where $\mu(f(Xi))$ is the mean of the frequency of all patterns of length exactly the same as that of pattern X. The measure of *surprise* of a pattern X is defined as one minus the ratio of the frequency of X over the average frequency of all patterns with same length as X

$$\text{surprise}(X) = 1 - f(X)/\mu(f(Xi)); \ \forall i \text{ such that } |Xi| = |X| \quad (2)$$

A candidate outlier pattern X is an *outlier periodic pattern* iff

$$\text{surprise}(X) > \text{surprisemin AND conf}(X, i\text{st}, i\text{end}, p) > \text{confmin} \quad (3)$$

4. Outlier Periodic Patterns Detection Algorithm

The process can be summarized in the following steps:

- build a suffix tree for the input sequence;
- annotate the suffix tree such that each internal node records the length of the substring it represents (the string obtained by tracing from the root till the node) and the frequency of the substring in the sequence;
- build a pattern frequency table (PFT) for recording the frequency of patterns of different length (up to the maximum pattern length);
- identify the candidate outlier patterns; and
- run STNR for all candidate outlier patterns to output valid periodic outlier patterns.

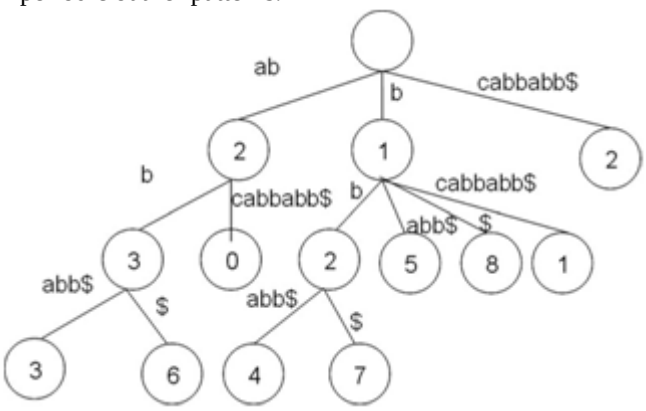

Figure 1: Suffix tree for the string abcabbabb$.

Table 1: PFT for string abcabbabb$

l	***fall*** *:Σ ∀f(Xi), |Xi| = l*	***Count***: *Σ ∀(Xi), |Xi| = l*	***mean***: fall/count
1	8	2	4
2	5	2	2
3	2	1	2

For periodicity detection, we use Algorithm (which includes the basics of STNR) to process the occurrence vectors. Briefly, STNR is a distance-based algorithm where a candidate period is the difference between two consecutive occurrences of a pattern. It traverses the occurrence vector once and records the test periods along with their frequency which keeps on updating as the occurrence vector is traversed. We have only presented the basic STNR algorithm to provide an idea of how STNR works; the complete algorithm (with time tolerance and maximum distance) can be found in [34].Since outlier patterns are expected to be rare and may appear with larger period values with nonstick periodic repetitions, the time tolerance window should also be specified larger than that for frequent periodic patterns. We believe that the time tolerance window concept is more handy when dealing with outlier periodic patterns. Similarly the minimum confidence value should also be set lower than that for the frequent periodic patterns.

Algorithm: Occurrence Vector Processing Algorithm

1: procedure PROCESSOCCURRENCEVECTOR (pattern *X*, list*occur*, int*minSegLen*, real*confmin*)
2: *ppre* = −5, *preCountPerCol*= *periodCol.Countpre*is previous period, *preCountPerCol* is previous countof period collection
3: for *m* = 0; *m*<|*occur*| − 1; *m*++ do
4: if *m*<|*occur*| − 1 then
5: *p* = *occur*[*m* + 1] − *occur*[*m*], *ist*=*occur*[*m*], *iend*= *occur*[|*occur*| − 1]
6: if *ppre* = *p* AND (*iend*+|*X*|−*ist*)>(*minSegLen*∗ |*s*|) AND Not *AlreadyThere*(*X, ist, iend, p*) then
7: *periodCol.add*(*X, ist, iend, p*) Add to testperiod list
8: end if
9: *ppre*= *p*
10: end ifVerify current occurrence against test period list
11: for *n* = *preCountPerCol*; *n* <*periodCol.count*; *n*++ do
12: if (*periodCol*[*n*].*ist*mod *periodCol*[*n*].*p*) ==(*occur*[*m*] mod *periodCol*[*n*].*p*) then
13: Increment period frequency: *periodCol*[*n*].*f*
14: *periodCol*[*n*].*iend*= *occur*[*n*]
15: end if
16: end for
17: end forRemove non-frequent and periods with shorter coverage
18: for *y* = 0, *k* =*preCountPerCol*; *k*<*periodCol.count*; *k* ++ do
19: *fmax*= *periodCol*[*k*].*iend*+1−|*X*|−*periodCol*[*k*].*isperiodCol*[*k*].*p* + 1
20: *conf*(*X, ist, iend, p*) = *ffmax*
21: if *conf*<*confmin*OR (*iend*+ |*X*| − *ist*)>(*minSegLen*∗ |*s*|) then
22: periodCol.remove(*X, ist, iend, p*)
23: end if
24: end for
25: end procedure

5. Use of Median Absolute Deviation

In existing algorithm mean value is used to find out outliers. But for various reasons this method is not efficient, this reasons are explained in the paper

1. Mean

Mean of any given data set is derived as follows:

$$\mu = \sum_{i=1}^{n} \text{yi/n}$$

It is the average value of any given data set. The reasons why it is considered a nonrobust estimator are as follows: 1) Mean value is highly biased even if there is a single outlier and 2) in a large data sets a mean value can be changed even though an outlier is removed. So, while using a mean value for detecting an outlier an outlier can be considered as a

normal data point. This reduces the efficiency of the method and makes it a nonrobust estimator.

2. The Median Absolute Deviation (Mad)

The MAD overcomes these problems. In [34], authors have illustrated the efficiency of MAD over mean and standard deviation with example. Which is given here as follows: The median (M) is, like the mean, a measure of central tendency but offers the advantage of being very insensitive to the presence of outliers. One indicator of this insensitivity is the "breakdown point" [35]. The estimator's breakdown point is the maximum proportion of observations that can be contaminated (i.e., set to infinity) without forcing the estimator to result in a false value (infinite or null in the case of an estimator of scale). For example, when a single observation has an infinite value, the mean of all observations becomes infinite; hence the mean's breakdown point is 0. By contrast, the median value remains unchanged. The median becomes absurd only when more than 50% of the observations are infinite. With a breakdown point of 0.5, the median is the location estimator that has the highest breakdown point. Exactly the same can be said about the Median Absolute Deviation as an estimator of scale (see the formula below for a definition). Moreover, the MAD is totally immune to the sample size. These two properties have led to describe the MAD as the "single most useful ancillary estimate of scale". It is for example more robust than the classical inter quartile range [37], which has a breakdown point of 25% only. To calculate the median, observation has to be sorted in ascending order. Let us consider the previous statistical series: 1, 3, 3, 6, 8, 10, 10, and 1000. The average rank can be calculated as equal to (n + 1) / 2 (i.e., 4.5 in our example). The median is therefore between the fourth and the fifth value, that is, between six and eight (i.e., seven). Calculating the MAD involves finding the median of absolute deviations from the median. the MAD is defined as follows [36]

$$MAD=b\ M_i(|x_i-m_j(x_j)|)$$

where the xj is the n original observations and Mi is the median of the series. Usually, b = 1.4826, a constant linked to the assumption of normality of the data, disregarding the abnormality induced by outliers (Rousseeuw & Croux, 1993). Calculating the MAD implies the following steps: (a) the series in which the median is subtracted of each observation becomes the series of absolute values of (1–7), (3–7), (3–7), (6–7), (8–7), (10–7), (10–7), and (1000–7), that is, 6, 4, 4, 1, 1, 3, 3, and 993; (b) when ranked, we obtain: 1, 1, 3, 3, 4, 4, 6, and 993; (c) and (d) the median equals 3.5 and will be multiplied by 1.4826 to find a MAD of 5.1891. To calculate MAD all the observations has to be sorted first. This can be a huge overhead in large data set. It requires preprocessing and it can become time consuming and in highly dynamic data it may become more difficult to hold a correct value.

6. Working

The algorithm is applied on the time series data set. The data values in data set are used to calculate MAD values. This calculated MAD value is used to determine surprising values from the given data set by comparing them with the MAD. Values which are 3 times away from the Median values are considered as Outliers. Periodicity of detected outliers are calculated by above mentioned Periodicity formula in Section IV

7. Conclusion

With this definition, it can also identify outlier patterns that may involve some (or all) frequent events, as it check the repetitions of combination of events and not just the individual events. The experimental results show that the proposedalgorithm consistently outperforms the existing approach InfoMiner. Additionally a novel algorithm for the periodicity detection of outlier, surprising, or unusual patterns is shown. It makes use of the MAD value to compare relative frequency of the outlier pattern instead of mean value which was previously used in the existing algorithm. As the mean method is not robust and do not give the accurate results. It can easily be affected with the presence of outlier. A new measure known as Median Absolute Deviation is used to detect outlier instead of mean, as it is more efficient compare to mean. It increases the accuracy of the existing algorithm. In the carried out experiments outliers detected by MAD are more accurate.

8. Acknowledgment

I take this opportunity to thank our guide "All staff" for placing this idea in my mind and giving marvelous suggestions from the platform of this project.I would not have been completed without the valuable guidance and encouragement of Prof. Mr. Kureshi, Prof. Mr. Kshirsagar, Mrs. Joshi madam, Prof. Mr. Jaypal, Prof. Mr. Prabhudeva, Prof. Mr. Natikar Sir and all my classmates. Finally, I would also like to thank my husband Mr. Prasad Kadam and my family members to their support.

References

[1] Pang-Ninh Tan (2006) Knowledge Discovery from Sensor Data. Sensors.
[2] Pei Sun (2006) Outlier detection in high dimensional, spatial and sequential data sets. Doctoral Thesis
[3] V. J. Hodge, J. Austin (2003) A survey of outlier detection methodologies. ArtificialIntelligence Review, vol. 22, pp 85-126
[4] D.M. Hawkins (1980) Identification of outliers. Chapman and Hall, Reading, London
[5] V. Barnett and T. Lewis (1994) Outliers in statistical data. John Wiley Sons, Reading,New York
[6] J. Han and M. Kamber (2001) Data mining: concepts and techniques. Morgan Kaufmann
[7] R. J. Bolton, D. J. Hand (2001) Unsupervised proofing methods for fraud detection. InProceedings of CSCC
[8] D.J. Marchette (2001) Computer intrusion detection and network monitoring: a statisticalviewpoint. New York: Springer
[9] G. M. Davis, K. B. Ensor (2006) Outlier detection in environmental monitoring networkdata: an application to ambient ozone measurements for Houston, Texas.

Journal of Statistical Computation and Simulation, vol. 76, no. 5, pp 407-422
[10] J. Lin, A. E. Fu and H. V. Herle (2005) Approximations to magic: Finding unusual medicaltime series. In: Proceedings of Symposium on Computer-Based Medical systems. Washington,DC, USA, pp 329-334
[11] W. Du, L. Fang and P. Ning (2005) LAD: localization anomaly detection for wirelesssensor networks. In: Proceedings of Parallel and Distributed Processing Symposium
[12] M. G. Elfeky, W. G. Aref, and A. K. Elmagarmid, "Periodicity detectionin time series databases," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 7,pp. 875–887, Jul. 2005.
[13] M. G. Elfeky,W. G. Aref, and A. K. Elmagarmid, "WARP: Time warpingfor periodicity detection," in *Proc. IEEE Int. Conf. Data Mining*, Nov.2005, pp. 8–15.
[14] J. Han, W. Gong, and Y. Yin, "Mining segment-wise periodic patternsin time related databases," in *Proc. ACM Int. Conf. Knowl. Discov. Data Mining*, vol. 8, no. 1, pp. 53–87, Aug. 1998.
[15] P. Indyk, N. Koudas, and S. Muthukrishnan, "Identifying representativetrends in massive time series data sets using sketches," in *Proc. Int. Conf.Very Large Data*, Sep. 2000, pp. 363–372.
[16] S. Ma and J. Hellerstein, "Mining partially periodic event patterns withunknown periods," in *Proc. IEEE Int. Conf.DataEng.*,Apr. 2001, pp. 205–214.
[17] J.Yang,W.Wang, and P.Yu, "InfoMiner+: Mining partial periodic patternswith gap penalties," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2002,pp. 725–728.
[18] C. Berberidis, W. Aref, M. Atallah, I. Vlahavas, and A Elmagarmid,"Multiple and partial periodicity mining in time series databases," in *Proc. Eur. Conf. Artif. Intell.*, Jul. 2002, pp. 79–94.
[19] C. Sheng,W. Hsu, andM.-L. Lee, "Mining dense periodic patterns in timeseries data," in *Proc. IEEE Int. Conf. Data Eng.*, 2005, p. 115.
[20] C. Sheng, W. Hsu, and M.-L. Lee, "Efficient mining of dense periodicpatterns in time series," Nat. Univ. Singapore, Singapore, Tech. Rep. 1,2005.
[21] J. Han, Y. Yin, and G. Dong, "Efficient mining of partial periodic patternsin time series database," in *Proc.IEEE Int. Conf.Data Eng.*, 1999, pp. 106–115.
[22] K.-Y. Huang and C.-H. Chang, "SMCA: Ageneral model for mining asynchronousperiodic patterns in temporal databases," *IEEE Trans. KnowlData Eng.*, vol. 17, no. 6, pp. 774–785, Jun. 2005.
[23] D. Gusfield, *Algorithms on Strings, Trees, and Sequences*. Cambridge,U.K.: Cambridge Univ. Press, 1997.
[24] R. Grossi and G. F. Italiano, "Suffix trees and their applications in stringalgorithms," in *Proc. South Amer. Workshop String Process.*, Sep. 1993,pp. 57–76.
[25] M. Dubiner, Z. Galil, and E. Magen, "Faster tree pattern matching," *J.ACM*, vol. 14, pp. 205–213, 1994.
[26] K. Huarng and T. H.-K.Yu, "Ratio-based lengths of intervals to improvefuzzy time series forecasting," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*,vol. 36, no. 2, pp. 328–340, Apr. 2006.
[27] K. Huarng, T. H.-K. Yu, and Y. W. Hsu, "A multivariate heuristic modelfor fuzzy time-series forecasting," *IEEE Trans. Syst., Man, Cybern. B,Cybern.*, vol. 37, no. 4, pp. 836–846, Aug. 2007.
[28] E. Keogh, S. Lonardi, and B. Y.-C.Chiu, "Finding surprising patterns in atime series database in linear time and space," in *Proc. 8th ACM SIGKDDInt. Conf. Knowl.Discov. Data Mining*, 2002, pp. 550–556.
[29] J. Yang, W. Wang, and P. S. Yu, "Infominer: Mining surprising periodicpatterns," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discov. DataMining*, 2001, pp. 395–400.
[30] J. Yang, W. Wang, and P. S. Yu, "STAMP: On discovery of statisticallyimportant pattern repeats in long sequential data," in *Proc. SIAM Int. Conf.Data Mining*, 2003, pp. 224–238.
[31] C. Shahabi, X. Tian, andW. Zhao, "Tsa-tree: A wavelet-based approach toimprove the efficiency of multi-level surprise and trend queries," in *Proc.12th Int. Conf. Sci. Statist. Database Manage.*, 2000, pp. 55–68.
[32] S.-T. Li andY.-C.Cheng, "A stochastic HMM-based forecastingmodel forfuzzy time series," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40,no. 5, pp. 1255–1266, Oct. 2010.
[33] J. Sullivan and W. H. Woodall, "A comparison of fuzzy forecasting andMarkov modeling," *Fuzzy Sets Syst.*, vol. 64, no. 3, pp. 279–293, 1994.
[34] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, Laurent Licata, Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median *Journal of Experimental Social Psychology*, *Volume 49, Issue 4, July 2013, Pages 764-766*
[35] Donoho, D. L., & Huber, P. J. (1983). In Bickel, Doksum, & Hodges (Eds.), The notion of breakdown point. California: Wadsworth
[36] Huber, P. J. (1981). Robust statistics. New York: John Wiley
[37] Rousseeuw, P. J., &Croux, C. (1993). Alternatives to the median absolute deviation.Journal of the American Statistical Association, 88(424), 1273–1283

Author Profile

Prof. Mrs. Sulochana Garage-Kadam Completed BE[IT] from P.D.V.V.P.COE Ahmednagar in 2012 Student Of ME II Computer engineering V..A.C.O.E. Ahmednagar. She is working as lecturer in Ashok institute of Engineering And Polytechnic Ashok nagar.