2009 International Conference on Computer Engineering and Applications
IPCSIT vol.2 (2011) © (2011) IACSIT Press, Singapore

Framework on Outlier Sequential patterns for Outbreak Detection

Zalizah Awang Long [1], Abdul Razak Hamdan [1] and Azuraliza Abu Bakar [1 +]

[1] Faculty Information Science and Technology
Universiti Kebangsaan Malaysia

Abstract. There are many outbreak detection that available with various techniques being introduced ranging from statistic to data mining including machine learning. With the direction of spatial-temporal data the research under public health surveillance especially outbreak detection or anomalies detection are promising research. In this paper we applied data mining techniques in detecting outbreak in public health surveillance. The phase involves learning, detecting and repository. An extracted sequential pattern method, outlier set was identified using outlier detection algorithm methods.

Keywords: data mining, surveillance, outlier, sequence pattern.

1.0 Introduction

The main objective of a health surveillance system is to reduce the impact of an outbreak by enabling officials to detect it quickly and implement timely, appropriate interventions. Identifying an outbreak days to weeks earlier than traditional surveillance will result in a reduction in morbidity, mortality, and its economic consequences. This is likely obtainable by improvements in data collection and associate analysis. Ideas express by Stoto in 'Syndromic surveillance: Is it Worth the Effort?' as to be useful for early detection of natural disease outbreaks [80]. With a spectrum of technologies from statistics, computer science and data mining that can help, there are numbers of analysis methods and informal reasoning for the early detection such that being uses in the Biosurveillance system. [56]

The field of surveillance, generally on biomedical informatics has drawn increasing popularity and attention, and has been growing rapidly over the past two decades. In particular, knowledge management, data mining, and text mining techniques have been adopted in various successful biomedical applications in recent years [60]. Paradigms for data mining or machine learning and data analysis including: probabilistic and statistical models, symbolic learning and rule induction, neural networks, evolution-based algorithms, and analytic learning and fuzzy logic to be merging into health informatics focusing on health surveillance.

This paper, which is divided into three main sections, presents sequence outlier detection based on data mining theory and surveillance perspective applied into public health domain. The first section discusses the basic concept and theory of data mining sequential pattern and outlier detection. The second section presents the framework of clustering-based outlier detection according to the outlier set produced by sequence pattern generated. The third section discusses the potential of propose techniques for outbreak detection.

2. Data mining

Data mining or well known as Knowledge Discovery in Database [28], [62], [17] has been evolve and intensive research focusing in various application domains. The finding in data mining research has motivate creation of new techniques to analyze, understand and visualize large amount of data that gathered from scientific, business and surveillance (e.g. network , medical records, etc.). Data mining involve the semiautomatic discovery of interesting knowledge, such as patterns, associations, changes, anomalies and significance structures from various kinds of database into information repositories [54].

+ Corresponding author. Tel.: + 603- 89216087; fax: +603-89256732.
E-mail address: zalizah@miit.unikl.edu.my; zalizahnmk@yahoo.com

2.1 Sequential Mining

Based on general classes proposed by [29] and [63], the views of sequential mining either apriori-like and pattern growth, this include the mining close sequential patterns. The main goal of mining sequential pattern is to detect patterns in database comprised of sequence of sets. Normally sequential pattern mining required support model from the complete set of frequent subsequence in the set of sequences. Other model is multiple alignment models, uses clustering as a preprocessing step to group similar sequences, and mines the underlying consensus pattern in each cluster directly through multiple alignments [45]. Generalized Pattern Mining (GSP) is the extension of apriori algorithm. GSP uses the horizontal data format [1], [2] and later come Sequential Pattern Discovery using equivalent classes (SPADE) uses vertical data format [88]. Comprehensive comparison studies of pattern growth-based by [29] shows that Prefixspan outperforms the GSP algorithm, Freespan and SPADE with integrated pseudo-projection in term of speeds. While Clopan seen as improvement efficiency over Prefixspan. The study was further explored by [63] by impose the multi constraint-based instead of mono constraint in the Prefixspan.

2.2 Outlier

Outlier detection focusing on finding some interesting patterns that out of norm. According to [16], anomalous patterns also referred as outlier, anomalies, discordant observations, exceptions, faults, defects aberrations, noise, errors, damage, surprise, novelty, peculiarities or contaminants in various applications domain. In such for public health, outlier detection being widely used to detect anomalous patterns in patient records which could be symptoms or disease.

Outlier detection had being used in broad domain and various approaches (outlier detection, novelty detection, anomaly detection, noise detection, variation detection or exception mining) [32].The exact definition of an outlier depends on hidden assumptions regarding the data structure and the applied detection method [6]. As quoted in [32], [6] definition derived from Bennett & Lewis (1994) that an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. Indicated in study outlier normally being considered as noise, and recently under data mining approach outlier are consider as important task to drill out important information. One of the steps towards obtaining a coherent analysis is detection of outlying observations [6].

Table 1 below listed the outlier mining being applied into various domains. The main interests are to look at the outlier being implemented in the medical and public health data. Found that limited number of research being explore in using outlier mining into surveillance. [21], [22] and [77] actively research conducted on outlier or namely as anomaly detection being applied into surveillance.

Table 1: Outlier Application

TECHNIQUES								DOMAIN	REFERENCES
A	B	C	D	E	F	G	H		
√		√	√	√	√			Intrusion detection	[25], [24], [73],[26], [36], [48]
			√		√	√		Fraud detection	[4], [8], [7]
	√		√		√			Industrial detection	[39], [50], [87]
				√		√	√	Image processing	[79], [64]
								Medical & health	[35], [46], [78], [69], [14], [40], [85], [77], [21] and [22]

Note : AStatistical profiling using histogram ,B-Parametric @ non parametric statistical modeling C-Markov models D-Neural Networks E-Support vector machine F-Rule-based systems G-Clustering, H-Nearest Neighbor approach

2.3 Public Health Surveillance

Epidemiologists refer to surveillance as the systematic collection, analysis, and interpretation of health data about a clinical syndrome that has a significant impact on public health. By definition surveillance can be define in general ways as continuous reporting system, collection, analysis and dissemination of information to authorities for further actions. Vary definitions define by [15], [10], [68], [67], [51] mostly highlighted definition of surveillance shall consists of reporting, analysis and dissemination for decision making. Reported in systematic review [9] reported 115 surveillance system including 9 syndromic, 13 collecting ILI, 23 laboratory. Some of the systems are expanded to include both early detection and situational awareness.

Routine surveillance depends on passive methods. Outbreaks need for active surveillance. Early detection can have a major impact in reducing the numbers of cases and deaths during the outbreaks. The amount of needed data for each outbreak varies with disease and the number of cases. In explosive outbreak with large numbers of cases there will be no time to collect detail information - more priority is to collect numbers of

cases. For outbreaks that are smaller in size to evolve slowly - require case investigation to obtain information [18].

The analysis of the outbreak detection algorithm methods can be divided according to disease parameter based on time, space, time-space and space-time-person [83] and according to weighted analysis such as time weight analysis, temporal-spatial, spatial clustering and data mining as in table 2. Regardless the analysis based on disease or weighted analysis; the analytic methods are often data source dependent [49].

The emerging disciplines of surveillance enable the real-time monitoring of pre-diagnosis information for the first signs of any disease outbreaks attack, with the data mining analysis techniques there are numbers of analysis methods and informal reasoning for early detection of disease outbreaks.

Table 2: detection methods

ANALYSIS METHODS	CATEGORY	TECHNIQUES
Time/Time weighted	Inductive	CuSum &EWMA ([59], [75], [53], [37], [82], [13]); Hidden Markov ([47], [65]);Time series [66]; Linear mix [41]; Negative binomial [30]; Bayesian ([52], [70]; Neural Network [27]; Case based [72]
Space/spatial clustering	Inductive	Spatial scan statistic [12]; Space-time scan statistic [57]
Time-space/Temporal-spatial	Deductive/ Hybrid	ST clustering [42]; ST permutation scan statistic [44]; M-statistic ,GAM, SSS [60]);ST scan statistic [71]; ARIMA(s) [76]; SS square grid [31]; Co-linearity [55] Farrington algorithm [33]; GLMM s [38]; Bayesian ([19],[58],[5],[20],[74]; Rule-based [86]; Concept based [11]; Ontology & knowledge based [23]

3. Framework

Our proposed recommendation mechanism is composed of three phases, as shown in figure 1. Namely the phase involves Phase I as learning phase, Phase II detecting phase and Phase III as repository. The first phase is involves extracting association rules from the potential hospital Emergency Department (ED) data source. In this case we propose data generator to generate data and impose outbreak into the candidate data for the outlier detection. Through the pattern mining on the ED data we are expecting to find hidden informative relationships between the attribute compose in the ED patient data. At this point we consider the patient records as potential transaction data based on the daily records. Detecting outlier, based on outlier detection module will determine whether normal or abnormal (in this case abnormal considered as outbreak) against normal sequence patterns that learnt. The confident evaluator will evaluate the sequence generated and outlier detection including similarity function and interestingness. Generally it consist of three main steps, The first will learn the synthetic data based on data generator, generated sequence pattern then the final step is to detect abnormal behavior to determine the outbreak.

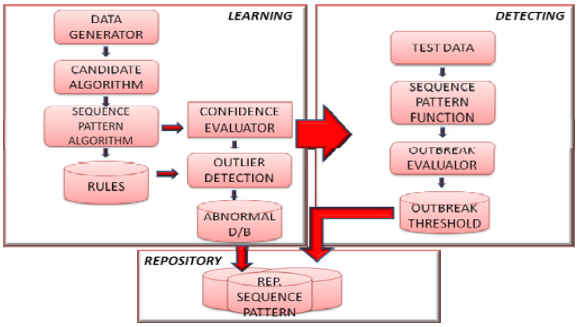

Figure 1 : Outbreak sequential Outlier Detection (OSO) Framework

4. Conclusion

In this paper, we propose the techniques for outbreak detection based on sequential mining and outlier detection. The framework developed to shown the concept of data mining being introduced in new spectrum in surveillance by incorporating both data mining tasks association and outlier into the outbreak detection for public health. There is no result in this paper; we leave it for future works.

References

1. Aggarwal, C. C. (2005). *On Abnormality Detection in Spuriously Populated Data Streams.* 5th SIAM Data Mining
2. Agrawal, R., & Srikant, R. (1994). *Fast algorithms for mining association rules*.

3. Agrawal, R., & Srikant, R. (1995). *Mining sequential patterns*.
4. Aleskerov, E., Freisleben, et.al. (1997). *Neural Network database mining system for credit card fraud detection*.
5. Banks, D., el. (2007). Bayesian Methods for Syndromic Surveillance. *Advances in Disease Surveillance, 2*(2), 41
6. Ben-Gal, I. (2005). Outlier Detection. *DM and KD Handbook Kluwer Academic Publishers*
7. Bolton, R. J., & Hand, D. J. (2001). *Unsupervised profiling methods for fraud detection*.
8. Brause, R., Langsdorf, T., & Hepp, M. (1999). *Neural data mining for credit card fraud detection.*
9. Bravata, D. M., McDonald, K. M., Smith, et al. (2004). Systematic Review: Surveillance Systems for Early Detection of Bioterrorism-Related Diseases. *Annals of Internal Medicine, 140*(11), 910.
10. Breslow, E. L. (2002). Information Systems." Encyclopedia of Public Health".
11. Buckeridge, D. L., Graham, J., Noy, N. F., Shahar, Y., & Henry, K. A. (2002). A Knowledge-based Approach to Temporal Abstraction of Clinical Data for Disease Surveillance
12. Burkom, H. S. (2003). Biosurveillance Applying Scan Statistics with Multiple, Disparate Data Sources. *Journal of Urban Health: Bulletin of the New York Academy of Medicine, 80*(1), i57-i65
13. Burr, T., Graves, T., Klamann, R., Michalak, S., Picard, et.al (2006). Accounting for seasonal patterns in syndromic surveillance data for outbreak detection. *BMC Medical Informatics and Decision Making, 6*(1), 40.
14. Campbell, C., & Bennett, K. P. (2001). A Linear Programming Approach to Novelty Detection. Advance in Neural Information Processing Systems, 395-401
15. CDC. (2007). Key facts about influenza and influenza vaccine. 2007.
16. Chandola, V., Banerjee, et.al (2007). *Outlier detection: a survey*: Technical Report. University of Minnesota, USA
17. Chen, J., Jin, H., He, H., O'Keefe, C. M., Sparks, R., Williams, G., et al. (2006). Frequency-based Rare Events Mining in Administrative Health Data. *Electronic Journal of Health Informatics, 1*(1).
18. Connolly, M. A. (2005). Communicable disease control in emergencies: a field manual. Geneva: WHO
19. Cooper, G. F., Dash, D. H., Levander, J. D et.al (2004). *Bayesian biosurveillance of disease outbreaks*
20. Cooper, G. F., Dowling, et.al. (2007). A Bayesian Algorithm for Detecting CDC Category a Outbreak Disease from Emergency Department Chief Complaints. *Advances in Disease Surveillance, 2*(2), 45.
21. Das, K., & Schneider, J. (2007). *Detecting anomalous records in categorical datasets.* KDD, San Jose, California,
22. Das, K., Schneider, J., & Neill, D. B. (2008). Anomaly pattern detection in categorical datasets. *KDD*
23. Doyle, T. J., Ma, H., Groseclose, S. L., & Hopkins, R. S. (2005). PHSkb: A knowledgebase to support notifiable disease surveillance. *BMC Medical Informatics and Decision Making, 5*(1), 27.
24. Eskin, E., Lee, W.,et.al (2001). *Modeling System Calls for Intrusion Detection with Dynamic Window Sizes*.
25. Forrest, S., Esponda, F., & Helman, P. (2004). A Formal Framework for Positive and Negative Detection Schemes. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 34*(1), 357-373.
26. Gao, B., Ma, H. Y., & Yang, Y. H. (2002). *HMMs based on anomaly intrusion detection method*.
27. Guthrie, et.al (2005). *Detection of disease outbreaks in pharmaceutical sales: NN and threshold algorithms*.
28. Han, J., & Kamber, M. (2001). *Data Mining: Concepts and Techniques*: Morgan Kaufmann.
29. Han, J., Pei, J., & Yan, X. (2005). Sequential Pattern Mining by Pattern-Growth: Principles and Extensions. *Studies Fuzziness and Soft Computing, 180*, 183.
30. Held, L., et.al (2006). A two-component model for counts of infectious diseases. *Biostatistics, 7*(3), 422-43
31. Higgs, B. W., Mohtashemi, M., et.al. (2007). Early Detection of Tuberculosis Outbreaks among the San Francisco Homeless: Trade-Offs Between Spatial Resolution and Temporal Scale. *PLoS ONE, 2*(12).
32. Hodge, V., & Austin, J. (2004). A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review, 22*(2)
33. Höhle, M., Paul,(2007). Statistical approaches to the surveillance of infectious diseases for veterinary public health.
34. Horn, P. S., Feng, L., Li, Y., & Pesce, A. J. (2001). Effect of Outliers and Nonhealthy Individuals on Reference Interval Estimation. *Clinical Chemistry, 47*(12), 2137-2145.
35. Hu, W., Liao, Y., & Vemuri, V. R. (2003). Robust Anomaly Detection Using Support Vector Machines. *IEEE Trans on Pattern Analysis and Machine Intelligence*
36. Hutwagner, L., Browne, T., Seeman, G. M., & Fleischauer, A. T. (2005). Comparing aberration detection methods with simulated data. *Emerging Infectious Diseases, 11*(2), 314-316.
37. Johnson, G. D. (2008). Prospective spatial prediction of infectious disease: experience of New York State (USA) with West Nile Virus. *Environmental and Ecological Statistics, 15*(3), 293-311
38. Keogh, E., & Lonardi, S. (2002). *Finding Surprising Patterns in a Time Series Database in Linear Time and Space.* Paper presented at the in proceedings of the eight ACM SIGKDD, New York, USA
39. Keogh, J. L. E., Fu, A., & Van Herle, H. (2005). Approximations to Magic: Finding Unusual Medical Time Series
40. Kleinman, K., Lazarus, R., & Platt, R. (2004). A Generalized Linear Mixed Models Approach for Detecting Incident Clusters of Disease in Small Areas. *American Journal of Epidemiology, 159*(3), 217.
41. Koch, M. W., et.al. (2001). Near-Real Time Surveillance against Bioterror Attack using Space-Time Clustering
42. Kulldorff, M., Heffernan, R., Hartman, J., Assuncao, R., & Mostashari, F. (2005). A Space-Time Permutation Scan Statistic for Disease Outbreak Detection. *PLOS MEDICINE, 2*(3), 216
43. Kum, H. C., Chang, J. H., & Wang, W. (2007). Benchmarking the effectiveness of sequential pattern mining methods. *Data & Knowledge Engineering, 60*(1), 30-50.
44. Laurikkala, J., Juhola, M., & Kentala, E. (2000). *Informal identification of outliers in medical data.* Fifth International Workshop on Inteligent Data Analysis in Medicine and Pharmacology

45. Le Strat, Y., & Carrat, F. (1999). Monitoring epidemiologic surveillance data using hidden Markov models. *Statistics in Medicine, 18*(24), 3463-3478
46. Lee, W., Stolfo, S. J., & Mok, K. W. (2000). Adaptive Intrusion Detection: A Data Mining Approach. *Artificial Intelligence Review, 14*(6), 533-567.
47. Li, C. S. (2005). Survey of Early Warning for Environmental and Public Health Applications
48. Li, Y., Pont, et.al(2002). Improving the performance of radial basis function classifiers in condition monitoring and fault diagnosis applications where unknown' faults may occur. *Pattern Recognition Letters, 23*(5), 569-577.
49. Liebert, M. A. (2003.). Syndromic Surveillance: A Case of Skillful Investment. Biosecurity and Bioterrorism *Biodefense Strategy, Practice and Science 1* (4)
50. Lombardo, J., Burkom, H., Elbert, et.al(2003). A Systems Overview of the Electronic Surveillance System for the Early Notification of Community-Based Epidemics *Journal of Urban Health, 80(1)* . 32-42.
51. Lumley, T., Sebestyen, K., Lober, W. B., & Painter, I. (2005). An Open Source Environment for The Statistical Evaluation Of Outbreak Detection Methods. *AMIA Annual Symposium Proceedings, 2005*, 1037.
52. Mohammadian, M. (2004). *Intelligent Agents for Data Mining and Information Retrieval*: Idea Group Publishing.
53. Mohtashemi, M., Yih, K., & Kleinman, K. (2007). Multi-Syndrome Analysis of Time Series: A new concept for outbreak investigation. *Advances in Disease Surveillance, 2*(2), 59.
54. Moore, A., Cooper, G., Tsui, R., & Wagner, M. (2002). Summary of Biosurveillance-relevant statistical and data mining technologies}. *Unpublished Internet Report, Feb*.
55. Neill, D. B., et.al. (2005). Anomalous Spatial Cluster Detection. *Data Mining Methods for Anomaly Detection*.
56. Neill, et.al(2006). A Bayesian Spatial Scan Statistic. Advances in neural information processing system*, 18*, 1003
57. O'Brien, S. J., & Christie, P. (1997). Do CuSums have a role in routine communicable disease surveillance? *Public Health, 111*(4), 255-258
58. Ozonoff, A., Webster, T., Vieira, V., et.al(2005). Cluster detection methods applied to the Upper Cape Cod cancer data. *Environmental Health: A Global Access Science Source, 4*(1), 19.
59. Pei, J., Han, J., & Wang, W. (2007). Constraint-based sequential pattern mining: the pattern-growth methods. *Journal of Intelligent Information Systems, 28*(2), 133-160.
60. Pokrajac, D., Lazarevic, A., & Latecki, L. J. (2007). *Incremental Local Outlier Detection for Data Streams*.
61. Rath, T. M., Carreras, M., & Sebastiani, P. (2003). Automated Detection of Influenza Epidemics with Hidden Markov Models. Lecture notes in computer science, 521-532
62. Reis, B. Y., & Mandl, K. D. (2003). *Integrating Syndromic Surveillance Data across Multiple Locations.*
63. Ritzwoller, D. P., Bridges, C. B., Shetterly, S., et.al. (2005). Effectiveness of the 2003-2004 Influenza Vaccine Among Children 6 Months to 8 Years of Age, With 1 vs 2 Doses. *Pediatrics, 116*(1), 153-159.
64. Ritzwoller, D. P., Kleinman,et al. (2005). Comparison of syndromic surveillance and a sentinel provider system in detecting an influenza outbreak-Denver, Colorado, 2003. *Morbidity and Mortality Weekly Report, 54*- 151-156
65. Roberts, S. J. (2000). *Extreme value statistics for novelty detection in biomedical dataprocessing*
66. Rolka, H., Burkom, H., Cooper, et.al (2007). Issues in applied statistics for public health bioterrorism surveillance using multiple data streams: research needs. *Statistics in Medicine, 26*(8), 1834
67. Sabhnani, M., Neill, D., Moore, A et.al. (2005). *Efficient Analytics for Effective Monitoring of Biomedical Security.*
68. Schmidt, R., & Gierl, L. (2002). Case-Based Reasoning for Prognosis of Threatening Influenza Waves. Lecture Notes in Computer Science, 99-108.
69. Sekar, R., Gupta, A. et al. (2002). *Specification-based detection: a new approach for detecting network intrusions*.
70. Shen, Y., & Cooper, G. F. (2007). A Bayesian Biosurveillance Method That Models Unknown Outbreak Diseases. Lecture Notes in Computer Science*, 4506*, 209
71. Shmueli, G. (2005). Current and Potential Statistical Methods for Anomaly Detection in Modern Time Series Data: The Case of Biosurveillance. *Data Mining Methods for Anomaly Detection*.
72. Shtatland, E., Kleinman, et.al(2006).Biosurveillance and outbreak detection using the arima and logistic procedures
73. Singliar, T., & Dash, D. H. (2007). COD: Online Temporal Clustering for Outbreak Detection. *22*, 633
74. Solberg, H. E., & Lahti, A. (2005). Detection of Outliers in Reference Distributions: Performance of Horn's Algorithm. *Clinical Chemistry*
75. Spence, C., Parra, L., & Sajda, P. (2001). *Detection, Synthesis and Compression in Mammographic Image Analysis with a Hierarchical Image Probability Model*
76. Stoto, M. A., Schonlau, M., & Mariano, L. T. (2004). Syndromic surveillance: is it worth the effort. *Chance, 17*(1),
77. Suzuki, E., Watanabe, T., Yokoi, H., & Takabayashi, K. (2003). *Detecting Interesting Exceptions from Medical Test Data with Visual Summarization*.
78. Watkins, R. E., Eagleson, S., et.al. (2008). Applying cusum-based methods for the detection of outbreaks of Ross River virus disease in Western Australia. *BMC Medical Informatics and Decision Making, 8*(1), 37.
79. Wong, W. K. (2004). Data mining for early disease outbreak detection
80. Wong, W. K., Moore, A., Cooper, G., et.al. (2005). What's Strange About Recent Events (WSARE): An Algorithm for the Early Detection of Disease Outbreaks. *The Journal of Machine Learning Research, 6*, 1961-1998.
81. Wong, W., Moore, A., Cooper, G., & Wagner, M. (2003). *Bayesian Network Anomaly Pattern Detection for Disease Outbreaks*
82. Zaki, M. J. (2001). SPADE: An Efficient Algorithm for Mining Frequent Sequences. *Machine Learning, 42*(1), 31-6