Semi-Supervised Outlier Detection

Jing Gao
Dept. of Computer Science
and Engineering
Michigan State University
East Lansing,MI 48824
gaojing2@msu.edu

Haibin Cheng
Dept. of Computer Science
and Engineering
Michigan State University
East Lansing,MI 48824
chenghai@msu.edu

Pang-Ning Tan
Dept. of Computer Science
and Engineering
Michigan State University
East Lansing,MI 48824
ptan@cse.msu.edu

ABSTRACT

Outlier detection has been extensively researched in the context of unsupervised learning. But the learning results are not always satisfactory, which can be significantly improved using supervision of some labeled points. In this paper, we are concerned with employing supervision of limited amount of label information to detect outliers more accurately. The key of our approach is an objective function that punishes poor clustering results and deviation from known labels as well as restricts the number of outliers. The outliers can be found as a solution to the discrete optimization problem regarding the objective function. By this way, this method can detect meaningful outliers that can not be identified by existing unsupervised methods.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications-Data Mining

General Terms

Algorithms

Keywords

Outlier detection, semi-supervised learning

1. INTRODUCTION

Semi-supervised learning, which uses both unlabeled and labeled data, has become a hot topic in recent years. Existing work includes semi-supervised classification [5], semi-supervised clustering [1], etc. This learning approach can improve the accuracy using supervision of some labeled data compared with that of unsupervised learning while reduce the need for expensive labeled data which is required in supervised learning.

In this paper, we explore the use of semi-supervised techniques on outlier detection. In many applications, such as network intrusion detection, fraud detection and criminal activities monitoring, cases that deviate significantly from majority are more interesting and useful than the common cases. Finding such outliers has begun to receive attention from researchers [3, 2]. Most of these algorithms fall into the category of unsupervised learning, which only take use of unlabeled data.

We show that by using known normal points, we can effectively determine the clusters in the dataset. On the other hand, utilization of known outliers prevent from misclassifying outliers as normal. Specifically, our idea is to capture the clustering of the normal points with the aid of labeled points thus remaining points are outliers. Instead of minimizing only the sum squared error, we try to optimize the objective function which incorporates outlier existence and labeled information into clustering. We propose an efficient algorithm based on K-means clustering method to solve this discrete optimization problem. This algorithm is computationally inexpensive. After several iterations, this algorithm obtains the local optimal result.

2. OBJECTIVE FUNCTION

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of data points drawn from R^m . Let the first $l < n$ points in X be labeled as shown in the indicator vector $F = \{u_1, u_2, \ldots, u_l\}$, where $u_i = 0$ if the i-th point is an outlier, and 1 otherwise. We wish to predict the labels of these points in X, i.e., whether a point is an outlier or not.

To this end, we assume normal points form K clusters and outliers do not belong to any clusters. Our aim is to find a $n \times K$ matrix $T = \{t_{ih} | 1 \leq i \leq n, 1 \leq h \leq K\}$, where $t_{ih} = 1$ if x_i is contained in the cluster C_h, and $t_{ih} = 0$ otherwise. Therefore for x_i,

$$\sum_{h=1}^{K} t_{ih} = \begin{cases} 1 & x_i \text{ is a normal point} \\ 0 & x_i \text{ is an outlier} \end{cases}$$

We may now define the optimization problem we want to solve: Minimize:

$$\begin{aligned} Q &= \sum_{i=1}^{n}\sum_{h=1}^{K} t_{ih} dist(c_h, x_i)^2 + \gamma_1 (n - \sum_{i=1}^{n}\sum_{h=1}^{K} t_{ih}) \\ &+ \gamma_2 \sum_{i=1}^{l} |u_i - \sum_{h=1}^{K} t_{ih}| \end{aligned} \quad (1)$$

SAC'06 April 23-27, 2006, Dijon, France
Copyright 2006 ACM 1-59593-108-2/06/0004 ...$5.00.

subject to:

$$t_{ih} \in \{0,1\} \text{ and for each } i \sum_{h=1}^{K} t_{ih} \leq 1$$

where c_h is the center of cluster $C_h(1 \leq h \leq K)$, *dist* is the standard Euclidean distance between two points and γ_1, γ_2 are two adjusting parameters. We aim at finding the indicator matrix T^* which minimize the objective function Q.

The first constraint on t_{ih} reflects that it is an indicator variable. The second constraint on the sum of t_{ih} with respect to i shows that each point should be assigned to at most one cluster. As for the objective function Q, the first term is directly inherited from the K means clustering objective function, which represents the sum squared error. But in our method, only normal points are partitioned into clusters, so outliers do not contribute to this term. We note that if only considering minimizing this term, it will classify every point as an outlier. Therefore we introduce the second term to constrain the number of outliers not to be too large. The third term maintains consistency of our labeling with existing labels. Q will be punished by mislabeled points. To make these three terms compete with each other, we incorporate two weighting parameters γ_1 and γ_2.

3. ALGORITHM

In this section, we propose an iterative algorithm based on K means clustering to solve this discrete optimization problem. Like the classic K means algorithm [4], we minimize the objective function through iteration of two steps. The specific algorithm is described as follows:

```
SSOD algorithm:
Input: Partially labeled data set $X$, number of
       clusters $K$
Output: Configuration matrix $T$, $K$ cluster centers
Method:
1. $s \leftarrow 0$.
2. Initialize $K$ centers as $c_1^0, c_2^0, \ldots, c_K^0$
3. Loop until algorithm converges
    3.1 With $c_h^s$ fixed, calculate the indicator matrix
        $T^s$ that minimizes $Q$ subject to the constraints.
    3.2 With $T^s$ fixed, compute the new $K$ centers
        $c_1^{s+1}, c_2^{s+1}, \ldots, c_K^{s+1}$ which minimizes $Q$.
    3.3 set $s \leftarrow s + 1$
```

Table 1: SSOD algorithm framework

Now we will discuss the details of the two steps. Let's look at 3.1 first. We construct a distance matrix D such that each unit $d_{ih} = dist(c_h, x_i)^2$. In order to minimize Q, we should minimize each point's contribution to this objective function. For each point x_i, to minimize the first term, we assign x_i to its closet centroid, i.e. set $t_{ih^*} = 1$ and $t_{ih} = 0(h \neq h^*)$ such that $h^* = \text{argmin}_h d_{ih}$. That will make every point to be identified as normal. Regarding the last two terms, we will discuss the possibility of some points to be reassigned as outliers when one of the following cases occurs:

1. When x_i is unlabeled, x_i does not contribute to the last term of Q and if it is classified as an outlier, it contributes γ_1 to the second term. If $d_{ih^*} > \gamma_1$,classifying x_i as an outlier will subtract d_{ih^*} from the first term and add γ_1 to the second term, which leads to the overall decrease of the objective function. So we flip the only 1 in row i of T into 0 to represent x_i as an outlier.

2. When x_i is labeled as a normal point, x_i contributes γ_1 to the second term in Q and γ_2 to the last term if it is classified as an outlier. Therefore if $d_{ih^*} > \gamma_1 + \gamma_2$, classifying x_i as an outlier will decrease the objective function.

3. When x_i is labeled as an outlier, if x_i is classified as an outlier, it only contribute γ_1 to the second term in Q. If x_i is regarded as normal, besides contribute d_{ih^*} to the first term, it is also counted in the last term. Therefore if $d_{ih^*} > \gamma_1 - \gamma_2$, classifying x_i as an outlier will decrease the objective function.

Now we will give the details of 3.2. Since T_s is fixed, the last two terms in Q are fixed. The problem of minimizing Q is equivalent to minimizing the sum squared error. Similar to K means clustering problem, it can be shown that by choosing c_h^{s+1} to be the center of the new cluster C_h^{s+1} , the objective function is minimized, i.e.

$$c_h^{s+1} = \sum_{i=1}^{n} t_{ih} x_i / \sum_{i=1}^{n} t_{ih} \quad (2)$$

4. CONCLUSIONS

In this paper, we have introduced a framework for semi-supervised outlier detection that employs a novel objective function which utilizes both unlabeled and labeled data to determine the outliers. By minimizing the objective function that takes clustering result, outlier assignments as well as mislabel punishment into consideration, we can find proper outliers that do not belong to any normal clusters. We also propose an efficient iterative algorithm to solve the optimization problem.

5. REFERENCES

[1] S. Basu, M. Bilenko, and R. J. Mooney. A probabilistic framework for semi-supervised clustering. In *KDD '04: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 59–68. ACM Press, 2004.

[2] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. In *SIGMOD '00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104. ACM Press, 2000.

[3] E. M. Knorr, R. T. Ng, and V. Tucakov. Distance-based outliers: Algorithms and applications. *VLDB Journal:Very Large Data Bases*, 8(3-4):237–253, 2000.

[4] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Symposium on Math, Statistics, and Probability*, pages 281–297, 1967.

[5] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39(2/3):103–134, 2000.