Data mining Application to Analyze Outbreak Surveillance and Response System: In case of Ethiopia

1st Author
1st author's affiliation
1st line of address
2nd line of address
Telephone number, incl. country code
1st author's E-mail address

2st Author
2st author's affiliation
2st line of address
2nd line of address
Telephone number, incl. country code
1st author's E-mail address

ABSTRACT
In the past, when sanitary conditions were poor, lifestyles were very traditional, and diseases were little understood, epidemics occurred periodically and killed thousands of people. In less developed countries, especially in the tropics, infectious diseases continue to be one of the commonest causes of death, particularly in children. During the past 70 years, there has been a dramatic decline in the incidence of infectious disease, mainly in the developed world, however, the problem is still very high in developing parts of the world, including Ethiopia. For example, communicable disease accounts about 80% of health problems of the country and it also becomes the major causes of morbidity, mortality, and disability to the people. In Ethiopia, there are public health sectors that work all over the country, but due to the lack of adequate performance assessment and data quality measure its emergence surveillance and response systems are still unproductive to deliver right evidence to tackle the problems aptly. The sector tries to limit the problems based on simple statistical data analysis approach using MS-Excel results, however, this approach couldn't support prediction of future occurrences. The aim of this study is, therefore, to show the applicability of data mining techniques and algorithms on the existing surveillance system databases using descriptive and predictive models. To do that, the study incorporated three data mining applications, including classification, clustering and association rules mining. Consequently, an attempt was made to investigate five chosen epidemic-prone disease outbreaks using 8796 usable records collected between the years 2004 to 2012. Based on the time and place dimensions, the test of two classification algorithms (i.e. decision-tree J48 with 87.44% and Naïve Bayes with 83.70% accuracy level) showed significant predictive potentials in case of Epidemic typhus occurrence and prevalence. However, in comparison, decision tree J48 algorithm was performing better than Naïve Bayes. Besides, association rule mining using Apriori algorithm showed that there was a correlation between some disease outbreaks regarding time and space perspectives. Finally, the results of the simple K-Means clustering method showed that there were clear groupings of disease outbreaks within the various time episodes across the country. Therefore, applying data mining techniques on emerging and re-emerging disease outbreak management activities are vitally important to develop well performing descriptive and predictive model to Ethiopian public health emergency management sectors. By suggesting the significant of quality data to be held by the sector, which was one of the major headache in the study, the study provided potential contributions for the planning, preparedness, decision-making, disease control and prevention measures of the sector and the domain experts using data mining applications.

Keywords: Data mining, Surveillance, KDD, Healthcare, Epidemic, Outbreak, association, cluster, classification, public health

INTRODUCTION
In the past, when sanitary conditions were poor and diseases were little understood, epidemics occurred periodically and killed thousands of people. One of the largest epidemics ever recorded was the outbreak of bubonic plague that raged throughout Europe, Africa and Asia from 1347 to 1350G.C. and killed one-third of the European population. An outbreak of influenza in 1918 also killed over 20 million people around the world. However, during the past 70 years, there has been a dramatic fall in the incidence of infectious diseases, particularly in developed countries [2]. This was because of several factors including: immunization, Anti-microbial chemotherapy, improved nutrition, better sanitation and housing. As stated in [7], the morbidity and mortality associated with infectious disease outbreaks, which are directly or indirectly linked to ecologic or climate events and trends pose a growing problem for global public health. In less developed countries, however, especially in the tropics, infectious diseases continue to be one of the commonest causes of death, particularly in children [2]. In addition, to that, Ethiopia, as part of the developing world, has big health problems. Most of the time, communicable diseases account about 80% of the health problem in the country. Unfortunately, such types of diseases could be prevented by simple sanitary measures and nutritional programs. However, they are still becoming the major causes of morbidity, mortality, and disability in the country [3]. In addition, communicable disease prevalence are high in the country, because of the poor socio-economic development, environmental factors, lack of access to safe and adequate sanitation facilities. As a result, the problems become inherent for centuries and affect three-fourth of the children in the county [4]. In 1996, as part of the response to the growing public health problem, especially to communicable diseases, Ethiopia introduced an Integrated Disease Surveillance and

Response (IDSR) strategy and later called Public Health Emergency Management (PHEM) focusing on 20 selected priority diseases in her surveillance and response interventions. Out of the 20 communicable diseases in the database, the weekly reportable diseases are only five. Surveillance, risk assessment and outbreak response capacity are a prerequisite for effective management of emerging disease outbreaks and other acute public health events. Effective national surveillance systems can generate reliable information for timely risk assessment strategy to respond to public health problems [8]. So, to better identify clusters, track trends, assess the effectiveness of interventions in explaining and predicting emerging and reemerging disease outbreaks require quality surveillance data-sets that could provide timely results [11]. One of the most challenging tasks facing an epidemiologist is working in a public health environment to investigate disease outbreak [13]. Proper collection and storage of disease outbreak data are very important to develop well-performing predictive model that can predict the future disease occurrences [14]. The PHEM sector under study stores data about disease outbreak using MS-excel work sheets and analyze through epi-info, and SPSS.

The rational was that, if there are adequate and relevant data sources, one could easily characterize disease outbreaks by time, place, and person [13]. However, characterizing an outbreak in that manner would only provide descriptive results that could only show what happened in the society after the disease had already occurred. To that end, the sector were attempting to perform interventions to limit the problems using descriptive statistical analysis of data, including percentage, mean, median, normal distribution, and standard deviation via simple visualization techniques like tables, graphs, charts, normal curves, which are entirely traditional data analysis and presentation approaches. However, the above statistical approaches are unproductive to analyze huge data-sets of several years as observed in Ethiopian emergency management and response sector. Therefore, the ability to extract useful knowledge hidden in data-sets of such type are becoming progressively essential in today's competitive world. To that end, applying data mining tools and techniques could provide better solutions to describe the existing situations and predict future occurrences of outbreak cases by suggesting a more rigorous descriptive as well as predictive models. To the current study, therefore, testing the application of data mining tools and techniques is unquestionable to generate strong decision-making alternatives in case of epidemic-prone outbreak diseases. Finally, the primary goal of this study was to analyze the outbreak surveillance and response system of Ethiopia using data mining applications to point out the possibilities of applying data mining techniques and tools to manage the problems by developing better descriptive and predictive model to escalate the decision making efforts of the sector in early warning, planning, preparedness, and response activities. In doing so, three data mining applications such as classification, clustering and association rules mining were tested to explore their applicability using PHEM data-set pools. Finally, the study would address three important questions shown as below: 1) to what extent data mining techniques and tools apply on the surveillance and response data-sets of Ethiopia? 2) Could it be possible to develop more efficient predictive and descriptive models to emerging and reemerging disease outbreak cases? 3) Are there any associations among emerging and reemerging disease outbreak cases in the PHEM data-set? To address the above research questions, a research method that could help in discovering important knowledge from the surveillance data-sets would be applied accordingly.

METHODOLOGY

This study considered all weekly reportable disease outbreak cases collected within Ethiopian public health emergency management (PHEM) sector. The target population were all reported disease outbreak cases gathered from all over the country. The study, particularly, would incorporate and analyze only five weekly reportable disease cases from the 20 epidemic-prone priority diseases identified in the surveillance system, which were collected between the years 2004 to 2012, including Malaria, Meningitis, Relapsing Fever, Typhoid Fever and Epidemic Typhus. The study used the Hybrid KDD data mining process model to analyze and evaluate the problems of the existing surveillance systems by understanding the data and discovering important patterns. It was because, the discovered knowledge and their usability could support to determine the applicability power of data mining techniques and tools to investigate outbreak surveillance data-sets. Accordingly, Apriori association rule mining was applied for pattern discovery to see the co-occurrence of outbreak disease cases; classification algorithms were used to predict the future occurrence of disease outbreaks with regard to time and place dimensions (here, the decision tree and Naïve Bayes classifiers were applied to predict Epidemic Typhus disease alone); and lastly Simple K-Means clustering algorithm was examined to see how disease outbreak cases were grouped together to describe outbreak prevalence across the country within the defined time period. Here, data transformation or pre-processing were performed in the first step after the data collection process to convert the data in to useable formats. Since, the surveillance system database were stored with MS-Excel file format while collecting data, we transformed them in to data mining tools acceptable format like Comma Separated Value (.CSV) text format and Attribute-Relation File Format (.ARFF). Consequently, the data prepared to be analyzed by applying data mining techniques using Weka 3.6 and Weka 3.7 software tools were executed to investigated the data-sets and discover the hidden patterns from the surveillance data. So, the findings of the study were planned to conduct more rigorous investigation on the surveillance and response system using recent data-sets to develop strong predictive and descriptive models. Because, the aim was to explain emerging disease surveillance and response systems of the country by supporting the decision making processes of the sector. Finally, the study would use the various steps of hybrid data mining process model as shown in the figure 1 to better understand the problems and discover important rules in the process.

DM, KD Process Model, and Hybrid KDD

KDD helps humans make sense of huge amounts of data by mining patterns [20]. In addition, KDD has evolved, and continues to evolve, from the intersection of research fields such as: machine learning, pattern recognition, databases, statistics, AI, knowledge acquisition for expert systems, data visualization, and high-performance computing to list some.

In general, the driving force behind KDD is the database field, however, combining different models together with KDD can provide better insights on a given problem area. Therefore, to this specific study, the combined educational and industrial knowledge process model called hybrid KDD is adapted for data analysis and knowledge discovery process. In one hand, the hybrid KD Process Model stretches from the process of understanding the problem domain and data, through data preparation and analysis, to evaluation, understanding, and application of the generated results. On the other hand, hybrid KDD model has been widely used in medicine where this study was coined. The proposed model emphasizes the iteration of activities within the various phases of data mining processes via many feedback loops that are triggered by a revision process as shown in figure 1.

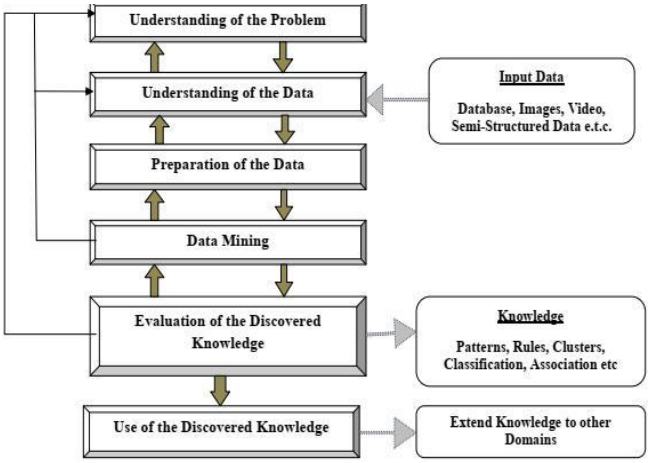

Figure 1: The six-step of hybrid KDD process model

APPLICATION OF DATA MINING IN HEALTHCARE

As stated in [35], Medical data mining applications has great potentials to explore hidden patterns of medical data, so that these patterns could be utilized for clinical diagnosis, pattern analysis, disease investigation, clinical decision making, disease outbreak investigation, and so on. The importance of decision support system in the delivery of managed healthcare can hardly be overemphasized [25]. Evidently, the existing raw medical data-sets are widely distributed, heterogeneous in nature, and voluminous in size. That means the healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden knowledge. However, medicine is highly sensitive to information distortion and its consequences with life threatening potentials [36]. In addition, medical diagnosis is regarded as one of an important yet complicated task that needs to be executed accurately and efficiently [35]. So, healthcare data in any form should be collected in an organized form and then integrated within the healthcare information system to discover important patterns, because the automation of such systems would be extremely advantageous. However, it needs a comparative study of the various available techniques to do so. Here, in this study the aim was to analyze the PHEM sector disease outbreak database by applying data mining algorithms and techniques to drive different predictive as well as descriptive results [35] to support the performance impacts of the healthcare systems. Furthermore, as stated in [38], the healthcare environment is generally perceived as being 'information rich' yet 'knowledge poor'. Therefore; there is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools and techniques to discover the hidden relationships and trends in the data. As already mentioned above, data mining applications can greatly benefit the healthcare industry without much limitation. However; according to [40][39], healthcare data mining applications are limited by about four major factors such as: 1) accessibility of compiled data, 2) data quality problems, 3) successful application of data mining requires knowledge of the domain area, and 4) data mining methodology and tools. Besides, lack of standard clinical vocabularies are serious hindrance to apply data mining on healthcare data [39].

Healthcare data mining projects can fail for a variety of reasons such as lack of management support, unrealistic user expectation, poor project management, and inadequate data mining expertise and many more [40]. So, information technology professionals and public health professionals have to work cooperatively to enhance the decision support power of health-related cares. Here, classification algorithms like decision tree (DT) and Naïve Bayes (NB) were used to forecast future occurrences of epidemic typhus disease outbreak. Besides, descriptive data mining modeling using association rule mining, specifically, Apriori algorithm were run to show the co-occurrence of diseases outbreaks. Lastly, disease segmentation via cluster analysis method, in particular, with Simple-K-Means algorithm was performed. According to WHO, effective communicable disease control relies on effective response systems and effective response systems rely on effective disease surveillance. As defined in [52] [55], Healthcare surveillance can be described as "the tracking and forecasting of any health event or health determinant through the continuous collection of high quality data, integration, analysis and interpretation of data results into surveillance products, thereby disseminating such products to whom they need to know [to address] as a means of "Information for Action" [52]. In general, Healthcare surveillance is an essential components of evidence-based decision-making practices in the public health systems. As mentioned above nowadays, investigations of diseases are more complex than they were in the past, because of emergency of new pathogens, risk factors and outbreaks, which cross jurisdictions and national boundaries, often raising political and economic burdens. The outbreaks of infectious disease and the recurrent incidences of natural disasters remind the importance of public health systems that encompass the government and private sector, academia, NGOs, associations and development partners. To that end, there were tremendous efforts performed to narrow down such public health gaps to maintain adequate intervention and limit the consequences of natural and human-made disasters [52], even if, the problems are still very high.

EXPERIMENTATION FOR MODEL DEVELOPMENT

Accessing and processing of the real emerging disease outbreak data for experimentation is a very difficult task because of national security and confidentiality problems.

However, based on the decision made by public health experts of the sector, the sector allowed us only weekly reportable routine surveillance data-sets for investigation. The data comprise five selected epidemic prone disease cases of outbreak nature such as Malaria, Typhoid Fever, Epidemic Typhus, Relapsing Fever, and Meningitis. So, we would develop the model of this study based on the data available on the surveillance database. A total of 9 years of 18,600 records were collected initially from PHEM and preprocessed using hybrid KDD process model. After the preprocessing phase a total of 8796 usable records were obtained to data mining experimentations. Out of the total usable records, the study used 4703 records from IDSR system database and 4093 records from PHEM system database recorded from the year 2004 to 2008 and from 2009 to 2012 respectively. For prediction purpose, Tenfold (i.e. 10-fold) cross-validation was used to test the performance of a decision tree and Naive's Bayes classification algorithm. In addition, the study tested simple K-means algorithm to group the data of disease outbreaks regarding their prevalence and occurrence through the lenses of place and time setting. Finally, the study examined the applicability of Apriori association rule mining algorithm to see whether there was an association between different diseases to occur together or not in the same time and place orientation. To see the association of disease co-occurrences, through Apriori algorithm rule interestingness and usefulness evaluation was performed with a minimum support of above 20% and a minimum confidence of above 90%. The reason of using a small support (i.e. 20%) when to develop interesting rules are to include even the rarely occurring disease cases (like meningitis) that were relatively less prevalent to occur but important to public health actions. As a result, strong rules are those rules that satisfy both minimum support threshold (min-sup) and minimum confidence threshold (min-conf). For the sake of convenience, we put support and confidence values to appear between 0% to100%, rather than 0 to 1.0 by choice to this study.

Experimentation on association rule mining

As indicated the study used two different Weka software versions to generate some interesting rules using Apriori association algorithm from unsupervised datasets. For experimentation, the study included all the five disease outbreak cases to show the co-occurrences and non-co-occurrences of such disease cases. Some rules that were generated from the experiment and repeated on the subsequent rules were automatically dropped out. To that, all the three datasets (i.e. IDSR, PHEM, and the combination of IDSR and PHEM) were involved. As a result, comparative results of the five subsequent experiments from the three data-sets indicated that there were some repeated rules to be eliminated. Even if, the purpose of the study was to see how data mining techniques and tools are applicable in the surveillance and response systems of Ethiopia, this approach allowed us to focus and discuss only on some best interesting rules that could perform well. Therefore, rules as shown below in table 1 were the most important once obtained from the five subsequent experiments of the three data-sets.

Exp. No.	Total Rules in the experiment	IDSR Datasets		PHEM Datasets		Combined Datasets (PHEM & IDSR)	
		Repeated Rules	Rules after removal of repeated Rules	Repeated Rules	Rules after removal of repeated Rules	Repeated Rules	After removal of repeated Rules
E1	10	-	10	-	10	-	10
E2	15	9	6	10	5	9	6
E3	20	14	6	13	7	16	4
E4	25	12	13	9	16	10	15
E5	30	29	1	27	3	27	3
Total Rules	100	64	36	59	41	62	38

Table1: Comparison among the three datasets for Apriori association rules mining result

The results of the association rule mining showed that disease outbreaks were occurring together and non-occurring together at some area at a point in time. So, the study projected which disease outbreak occurred with the occurrence of which types of other disease outbreaks. In addition, the experiments have also revealed that when there were non-occurrences of some disease outbreak cases some another disease outbreaks were not occurring. As a result, some best common rules, which were getting greater acceptance from the domain experts, were considerately chosen for the sake of rule comparison purpose shown as follows in the table 2 below.

Dataset Name	**Rule No.**	**Best Rules Produced**				**Measure of Association**		**Remark**
		Antecedence		**Consequence**				
		Occur	**Not occur**	**Occur**	**Not occur**	**Confidence**	**Support**	
IDSR	1.		TF, RF		ET	99%	18.71%	**Not accepted**
	2.	ET		TF		99%	21.01%	Accepted
	3.	Mal, ET		TF		98%	18.86%	**Not accepted**
	4.	Mal, RF		TF		94%	28.86%	Accepted
	5.	RF		TF		94%	32.87%	Accepted
	6.	TF		Mal		91%	72.12%	Accepted
PHEM	1.		TF, RF		ET	99%	26.34%	Accepted
	2.	ET		TF		99%	31.57%	Accepted
	3.	Mal, ET		TF		99%	31.3%	Accepted
	4.	Mal, RF		TF		98%	29.12%	Accepted
	5.	RF		TF		96%	29.83%	Accepted
	6.	TF		Mal		97%	70.29%	Accepted
(IDSR+ PHEM) Datasets	1.		TF, RF		ET	99%	22.27%	Accepted
	2.	ET		TF		99%	25.97%	Accepted
	3.	Mal, ET		TF		99%	23.31%	Accepted
	4.	Mal, RF		TF		96%	29.02%	Accepted
	5.	RF		TF		95%	31.46%	Accepted
	6.	TF		Mal		94%	71.29%	Accepted

Table2: Selected rules and their respected level of support and confidence

N.B: - ET = Epidemic Typhus disease case, Mal = Malaria disease case, Men = Meningitis disease case, RF = Relapsing Fever disease case, TF = Typhoid Fever disease case.

N.B: Rule numbers were given manually for the sake of discussion purpose alone

The remark part in the table above showed that rules were accepted or not accepted according to their interestingness measures regarding the minimum support and minimum confidence threshold of the rules as defined in rules' interestingness measure. The results of the study implied that rules were accepted whenever they were interesting and not accepted whenever they were not interesting based on the given defined acceptance threshold. Even though, there were variations among the three datasets in providing interesting rules, the average support values of over 20% to all the data-sets were accepted. In short, the average support and confidence values of the six chosen rules were greater than the required minimum defined threshold (i.e. Min-Sup=20% and Min-Conf=90%). Finally, the researchers had manually removed rules that were repeated and again appeared on the subsequent iterations to make proper analysis decisions on the remaining rules. Data mining applications, in general, association rule mining techniques, in particular, were highly significant and very applicable to analyze disease outbreak data-sets of Ethiopian surveillance and response system database.

Experimentation on Epidemic Typhus classification

Classification algorithms were executed to classify disease outbreaks regarding their current occurrences and future incidences at a point in time within a certain area. Here, the aim of the study was only to predict the occurrence of disease cases to show the future occurrence of new or reemerging incidences. To that end, the study incorporated decision tree with J48 and naïve Bayes classifiers on the newly established PHEM datasets using the 10-fold cross-validation. For prediction purpose, attribute evaluator (i.e. feature selector) called Gain Ratio feature evaluator was used for testing by the search method of Attribute Ranker to select the best attribute. So, the study incorporated ranked list of attributes according to their gain ratio of each attribute's value to predict the chosen class. After pre-processing phase, we analyzed around 4093 usable records using seven chosen attributes (Region, Zone, Woreda, Year, Month, and Epidemic Weeks) to predict the class called occurrence and non-occurrence of Epidemic typhus. Finally, the occurrence of an outbreak was represented by 1 and non-occurrence by 0. Later on, we converted the value 1 to 'YES' for the occurrence and 0 to 'NO' for the non-occurrence for convenience. Thus, the analysis results of decision tree clarifiers showed better statistical values than the naïve Bayes classifiers as presented in the table 3 below.

Classification Metrics for predictive model development

After a model has been developed to predict disease outbreaks using the current PHEM datasets, performance measure was done using the training data-sets so as to utilize model testing activities in the upcoming data of the surveillance system. In that regard, the future epidemic typhus disease outbreaks would be accurately predicted when and where outbreaks could occurs. To that end, the predictive performance of decision tree and Naive Bayes Classifier using confusion matrix was empirically depicted as shown in table 3. As described in [50], classification models were evaluated using the conventional machine learning metrics such as Precision, Recall, F-Measure, TP Rate, FP Rate, and ROC Area. Because, as of [26], the *confusion matrix* is a useful tool for analyzing how well your classifier can recognize tuples of different classes. In addition, the confusion matrix is more commonly named contingency table [51] like table 3. It states that the number of correctly classified instances are the sum of diagonals in the matrix; all the others are incorrectly classified.

No.	Type of Measure	Epidemic Typhu disease Case occurrence	Naïve Bayes	Decision Tre with J48 Algorithm
1	TP Rate	No	0.798	0.915
		Yes	0.923	0.795
		Average	0.838	0.877
2	ROC Area	No	0.939	0.935
		Yes	0.939	0.935
		Average	0.939	0.935
3	Sensitivity		0.6814	0.8141
4	Specificity		0.9570	0.9049

Table3: Measurement evaluation for the two decision tree classifiers for comparison

As shown in the table above, one could see the occurrence of disease cases were measured by four important prediction performance tests. Since the study was planned to see the best performing classification algorithms to predict the occurrence of *Epidemic typhus* outbreak at a certain place and a future point in time, performance was evaluated by average true positive (TP) rate to evaluate Decision Tree with J48 classifier and the Naïve Bayes classifier. From the experiment, the average TP rates were 83.8% and 87.7% for Naïve Bayes and Decision Tree J48 classifiers respectively. In principle, the better the average TP rate the better the performance will be. The second one was the ROC area which is the combination of the sensitivity and specificity measure. The average values of ROC area were 93.9% for Naïve Bayes and 93.5% for Decision Tree classifiers, which was greater than 90% and nearly equal, showed that both algorithms generated better performance. Since specificity and sensitivity are the most important performance measures in the field of healthcare related studies. As a result, the two measures were compared and found that the sensitivity of Naïve Bayes algorithm was 68.14% and that of the Decision tree j48classifier was 81.41%. Based on the measures of sensitivity, the true occurrences of disease cases were to be classified as an occurrence. From the above test results, one could understand that decision tree had a better performance impact than Naïve Bayes classifier to predict the future occurrences of Epidemic Typhus outbreaks. In contrast, specificity measure showed that the non-occurrences of Epidemic Typhus outbreaks were classified as non-occurrences with the test results of 95.7% and 90.49% for Naïve Bayes and decision tree classifiers respectively. Even though, both classifiers had shown high specificity measures with a relative higher performance measure of Naïve Bayes classifier, as per the objective of the study, outbreak disease cases were evaluated not to test the non-occurrence as can be measured by specificity, but the occurrences as can be measured by sensitivity. So, sensitivity, TP rate and the ROC

area measures were taken in to consideration to make decisions on what classification algorithm better performance was attained. Moreover, Decision tree classifier (using an IF-THEN rules representation) had classified data instances correctly with 87.6619% accuracy rate, but that of the Naïve Bayes classifier was with 83.7772%. Even if, the aim of the study was to show and understand how data mining techniques and tools are applicable in case of disease outbreak surveillance and response databases, based on the discussions and justifications above decision tree with j48 was selected for model development as having better performance than Naïve Bayes classifier.

Experimentation to analyze outbreak clusters

As per the objective of the research, better results and interesting clusters were observed from Simple K-means clustering algorithm experiment results. So, data mining have brought greater applicability to correctly describe the data-sets of disease outbreak surveillance and response system of the country. The study used the combined data-sets of 8796 usable records for clustering techniques. The results were also interesting to design new hypotheses and conduct empirical studies in the field of machine learning applications, including data mining techniques within the public health emergency management (PHEM) activities of the country. Because, the above implication could provide a measurable plan of action for decision makers to better prepare for and respond to outbreak incidences. In fact, we had two options to find the number of clusters (i.e. K) from the whole data-sets: 1) using the total numbers of years in which data collection process in the sector was performed (i.e. 9 years of data-sets); and 2) the total number of regions in the country in which data were reported (i.e. 11 total regions indicated in the data-sets). To that end, we discussed with domain experts to productively decide the number of Ks and we agreed to cluster the data-sets based on the number of regions in the country (i.e. 11 clusters). Cluster analysis showed that the year 2011 was indicated as having more outbreak disease occurrences than the other sampled years and appeared within 4 different places across the country. Except the years 2004 and 2012 were we could find any cluster, all the other years in the experiment appeared exactly one each. So, the surveillance and response system experts should have to cross-check the prevalence of outbreak disease cases against their plan of actions on that specified year and places of occurrences.

STUDY LIMITATION

Even though, there were promising results from the study that meet its objective, there were some important limitations as other studies have. The first limitation is that the data-sets were poor quality and having very shallow dimensionalities. So, using such datasets to provide significant implications may not be fully applicable, since data quality and dimensionality are the upmost important things to use machine learning, including data mining. In addition, the datasets were collected many years back from 2004 to 2012 with a lot of missing values, so, the current world reality dynamics might be different from the time when the data were collected. So, providing recommendation based on such data-sets might not work well today. The surveillance system was lacking consistent and proper data reporting format while gathering data from all over the country. For example, outbreak data reporting formats of different regions were very different. In addition to that, there are diverse data mining algorithms and techniques with different predictive and descriptive powers and applicability, but this specific study was only relied on some common algorisms, including decision tree with j48, Naïve Bayes, Apriory, and Simple-Kmeans from the set of several techniques. Therefore, the potential power of data mining applications might not be fully utilized.

CONCLUSION

The PHEM surveillance and response system were suffered from a serious shortage of quality data and consistent data reporting mechanisms across the various regions of the country. In addition, number of attributes (e.g. lack of unseen variables) in the PHEM data-sets were not adequate for applying data mining. So, the absence of quality data in the sector might adversely affect the quality of both the predictive and descriptive power of the models. Disease prediction or classification has shown that some places of the country were more vulnerable than others for the incidence of Epidemic typhus disease. Decision tree J48 algorithm was better than Naïve Bayes classifier for the data at hand to prepare prediction model in relation to time and place settings. It was found that diseases were associated to occur together and not occur together. Therefore, association rule mining with Apriori algorithm was an important means to show the real association of disease cases (but here the associations didn't show causalities). Based on the data-sets, the more affected and the less affected areas of the country were clearly identified via clustering techniques for 9 successive years. In general, data mining techniques were important and highly applicable in the classification, clustering and association rule mining for emerging and reemerging disease outbreak cases. Having that in mind, outbreak disease planning, preparedness, and response would be much easier if data mining applications are becoming a means of data analysis in the face of quality and voluminous data-sets as expected from such sectors. Apriori algorithm couldn't detect the infrequently occurring rear disease cases like meningitis even with the defined minimum support value of 20%. The simple K-Means clustering algorithm showed that some disease cases such as: Malaria and typhoid fever were more frequently occurring than others throughout all regions of the country. From the two data-sets incorporated, one can conclude that most prevalent disease cases such as: malaria and typhoid fever were also highly prevalent in the former IDSR data-sets. It indicated that the two outbreaks were not effectively controlled by the newly established PHEM surveillance and response system. However, in general, the current PHEM surveillance system of the country was better than the former IDSR system in providing interesting patterns with data mining techniques and tools.

REFERENCES

[1] W.A. Redmond, "Epidemic," Microsoft Corporation, Microsoft® Encarta® 2009, 2008.

[2] A. Mulugeta, "Communicable Disease Control," Ethiopian public health training initiatives, Hawassa University, Ethiopia, 2004.

[3] K. Abera, A. Ahmed, "An overview of environmental health status in Ethiopia with particular emphasis to its organization, drinking water and sanitation," Department of Community Health, Medical Faculty, Addis Ababa University, Ethiopia, 2005.
[4] Krusche, OtterWasse, "Introduction of Ecological Sanitation for Large Scale Housing Programs in Ethiopia," GTZ, 2007.
[5] B. Tsedaye, "Epidemiology and Disease Surveillance," FMOH/TUTAPE, 2009
[6] WHO regional office for Africa, "WHO country Cooperation strategy 2008–2011," Ethiopia, India, 2009.
[7] R. Grant, S. Spring, "A multidisciplinary approach for the early detection and response to disease outbreaks," Armed Forces Health Surveillance Center, USA, 2011.
[8] WHO, "Asian pacific strategy for Emerging Diseases," Library Cataloguing in Publication Data, 2010.
[9] G. W. Richard, "Reviewing Ethiopia's Health System Development," International Medical Community, JMAJ 52(4): 279–286, 2009.
[10] R. Freeman, "Evolution of healthcare-associated infection surveillance in England: initiatives, implementation and opportunities for innovation," 2011.
[11] Communicable Disease Prevention and Control (CDC), "Communicable Disease Prevention and Control (Focus Area Profile) plan," Wisconsin Department of Health, *Healthiest Wisconsin*, 2010.
[12] J. A. Patz, A. K. Githeko, J. P. McCart, S. Hussein, U. Confalonieri, N. de Wet, "Climate change and infectious diseases".
[13] U.S. Department of Health and Human Services, Principles of Epidemiology, 2nd Edition, "An Introduction to Applied Epidemiology and Biostatistics," Centers for Disease Control and Prevention (CDC), Epidemiology Program Office, Public Health Practice Program Office, Atlanta, Georgia 30333, 2005.
[14] Institute of Environmental Science & Research, "Disease Outbreak Manual," Porirua, New Zealand, 2002.
[15] A. Ghosh, B. Nath, "Multi-objective rule mining using genetic algorithms", Information Sciences, 123–133, 2004.
[16] M. Kantardzic, "Data Mining: Concepts, Models, Methods, and Algorithms," University of Louisville, The Institute of Electrical and Electronics Engineers, Inc., 2003.
[17] S. Chakrabarti, E. Cox, E. Frank, RH. Güting, J. Han, X. Jiang, et al., "Data Mining knows it all," Morgan Kaufmann, Elsevier Inc., 2009.
[18] M. Refaat, M. Schneider, T J. Teorey, I H. Witten, "Data Mining knows it all," Morgan Kaufmann, Elsevier Inc., 2009.
[19] A. Omari, "Data Mining for Retail Website Design and Enhanced Marketing," Inaugural- Dissertation, Mathematisch-Naturwissenschaftlichen Fakult¨at der Heinrich-Heine- Universit¨at D¨usseldorf vorgelegt von, 2008.
[20] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," American Association for Artificial Intelligence, Vol. 17 no. 3, 1996); 1996.
[21] En. Tayfun, "Parallel Closet+ Algorithm for Finding Frequent Closed Itemsets," Masters' thesis, Middle East Technical University, 2009.
[22] K.J. Cios, W. Pedrycz, R.W. Swiniarski, L.A. Kurgan, "Data mining, A knowledge discovery approach,", Springer, 2007
[23] J. Jackson, "DATA MINING: A CONCEPTUAL OVERVIEW," Communications of the Association for Information Systems vol. 8, 267-296, Management Science Department, University of South Carolina, 2002.
[24] J. Ponce, A. Karahoca, "Data Mining and Knowledge Discovery in Real Life Applications," Croatia, 2009.
[25] J. Ranjan, R. Nagar, Ghaziabad, "Applications of Data Mining Techniques in Pharmaceutical Industry," Information Management and Technology Area, Institute of Management Technology, Journal of Theoretical and Applied Information Technology, Uttar Pradesh, India, 2007.
[26] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," 2nd Edition, the University of Illinois at Urbana-Champaign, Morgan Kaufmann, 2006.
[27] G. Shmueli, N R. Patel, P.C Bruce, "Data Mining in Excel: Lecture Notes and Cases," USA, 2005.
[28] L. Ma1, F. Tsui, HR. William, MW. Wagner, Ma. Haobo, "A Framework for Infection Control Surveillance Using Association Rules," RODS Laboratory, Center of Biomedical Informatics and Intelligent Systems Program, University of Pittsburgh, Pittsburgh, 2003.
[29] S.N. Sivanandam, S. Sumathi, P. K Madhavan, "Data Mining Concepts, Tasks and Technique," 2006.
[30] S. Vinnakota, N.S Lam, "Knowledge Discovery from Mining the Spatial Associations Between Cancer Mortality and Socioeconomic Characteristics," Department of Geography and Anthropology, Louisiana State University, Baton Rouge, La 70803, USA.
[31] S. Jain, M. Aalam Afshar, M.N. Doja, "K-MEANS CLUSTERING USING WEKA INTERFACE," The 4th National Conference, New Delhi, India, 2010.
[32] A.K. Jain, "Data Clustering: 50 Years beyond K-Means," Department of Computer Science & Engineering, Michigan State University, USA, *2009*.
[33] D.T. Larose, "Discovering Knowledge in Data, an Introduction to Data Mining," A John Wiley & Sons, Inc., Publication, 2005.
[34] S. Nasser, R. Alkhaldi, G. Vert, "A Modified Fuzzy K-means Clustering using Expectation Maximization," Department of Computer Science and Engineering, 171, University of Nevada Reno, Reno NV 89557, USA.
[35] J. Soni, U. Ansari, D. Sharma, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction," International Journal of

Computer Applications, vol. 17, no. 8, 0975 – 8887, 2011.
[36] A. Shillabeer, J.F. Roddick," Establishing a Lineage for Medical Knowledge Discovery," Carnegie Mellon University, Australia, 2007.
[37] L. A. Mei Yin, "Prediction Model for HlNl Disease," University Utara, Malaysia, 2011.
[38] K. Srinivas, B.R. Kavihta, A. Govrdhan, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks," International Journal on Computer Science and Engineering, vol. 02, no.02, 250-255, 2010.
[39] H.C. Koh and G.Tan, "Data Mining Applications in Healthcare," Journal of Healthcare Information Management, vol. 19, no. 2, 2005.
[40] B. Amanuel Dibaba, "Application of data mining techniques to predict household health seeking behavior: the case of butajira rural health project (BRHP)," Master's thesis, Addis Ababa University, Ethiopia, 2010.
[41] W.W. Chapman, J.N Dowling, O. Ivanov, P.H. Gesteland, R.T. Olszewski, Jeremy, et al., "Evaluating Natural Language Processing Applications Applied to Outbreak and Disease Surveillance," Pittsburgh, USA, 2004.
[42] M. Last, R. Carel D. Barak, "Utilization of Data-Mining Techniques for Evaluation of Patterns of Asthma Drugs Use by Ambulatory Patients in a Large Health Maintenance Organization," 2007.
[43] T. Wongstitwilairoong, J. Gaywee, N. Sirisophana, C.J. Mason, J.A. Pavlin, "Mining Pattern Model of Influenza Surveillance," Armed Forces Research Institute of Medical Sciences, Bangkok, Thailand, 2008.
[44] J. Erman, M. Arlitt, A. Mahanti, "Traffic Classification Using Clustering Algorithms, University of Calgary," Canada, 2006.
[45] S.B. Aher, L.M.R.J Lobo, "A Comparative Study of Association Rule Algorithms for Course Recommender System in E-learning," International Journal of Computer Applications, 2012, vol. 39, no. 1, 0975 – 8887, 2012.
[46] P.J. Azeve and A.M. Jorge, "Comparing Rule Measures for Predictive Association Rules," department of informatics, University of Minho and University of Porto, Portugal
[47] J. Shawe-Taylor, T. Diethe, "Analysis of Complex Data," 2009.
[48] D. Soria, J.M. Garibaldi, F. Ambrogi, E.M. Biganzoli, I. O. Ellis, "A 'non-parametric' version of the naive Bayes classifier," Knowledge-Based Systems24, 2011.
[49] M. Bramer, "Principles of Data Mining," Springer-Verlag, University of Portsmouth,UK, 2007.
[50] K. Chaudhary, I. Papapanagiotou, and M. Devetsikiotis, "Flow Classification Using Clustering and Association Rule Mining," Electrical and Computer Engineering, North Carolina State university, Raleigh, USA, 2010.
[51] P. Reutemann, A. Seewald, D. Scuse, "WEKA Manual for Version 3-6-0", 2008.
[52] Federal Ministry of Health (FMoH), "Public Health Emergency Management Guideline," Ethiopian Public Health Institute, Addis Ababa, Ethiopia, 2009.
[53] A. Garcia-Abreu, W. Halperin, I. Danel, "Public Health Surveillance Toolkit, a guide fo busy task managers," World Bank, 2002.
[54] Institute of Environmental Science & Research (ESR), "Disease Outbreak Manual, Porirua," New Zealand, 2002.
[55] WHO, "Outbreak surveillance and response in humanitarian emergencies," Geneva, Switzerland, 2012.
[56] Federal Democratic Republic of Ethiopia Ministry of Health (FDRE MoH), "Public Health Emergency Management Core Process," Ethiopian Public Health Institute, Addis Ababa, 2008.
[57] WHO, "Protocol for the Assessment of National Communicable Disease Surveillance and Response Systems," 2001.
[58] M. Tsehaynesh, "A Five Year, Balanced Score Card Based Strategic Plan (2010-2015G.C)," Ethiopian Health and Nutrition Research Institute, Addis Ababa, Ethiopia, 2010.
[59] U.S. Department of Health and Human Services, CDC, "an Introduction to Applied Epidemiology and Biostatistics," 2nd Edition, Principles of Epidemiology, Atlanta, Georgia, 2005.
[60] S. Gao, "Advanced Health Information Sharing with Web-Based GIS," Department of Geodesy and Geometrics Engineering University of New Brunswick, Canada, 2010.
[61] K. Srinivas, G.R Rao, A. Govardhan, "Survey on Prediction of Heart Morbidity Using Data Mining Techniques," International Journal of Data Mining & Knowledge Management Process (Ijdkp) vol.1, no.3, India, 2011.
[62] Johns Hopkins and Red Cross Red Crescent, "Public health guide in emergencies," International Federation of Red Cross and Red Crescent Societies, 2nd edition, Geneva, Switzerland, 2008.
[63] M. Kamber, J. Han, "Data Mining: Concepts and Techniques 2nd edition," University of Illinois at Urbana-Champaign, 2006.
[64] G. Diansheng, M. Jeremy, "Spatial data mining and geographic knowledge discovery—an introduction," Journal of Computers, Environment and Urban Systems, USA, 2009.