International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:7, No:3, 2013

Anomaly Based On Frequent-Outlier for Outbreak Detection in Public Health Surveillance

Zalizah Awang Long, Abdul Razak Hamdan, and Azuraliza Abu Bakar

***Abstract*—**Public health surveillance system focuses on outbreak detection and data sources used. Variation or aberration in the frequency distribution of health data, compared to historical data is often used to detect outbreaks. It is important that new techniques be developed to improve the detection rate, thereby reducing wastage of resources in public health. Thus, the objective is to developed technique by applying frequent mining and outlier mining techniques in outbreak detection. 14 datasets from the UCI were tested on the proposed technique. The performance of the effectiveness for each technique was measured by t-test. The overall performance shows that DTK can be used to detect outlier within frequent dataset. In conclusion the outbreak detection technique using anomaly-based on frequent-outlier technique can be used to identify the outlier within frequent dataset.

***Keywords*—**Outlier detection, frequent-outlier, outbreak, anomaly, surveillance, public health.

I. Introduction

PUBLIC health agencies require physical evidence in making decisions concerning the public health system. Generally, diseases are divided into two major groups under the public health system; communicable diseases (CDs) and non-communicable diseases (NCDs). The term non-communicable diseases (NCDs) refer to major chronic diseases including cardiovascular disease, diabetes, cancer and chronic respiratory disease. According to the statistics, NCD accounts for 80% of disease burden for low and middle income countries. In 2008, 57 million deaths and 36 million deaths occurred globally were due to NCD [1] as shown in Fig 1. NCD cases are estimated to reach up to 60% in 2020 and 70% are cases of death [2]. While in infectious disease (CD) are brought by microorganisms and transmitted by people, animals, environment, food or air. An infectious disease depends on the transfer of liquids, contaminated materials, or contacts the carrier to the other healthy individuals [1].

Reference [3] reports that infectious diseases remain a challenge in Southeast Asian countries (SEAR). It is estimated that approximately 40% of the 14 million deaths a year in the SEAR area is caused by CD, by [4].

Infectious disease (CD) and non-communicable diseases (NCD) are increasing each year in the WHO reports. Moderate to extreme pandemic has resulted loss in terms of mortality and mobility. The development of medical facilities and better health care, allow the effort to reduce infectious cases and the spread of information to authorized agencies enables the pandemic movement to be monitored in the world.[5]-[12].

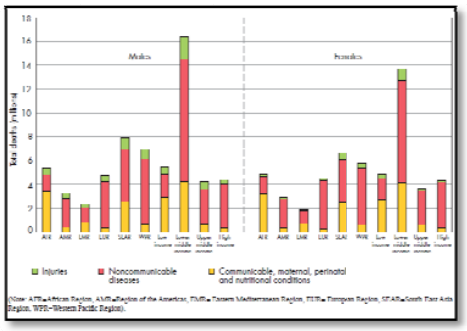

Fig. 1 Total deaths by broad cause group, by WHO Region, World Bank income group and by sex, 2008 (Source: Global Status Report for NCD 2010 WHO 2011)

Based on the above statistics, public health surveillance systems generally work as a support tool in public health agencies. The information required for decision on the dissemination of the possibility of a pandemic occurs and controlling the distribution of vaccine [13]–[15].

Normally the surveillance system is in the form of a passive system. A passive surveillance system depends on the medical report and disseminates information to public health information system voluntarily. Information such as signs and symptoms related to past data and trends that can be used as a comparison of baseline and long-term comparison in outbreaks monitoring [16], [15].

Outbreaks refer to occurrences of one or more cases which are above the norms in the area in a given period of time. The determination of an epidemic outbreak relies on the source of the outbreak and the reasons for the epidemic with analysis carried out using sophisticated laboratory techniques, combined with field investigations to find the cause of the disease, how the disease spreads and developing a measure to control the outbreak. Confirmation of the possibility of an outbreak is important in the determination of the epidemic. The definition of an outbreak is based on common definitions obtained from the Center Disease Control (CDC), online dictionaries and other health department's website. Table I shows the main concepts in defining outbreak [17]-[23].

There is diversity within the definition of an outbreak because of factors such as the environment, government policy

Zalizah Awang Long is with the Malaysia Institute Information Technology, Universiti Kuala Lumpur, Malaysia (e-mail: zalizah@miit.unikl.edu.my).

Abdul Razak Hamdan, Azuraliza Abu Bakar, was with Center for AI, UKM , Malaysia (e-mail: arh;aab@ftsm.ukm).

and the rate the disease spreads [24]- [26]. As in Table I, an outbreak is defined depending on the type and the carrier of the disease. The overall definition is a widely accepted definition of the CDC, the occurrence of more than one case of normal conditions in certain locations within a range of time.

TABLE I
OUTBREAK CONCEPT DEFINITION

Definition	The outbreak key word			
	Data anomaly	**Data type**		**Event**
		Spatial	**Temporal**	
Center Disease Control (CDC) www.cdc.gov	**X**	**X**	**X**	-
Dictionary.com (online) http://dictionary.reference.com/browse/outbreak	**X**	**X**	-	-
Columbia Electronic Encyclopedia (online) www.questia.com/library/encyclopedia	**X**	**X**	**X**	**X**
ESR – Manual public health New Zealand www.surv.esr.cri.nz	**X**	-	-	**X**
Minnesota Dept of Health www.health.state.mn.us	**X**	**X**	**X**	**X**
Island Country Dept of Health www.islandcountry.net/health/outbreak.htm	**X**	-	-	**X**

The objective of surveillance system is to reduce the impact of the epidemic outbreak effects by allowing the authorities to detect cases earlier and thus, enabling the authorities to plan and act accordingly. The surveillance systems act to facilitate outbreak identified detection by allowing the use of a combination of various sources of data. [27]–[30] In addition to the outbreak detection, the surveillance systems are also used in identifying network intrusion, stock market fraud, abuse of power and fraud in the system call in which outlier detection in applied.

There are many studies related to the detection of outlier detection in over 30 decades, beginning with the study of Hawkins (1980). The discussion on the comprehensive study reviewed by [31]-[35] and [36],[37] will be further discussed in the next session.

II. RELATED STUDY

The simplest definition to describe the outlier is "data that is isolated from data". There are a few definitions used to describe an outlier in a variety of outlier detection. Despite the different definitions used, it has the same goal, which is to find something strange or isolated when compared with the comparison group. [33] and [34],[35] classify outlier detection techniques through four main methods: distribution, distance, clustering and density based.

The detection of outlier detection covers various techniques with a broad spectrum of technology. According to [31], the various techniques used in the detection of outlier data is similar, but the introduction used by the authors is diverse. [31] and [38] quoted from Bennett & Lewis (1994) define an outlier as "…that an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occur..". [31], [34], [35] An Outlier, also known as anomaly detection is the phenomenon of viewing patterns within a dataset which does not follow the normal behavior. The situation is referred to as anomalies, isolated data, conflicting observations, and aberration, strange or corrupted in the various domains. The outlier detection has been applied in different domains with various naming such as novel detection; anomaly detection; noise data detection and deviation detection.

The development of the outlier detection techniques are based on statistics and also on data mining. The study conducted by [39] and [40], [41] utilizes a technique known as pattern anomaly detection (APD). The APD technique uses likelihood ratios based on the probability of an odd list of data. Bayesian networks are used to generate probability density model, as well as optimization techniques for studying the structure and estimating the network parameters. The pattern frequency technique (FOFP) [42] is an Apriori-based technique to find common patterns in the data set. FOFP analyzed data point contains at least the pattern and is referred to as outlier data. Reference [42] report, the results achieved by FOFP performs well when compared to clusters-based (CLBOF) and RNN. Based on the distribution method, the AVF technique identified the outlier by analyzing each attribute including the smallest value within the dataset. The overall detection of AVF is equivalent to the FOFP which is fast in detection rate.

Referring to [35] and [31], the terms outlier and anomaly are often used in mutual exchange. The main purpose of this study is to use the outlier technique for outbreak detection. To facilitate the use of the term deviation, the isolated data anomalies and outbreaks in subsequent discussions will refer to the same terms.

References [35], [32], [33] and [31] mainly focused their research on the detection of domain aspects, design

supervision, the type of data used and the techniques used. The table below attempts to formulate a general use of isolated or anomalous data for the concept of an outlier data in the outbreak detection.

Referring to the Table II, the overview of the relation between the detection techniques with multi-domains is shown. Fraud, image processing and the destruction of the industry domain seem to explore various forms of supervision using high and low dimensional data types. While on the medical and public health domain, the exploration of the unsupervised and high-dimensional data has not been explored thoroughly.

Table II also shows the statistical techniques and classification-based techniques used in all domains listed. The exploration of the various techniques in data mining still leaves plenty of room to be explored. The nature of most of the data in the surveillance system is high dimensional and unsupervised data. In addition, large amounts of data often exist in the surveillance system for the data mining exploration techniques. This has led to research opportunities to detect outbreaks using outlier detection in this research.

TABLE II
TECHNIQUE , DATA AND DOMAIN RELATIONS

Domain	Supervision			Data type		Techniques-based					
	A	B	C	1	2	CL	GB	NN	ST	IT	SP
Intrusion		√	√		√	√			√		
Fraud	√	√		√	√	√	√	√	√	√	
Medical		√			√	√		√	√		
Industry		√	√	√	√	√			√		√
Image	√			√	√	√	√	√	√		

NOTA : A=Supervised, B=Semi supervised, C=Unsupervised, 1=high dimension, 2=low dimension, CL=Classification-based , GB=Clustering-based, NN=Nearest Neighour , ST=Statistic , TM=Information theory, SP= Spectral

III. ANOMALY DETECTION

The outlier detection techniques have found a place in research and are being adopted in various domains including intrusion, fraud, and destruction of industry, health and image processing. There is a variety of approaches in the outlier detection techniques such as distribution-based, distance-based, cluster-based and density-based approach.

A. Anomaly-Based Frequent-Outlier (DTK) for Outbreak Detection

DTK defines the combination of an increase of health related events and the behavior of abnormalities as an outbreak. The proposed model here involves steps such as depicted in fig. 2.

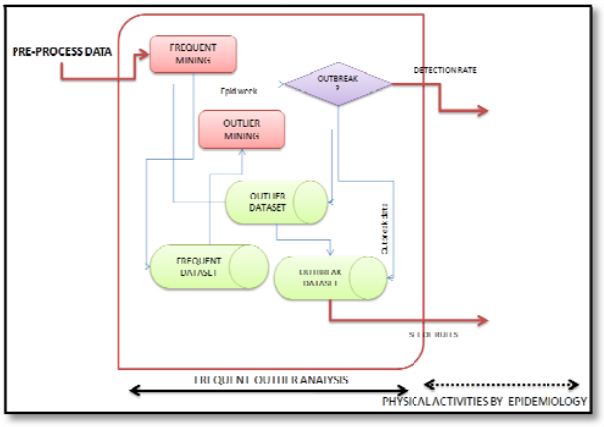

Fig. 2 Anomaly-based Frequent-Outlier (DTK) model

Frequent Outlier (DTK) involves a combination of two data mining techniques consisting of:

– Mining the frequency of item is based on the attribute value (MAV).
– Score generated to identify the outlier records, based on the density of element within the attribute.

Both developments of DTK and MAV techniques are inspired by the concept applied in information retrieval based on Vector Space Model (VSM). SVM is often found in the

information retrieval techniques to represent the p-dimensional column vector (Afred 2008). The documents and queries are represented in vector form.

$$d_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j}) \quad (1)$$

$$q = (w_{1,q}, w_{2,q}, \ldots, w_{t,q}) \quad (2)$$

Each wt, j is the weight of the word j in document t or known as representation of the word bag. Weights are often used in information retrieval and text mining is tf-idf weighting. These weights are used in statistical analysis to explore the frequency of words found in the document and also the frequency of words in all documents.

The calculation of the words found in the documents refers to ti and was found in the document dj. The frequency of words can be represented as below:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (3)$$

The variables ni,j is the number of words (ti) in document dj The devisor is the sum of all words found in the document dj. The idea is that more frequently found words are more important than the less frequently found words within the document. This leads to the deriving measurement called the inverse document frequency idfi in the thesis of Alfred (2008), it is proven that tf-idf is a good weighting scheme in document clustering.

The development of (idf) attempts to develop the frequent mining technique using Multiple Attribute Value (MAV) [43]. The concept of inverse document frequency (tf) was inspired to develop of Frequent-Outlier (DTK) techniques. The combination of frequent and outlier mining techniques led to the development of the outbreak detection.

B. Frequent Mining

Apriori technical developments are fundamental to the development of MAV techniques. Literature review found that modification of Apriori algorithm is necessary to meet the research domain. The modifications of the Apriori algorithm basis were used to meet the need of frequent mining using MAV.

Frequent mining is part of the association mining in generating the frequent item sets and in identifying relationships between the frequent generated items.

The relationship between the items and set items provide useful information in various domains such as marketing analysis, web, networking to detect patterns from the dataset.

The study found that the frequencies of attribute values are unique and different from other attributes in the same transaction. This technique is known as a frequent technique with multiple attribute values (MAV).

C. Outlier Mining

Reference [42] suggested FOFP technique based on the density approach. The FOFP detects an outlier data by viewing the pattern in a set of items. The FOFP identifies the outlier data based on the concept of 'outlier-ness'. While in [44], the AVF uses a method based on identifying the distribution of the outlier. AVF score is generated according to the number of rare values in the attribute in identifying the outlier data.

Reference [42] (FOFP) and [44] (AVF) refer to the calculation of the scores in determining the outlier data. The score used by the researchers are as follows:

$$FOFP\ Score\ (x) = \frac{\sum_{F \subset x,\ F \in FIS} Support\ (F)}{\|FIS\|} \quad (4)$$

$$AVF\ Score\ (x_i) = \frac{1}{m}\sum_{i=1}^{m} f(x_{il}) \quad (5)$$

The FOFP score refers to each frequency subset that exists in record x and is divided with the sum of all frequencies in the whole dataset D in which the data with the lowest score is considered as outlier data. While AVF score refers to the number of I-th attribute value that exists in the dataset. Xi. Low score value indicates the data point as an outlier data.

D. Frequent-Outlier (DTK)

DTK has inspired score based on FOFP score and also AVF score. The approach in developing DTK is derived from the calculation of the frequency of words used in information retrieval and is adopted in obtaining frequent dataset. This refers to the situation where the number of word within the document (ti) is found. The concept used in the DTK score describes using suitable examples as below.

Equation 3 was developed to illustrate the calculation in obtaining the DTK score for detecting outlier in the frequent set.

$$\text{DTK Score } (t_i) = \frac{\sum \frac{DK_{ij}}{x}}{y}, \text{ where } \sum t_i = x\ , \sum P_j = y \quad (6)$$

As for example in Table I, T1 {b}{a}{b}{b}{d}. The total Tn = 10 with Pi = 5. Using T1 as example, the score for T1 isT1 = 3/10, 3/10, 6/10, 2/10 and 4/10. The summation of T1 divided by the number of attributes in the dataset (∑Pi) = (5). A weighted value generated from equation 3 known as DTK Score. The DTK score generated is referred to in identifying the outlier data. The sample calculation for the DTK scores can be seen in Table III.

TABLE III
DTK SCORE IN NEWTRANSTAB

T	P_1	P_2	P_3	P_4	P_5	**DTK Score**
T_1	b	a	b	b	d	0.36
T_2	a	b	b	d	d	0.36
T_3	b	a	b	d	d	0.36
T_4	a	b	b	b	b	0.32
T_7	a	a	b	a	d	0.36
T_{10}	b	b	b	a	b	0.32

Based on equation 3, the T4 and T10 has produced the lowest score value. According to [44], the lowest score refers to the outlier data.

TABLE IV
SCORE VALUE FOR DTK, FOFP AND AVF

Trans	Attribute					Skor data terpencil		
T_n	P_1	P_2	P_3	P_4	P_5	DTK	FOFP	AVF
T_1	b	a	b	b	d	0.36	0.20	0.5
T_2	a	b	b	d	d	0.36	0.20	0.5
T_3	b	a	b	d	d	0.36	0.22	0.5
T_4	**a**	**b**	**b**	**b**	**b**	**0.32**	0.20	0.5
T_7	a	a	b	a	d	0.36	0.25	0.54
T_{10}	**b**	**b**	**b**	**a**	**b**	**0.32**	**0.14**	**0.42**

P1 to P5 are attributes and Tn is the records with elements {P1j , P2j, … Pij}. FOFP score and AVF score referred to as in Table II. The comparison in Table II without taking into account the attributes that is infrequent. The DTK technique is used to identify the frequency of the attribute by applying MAV techniques [45], [43]. The record less than 20% min_supp was not considered during the calculation. The comparison was made based on only frequent dataset identified by the DTK using the MAV technique.

The purpose of this was to review the outlier in the frequent dataset, taking into account only the generations of frequent dataset which are based on the records that meet specified constraints (current min_supp). DTK, FOFP and AVF score managed to trace the T10 as an outlier record. DTK technique is able to detect T4 as outlier record, and show that DTK is able to detect more outlier records in frequent dataset

Based on the example above, the algorithm was developed as shown in Fig. 3 below:

```
For every frequent set (D')/*NewTransTab
Count attribute (y)
For each attribute (y)
Count attribute element (x) for each (y)
Store as xy
End for
End for
For each record (T)
Count attributes value (z)
Skor (z) = Attribute element (xy) / total record (T) / attribute count(Y)
End for
Sort ascending order score (z)
END
```

Fig. 3 DTK algorithm

According to Fig. 3, the algorithm was developed using Java. The dataset was analyzed and the preprocessing was conducted as shown on the following table.

TABLE V
UCI DATASET

Set data	*Outlier technique*	*Structure*		
		Actual	**Remove**	**Total**
Iris Plant (IRP)	RS	150	47	58
Zoo (ZOO)	RS	101	Nil	101
Australia Credit Card (ACC)	RS	690	362	328
Glass (GLA)	RS	214	Nil	214
Coil2000 (COL)	FOFP	160	Nil	160
Lymphography (LYM)	FOFP/AVF	148	Nil	148
Cleaveland (CLV)	FOFP	302	Nil	302
Echoli (ECO)	FOFP	336	Nil	336
Horse Colic (HORSE)	NEW	300	Nil	300
Dermatology (DMT)	NEW	366	Nil	366
Contraceptive Prevalence (CONTRA)	NEW	1472	200	1272
Hayes Roth (HYR)	NEW	132	30	102
Monk (MONK)	NEW	432	Nil	432
Balance (BSWD)	NEW	625	Nil	625

RS = Rough Set Faizah 2008, FOFP = (Hawkins et al 2002; Williams et al 2002; He et al 2003; He et al 2004a; He et al 2005), AVF = Koufakou et al. 2007

IV. RESULTS AND DISCUSSION

The Frequent-Outlier (DTK) was designed to detect an outlier record within a frequent record set. The DTK algorithm is able to evaluate the outlier based on the smallest score generated. Fig. 4, Fig. 5 and Fig. 6 illustrate the performance of DTK.

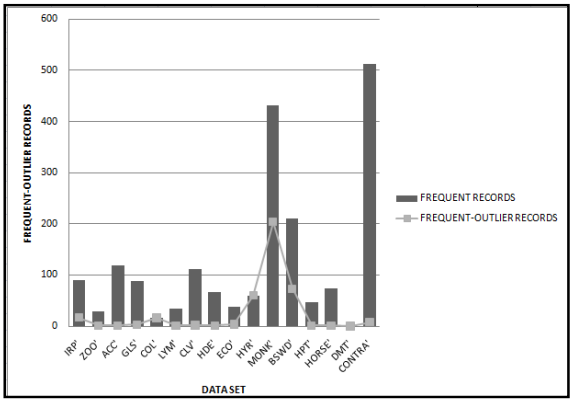

Fig. 4 Outlier records within frequent set records

0.8% to 47.2% was identified as outlier records based on frequent record sets. According to Fig. 4, COL`, HYR` and DMT` no outlier record were identified. This is due to the computation of the score which indicated that there were no smallest values being generated from the given set. Dataset MONK` and BSWD` showed higher percentage in identifying outlier records as in 35% and 47% from frequent records. The

initial structure for MONK` is 432 before and after executed with frequent mining. This can be formulated that MONK` dataset has a uniform distribution of data elements. While in BSWD`, a reduction up to 66% during the process identifying frequent records led to difficulties in identifying outlier records in BSWD` dataset. 1% to 3% outlier records were identified in the dataset IRP`, ACC`, GLS`, LYM`, CLV`, HDE`, HPT`, HORSE` and CONTRA` based on original dataset.

A total of eight datasets were chosen based on the percentage range for the outlier detected in the datasets. Fig. 5 and Fig. 6 shows the DTK techniques perform better than FOFP and AVF.

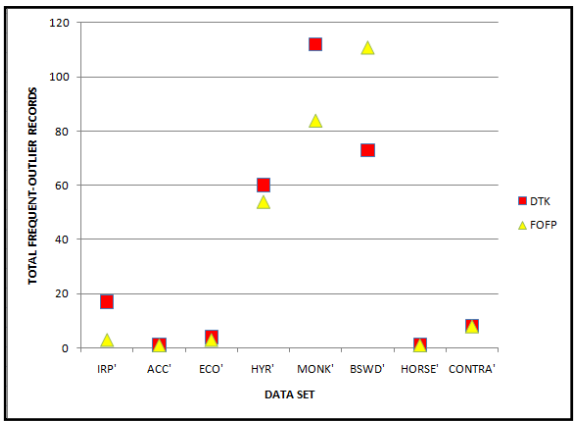

Fig. 5 DTK vs FOFP

The FOFP was developed based on the density method in securing the outlier data. The score generated by FOFP were used in the detection of the outlier data. DTK viewed the density method based on the frequency of the combination of the attribute value as element of attribute value (Multiple Attribute Value). The experiment aimed to explore the possibilities of DTK to detect outlier records based on the density method in terms of frequency as in FOFP.

Based on Fig. 5, DTK is able to track down outlier records more than FOFP in the dataset IRP`, ECO`, HYR`, and MONK`. Unfortunately, DTK was not successful in tracking outlier records in BSWD`. This is due to BSWD` having a complex data structure with a variety of elements in a single attribute. As in dataset ACC`, HORSE`, and CONTRA` the experiment indicated the performance of DTK produced the same result as in FOFP.

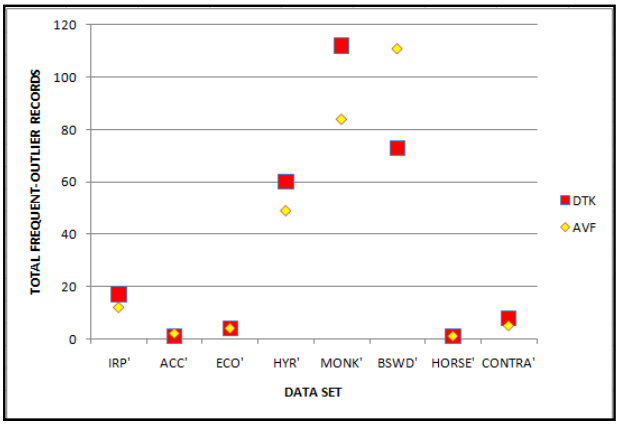

Fig. 6 DTK vs AVF

Based on ideas of every single attribute contain an outlier, AVF was developed according to the distribution method. DTK is improves AVF by viewing the attribute value and identifying common elements and attributes. Fig. 6 shows the DTK doing well in detecting outlier records on dataset IRP`, HYR`, MONK`, and CONTRA`. While in the dataset HORSE` and ECO` the performance recorded an equal results. When comparing results obtain by DTK with AVF or FOFP, the DTK recorded lower results in the dataset BSWD`.

The DTK technique developed a score in identifying outlier records by referring to the lowest score value. In this experiment, the DTK technique successfully identified the outlying records by taking into account various data structures. The comparison involved the ability of techniques to identify outlier records in term of volume detected as in Table VI. Significance tests were also performed and reported in Table VII. The overall performance shows that the DTK technique is able to detect more outlier records in certain dataset. This could be due to the fact that complex data structures resulted in the DTK performing lesser in dataset BSWD`.

TABLE VI
COMPARISON DTK OVER AVF AND FOFP

Comparison Techniques	AVF	FOFP
DATASET		
IRP'	√	√
ACC'	**X**	=
ECO'	=	√
HYR'	√	√
MONK'	√	√
BSWD'	**X**	**X**
HORSE'	=	=
CONTRA'	√	=

Note: √ referring to DTK able to detect higher volume of outlier record; = referring to an equal numbers of outlier records detected; x indicated DTK less performance than comparison techniques.

The propose technique was implemented and the results were recorded and performance significance tests based on

volume outlier records were detected. The results shown that the DTK is more significant based on the sig. value when the value is lesser than AVF and FOFP as in Table VII.

TABLE VII
T-TEST FOR DTK, AVF AND FOFP

	Test Value = 0					
	t	Df	Sig.	Mean diff.	Asymptotic 95% confidence interval	
					Lower bound	Upper bound
DTK	2.326	7	**0.53**	34.5	-.58	69.58
AVF	2.200	7	0.64	33.5	-2.51	69.51
FOFP	2.126	7	0.71	33.1	-3.72	69.97

The main purpose of the development of the DTK technique is to detect outbreak cases in public health. Based on studies, outlier detection is often used in networking, banks, industry, image processing and even in text processing to detect unusual or anomaly activities for the purpose of analyzing and controlling the current situation.

The exploration of outlier detection concept were not implemented directly into outbreak detection, thereby the AVF and FOFP are not tested for the outbreak detection. The DTK concept is based on the definition for outbreak in public health which focuses on the increasing cases in health which is related to the generation of regular data (frequent). In this case, outbreak means is an isolated case or rare event happen in health data. This situation is regarded as outlier detection based on frequent set (frequent-outlier). So, DTK techniques are the combination of frequent and outlier techniques in outbreak detection.

V. CONCLUSION

Frequent-Outlier (DTK) was developed based on density of elements in the attributes that meet the constraints set (min_supp) in generating the frequent dataset. The development of DTK involved the calculation of DTK score in determining the most outlier records based on the smallest score obtain. A FOFP technique was developed based on density concepts. One of the limitations of the density based is the updating of the outlier-ness. DTK was inherent to the capabilities of AVF in identifying an outlier by observing the concept of each attribute which had a strange element. Thus, in DTK, updating the outlier-ness is not required. DTK explored the capacity density of the attribute and the attribute elements for identifying outlier record rather than view the attributes that are peculiar only as in AVF.

Techniques developed are generally intended for the use in detecting the outbreak in public health. Research shows that DTK technique can be applied to general data such as tested with data from UCI with different purposes. Table VIII displays the comparison of detection features for the purposes of outbreak detection.

TABLE VIII
DETECTION FEATURES FOR OUBREAK

Detection features	**Outlier Detection Techniques**		
	FOFP	**AVF**	**DTK**
Frequent detection			√
Outlier detection	√	√	√
Outbreak detection			√
Outlier method	density	distribution	Density & distribution

REFERENCES

[1] WHO. 2011. Global Status Report on non-communicable disease 2010. ISBN 978 92 4 068645 8 (online version) www.bmj.com/content/343/bmj.d4888.extract
[2] WHO. 2002. Innovative Care for Chronic Conditions Building Block for Action. www.who.int/chp/knowledge/publications/icccglobalreport.pdf [2007]
[3] WHO. 2010. Regional Office for South-East Asia Region. http://www.searo.who.int/index.htm . [2010]
[4] Setel, P. W., Saker, L., Unwin, N. C., Hemed, Y., Whiting, D. R. & Kitange, H. 2004. Is it time to reassess the categorization of disease burdens in low-income countries? American journal of public health 94(3): 384.
[5] Cox, N. J. & Subbarao, K. 2000. Global epidemiology of influenza: past and present. Annual Review of Medicine 51(1): 407-421.
[6] Potter, C. W. 2001. A history of influenza. Journal of applied microbiology 91(4): 572-579.
[7] Chan, P. K. S. 2002. Outbreak of avian influenza A (H5N1) virus infection in Hong Kong in 1997. Journal Clinical infectious diseases 34(S58-S64).
[8] Sarubbi, F. A. 2003. Influenza: A Historical Perspective. Southern Medical Journal 96(8): 735.
[9] Rahmat, R.B.H. 2004. Ministry of Health Malaysia Influenza System (M.I.S.S) Clinical & Laboratory Surveillance. Ministry of Health Malaysia
[10] CDC. 2007. Key facts about influenza and influenza vaccine. www.cdc.gov/flu/keyfacts.htm. [Sept 2010].
[11] Lau, E. H. Y., Cowling, B. J., Ho, L. M. & Leung, G. M. 2008. Optimizing Use of Multistream Influenza Sentinel Surveillance Data. Emerg Infect Dis 14(7): 1154-7.
[12] Zimmer, S. M. & Burke, D. S. 2009. Historical perspective--Emergence of influenza A (H1N1) viruses. The New England journal of medicine 361(3): 279.
[13] German, R. R., Armstrong, G., Birkhead, G. S., Horan, J. M. & Herrera, G. 2001. Updated guidelines for evaluating public health surveillance systems. MMWR Recomm Rep 50(1-35).
[14] Buehler, J. W., Berkelman, R. L., Hartley, D. M. & Peters, C. J. 2003. Syndromic surveillance and bioterrorism-related epidemics. Emerging Infectious Diseases 9(10): 1197-1204.
[15] Connolly, M. A. 2005. Communicable disease control in emergencies: a field manual. Geneva: WHO
[16] Pavlin, J. A., F. Mostashari, et al. (2003). Innovative Surveillance Methods for Rapid Detection of Disease Outbreaks and Bioterrorism: Results of an Interagency Workshop on Health Indicator Surveillance, Am Public Health Assoc. 93: 1230-1235.
[17] Axcite.2004. "Introduction to Investigating an Outbreak". Excellence in Curriculum Innovation through Teaching Epidemiology and the Science of Public Health. http://www.cdc.gov/excite/classroom/outbreak/objectives.htm. [May 2008]
[18] MDH, Minnesota Department of Health .2007. "Food Borne Outbreak". http://www.health.state.mn.us/ .[20 July 2009].
[19] ICPH, Island Country Public Health. 2009. " Outbreak Investigation Procedure Island County Public Health Personnel". http://www.islandcounty.net/health/outbreak.htm. [Jan 2010].
[20] Global Alert and Respond. 2009. " What is the pandemic (H1N1) 2009 virus?" http://www.who.int/csr/disease/swineflu/frequently_asked_questions/about_disease/en/index.html [16 April 2010].

[21] CEE, The Columbia Electronic Encyclopedia . 2004. "Influenza Outbreak". Columbia University Press. http://www.questia.com/library/encyclopedia/columbia_university.jsp. [July 2009].
[22] Dictionary, A. H. 2000. American heritage dictionary. The American Heritage Dictionary of the English Language
[23] NZPHA, Public Health Surveillance Information for New Zealand Public Health Action. 2011. "What is Public Health Surveillance". http://www.surv.esr.cri.nz/ . [11 Jan 2011]
[24] Seng, S. B., Chong, A. K. & Moore, A. 2005. Geostatistical modelling, analysis and mapping of epidemiology of Dengue fever in Johor State, Malaysia.
[25] Ooi, E. E., Gubler, D. J. & Nam, V. S. 2007. Dengue research needs related to surveillance and emergency response. Report of the Scientific Working Group Meeting on Dengue. Geneva: World Health Organization. hlm. 124-33.
[26] Runge-Ranzinger, S., Horstick, O., Marx, M. & Kroeger, A. 2008. What does dengue disease surveillance contribute to predicting and detecting outbreaks and describing trends? Tropical Medicine & International Health 13(8): 1022-1041.
[27] Wagner, M. M., Tsui, F. C., Espino, J. U., Dato, V. M., Sittig, D. F., Caruana, R. A., McGinnis, L. F., Deerfield, D. W., Druzdzel, M. J. & Fridsma, D. B. 2001. The emerging science of very early detection of disease outbreaks. J Public Health Manag Pract 7(6): 51-9.
[28] Wagner, M. M., Robinson, J. M., Tsui, F. C., Espino, J. U. & Hogan, W. R. 2003. Design of a national retail data monitor for public health surveillance. Journal of the American Medical Informatics Association 10(5): 409-418.
[29] Wagner, M. M., Espino, J., Tsui, F. C., Gesteland, P., Chapman, W., Ivanov, O., Moore, A., Wong, W., Dowling, J. & Hutman, J. 2004. Syndrome and Outbreak Detection Using Chief-Complaint Data—Experience of the Real-Time Outbreak and Disease Surveillance Project. Syndromic Surveillance
[30] Widdowson, M. A., Bosman, A., van Straten, E., Tinga, M., Chaves, S., van Eerden, L. & van Pelt, W. 2003. Automated, laboratory-based system using the Internet for disease outbreak detection, the Netherlands. Emerg Infect Dis 9(9): 1046-1052.
[31] Hodge, V. & Austin, J. 2004. A Survey of Outlier Detection Methodologies. Artificial Intelligence Review 22(2): 85-126.
[32] Patcha, A. & Park, J. M. 2007. An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer Networks 51(12): 3448-3470.
[33] Zhang, Y., Meratnia, N. & Havinga, P. J. M. 2007. A taxonomy framework for unsupervised outlier detection techniques for multi-type data sets.
[34] Chandola, V., Mithal, V. & Kumar, V. 2008. A Comparative Evaluation of Anomaly Detection Techniques for Sequence Data. 2008 IEEE International Conference on Data Mining (ICDM). hlm.
[35] Chandola, V., Banerjee, A. & Kumar, V. 2009. Anomaly detection: A survey. ACM Computing Surveys (CSUR) 41(3): 15.
[36] Faizah, Shaari. 2008. Outlier Detection Method Based on Non-Reduct Computation using Rough Sets Theory. Tesis Dr. Falsafah, Fakulti Teknologi dan Sistem Maklumat, Universiti Kebangsaan Malaysia.
[37] Zhang, J. 2008. Towards outlier detection for high-dimensional data streams using projected outlier analysis strategy. Tesis Ph.D. Dalhousie University.
[38] Ben-Gal, I. 2005. Outlier Detection. Dlm Maimon. O & Rockach. H. (pnyt). Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Tel Aviv, Israel. Kluwer Academic Publishers.
[39] Das, K. & Schneider, J. 2007. Detecting anomalous records in categorical datasets. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD. hlm. 220-229.
[40] Das, K., Schneider, J. & Neill, D. B. 2008. Anomaly pattern detection in categorical datasets. Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM . hlm. 169-176.
[41] Das, K., Schneider, J. & Neill, D. B. 2009. Detecting Anomalous Groups in Categorical Datasets. Submitted to the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. hlm.
[42] He, Z., Xu, X., Huang, Z. & Deng, S. 2005. FP-outlier: Frequent pattern based outlier detection. Computer Science and Information Systems/ComSIS 2(1): 103-118.
[43] Zalizah A.L, A.R.H, Azuraliza A.B 2011. Frequent Pattern using Multiple Attribute Value for Frequent Itemset Generation. IEEE, Data Mining & Optimization (DMO), UKM.
[44] Koufakou, A., Ortiz, E., Georgiopoulos, M., Anagnostopoulos, G. & Reynolds, K. 2007. A Scalable and Efficient Outlier Detection Strategy for Categorical Data. IEEE International Conference on Tools with Artificial Intelligence (ICTAI), Patras, Greece. hlm.
[45] Zalizah A. L, A. R. H., Azuraliza A.B. 2010. Multiple Attribute Frequent Mining-based for Dengue Outbreak. Lecture Notes In Computer Science.