ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

OUTLIER DETECTION USING INNER AND OUTER RADIUS BASED METHOD

Vikas Kumar*, Ankur Singh Bist, Rudranshu Sharma
* U.P.T.U.
U.P.T.U.

ABSTRACT

Outlier detection is a fundamental issue in data mining, specifically it has been used to detect and remove anomalous objects from data mining. The proposed approach to detect outlier includes two distances which are inner radius and outer radius. Inner radius is calculated from the global centroid distance to the nearest cluster distance minus the radius of that cluster. Similarly we calculate outer radius which is the maximum distance between global centroid and any one of the cluster plus that cluster radius. For clustering FCM algorithm is used which partition the dataset into given number of clusters. The clustering is done only on useful data points. This will act as a model of my project on the basis of these clusters we will point out outlier. These two radius we will point out the outlier points. While pointing any point to be an outlier we will also check, are there any groups of points which form another cluster, for that case we have to check that condition separately. Those points which are outside outer radius are outlier points and those points which are less then inner radius are outlier points.

KEYWORDS: Data Set Information, Iris Data Set & ABALONE DATA SET.

INTRODUCTION

Outlier detection is a fundamental step in a large number of data quality, data management, and data analysis tasks. Stephen Hawkins defines an outlier as “an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”. The problem of outlier detection has been around for over a century and has been the focus of much research in the statistics literature. Here, data is assumed to follow a radius distribution and the objects that do not fit properly the model are considered outliers.

Double Radius-based detection is a good style outlier detection approach. We denote K-Nearest Neighbor distance between the centroid of clusters in this paper. The global inner distance from any centroid to global centroid minus centroid radius denoted as G_in radius. Similarly find out global outer radius denoted as G_out.

PROBLEM IN MEAN DISTANCE BASED OUTLIER DETECTION ALGORITHM

However, it mainly has two shortcomings when it is applied for outlier detection: the first one is that it distinguishes the normal and abnormal dataset just by a value of delta. So the clustering accuracy is far from enough. Second, it doesn’t offer a reasonable method to address outliers, but just simply throw it away. With this coarse granularity partition, it can’t receive a satisfied detection rate.
In distance based outlier detection, time computation is too high because each data set is compared with the delta value.

PROPOSED INNER AND OUTER RADIUS BASED OUTLIER DETECTION ALGORITHM

By this algorithm we can easily determine outlier based on the value of G_out and G_in .
Algorithm:
INPUT: a set S of points

1) apply FCM on S which divides the points into k clusters
2) Calculate global centroid C_global
3) Calculate radius of each cluster
4) C_out is the outer radius from global centroid

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

5) C_in is the inner radius from global centroid
6) If point lie between C_out and C_in then are normal point
7) Else
8) Detected outlier

EXPERIMENTAL RESULT ON DATA SETS

9.1. Diagnostic Wisconsin Breast Cancer Database Data Set:

Data Set Characteristics	Multivariate	Number of Instances	569	Area	Life
Attribute Characteristics	Real	Number of Attributes	32	Date Donated	1995-11-01
Associated Tasks	Classification	Missing Values?	No	Number of Web Hits	276511

Data Set Information:
Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.
Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

Results for WDBC Data Set:

Total Number Of Points	Model Built From	Number Of Cluster	Number Of Benign Points	Number Of Malignant Points	Outlier Detected
200	100	5	80	20	18
200	100	6	80	20	18
300	150	10	100	50	46
350	150	11	150	50	45
400	100	15	250	50	43
400	100	20	200	100	85
400	150	25	200	100	86
400	200	30	150	50	28
450	200	30	150	100	52
500	100	35	300	100	41

Iris Data Set:

Data Set Characteristics	Multivariate	Number of Instances	150	Area	Life
Attribute Characteristics	Real	Number of Attributes	4	Date Donated	1998-07-01
Associated Tasks	Classification	Missing Values?	No	Number of Web Hits	562966

Data Set Information:
This is perhaps the best known database to be found in the pattern recognition literature. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class: Iris Setosa , Iris Versicolour , Iris Virginica

Result for Iris Data Set:

Total Number Of Points	Model Built From	Number Of Cluster	Number Of Inliers Points	Number Of Outliers Points	Outlier Detected
50	20	4	20	10	9
50	20	6	20	10	8
60	30	5	20	10	9
60	30	6	20	10	9
70	30	10	30	10	8
70	30	15	25	15	13
75	30	15	20	25	23
75	35	15	20	15	14
80	30	15	40	10	9

ABALONE DATA SET:

Data Set Characteristics	Multivariate	Number of Instances	4177	Area	Life
Attribute Characteristics	Categorical, Integer, Real	Number of Attributes	8	Date Donated	1995-12-01
Associated Tasks	Classification	Missing Values?	No	Number of Web Hits	219061

Data Set Information:
Predicting the age of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.

Attribute Information:
Given is the attribute name, attribute type, the measurement unit and a brief description. The number of rings is the value to predict: either as a continuous value or as a classification problem.
Name / Data Type / Measurement Unit / Description

Sex / nominal / -- / M, F, and I (infant)
Length / continuous / mm / Longest shell measurement
Diameter / continuous / mm / perpendicular to length
Height / continuous / mm / with meat in shell
Whole weight / continuous / grams / whole abalone
Shucked weight / continuous / grams / weight of meat
Viscera weight / continuous / grams / gut weight (after bleeding)
Shell weight / continuous / grams / after being dried
Rings / integer / -- / +1.5 gives the age in years

Result for Abalone Data Set:

Total Number Of Points	Model Built From	Number Of Cluster	Number Of Benign Points	Number Of Malignant Points	Outlier Detected
700	300	10	300	100	86
700	400	15	200	100	80
900	500	15	300	100	90
900	500	17	300	100	90
1000	600	20	250	150	75
1100	700	25	300	100	69
1100	700	30	300	100	60

CONCLUSIONS
In simply mean distance measures will not provide the good result. By using the double radius technique it will provide the good result for detecting outlier. This Radius Based algorithm presented in this paper may overcome some disadvantages of the Mean Distance Based algorithm for intrusion detection, because in Mean Distance Based algorithm we will discard points on the bases of mean value, this will not detect boundary outlier. So to overcome these problem we use double radius for outlier detection.

REFERENCES
1. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. Communications of the ACM, 51(1):117–122, 2008.
2. F. Angiulli and F. Fassetti. Very efficient mining of distance-based outliers. In M. J. Silva, A. H. F. Laender, R. A. Baeza-Yates, D. L. McGuinness, B. Olstad, Ø. H. Olsen, and A. O. Falc˜ao, editors, CIKM, pages 791–800. ACM, 2007.
3. F. Angiulli and C. Pizzuti. Fast outlier detection in high dimensional spaces. In PKDD ’02: Proc. of the 6th European Conf. on Principles of Data Mining and Knowledge Discovery, pages 15–26, London, UK, 2002. Springer-Verlag.
4. S. D. Bay, D. Kibler, M. J. Pazzani, and P. Smyth. The uci kdd archive of large data sets for data mining research and experimentation. SIGKDD Explor. Newsl., 2(2):81–85, 2000.

5. S. D. Bay and M. Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In 9th ACM SIGKDD Int. Conf. on Knowledge Discovery on Data Mining, 2003.
6. M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: Identifying density-based local outliers. In W. Chen, J. F. Naughton, and P. A. Bernstein, editors, Proc. of the 2000 ACM SIGMOD Int. Conf. on Management of Data, May 16-18, 2000, Dallas, Texas, USA, pages 93–104. ACM, 2000.
7. M. Ester, J. Kriegel, H. P.and Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial fatabases with noise. In In Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. AAAI Press, 1996.
8. C. Faloutsos and K. Lin. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In Proceedings of the 1995 ACM SIGMOD international conference on Management of data, pages 163–174. ACM New York, NY, USA, 1995.
9. A. Ghoting, S. Parthasarathy, and M. E. Otey. Fast mining of distance-based outliers in high-dimensional datasets. 6th SIAM Int. Conf. on Data Mining, April 2005.
10. A. Ghoting, S. Parthasarathy, and M. E. Otey. Fast mining of distance-based outliers in high-dimensional datasets. Data Min. Knowl. Discov., 16(3):349–364, 2008.
11. S. Guha, R. Rastogi, and K. Shim. Cure: an efficient clustering algorithm for large databases. In SIGMOD ’98: ACM SIGMOD Int. Conf. on Management of data, pages 73–84, New York, NY, USA, 1998. ACM.
12. Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min. Knowl. Discov., 2(3):283–304, 1998.
13. E. M. Knorr and R. T. Ng. Finding intensional knowledge of distance-based outliers. In VLDB ’99: 25th Int. Conf. on Very Large Data Bases, pages 211–222, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.
14. H. Kriegel, P. Kroger, and A. Zimek. Outlier Detection Techniques. In Tutorial at the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2009.
15. J. Laurikkala, M. Juhola, and E. Kentala. Informal identification of outliers in medical data. In The Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology. Citeseer, 2000.
16. M. Mahoney and P. Chan. Learning nonstationary models of normal network traffic for detecting novel attacks. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 376–385. ACM New York, NY, USA, 2002.
17. M. Mahoney and P. Chan. Learning rules for anomaly detection of hostile network traffic. In Proceedings of the Third IEEE International Conference on Data Mining, page 601. Citeseer, 2003.