International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 3

Outlier Detection using Reverse Neares Neighbor for Unsupervised Data

[1]V. V. R. Manoj, V. Aditya Rama Narayana,
A. Lakshmi Prasanna, A. Bhargavi, Md. Aakhila Bhanu

[1]Assistant Professor,
Dhanekula Institute of Engineering and Technology, Ganguru, Vijayawada, Andhra Pradesh, India

ABSTRACT

Data mining has become one of the most popular and new technology that it has gained a lot of attention in the recent times and with the increase in the popularity and the usage there comes a lot of issues/problems with the usage one of it Outlier detection and maintaining the datasets without the expected patterns. To identify the difference between Outlier and normal behavior we use key assumption techniques. We Provide the reverse nearest neighbor technique. There is a connection between the hubs and antihubs, outliers and the present unsupervised detection methods. With the KNN method it will be possible to identify and influence the outlier and antihub methods on real life datasets and synthetic datasets. So, From this we provide the insight of the Reverse neighbor count on unsupervised outlier detection.

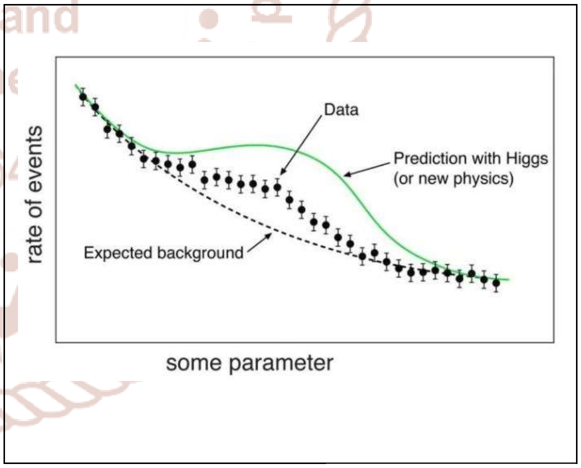
***Keywords**: Reverse nearest neighbor; Outlier detection*

INTRODUCTION:

Outliers are huge values that differentiate from other observations on data; they may showcase difference in measurements and experimental errors. That is an outlier is an observation which separates from an overall pattern. These outliers can be divided into two types. Those are univariate and multivariate.

Univariate outliers can be found in a single feature space having a lot of values. Multivariate outliers can be found in a multi-dimensional space. Identifying the multi-dimensional distributions can be very difficult for the human brain, that is why we need to train a model to do it for us.

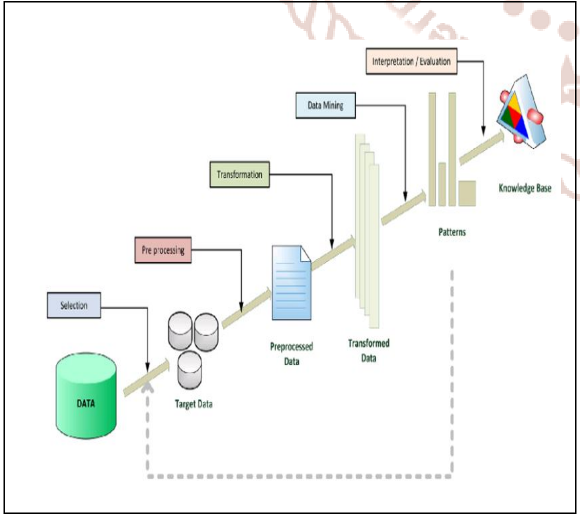
With the decrease in the rate of events against the parameters that are present, the expected background data is very much less compared to the prediction with the higgs theorem.



Detecting outliers can be divided into three different and effective ways. Those ways are supervised, semi-supervised, and unsupervised; the outliers are divided into those categories depending on the labels for outliers. From the above given categories, unsupervised methods are the ones that are mostly used as the other categories require accurate and representative labels that are expensive to obtain. Unsupervised methods also include distance-based methods that depend on a measure of distance or similarity to detect outliers. It is known that with the dimensionalities curse, distance becomes meaningless, that means pair wise distances become impossible to see as dimensionality increases. Distance on unsupervised outlier detection becomes good when linked to the high dimensionalities.

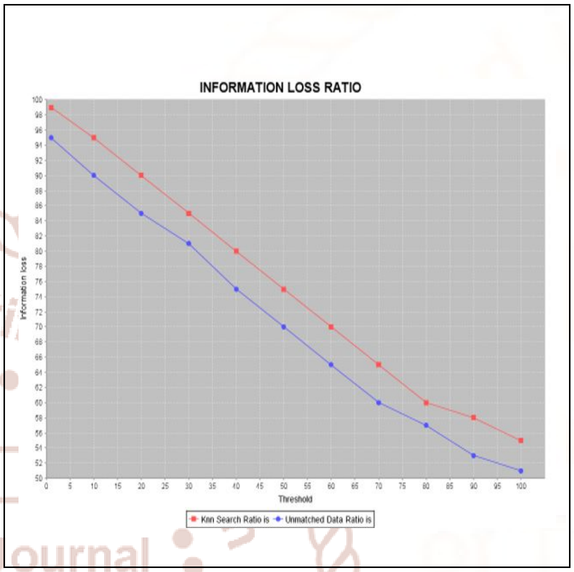
Outlier Detection from Antihubs:
Antihubs search and find the nearest neighbor and outlier can be detected using distance methods and for example KNN identifies the last nearest neighbor.

System Architecture:
System Architecture is a conceptual model that defines the structure, behavior, and more views of a system. An architecture description is a description that represents the system and is organized in a way such that it helps to evaluate the process that is present in the system and helps to understand reasoning about the structures and behavior of system. System Architecture describes that from the data dump we select some specific data that we require which is converted into the target data on which preprocessing is applied. Preprocessing is converting the raw data collected into some understandable format. As the data maybe incomplete or insufficient applying preprocessing technique may help to resolve the issue. After the preprocessed data some transformation is done, and the transformed data is then mined. Now applying the data mining technique to the transformed data, we get the patterns from the data acquired. Evaluating the acquired patterns, we can get the actual and factual data that we require.



CONCLUSION:

With this we like to say that we can calculate the ratio for the change in the normal data to the preprocessed data in a large dimensional dataset. When a data is preprocessed data is formatted it changes from the dump data to make it understandable. With the help of the above methods i.e., distance methods, KNN, and other different methods.



From the above figure we can clearly see that the original data lost the data but with the KNN approach we can see the all the formatted data that can be understandable to everyone and can be read.

REFERENCES:

1. M. Newman and Y. Rinott, "Nearest neighbors and Voronoivolumes in high-dimensional point processes with various distancefunctions," Adv. Appl. Probab., vol. 17, no. 4, pp. 794–809,1985.
2. M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander, "LOF: Identifyingdensity-based local outliers," in Proc. ACM Int. Conf. Manage.Data, 2000, pp. 93–104.
3. E. Achtert, S. Goldhofer, H.-P. Kriegel, E. Schubert, and A. Zimek,"Evaluation of clusterings—metrics and visual support," in Proc.28th Int. Conf. Data Eng., 2012, pp. 1285–1288.
4. E. M€uller, M. Schiffer, and T. Seidl, "Statistical selection of relevantsubspace projections for outlier ranking," in Proc. 27th IEEEInt. Conf. Data Eng., 2011, pp. 434–445.
5. J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders, "TheAmsterdam library of object

images," Int. J. Comput. Vis., vol. 61,no. 1, pp. 103–112, 2005.

6. E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers:Algorithms and applications," VLDB J., vol. 8, nos. 3–4, pp. 237–253, 2000.
7. K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is"nearest neighbor" meaningful?" in Proc. 7th Int. Conf. Database Theory, 1999, pp. 217–235.
8. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprisingbehavior of distance metrics in high dimensional spaces," inProc. 8th Int. Conf. Database Theory, 2001, pp. 420–434.
9. D. Franc¸ois, V. Wertz, and M. Verleysen, "The concentration offractional distances," IEEE Trans. Knowl. Data. Eng., vol. 19, no. 7,pp. 873–886, Jul. 2007.
10. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensionaldata," in Proc. 27th ACM SIGMOD Int. Conf. Manage. Data,2001, pp. 37–46.
11. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervisedoutlier detection in high-dimensional numerical data," Statist. Anal. Data Mining, vol. 5, no. 5, pp. 363–387, 2012.
12. T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Introductionto Algorithms, 3rd ed. Cambridge, MA, USA: MIT Press, 2009.
13. N. Toma_sev and D. Mladeni_c, "Nearest neighbor voting in highdimensional data: Learning from past occurrences," Comput. Sci.Inform. Syst., vol. 9, no. 2, pp. 691–712, 2012.
14. N. Toma_sev, M. Radovanovi_c, D. Mladeni_c, and M. Ivanovi_c,"The role of hubness in clustering high-dimensional data," IEEETrans. Knowl. Data Eng., vol. 26, no. 3, pp. 739–751, Mar. 2014.
15. M. E. Houle, H.-P. Kriegel, P. Kr€oger, E. Schubert, and A. Zimek,"Can shared-neighbor distances defeat the curse of dimensionality?"in Proc 22nd Int. Conf. Sci. Statist. Database Manage., 2010, pp. 482–500.
16. Singh, H. Ferhatosmano_glu, and A. ¸SamanTosun, "Highdimensional reverse nearest neighbor queries," in Proc 12th ACM Conf. Inform. Knowl. Manage., 2003, pp. 91–98.