Evaluating the Performance of Classification Algorithms using Ebola Virus Dataset

Kanika Chuchra
Department of Computer Science and Engineering
Chandigarh University, Mohali

Richa Vasuja
Department of Computer Science and Engineering
Chandigarh University, Mohali

Ayesha Bhandralia
Department of Computer Science and Engineering
Chandigarh University, Mohali

ABSTRACT
Highly fatal Ebola virus disease has emerged in Africa and got declared as public health emergency by W.H.O. many humans got infected by the virus so mining of disease is done using WEKA tool to predict whether is died or not by analyzing various symptoms. Various classification algorithms have been used. Further to improve the accuracy rate fusion of algorithm is done using unsupervised filter in MATLAB.

Keywords
Ebola virus, WEKA, big data, classification Algorithms, filter

1. INTRODUCTION
Occurrences of infectious diseases direct to strict disruptions of civilization, not to state the defeat of existence in each and every one its misfortune. The past of the clash beside diseases shows a few successes, like against the epidemicas well as pox, however there are at rest several diseases so as to we comprise not been capable to wipe out, like malaria, tuberculosis, and influenza. Ebola is solitary such unsettled disease [1]. The problem is thinking to exist in fruit bats. Because of this inscription, here are no accepted vaccines or sufficient treatments meant for Ebola disease, even if trials bebeneath way. The bug spreads among humans via get in contact with corporal fluids, like blood, or sweat. Incubation time is long, among one and three weeks. The deficiency of consistent information is a solemn contributing issue to the 2014 Ebola outburst, according to the World Health Organization [2].

1.1 Knowledge representation techniques
In diagnosing Ebola statistics from diverse sources is together in dissimilar statistics sets [3]. It is within combining statistics from dissimilar areas wherever the genuine power of facts knowledge lies: triangulating statistics to advance data superiority, and too, verdict unforeseen model [4]. To be capable to compare objects from dissimilar statistics sets, the data have to be represented in a planned and equivalent way. The ground of awareness illustration studies this feature [5]. It uses techniques like semantic networks as well as computerized differencing to arrange information in taxonomies as well as ontologies [6]. Semantic network techniques for connected open data permit automatic conjecture of varied kinds of statistics, for instance social complex statistics [7]. Communal and semantic system techniques are areas of vigorous study [8]. They apply in serving to analyze Ebola cases illustrates how basic study and actual global challenges be able to go mutually [9].

1.2 Data Science: Three Techniques for Ebola
Three techniques with the purpose of to resolve the Ebola calamity are [10, 11]

- ✓ Gathering of high class statistics plus organizing the information so as to made combinations among statistics sets of a varied structure
- ✓ meant for the study of vast and varied statistics sets, by adaptive techniques used for elevated dimensional statistics sets , also
- ✓ High presentation of drug detection techniques. High presentation techniques are essential since the dimension of the information, especially at what time combinations are prepared, rapidly becomes too giant for normal computers [12].

Dynamic data: Dynamic statistics contains historical information, date is to be created and information will be created, which produce single of a set of information [5].Locate the existing time is S plus the date locate produced previous to the S is chronological information set, denoted D_{old}, information set as of epoch of time earlier than S to S is known as the present information set, denoted $D_{Current}$. Information set created later than the S is known as the subsequent information set, denoted $D_{new.}$ Procedure for using information set $D_{Current}$ D_{old} and D_{new} to take out the facts and regulations in active data source known as dynamic data mining (DDM) [6].

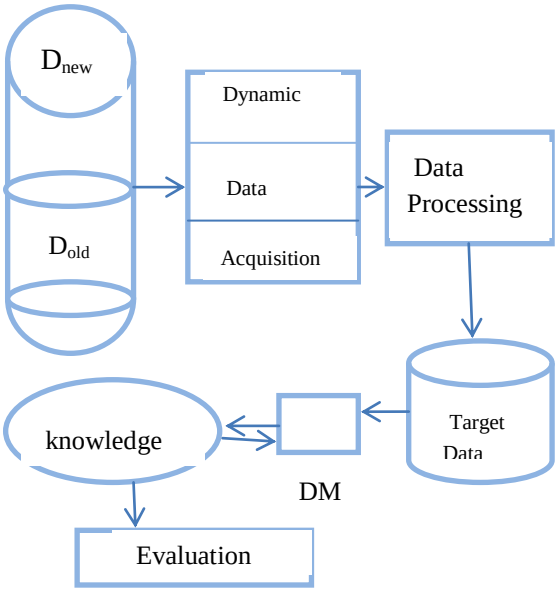

Fig 1. The Process of Dynamic Data Mining

1.3 Ebola and Big data
This execution is to classify the Ebola bug information by reason of continents. The African, European along with North

American continents is being analyzed below the statistics for the reason of categorization of Ebola disease decease rates crosswise the planet [11,15]. The Ebola disease is intensified the result in the African, European as well as North American contents [16] therefore it becomes extremely indispensable to group the information of Ebola disease so as to discover the section wise concentration of the Ebola disease decease rates. Subsequent algorithm is being used for Ebola disease categorization [17].

BIG DATA and its Issues: At present, high amount of priceless data such as network logs, texts as well as credentials, business dealings, banking financial records, economic charts, health images, genetic, life science, as well as communal media information can be simply collected or created from diverse sources, in diverse formats, as well as a high rate in lots of real-life uses in current organizations and civilization [10].

- ✓ Issues connected to characteristics
  - Data Capacity
  - Data Speed
  - Data diversity
  - Data significance
  - Data complication
- ✓ Storage as well as Transfer Issues
- ✓ Data supervision (Structure) Issues
- ✓ Processing issues

2. RELATED WORK

Huan Wang et al. [1] paying attention on investigation for locating a secondary treatment representation based on information pulling out technique which utilize the connection among the treatment outcomes and the features of diverse patients. R et al. [2] gave the assessment about diverse categorization techniques used in predicting the danger stage of each individual based on age, masculinity, Blood stress, cholesterol, pulsation pace. The patient danger level is classify using datamining cataloging techniques like Naïve Bayes, KNN, Decision Tree Algorithm, and Neural Network. etc., Correctness of the peril level is elevated when using additional quantity of attributes.

Modest von Korffet al. [3] said that every ground in vector presented a genetic material and protein. The importance in the ground was resulting from the numeral publications in which this genetic material occurred mutually with the virus term. Virus relations were calculating by vector-similarity computation. Five viruses were examined jointly with their nearby neighbor viruses to explain the authority of our advance.

Chang Sheng et al. [4] meant to sense both single-point mutations and k-mutations in the viral series. We classify the difficulty of mutation series pattern withdrawal and plan algorithms to determine applicable mutation chains. Compressed information structures to assist the withdrawal process and pruning stratagem to boost the scalability of algorithms are work out.

Hongmei Chen et al. [5] document offered an advance for forceful maintenance of estimations w.r.t. substance and quality added at the same time under the construction of Decision-Theoretic Rough Set (DTRS). Correspondence characteristic vector as well as matrix is clears firstly to modernize estimations of DTRS in diverse stages of granularity. After that, information scheme is rotten in subspaces and the sameness characteristic matrix is modernized in diverse subspaces incrementally.

Sunaina Sharma et al. [7] projected information withdrawal approach for the categorization of huge dataset stand on death fee by plague outburst of Ebola disease and evaluate its significance with further epidemic viruses and simplify error and infraclass reparability using relevance vector machine classifier.

Gire et al. [17] represented a largest outburst, and has demonstrated continual human-to-human broadcast afterwards, with no verification of extra zoonotic sources. As several mutations modify protein sequences as well as other organically significant targets, they ought to be monitored meant for effect on diagnostics, vaccines, as well as therapies serious to outburst reply.

3. METHODOLOGY

Ebola virus dataset is loaded into WEKA tool and various classification algorithms likes J48, LMT, random tree, REP are applied and results are evaluated. We analyzed the accuracy rate is how correctly results are predicted In order to improve the results unsupervised filter is used. Data is loaded and unsupervised filter is called in Matlab and results are compared before and after filtering. After using filter there is significant improvement in the results. Further fusion of algorithm i.e. In this we combine the algorithms and results are evaluated. The following table is showing the results .

4. CONCLUSION AND FUTURE SCOPE

In this paper analysis of ebola virus disese is done which is very fatal. By analysis the symptoms analysis is made whether a person will died of ebola virus or not. In this various algorithms are used and further filters are used to improve the results. Fusion of algorithm is done for better classification.

In this future we can use certain algorithms to remove the noise and that will analyse in better way.

Algorithm	TP rate	FP rate	Precision	Recall	F-Measure	ROC Area	Accuracy
J48(Without filtering)	0.849	0.407	0.843	0.849	0.846	0.727	84.8739 %
LMT(without filtering)	0.908	0.282	0.904	0.908	0.905	0.855	90.7563 %
REP(Without filtering)	0.866	0.59	0.864	0.866	0.836	0.77	86.5546 %
RandomTree(without filtering)	0.857	0.405	0.85	0.857	0.853	0.764	85.7143 %
J48(with filtering)	0.941	0.275	0.945	0.941	0.936	0.835	94.1176 %

LMT(with filtering)	0.966	0.157	0.968	0.966	0.965	0.994	96.6387 %
REP(with filtering)	0.933	0.276	0.933	0.933	0.928	0.929	93.2773 %
LMT+RandomTree	0.983	0.078	0.984	0.983	0.983	0.984	98.3193 %
J48+Random Tree	0.971	0.137	0.972	0.971	0.97	0.944	97.0588 %
REP+Random Tree	0.966	0.138	0.967	0.966	0.965	0.966	96.6387 %
LMT+J48	0.954	0.216	0.956	0.954	0.951	0.926	95.3782 %
LMT+REP	0.95	0.217	0.951	0.95	0.947	0.96	94.958 %
LMT+REP	0.937	0.275	0.939	0.937	0.932	0.898	93.6975 %

5. REFERENCES

[1] Wang, Huan, et al. "A medical support treatment model based on data mining." Computer Science & Education (ICCSE), 2015 10th International Conference on.IEEE, 2015.

[2] Thomas, J., and R. Theresa Princy. "Human heart disease prediction system using data mining techniques." Circuit, Power and Computing Technologies (ICCPCT), 2016 International Conference on.IEEE, 2016.

[3] Von Korff, Modest, Bernard Deffarges, and Thomas Sander. "Data Mining in MEDLINE for Disease-Disease Associations Via Second Order Co-Occurrence." *Computational Intelligence, 2015 IEEE Symposium Series on*.IEEE, 2015.

[4] Sheng, Chang, et al. "Mining mutation chains in biological sequences." *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*.IEEE, 2010.

[5] Chen, Hongmei, et al. "A decision-theoretic rough set approach for dynamic data mining." *IEEE Transactions on Fuzzy Systems* 23.6 (2015): 1958-1970.

[6] Jingxin Du, Jun Zhou et al. '*An Overview of Dynamic Data Mining'* 2016 3rd International Conference on Informative and Cybernetics for Computational SocialSystems (ICCSS)

[7] Sharma, Sunaina, and VeenuMangat. "Relevance vector machine classification for big data on Ebola outbreak." *Next Generation Computing Technologies (NGCT), 2015 1st International Conference on*.IEEE, 2015.

[8] Team, WHO Ebola Response. "Ebola virus disease in West Africa—the first 9 months of the epidemic and forward projections." *N Engl J Med* 2014.371 (2014): 1481-1495.

[9] Leroy, Eric M., et al. "Fruit bats as reservoirs of Ebola virus." *Nature*438.7068 (2005): 575-576.

[10] Team, WHO Ebola Response. "West African Ebola epidemic after one year—slowing but not yet under control." *N Engl J Med* 2015.372 (2015): 584-587.

[11] Sullivan, Nancy J., et al. "Development of a preventive vaccine for Ebola virus infection in primates." *Nature* 408.6812 (2000): 605-609.

[12] Simmons, Graham, et al. "DC-SIGN and DC-SIGNR bind ebola glycoproteins and enhance infection of macrophages and endothelial cells." *Virology* 305.1 (2003): 115-123.

[13] Feldmann, Heinz, and Thomas W. Geisbert. "Ebola haemorrhagic fever." *The Lancet* 377.9768 (2011): 849-862.

[14] Leroy, Eric M., et al. "Multiple Ebola virus transmission events and rapid decline of central African wildlife." *Science* 303.5656 (2004): 387-390.

[15] Baize, Sylvain, et al. "Emergence of Zaire Ebola virus disease in Guinea." *New England Journal of Medicine* 371.15 (2014): 1418-1425.

[16] Sullivan, Nancy J., et al. "Accelerated vaccination for Ebola virus haemorrhagic fever in non-human primates." *Nature* 424.6949 (2003): 681-684.

[17] Gire, Stephen K., et al. "Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak." *science* 345.6202 (2014): 1369-1372.