Outlier Detection Algorithms in Data Mining

Sharmila Shinde[1] *Madhu M.N.*[2]

Abstract

The identification of outliers can lead to the discovery of truly unexpected knowledge in areas such as electronic commerce, credit card fraud; Outlier is defined as an observation that deviates too much from other observations. Existing methods that we have seen for finding outliers can only deal efficiently with two dimensions/ attributes of a dataset. Many recent algorithms have been proposed for outlier detection that uses several concepts of proximity in order to find the outliers based on their relationship to the other points in the data. However, in high-dimensional space, the data are sparse and concepts using the notion of proximity fail to retain their effectiveness This paper mainly discusses and compares approach of different outlier detection from data mining perspective, which can be categorized into two categories: classic outlier approach and spatial outlier approach

I. Introduction

Knowledge discovery tasks can be classified into four general categories: (a) dependency detection, (b) class identification, (c) class description, and (d) exception/ outlier detection. The first three categories of tasks correspond to patterns that apply to many objects, or to a large percentage of objects, in the dataset. Most research in data mining – such as association rules [1], classification [5], and data clustering [10, 24] – belongs to these three categories. The fourth category, in contrast, focuses on a very small percentage of data objects, which are often ignored or discarded as noise. For example, some existing algorithms in machine learning and data mining have considered outliers, but only to the extent of tolerating them in whatever the algorithms are supposed to do [10, 24]. It has been said that "One person's noise is another person's signal." Indeed, for some applications, the rare events are often more interesting than the common ones, from a knowledge discovery standpoint. Sample applications include the detection of credit card fraud and the monitoring of criminal activities in electronic commerce. For example, in Internet commerce or smart-card applications, we expect many low-values

transactions to occur. However, it is the exceptional cases – exceptional perhaps in monetary amount, type of purchase, timeframe, location, or some combination thereof – that may interest us, either for fraud detection or marketing reasons.

II. Previous Work

The existing work on outlier detection lies in the field of statistics [3, 17]. While there is no single, generally accepted, formal definition of an outlier, Hawkins' definition captures the spirit: "an outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism"

[1] IT Department, BVCOE, Navi Mumbai ,India ,PIIT, New Panvel,India. sharmi_anant@yahoo.co.uk,

[2] Computer Department, BVCOE, Navi Mumbai ,India ,PIIT, New Panvel,India. madhu.nashipudi@yahoo.in

[17]. Outlier detection techniques can be categorized into several groups: (1) statistical or distribution-based approaches; (2) geometric-based approaches; (3) profiling methods; and (4) model-based approaches. In statistical techniques [2, 3], the data points are typically modeled using a data distribution, and points are labeled as outliers depending on their relationship with the distributional model. Geometric-based approaches detect outliers by (i) computing distances among points using all the available features [4, 5] or only feature projections [6]; (ii) computing densities of local neighborhoods [7, 8]; (iii) identifying side products of the clustering algorithms (as points that do not belong to clusters) [9] or as clusters that are significantly smaller than others. In profiling methods, profiles of normal behavior are built using different data mining techniques or heuristic-based approaches, and deviations from them are considered as outliers. Finally, model-based approaches usually first characterize the normal behavior using some predictive models (e.g., replicator neural networks [10] or unsupervised support vector machines [11]), and then detect outliers as deviations from the learned model. all of those tests suffer from the following two serious problems. First, almost of them are univariate(i.e., single attribute). This restriction makes them unsuitable for multidimensional datasets

III. Outlier Detection Approach

These approaches can be mainly classified into two categories: classic outlier approach and spatial outlier approach. The classic outlier approach analyzes outlier based on transaction dataset, which can be grouped into statistical-based approach, distance-based approach, deviation-based approach, density-based approach. The spatial outlier approaches analyze outlier based on spatial dataset, which can be grouped into space-based approach and graph-based approach, as illustrated in Figure 1.

3.1. Classic Outlier

Classic outlier approach analyzes outlier based on transaction dataset, which consists of collections of items. A typical example is market basket data, where each transaction is the collection of items purchased by a customer in a single transaction. Such data can also be augmented by additional "items" describing the customer or the context of the transaction. Commonly, transaction data is relative to other data to be simple for the outlier detection. Thus, most outlier approaches are researched on transaction data.

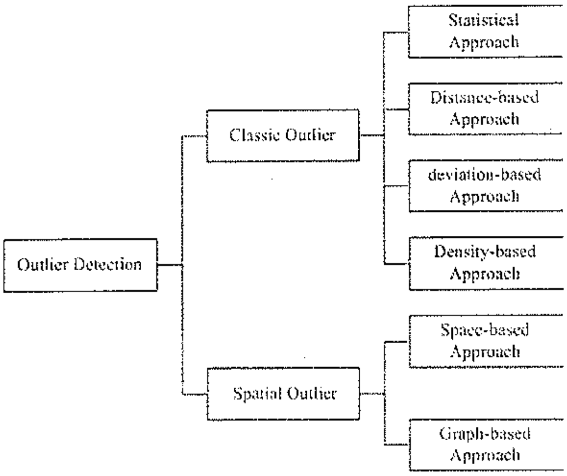

Figure 1 Outlier Detection Approach

Statistical Approach Statistical approaches were the earliest algorithms used for outlier detection, which assumes a distribution or probability model for the given data set and then identifies outliers with respect to the model using a discordancy test.In fact, many of the techniques described in both Barnett and Lewis [15] and Rousseeuw and Leroy [16] are single dimensional. However, with the dimensions increasing, it becomes more difficult and inaccurate to make a model for dataset.

Distribution-based approaches (Hawkins, 1980; Barnett and Lewis, 1994; Rousseeuw and Leroy, 1996) develop statistical models (typically for the normal behavior) from the given data and then apply a statistical test to determine if an object belongs to this model or not. Objects that have low probability to belong to the statistical model are declared as outliers. However, Distribution-based approaches cannot be applied in multidimensional scenarios because they are univariate in nature. In addition, a prior knowledge of the data distribution is required, making the distribution-based approaches difficult to be used in practical applications (Zhang and Wang, 2006). In the distance-based approach [(Knorr, 1998; Knorr, 2000; Ramaswami, 2000; Angiulli and Pizzut, 2005), outliers are detected as follows. Given a distance measure on a feature space, a point q in a data set is an outlier with respect to the parameters M and d, if there are less than M points within the distance d from q, where the values of M and d are decided by the user. The problem with this approach is that it is difficult to determine the values of M and d.

Density-based approaches (Breunig, 2000; Papadimitriou, 2003) compute the density of regions in the data and declare the objects in low dense regions as outliers. In (Breunig, 2000), the authors assign an outlier score to any given data point, known as Local Outlier Factor (LOF), depending on its distance from its local neighborhood. A similar work is reported in (Papadimitriou, 2003).

Clustering-based approaches (Loureiro, 2004; Gath and Geva, 1989; Cutsem and Gath, 1993; Jiang, 2001; Acuna and Rodriguez, 2004), consider clusters of small sizes as clustered outliers. In these approaches, small clusters (i.e., clusters containing significantly less points than other clusters) are considered outliers. The advantage of the clustering-based approaches is that they do not have to be supervised. Moreover, clustering-based techniques are capable of being used in an incremental mode (i.e., after learning the clusters, new points can be inserted into the system and tested for outliers).

3.2. Spatial Outlier

For spatial data, classic approaches have to be modified because of the qualitative difference between spatial and non-spatial attributes. Spatial dataset could be defined as a collection of spatially referenced objects, such as roads, buildings and cities. Attributes of spatial objects fall into two categories: spatial attributes and nonspatial attributes. The spatial attributes include location, shape and other geometric or topological properties. Nonspatial attributes include length, height, owner, building age and name. A spatial neighborhood of a spatially referenced object is a subset of the spatial data based on the spatial dimension using spatial relationships, e.g., distance and adjacency. Comparisons between spatially referenced objects are based on non-spatial attributes [21]. Spatial outliers are spatially referenced objects whose non-spatial attribute values are significantly different from those of other spatially referenced objects in their spatial neighborhoods. Informally, a spatial outlier is a local instability, or an extreme observation with respect to its neighboring values, even though it may not be significantly different from the entire population.

Detecting spatial outliers is useful in many applications of geographic information systems and spatial dataset [11, 21, 22]. The identification of spatial outliers can reveal hidden but valuable information in many applications, For example, it can help locate severe meteorological events, discover highway congestion segments, pinpoint military targets in satellite images, determine potential locations of oil reservoirs, and detect water pollution incidents. (1) Space-based Approach Space-based outliers use Euclidean distances to define spatial neighborhoods. Kou et al. developed spatial weighted outlier detection algorithms which use properties such as center distance and common border length as weight when comparing non-spatial attributes [12]. Adamet al. proposed an algorithm which considers both spatial relationship and semantic relationship among neighbors [13]. Liu et al. proposed a method for detecting outliers in an irregularly-distributed spatial data set [14].

(2) Graph-based Approach Graph-based Approach uses graph connectivity to define spatial neighborhoods. Yufeng Kou et al. proposed a set of graph-based algorithms to identify spatial outliers, which first constructs a graph based on k-nearest neighbor relationship in spatial domain, assigns the non-spatial attribute differences as edge weights, and continuously cuts high- weight edges to identify isolated points or regions that are much dissimilar to their neighboring objects. The algorithms have two major advantages compared with the existing spatial outlier detection methods: accurate in detecting point outliers and capable of identifying region outliers [24].

IV. Recent Advances in Outlier Detection

Along with the fast development of data mining technique, identification of outliers in large dataset has received more and more attention. Traditional outlier detection methods may not be efficiently applicable to large dataset. So some new methods are specially designed for special background.

(1) High Dimension-based Approach High dimension space is a difficult problem for outlier detection. According to the criterion of the technique designed for high dimension proposed in the literature [22] , a new method ODHDP based on the concept of projection is proposed in this paper, it can well deal with the sparsity of high dimensional points. The basic idea of the approach is to find the outliers by clustering the projections of data set. So, firstly, clustering the projections of data set in each dimension, and putting different weight to each dimension; secondly, selecting the dimension which has the maximum weight in the rest of dimensions for Descartes combination clustering in turn, then pruning the candidate clusters in which the number of the points is less than threshold, until all dimensions are scanned; thirdly, computing the similarity of the points in the remains based on their relationship with the clusters in full dimension, by which the outliers is distinguished from the remaining [23].

(2) SVM-based Approach

A SVM-based outlier detection approach was proposed [25]. The method uses several models of varying complexity to detect outliers based on the characteristics of the support vectors obtained from SVM-models. This has the advantage that the decision does not depend on the quality of a single model, which adds to the robustness of the approach. Furthermore, since it is an iterative approach, the most severe outliers are removed first. This allows the models in the next iteration to learn from “cleaner” data and thus reveal outliers that were “masked” in the initial model. Other outlier detection efforts include Support Vector approach [26], using Replicator Neural Networks (RNNs) [27], and using a relative degree of density with respect only to a few fixed reference points [28].

V. Conclusions

This paper mainly discusses about outlier detection approaches from data mining perspective. Firstly, we reviews related work in outlier detection. Next, we discuss and compare algorithms of outlier detection which can be categorized into two categories: classic outlier approach and spatial outlier approach. The classic outlier approach analyzes outlier based on transaction dataset, which can be grouped into statistical-based approach, distance-based approach, deviation-based approach, density-based approach. The spatial outlier approach analyzes outlierbased on spatial dataset, which can be grouped into space based approach, graph-based approach. Thirdly, we conclude some advances in outlier detection recently.

References

[1] Yu, D., Sheikholeslami, G. and Zang, "A find out: finding outliers in very large datasets". *In Knowledge and Information Systems*, 2002, pp.387-412.

[2] Breunig, M.M., Kriegel, H.P., and Ng, R.T., "LOF: Identifying density-based local outliers." *ACM Conference Proceedings*, 2000, pp. 93-104.

[3] D. M. Hawkins, "Identification of Outliers". *Chapman and Hall*, London, 1980.

[4] Aggarwal, C. C., Yu, S. P., "An effective and efficient algorithm for high-dimensional outlier detection, *The VLDB Journal*, 2005, vol. 14, pp. 211-221.

[5] Yamanishi. K, Takeuchi. J ,and Williams. G On-line, "unsupervised outlier detection using finite mixtures with discounting learning algorithms". *In Proceedings of the Sixth ACM SIGKDDOO*, Boston, MA, USA, pp.320-324.

[6] Knorr, E.M., Ng, R.T., "Finding Intentional Knowledge of Distance-Based Outliers", *Proceedings of the 25th International Conference on Very Large Data Bases*, Edinburgh, Scotland, pp.211-222, September 1999.

[7] Agarwal, D., Phillips, J.M., Venkatasubramanian, "The hunting of the bump: on maximizing statistical discrepancy". *In: Proc. 17th Ann. ACM-SIAM Symp.* on Disc. Alg. pp. 1137–1146 (2006).

[8] Berchtold, S., Keim, D., Kriegel, H.-P, "The X-tree: An efficient and robust access method for points and rectangles". *In: VLDB (1996).*

[9] Jin, W., Tung, A.K.H., Han, J.W. "Mining Top-n Local Outliers in Large Databases". *In: KDD (2001).*

[10] Lazarevic, A., Kumar" Feature Bagging for Outlier Detection". *In: KDD (2005).*

[11] S. C. Shashi Shekhar, "Spatial Databases: A Tour. Prentice Hall", 2003.

[12] Y. Kou, C.-T. Lu, and D. Chen. "Spatial weighted outlier detection". *In Proceedings of the Sixth SIAM International Conference on Data Mining*,pp. 614–618, Bethesda, Maryland, USA, 2006.

[13] N. R. Adam, V. P. Janeja, and V. Atluri., "Neighborhoodbased detection of anomalies in high-dimensional spatiotemporal sensor datasets". *In Proceedings of the 2004 ACM symposium on Applied computing*, Nicosia, Cyprus, 2004. pp. 576–583

[14] H. Liu, K. C. Jezek, and M. E. O'Kelly, "Detecting outliers in irregularly distributed spatial data sets by locally adaptive and robust statistical analysis and gis". *International Journal of Geographical Information Science*, 15(8), 2001. pp.721–741.

[15] Barnett, V. & Lewis, T. (1994).,"Outliers in Statistical Data", *3rd edn.* John Wiley & Sons.

[16] Rousseeuw, P. & Leroy, A. (1996).,"Robust Regression and Outlier Detection", *3rd edn.* John Wiley & Sons.

[17] E. Knorr, R. Ng, and V. Tucakov, "Distance-Based Outlier: Algorithms and Applications," *VLDB* J., vol. 8, nos. 3-4 2000, pp. 237-253.

[18] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient Algorithms for Mining Outliers from Large Data Sets," *Proc. Int'l Conf. Management of Data (SIGMOD '00)*, 2000, pp. 427-438.

[19] A. Arning, R. Agrawal, and P. Raghavan, "A Linear Method for Deviation Detection in Large Databases," Proc. Int'l Conf. Knowledge Discovery and Data Mining, 1996, pp. 164-169.

[20] Papadimitriou, S., Kitawaga, H., Gibbons, P., Faloutsos, C., "LOCI: Fast outlier detection using the local correlation integral", *Proc. of the Int'l Conf. on Data Engineering*, 2003.

[21] Chang-Tien Lu, Dechang Chen,Yufeng Kou, "Detecting spatial outliers with multiple attributes", *Tools with Artificial Intelligence, 2003.* Proceedings. 2003, pp.122–128.

[22] Aggarwal, C.C, Yu, P. "Outlier detection for high dimensional data", *Proceedings of the ACM SIGMOD International Conference on Management of Data*. Santa Barbara, CA, 2001, pp. 37-47.

[23] Ping Guo, Ji-Yong Dai, Yan-Xia Wang, "Outlier Detection in High Dimension Based on Projection", Machine Learning and Cybernetics, 2006 International Conference, 2006,pp.1165 – 1169.

[24] Yufeng Kou, Chang-Tien Lu, Dos Santos, R.F." Spatial Outlier Detection: A Graph-Based Approach", ICTAI 2007, Volume 1, 2007,pp.281 – 288.

[25] E.M. Jordaan. Deployment of Robust Inferential Sensors, "Irtdusrriol application of Supper Vector Machines for Regression", Ph. D. thesis. Eindhaven University of Technology, 2002.

[26] Jordaan, E.M.;,Smits, G.F., " Robust outlier detection using SVM regression", Neural Networks, 2004. Proceedings. 2004,pp.2017 – 2022.

[27] Harkins, S., He, H., Williams, G., Baster, R., "Outlier Detection Using Replicator Neural Networks", *DaWaK'02*, 2002, pp. 170-180.

[28] Pei, Y., Zaiane, O., Gao, Y., "An Efficient Reference based Approach to Outlier Detection in Large Dataset", *IEEE Int'l Conference on Data Mining*, 2006. 97