Knowledge-Based Systems 99 (2016) 1–9

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A relevant subspace based contextual outlier mining algorithm

Jifu Zhang[a,*], Xiaolong Yu[a], Yonghong Li[a], Sulan Zhang[a], Yaling Xun[a], Xiao Qin[b]

[a] School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024 PR China
[b] Department of Computer Science and Software Engineering, Auburn University, Auburn, AL 36849-5347 USA

ARTICLE INFO

Article history:
Received 17 May 2015
Revised 7 January 2016
Accepted 9 January 2016
Available online 23 February 2016

Keywords:
Contextual outlier
Relevant subspace
Interpretability and comprehensibility
Local sparsity
Probability density

ABSTRACT

For high-dimensional and massive data sets, a relevant subspace based contextual outlier detection algorithm is proposed. Firstly, the relevant subspace, which can effectively describe the local distribution of the various data sets, is redefined by using local sparseness of attribute dimensions. Secondly, a local outlier factor calculation formula in the relevant subspace is defined with probability density of local data sets, and the formula can effectively reflect the outlier degree of data object that does not obey the distribution of the local data set in the relevant subspace. Thirdly, attribute dimensions of constituting the relevant subspace and local outlier factor are defined as the contextual information, which can improve the interpretability and comprehensibility of outlier. Fourthly, the selection of N data objects with the greatest local outlier factor value is defined as contextual outliers. In the end, experimental results validate the effectiveness of the algorithm by using UCI data sets.

1. Introduction

Outlier is a general pattern or behavior, which significantly deviates from the other data. This type of data are inconsistent with other existing data [1], and often contain a large number of information which is not easily to be found but offers a very high level of research and application value. Outlier mining, an important branch of data mining and data analysis, has been widely used in astronomical spectroscopy [2], credit card fraud [3], network intrusion mining [4,5], data cleaning [6] and so on. At present, outlier mining algorithms can be broadly classified into such methods as statistics-based [7], distance-based [8], density-based [9,10], deviation-based [11], angle-based methods [3,12] and so on. However these methods assume that each dimension plays a uniform contribution for detecting outliers [13]. The existence of some irrelevant dimensions will increase effects of "disaster dimension" on mining results, and lead to unable to find some outliers hidden in the subspace.

For the high-dimensional and massive data, the effectiveness and efficiency of outlier mining have been seriously influenced and some outliers hidden in the subspace may not be found. In most cases, the outliers are such data objects which are apparently inconsistent with the distribution features of the local dataset. But the inconsistent valuable information can be given in some attribute dimensions, while nothing important information is afforded on the other attribute dimensions and these dimensions are nearly irrelevant to detect the outliers [14]. Therefore, it can effectively reduce the disturbance of "disaster dimension" and find the hidden outliers by searching and deleting the attribute dimensions that cannot provide some valuable information in the high dimensional datasets. Aiming at the problem of high-dimensional and massive data, a new contextual outlier mining algorithm based on relevant subspace is proposed in this paper. From the perspective of the local sparsity, the relevant subspace is redefined by making use of the local sparsity factor in the local dataset, and thus effectively described the distribution characteristics of all the local datasets. For the data objects in the relevant subspace, the calculation formula of the local outlier factor is given by using the probability density and Gaussian error function of the local dataset. Then the data attributes constituting the relevant subspace are regarded as the data objects' contextual information, and a contextual outlier mining algorithm is proposed, and accordingly it effectively reflects the inconsistency degree of the distribution in the local dataset and data objects in the relevant subspace, and provides the origin, meaning and features of outliers. The selection of n data objects which are of the most outlier, are identified as the contextual outlier. Finally, the interpretability and performance of the algorithms are validated by taking UCI data as experimental datasets. In this paper, the main highlights and contributions are summarized as follows:

- The relevant subspace, which can effectively describe the local distribution of the various data sets, is redefined by using the local sparseness of attribute dimensions.

* Corresponding author. Tel.: +86 3516998016.
E-mail address: jifuzh@sina.com, zjf@tyust.edu.cn (J. Zhang).

http://dx.doi.org/10.1016/j.knosys.2016.01.013
0950-7051/

- We proposed that the attribute dimensions and outlier factor of data object in the relevant subspace are regarded as the contextual information, which can effectively explain the outlier's meaning and cause, and is benefit for the understanding of outlier.
- The computing formula of outlier factor in data object' relevant subspace is presented to define the object's outlier degree.

2. Related work and analysis

Since mostly traditional outlier mining methods [7–12], detect outliers from all the dimensions of the data space, some irrelevant dimensions would have an effect on the efficiency and precision of outlier mining. It is one of the hot spots of high-dimensional outlier mining to project the data onto a subspace for outliers mining [14–16], however, it is more difficult to search for the meaningful subspace [15], and has the dimension-exponential complexity [14]. Aiming at outlier mining tasks, the typical methods of selecting meaningful subspace in the high-dimensional data set mainly include two categories: one is sparsity subspace method [2,17,18] and the other is relevant subspace method [13,19,20]. The sparse subspace method projects all overall the data onto the sparse subspace according to sparse coefficient threshold given by the user, and the data objects which are contained in the subspace are determined as outliers. Therefore, the sparse subspace method is a type of global subspace method. Typical work are: Agarwal et al. used genetic algorithm to search for sparse subspace [17], but the algorithm performance was affected by the initial population, the completeness and accuracy of outlier mining's results cannot be guaranteed. Aiming at the deficiencies in [17], Zhang et al., took concept lattice as a description tool for the subspace [2,18], and determined the sparse subspace through introducing the density coefficient. The method further improves the accuracy and completeness of the mining results, and has obtained better application in the astronomical spectrum data. However, the mining efficiency is lower because the construction of concept lattice is complex.

The basic idea of the relevant subspace method [13,14,16,19,20] is to look for the relevant subspace composed of some meaningful attribute dimensions in the datasets, from which the outliers are determined. The mainly used method is based on linear correlation of local reference datasets [13,19,20], as well as statistical models of local reference datasets [14,16] and so on. Typical work includes: Kriegel et al. [19] proposed an outlier detection method in axis-parallel subspaces of high dimensional data (SOD). By sharing the nearest neighbor (SNN), the method looks for a similar subset for each data object, then determines the axis-parallel linear correlation subspace in the similar subset based on the theory of attribute-based dimensions with low variance in the relevant subspace [19]. In the SOD, since the outlier degree of data objects is depicted with only one relevant subspace, this method will not be able to distinguish their outlier degree when the data objects are located in two or more than two subspace [14]. Clearly, it has some obvious deficiencies to detect outliers in the relevant subspace by using the mean-dimension of the Euclidean distance method. Muller et al. [14] proposed a method to select the meaningful subspace from datasets by using Kolmogorov–Smirnov test statistics method, and search for non-uniform distribution subspace (relevant subspace)from low-to high-dimensional in recursive ways, then regard the multiplicative of local outlier in each relevant subspace of data objects, as the ultimate outlier degree of the data objects, and accordingly solved the problem of outlier coefficient comparability in different subspaces. However, the complexity of determining all the relevant subspace in this algorithm is exponential time of the dimension [3], and the efficiency is low. Furthermore, the scalability of dimension for high dimensional datasets is poor, and unable to be adapted to the high-dimensional and massive data. Keller et al. [16] using Monte Carlo methods to find relevant subspace set, and this method measures the outlier degree of objects from the local subset which the data objects correspond to each relevant subspace, but the relevant subspace are actually achieved from the point of global view. Kriegel et al. [13], on the basis of the theory of principal component analysis, detected outlier by using of Mahalanobis distance with gamma distribution methods. Relevant subspace achieved by this method is arbitrary subspace of linear correlation, and the relevant subspace has good adaptability to linear data. However, this method is based on the deviation from the local data distribution, and it is liable to make mistakes in the relevant subspace obtained from maximum likelihood estimation when outliers was in the regional which not obviously reflects of correlation. Therefore, the method needs enough local data to reflect the obviously deviation trend, and the complexity of this algorithm is the cube of the dimension and difficult to fit for the high dimensional data. In addition, Mohamed et al. [20] find the relevant subspace by using of sparse density matrix, and make clustering analysis in the relevant subspace.

Most of outlier mining algorithms only to focus on the mining methods, and the reasons with the interpretation of outlier's generation are relatively few. So, the origins, meaning and characteristics of the outlier's generation have not been able to effectively explain. Tang et al. proposed an outlier model with contextual information in [21]. On this basis, an outlier mining algorithms with the contextual information is presented, but the mining efficiency is low because the sphere of searching outlier is first narrowed, then outlier is searched with enumerating method. Wang and Davidson proposed a probabilistic approach based on random walks graph, which is essentially a homogeneous Markov chain as characterized by a transition matrix [22]. The principal eigenvector of the transition matrix gives the stationary distribution of the nodes being visited in the graph under a global random walk, so the method can simultaneously explore meaningful contexts and score contextual outliers.

3. Relevant subspace and local outlier probability

According to the mining tasks on the different data sets, the data objects in the datasets only reflect the valuable information in some specific attribute dimensions, while there may little or no valuable information in other attribute dimensions [14]. Suppose DS be an arbitrary d-dimensional dataset, the attribute set F, $FS = \{A_1, A_2, \ldots, A_d\}$, and $x_{ij}(i = 1,2, \ldots n, j = 1,2, \ldots d)$ denote the jth attribute value of the ith data object. Referring to the reference [14] and [23], the basic concepts of the relevant subspace and local outlier probability are described as follows:

For an arbitrary subspace S, $S \subseteq FS$, data object o, $o \in DS$, and N(o, S) is the nearest neighbor of o. If S is a relevant subspace, then N(o, S) is non-uniform distribution. If S is not a relevant subspace, then N(o, S) is uniform distribution.

In [14], the non-uniform distribution of the subspace can effectively embody the valuable information of "outlier ", while the uniform distribution of attribute dimensions cannot reflect the valuable information of "outlier". Therefore, the measurement and searching for the relevant subspace has become the key to detect outliers since the relevant subspace can effectively reflect the valuable information of "outlier".

In [23], for an arbitrary data object o, $o \in DS$, suppose S be the correlation dataset, $S \subseteq DS$ and $o \in S$, then the probability distance pdist of o in s, satisfy the condition that $\forall s \in S$: $[d(o, s) \leqq pdist(o, S)] \geq \varphi$.

Intuitively, pdist (o, S) is the probability of the data objects' number contained in S, which is in the spherical with the center of o and the radius of pdist. Pdist (o, S) can be indirectly used to