Identifying erroneous data using outlier detection techniques

Wei Zhuang[1], Yunqing Zhang[2] and J. Fred Grassle[2]

[1]Department of Computer Science, Rutgers, the State University of New Jersey, Piscataway, NJ 08854-8019, USA E-mail: weiz@paul.rutgers.edu

[2]Institute of Marine and Coastal Sciences, Rutgers, the State University of New Jersey, New Brunswick, NJ 08901, USA

Abstract

Common data quality problems observed in OBIS data integration processes are described. DBSCAN, a density-based clustering algorithm for large spatial databases is employed to identify geographical outliers in federated data from a public Web service on the OBIS Portal. The algorithm is shown to be effective and efficient for this purpose. The relationship between outliers and erroneous data points are discussed and the future plan to develop an operational data quality checking tool based on this algorithm is discussed.

Keywords: QA/QC; outliers; clustering; data quality solving.

Introduction

Federated scientific databases such as the Ocean Biogeographic Information System (OBIS, http://www.iobis.org) and the Global Biodiversity Information Facility (GBIF, http://www.gbif.org), have solved the data heterogeneity problem by employing open communication protocols and data exchange standards (Grassle and Stocks, 1999; Zhang and Grassle, 2003). For the first time in science history, tens of millions of records about our shared biodiversity heritage have been made publicly accessible on the World Wide Web. Although scientists and data managers have carefully performed quality checking over individual datasets and collections, data corruptions can still occur during data compilation, re-entry, conversion, and the transfer process. For example, default null database objects can turn into string "null"s; latitudes and longitudes are reversed; non-ascii characters are mistakenly encoded, etc. These data quality concerns have been familiar to data warehousing communities and a great deal of research and development has been carried out in this area. Our problem, which is data quality checking over federated ocean biodiversity information, is unique in at least two aspects:

- DQ solving has to be efficient for large spatial datasets
- The domain knowledge is highly specialized and may not be translated into simple database constraints in many cases (this is indeed a common problem for scientific data management).

Outliers are commonly defined as rare or atypical data objects that do not behave like the rest of the data. Often, erroneous data points appear as outliers in a database. Scientists and data managers have used visualization tools to identify outliers in datasets. When the dimension of the feature space is more than two, visual identification becomes challenging. Moreover, when the database multiplies in content, manual identification by naked eyes becomes infeasible. Henceforth what is needed here is an automatic outlier detection tool that can efficiently handle large, high dimensional databases. It should be made clear that automatic tools are not to replace domain scientists' opinion in data quality checking, rather, they are "pre-processors" to provide assistance to domain scientists. The question of how to integrate domain knowledge in automatic outlier identification tools will be discussed elsewhere. In this paper we will concentrate on the algorithm testing aspect of the development.

Section II describes the method. In section III we report and analyze the results. We discuss the results and direction for future work in Section IV.

Method

There is a considerable body of research on outliers by statisticians (Barnett and Lewis, 1994; Hawkins, 1980). Fitting databases with parametric models requires prior knowledge of data distribution and using parametric models in the data processing stage may lead to circular arguments and produce spurious patterns when doing data analysis. Non-parametric clustering algorithms are attractive for grouping objects in a database into subclasses and, intuitively, small clusters, or classes with few members, are where outliers are. Computer scientists have been conducting extensive research to develop efficient clustering algorithms for large databases (Berkhin, 2002; Guha *et al.*, 1998; Zhang *et al.*, 1996; Ng and Han, 1994,).

The well-known K-means algorithms partition a dataset into a set of k clusters in two steps: firstly it determines the k cluster centers by minimizing an object function; secondly it assigns a cluster membership based on the distance of the data object to the cluster centers. In these partition-based algorithms, the number of clusters, k, is an input parameter provided by the user while in many cases the user has no idea of the number of clusters. These algorithms are sensitive to noise. We then investigate a different, density-based family of clustering algorithms. In these algorithms, parts of the feature space with dense data points form clusters while outliers have a much lower density and are further away from the clusters. DBSCAN (Ester *et al.*, 1996), DENCLUE (Hinneberg and Kleim, 1998) and WaveCluster (Sheikholeslami *et al.*, 1999) are well known algorithms in this family. It has been demonstrated that DBSCAN requires minimal domain knowledge, can discover clusters with arbitrary shapes and is efficient on large databases. Most importantly, it can separate "noise" (*i.e.* outliers) while performing clustering. Details of the algorithm can be found in Ester *et al.* (1996).

We obtained the software from the first author as a C++ package and adapted it for OBIS data. The clustering was run on a Sun Solaris 9 machine. The experimental data are provided by the OBIS portal as a Web service. We tested the algorithm by

identifying geographical outliers and great circle distance is used to define the distance function between two data points.

Results

Here we report the results for three species: *Euthynnus alletteratus*, *Albula vulpes* and *Balaenoptera borealis*. The time complexity of DBSCAN is O (nlogn) where n is the number of data points. In table I we list the run time for these three experiments and the number is consistent with the O (nlogn) estimation.

Table I: Runtime for clustering and identifying outliers using DBSCAN.

Dataset	Number of records	Runtime (in milliseconds)
Euthynnus alletteratus	338	1780
Albula vulpes	840	5693
Balaenoptera borealis	7125	424910

In Fig. 1-3 we show clustering results for the three species where outliers are represented by round dots and non-outliers triangles. Examining the three figures together with the underlying datasets, we see that this algorithm correctly identifies all the single records far away from data clusters. Some non-outliers may appear to be outliers to the naked eye. For example, the triangle at (44.15°N, 6.03°E) in Fig. 2 is far away from the other clusters but in fact it represents 12 individual data records and thus is not an outlier in its common definition. One could visit the OBIS Portal to look up the interactive maps and download the datasets for further confirmation of our results.

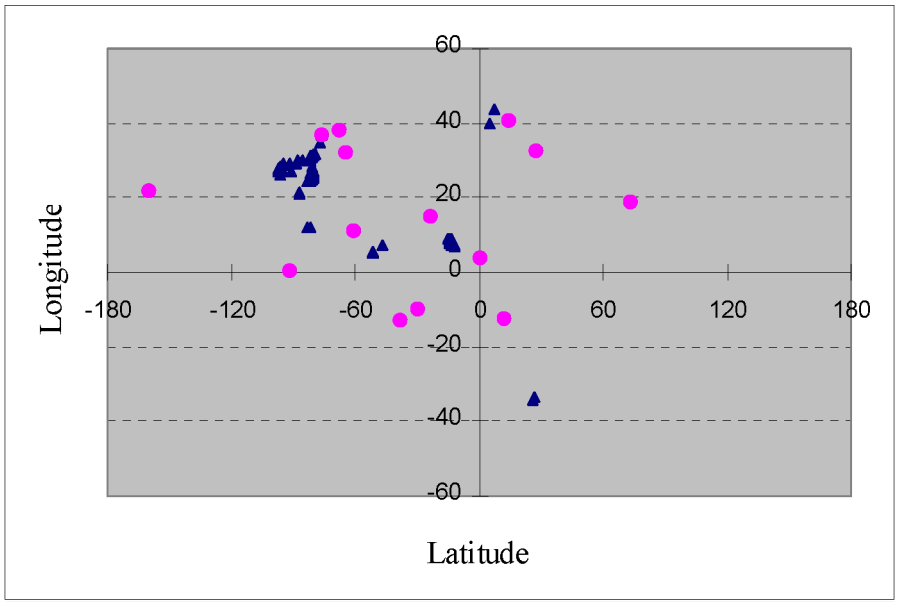

Fig. 1. Result set for Euthynnus alletteratus.

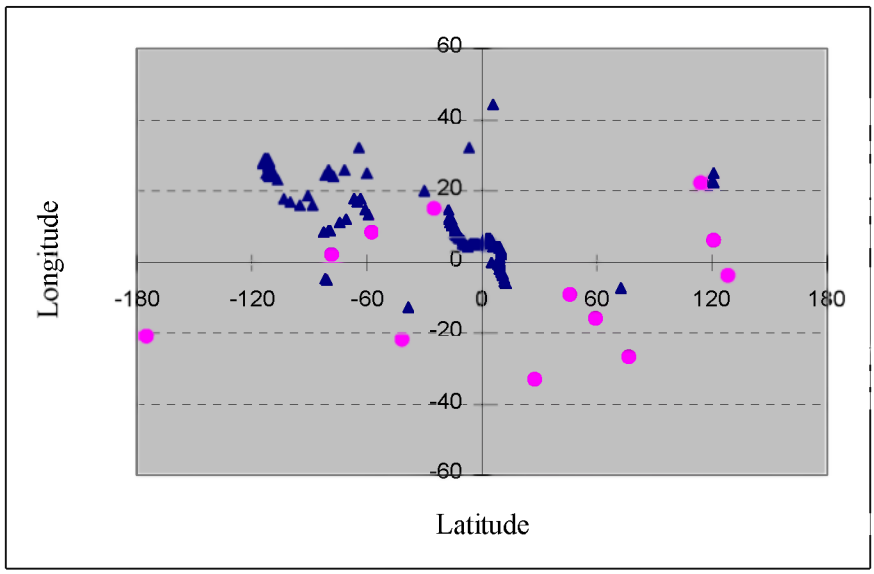

Fig. 2. Result set for Albula vulpes.

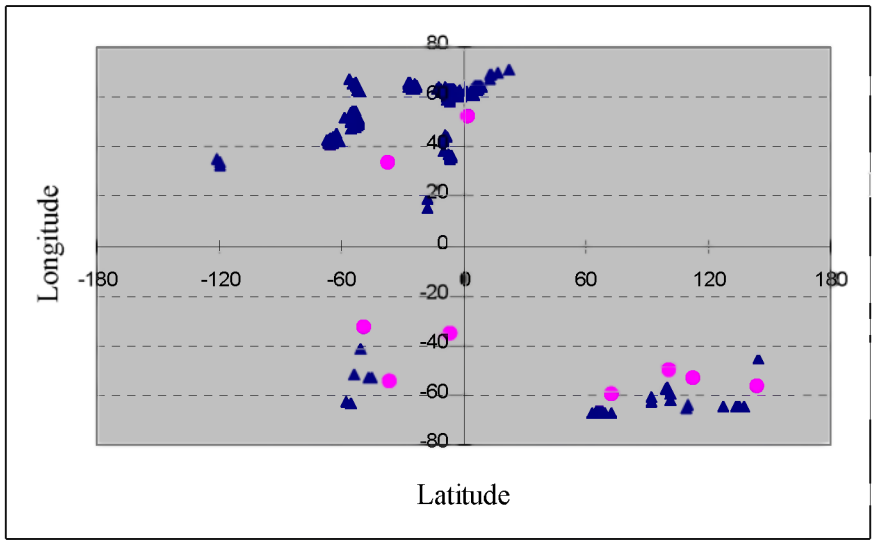

Fig. 3. Result set for Balaenoptera borealis.

Discussion and Future Direction

In this work we have demonstrated the usability of the density-based clustering algorithm –DBSCAN– in identifying geographical outliers. Because sampling is not complete yet, the outliers are not necessarily erroneous data points. Sometimes they are rare sightings or a single specimen in museum collections. Under these circumstances an expert has to examine the outliers and identify the actual erroneous data. On the other hand, the other features in the data space (temperature, salinity, etc…) may have been better sampled and outliers identified in those feature spaces may be more an indicator of data errors. In fact we have performed preliminary studies on temperature space and the results are promising. In the next step, we will develop an incremental learner where outlier detection results obtained from different feature spaces are combined. Domain scientists will play an active and critical role in this learner because:

- they will be prompted with candidates produced by the outlier detection program and select the erroneous data from the candidates
- their decision will be fed back to the learner where the relative weights assigned to individual learners will be readjusted.

Conclusion

The clustering algorithm —DBSCAN— has been successfully applied to identifying geographical outliers in OBIS point data on a species-by-species basis. The algorithm is efficient enough to scan large spatial databases such as OBIS. With more samples coming into OBIS, the outlier detection technique can be used to identify erroneous data points and be part of an operational data quality checking tool where domain knowledge and automatic learners are integrated in a dynamic way.

References

Barnett V. and T. Lewis. 1994. Outliers in Statistical Data. John Wiley & Sons, Chichester, New York. 608p.

Berkhin P. 2002. Survey of Clustering Data Mining Techniques. Accrue Software. http://citeseer.ist.psu.edu/berkhin02survey.html.

Ester M., H.-P Kriegel., J. Sander and X. Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. Portland, OR, 226-231.

Grassle J.F. and K.I. Stocks. 1999. A Global Ocean Biogeographical Information System (OBIS) for the Census of Marine Life. Oceanography 12(3):12-14.

Guha S., R. Rastogi and K. Shim. 1998. Cure: An Efficient Clustering Algorithm for Large Databases. Proc. of ACM SIGMOD Int'l Conf. on Management of Data, 73-84. ACM Press

Hawkins D. 1980. Identification of outliers. Chapman and Hall, London. 188p.

Hinneburg A. and D.A. Kleim. 1998. An Efficient Approach to Clustering in Large Multimedia Databases with Noise, KDD'98, New York, Aug. 1998.

Ng R.T., J. Han. 1994. Efficient and Effective Clustering Methods for Spatial Data Mining 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago, Chile, Proceedings. 144-155.

Sheikholeslami G., S. Chatterjee, A. Zhang. 1999. WaveCluster: A Wavelet Based Clustering Approach for Spatial Data in Very Large Databases. 289-304.

Zhang T., R. Ramakrishnan, M. Livny. 1996. BIRCH: An Efficient Data Clustering Method for Very Large Databases. ACM SIGMOD International conference on Management of Data, Montreal, Canada. 103-114.

Zhang Y. and J.F. Grassle, 2003. A Portal for the Ocean Biogeographic Information System, Oceanologica Acta. 25(5):199-206.