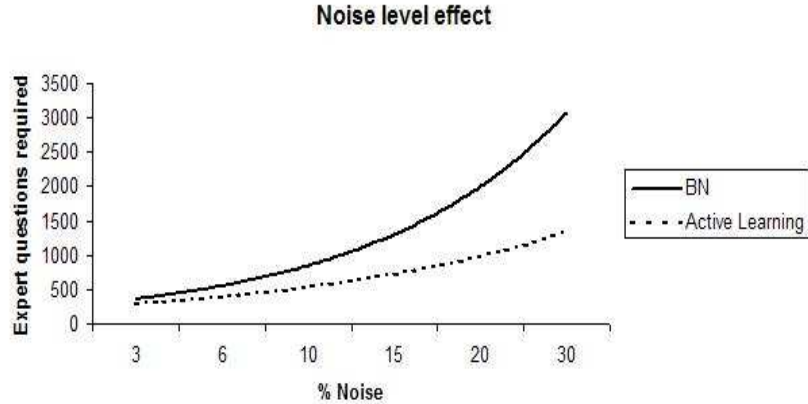


Fig. 3. How the Noise level increase the time required for anomaly detection. Active Learning is less sensible to noise.

It is relevant to note that the previous experiments were conducted under different conditions of database size, dimension, and number of anomalies. In all tests, we observed similar results to the ones shown in this paper.

3.4 Anomaly detection in a real dataset

We test our active learning approach and the BN-detection scheme with a real database containing 9000 records with 16 attributes. Each record corresponds to a handwritten character of one of 10 different classes. Here, two of these classes are considered as anomalous because they correspond to only 2% of the database. Figure 4 shows the number of detections with both approaches. Our algorithm detects 90% of the anomalies analyzing around 20% of the total dataset. Particularly, in the case of the first anomalous class, the number of anomalies detected is highly improved with the active learning scheme. For this case, the active learning based approach detects 150% more anomalies than the BN-detection scheme among the first 25% of the objects shown to the expert.

4 Previous Work

Given space constraints we just briefly review some relevant related work in the area of anomaly detection and active learning. The AI [7] [1] and the related

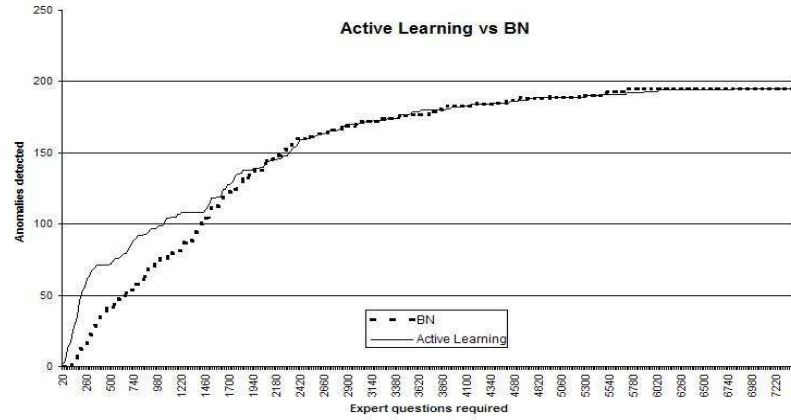


Fig. 4. Anomaly detection performance using the active learning scheme and the BN detection for both anomalous class on a real database.

Machine Learning communities [8] [6] have tackled the problem of anomaly detection, motivated mainly by applications on fraud detection. In contrast to our approach, most of these applications are based on supervised learning techniques. Unsupervised learning, such as clustering techniques, have also been used to detect anomalies. Using clustering, anomalies are detected as micro clusters or isolated points located in low density regions of the features space [8].

In the context of active learning, mainly in problems related to classification, there have been considerable research about the problem of deciding how to improve the accuracy of a classifier by actively deciding what instance to label [14] [9] [13] [16]. In a work closely related to our application domain, Pelleg and Moore [12] propose an active learning strategy to find anomalies in a database. The main difference with our approach is that they do not explicitly search for relevant anomalies in selective subspaces.

5 Conclusions

This work contributed with an algorithm based on the active learning paradigm that tackles the problem of detecting anomalous records in large datasets. Using the factorization of the joint pdf provided by a BN and the properties of Dirichlet distributions to model a Polya urn process, we were able to use the feedback from the user to speed up the selection of relevant anomalies that exhibit regularities or patterns in selective subspaces.

Our results indicated that with respect to a baseline method that do not incorporate active learning, the approach presented in this work was able to significantly decrease the time to reach the relevant anomalies. Furthermore, by providing a set of specific attributes corresponding to the subspace used to detect

the anomaly, the method proposed here was also able to provide an explanation of the main sources of the anomaly.

As future work, we believe that the incorporation of previous knowledge in the modeling steps can improve the detection. Also a most exhaustive experimental analysis in datasets coming from different domains is also a valuable step forward. Given that the expert time is usually the most valuable resource in the loop, we believe that tools as the one presented here may be of great help as a filtering step to help and guide the search in datasets where an exhaustive analysis is not possible.

References

1. A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *Artificial Intelligence Communications*, 7(1):39–59, 1994.
2. A. Asuncion and D.J. Newman. UCI machine learning repository, <http://www.ics.uci.edu/~mllearn/MLRepository.html>, 2007.
3. D. Blackwell and J. MacQueen. Ferguson distribution via polya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.
4. A. Cansado and A. Soto. Unsupervised anomaly detection in large databases using bayesian networks. *Applied Artificial Intelligence*, 22(4):309–330, 2008.
5. T. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.
6. V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.
7. P. Jackson. *Introduction to Expert Systems*. Addison Wesley, 1998.
8. Y. Kou, C. Lu, S. Sirwongwattana, and Y. Huang. Survey of fraud detection techniques. In *Proc. of the IEEE Int. Conf. on Networking, Sensing and Control*, pages 749–754, 2004.
9. D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *Proc. of 17th Int. Conf. ACM SIGIR*, pages 3–12, 1994.
10. R. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, 2004.
11. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.
12. D. Pelleg and A. Moore. Active learning for anomaly and rare-category detection. In *Proc. of the 18th Conf. on Advances in Neural Information Processing Systems, NIPS*, 2004.
13. N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proc. of 18th Int. Conf. on Machine Learning, ICML*, pages 441–448, 2001.
14. S. Seung, M. Opper, and H. Sompolinski. Query by committee. In *Proc. of 5th Annual ACM Workshop on Computational Learning Theory*, pages 287–294, 1992.
15. A. Soto, F. Zavala, and A. Aranedá. An accelerated algorithm for density estimation in large databases using Gaussian mixtures. *Cybernetics and Systems*, 38(2):123–139, 2007.
16. S. Tong and D. Koller. Active learning for parameter estimation in bayesian networks. In *Proc. of the 13th Conf. on Advances in Neural Information Processing Systems, NIPS*, pages 647–653, 2001.