Advanced Engineering Forum *Online: 2012-09-26*
ISSN: 2234-991X, Vols. 6-7, pp 621-624
doi:10.4028/www.scientific.net/AEF.6-7.621

Outlier Detection Algorithm Basing on Similarity Measurement Relation

Hong -Bin Fang

Wuxi City College of Vocational Technology , Wuxi Jiangsu 214153, China

email,: lq139666@sohu.com

Keywords: Outlier, data mining, similarity relation

Abstract. Outlier detection is an important field of data mining, which is widely used in credit card fraud detection, network intrusion detection ,etc. A kind of high dimensional data similarity metric function and the concept of class density are given in the paper, basing on the combination of hierarchical clustering and similarity, as well as outlier detection algorithm about similarity measurement is presented after the redefinition of high dimension density outliers is put. The algorithm has some value for outliers detection of high dimensional data set in view of experimental result.

Introduction

Outlier detection is important field in data mining after Outlier was defined by Hawkings[1], namely An outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by different mechanism. Outliers are probably generated because of measurement or executed error , etc, as the same time it could be considered as noise generally or decreased its effect in view of revised outlier value in order to pretreat data set. Outlier data is not error data because it probably contains important information, so that it is key object in data analysis, such as credit card fraud detection and disaster prediction as well as terror activity security[2,3]. Outlier detection firstly appeared in statistics[4], then was introduced to data mining by Knorr [2].At present Outlier detection methods have mainly five kinds ,namely method basing on statistics, method basing on depth, method basing on clustering, method basing on density and method basing on distance

At present outlier detection, integrating cluster analysis and distance, is ordinary. Clustering algorithms about high dimension data set and large-scale data set are mainly subspace cluster algorithm and object similarity algorithm, such as CLIQUE[5] and PROCLUS[6]. Subspace cluster algorithm is a valid approach to carry out clustering for high dimension data set , namely a kind of extension to traditional clustering algorithm , and its kernel method is locally search in part dimension.

On the other hand, clustering algorithms basing on object similarity have mainly graph partition algorithm basing on SL tree and HETIS algorithm[3,8]. Their defects produce more noise in high dimension harnessing distance measure to similarity among objects in light of point sparseness of high dimension space[2,3,8].

A new method of outlier detection basing on similarity measure is given in this paper after the combination of hierarchical clustering and similarity are considered. In light of the method, a new concept of similarity measure function and genus density in high dimension data is given so that outlier basing on genus density in high dimension data space is defined and accordingly a new outlier detection algorithm about high dimension data space, in which clustering of high dimension data is put up and outlier detection is based on genus density, is designed..

Similarity measure function

Similarity among data is given in light of distance among data traditionally, but Similarity among data in light of distance is invalid in high dimension space because of dimensionality disaster problem[9]. Many scholars considered similarity function Lately, such as literature [2] .In the literature [2], similarity function is defined as $F\sin(X_1, X_2)$, namely

$$H\sin(X_1, X_2) = \frac{\sum_{i=1}^{d} \frac{1}{1+|x_{1i} - x_{2i}|}}{d} \quad (1)$$

In the paper, a new similarity function is considered as $F\sin(X_1, X_2)$, namely

$$F\sin(X_1, X_2) = \frac{\sum_{i=1}^{d} \frac{m_i}{|x_{1i} - x_{2i}| + m_i}}{d} \quad (2)$$

Function $F\sin(X_1, X_2)$ is validated similarity function as following:

$0 \le F\sin(X_1, X_2) \le 1$;

iff $F\sin(X_1, X_2) = 1$, $X_1 = X_2$;

$F\sin(X_1, X_2) = F\sin(X_2, X_1)$;

Basing on the definition of similarity function of literature [8], Function $F\sin(X_1, X_2)$ is similarity function. $X_1 = (x_{11}, x_{12} \cdots, x_{1d})$ and $X_2 = (x_{21}, x_{22}, \cdots, x_{2d})$, $M = (m_1, m_2 \cdots, m_d)$, $m_i = \left|\sum_{j=1}^{N} x_{ji}\right|$. data set $S = \{X_1, X_2, \cdots\cdots, X_N\}$

Genus density definition

Data X_i （$i = 1, 2, \cdots, N$） is d dimension data and clustered k clusters $LC = \{C_1, C_2, \cdots C_k\}$,then genus density of C_i is defined as $dest(C_i) = \frac{|C_i|}{N}$, $|C_i|$ is the number of data in C_i. outlier basing on genus density is data in C_i when dest(C_i)<t and t is a given threshold value .

Outlier Detection Algorithm Basing on Similarity Measurement and genus density

Data $X_i (i = 1, 2, \cdots N)$ is d dimension data, compute the value of $F\sin$ between X_i and X_j . A formula of similarity among clusters is the following as: $s(P, Q) = \min\{s(x, y) \mid x \in P, y \in Q\}$ [3] , s(x,y) is the value of similarity between data X and data Y. Concrete algorithm is the following:

Step1 classify data set

1 calculate compute the value of $F\sin(X_1, X_2)$

2 initialize every data into a class

3 unite two classes when $s(P, Q) = \min\{s(x, y) \mid x \in P, y \in Q\}$ >=r, r is is a given threshold value.

4 repeat 3 until all $s(P, Q) = \min\{s(x, y) \mid x \in P, y \in Q\}$ <r, r is a given threshold value.

5 output all classes.

Step2 t is a given threshold value

For i=1:k

$dest(C_i) = \frac{|C_i|}{N}$

If $dest(C_i)$ <t

Then C_i is outlier class and output it

Endif

Endfor

Algorithm complexity analysis

Time complexity about clustering algorithm of similarity measure in literature [11] is $O(n^3)$， but time complexity about outlier detection algorithm basing on similarity measurement and genus density is $O(n)$ at the same time it can detect outliers of data set basing on threshold value easily.

Experiment data analysis

Experiment program in the paper is written by Matlab7.0 and realized through CPU of AMD Anthon 64 2.91GHz, at the same time experiment data is breast-cancer-wisconsin data from UCI , especially 0 represents special symbols. In the paper, outlier detection is discussed under the condition of threshod value r given. Experiment result is the following table 1 and table2

Table 1 r=0.8, t=0.05

datasize	number of clustering	number of outliers	runtime（ms）
10	9	0	47
20	15	11	609
50	35	38	10359
100	68	85	102172
150	102	129	353078
200	128	172	750531

Table 2 r=0.8, t=1/N, N is the number of data in experiment set

data size	number of clustering	number of outliers	runtime（ms）
10	9	0	47
20	15	11	625
50	35	26	10328
100	68	55	98297
150	102	85	340937
200	128	107	814468

The threshold value t has large influence on outlier detection in view of list 1 and list 2 under the condition of the same dimension and threshold value, on the other hand threshold value t is bigger when experiment set is bigger, vice versa. If t=1/N and N is the number of data in experiment set, the ratio of number of outlier is relatively stable with addition of the number of data in experiment set.

Summary

A new method of outlier detection basing on similarity measure is given in this paper after the combination of hierarchical clustering and similarity are considered. The number of Outliers in data set varies with threshold value t . In the future, similarity measure function must be optimized in order to detect outliers accurately.

Acknowledgement

This work was supported by the philosophic and social science foundation of university in JiangSu province of China(No. 2012SJD790059),and key science foundation of wuxi city college of vocational technology

References

[1] D Hawkins,Identifications of Outliers[M].London:Chapman and Hall,1980.

[2] EKnorr,RNg.Algorithms for mining distance-based outliers in large datasets[A].In Proc of the24th VLDB Conf[C].NewYork:Morgan Kaufmann,1998.392-403.

[3] J W Han,M Damber.Data Mining:Concepts and Technologies [M].San Francisco:Morgan Kaufmann,2001

[4] P J Rousseeuw,A M Leroy.Robust Regression and Outlier Detection[M].New York:John Wiley& Sons,1987

[5] Rakes h Agra w a, l JohannesGehrke , D m i it rios Gunopulos , et al. Au to m at ic Subspace Clustering of H i gh D i m ens i ona lDat a for Data Mining Applicati on [C] / /Proceed i ngs of the 1998 ACMSIGMOD Internation a Conference on Management of Data, Seattle, Washington , 1998.

[6] A ggarwal C C, P rocopiuc C, Wolf JL, etal Fast al gorithmsf or projected clustering [C]//Proc.of the ACM SIGMOD Conference Philadel Phia ,P A,1999: 61-72.

[7] A gra w alR, Geh rke J . Gun opol os D, et a. l Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications . In ACM SIGMOD Con f eren ce, 1998.

[8] Zenshui Xu,Meimei Xia.Distance and similarity measures for hesitant fuzzy sets[J].Information Sciences,2011.2128-2138.