RKOF: Robust Kernel-Based Local Outlier Detection*

Jun Gao[1], Weiming Hu[1], Zhongfei (Mark) Zhang[2], Xiaoqin Zhang[3], and Ou Wu[1]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
{jgao,wmhu,wuou}@nlpr.ia.ac.cn
[2] Dept. of Computer Science, State Univ. of New York at Binghamton, Binghamton, NY 13902, USA
zhongfei@cs.binghamton.edu
[3] College of Mathematics & Information Science, Wenzhou University, Zhejiang, China
xqzhang@wzu.edu.cn

Abstract. Outlier detection is an important and attractive problem in knowledge discovery in large data sets. The majority of the recent work in outlier detection follow the framework of Local Outlier Factor (LOF), which is based on the density estimate theory. However, LOF has two disadvantages that restrict its performance in outlier detection. First, the local density estimate of LOF is not accurate enough to detect outliers in the complex and large databases. Second, the performance of LOF depends on the parameter k that determines the scale of the local neighborhood. Our approach adopts the variable kernel density estimate to address the first disadvantage and the weighted neighborhood density estimate to improve the robustness to the variations of the parameter k, while keeping the same framework with LOF. Besides, we propose a novel kernel function named the Volcano kernel, which is more suitable for outlier detection. Experiments on several synthetic and real data sets demonstrate that our approach not only substantially increases the detection performance, but also is relatively scalable in large data sets in comparison to the state-of-the-art outlier detection methods.

Keywords: Outlier detection, Kernel methods, Local density estimate.

1 Introduction

Compared with the other knowledge discovery problems, outlier detection is arguably more valuable and effective in finding rare events and exceptional cases from the data in many applications such as stock market analysis, intrusion detection, and medical diagnostics. In general, there are two definitions of the

* This work is supported in part by the NSFC (Grant No. 60825204, 60935002 and 60903147) and the US NSF (Grant No. IIS-0812114 and CCF-1017828).

J.Z. Huang, L. Cao, and J. Srivastava (Eds.): PAKDD 2011, Part II, LNAI 6635, pp. 270–283, 2011.

outlier detection: Regression outlier and Hawkins outlier. Regression outlier defines that an outlier is an observation which does not match the predefined metric model of the interesting data [1]. Hawkins outlier defines that an outlier is an observation that deviates so much from other observations as to arouse suspicion that this observation is generated by a different mechanism [2]. Compared with Regression outlier detection, Hawkins outlier detection is more challenging work because of the unknown generative mechanism of the normal data. In this paper, we focus on the unsupervised methods for Hawkins outlier detection. In the rest of this paper, outlier detection refers particularly to Hawkins outlier detection.

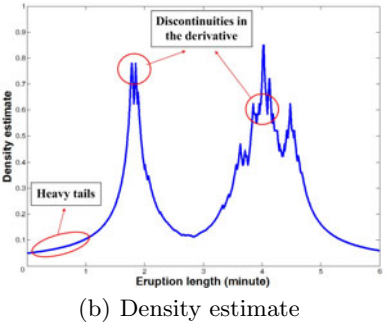
Over the past several decades, the research on outlier detection varies from the global computation to the local analysis, and the descriptions of outliers vary from the binary interpretations to probabilistic representations. Breunig et al. propose a density estimation based Local Outlier Factor (LOF) [4]. This work is so influential that there is a rich body of the literature on the local density-based outlier detection. On the one hand, plenty of local density-based methods are proposed to compute the outlier factors, such as the local correlation integral [5], the connectivity-based outlier factor [8], the spatial local outlier measure [9], and the local peculiarity factor [7]. On the other hand, many efforts are committed to combining machine learning methods with LOF to accommodate the large and high dimensional data [10,14].

Although LOF is popular in use in the literature, there are two major disadvantages restricting its applications. First, since LOF is based on the local density estimate theory, it is obvious that the more accurate the density estimate, the better the detection performance. The local reach-ability density used in LOF is the reciprocal of the average of reach-distances between the given object and its neighbors. This density estimate is an extension of the nearest neighbor density estimate, which is defined as

$$f(p) = \frac{k}{2n} \cdot \frac{1}{d_k(p)} \tag{1}$$

4.7	3.87	4.12	2.72	4.58	1.9
1.68	1.73	4	4.03	3.5	4.08
1.75	3.92	4.93	1.73	4.62	3.43
4.35	3.2	3.68	3.1	4.03	1.77
1.77	2.33	1.85	4.62	1.97	4.5
4.25	4.57	3.83	1.88	4.6	1.8
4.1	3.58	1.85	3.52	4	3.7
4.05	3.7	3.8	3.77	3.75	2.5
1.9	4.25	3.8	3.43	4	2.27
4	3.58	3.33	2	4.33	2.93
4.42	3.67	3.73	3.73	1.82	4.63
1.83	1.9	1.67	4.6	1.67	4
1.83	4.13	4.63	2.93	3.5	1.97
3.95	4.53	1.83	4.65	4.2	3.93
4.83	4.1	2.03	4.18	4.43	4.07
4.5	4.25	4.73	4.4	4.08	4.33
2.25	3.92	3.72	4.58	4.28	4.5
4.13	1.95	3.5	1.8		

(a) Old Faithful data



(b) Density estimate

Fig. 1. (a) Eruption lengths of 107 eruptions of Old Faithful geyser. (b) The density of Old Faithful data based on the nearest neighbor density estimate, redrawn from [3].