New Trends in Data Mining

Association Rules and Anomaly detection

Literary review of anomaly detection methods

A Novel Anomaly Detection Scheme Using Principal Components Classifier

Introduction

Anomaly detection is an application of machine learning techniques that targets observations that do not fit well within the data and thus, are labeled as anomalous.
These methods have several use cases such as fraud detection, predictive maintenance and data cleanup. They can also be used both in supervised and unsupervised detection problems, although some limitations apply, as it is assumed that in the latter situation

Anomalies can be of different natures:

- Point anomalies, which are outliers with respect to the rest of the data.
- Contextual anomalies, which are anomalous within a specific context or subset of the data.
- Collective anomalies, when a group of similar data points are anomalous with respect to the entire dataset, but not to individual instances.

Contextual anomalies can be detected by isolating one segment of the data and perform detection techniques on that subgroup.

There many methods and different models that address this problem, such as the use of local outlier factor theory (LOF), association rules and geometric methods like distance-based models, as well as applications of information theory.
Machine learning techniques can also be applied to the problem of anomaly detection by approaching it as a classification task. Models such as support vector machines and recursive neural networks are commonly used in this context.

This essay will focus on a novel approach to anomaly detection pioneered by Shyu *et al*. (n.d) in YEAR, consisting in the application of Principal Component theory. The paper will be closely followed in order to document the technique.

One of the most attractive attributes of this technique is the fact that, since it reaches a model with much lower dimensionality, it lowers computation time and allows the technique to be used in real-time. This is specially interesting in industry, where it can be applied to solve many business problems.

When compared against other anomaly detection techniques, the Principal Component Classifier (PCC) model outperforms the k-nearest neighbour approach, the LOF method and the Canberra metric.
Moreover, the PCC method features an excellent false alarm rate. This metric is defined as the ratio of false positives divided by the total number of observations. That is, the percentage of times in which the algorithm wrongly classifies a data point as an anomaly.

Principal Component theory

Principal components are linear combinations of the random variables in the data, with three important properties:

1. The components are uncorrelated, orthogonal to each other.
2. The first principal component has the highest percentage of explained variance
3. The total variance of the principal components is the same as that of the original variables.

In some instances, it is more beneficial to perform the PC analysis on the correlation matrix rather than the covariance matrix. This is the case when the variables are measured on scales with considerably different ranges.

Outlier Detection

Data points are considered outliers when their distance from the center of the data is very high. Many distance metrics can be used, such as the Euclidean, Manhattan or Canberra metrics.
For the PCC technique, we will focus on the Mahalanobis distance. The reason behind this is that the sum of squares of the standardized principal component scores is equivalent to the Mahalanobis distance of an observation from the average of the sample. (Jobson, 1992)

When working with high-dimensional data, some anomalous instances do not appear to be outliers when considering each dimension separately. For that reason, it is crucial to use all features in the the multivariate approach.

Principal Component Classifiers for Anomaly Detection

The presented anomaly detection method has one peculiarity in its computation of principal components. Not only does it take into account the traditional major principal components, it also considers the minor principal components.
The minor components are the last few principal components, which represent linear functions of the original features with the minimum variance. The intuition behind this

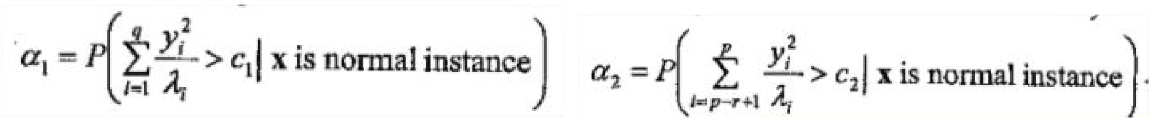
technique is that these components are sensitive to observations that are inconsistent with the correlation structure but not with respect to the original variables.

Therefore, the largest values of the observations within the minor components will reflect multivariate outliers which could not be detected by the major components alone. This premise tested and proven right later in the paper, by comparing the performance of 3 different models: one with only major components, one with only minor and one with both combined. The results show that the minor components greatly outperform the major ones when applied separately for specific subgroups of the anomalous data, which is used as example to demonstrate the contribution of the minor components to the anomaly detection scheme.

In summary, the major components aim to identify extreme observations with large values on the original features, while the minor components help detect the observations that do not conform to the correlation structure.
The criteria to choose the number of components to use is largely based on performance, although there is a general rule of thumb:

- Use the *q* major components that can explain about 50% of the total variance in the data.
- Use the *r* minor components whose eigenvalues are less than 0.2.

The classification framework of anomalous instances is as follows:



$$\alpha_1 = P\left(\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} > c_1 \middle| \text{x is normal instance}\right) \quad \alpha_2 = P\left(\sum_{i=p-r+1}^{p} \frac{y_i^2}{\lambda_i} > c_2 \middle| \text{x is normal instance}\right).$$

Source: Shyu *et a*l. (n.d)

Where c_1 and c_2 are the defined outlier thresholds such that the model produces a specified false alarm rate.
Thus, if we assume the data is normally distributed, the false alarm rate of the classifier is:

$$\alpha = \alpha_1 + \alpha_2 - \alpha_1\alpha_2$$

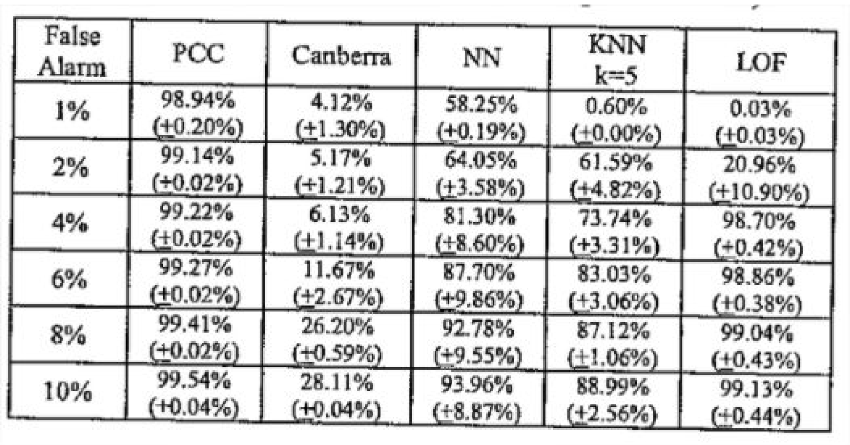
The researchers in Shyu *et al.* (n.d) set different levels of the false alarm parameter to test the model in different conditions, as well as comparing it with the LOF approach, Canberra metric and Euclidean distance, which is in fact the k-nearest neighbor method.
When evaluating the models, their accuracy is measured by its rate of misclassification, as well as precision and recall:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \qquad Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Another evaluation metric is the receiver operating characteristics (ROC) curve, which is the plot of the detection rate against the false alarm rate, and its associated area under the curve (AUC).

Results and conclusion

After running the experiments, the authors come up with the follow results matrix:

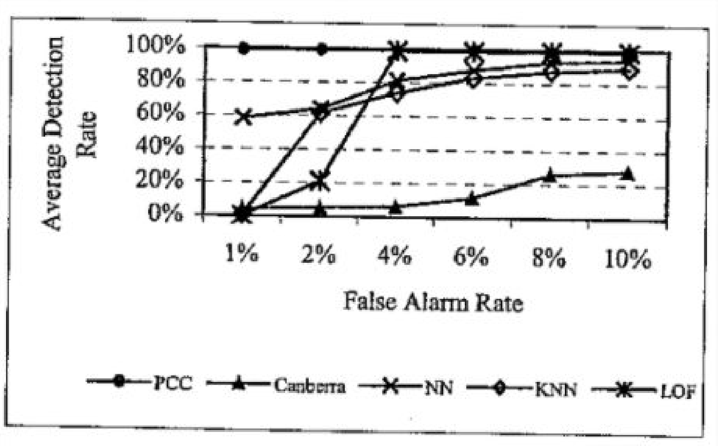


False Alarm	PCC	Canberra	NN	KNN k=5	LOF
1%	98.94% (±0.20%)	4.12% (±1.30%)	58.25% (±0.19%)	0.60% (±0.00%)	0.03% (±0.03%)
2%	99.14% (±0.02%)	5.17% (±1.21%)	64.05% (±3.58%)	61.59% (±4.82%)	20.96% (±10.90%)
4%	99.22% (±0.02%)	6.13% (±1.14%)	81.30% (±8.60%)	73.74% (±3.31%)	98.70% (±0.42%)
6%	99.27% (±0.02%)	11.67% (±2.67%)	87.70% (±9.86%)	83.03% (±3.06%)	98.86% (±0.38%)
8%	99.41% (±0.02%)	26.20% (±0.59%)	92.78% (±9.55%)	87.12% (±1.06%)	99.04% (±0.43%)
10%	99.54% (±0.04%)	28.11% (±0.04%)	93.96% (±8.87%)	88.99% (±2.56%)	99.13% (±0.44%)

Source: Shyu *et al.* (n.d)

These are the average detection rates of each of the models for specified levels of false alarm rate, after running 5 independent experiments for each method. The standard deviation of each statistic in shown between brackets below the average.

In general, the Canberra metric performs poorly, which is consistent with previous research in the field. (Emran and Ye, 2001). However, the novel PCC technique has an exceptionally good performance at all false alarm levels. As seen below, the method greatly outperforms all other detection methods:



Source: Shyu *et al.* (n.d)

In summary, and as can be seen in the ROC curve chart above, the PCC based anomaly detection scheme is a novel technique ranked among the most accurate in the intrusion detection problem.

References

J.D. Jobson, “Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods.” Springer-Verlag, NY. 1992

S.M. Enran and N. Ye. Robustness of Canberra Metric in Computer Intrusion Detection. Proceedings of the 2001 IEEE Workshop on Information Assurance and Security, United States Military Academy, West Point, NY. June 5-6, 2001.

Shyu, M., Chen, S., Sarinnapakorn, K., & Chang, L. (n.d.). Principal Component-based Anomaly Detection Scheme. Foundations and Novel Approaches in Data Mining Studies in Computational Intelligence, 311-329.