ISSN: 0975-766X
CODEN: IJPTFI
Available Online through *Research Article*
www.ijptonline.com

UNIVARIATE ANALYSIS FOR INFORMATION AGGLOMERATION

M.Sriram[1], Dr.R.M.Suresh[2]
Research Scholar, Department of CSE, Bharath University, Chennai[1]
Professor, Department of Computer Science and Engineering, SLACE, Chennai[2]
Email: msr1sriram@gmail.com

Received on: 15.10.2016 *Accepted on: 22.11.2016*

Abstract

Among the growing range of knowledge mining techniques in varied application areas, outlier detection has gained importance in recent times. Police investigation the objects during a information set with uncommon properties is important; intrinsically outlier objects typically contain helpful info on abnormal behavior of the system or its elements delineated by the information set. Outlier detection has been popularly used for detection of anomalies in pc networks, fraud detection and such applications. although variety of analysis efforts address the matter of police investigation outliers in information sets, there square measure still several challenges faced by the analysis community in terms of characteristic an appropriate technique for addressing specific applications of interest. These challenges square measure primarily attributable to the big volume of high dimensional information related to most data processing applications and conjointly attributable to the performance necessities. during this paper, we have a tendency to applied boxplot- a applied math tool to spot and eliminate the univariate outliers for information agglomeration method.

Keywords: Boxplot, Clustering, data processing, K- Means, Fuzzy C-Means, Outlier Detection, Univariate outliers.

I. Introduction

The recent developments within the field of knowledge mining have result in the outlier detection method mature jointly of the popular data processing tasks [7]. attributable to its significance within the process, outlier detection is additionally referred to as outlier mining. Typically, outliers square measure information objects that square measure considerably totally different from the remainder of the information. Outlier detection or outlier mining refers to the method of characteristic such rare objects during a given information set. though rare objects square measure renowned to be fewer in numbers, their significance is high compared to different objects, creating their detection a crucial task. The final demand of this task is to spot and take away the contaminating result of the far objects on the

information and intrinsically to purify the information for more process. a lot of formally, the outlier detection drawback is outlined as follows: Given a group of knowledge objects, realize a selected range of objects that square measure significantly dissimilar, exceptional and inconsistent with reference to the remaining information. variety of latest techniques are planned recently within the field of knowledge mining to unravel this drawback. during this paper, we have a tendency to applied Boxplot model to spot the univariate outliers. to indicate the impact of outliers within the agglomeration method, K-Means and Fuzzy C-Means algorithms square measure applied. The rest of the paper is organized as follows. The summary of outliers is mentioned in section II. Section III deals with Boxplot model. The summary of agglomeration is mentioned in section IV. The planned work and experimental analysis is explored in section V. Section VI concludes the paper with direction for future analysis work.

II. Outlier

Although there square measure variety of ways for police investigation outliers during a given dataset, no single methodology is found to be the universal selection. looking on the character of target application, totally different applications need use of various detection ways. in keeping with the taxonomy brought get into [8], the outlier detection techniques is generally divided into constant quantity and non-parametric varieties. The sttistics-based ways that assume some model for a given information square measure the constant quantity selection. Typically, the user must model a given information set employing a organisation, and information objects square measure determined to be outliers looking on however they seem in respect to the postulated model. On the opposite hand, most of the non-parametric ways accept some well-defined notion of distance to live the separation between 2 information objects. The non-parametric selection includes distance-based, density-based, and clustering-based ways, conjointly referred to as the information mining ways. A taxonomy of the present outlier detection ways The distance-based ways square measure one among the first techniques that were planned for outlier detection below the information mining selection. so as to beat the issues related to applied math ways, a distance-based methodology was planned in [ ] employing a straightforward and intuitive definition for outliers. in keeping with this methodology, Associate in Nursing object {in a|during a|in Associate in Nursing exceedingly|in a very} information set is an outlier with reference to parameters k and d if no over k objects within the information set square measure at a distance of d or less from that object.

III. Clustering: Clustering (or cluster analysis) aims to prepare a set of knowledge things into clusters, specified things inside a cluster square measure a lot of "similar" to every apart from there to things within the different

clusters[1]. This notion of similarity is expressed in terribly alternative ways, in keeping with the aim of the study, to domain-specific assumptions and to previous data of the matter. agglomeration is sometimes performed once no info is obtainable regarding the membership of knowledge things to predefined categories. For this reason, agglomeration is historically seen as a part of unattended learning. To support the in depth use of agglomeration in pc vision, pattern recognition, info retrieval, data processing, etc., terribly many various ways were developed in many communities.

K-Means algorithmic program

K-means [2] is one among the best unattended learning algorithms that solve the renowned agglomeration drawback. The procedure follows a straightforward and straightforward thanks to classify a given information set through a precise range of clusters (assume k clusters) mounted a priori. the most plan is to outline k centroids, one for every cluster. These centroids ought to be placed during a crafty approach thanks to {different|totally totally different|completely different} location causes different result. So, the higher selection is to position them the maximum amount as doable isolated from one another. successive step is to require every purpose happiness to a given information set and associate it to the closest centre of mass. once no purpose is unfinished, the primary step is completed Associate in Nursingd an early groupage is finished. At this time we'd like to re-calculate k new centroids of the clusters ensuing from the previous step. when we've got these k new centroids, a brand new binding must be done between an equivalent information set points and also the nearest new centre of mass. A loop has been generated. As a results of this loop we have a tendency to might notice that the k centroids modification their location step by step till no a lot of changes square measure done. In different words centroids don't move any longer.

(1) choice of the initial k suggests that for k clusters,

(2) Calculation of the difference between Associate in Nursing object and also the mean of a cluster,

(3) Allocation of Associate in Nursing object to the cluster whose mean is nearest to the item,

(4) Re-calculation of the mean of a cluster from the objects allotted thereto in order that the intra cluster difference is decreased.

Except for the primary operation, the opposite 3 operations square measure repeatedly performed within the algorithmic program till the algorithmic program converges. There exist many variants of the k-means algorithmic program that disagree in choice of the initial k suggests that, difference calculations and techniques to calculate cluster suggests that.

Most k-means kind algorithms are well-tried focused. The k-means algorithmic program has the subsequent vital properties.

1. it's economical in process giant information sets. The procedure complexness of the algorithmic program is O (tk m n), wherever m is that the range of attributes, n is that the range of objects, k is that the range of clusters, and t is that the range of iterations over the full information set. In agglomeration giant information sets the k-means algorithmic program is far quicker than the graded agglomeration algorithms.

2. It typically terminates at an area optimum. to search out out the worldwide optimum, techniques like settled tempering and genetic algorithmic programs is incorporated with the k-means algorithm.

3. It works solely on numeric values as a result of it minimizes a value operate by shrewd the suggests that of clusters.

4. The clusters have protrusive shapes.

V. Planned Work and Experimental Analysis

The objective of the planned work is to analyse the existence of univariate outliers and have with most outliers square measure eliminated. The remaining options square measure then applied to K-Means and Fuzzy C-Means algorithms. The results square measure given in Table I. For experimental purpose Australian dataset from UCI machine learning repository is taken into account.

VI. Conclusion

In this planned work individual attributes of a dataset is analyzed by victimisation applied math tool – boxplot. The options that contain most outlier score square measure thought-about as inapplicable options, and that they square measure eliminated. The agglomeration is performed solely with the relevant options. The experimental result shows that, the planned methodology improves the performance of the agglomeration results. The identification of outliers with totally different criteria desires future analysis.

References

1. Jain, A.K., Murty, M.N., Flynn, P.J.(1999). information Clustering: A Review, (Ed.), ACM Computing Surveys, 264-323.
2. Ahmad A., Dey L., "A K-Means agglomeration algorithmic program for Mixed Numeric and Categorical Data", information and data Engineering, Vol. 63, pp. 503-507, 2007.
3. Kim D., Kwant H.L., Lee D., "A Novel initialisation theme for Fuzzy C-Means algorithmic program for Color Clustering", Pattern Recognition Letters, Vol. 25, pp. 227-237, 2004.

4. Lingras P., Yan R., West C., "Fuzzy C-Means agglomeration of internet Users for Education Sites", In: Advances in computer science, Y. Xiang, B. Chaib draa (Eds.), LNCS, Springer, Vol. 2671, pp. 557-562, June 2003
5. P. Alagambigai, K. Thangavel, N. Karthikeyani Visalakshi, "Improved Visual Cluster Rendering System", in: K. Thangavel.(ed.), Intelligent and Computing Model, Narosa publishing company, New Delhi, pp. 16-23, 2009
6. Alagambigai, P., Thangavel, K., "Feature choice for Visual Clustering", Proceedings of International Conference on Advances in Recent Technologies in Communication and Computing, IEEE pc Society, pp.498-502, 2009
7. Han, J., & Kamber, M. (2000). information mining: ideas and Techniques. Morgan George S. Kaufman Publishers.
8. Ben-Gal, I. (2005). Outlier Detection. In Maimon, O., & Rockack, L (Ed.) data processing and data Discovery Handbook: an entire Guide for Practitioners and Researchers. Kluwer educational Publishers.