Outlier Detection Model Based on SOM for Classification Problem

Jin Xiao\textsuperscript{1,a}, Qin Lei\textsuperscript{1,b}, Dunhu Liu\textsuperscript{2,c} and Ling Xie\textsuperscript{1,d,*}

\textsuperscript{1}Business School, Sichuan University, Chengdu 610064, China
\textsuperscript{2}Management Faculty, Chengdu University of Information Technology, Chengdu 610103, China
\textsuperscript{a}xiaojin@scu.edu.cn, \textsuperscript{b}18382326524@163.com, \textsuperscript{c}264885613@qq.com, \textsuperscript{d}xie_ling0101@126.com

*Corresponding author

Keywords: Classification problem, Outlier detection, Self-organizing mapping network, Pruning, Local outlier factor.

Abstract. In many practical classification problems, it often contains some outliers in the data set, which may affect the performance of classification model. To solve this problem, this paper combines the self-organizing mapping network (SOM), the pruning technique and the local outlier factor (LOF), constructs the outlier detection model based on SOM (SOD). Firstly, it clusters with SOM on the training set, and then obtains the new training set by pruning the clustering results. Finally, it detects the outliers by the local outlier factor of each sample on the new training set. The empirical results show that the SOD model has better detection performance compared with some existing outlier detection models, and it can improve the classification accuracy more efficiently through the models trained without the outliers.

1. Introduction

With the development of computer network technology, communication technology and database technology, data mining technology has also emerged. Data mining refers to the process of extracting the potential useful information from a large amount of data by statistics, computer technology and comprehensive analysis. Clustering analysis, data classification, association rules analysis and prediction are several branches of data mining technology, and data classification is one of the important directions [1]. Data classification constructs a classification model based on the characteristics of the data, which can map the class-unknown samples to some given class. The process of constructing the classification model contains two steps: training and testing. In the training stage, it generates one corresponding accurate description or model for each class according to their characteristics. In the testing phase, it classifies with the class description or model, tests its classification accuracy. Many problems in reality, such as network text classification, face recognition, medical disease diagnosis, credit risk evaluation and customer churn prediction are all classification problems [2]. However, in the actual classification problem, we often encounter the data that do not correspond to the behavioral characteristics and patterns of most data, which we call outliers. There is a significant difference between the outlier data and most data in a data set, so that it is suspected that they may be generated by another completely different mechanism. The existence of these outliers will slow down and mislead the learning process and may seriously affect the classification performance of trained model [3,4]. Therefore, to improve the accuracy and validity of classification, it is necessary to carry out outlier detection before classifying.

The reasons for generating outliers are generally divided into the following three reasons. Firstly, the data come from different classes. If an object is derived from a class different from other data objects, it is obvious that it will be different from other samples. Secondly, the data variables change inherently, which reflects the characteristics of the data set. Thirdly, there are errors in data measurement and collection, which may be man-made or caused by the measurement equipment breakdown, it is also one of the outliers’ sources.
Many scholars have proposed a lot of methods to detect outliers. These methods can be roughly divided into the following five categories: 1) Statistical method, the main idea is to assume that the data set obeys a distribution or one certain probability model, and regard those records seriously deviated from the distribution as the outliers by the inconsistent test [5], this method relies heavily on whether the data set satisfies a certain probability distribution model. However, in actual activity, the probability distribution for the data set is mostly unknown, which is the biggest drawback of statistical methods. 2) Distance-based method [6], it is relatively easy to understand, but for high-dimensional large data set, this method has the disadvantage of high time complexity; 3) Density method, it is applicable to a data set with unbalanced density, that is, the data set has clustering distribution with sparse type and dense type [7]. 4) Clustering method, it defines the outlier as not belong to any cluster [8], but the main goal of the clustering algorithm is to discover the clusters rather than discover the outliers, so it is less efficient to mine the outliers. 5) Classification method, the outlier detection algorithm based on classification depends on not only the classification algorithm, but also a wide range of representative and the correct category label.

In the above five outlier detection methods, the density-based detection method is fundamentally different from other detection methods. To begin with, it proposes the definition of local outliers and the solution to the mining problem for the data sets with unbalanced density distribution. Next, it changes the outliers from either-or to degree concept, and each sample has its own outliers, which may be high or low. Finally, it can detect the characteristics of outliers locally, to avoid missing any outlier [9]. Therefore, it has become the most commonly-used method in outlier detection. To solve the problem of local outlier mining, Breunig [7] proposed the concept of local outlier factor (LOF). In the real classification problem, the data set is often large scale, so we hope to not only ensure the detection performance, but also reduce the calculation time. In order to solve this issue, we combine the clustering algorithm SOM with the LOF outlier detection method and propose the SOM-based outlier detection model (SOD), so as to effectively improve the outlier detection performance in large-scale data set.

The structure of this study is as follows. In Section 2, we introduce the related theories, including SOM and LOF. In Section 3, we present the outlier detection model SOD. Section 4 is the empirical research. By the experiments on the classification data sets, the experimental results show that the SOD model proposed in this paper has better outlier detection performance compared with some existing models. Section 5 conclusions this paper.

2. Related Work

2.1 Self-organizing Mapping network (SOM)
Kohonen has proposed a self-organizing mapping network (SOM) [10-11], which is an unsupervised training neural network. The topology diagram is shown in Figure 1.

![Figure 1. SOM topology diagram.](image)

SOM is a type of artificial neural network (ANN), which is trained through unsupervised learning to produce a low-dimensional (typically two-dimension) network, discretized representation of the
input space of the training samples. SOM differs from other artificial neural networks as it applies competitive learning as opposed to error-correction learning, and they use a neighborhood function to preserve the topological properties of the input space. SOM operates in two modes: training and mapping. "Training" builds the map using input examples, while "mapping" automatically classifies a new input vector.

In SOM, an output node can react specifically to a class of patterns to represent the class, and adjacent nodes on the output layer can make a special reflection of the similar classes in the actual distribution. When a sample is input, the maximum stimulus (winning node) is generated for an output node, and a large stimulus is generated for some nodes around the winning node. In the training, the connection weight between the winning node and the input node is constantly adjusted, the connection weight of the neighborhood node of the winning node is also constantly adjusted. In the training process, the neighborhood of the winning node is narrowed, until only slightly adjust the connection weight between the winning node and the input node.

Assuming that the input vector \( X = (x_1, x_2, \ldots, x_m) \), the dimension is \( m \), the number of output nodes is \( N \), the following are the steps of SOM algorithm:

1. Randomize the initial weight vectors \( \mathbf{W}_j(0) \) \( (j = 1, \ldots, N) \).
2. Calculate the Euclidean distance between \( X \) and \( \mathbf{W}_j \), and select the winning node \( i(x) \).
   \[
   i(x) = \arg\min_j ||X - \mathbf{W}_j||. \tag{1}
   \]
3. Define the neighborhood of the winning node \( \mathcal{N}_{i(x)}(t) \), there is usually choose the Gaussian function \( h_{Gauss} \), the initial area is larger, it will gradually narrowing in the training process.
4. Update the connection weight in the neighborhood of the winning node (including the winning node itself), where \( \zeta_\varepsilon \) is the learning rate.
   \[
   \mathbf{W}_j(t+1) = \mathbf{W}_j(t) + \zeta_\varepsilon \cdot h_{Gauss}(t) \left( \frac{X(t) - \mathbf{W}_j(t)}{||X(t) - \mathbf{W}_j(t)||} \right) \quad j \in \mathcal{N}_{i(x)}(t). \tag{2}
   \]
5. Repeat Steps (2) – (4) until only slightly adjust the connection weight between the winning node and the input node.

### 2.2 Local Outlier Factor (LOF)

When the data set has clustering distribution with sparse type and dense type at the same time, that is, the density distribution is not balanced, there will be local outliers. Distance-based outlier detection method cannot effectively detect these local outliers. For the local outliers, we use Figure 2 to explain it. There is a two-dimensional data set, which contains two clusters \( C_1, C_2 \) and two outliers \( O_1, O_2 \), where \( C_1 \) is a cluster with dense type, \( C_2 \) is sparse, \( O_1 \) is the global outlier, \( O_2 \) is a local outlier. By distance-based detection method, \( O_1 \) is easy to be detected, but \( O_2 \) is difficult to be detected, if in order to dig out \( O_2 \) and change the parameters, maybe the most of the data points in \( C_1 \) will be identified as outliers. For this reason, Breunig [7] proposed the concept of local outlier factor and used it to detect local outliers.

![Figure 2. The unbalanced density data set.](image)

Given an integer parameter \( \text{MinPts} \), The LOF of object \( p \) is defined as,
represents $\text{MinPts}$-nearest neighbors of $p$, $\text{ind}_{\text{MinPts}}(p)$ is the local reachability density of $p$, which is defined as follow,

$$\text{ind}_{\text{MinPts}}(p) = \frac{1}{N_{\text{MinPts}}(p)}$$

Let $\text{MinPts-distance}(o)$ be the distance of the object $o$ to the $\text{MinPts}$-nearest neighbor. The reachability distance between object $p$ and object $o$ is defined as,

$$\text{reach-dist}(p, o) = \max\{\text{MinPts-distance}(o), \text{dist}(p, o)\}$$

3. SOD Model

3.1 Basic idea

In the actual classification problem, the data set is often a large-scale data set, LOF can effectively detect outliers in the data set with unbalanced density, but the algorithm has high time complexity. So in this paper, we propose outlier detection model SOD, clustering the samples by using SOM, and then pruning clusters, which greatly reduces the data size.

The modeling process of the SOD model mainly includes the following three steps: 1) Clusters with SOM on the training set; 2) Obtains the new training set by pruning the clustering results; 3) Detects the outliers by calculating the LOF of each sample on the new training set.

3.2 Detailed modeling steps

Before modeling the SOD, we need to divide the data sample set into training set $L$ and testing set $T$. The modeling process of the SOD is shown in Figure 3.

**Step 1.** On the training set $L$ ($n$ data samples), it clusters with SOM, and then get $C$ clusters.

**Step 2.** Pruning.

In this paper, through experiments, we use the number of samples in the cluster and the distance between the samples in the cluster as the pruning judgment criteria.

1) If the number of the cluster $|\mathcal{C}_i|$ is less than a certain threshold $\theta$, all samples in the cluster are retained in the new training set $L'$;

2) If the number of the cluster $|\mathcal{C}_i|$ is more than a certain threshold $\theta$, then calculate the distance $d$ between the sample $\mathbf{x}_i$ and the center of the cluster $\mathbf{z}_i$, if $d$ is more than $\bar{d} = \frac{1}{n} \sum_{i=1}^{n} \text{dist}(\mathbf{x}_i, \mathbf{z}_i)$, the $\mathbf{x}_i$ is also stored in the new training set $L'$, where $\mathcal{C}_i$ is the $i$-th cluster.

**Step 3.** Outliers detection.

Calculate the LOF values of all data samples in the new training set $L'$, and the samples with higher LOF values are detected as outliers.

![Figure 3. The modeling process of the SOD.](image-url)
4. Summary Experiment

In order to analyze the outlier detection performance of the SOD model proposed in this paper, we have conducted empirical research. We compare the SOD model with the following three outlier detection models: 1) Distance-based outlier detection model (PLDOF) proposed by Pamula et al. [12]; 2) LOF model without clustering and pruning [7]; 3) Local outlier detection model (CBLOF) proposed by He et al. [13].

4.1 Data sets

The experiment is conducted on three customer classification data sets: 1) German data set. The data set is from the international public database of the UCI, which describes credit scoring of 1,000 german customer samples. Every customer sample contains 20 features and one class label. The class labels are divided into two kinds. Correspondingly, all customers are divided into customers with a good or poor credit scoring. Among them, there are 700 customer samples with a good credit scoring, and 300 customer samples with a poor credit scoring. The ratio between the two kinds of samples is 2.33:1. The 20 features of the data set include seven numerical features and thirteen qualitative features. 2) Churn data set. The data set also comes from the UCI, which deals with cellular service provider customers and the data pertinent to the voice calls they make. There are 3,333 samples, among which 2,850 are non-churn customers and 483 are churn customers (meaning they changed service providers). The ratio between the two kinds of samples is 5.90:1. The data set includes 20 features. 3) Australia data set. The data set also comes from the UCI, which describes credit scoring of 690 Australian customer samples. Every customer sample contains 14 features and one class label. The class labels are divided into two kinds. Correspondingly, all customers are divided into customers with a good or poor credit scoring. Among them, there are 383 customer samples with a good credit scoring, and 307 customer samples with a poor credit scoring. The ratio between the two kinds of samples is 1.25:1.

4.2 Experimental setup

In this experiment, we need to divide the data sets into two subsets: the training set $L$ and the testing set $T$. We should ensure that the proportion of positive samples and negative samples in $L$ and $T$ is the same as that of the original data set. Firstly, according to the class label, the data sets are divided into two parts: positive class and negative class. Then, the two classes are randomly divided into two subsets according to the ratio of 7:3. Finally, the first subset of the two classes is merged to get the training set $L$, the second subset of the two classes is merged to get the testing set $T$.

And we use the model to detect the outliers on the training set $L$, and then train the classification model after removing the outliers. Finally, we use the classification model trained to classify the samples in the testing set $T$. The performance of the outlier detection model is evaluated by the change of the evaluation criteria before and after removing the outliers. In the experiment we choose BP neural network as the basic classification algorithm. In the BP neural network, the number of hidden nodes is a very important parameter, the optimal value is often different on different data sets. Through several experiments, we find that the optimal hidden nodes on the Australia, German and churn data sets are 4, 5 and 5, respectively. All the experimental results are averaged to take ten experiments and are programmed on the MATLAB R2014a software to achieve.

4.3 Evaluation Criteria

To evaluate the performance of the models referred to in this study, we introduced the confusion matrix in Table 1. On this basis, two commonly used evaluation criteria were adopted.
Table 1. Confusion matrix.

<table>
<thead>
<tr>
<th></th>
<th>Predicted positive</th>
<th>Predicted negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual positive</td>
<td>TP (the number of true positives)</td>
<td>FN (the number of false negatives)</td>
</tr>
<tr>
<td>Actual negative</td>
<td>FP (the number of false positives)</td>
<td>TN (the number of true negatives)</td>
</tr>
</tbody>
</table>

(1) Accuracy(ACC) = ((TP + TN)/(TP + FN + FP + TN))×100% .
(2) The area under the receiver operating characteristic curve (AUC). The receiver operating characteristic (ROC) curve is an important evaluation criterion of classification model in the data with imbalanced class distribution. For an issue of two classes, the ROC graph is a true positive rate-false positive rate graph, where y-axis is true positive rate (TP/(TP + FN)×100%) and the x-axis is false positive rate (FP/(FP + TN)×100%). However, sometimes it is difficult to compare ROC curves of different models directly, so AUC is more convenient and popular.

4.4 Experimental results

Table 2 shows the ACC values of the SOD model and the other three outlier detection models on different data sets. Table 3 shows the AUC values of the SOD model and the other three outlier detection models on different data sets. And the bold in the table indicates the maximum value. The last line in the table also shows the classification results without using the outlier detection models. From Table 2 and Table 3 we can get the following conclusions:

1) The performance of the classification models have been improved after eliminating the outliers. The reason may be that the original data set contains a certain amount of noise and outliers, which can affect the performance of the classification model when training the classification models. Through the outlier detection, it can eliminate a part of the noise samples and outliers, which improves the performance of the classification models significantly. Thus, we can conclude that outlier detection is necessary and meaningful in the classification problem.

2) Compared to the classification results without the outlier detection, the ACC values and the AUC values of the SOD model have improved in the three data sets, and this improvement is superior to the other three outlier detection models. This indicates that the outlier detection performance of the SOD model is better than the other three outlier detection models.

3) After detecting the outliers and removing the outliers, the performance of some models has not been improved, but there has been a decline, such as the PLDOF model used on the churn data set. This may be because, when detecting the outlier, the outlier detection model incorrectly detects some normal samples as outliers, thus this samples be removed, so the amount of useful and normal samples are become less in the remaining data set, the classification model performance becomes worse. But from these tables we can see the performance of the SOD model is not become worse in the three data sets. The above situation shows that the SOD model proposed in this paper has good stability, and it is further proved that the Step 1 outlier detection performance of SOD model is superior to other models.

4) In order to determine whether there is a statistically significant difference between the SOD outlier detection model and other outlier detection models, we performed the t-test for the ACC values and AUC values in the SOD model and the other three outlier detection models, test results are shown in Table 2 and Table 3. At the 95% confidence level, if there is a significant statistical difference between the SOD model and the other models, we use "***" to represent, and at the 90% confidence level, if there is a significant statistical difference between the SOD model and the other models, we use "**" to represent. It can be seen from Table 2 and Table 3, the ACC value and AUC value of the SOD model are significantly superior to other models at 95% or 90% confidence level. This further shows that the proposed SOD outlier detection model is more efficient and superior.
Table 2. ACC values for different models.

<table>
<thead>
<tr>
<th>Model/ Data set</th>
<th>Australia</th>
<th>German</th>
<th>Churn</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOD</td>
<td>0.8776</td>
<td>0.7362</td>
<td>0.8918</td>
</tr>
<tr>
<td>PLDOF</td>
<td>0.8422**</td>
<td>0.7308*</td>
<td>0.8230**</td>
</tr>
<tr>
<td>LOF</td>
<td>0.8522**</td>
<td>0.7259**</td>
<td>0.8818**</td>
</tr>
<tr>
<td>CBLOF</td>
<td>0.8577**</td>
<td>0.7288**</td>
<td>0.8839*</td>
</tr>
<tr>
<td>Without detecting outliers</td>
<td>0.8333**</td>
<td>0.7043**</td>
<td>0.8803***</td>
</tr>
</tbody>
</table>

Table 3. AUC values for different models.

<table>
<thead>
<tr>
<th>Model/ Data set</th>
<th>Australia</th>
<th>German</th>
<th>Churn</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOD</td>
<td>0.9166</td>
<td>0.7417</td>
<td>0.8493</td>
</tr>
<tr>
<td>PLDOF</td>
<td>0.8867**</td>
<td>0.7259**</td>
<td>0.8311**</td>
</tr>
<tr>
<td>LOF</td>
<td>0.9000*</td>
<td>0.7396*</td>
<td>0.8358**</td>
</tr>
<tr>
<td>CBLOF</td>
<td>0.8945**</td>
<td>0.7354*</td>
<td>0.8378*</td>
</tr>
<tr>
<td>Without detecting outliers</td>
<td>0.8482**</td>
<td>0.6991**</td>
<td>0.8344**</td>
</tr>
</tbody>
</table>

5. Conclusion

In many practical classification problems, it often contains some outliers in the data set, which may affect the performance of classification model. To solve this problem, this paper combines SOM, the pruning technique and the LOF, constructs the outlier detection model based on SOM (SOD). The empirical results show that the SOD model has better detection performance compared with some existing outlier detection models, and it can improve the classification accuracy more efficiently through the models trained without the outliers. Although this paper focuses on outlier detection in classification problems, the proposed SOD model is also important for outlier detection in other areas of data mining.

Acknowledgement

This research is partly supported by the Natural Science Foundation of China under Grant Nos. 71471124, 71273036 and 71673011, Excellent Youth Fund of Sichuan University under Grant No. skqx201607 and skzx2016-rcrw14. Ling Xie is the corresponding author.

References


