Journal of Information Technology and Education Science

JITES

Published by:
University of Computer Studies (Taungoo)

December, 2020
Volume-02, Issue-01

JOURNAL OF INFORMATION TECHNOLOGY AND EDUCATION
SCIENCE VOL. 02, ISSUE. 01, 2020

ISBN: 978-99971-0-928-6
Publisher: University of Computer Studies (Taungoo)

University of Computer Studies (Taungoo)
Yangon-Mandalay Road,
Taungoo Township,
Bago Region, Myanmar.

JOURNAL OF

INFORMATION TECHNOLOGY AND EDUCATION SCIENCE

JITES2020 VOL. 02, ISSUE. 01, 2020

UNIVERSITY OF COMPUTER STUDIES (TAUNGOO)

Journal of Information Technology and Education Science

Vol. 02, Issue. 01, 2020

Content

Information Technology

1. Analysis of Routing Protocols over TCP in Mobile Ad-hoc Networks using Random Way Point Model 1
Khaing Su Wai, Hsu Mon Maung

2. Multilevel Association Rules Mining using Apriori Algorithm 7
Nang Khin Pyone Myint, Thin Nu Nu Win, Nang Sabae Phyu

3. Genetic Algorithm-Based Feature Selection and Classification of Breast Cancer Using Bayesian Network Classifier 14
Yi Mon Aung, Nwet Nwet Than, Linn Linn Htun

4. Sentiment Analysis of Students' feedback from Coursera Online Learning Using Bernoulli Naïve Bayes Classifier 20
Nilar Htun, Nang Seint Seint Soe

5. Detection of Diabetes Using Classification Methods 26
Phyu Thwe, Cho Cho Lwin, Hnin Pwint Myu Wai

6. Performance Analysis of Classification Algorithms in Data Mining Technique 31
Aye Mon Win, Yu Yu Khaing, Lei Yi Htwe

7. Introduction to Popular Hadoop Analytics Tools for Big Data 36
Aye Myat Nyo, Day Si Win, Khin Nyein Myint

8. Documents Retrieval using Solr 41
Ei Marlar Win, Shwe Thinzar Aung

Education Science

9. An Investigation into Non-verbal Intelligence of Primary Students 46
Su Myat Aye , Naing Naing Maw

10. An Investigation into Adolescents' Emotional Creativity and the Influence of Personality Traits on it 50
May Oo Mon, Su Myat Aye, Naing Naing Maw

11. An Investigation into Personality and Career Interest of High School Students 58
May Wah Linn, Naing Naing Maw

12. Analysis of Gender and Grade Differences on Abstract Reasoning Test for High School Students 66
Phue Wai Ko Ko, Su Myat Aye, Naing Naing Maw

13. Development of Numeracy Test For Adolescents 73
Myat Yu Yu Mon , Su Myat Aye

14. Student-Centered Approach is more Effective than Teacher-Centered Approach on Mathematics 80
San San Nwe, Hla Yin Moe, Lin Lin Aye

15. Evaluating the Effect of Environmental Education and Awareness on Solid Waste Management within Elementary Students 85
San Myint Yi

Computational Mathematics

16. Generation of Orthogonal Polynomials in Least-Squares Approximations 91
Tin Tin Maw

17. Applying the Queue Theory in Bank Service Centers 98
Khin Moh Moh Thin, Hla Yin Moe, Lin Lin Aye

18. Optimizing Integer Programming Problem Using Branch and Bound Method 103
Zin Nwe Khaing, Hla Yin Moe, Aye Mya Mya Moe

Natural Science

19. Design And Construction of Android Phone Controlled Electrical Household Appliances 109
Thet Lwin Oo , Win San Win, Hnin Ei Maung, Lwin Lwin Soe

20. Construction of Motor Control System Using Fingerprint Sensor 115
Hla Thein Maung, Aung San Min, Aye Aye Khine

21. Study on Neutron-proton Scattering using CD-Bonn Potential 121
Thandar Kyi, Htun Htun Oo

22. Characterization of Advanced Superionic Conducting Materials of Lithium Cobalt-Nickel Oxides for Solid Oxide Fuel Cell (SOFC) Application 128
Aye Aye Lwin, Win Kyaw, Thaik Thaik, San San Wai

23. Scanning System for Light Transmission of Glasses 133
Shwe Zin Aung

Language

24. The Concept of Speech Function in the Articles in the Reader`s Digest 139
Aye Lae Maw

25. Grammatical Parallel Structure In Inaugural Addresses By George W. Bush 143
Myat Theingi Kyaw, Wai Wai Phyo

26. Reducing Students' Anxiety in Learning Reading Passages through Reciprocal Teaching 149
Naing Naing Maw

27. Grammatical Collocations Found In The Selected Academic Texts 155
Wai Wai Phyo, Myat Theingi Kyaw

28. Illocutionary Acts Of The Main Character’s Utterances In The Movie “Beyond Rangoon” 161
Moe Ei Swe, Mar Mar Lwin, Than Than Sint

29. A Survey On Preferences Towards Learning Styles Of Undergraduate Arts And Science Students 166
Tint Tint Ei, Mar Mar Lwin, Hnin Lae Win

30. A Study of Reading Comprehension Questions From Matriculation English Question Papers By Using Bloom's Taxonomy 171
Aung Kyi, Zaw Moe Aung

31. ရှင်ဥတ္တမကျော်တောလားလာ ရာသီဘွဲ့၏ သဘောသဘာဝများ လေ့လာချက် 177
Khin Pyone, Phyu Phyu Khaing

32. မြိတ်ဒေသိယစကားကို အတ္ထဗေဒရှုထောင့်မှလေ့လာချက် 181
Shine Maung Maung

33. တရုတ်သံရောက်မော်ကွန်းမှ နန်းဓလေ့ယဉ်ကျေးမှုများ 186
Wai Wai Tin

34. မြန်မာသဒ္ဒါသမိုင်းကြောင်းလေ့လာချက် 191
Khine Khine

35. မြန်မာဝါကျရိုးဖွဲ့စည်းပုံရှိ စကားမြှုပ်စနစ်နှင့်ဘာသာဗေဒအမြင် 198
Ye Ye Cho

36. ငြိမ်းချမ်းချိုးဖြူဆက်၍ကူ"ကဗျာမှဘာသာစကားတာဝန်များစိစစ်ချက် (လူမှုဘာသာဗေဒ) 204
Yu Yu Tun, Aye Thandar Win, Htar Hlaing Soe

37. ဦးပုည၏ တေးထပ်များမှ နှစ်သက်မှုသက်သက်ကိုသာပေးသော ရသမြောက်အဖွဲ့များ လေ့လာချက် 209
Khin Khin Maw

38. တ-ဝဂ်အတွင်းရှိအသံစွဲစကားလုံးများ၏အနက်ကိုစိစစ်ခြင်း 215
Than Than Naing

39. ကယန်း (ပဒေါင်) တိုင်းရင်းသားတို့၏ ကြေးပတ်ခြင်းဓလေ့နှင့် ဘာသာစကားသဘောလက္ခဏာ 220
Mya Mya Win

40. ပန်းမွေ့ရာ ရွှေကော်ဇော်ဝတ္ထုတိုကန့်သတ်သိ ရှုထောင့်များ 225
Phyo Wai Hlaing

41. ဦးကြီး၏လွမ်းချင်းကဗျာများမှ မြန်မာ့ကျေးလက်ဓလေ့များ 230
Sandar Shein

42. နည်းပညာဆိုင်ရာဝေါဟာရများနှင့် မော်ဒန်ကဗျာ 235
Zaw Naing

43. မောင်ချောနွယ်၏ရထားကဗျာပေါင်းချုပ်မှုနိမိတ်ပုံအသုံးများလေ့လာချက် 240
Nyein Ei Lwin, Thi Thi Swe, Khaing Khin Aye

44. မိုးမိုး(အင်းလျား)၏ "မေတ္တာကမ်းနားအချစ်သစ်ပင်"ဝတ္ထုမှ ဇာတ်ဆောင်စရိုက်ဖန်တီးမှု အတတ်ပညာ 244
Khaing Khin Aye, Thi Thi Swe, Nyein Ei Lwin

45. ကာတွန်းများမှ ထုတ်ဖော်မပြောသောအနက် 250
Shwe Sin Win, Khin Htwe Myint, Zar Zar Thin

Information Technology

Analysis of Routing Protocols over TCP in Mobile Ad-hoc Networks using Random Way Point Model

Khaing Su Wai[1], Hsu Mon Maung[2]
[1,2]*Myanmar Institute of Information Technology, Mandalay*
[1]*khaingsuwai3919@gmail.com*, [2]*hsumon77@gmail.com*

ABSTRACT: *Mobile ad hoc network, MANET is an impartial disseminated wi-fi structure inclusive of unbound nodes due to the fact, each node may communicate with each other at random, and then, redirecting data like a router for exclusive nodes. Packet forwarding in MANET is challenging assignment, furthermore, the presence of malicious nodes makes the general platform very insecure and the unpredictable existence of the changing nodes adds complexity. Shifting of the nodes has massive influence at the network performance. This paper offers the overall performance comparison between dynamic source routing, DSR, ad hoc on demand distance vector routing, AODV and destination sequenced distance vector, DSDV [7]. And then, precisely determine which routing mechanism is greater powerful. The objective of this paper is to review routing mechanisms in MANET to get a perfect performance of the factors influencing the actual quality among these network applications. This analysis of routing mechanisms is useful in know-how of the requirements and challenging circumstances for routing mechanisms in MANET and procedures relating to premise of developing a new routing protocol which we expect to supply in the future. The overall performance measurement of three routing protocols using the random way point mobility over tcp was performed and assessed the measurement of those protocols in phrases of the window size of tcp, packet loss, jitter, average throughput, average delay and packet delivery ratio with regard to the alterable number of nodes. In this paper, we are able to simulate MANET using network simulator NS2 and then make a result-primarily based assessment by using NS2 visual trace analyzer.*

Keywords: MANET; AODV; DSDV; DSR

1. INTRODUCTION

MANET generally is a sequence of freedom to integrated progressive wireless nodes' framework because there are no a centralized rule or platform. These dynamic nodes are more effective either directly or with the aid of intervening nodes to establish a correspondence with each other due to there is no core management using radio connections or wireless multi-hop networks. Such systems are called dynamic topology. This network topology is not rigid and the entire nodes work as access points because there is no the need of any base stations. Home and industrial communication, military environments, moving vehicles, and IOT systems are some applications where the ad hoc service is used. The routing mechanisms for ad hoc networks are broadly categorized into three classes based upon the update mechanism of the routing information: proactive, reactive, and hybrid. Each node updates and retains the routing tables in proactive protocols to keep track of all potential targets for the instant availability of the routes for future use. DSDV is the sample of proactive protocol. Reactive protocols set up routes only when routes are essential by an initial node and every node sustains individual routing data to targets but would not own an exhaustive topological view of the system. In reactive protocols routes are found on-call and for locating a route to target, a route request is initiated. The samples of reactive protocols are DSR and AODV. The main purpose of this paper is to post a detailed analysis of MANET protocols. In this analysis, we elected three routing protocols AODV, DSR, and DSDV then, compared their results. First we illustrate the particular features of routing protocols for MANET and describe their strengths and limitations. We used network simulator NS2 for simulations and it is the most common wireless network testing simulator supporting many MANET routing protocols. Ns2 visual trace analyzer has used to assess the performances of these protocols. The parameters are window size of tcp, packet loss, jitter, average delay, average throughput, and packet delivery ratio. The analysis is reviewed by simulating networks with disparate variables of nodes, tcp traffic, and random way point mobility model. The remaining paper is standardized as follows: section 2 discusses the work related to MANET routing simulations, section 3 briefly describes the basic MANET architecture and the features of each of the three routing protocols, the performance assessment metrics are discussed in section 4 and the results of the study are analyzed, and lastly, section 5 terminates the paper.

2. RELATED WORK

This section discusses the reviews related to MANET routing protocols by researchers. H. Ehsan and Z.A. Uzmi [3] contrasted AODV, DSR, DSDV and TORA to ad hoc routing protocols. Their research indicates that DSR is outperforming other routing protocols due to its capacity to efficiently use caching and support various paths to target. S. S. Tyagi and R. K. Chauhan conducted a related survey [8]. They analyze protocols using PDR, average delay, packet loss, and overhead routing. The number of nodes, speed, time of pause and time of simulation varies. In large environments, they reason that AODV performs better than DSR, and that both AODV and DSR perform better than DSDV. The authors undertake a related survey in

[6]. They assess similar protocols by shifting the quantity of sources, the pause time, the quantity of nodes and speed. They presume that in high mobility scenarios AODV and DSR perform better than DSDV, and that AODV beats DSR in higher load scenarios. N.Vetrivelan and A V Reddy[4] assess average delay, fraction of packet delivery and load of routing for AODV , DSDV and TORA. They shifted the quantity of nodes and held up to 100sec simulation time. Their findings show that AODV outperforms the other two routing protocols as far as average delay is concerned but TORA provides better performance in terms of packet transmission fraction and DSDV performs best in less stressful situations. DSDV performs best in less distressing circumstances. DSDV is best in upsetting conditions followed by TORA for the standardized routing load.

3. MANET AND ROUTING PROTOCOLS

3.1. MANET

MANETs are very useful when networks dependent on the infrastructure are not accessible, inefficient or costly. That is not always feasible to set up fixed access points and backbone networks. MANET is a network where no wireless or cellular networking infrastructure exists. MANETs require no backbone infrastructure support. Instead MANET is a network in which frequent host, frequent movement, topology changes and wireless multi-hop links occur. Any applications of MANET are smart phone, laptop, wrist watch, military environments, vehicles, aircraft, civil environments, taxi-cab network, conference rooms, sports facilities, emergency operations. Mobile nodes can play the features of hosts and routers in this environment and are free to transfer and manage arbitrarily. The mobile nodes within the radio range can interact directly with one another. In this environment, data must be routed across intermediate nodes. These networks are fully dispersed, can be established anywhere and at any time and then contribute access to message and resources without any infrastructure support. Some of MANET's challenges are packet loss because of transmission faults, variable capacity link, frequent disconnections or

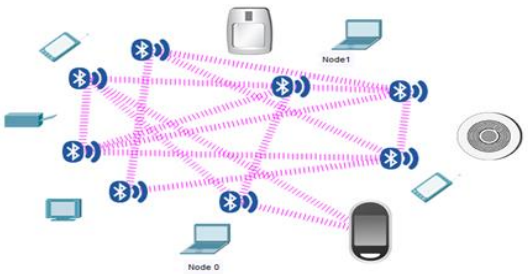

Figure 1. Our model for MANET

partitions, restricted communication bandwidth, communications broadcast nature, mobility-imposed constraints, dynamically changing topologies, paths lack of system and application mobility awareness. Figure 1 demonstrates our model for MANET in this paper.

3.2. AODV

AODV is one of a classification of demand-driven routing protocols intended for use in MANETs [1]. The mechanism for route discovery is only requested when a node wants to transmit information to a different node. And this protocol is reactive. For every route entry, AODV utilizes a destination sequence number. Considering the option between two routes to a target, a requesting node always chooses the one with the highest number of sequences. To discover and preserve connections, the protocol uses various messages. If a source node needs to discover a route to target, it will transmit a route request, RREQ message to the whole network [9]. When an RREQ message reaches a target, the target will send a route replies, RREP message by unicasting back to the source route and then the target route is made available. An interceding node can also respond with an RREP message if the route to the target is fresh sufficiently. As the RREP propagates again to the source node and their routing tables are modified by interceding nodes. Nodes of an active route can deliver connectivity information to their instant neighbors by periodically transmitting local Hello messages. If Hello messages prevent originating from a neighbor within a given time interval, it is presumed the communication is fail. When a node observes that a path to a neighbor is presently false, it eliminates the routing entry and transmits a route error, RERR message to active neighboring nodes the utilization of the path.

3.3. DSR

The DSR protocol is a reactive or on-demand routing mechanism designed for wireless communication systems [1]. In DSR, routing entries at the interceding nodes are not organized and use the source routing mechanism. The source of a packet specifies the full sequence of a route that packets of data are forwarded. A route discovery packet, RREQ is transmitted to all neighbors by the source. This packet involves the target host address, the source address, a route record field and a unique identifier. Every node that receives this packet retransmits it, except it is the target or there is a path in its cache to the target. Only when the RREQ message hits the final destination, it will send back an RREP message to source by going backward. The route that RREP packet conveys back is stored for later use at the source. If any connection on a source route is broken, a Route Error (RERR) packet is used to alert the source node. The source eliminates any route utilizing this connection from its store. The routes to some random node are stored at the source in a route cache and hence routing loops can't be made as they will be detected immediately. This protocol diminishes transmission capacity squandered in remote systems which the control packets and erases the periodic routing table update messages.

3.4. DSDV

The DSDV routing protocol is a table-driven routing scheme intended for MANETs. Every node has a routing table showing the next hop and quantity of hops to the target and regularly forwarding the routing table to neighbors [1]. A sequence number is utilized to label each route when every node advertises its own routing information to each neighbor, and routes with greater number of sequences are more desirable. Moreover, the one with better metric is more desirable among two routes with an equal number of sequences. In the event that a node identifies that a path to a target has fallen, then it will set its hop number to infinity and change its sequence number. Information about new paths, broken connections, metric change is propagated to neighbors immediately. By exchanging that refurbished routing information, every node updates its own routing tables. Because of there is some alteration, and all nodes share the routing information changes, the overhead is more encouraged in DSDV protocol.

4. RESULT AND DISCUSSION

Table 1. Simulation parameters

Parameter	Value
Routing Protocols	AODV, DSR, DSDV
Simulation Duration	150 seconds
Number of Nodes	10,30,100,150
Simulation Area	950 X 700 meters
Antenna	Omni-directional
MAC	IEEE802.11
Traffic Agent	TCP
Traffic Type	FTP
Packet Size	512 bytes
Channel Type	Wireless
Propagation Model	Two ray ground reflect
Mobility Model	Random Way Point
Node 0 position	(5,5)
Node 1 position	(490,285)
Mobility Speed	3m/s

We tend to compare the performance of AODV, DSR and DSDV over TCP in ad hoc wireless networks using Random way point mobility regarding window size of tcp, throughput, average delay and packet delivery ratio, whereas, variable the network size. The source is node 0 and also the destination is node 1. As shown in table 1, the first location of node 0 and 1 are severally (6,6), (489,284) and also the z coordinate is 0. At time 1s, node 0 begins to move towards point (249,249) at a speed of 3 m / sec and node 1 also starts to move towards point (44, 258) at a speed of 3 m / sec. The other nodes are randomly moved by Random way point mobility at time 1s and a speed of 3 m/sec. A tcp link is initiated between node 0 and node 1 at time 1s, using a routing protocol and also the IEEE802.11 mac protocol. During this tcp

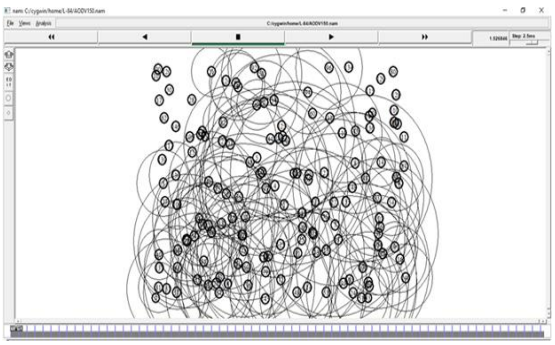

Figure 2. Simulation of our model

protocol, our application is file transfer protocol from source to destination. The form of channel will be set to wireless internet. The Two Ray Ground model is designed to be radio-propagation. The function of the network interface is set as Wireless. The mac type is set to suit in IEEE protocol 802 11 mac. The interface queue type for AODV and DSDV is ready to be Queue/DropTail/PriQueue. DSR type of interface queue is configured to be CMUPriQueue. The layout antenna is designed to be OmniAntenna. The maximum packet is set to 50 in interface queue. We will compare multiple protocols using different sets of mobile nodes. The small variety of node is configured as 10, the medium range of node is configured as 30 and the large wide variety of node is configured as 100 and 150. AODV, DSDV, and DSR are set to the routing protocol. The topography dimension X is set at 950. The topography Y-dimension is set at 700. Simulation end time is set to be 150s. Figure 2 illustrates our simulation environment in this analysis.

Window size of tcp is the size of the receiver's buffer which will influence the flow of transmission. To evaluate the window size of tcp for each protocol is based on the total number of TCP transferred packets.

Table 2. AODV result of simulation

	AODV			
	10 nodes	30 nodes	100 nodes	150 nodes
Generated packets	3801	5298	7717	5944
Lost packets	38	15	48	33
Transferred packets	3763	5283	7669	5911
Jitter	0.013072 s	0.024729 s	0.009754 s	0.013046 s
Average Delay	0.231268 s	0.193891 s	0.184348 s	0.251609 s
Average Throughput	17 KB/s	26 KB/s	28 KB/s	21 KB/s
Packet Delivery Ratio	1.00%	0.28%	0.62%	0.56%

Table 3. DSR result of simulation

	DSR			
	10 nodes	30 nodes	100 nodes	150 nodes
Generated packets	8102	9276	9181	9036
Lost packets	18	8	8	3
Transferred packets	8084	9268	9173	9033
Jitter	0.010714 s	0.011089 s	0.009848 s	0.009973 s
Average Delay	0.164771 s	0.204117 s	0.197196 s	0.190531 s
Average Throughput	40 KB/s	34 KB/s	33KB/s	33 KB/s
Packet Delivery Ratio	0.22%	0.09%	0.09%	0.03%

Table 4. DSDV result of simulation

	DSDV			
	10 nodes	30 nodes	100 nodes	150 nodes
Generated packets	6800	7772	7316	6625
Lost packets	20	12	9	32
Transferred Packets	6780	7760	7307	6593
Jitter	0.008149 s	0.007486 s	0.008523 s	0.010002 s
Average Delay	0.123940 s	0.134735 s	0.132913 s	0.146609 s
Average Throughput	41 KB/s	40 KB/s	38 KB/s	34 KB/s
Packet Delivery Ratio	0.29%	0.15%	0.12%	0.48%

The protocol that can transmit most packets has the best window size. Table 2,3,4 demonstrate the effect of network size on the total number of TCP transferred packets between AODV, DSR and DSDV routing mechanisms, separately. Comparing these total transferred packets, it is easy to know that in the DSR, it has transferred most packets. DSR protocol most suits highly mobile systems among the DSDV, DSR, and AODV. Because according to the simulation results shown in Figure 3, the DSR can transmit most packets in scenario, which is best to highly mobility systems. So the windows size evolution of DSR is better than other protocols when the size of network is enormous. There is negligible impact on window size in DSDV protocol as the size of network extends. But the number of transferred packets is slightly decrease when the size of network increments. The total amount of TCP transferred packets on AODV is less than the others, but increases when number of nodes is enlarged up to 100. We presume that window size in DSR outflanks the other two protocols when organization size is huge.

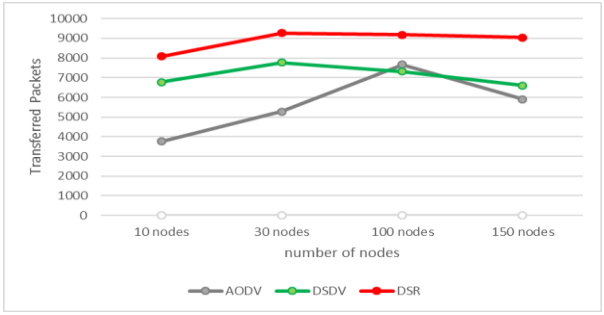

Figure 3. Transferred packets

Packet loss occurs when one, or extra packets of data traveling throughout a network fail to attain their destination node. Packet loss is measured with respect to packets sent as a percentage of packets lost. If a path to the destination is not available or the buffer that stores pending packets is complete, a packet may be dropped at the source. If the connection to the subsequent hop is broken, it can also be dropped on an intermediate host. Wireless link transmission errors, host mobility, traffic load and buffer overload (congestion) are key causes for packet loss in mobile ad hoc networks. Protocol efficiency will improve if the loss of the packet is low. From Table 2,3,4, AODV has higher packet loss for a few nodes, and the packet loss is significantly decrease up to 30 nodes then the packet loss is highest when the size of network set to 100. But, the packet loss is moderately decrease when the size of network is huge. DSR has higher packet loss for a few nodes. But the packet loss is significantly decrease when the size of network enlarges. The packet loss in DSDV is slightly decrease when the size of network grows. But the packet loss is significantly increase when the size of network enlarges. DSDV outperforms AODV due to the fact the packet loss for DSDV is much less than AODV. In AODV, the packet loss is higher than the other two protocols. In DSR, the packet loss is less than that of AODV and DSDV for all different sets of mobile nodes. So, DSR is the most efficient option at packet loss metric. DSR is the most suitable protocol for real-time applications where packet loss is an important consideration.

Jitter is a latency that varies over time, or when packets are not sent in the same order. Jitter is the variation in the time of arrival of the packet in another phrase. There are no variations or jitters in a network with constant latency. The packet jitter is expressed as an average of the network's mean latency variation. The performance of protocol is better efficiency if the latency between various packets is short. From the results in table 2,3,4, AODV has the highest jitter when the network size

is 30 and the lowest jitter when the network size is set to 100. In DSR, jitters are higher for small network sizes. But as the network grows, jitter of DSR decreases. The jitter in DSDV is lower on 10 to 100 nodes. But, there is minimal increase in jitter when the size of network set to 150. However, DSDV gives better jitter performance than AODV and DSR. With the different range of nodes, DSR is more fitting than AODV. AODV shows higher degree of jitter than that of the other protocols.

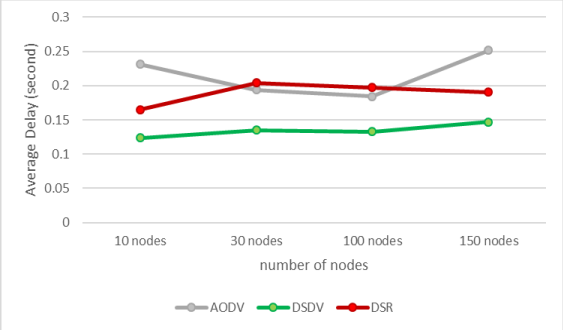

Figure 4. Average delay

The average delay can be characterized as the average time it takes for data packets to reach the destination through the network from the source [5]. It consists of the queue in the transfer of data packets, the delay caused by route discovery process, MAC retransmission delays, packet propagation and transfer times. In a routing protocol, a lower average delay value represents powerful protocol, fast route convergence, and packets transiting the optimal path. Table 2,3,4 display the effect of network size on average delays for AODV, DSR and DSDV, respectively.

Figure 4 displays average delay outcomes of our simulation for the routing protocols. It indicates that AODV has higher delay for a small number of nodes and the delay is moderately decrease up to 100 nodes then the delay starts significantly increase when the size of network increases. DSR has higher delay for a small number of nodes due to the delay of DSR significantly increases when the network size is extended to 30 nodes. But the delay is slightly decrease when the size of network enlarges. DSR become more suitable in a large number of nodes compared to AODV. DSDV protocol outperforms the other two routing protocols because it has lower delay. But the delay is slightly decrease when the size of network grows. In average delay metric, AODV gives worst form comparing with other protocols and is suitable for medium sized network. DSDV is suitable for applications where delay is an important consideration.

Throughput is the proportion of how quick we can really send packets through the network [2]. The quantity of packets that are sent to the destination gives network throughput. The proportion of the aggregate sum of data that a source gets to a destination to the time it takes for the destination to get the final packet is known as throughput. The efficiency is better when it's higher throughput. It represents to a powerful throughput network. For AODV, DSR and DSDV routing protocols,

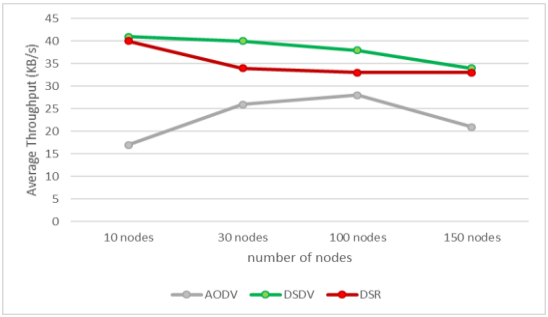

Figure 5. Throughput

respectively, Table 2,3,4 display the effect of network size on the throughput.

From Figure 5 we noticed the DSDV has higher throughput, AODV has lower throughput, and both of DSDV and DSR act the same performance. Throughput in DSDV and DSR declines moderately as a number of nodes increase, but the better impact is observed in AODV where throughput increases appreciably as network size increases. But the throughput in AODV is slightly decrease when the size of network is over 100 nodes. We presume that DSDV outflanks the other two routing protocols and it is generally appropriate for small networks. For large networks, DSR is generally suitable because throughput in DSR does not decline at over 100 nodes.

The packet delivery ratio is the ratio of packets delivered successfully to the destination to the packets generated by the source. It reflects the success rate of packet transmission that is in an exceedingly given period, what percentage packets out of the overall packets that were transmitted can reach the destination. It is a process of packet loss because of route congestion, network queuing delays, and efficiency of routing algorithms. An effective routing protocol guaranteeing a large proportion of the packet transmission. Performance is higher when the delivery ratio for the packets is closer to one. Table 2,3,4 demonstrate the effect of the packet delivery ratio for the routing protocols AODV, DSR and DSDV, severally.

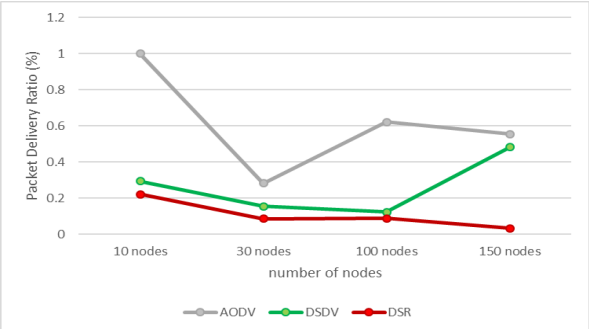

Figure 6. Packet delivery ratio

From Figure 6, we know that AODV has higher PDR for small number of node, and it considerably declines at 30 nodes and considerably increases at 100 nodes. However, AODV has had better packet delivery ratio than DSDV and DSR. Packet delivery ratio of DSR is extremely less than compared to AODV and it slightly

declines as network size increases. Each of DSDV and DSR act a similar performance on packet delivery ratio and DSDV declines slightly as a number of nodes increase. But the better impact is observed in DSDV where packet delivery ratio increases appreciably as network size increases over 100 nodes. We have a tendency to conclude that AODV achieves the best packet delivery ratio performance but, if we consider the impact of large network size, DSDV also achieves optimum performance.

5. CONCLUSIONS

In future networking, wireless networks are anticipated to work an essential role. Because of common properties with respect to connection characteristics, node mobility and alterable network size, routing protocols in wireless networks are complicated whenever contrasted with wired networks. There are a variety of routing protocols being approved in ad hoc wireless networks that are absolutely different within the results from one another. The performance comparison between DSR, AODV and DSDV has been proposed in this paper to verify exactly which protocol is more powerful. We have used ns2.35 for simulations. NS2 visual trace analyzer has used to assess the performance of these protocols with regard to the variable number of nodes in relation to the performance metrics. The simulation results for window size in tcp, packet loss, jitter, throughput, average delay and packet delivery ratio show that with increase in networks size, Random way point mobility model and transmission control protocol as type of traffic. From the analysis of the graphs obtained from the simulation of the protocols shows that, window size in DSR outperforms than the other two routing protocols when the network size is huge. The most successful option for packet loss metrics is DSR, since the packet loss for all separate sets of mobile nodes is lower than that of AODV and DSDV. In jitter performance, DSDV achieves greater effectiveness than AODV and DSR. At average delay metric, DSDV is ideal for applications wherever delay is a critical factor. In throughput metric, DSDV outflanks the other two routing protocols and it is especially perfect for smaller networks. DSR is typically ideal for large networks because throughput in DSR does not decline to more than a hundred nodes. AODV achieves the most efficient results on the packet delivery ratio. If we prefer to observe the combined effect of network size, window size in tcp, throughput, average delay and packet delivery ratio, DSR is the most efficient option for large networks.

REFERENCES

[1] Cana, Erion "Comparative Performance Simulation of DSDV, AODV and DSR MANET Protocols in NS2," International Journal of Business and Technology: Vol. 2 : Iss. 1, Article 4, 2013. <https://knowledgecenter.ubt-uni.net/ijbte/vol2/iss1/4>

[2] Gouda, B. S., C. K. Behera, and R. K. Behera. "A scenario based simulation analysis and performance evaluation of energy efficiency enhancement of routing protocols in MANET", International Mutli-Conference on Automation Computing Communication Control and Compressed Sensing (iMac4s), 2013

[3] H. Ehsan and Z. A. Uzmi, “Performance comparison of ad hoc wireless network routing protocols,” Proc. of the 8th International Multitopic Conference (INMIC 2004), Pakistan, Dec 2004, pp. 457-465.

[4] N. Vetrivelan, and A. V. Reddy, “Performance analysis of three routing protocols for varying MANET size,” Proceeding of the International Multi Conference of Engineers and Computer Scientists Vol. II, (IMECS ‘08), Hong Kong, March 2008, pp. 19-21.

[5] Piyush Yadav, Rajeev Agrawal, Komal Kashish. "Performance Evaluation of ad hoc Wireless Local Area Network in Telemedicine Applications", Procedia Computer Science, 2018

[6] Shah, S., Khandre, A., Shirole, M., Bhole, G.: Performance Evaluation of Ad Hoc Routing Protocols Using NS2 Simulation, Mobile and Pervasive Computing (CoMPC-2008).

[7] Sunil Kumar Singh, Rajesh Duvvuru, Jyoti Prakash Singh. "Chapter 90 Performance Impact of TCP and UDP on the Mobility Models and Routing Protocols in MANET", Springer Science and Business Media LLC, 2014

[8] Tyagi, S.S., Chauhan, K.R.: Performance Analysis of Proactive Routing Protocols for Ad hoc Networks. International Journal of Computer Applications (0975 – 8887),2010

[9] Vandana Dubey, Nilesh Dubey. "Performance Evaluation of AODV and AODVETX", 2014 International Conference on Computational Intelligence and Communication Networks, 2014

Multilevel Association Rules Mining using Apriori Algorithm

Nang Khin Pyone Myint[1], Thin Nu Nu Win[2], Nang Sabae Phyu[3]
[1,2,3]*Computer University, Pinlon*
[1]*nannkhinpyonemyint@gmail.com*, [2]*thinnunuwin@ucspinlon.edu.mm*, [3]*nangsabaephyu@gmail.com*

ABSTRACT: *Association rule mining is a significant component of data mining. Multilevel association rules give the more accurate and specific information. The system that searches association rules from medical store dataset is implemented. This system, in a specified transaction dataset, a model of mining multilevel association rules that accepts the various minimum supports for each step, finds multilevel association rules. To seek multilevel association rules in a describe transaction dataset, it is used encode taxonomy and dissimilar minimum supports and confidences. To search frequent itemset in each level, Apriori algorithm is used, and then fast algorithm generates association rules in this model. Experiments are conducted using medicine store transaction dataset. Execution time of the system is compared with that without using encode taxonomy. The experimental results show that generating association rules using the encode taxonomy is faster than generating association rules without using encode taxonomy.*

Keywords: multilevel association rules; apriori algorithm; frequent itemset; fast algorithm

1. INTRODUCTION

Among the most important branches of data mining, the association rule is a technique for data mining that allows conditional statements to be formulated, such as if clients purchase item x, then they also purchase item y. the purpose of association rule mining is to find significant relationships between items, so that the existence of certain items in a transaction would indicate the presence of another items. Agrawal and colleagues proposed multiple mining algorithms relying on the principle of huge item sets to discover association rules in transaction data to achieve this goal [1]. They separated the mining process into two steps. First step, frequent (large) itemset can be discovered depend on the counts by examining the transaction data. Second step, association regulations were caused from the large itemset discovered in the first.

Basically association rules follow Apriori principle. Apriori algorithm is one of the fastest known data mining algorithms to find *all* frequent itemset in a large database, i.e., all sets are contained in at least *minsup* transactions from the original database [2]. Apriori algorithm analyzes in a bottom-up, breadth-first search approach. The calculation begins from the smallest collection of frequent itemset and shifts upward until the largest frequent itemset is reached. Apriori is a modern algorithm for mining frequent itemset and learning association rules of single level.

The purpose of this system is to help the managers or owners of medical stores for to do selective marketing and planning their shelf spaces. This paper is also focused on the generation of association rules in multilevel from medicine transactions. In the following section, multilevel association rules are introduced. Section 3 are described A Concept Hierarchy for Medical Store, the Implementation of the System, and Experiments. Finally, we present the conclusions.

2. MULTILEVEL ASSOCIATION RULES

It is not easy for most of the applications to search strong similarities between data items at low or basic levels of abstraction according to the poorness of data at those levels [3]. Strong associations found at high levels of abstraction indicate common sense information. In addition, to one person, what might reflect common sense may appear unique to the other. Thus, data mining systems apply facilities for mining association rules at multi-levels of extraction, including efficient capability for simple entry between dissimilar abstraction areas. Multilevel association rules are known as association rules created at multiple abstraction levels from mining data.

2.1. The Apriori Algorithm: Finding Frequent Itemset Using Candidate Generation

Apriori is an important algorithm for the mining of frequent Boolean association rules itemset. The algorithm is focused on the assumption that the algorithm applies previous information of frequent itemset properties. Two-stage process of association rule mining are [4] (i) Find all frequent itemset: By definition, Both of these item sets will occur at least as commonly as a pre-determined minimum support count and (ii) Generate strong Association rules from the frequent itemset: By definition, these rules must be fulfilled minimum confidence.

An iterative approach known as a level-wise finding is used by Apriori employees, where k-itemset are used to explore (k+1)-itemset. First, the collection of frequent 1-itemset is discovered. This collection is defined L_1. L_1 is used to find L_2, the group of frequent 2-itemset, which is applied to search L_3, and so on, until no more frequent k-itemset can be searched. The occurrence of each L_k needs one complete scan of the database. Based on the analysis, an important property called Apriori property is that if an itemset I is not frequent, i.e. P(I) < min-sup, then if an item A can be added to the itemset I, the output itemset (i.e. I∪A) cannot appear sometimes than I. So, I∪A is not frequent either, i.e. P (I∪A) < min_sup. To realize how Apriori property is

applied in the algorithm, we can see at how L_{k-1} is utilized to discover L_k. A two-key process is followed, involving of join and prune actions.

The join step: A collection of candidate k-itemset is produced by connecting L_{k-1} with itself in order to discover L_k. This collection of candidates is represented by C_k. Let l1 and l2 be itemset in Lk-1 then l1 and l2 are concerned if their first (k-2) items are in usual, i.e., (l1[1]=l2[1]). (l1[2]=l2[2]) ..... (l1[k-2]=l2[k-2]). (l1[k-1]<l2[k-1]).

The prune step: C_k is the subset of L_k. The resolution of L_k (itemset having a count of no less than minimum help in C_k) will result in a search of the database to decide the count of each candidate in Ck. But by using the Apriori property, this scan and calculation can be reduced. Any non-frequent (k-1)-itemset is not a subset of a frequent k-itemset. Thus, if a k-itemset candidate's is not in (k-1)-subset is not in L_{k-1}, then the candidate is not frequent either and can be eliminated from C_k.

2.1.1. Apriori Algorithm

The following is the summary of the algorithm is being optimized.

Algorithm Apriori: Find a frequent itemset based on candidate generation using an iterative level-wise approach.

Input: *D*, a database of transactions; *min_sup*, the threshold of the minimum support count

Method:

1) L_1=find_frequent_1-itemset(*D*);
2) for (*k*=2; $L_{k-1}\neq\emptyset$; *k*++) {
3) C_k=apriori_gen (L_{k-1});
4) for each transaction $t \in D$ {
5) C_t=subset (C_k, *t*);
6) for each candidate $c \in C_t$
7) *c*.count++;
8) }
9) L_k= {$c \in C_k$ | *c*. count ≥ min-sup}
10) }
11) return $L=U_k L_k$;

procedure apriori_gen (L_{k-1}: frequent (*k*-1)-itemset)

1) for each itemset $l_1 \in L_{k-1}$
2) for each itemset $l_2 \in L_{k-1}$
3) if(l_1[1]=l_2[1])^(l_1[2]=l_2[2])^…^(l_1[k-2]=l_2[k-2])^(l_1[k-1]<l_2[*k*-1]) then {
4) $c = l_1$ join l_2;
5) if has_infrequent_subset (*c*, L_{k-1}) then
6) delete *c*;
7) else add *c* to C_k;
8) }
9) return C_k;

procedure has_infrequent_subset (c: candidate k-itemset; L_{k-1}: frequent (k-1)-itemset);

1) for each (*k*-1)-subset *s* of *c*
2) if $s \notin L_{k-1}$ then
3) return TRUE;
4) return FALSE;

2.2. Generating Association Rules from Frequent Itemset

If the frequent itemset from transactions in a D database has been discovered, it is easy to generate high association rules from them where both minimum support and minimum trust are satisfied with high association rule s. It can also be achieved using the confidence Equation (1), which is showed as a supplement:

$$confidence(A \Longrightarrow B) = P(B|A) = \frac{support(A \cup B)}{support(A)} = \frac{support\_count(A \cup B)}{support\_count(A)} \quad (1)$$

The provisional possibility is shown in terms of itemset support count, where $support\_count(A \cup B)$ is the number of transactions including the itemset $A \cup B$, and $support\_count(A)$ is the number of transactions consisting of the itemset *A*. Focused on this equation, it is possible to produce association rules as follows:

1. In every frequent itemset *l*, all nonempty subsets of *l* must be generated.
2. For each nonempty subset *s* of *l*, generate the rule,

$$s \Longrightarrow "(l\text{-}s)" \; if \; \frac{support\_count(l)}{support\_count(s)} \geq \min\_conf$$

where min_conf is the minimum confidence threshold.

As the rules are produced from frequent itemset, each one immediately completes minimum support. In hash tables with their counts, frequent itemset is put ahead of time so that they can be rapidly accessed.

By producing the subsets of a large item set in a recursive depth-first function, the above method may be improved. For instance, if you have an itemset *ABCD*, initial identify the subset *ABC* and then *AB*, etc. Then if a subset a of a big itemset l does not produce a rule, the subsets of a requirement may not be examined for creating rules using *l*. For instance, if $ABC \Rightarrow D$ does not have sufficient confidence, there is not require to determine whether $AB \Rightarrow CD$ holds. The support of any subset $\bar{a}$ of a must be as important as the support of a. Some rule is not skipped. Furthermore, the confidence of the rule $\bar{a} \Rightarrow (l - \bar{a})$ may not be enough the confidence of $a \Rightarrow (l - a)$. Therefore, if a did not offer a rule containing all the items in l with a as the antecedent, neither would be $\bar{a}$.

2.2.1. Fast Algorithm

If $a \Rightarrow (l - a)$ unable store, neither does $\bar{a} \Rightarrow (l - \bar{a})$ for any$\bar{a} \subset a$. By writing again, it follows that all principles of the form $(l - \bar{c}) \Rightarrow \bar{c}$ must also keep for a rule $(l - \bar{c}) \Rightarrow \bar{c}$ to hold, where $\bar{c}$ is a non-empty subset of c. For instance, if the rule $AB \Rightarrow CD$ holds, then the rules $ABC \Rightarrow D$ and $ABD \Rightarrow C$ must also hold. Think the above property that if a rule with consequent *c* holds for a given large item set, then so perform rules with impacts which are subsets of *c*. Such the property, if an itemset is not small then so is all its subsets. So, from a large itemset *l*, all rules require to be created first with one item in the subsequent. Use the subsequent of these rules and

the ***apriori_gen*** function in Section 2.1.1 to produce all possible subsequent with two items that can occur in a rule produced from *l*, etc. An algorithm is shown below using this concept. In step 2 of this algorithm, rules with one-item consequences can be discovered by applying a special version of the introduction function of ***ap-genrules,*** in which step 8 and 9 are removed to ignore the recursive call.

Fast Algorithm

1) For all large *k*-itemset l_k, $k \geq 2$ do begin
2) H_1 = {consequents of rules extracted from l_k for one item in the consequent};
3) call ap-genrules (l_k, H_1);
4) end

procedure ap-genrules (l_k: large *k*-itemset, H_m: set of *m*-item consequents)

1) if ($k > m+1$) then begin
2) H_{m+1} = apriori-gen (H_m);
3) forall h_{m+1} to H_{m+1} do begin
4) conf = support(l_k)/support (l_k - h_{m+1});
5) if (conf _ minconf) then
6) output the rule (l_k - h_{m+1}) => h_{m+1} with confidence = conf and support = support(l_k);
7) else
8) delete h_{m+1} from H_{m+1};
9) end
10) call ap-genrules (l_k, H_{m+1});
11) end

3. IMPLEMENTATION

The multilevel association rules for medical sales transactions are mind-effective in this system, using definition hierarchies under a reduced support-confidence framework. Concept hierarchies describe the relationships of generalization and specification among the items, and classify them at multiple levels of abstraction. This system describes how the process of the producing multilevel association rules mining is implemented step by step using Apriori algorithm. The Fast algorithm is used to generate strong association rules.

3.1. A Concept Hierarchy for Medical Store

An illustrative example is given how the operate of the producing multilevel association rules mining is implemented successive in this system. To extract multilevel association rules, concept hierarchy or items taxonomy are needed. "Figure **1"** illustrates an example of a concept hierarchy on sales of medical products.

A definition hierarchy describes a collection of mappings from a set of low level concepts to higher level, more general concepts. Data can be analyzed by replacing low level concepts within the data by their higher level concepts, or ancestors, from a concept hierarchy. The concept hierarchy of "Figure1" has three levels, referred to as levels 1, 2 and 3. By convention, levels within a concept hierarchy are numbered from top to bottom, starting with the root node for all (the most general abstraction level). The group (such as Vitamins, minerals and nutrition, Relief of fever and pain) represents the first level (level 1) concept, the category (such as for Amino acid, Glucose, Analgesic) for the second level (level 2), and the brand (such as Amino inj F.BC, Celemin 200ml, Celcox-100, Celcox-200, Decolgel Cream, Dits Gel) for the third level one. Level-1 represents more general concepts whereas level 3 represents the most exact abstraction level of this hierarchy.

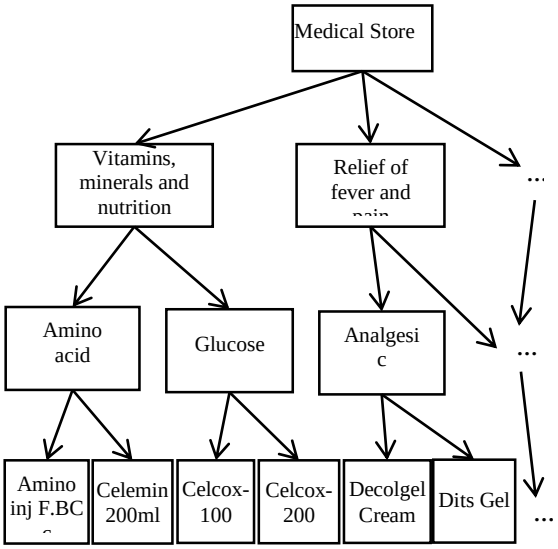

Figure 1. A Concept hierarchy for medical store's medicines

Table1, Table 2, and Table 3 show some of the item codes and its description of each level.

Table 1. Encode medical transaction database for level 1

Code for Level 1	Group Name
I001_000_000	Vitamins, minerals and nutrition
I002_000_000	Relief of fever and pain
I003_000_000	Heart and blood vessels
…	…
I018_000_000	Eye

Table 2. Encode medical transaction database for level 2

Code for Level 2	Category Name
I001_001_000	Amino
I001_002_000	A/S
…	…
I018_001_000	Eye drop (Bangladesh)

Table 3. Encode medical transaction database for level 3

Code for Level 3	Product Name
I001_001_001	Unimin Syr
I001_001_002	Unimin-G Syr
I001_001_003	Astymin Forte
…	…
I018_001_011	Deflux drop

3.2. Implementation of the System

The system presents for producing association rules from the sale data of medical store using NetBeans IDE 7.4 and MySQL Server 5. Apriori algorithm discovers frequent *n*-itemset for each level and Fast algorithm generates association rules for each level after discovering frequent *n*-itemset.

3.3. Experiments

In this experiment, the length of transaction data is 2009. The number of the items for each level for transaction data is shown in Table 4.

Table 4. Number of items in each level

Level	Number of items
Level-1	18
Level-2	65
Level-3	3570

3.3.1. An Illustrative Example

Step1: Transaction Data

The system reads only the transaction data that is a Microsoft Excel file with file extension .xls. The transaction data in the file is represented with two columns, the first column is transaction ID (TID) and the second is a medicine name of sale item is shown in Table 5. In real there is no need column name in transaction file.

Step2: Encode Transaction Data

The Apriori algorithm mines the frequent patterns from a collection of transactions in the form of the TID-itemset, where TID is a transaction ID, and the set of items obtained in TID transaction. This data form is called the horizontal data format. So the above transaction data Table 5 is converted to the horizontal data format with their receptive codes figured in Table 6.

Step3: Encode Taxonomy for Transaction Items

The encode Taxonomy for Transaction Items in transactional database are displayed in Table 7.

Step4: Finding frequent itemset for each level using Apriori Algorithm

Finding frequent itemset for each level using apriori algorithm is described in Figure 2.

Step5: Generating Association Rules for each level using Fast Algorithm

An algorithm named fast algorithm developed by Rakesh Agrawal and Ramakrishnan Srikant is applied to discover the association rules from frequent itemset for each level in this system.

Generating association rules for each level using fast algorithm is described in Figure 3.

Table 5. Medical store's transaction data

TID	Medicine Name
1	Diclogel Cream
1	Unimin Syr
1	Diclogel Cream
2	Dits Gel
2	Novonorm 1g
2	Green Tea(Fame)
3	Invoril 5
3	Diclogel Cream
3	Nifidepine tab (China)
4	Unimin-G Syr
4	Invoril 5
4	Diclogel Cream
4	Nifidepine tab (China)
5	Unimin-G Syr
5	Diclogel Cream
5	Invoril 5
6	Unimin Syr
6	Dits Gel
7	Invoril 5
7	Diclogel Cream
7	Novonorm 1g

Table 6. Encode transaction table

TID	Medicine
1	I001_001_001, I002_001_001
2	I002_001_002, I001_002_001, I001_002_002
3	I003_002_001, I002_001_001, I003_001_001
4	I001_001_002, I003_002_001, I002_001_001, I003_001_001
5	I001_001_002, I002_001_001, I003_002_001
6	I001_001_001, I002_001_002
7	I003_002_001, I002_001_001, I001_002_001

Table 7. Encode taxonomy for transaction items

Codes	Level	Description
I001_000_000	Level 1	Vitamins, minerals and nutrition
I002_000_000	Level 1	Relief of fever and pain
I003_000_000	Level 1	Heart and blood vessels
I001_001_000	Level 2	Amino acid
I001_002_000	Level 2	A/S
I002_001_000	Level 2	Analgesic
I003_001_000	Level 2	Antihypertensive
I003_002_000	Level 2	Antiangina
I001_001_001	Level 3	Unimin Syr
I001_001_002	Level 3	Unimin-G Syr
I001_002_001	Level 3	Novonorm 1g
I001_002_002	Level 3	Green Tea(Fame)
I002_001_001	Level 3	Diclogel Cream
I002_001_002	Level 3	Dits Gel
I003_001_001	Level 3	Nifidepine tab (China)
I003_002_001	Level 3	Invoril 5

3.3.2. Frequent Pattern for Level 1

In this system, a top-down strategy is employed to find association rules at three levels of abstraction. So, the association rules for the concept level 1 are firstly discovered. To find frequent items from transactions, the user has to enter the minimum support count and minimum confidence threshold for the system.

3.3.2.1. Finding Frequent Itemset for Level 1

Before generating the association rules from the encode transaction data, the frequent itemset must be discovered by using apriori algorithm.

Table 8. Frequent 5-itemset for level 1

No.	Frequent 5-itemset	Absolute Support count
1	I001_000_000, I002_000_000, I003_000_000, I004_000_000, I008_000_000	669

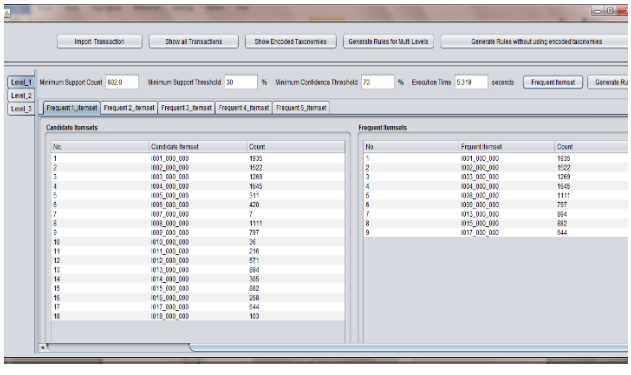

Figure 2. Finding frequent n-itemset for level 1

The minimum relative support at the top-most concept level is taken as 30% to find frequent *n*-itemset. With 30% support count, n-itemset of occurrence greater than or equal 602 times (30/100 * 2009) are considered as frequent n-itemset. With 30% support count, only one frequent 5-itemset is generated with absolute support count 669. Each a frequent 5-itemset is found 602 times in transactions. These frequent 5-itemset that carry for next level are shown in Table 8.

3.3.2.2. Generating Association Rules from Level-1 Frequent *n*-Itemset

In this system, the fast algorithm is applied to produce rules from observed frequent *n*-itemset.

If the minimum confidence threshold for the association rules that can be produced from 1 frequent 5-itemset shown in Table 12 is 70%, then the 11 above rules are generated. These are the strongly generated ones. Table 13 shows the list of these rules with its confidence. The following rule has a 99% confidence.

$$\text{buys}\left(\begin{matrix}\text{X, "Relief of fever and pain, Heart and blood vessels,}\\ \text{Infection control, Stomach"}\end{matrix}\right)$$

$$\Rightarrow \text{buys(X, "Vitamins, minerals and nutrition")}$$

The generated rules at this level are with the medicine's group name because the top-most level of a concept hierarchy is more general concepts.

3.3.3. Frequent Pattern for Level-2

Before entering the reduced support to the system for level 2, the user must remember that this system is used reduced support to find frequent itemset. Therefore, the relative support for level 2 is less than level 1's. There is no problem whatever confidence threshold is.

3.3.3.1. Finding Frequent Itemset for Level-2

Only the descendants of the items from the largest frequent *n*-itemset at level-l are considered as candidates in the level 2 large 1-itemset. There are 40 unique items descendants from 5 unique items from Table 8. These items are used to find the frequent *n*-itemset for level 2. The minimum relative support at this level is 20% (absolute support count = 401). Three frequent *5*-itemset are generated as listed in Table 9.

Table 9. Frequent 5-itemset for level 2

No.	Frequent *5*-itemset	Absolute support count
1	I001_004_000, I001_009_000, I001_014_000, I002_001_000, I004_002_000	516
2	I001_004_000, I001_009_000, I001_014_000, I002_001_000, I004_004_000	433
3	I001_004_000, I001_009_000, I001_014_000, I004_002_000, I004_004_000	419

3.3.3.2. Generating Association Rules from Level-2 Frequent *n*-Itemset

At the level 2, the 23 rules are generated as strong rules with confidence 70. These rules are listed in **Table 10**. The following rule has a 93.11% confidence. The rules with category name at level 2 describe more specific concepts than those with group name at level 1.

$$\text{buys}(X, \text{"ANA, Vitamin, Analgesic, Antibiotic"}) \Rightarrow \text{buys}(X, \text{"Supplement"})$$

Table 10. Rules with frequent 5-itemset at level 2

No.	Rule	Confidence threshold (%)
1	ANA∧Vitamin∧Antibiotic∧Antiseptic=>Supplement	93.94
2	ANA∧Vitamin∧Analgesic∧Antiseptic=>Supplement	93.11
3	Vitamin∧Supplement∧Antibiotic∧Antiseptic=>ANA	89.91
4	Vitamin∧Supplement∧Analgesic∧Antibiotic=>ANA	88.81
5	ANA∧Vitamin∧Analgesic∧Antibiotic=>Supplement	88.05
6	ANA∧Supplement∧Antibiotic∧Antiseptic=>Vitamin	86.03
	: : : :	
23	ANA∧Vitamin∧Antiseptic=>Supplement∧Antibiotic	71.01

3.3.4 Frequent Pattern for Level-3

Level-3 is the lowest level of the concept hierarchy of "Figure **1".** The data at this level is the most specific concepts. The minimum relative support must be less than level 1 and level 2.

3.3.4.1 Finding Frequent Itemset for Level-3

There are 1287 unique items descendants from 6 unique items from **Table 9**. By using these items, the frequent 6-itemset shown in **Table 11** are obtained with minimum relative support at this level is 5% (absolute support count = 100).

Table 11. Frequent 6-itemset for level 3

No.	Frequent 6-itemset	Absolute support count
1	I001_009_157, I001_014_068, I001_014_138,I001_014_139, I002_001_155, I004_004_077	122

3.3.4.2 Generating Association Rules from Level 3 Frequent *n*-Itemset

The only 41 rules above the confidence 70% are generated from frequent 6-itemset.

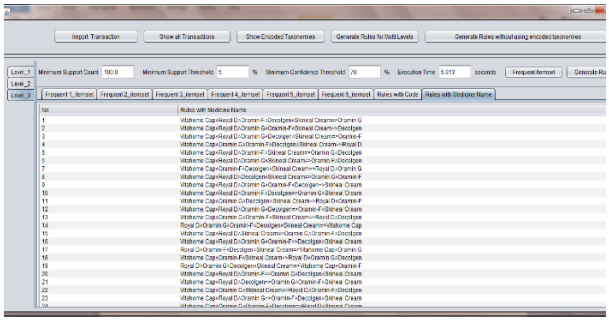

Figure 3. Finding association rules for level 3

Table 12. Rules with medicine name from frequent 6-itemset for level 3

No.	Rules	Confidence (%)
1	Vitahome Cap ∧ Royal D ∧ Oramin-F ∧ Decolgen ∧ Skineal Cream => Oramin G	99.18
2	Vitahome Cap ∧ Royal D ∧ Oramin G ∧ Oramin-F ∧ Skineal Cream => Decolgen	98.38
	: :	
41	Decolgen ∧ Skineal Cream => Vitahome Cap ∧ Royal D ∧ Oramin G ∧ Oramin-F	73.93

Table 13. Rules with frequent 5-itemset at level 1

No.	Rule	Confidence threshold (%)
1	Relief of fever and pain∧Heart and blood vessels∧Infection control ∧Stomach=> Vitamins, minerals and nutrition	99.85
2	Vitamins, minerals and nutrition∧Relief of Fever and pain∧Heart and blood vessels ∧ Stomach=>Infection control	95.43
3	Relief of fever and pain∧Heart and blood vessels∧Stomach=>Vitamins, minerals and nutrition∧Infection control	94.89
	: : :	
11	Relief of fever and pain∧Stomach=> Vitamins, minerals and nutrition∧Heart and blood vessels∧ Infection control	73.19

The strong association rules at level 2 and level 3 are the descendants of level 1 and level 2 respectively. At level 3, the rules are described with the medicine's product name. An example is a rule (Vitahome *Cap* ∧ *Royal D* ∧ *Oramin-F* ∧ *Decolgen* ∧ *Skineal Cream* => *Oramin G)* with medicine's product name at level 3. It cannot know clearly what kind of medicine it is. To know a rule with common sense, level 2 presents for the

category of medicine in this system. Comparing a rule (*ANA* ∧ *Vitamin* ∧ *Antibiotic* ∧ *Antiseptic* => *Supplement*) at level 2 with a rule (*Vitahome Cap* ∧ *Royal D* ∧ *Oramin-F* ∧ *Decolgen* ∧ *Skineal Cream* => *Oramin G*) at level 3, the relationship of medicines of the rule at level 2 can be recognized obviously rather than at level 3. The rules at level 1 show with the most general group name of medicines. For example (Relief *of fever and pain* ∧ *Heart and blood vessels* ∧ *Infection control* ∧*s Stomach* => *Vitamins, minerals and nutrition)* is a rule at level 1.

Now, the items at each level obtained encode taxonomy for transaction items in transactional database are used to discover the frequent *n*-itemset.

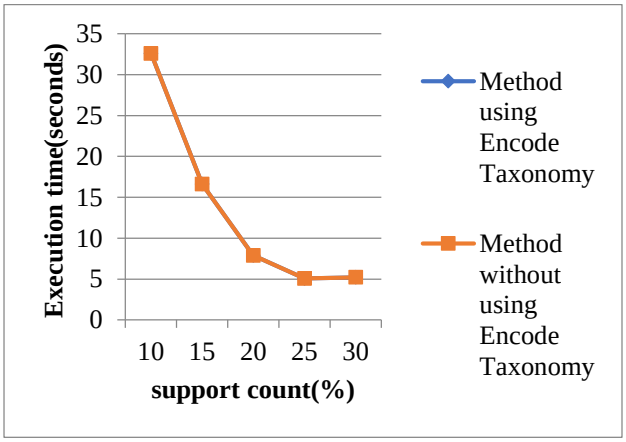

Figure 4. Graph of execution time for level 1

The association rules at level 1 are the same with the rules of method without using Encode Taxonomy of the higher level but at level 2 and level 3 are not same because of using reduced minimum support count at lower levels. Their execution times are different. The execution times of both methods are described Figure 4 to Figure 6 according of different support counts for each level. By comparing these times, the method using Encode Taxonomy is less than other method without using Encode Taxonomy according to the filtering items. The method without using Encode Taxonomy is more taken longer than the other because it scans the database more time for all items at level 2 included in Encode Taxonomy.

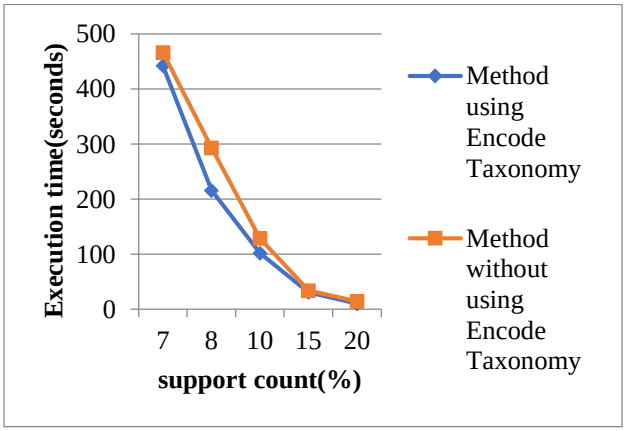

Figure 5. Graph of execution time for level 2

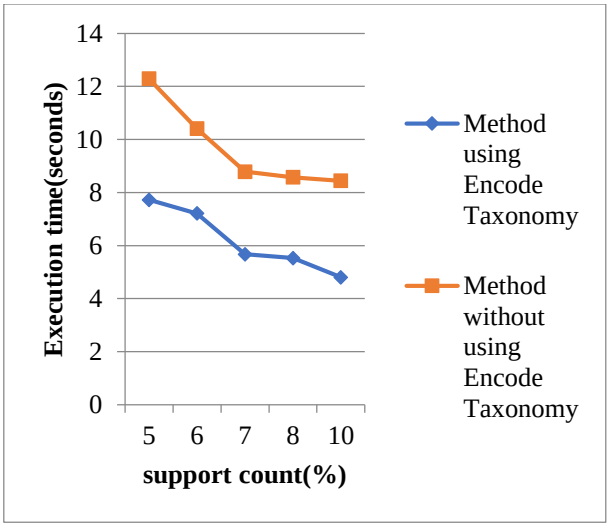

Figure 6. Graph of execution time for level 3

4. CONCLUSIONS

In this work, the sales transaction data of a medical store is used to discover the interesting correlation relationships among people. This system can help people who work in pharmacies. The frequent itemset under support threshold are searched from sale transaction data and from which strong association rules are generated under confidence threshold.

Our proposal generates the association rules for three levels. The rules generated from multilevel mining will assist large data for the users and improve the flexibility and efficiency of the systems. If an item occurs occasionally, its descendants will occur even less frequently. So only the descendants of the frequent items carry to the next level. Multilevel association rules give more correct and specific information. It can assist organizations to construct promotional strategies and help enhancing the sales and setting the future plans.

REFERENCES

[1] C. Agarwal, "A Tree Projection Algorithm for Generation of Frequent Itemset", JPDC, 2001.

[2] G. Pratima, "An Efficient Algorithm for Mining Multilevel Association Rule Based on Pincer Search", Computer Application, MANIT, Bhopal, M.P. 462032, India.

[3] X. Yuan, "An improved Apriori Algorithm for Mining Association Rules", AIP conference, 2017.

[4] Y. Jiao, "Research for an Improved Apriori Algorithm in Data Mining Association Rules" IJCCE, Vol. 2, No. 1, Jan 2013

Genetic Algorithm-Based Feature Selection and Classification of Breast Cancer Using Bayesian Network Classifier

Yi Mon Aung[1], Nwet Nwet Than[2], Linn Linn Htun[3]
[1,3]*Faculty of Information Science, University of Computer Studies (Magway)*
[2]*Information Technology Supporting and Maintenance, UCSY*
[1]*yimonaung@ucsmgy.edu.mm*, [2]*nwenwethan@ucsy.edu.mm*, [3]*linnlinntun@ucsmgy.edu.mm*

ABSTRACT: *Cancer is one of the fastest growing and most dangerous diseases in the healthcare sector. Early diagnosis of this disease is very important because the success of your treatment depends on how quickly and accurately it is diagnosed. Data mining technology can help clinicians make diagnostic decisions. To improve the efficiency of these algorithms, the best features are needed to choose. To exclude non-essential attributes, genetic algorithms are used to extract useful and important attributes. This process speeds up the data mining process and reduces computational complexity. Therefore, this study uses genetic algorithms to select the best features before applying the classification algorithms to three breast cancer datasets retrieved from the UCI repository. The process also used several single and multiple classifier systems to build a precise system for breast cancer diagnosis. This approach was useful for early forecasting, and the results show that the Genetic Algorithm based features selection performs better accuracy than other classifiers without features selection.*

Keywords: Genetic Algorithm; Classification; Bayesian Network; Feature Selection.

1. INTRODUCTION

Breast cancer is a major growth for women in both developed and developing countries. Breast cancer rates are increasing due to expansions in urban improvement and Western way of life. Worldwide, more than women died for breast cancer in 2011 in number of 508,000. Breast cancer survival proportions vary commonly around the world. North America, Sweden and Japan account for more than 80%, middle-income countries 60% and low-income republics less than 40%. In developing countries, low being is primarily due to the lack of early finding programs and the lack of high occurrence women and health care accommodations and appropriate diagnosis and treatment.

Data mining is dominant. This is a new field, and various approaches for examining real-life problems have freshly been developed. Change raw data into valuable data in a diversity of research self-controls and explore designs to control future instructions in the health field. There are various major data mining methods recently developed in the data mining industry to gain knowledge from databases. Breast cancer is the leading cause of death in women in developing countries and the second highest cancer incidence in developed countries, according to the National Cancer Organization. The highest occurrence in women worldwide is found in the most common form of breast cancer mammogram, asymptomatic knobs. Breast cancer can occur at virtually any age, so patients and surgeons need to be aware of new breast cancer indications [1].

This paper presents a system that improves the accuracy of classification of breast cancer patients. The performance of generalized Bayesian Networks, Naive Bayesian, Logistic Regression, Artificial Neural Networks, and Random Forests to improve the predictive model of breast cancer identification decisions based on three breast cancer datasets.

2. RELATED WORKS

A study by Abdelghani Bellachia and Erhan Guven uses information retrieval methods to estimate the survival rate of breast cancer patients [2]. In this paper, the SEER Public-Use Data is applied and available with 16 attributes from the SEER database, the preprocessed data set consists of 151,886 records. Three data mining methods is used to conduct SEER data group. The C4.5 decision tree algorithm and backpropagated neural network were examined. Many experiments practice these algorithms. It was carried out and finally, it was decided that the C4.5 algorithm did much better than the other two methods.

Chandra Prasetio Utomo's research paper uses an artificial neural network with exciting technology to detect breast cancer based on the Wisconsin breast cancer dataset. Breast Cancer Screening [3], in this study, they applied ANN to extreme breast cancer studies based on the Wisconsin Breast Cancer Database. The capabilities of this technology have grown to become an intelligent section of medical decision support systems. They linked the two methods and determined that the neural network of the artificial learning machine was better than BPANN.

The Kharya and Shweta readings used fact-finding techniques to identify and predict cancer [11]. They claim that predicting the consequence of illness is one of the most stimulating and difficult tasks in emerging data mining tools. This thing outlines current studies that have used data analysis techniques to recover the diagnosis and scenario of breast cancer data. In this paper, the accuracy of the three data acquisition methods will be linked. Applied and introductory results of their method can be applied to information retrieval methods in predicting survival in databases. The resulting acting is compared to current technology.

Khan Muhammad Umer et al further research [10], in their study, they deliberated the expectedness of breast cancer survival using an uncertain choice sapling for

personal health care. They provide an information system aimed at serving physicians provide a certain side by side of reliability and significance to their final choices in order to be reliable and precise with human decision. Studying non-deterministic decision tree-based hybridization systems is recommended as a viable alternative to explicit identifiers for the development of such analytical systems. Various decision rules and experiments were performed using indeterminate membership types and demo techniques. In this paper, the forecast of breast cancer will be liken. The recommended classification of the hybrid fuzzy permission tree is more reliable and fair than the freely available network structure. In accumulation, it has the likely to adapt to noteworthy performance gains.

3. MATERIALS AND METHODS

In this paper, the Bayesian Network, Bayes algorithm was instigated using the Java Netbeans interface to predict the type of breast cancer. The results with other algorithms were compared by using Weka. The datasets from the UCI Machine Learning repository are used.

3.1. Dataset Description

In this study, three dataset are used to analyse the breast cancer. One of the datasets is from UCI Machine Learning Repository is used. They have been collected by M. Zwitter and M. Soklic at the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia [13]. The other two dataset is from Wisconsin breast cancer dataset. The detail of these datasets is shown in table 1.

Table 1. Breast cancer datasets

Name	No of instances	No of attributes
Breast cancer	10	286
Wisconsin prognosis breast cancer (WPBC)	35	198
Wisconsin diagnosis breast cancer (WDBC)	33	569

3.2. Feature Selection

The general feature selection process consists of four simple steps. This means creating a subset, evaluating the subset, stopping the condition, and checking the results. Subset generation is the search method chosen to evaluate based on a particular search strategy. Each is compared to the best previous division according to the ranking and reliability criteria of the candidate division [7]. If the new one is better than the old one, delete the old one. The process of designing and evaluating most components is often tedious until a particular shutdown decision is made. In that case, the user usually needs to use prior knowledge or various tests to recognize the best part. Depends on the composite and / or the actual dataset. Features can be selected in many areas of data collection, such as group rules and regulations. The procedure for collecting design elements from different standards can be divided into three main groups.

3.3. Genetic Algorithm

Genetic Algorithms (GA) are widespread adaptive study systems that livelihood the assortment of Darwinian and the assessment of inherited organic systems. GA works with candidate solutions called populations. When iterative controls are functional, GA gets the best explanation. GA hunts for other pictures of chromosomes until satisfactory results are obtained [12].

The Genetic Algorithm (GA), originally established in the Netherlands, is a computational optimization paradigm demonstrated on the concept of genetic evolution. GA is an optimization procedure that runs in the twofold search space and operates a group of possible solutions. Points in the exploration space are characterized by a finite arrangement of zeros and ones called genes. The quality of possible solutions is assessed by the flexibility function. The survival rate is relative to the total of chromosomal adaptableness. In GA, the initial populace is randomly produced by three operators: selection, crossover, and mutation. The selection operator selects the leading to pass indirectly to the next group. The crossover operator randomly swaps parts of a chromosome between two designated parents to produce descendants chromosomes. The mutation operator arbitrarily warns about little in the chromosome.

3.4. Bayesian Network

The Bayesian network encrypts a joint probability distribution of variables $\{x_1, \ldots, x_v\}$, in the form of aperiodic charts and immunity tables. The purpose of the organization is to precisely estimation the value of a discrete class variable called $y = x_v$ given by the weaknesses of the estimator or attribute. An algorithm that learns the structure of Bayesian network isolation by increasing the likelihood of exceptions. This is similar to the acceleration algorithm of Heckerman et al. Except for using the logarithmic probability of the class as the main objective goal. It all starts with an unavailable network. At each stage, delete the new semicircle (that is, the arc that does not create a path) and each existing arc [5].

3.5. Naïve Bayes Classifier

The Naive Bayes (NB) classification is a possible arrangement grounded on Bayesian theory. The naive Bayes classifier yields probability approximations, not forecast. For each class worth, estimate the probability of owning that class according to the example given. The effect of attribute values in a precise class is considered independent of the values of other characteristics. This perception is called excellent class individuality. The Naive Bayesian model take up that all variables are self-governing of each other [8].

3.6. Logistic Regression

Logistics regression were native since models of posterior probabilities for rows K via a linear meaning of x, but they are tried to keep within the variety of sums and

distances. This design can be denoted by a logical variation of K-1 or a logarithmic argument. For odds ratios, the model uses the last class as the denominator, but the choice of splits is random and the forecasts are distributed under that option. The model is simple because K = 2 has only one linear function. This pattern is widely used in biostatic presentations where binary responses (only two classes) occur on a regular basis [6].

3.7. Artificial Neural Network

A neural network (ANN) is an interrelated network of measuring device cells that emit complex patterns of electrical signals. The human brain is a kind of biological neural network. Artificial neural networks are knowledge models based on biological neural networks. Advantages of Neural Network:

- Neural networks are well-matched and flexible in altering the surroundings.
- Neural networks can examine and be aware of patterns in datasets.
- Neural networks can easily generate statistics and handle very complex circumstances [9].

3.8. Random Forest

The Random Forest (RF) algorithm procedures a classification method that relies on an arrangement of decision trees. The individuality of this classification is that those tree portions are developed from random quantities. Based on this idea, RF determination is defined as a common principle of chance tree sets. Proper binary excruciating sends information from the parent node to the two nodes. Only then will the node continue from node to parent node. RF is a gathering of hundreds of thousands of trees. Each tree was established using a bootstrap sample of the unique substantial [4].

4. EXPERIMENTAL RESULTS

All experiments defined in this paper were accomplished using the Weka machine learning framework library (version 3.8.4). Weka is a machine learning toolbox that aims to help the system put on machine learning techniques to a diversity of real-world complications. Firstly, Genetic Algorithm (GA) Feature Selection are applied to select the best attributes, and then used five data mining techniques. In this system, tenfold cross-validation was used. The selection of GA features with the best attributes selected was used, and then tenfold cross-validation was used for those selected attributes. Breast cancer datasets have 10 attributes: age, menopause, tumor size, inv-nodes, node caps, deg-malig, breast, breast-quad, irradiation and class. Among them, GA-based function selection extracts 6 features. These are tumor size, inv-nodes, node caps, deg-malig, irradiation and class. The system processes Bayesian networks, naive Bayesian algorithms, logistic regression, artificial neural networks, and random forests to identify signs of potential breast cancer patients.

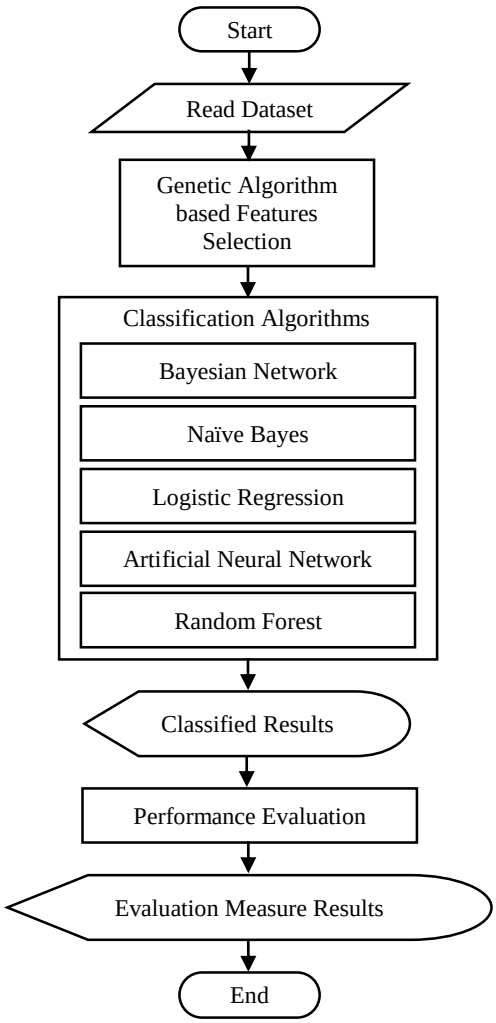

Figure 1. System flow diagram

The system is evaluated by using the following three equations.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Where,

TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

Table 2 shows the information about the accuracy of all algorithms with precision and recall.

Table 2. Precision, recall and accuracy of 5 classifiers without features selection using breast cancer dataset

Classifiers	Precision	Recall	Accuracy (%)
Bayesian Network	0.707	0.720	72.03
Naïve Bayes	0.704	0.717	71.68
Logistic Regression	0.668	0.689	68.88
Artificial Neural Network	0.648	0.647	64.69
Random Forest	0.664	0.696	69.58

Table 2 presents a comparison of the five classifiers without feature selection performed on the breast cancer data. The Bayesian network has the highest accuracy in all measurements 72.03%. The information on the five classifiers and the precision and recall tables 2 and 3 compare the selection of the characteristics based on the genetic algorithm.

Table 3. Precision, recall and accuracy of 5 classifiers with genetic algorithm based features selection using breast cancer dataset

Classifiers	Precision	Recall	Accuracy (%)
Bayesian Network	0.723	0.734	73.43
Naïve Bayes	0.713	0.724	72.38
Logistic Regression	0.693	0.717	71.68
Artificial Neural Network	0.701	0.717	71.68
Random Forest	0.675	0.706	70.63

In this study a feature selection model with GA-based feature selection is designed to classify appropriate features. The association of average accuracies for the five classifiers (Bayesian Network, Naïve Bayes, Logistic Regression, Artificial Neural Network, and Random Forest) with and without feature selection on breast cancer dataset displayed that without feature selection the accuracy of Bayesian Network (72.03%) is the best and the accuracy obtained by Bayesian Network is better than that produced by GA-classifier (73.43%). Also it is apparent from results attained that precision and recall has been approximately developed by feature selection on breast cancer dataset shown in figure 2.

In table 4 and 5, classification of breast cancer without and with GA-base feature selection is shown. They are using wisconsin prognosis breast cancer (wpbc) dataset. Feature selection based classification is good for overall classifiers except Random Forest and Logistic Regression. Logistic Regression has the same accuracy results for both classifications. However, Random Forest classifier has the best accuracy without features selection.

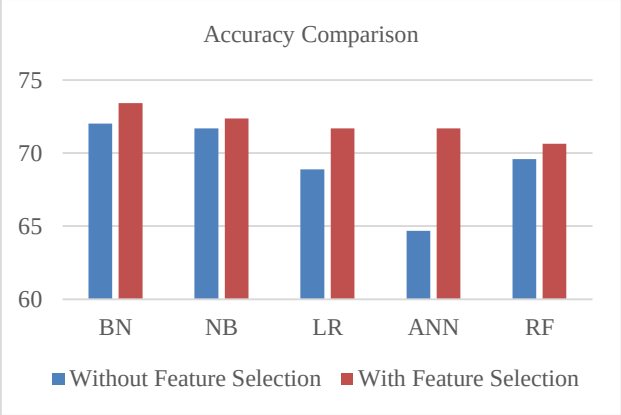

Figure 2. Accuracy Comparison on Breast Cancer Dataset

Table 4. Precision, recall and accuracy of 5 classifiers without features selection using WPBC dataset

Classifiers	Precision	Recall	Accuracy (%)
Bayesian Network	0.628	0.747	74.75
Naïve Bayes	0.716	0.672	67.17
Logistic Regression	0.765	0.783	78.28
Artificial Neural Network	0.726	0.737	73.74
Random Forest	0.821	0.813	81.31

Table 5. Precision, recall and accuracy of 5 classifiers with genetic algorithm based features selection using WPBC dataset

Classifiers	Precision	Recall	Accuracy (%)
Bayesian Network	0.763	0.763	75.76
Naïve Bayes	0.748	0.763	76.26
Logistic Regression	0.765	0.783	78.28
Artificial Neural Network	0.722	0.758	75.76
Random Forest	0.736	0.763	76.26

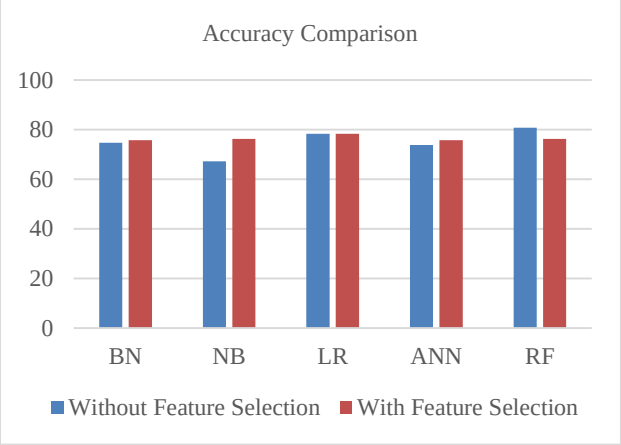

Figure 3. Accuracy comparison on (WPBC) breast cancer dataset

Table 6. Precision, recall and accuracy of 5 classifiers without features selection using WDBC dataset

Classifiers	Precision	Recall	Accuracy (%)
Bayesian Network	0.953	0.953	95.2548
Naïve Bayes	0.926	0.926	92.6186
Logistic Regression	0.943	0.942	94.2004
Artificial Neural Network	0.956	0.956	95.6063
Random Forest	0.965	0.965	96.4851

Table 7. Precision, recall and accuracy of 5 classifiers with genetic algorithm based features selection using WDBC dataset

Classifiers	Precision	Recall	Accuracy (%)
Bayesian Network	0.951	0.951	95.08
Naïve Bayes	0.940	0.940	94.02
Logistic Regression	0.961	0.961	96.13
Artificial Neural Network	0.961	0.968	96.84
Random Forest	0.965	0.965	96.49

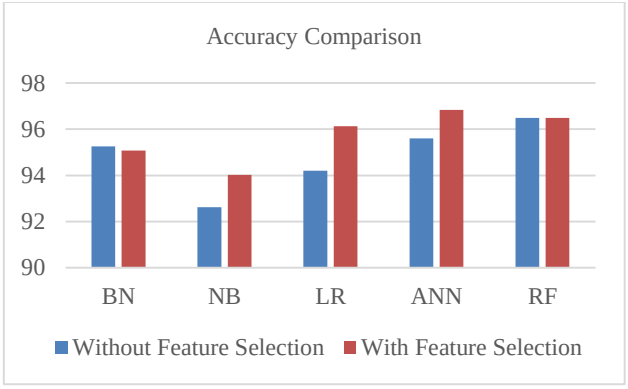

Figure 4. Accuracy comparison on (WDBC) breast cancer dataset

It is detected that feature selection better-quality the accuracy of all classifiers and the best accuracy with feature selection accomplished by Artificial Neural Network (96.84%). Overall, classification of breast cancer with feature selection is the best accuracy except Bayesian Network and Random Forest. Bayesian Network has slightly decreased the accuracy value when the system used the GA-based features selection. Random Forest has the same accuracy value with and without features selection.

5. CONCLUSIONS

In this paper, a feature selection method using GA were proposed for choosing the greatest subset of features for breast cancer diagnosis system. Bayesian networks, Naïve Bayes classifier, Logistic regression, Artificial Neural Network and Random Forest were used to assess Genetic Algorithm based feature selection method on Breast Cancer Datasets. In this paper, the classification using Bayesian Network is superior to other classification algorithms and achieved the best accuracy in breast cancer dataset. In WPBC dataset, Random Forest is the best accuracy value without feature selection. Artificial Neural Network is the best accuracy value with GA-based feature selection in WDBC dataset. To sum up, the results obviously show that features selection significantly improved fit and sensitivity in accuracy of all others classifiers with GA-based feature selection.

REFERENCES

[1] Aarti Sharma, Rahul Sharma,Vivek Kr. Sharma,Vishal Shrivatava, “Application of Data Mining A Survey Paper”, International Journal of Computer Science and Information Technologies, Vol. 5, Issue 2, 2014.

[2] Abdelghani Bellaachia, Erhan Guven, “Predicting Breast Cancer Survivability Using Data Mining Techniques”, January, 2006.

[3] Chandra Prasetyo Utomo, Aan Kardiana, Rika Yuliwulandari, “Breast Cancer Diagnosis using Artificial Neural Networks with Extreme Learning Techniques”, International Journal of Advanced Research in Artificial Intelligence, Vol. 3, No. 7, 2014.

[4] Cuong Nguyen, Yong Wang, Ha Nam Nguyen, “Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic”, Journal of Biomedical Science and Engineering, 2013.
[5] Daniel Grossman, Pedro Domingos, “Learning Bayesian Network Classifiers by Maximizing Conditional Likelihood”, Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 2004.
[6] Emina Alicˇkovic, Abdulhamit Subasi, “Breast cancer diagnosis using GA feature selection and Rotation Forest”, Neural Comput & Applic, Springer, Nov, 2015.
[7] Kathija, Shajun Nisha, “Breast Cancer Data Classification Using SVM and Naïve Bayes Techniques”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 12, December 2016.
[8] Megha Rathi, Arun Kumar Singh, “Breast Cancer Prediction using Naïve Bayes Classifier”, International Journal of Information Technology & Systems, Vol. 1, No. 2, ISSN: 2277-9825, 2012.
[9] Mr. Akshay Jadhav, Ms. Jennifer D’Cruz, Mr. Virendra Chavan, Ms. Ashvini Dighe, Prof. Jayashree Chaudhari, “Detection of Lung Cancer Using Backpropagation Neural Networks and Genetic Algorithm”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 4, 2016.
[10] Khan, Muhammad Umer, Jong Pill Choi, Hyunjung Shin, and Minkoo Kim, “Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare”, IEEE Int. Conference on Engineering in Medicine and Biology Society, 2008.
[11] Shweta Kharya, “Using Data Mining Techniques For Diagnosis and Prognosis of cancer disease”, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012.
[12] Tanzeem Khan Mansoori, Amrit Suman, Dr. Sadhana K. Mishra, “Feature Selection by Genetic Algorithm and SVM Classification for Cancer Detection”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 9, September 2014.
[13] Breast Cancer Dataset from UCI Machine Learning Repository.
https://archive.ics.uci.edu/ml/datasets/breast+cancer

Sentiment Analysis of Students' Feedback from Coursera Online Learning Using Bernoulli Naïve Bayes Classifier

Nilar Htun[1], Nang Seint Seint Soe[2]
[1,2]*Faculty of Computer Science, University of Computer Studies (Taungoo)*
[1]*nilarhtun@ucstaungoo.edu.mm,* [2]*nangseintseintsoe@ucstaungoo.edu.mm*

ABSTRACT: *Nowadays, online learning is popular and the effective way in our society. Therefore, sentiment analysis becomes a useful tool to listen the feedback from the students. Sentiment analysis, also known as the opinion mining, is the process of analyzing, processing, and classifying subjective texts especially in the application of market research, customer service, brand monitoring and social media monitoring. Conventional sentiment classification task classifies the emotions in 3-way classification (positive, negative or neutral). The system aims to classify the students' opinion from Coursera's online learning into positive or negative classes (binary classification) by using the Bernoulli Naïve Bayes (BernoulliNB) which is a variant of Naive Bayes and predicts the probability of the input review for binary classification. The system consists of 4 main parts: (1) preprocessing the raw data (2) POS tagging (3) sentiment word selection and (4) sentiment classification with Bernoulli Naïve Bayes classifier. The system uses the open NLP parser for POS tagging and is made for the building of sentiment words. For the experiments, the system uses students' feedback reviews from Coursera online learning website. Finally, the system compares the results with other methods: MultinomialNB, BernoulliNB and KNeighbors Classifiers and proves the performance of the system is higher than the other methods. Precision, recall, f1-score and accuracy measurement are used for the performance evaluation.*

Keywords: Online learning; Sentiment analysis; POS tagging; Bernoulli Naïve Bayes classifier.

1. INTRODUCTION

The advent of the Internet has been leading to a wealth of benefits in the area of resource utilization such as review sites, online forums, and web blogs [1]. Online reviews by users are mostly available on the Internet and it helps to decide and ensure the quality of any item, product or entity. With the proliferation of universities around the world, there are fewer courses available for students to attend. Online learning is expected to grow at a faster rate as the Internet and information technology infrastructure grow rapidly. Users get online reviews on the Internet, which can help to determine their quality of knowledge.

Sentiment analysis or emotional analysis, also known as opinion mining, is a computer system that seeks to understand and explain opinions and attitudes by analyzing large amounts of opinion information in a useful way, such as assisting in human decision-making. It plays a vital or main role in development for the developers to improve the service that they offered, reveals the quality of online learning system and new users' (learners and teachers) opportunities. It helps to classify the opinion of the users based on reviews or comments as positive or negative, consistent with the overall sentiment expressed within or via those reviews. Coursera is one of the online training platforms and aims to provide a lifelong learning experience anywhere in the world.

In this system, users' reviews from Coursera are classified as positive or negative class using Bernoulli Naïve Bayes Classifier which is a classification algorithm of Machine Learning based on Bayes theorem and it gives the likelihood of occurrence of the review. They are extremely fast as compared to other classification models. The performances of the system with other methods are explained detail in Section 4.

Full details of this paper are organized in the following sections. The related works and background information are described in Section 2. In Section 3, the system is explained in detail. Experiments and evaluation for the results are shown in Section 4 and finally, Section 5 concludes the presented study.

2. RELATED WORK

According to the previous studies, online learning has become the largest sector of the distance education in recent years. Actually, its basic definition is a form of distance learning (i.e. learning acquired at a distance via materials placed on the Web and via Internet services).Hence, online-learning [2] is considered as a type of education, where students can self-study at anytime and anywhere, and communicate with teachers and other students (leaners) by the use of digital technologies such as e-mail, electronic forum, videoconferencing or videoconference, chat rooms, bulletin boards along with other computer-based communication methods. The authors-built AI based computer assisted learning system that gives the student a specific framework to solve exercises and teaches a general problem-solving method. They also discuss the benefits of using system in a learning process, in relation to standard teaching in the classroom [3]. The authors in [4] present the sentiment classification using Machine learning techniques like Naïve Bayes (NB), Maximum Entropy (ME) and Support Vector Machines (SVMs). The performance shows that the Naïve Bayes tends to do the worst and SVMs tend to do the best.

As already mentioned, there are currently a lot of online reviews are available on the various kinds of study blogs and forums that go beyond the reach of any human's opinion and visually impaired. Therefore, an urgent need arises for innovative technologies that can automatically analyze the attitudes of the users mentioned in the reviews. Fundamentally, practical sentiment classification techniques can automatically classify, analyze based on

online learning blogs. Here, overview of specific online learning systems is useful for those who build positive or negative feedback. Sentiment classification is being investigated in a wide range of domains such as film reviews, product reviews, travel reviews, and online-learning reviews. The system is to examine sentiment analysis when accepting it as a two-topic document analysis; positive and negative.

Essentially, the authors present the results for experiments using: Naïve Bayes arrangement [5]. Furthermore, Bernoulli Naïve Bayes (NB) proved to be the fastest. Many have been studied about sentiment analysis. Sentiment analysis includes text analytics, NLP, and Linguistics calculation to organize sentiment polarities. Moreover, it is responsible for digging into the special text, which is used to determine and extract people's personal opinions or opinions from the text. A sentimental distinction is the primary function of the sentiment analysis to distinguish people's opinions which are expressed in the way of text format into the diverse sentiment polarity programs that are the positive class and negative class [6].

The authors proposed to create a Myanmar spirit dictionary for food and restaurant domains. The approach in the research is based on the dictionary machine learning process [7]. It used a Naïve Bayes classifier to shape the classifier. According to test results, positive and negative data are needed to use in similar proportions to practice the classification application for effective results.
Comparison tests among Naïve Bayes, SVM, Decision Tree, and KNN are performed.

In this paper, pre-processing on raw data reviews are selected basically from Coursera, online courses. The aim of the study on the paper is a famous online learning platform; Coursera is an education stage that partners with top universities and organizations around the world and offer online courses in many different fields for everyone to take. And next, a token is built for classification works. For creating sentiment tokens, word correlation is applied. Coursera uses traditional text-mining and sentiment token constructions to extract functionality.

3. METHODOLOGY

In this section, pre-processing, POS tagging, sentiment classification with Bernoulli Naïve Bayes algorithm is discussed. Before the preprocessing step, the data are firstly collected from the Coursera website. After data collection and pre-processing steps, the system predicts the positive and negative sentiment polarity by Bernoulli NB. The system design is shown in Figure 1.

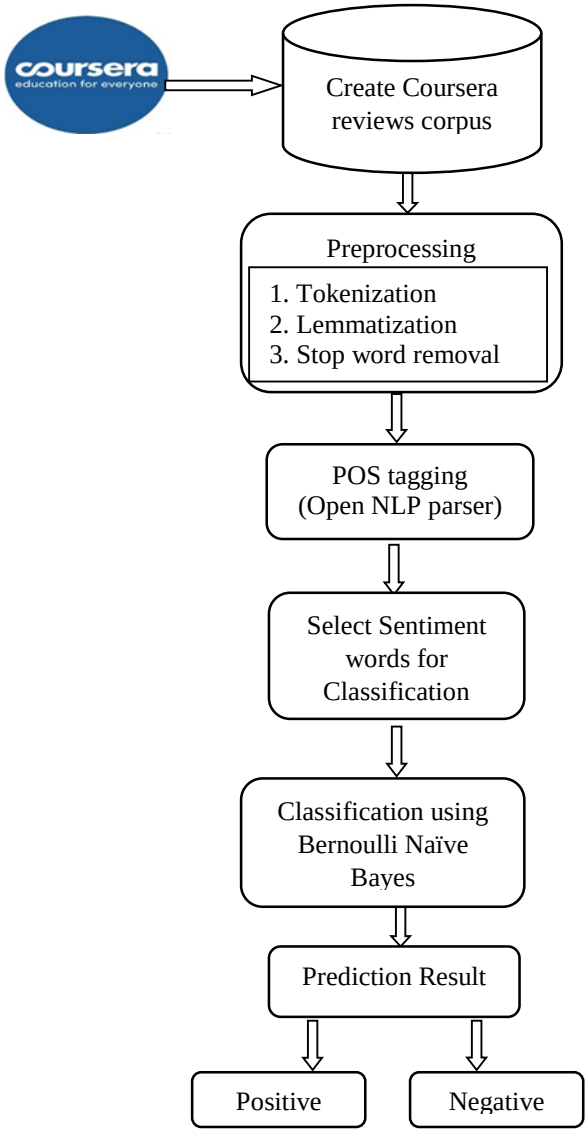

Figure 1. System design

3.1. Creating Coursera Reviews Corpus

One of the important tasks of system is the collection of the data. Similarly, important is to identify sources of the data. This system considered Coursera as a source of data, we have used the same data set of the University as used in state-of art work by Mohammad Aman Ullah, where they have collected data from different universities and the students are allowed to express their opinions in free of context about the lectures [8].

Experts made to gather positive and negative reviews from various sources that meet the criteria. According to we create our custom corpus for the review data of Coursera online learning by using NLTK. NLTK supports many corpora (for example, “Movie Reviews” corpus provides the 2K movie reviews for the sentiment polarity classification). But there is no corpus for the Coursera data, so we create our own corpus. To create the custom corpus, the path for the review must be within the *nltk. data, path,*

so that, NLTK can found the review data to be created. We firstly download the NLTK and then give the path to create the corpus in the NLTK. We added positive and negative reviews to the corpus as the .csv format.

3.2. Preprocessing

After data collection, the preprocessing step is processed. In the preprocessing, NLTK (Natural Language Toolkit) is used for tokenizing and stop word removal. Pre-arranging data can reduce the complexity of computational changes and produce higher-quality text classification. Procedures for preprocessing include the following steps:

3.2.1. Tokenization

In the tokenization step, the PunktSentence Tokenizer from NLTK is used to tokenize the sentences from raw data on the whole corpus. It divides the review document into small parts called tokens. For example,the review sentence, “I love Coursera. It is a great learning platform for the learners”.

The PunktSentenceTokenizer tokenizes the sentence and adds to the list and name the list name into *tokenize_word*:

tokenize_word= [‘I’, ‘love’, ‘Coursera’, ‘.’, ‘It’, ‘is’, ‘a’, ‘great’, ‘learning’, ‘platform’, ‘for’, ‘the’, ‘learners’, ‘.’]

3.2.2. Lemmatization

In this step, WordNet Lemmatizer is used for grouping the different inflected forms of a word to be analyzed as a single item. Lemmatization is similar to stemming and it does the morphological analysis of the words.For example, we lemmatized the word learn: the original word maybe learn, learning, learned, learned.

3.2.3. Stop Word Removal

It needs to remove stop words with the numbers, punctuation marks, and other elements removed from the dataset. Since it can help to reduce the time complexity of system processing.

Some of the extracted stop words are listed below:

['i', 'me', 'my', 'myself', , 'it', "it's", 'its', 'itself', 'they', 'them' , 'their', 'theirs', 'what', 'which', 'who', 'whom', 'this', 'am', 'i s', 'are', 'was', 'were', 'be', 'been', 'being', 'do', 'a', 'an', 'the', ' and', 'if', 'into', 'before', 'after', 'again', 'further', 'then', 'here ', 'there', 'more', 'other', 'some', 'such', 'so', 'can', 'will', 'just'].

3.3. POS Tagging (Open NLP Parser)

Opinion mining offers several ways to estimate the products, services, or many other areas via feedback statements or reviews. The Open NLP parser can be responsible for defining the shape of a word in a response sentence. It is also called Part of Speech (POS) tagging. Open NLP analyzes each sentence and defines the type or format of each word. A word can be categorized into one or more of a set of lexical or part-of-speech. A response sentence contains many words or patterns such as nouns, verbs, adjectives, and adverbs, articles, etc. The function of the Part of Speech (POS) tagging is to identify each symbol in the response sentence with the corresponding index or word type. POS tagging is a grammatical tagging model. For example, "The Coursera is good", the output of a POS tagger would be the /AT Coursera/NN is/VB good/JJ. The detail description of POS tagging is shown in Table 1 [9].

Table 1. POS tagging description

Tag	Group	Description
NN	noun	Noun, singular
NNS	noun	Noun, plural
NNP	noun	Proper noun, singular
NNPS	noun	Proper noun, plural
VB	verb	Verb, base form
VBG	verb	Verb, present participle
VBD	verb	Verb, past tense
VBN	verb	Verb, past participle
JJ	adjective	Adjective
JJR	adjective	Adjective, comparative
JJS	adjective	Adjective, superlative
ADVP	Adjective	Adverb phrases
RD	Adjective	Adverb
AT	Article	Article

3.4. Sentiment Words Selection

After removing the stop words, sentiment words selection is completed. All the words of reviews are compared to vocabulary for mood signals, built to determine whether or not the mood is a word for reviews. Even if the review contains at least one sentiment word, that review is stored for processing. If there is no sentiment word in the reviews, that review is rejected.

The extraction of words from the created coursera_review corpus and it shows that there are 1583820 words for the 3000 reviews and contains two categories for positive sentiment words and negative sentiment words. Example positive and negative sentiment words in coursera_review corpus is shown in Table 2.

Table 2. Example sentiment words of Adjective, Verb, Adverb, and Noun

	Positive	Negative
Adjective	good, interesting, creative, excellent, helpful, free of charge, thank, useful	bad, boring, lazy, complicated, difficult, useless,
Verb	attract, encourage	fail
Adverb	much	hardly, not, never, unfortunately
Noun	enthusiasm	waste

3.5. Bernoulli Naïve Bayes Classifier

The Bernoulli Naïve Bayes classifier is a part of the family of Naïve Bayes and its main feature is that it assumes all the features are binary that takes only two values (0s and 1s). It is especially popular for classifying

short texts. It is fast and accurate and able to make real-time predictions. It is used for discrete data and works on Bernoulli distribution.

For a random variable 'x' in Bernoulli distribution:

$$p(x) = \begin{cases} q = 1 - p, & x = 0 \\ p \quad , & x = 1 \end{cases}$$

Where 'p' is the probability of success and q is the probability of failure and x can have two values (0 or 1). In this system, the two values represent the sentiment polarity as negative or positive in the given review.

Bernoulli Naïve Bayes Classifier is based on the Bernoulli distribution and it estimates the probability of whether a sentiment polarity is or not in the given review. The Bernoulli Naïve Bayes task is to calculate the probabilities of each word found in the review and it is shown in equation (1).

$$p(x_i|y) = p(i|y)x_i + (1 - \mathrm{p(i|y)})(1 - x_i) \quad (1)$$

Therefore, the Bernoulli Naïve Bayes classifier calculates the probability $p(x_i|y)$ for whether a sentiment words are in the review or not. It classifies all the words in the review sentence and predicts the relevant class: "Positive" or "Negative" according to the final calculation of the result. If a sentiment polarity gives a greater positive value than a negative value, it is classified as a positive review or negative otherwise.

4. RESULT AND EVALUATION

In this section, the dataset, experiments, evaluation, and experimental setup are explained.

4.1. Dataset

The review data in this paper is collected from thirteen free online courses of Coursera websites that we have attended during the COVID-19 pandemic in 2020. Example sentences and their polarity are shown in Table3.

Table 3. Example review sentences and their polarity

No	Review text	Sentiment polarity
1	This course is excellent and refreshes the grammar and punctuation.	positive
2	It was good and easy.	positive
3	Very bad quality content and very short.	negative
4	The Course content is creative and the lecturer is very knowledgeable.	positive
5	This course is very useful to design webpages with JavaScript, CSS, and HTML. I'm interested in this course.	positive

The data manually collected only the year 2020 period among several years and stored with .csv format. It contains 3000 reviews in Table 4 and is divided into training and testing datasets.

Table 4. Number of reviews from different courses

No.	Course name	No. of review
1	Intermediate Relational Database and SQL	34
2	Grammar and Punctuation	536
3	Object-Oriented Programming with Java	129
4	Big Data, Artificial Intelligence, and Ethics	54
5	Advance Relational Dataset and SQL	30
6	Programming Foundations with JavaScript, HTML and CSS	506
7	Data Science in real life	20
8	Write Professional Emails in English	630
9	Introduction to Relational Database and SQL	50
10	Programming for every day (Getting Started with Python)	700
11	Programming fundamentals	200
12	Image Compression with K-Means Clustering	50
13	Computer Vision-Image Basics with Open CV and Python	61

4.2. Experimental Setup

For implementation, the system runs on Google Colaboratory (Colab) which is a python development environment using Google Cloud and it can support free GPU access. Other system requirements for implementation are scikit-learn, nltk, pandas, numpy, Python, RAM, System type, and Processor. In those systems, they have a different version and different descriptions. The version of scikit-learn is 0.22.2. post1 and it is an efficient tool for predictive data analysis. Version 3.2.5 of nltk describes a natural language toolkit to build Python programs and it works on human language data. In panda's version 1.1.2, the Software library was written in Python for data manipulation and analysis. Version 1.18.5 of NumPy defines an open-source numerical Python library and is used for mathematical operations on arrays. Python's version 3.6.9 interprets high-level general-purpose programming language. As for RAM version 8.0GB, System type version 64-bit OS, Windows 10, and Processor version Intel®Core™ i7-4770 CPU@3.40GHz have no more description in the experiment.

4.3. Evaluation Metrics

To test the performance of the implementation methods of the system, the positive and negative reviews are calculated on precision, recall, and f1-score values using the equations given in (2), (3), and (4), to know the

performance of each model. Confusion matrix for binary classification is presented in Table 5, where TP is True positive, FP – False positive, FN – False negative, and TN– True negative.

Table 5. Confusion matrix for binary classification

	Positive	Negative
Positive	TP	FP
Negative	FN	TN

$$Precision(P) = \frac{TP}{TP+FP} \quad (2)$$

$$Recall(R) = \frac{TP}{TP+FN} \quad (3)$$

$$F1 = 2.\frac{P.R}{P+R} \quad (4)$$

Figure 2 shows that the confusion matrix for the prediction for the positive and negative sentiment polarity. '0' means for negative sentiment and '1' for positive sentiment. 0-0 in the confusion matrix means prediction sentiment is negative where true label for the sentiment is also negative. So, the sentiment classification for the review is correct and the accuracy will increase.

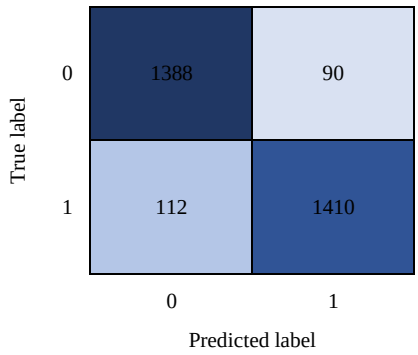

Figure 2. Confusion matrix of prediction on sentiment polarity of 0 (negative) or 1 (positive)

4.4. Evaluation

For evaluation, the system tested with different classifiers like BernoulliNB, MultinomialNB, and KNeighbors Classifiers on Coursera dataset and obtained results (such as precision, recall, and f1-score values) of the system are shown in Table 6, Table 7, and Table 8, where positive sentiments have a label of 1, and negative sentiments have a label of 0. We evaluate the training (80%) and testing (20%) datasets.

Table 6. Positive and negative classes on BernoulliNB classifier

Label	Precision	Recall	F1-score
0	0.90	0.99	0.94
1	0.99	0.89	0.94

Table 7. Positive and negative classes on MultinomialNB classifier

Label	Precision	Recall	F1-score
0	0.95	0.90	0.93
1	0.91	0.95	0.93

Table 8. Positive and negative classes on KNeighbors classifier

Label	Precision	Recall	F1-score
0	0.51	0.64	0.57
1	0.55	0.42	0.47

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+TN+FP} \quad (5)$$

According to the comparison using the equation given in (5), the accuracy of BernoulliNB reaches 94%, MultinomialNB is 93% and KNeighbors Classifiers gets 52%, respectively. Therefore, the accuracy for the BernoulliNB classifier on Coursera data gets higher than MultinomialNB and KNeighbors. Different accuracies on these classifiers are shown in Figure 5.

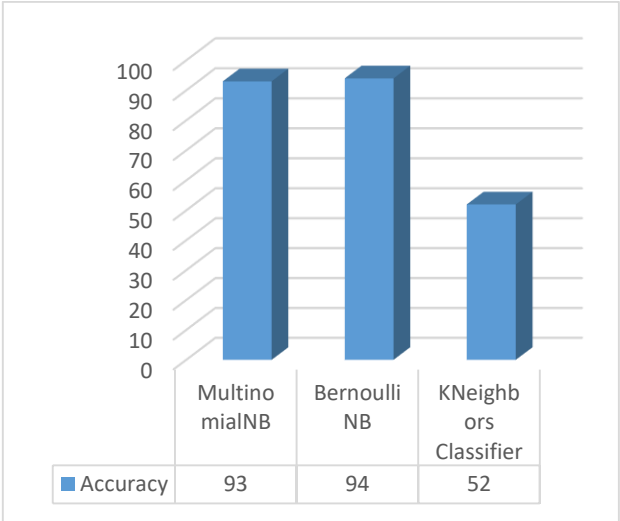

Figure 3. Accuracy Comparison

5. CONCLUSION

Today, using social media use of the internet is growing rapidly in the daily lives of almost everyone across Myanmar. These are also great ways to use the resources of technology for development, especially in education. The system develops a sentiment model by using Bernoulli Naïve Bayes classifier. The system especially aims for the students who are learning in the online learning system. By developing the sentiment tool for this purpose, they can know the opinion of the student and can decide whether to learning or not in this on the specific course. Bernoulli Naïve Bayes classifier is constructed for sentiment classification of online' user reviews into the negative or positive class and it is better than other methods. Therefore, the system develops the

sentiment model for learning and it will be very helpful and effective for the future development of an online learning system. For the future work, we aim to develop sentiment model on different domain such as movie domain, product domain and political domain by using different machine learning algorithms.

ACKNOWLEDGMENT

I would like to special thanks to our Rector Dr. Ei Ei Hlaing, our head of department Dr. Hnin Pwint Phyu, and deputy of the head of the department Dr. Shwe Thinzar Aung (Faculty of Computer Science) for their invaluable support, encouragement, supervision, and useful suggestions throughout this paper. And I would like to appreciate to my sister Daw Win Lei Kay Khine, for guiding and giving me the good idea, strength, and motivation to continue this paper.

REFERENCES

[1] Z. Kechaou, M. B. Ammar, A. M. Alimi, "Improving e-learning with sentiment analysis of user's opinions", 2011 IEEE Global Engineering Education Conference (EDUCON), 2011.

[2] L. C. Seng, T. T. Hok, "Humanizing E-learning", International Conference on Cyberworlds, Singapore 2003.

[3] H. Giroire, F. Le Calvez, G. Tisseau, "Benefits of knowledge-based interactive learning environments: A case in combinatorics", Proceedings of the Sixth International Conference on Advanced Learning Technologies, 2006, pp- 285-289.

[4] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques", Proceedings of the conference on empirical methods in natural language processing (EMNLP 2002) Philadelphia, PA, USA, 2002, pp- 79-86.

[5] H. Parveen and S. Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm", 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2016, pp: 416-419.

[6] N. V. Kolekar, G. Rao, S. Dey, "SentimentAnalysis and Classification using Lexicon-based approach and Addressing Polarity Shift Problem", JATIT, Vol 90, No.1, 15th August 2016.

[7] Y. M. Aye and S. S. Aung, "Senti-Lexicon and Analysis for Restaurant Reviews of Myanmar Text", IJAEMS [Volume 4, issue-5, May 2018], ISSN: 2454-1311.

[8] M.A. Ullah. "Sentiment Analysis of Students feedback: A Study towards Optimal tools", International Workshop on Computational Intelligence (IWCI). 12-13 December 2016, pp-175-180.

[9] N. Soe, P. T. Soe, "Domain Oriented Aspect Detection for Student Feedback System", 2019 International Conference on Advanced Information Technologies (ICAIT), 2019, pp-90-94.

Detection of Diabetes Using Classification Methods

Phyu Thwe[1], Cho Cho Lwin[2], Hnin Pwint Myu Wai[3]
[1]*Myanmar Institute of Information Technology,* [2,3]*University of Computer Studies (Myitkyina)*
[1]*phyu_thwe@miit.edu.mm,* [2]*cho_cho_lwin@miit.edu.mm,* [3]*hninpwintmyuwai14@gmail.com*

ABSTRACT*: Early diagnosis of diabetes is possible only with an adequate assessment of the symptoms of the most frequent and less frequent symptoms that arise at different stages from the onset of the disease to the analysis. Data mining classification techniques are widely used in diabetes prediction system. In this paper, the performance of the system is tested using four classifiers: decision tree, support vector machine, random forest, and multilayer perceptron. Bangladesh Diabetes Dataset and Pima Indians Diabetes Dataset are used to validate the performance and are from the UCI Machine Learning Repository. Various data mining algorithms have been used to assess the accuracy of diabetes detection. Classification methods are examined and compared based on classification accuracy, precision, recall and f-measures. The results are promising, with 98.08% and 76.30% accuracy on a 10-fold cross validation over dataset. These results are comparable to those obtained by other methods. This study suggests that the random forest algorithm is the best algorithm and outperforms it in accuracy.*

Keywords: Classification; Decision Tree; Multilayer Perceptron, Random Forest; Support Vector Machine.

1. INTRODUCTION

Diabetes is a chronic condition in which the body does not produce the right amount of insulin. Diabetes increases the risk of many illnesses, including heart disease, kidney dysfunction, and nervous system damage. The expert relies on past experience to diagnose a patient with diabetes. Several studies have been shown to analyze diabetes based on different parameters. There are three main types of diabetes: diabetes 1, 2, and gestational diabetes [7].

In this three type of diabetes, insulin-producing pancreatic cells are demolished by the body's defense system. Patients with type 1 diabetes should be prearranged after insulin injections, frequent blood tests, and dietary limitations. Obesity, overweight, and physical inactivity can lead to type 2 diabetes. It is also understood that the threat of diabetes increases with developing age. The majority of people with type 2 diabetes have borderline diabetes or pre-diabetes. This is a condition in which blood sugar levels are higher than normal, but not as high as in diabetics. Diabetes during pregnancy-a type of diabetes that inclines to occur in pregnant women due to their high sugar at ease because the pancreas does not produce enough insulin. Diabetes, for the measurement of blood glucose levels are not controlled, has been considered a serious health problem. Diabetes is not only affected by a variety of factors such as height, weight, genetic factors, and insulin, but sugar levels are the main reason for all of these factors. Early detection is the only way to avoid complications.

In this paper, a comparative study is introduced that improves the accuracy of classification of diabetics in Bangladesh. In this system, the generalized performance of Decision Tree, Random Forest, Support Vector Machine, and Multilayer Perceptron is presented to enhance the classification model of decision-making systems in identifying diabetes. The rest of this work consists of: the data mining classification algorithms are shown in Section 2. The methodology for this work is then shown in Section 3. Section 4 shows the results of the model, Section 5 shows the discussion of the system and the conclusions are shown in Section 6.

2. RELATED WORKS

Diabetes caused long-term complications and serious health problems called a non-communicable disease. The World Health Organization report [11] spoke of diabetes and its complication that have physically, economically and financial significant diabetes and its complications for families. Studies showed that about 1.2 million people die during unmanaged health stages. Approximately 2.2 million people had died from risk factors for diabetes such as cardiovascular disease.

This paper provided an overview and description of the various methods used to classify diabetes diagnoses across different datasets. These methods were researched and considered taking into account their advantages, problems and classification accuracy [2]. In this paper, a combined hybrid classification method for fuzzy Min-Max neural network, regression tree and random forest (FMM-CARTRF) was proposed, which achieved a PIDD accuracy of 78.39% [6]. This study showed three machine learning classification algorithms: SVM, Naive Bayes, and Decision Tree. They were used for diabetes detection at early stage.

In this paper [8], study of the Indian Pima dataset was accomplished using a variety of classification algorithms such as Naive Bayes, Logistic Regression, Zero R, MLP, Random Forest and J48. Assessment and likelihood of positive and negative diabetes from an accuracy and performance perspective, it was better to diagnose diabetes using a data mining tool that used the WEKA tool. This means that MLP was better in accuracy and performance.

People with diabetes should constantly monitor their blood glucose levels and change their insulin readings to get as close to normal as possible [4]. Blood levels that deviate from the typical collection can cause real complications now and in the long term. Scheduled waits

indicate that individuals who were warned of forthcoming changes in blood glucose would be authorized to take cautionary measures. This paper referred to a response that uses a mild physical model of blood glucose development to make bright places of interest of support vector regression representations created using specific acceptance information.

3. MATERIALS AND METHODS

In this paper, Decision Tree, Random Forest, Support Vector Machine and Multilayer Perceptron was taken about using the python programming language to predict the diabetes based on two diabetes dataset. The results then matched with other algorithms used by scikit-learn machine learning library.

3.1. Dataset Description

In this study, the two diabetes dataset from UCI Machine Learning Repository is used and maintained it with 520 instances and 17 attributes [10]. This dataset comprises the sign and indications data of newly diabetic or would be diabetic patient. They have been collected by via direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and accepted by a doctor. The detail of these datasets is shown in table 1. The other dataset is 768 instances and 9 attributes. This is Pima Indians diabetes dataset. The attributes name are pregnancies, glucose, blood pressure, skin thickness, insulin, bmi, diabetes pedigree function, age and outcome. All attributes are numeric values and class label, outcome, is 0 and 1.

Table 1. Attribute information

No	Attributes	Description
1	age	20-65
2	sex	No, Yes
3	polyuria	No, Yes
4	polydipsia	No, Yes
5	sudden weight loss	No, Yes
6	weakness	No, Yes
7	polyphagia	No, Yes
8	genital thrush	No, Yes
9	visual blurring	No, Yes
10	itching	No, Yes
11	irritability	No, Yes
12	delayed healing	No, Yes
13	partial paresis	No, Yes
14	muscle stiffness	No, Yes
15	alopecia	No, Yes
16	obesity	No, Yes
17	class	Positive, Negative

3.2. Support Vector Machines (SVM)

Support Vector Machines (SVMs) [5] is a classifier that classifies into examined machine learning models. SVM discovers the best hyper plane using up to two separating lines that are distinguishable from each other. The distance in the middle of them is the maximum possible distance, and the intermediate between these two margins is measured the final hyper plane. In SVM, the support vector is the fact that traces these boundaries. Noticeably, a good model will have smaller amount vectors to support. As a consequence, dependencies can be reduced if there is variability between data points.

3.3. Decision Trees (DT)

The decision tree has the highest simplicity and elasticity in the classification method, and this feature makes decision making easier to comprehend and makes it the most recognized classification method for information detection [9]. The decision tree generates a classification and regression tree model in the form of a tree structure by allocating the dataset into less significant subsets and at the same time spread out the related decision tree. A decision tree is a top-down construction with root nodes that boundaries branches that have a parent-child correlation. The tree consists of a root node, a leaf node that characterizes all classes, and an internal node that represents test conditions.

3.4. Random Forest (RF)

There was an explanation of a random forest system. This is part of the decision tree approach, as there is a random tree that is an ad hoc classification. At each sequence of the carrying method, a typical natural tree drive produces its own range matrix, often creating a important risk factor. The tree is as a final point fully sophisticated and is not cut. For new datasets, the tree structure is pushed down. When the command line node exits, training samples are allocated to tags. This procedure is recognized by the name of Random Forest invention and is accomplished in all forests [3].

3.5. Multilayer Perceptron (MLP)

MLP is a keep an eye on machine learning algorithm. As the name point toward, it has multiple layers. If the problem is direct, only one layer is needed, but for complex, non-linear problems, one layer of perceptron enhances more layers. This network is called a multi-layer perceptron. MLP is a direct opinion neural network with one or more hidden layers. The MLP contains three or more layers: an input layer, one or more hidden layers, and an output layer. The input layer uses a linear activation function with no starting point to cause input for following layers. However, the hidden and output layers have non-linear thresholds and activation functions [1].

4. EXPERIMENTAL RESULTS

The empirical evaluation is performed with different machine learning algorithm like Decision Tree, Random Forest, Support Vector Machine and Multilayer Perceptron. It is observed that in terms of performance, Random Forest classifier is more efficient than the other machine learning algorithms. The advantages of classification methods in the diagnosis of diabetes provide

in classifying multiple datasets to prove that the methods used can provide the best accuracy.

4.1. Classification Accuracy

The classification accuracy is one of the performance evaluation measures. Accuracy shows how well the classifier works when predicting instances based on training data.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

- TP (True Positive): No. of occurrences anticipated positive and are actually positive
- FP (False Positive): No. of occurrences anticipated negative and are actually negative
- TN (True Negative): No. of occurrences anticipated positive but are actually negative
- FN (False Negative): No. of occurrences anticipated negative but are actually positive

$$TP - rate = \frac{TP}{TP+FN} \quad (2)$$

$$FP - rate = \frac{FP}{FP+TN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F - Measure = \frac{2*Precision*Recall}{Precision+Recall} \quad (6)$$

$$weighted\ average = \frac{\text{Cyes}*(\text{TP}+\text{FN})+\text{Cno}*(\text{TN}+\text{FP})}{TP+FN+TN+FP} \quad (7)$$

Where Cyes represents class label yes and Cno represents class label no.

Figure 1 shows the design of a system for early detection of diabetes. The main process consisted of two stages: a classification stage and an evaluation stage. Data entry is a set of diabetes data. Diabetes detection model in the classification phase is run using a decision tree, support vector machine, random forest, and multilayer perceptron. The evaluation phase aims to evaluate and improve the accuracy of the model. A 10-fold cross-validation was applied and a model was created.

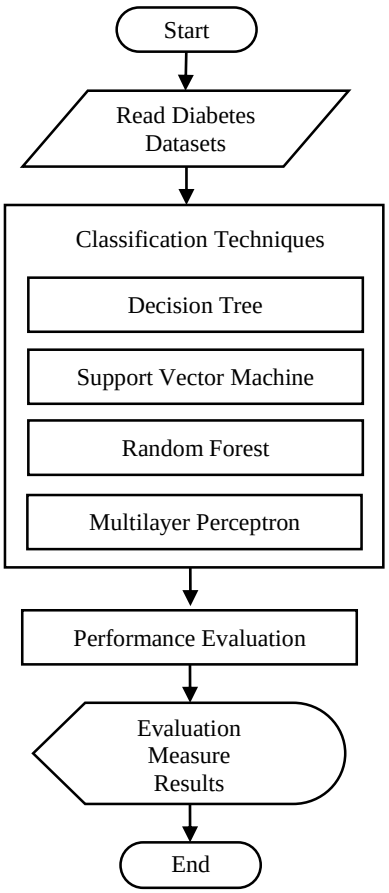

Figure 1. System flow diagram

Table 2 represents the classification accuracy results of four classifiers that is defined in equation 1.

Table 2. Classifiers accuracy

Classifiers	Accuracy Value (Sylhet Diabetes)	Accuracy Value (Pima Indians Diabetes)
Decision Tree	97.50	70.05
Support Vector Machine	61.54	76.17
Random Forest	98.08	76.30
Multilayer Perceptron	93.08	69.01

Figure 2. Performance of classification algorithms for sylhet diabetes

Table 3. TP rate of classifiers

Classifiers	Positive	Negative	Weighted Average
Decision Tree	0.981	0.965	0.975
Support Vector Machine	0.489	0.908	0.762
Random Forest	0.981	0.980	0.981
Multilayer Perceptron	0.941	0.915	0.931

Table 4. FP rate of classifiers

Classifiers	Positive	Negative	Weighted Average
Decision Tree	0.035	0.019	0.029
Support Vector Machine	0.092	0.511	0.365
Random Forest	0.020	0.019	0.020
Multilayer Perceptron	0.085	0.059	0.075

Table 5. Precision of classifiers

Classifiers	Positive	Negative	Weighted Average
Decision Tree	0.978	0.970	0.975
Support Vector Machine	0.740	0.768	0.758
Random Forest	0.987	0.970	0.981
Multilayer Perceptron	0.947	0.906	0.931

Table 6. Recall of classifiers

Classifiers	Positive	Negative	Weighted Average
Decision Tree	0.981	0.965	0.975
Support Vector Machine	0.489	0.908	0.762
Random Forest	0.981	0.980	0.981
Multilayer Perceptron	0.941	0.915	0.931

Table 7. F-measure of classifiers

Classifiers	Positive	Negative	Weighted Average
Decision Tree	0.980	0.967	0.975
Support Vector Machine	0.589	0.832	0.747
Random Forest	0.984	0.975	0.981
Multilayer Perceptron	0.944	0.910	0.931

5. DISCUSSIONS

To test the performance of the random forest and to make sure consistent performance, a k-fold test for 10 iterations was implemented to determine the robustness of the proposed structure. The data in Table 2 show the accuracy and show that the performance of the random forest is very reliable and can be chosen to predict diabetes using the proposed predictive system.

Table 3 shows the true positive rate of four classifiers using equation 2 and weighted average of both classes defined in equation 7. Random forest classifier has acceptable true positive rate. Table 4 shows the false positive rate of four classifiers using equation 3 and weighted average of both classes defined in equation 7. Table 5 shows the precision of four classifiers using equation 4 and weighted average of both classes defined in equation 7. Table 6 shows the recall of four classifiers using equation 5 and weighted average of both classes defined in equation 7. Table 7 shows the F-measure of four classifiers using equation 6 and weighted average of both classes defined in equation 7.

For Sylhet diabetes dataset, random forest classifier shows the best accuracy (98.08%). From Table 2 it is obvious that the random forest accuracy with pima Indians diabetes dataset is well than other classifiers accuracies namely (76.3%). Results show that the accuracy of all four classifiers and the best accuracy with random forest achieved by both dataset (98.03% and 76.30%).

6. CONCLUSIONS

This paper presents a case study in which machine learning algorithms have been trained to better predict diabetes based on patient records. In this paper, an early diabetes prediction experiment and Pima Indians diabetes are conducted using four classifiers based on a decision tree of a machine learning algorithm, a support vector machine, a random forest, and a multilayer perceptron classifier. The four classifiers are compared based on accuracy values. Another method of evaluating effectiveness was to measure the accuracy of classifiers including TP rate, FP rate, precision, recall, and f-measure. The overall effectiveness of the random forest classifier for predicting diabetes outperforms all other classifiers based on two diabetes datasets.

REFERENCES

[1] Garima Verma, Hemraj Verma, "A Multilayer Perceptron Neural Network Model For Predicting Diabetes", International Journal of Grid and Distributed Computing Vol. 13, No. 1, 2020.

[2] Dilip Kumar Choubey, Sanchita Paul, "Classification techniques for diagnosis of diabetes: A review", International Journal of Biomedical Engineering and Technology, 2016

[3] K.Koteswara Chari, M.Chinna babu, Sarangarm Kodati, "Classification of Diabetes using Random Forest with Feature Selection Algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-1, November 2019

[4] Kevin Plis, Razvan Bunescu, Cindy Marling, Jay Shubrook, Frank Schwartz, "A Machine Learning Approach to Predicting Blood Glucose Levels For Diabetes Management", AAAI Publications, Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

[5] Manisha Singla, Samyak Soni, Prateek Saini, Abhishek Chaudhary, K. K. Shukla, "Diabetic Retinopathy Detection Using Twin Support Vector Machines", Advances in Bioinformatics, Multimedia, and Electronics Circuits and Signals, Proceedings of GUCON 2019, Springer, 2019.

[6] Manjeevan Seera and Chee Peng Lim, "A hybrid intelligent system for medical data classification", Expert System with Applications, vol. 41, Issue 5, 2014.

[7] N. Sneha and Tarun Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection", Journal of Big Data, Springer, 2019

[8] Saman Hina, Anita Shaikh, Sohail Abul Sattar, "Analyzing diabetes Datasets Using Data Mining", Journal of Basic and Applied Sciences, 2017.

[9] Saman Hina, Anita Shaikh, Sohail Abul Sattar, "Analyzing diabetes Datasets Using Data Mining", Journal of Basic and Applied Sciences, 2017.

[10] Diabetes Dataset from UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset

[11] "Global report on diabetes", World Health Organization, 2016. https://www.who.int/diabetes/global-report/en/

Performance Analysis of Classification Algorithms in Data Mining Technique

[1]Aye Mon Win, [2]Yu Yu Khaing, [3]Lei Yi Htwe
[1, 2, 3] *Faculty of Information Science, University of Computer Studies, Pathein*
[1]*ayemonwin@ucspathein.edu.mm*, [2]*yuyukhaing@ucspathein.edu.mm*, [3]*leiyihtwe@ucspathein.edu.mm*

ABSTRACT: *In this information age we are in now, a large amount of data is gathered everyday. Analysis of such data has become an important task. Data mining is the process of discovering the knowledge from a huge amount of data. Different algorithms and techniques are now available in data mining. It is very important to extract these valuable pieces of information and classification is one important technique, which can help categorize the data in some predefined classes. Different ways of data mining can be accessed for classification algorithms. This paper discussed various classification algorithms, such as Naive Bayes classifier, Bagging Classifier, OneR rule-base classifier and Decision Tree (J48) classifier. The main factor of this paper is to present how the outcome is different when working with binary class or multiclass dataset. This paper will also present along with how time and accuracy can affect their performance when the number of instances is decreased.*

Keywords: Data Mining, Decision Tree, Naive Bayes, Bagging, Rule-base

1. INTRODUCTION

The increase in data has become more rapid in these days. Data Mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems [9]. It, therefore, extracts the meaningful information from the dataset provided. Although there is an increase in the amount of available information, there has been a decline in understanding that data since the useful information has to be looked deeper in those irrelevant layers of data and we have to trust our intuition instead of informed logistics. However, data mining helps us extract useful data from a large amount of data, which leads to making informed decisions. When analyzing data, tools for data mining can unveil data patterns and their RELATION which are of great help to knowledge base, business strategies, and medical and scientific research. Because of the broadening gap between data and information, data mining tools have been developed systematically and they will transform data tombs into 'golden nuggets' of knowledge [2]. This paper will present two datasets such as Diabetes and Soybean. Diabetes is a group of disease that influence how blood sugar (glucose) is used by the human body because glucose is an essential source of energy for the muscle and the tissue producing cells and is therefore necessary for health [1].Soybean is one of the crops cultivated. Soybean seeds are rich in amino acids, lipids, vitamins, and minerals and are an abundant source of proteins, constituting a key crop for global food security. Achieving high soybean yield depends on successful establishment of soybean plants, which requires use of high quality seeds [13].

2. RELATED WORK

In real world, many classification algorithms are used in different applications. A number of researchers have implemented and developed on the use of data mining techniques in the various classification algorithms. Some works are reviewed in the following.

The author, Amit Gupta, Ali Syed, Azeem Mohammad, Malka N. Halgamuge [4] proposed about analysis of application and performance of six classification algorithms that produce different results. This experiment showed that JRip is good in accuracy and Naive Bayes is good in training time than the other classification algorithms. S. Asha Kiranmail and A. Jaya Laxmi [14] they presented that the classification algorithms are analyzed on the basis of accuracy and precision by taking the dataset and also present the comprehensive evaluation of different classifiers of WEKA. Dr Hemlata Chahal [5], Dr. Vaishali S. Parsania, Dr. N. N. Jani and Navneet H Bhalodiya [6] the authors discussed the Naive Bayes, Bayes Net, PART, JRip and OneR learner and present the accuracy is important for the classification algorithm. They proposed the comparison of the various classification algorithm of the data mining.

3. MACHINE LEARNING

Machine Learning is a method of data analysis that investigates how the machine automatically learns in order to predict the result of an unseen data based on previous observation. In data mining, various machine learning algorithms are used for the classification. Association, classification and clustering are popular in machine learning. Some of the machine learning strategies are:

Supervised learning strategies: Supervised learning is basically classification. It is a function of data mining from labeled data on **training**. The monitoring of the learning outcome associated with the training data set on the labels [2].

Unsupervised Learning Strategy: Unsupervised learning is a clustering algorithm, that to find hidden structure in unlabeled data [2].

4. CLASSIFICATION IN DATA MINING

The most important data mining technique is classification which works on the basis of machine learning and it assigns items to aim for groups or classes.

The aim of classification is to be able to tell the target class correctly for every single case of the data. This technique is also useful for categorizing the data item set in the predefined set of classes or categories. In this method, a model which can be taught to categorize the data items is built, which is the first step of classification and the construction is based on some training data set. This step is followed by the step where the model classifies an unknown tuple into a class label. Several model techniques are used in classification and some of the most well-known algorithms are Decision Tree, Naive Bayes, bagging and rule base classifier.

4.1. Decision Tree

A decision tree is a supervised machine learning Algorithm. It is a flow-chart-like tree structure. Here, each of the internal nodes represents a test on an attribute and each branch is the result of this test [8]. The decision tree structure provides an explicit set of if-then rules rather than abstract mathematical equations, making the results easy to interpret [7]. In decision tree all the attributes are organized in the form of a tree. The leaf nodes in the tree represent the classes where the instances are grouped into after being taken one by one. Therefore, decision tree divides the entire input space into different cells in which every cell becomes a part of one class as defined in the class label. The division is called a series of tests. The popular Decision Tree algorithms are ID3, C4.5, J48 and CART. J48 decision tree algorithm will propose in this system.

4.1.1. J48 Algorithm

J48 is the improved versions of C4.5. The output result of J48 is the decision tree. A tree structure has different nodes, such as root node, intermediate nodes and leaf node [10]. Each node contains a decision and that decision gives on to decision tree.

Advantages

1. A Decision trees model makes understand the classification process easily and the classification rules.
2. Decision trees are not cost efficient.
3. Data preparation needs less effort during pre-processing compared to other algorithms.
4. NO impact of missing values is caused on the decision tree construction to a large extent.

Disadvantages

1. Decision Trees cannot be used directly when information is provided in a series of amounts.
2. Larger number of classes with a large amount of data that leads to large decision tree can hinder comprehensibility.
3. Calculation can be more complicated in Decision Tree compared to other algorithms.
4. It takes time to train the model.

4.2. Naive Bayes Classifier

The Naive Bayes classifier is a simple and supervised machine learning algorithm. It is based on Bayes' theorem with the independence assumptions between predictors. Moreover, this model is easy to build and the value of attribute on a given class is independent of the other attributes value. It is useful for very large datasets and it can calculate posterior probabilities for hypothesis and it is robust to noise in input data. In spite of its clarity, the Naive Bayesian classifier is often amazingly well and it is widely used because it performs advanced classification methods.

Advantages

1. Naive Bayes is simple and easy to implement.
2. It does not require a large amount of training data to estimate the test data. So, the training time is less.
3. When decision making, Naive Bayes classifiers are computationally fast.

Disadvantages

1. The presumption of independence between attributes is not always true and therefore the accuracy of Naive Bayes classifier is unstable [3].

4.3. Bagging Classifier

Bagging is a machine learning algorithm. It is designed to improve the stability and accuracy of machine learning algorithms used in classification and regression [11]. Bagging is usually applied to decision tree procedure. This algorithm can be an effective approach to reduce the variance of a model, to prevent over fitting and to increase the accuracy of unstable models [11]. It is suitable for the training data.

Advantages

1. Bagging takes ensemble learning where in multiple weak learners outperform a single strong learner [12].
2. It helps reduce variance and thus helps us avoid over fitting [12].

Disadvantages

1. It is loss of interpretability and Computational complexity.

4.4. Rule-based classifier

Rule-based classifiers are another form of classification method. It identifies the data by using various "if..else" rules. The rule is made by the combination of attributes. The result of rule is a positive or negative classification. This system will present OneR rule-based classifier. It is a simple classification algorithm. OneR is able to deduce typically simple, yet precise, classification rules from a set of instances. OneR is also able to handle missing values and numeric attributes showing flexibility in spite of simplicity [6].

Advantages

1. It is easy to interpret and generate.
2. It can classify neo instances rapidly.
3. It can easily handle missing values and numeric attributes [1].

Disadvantages

1. It can produce rules with very small coverage.

5. EXPERIMENT AND RESULTS

This paper focuses on the classification of algorithms using Waikato Environment for Knowledge Analysis or in short, WEKA tool. The result of this experiment proves that the numbers of instances are decreased in a dataset over the performance accuracy of various classification algorithms. The datasets are taken from UCI Machine Learning Repository library. The "Decision Tree (J48)" algorithm, "Bagging" algorithm "Naive Bayes" classification algorithm and "OneR" algorithm were used for this experiment. In this experiment, classification algorithms evaluated using the accuracy and training time. Accuracy is defined as the percentage of correct predictions made by the total number of instances in a classification algorithm.

The formula of accuracy is

$$\text{Accuracy}= \frac{TP+TN}{P+N}$$

where, TP = True Positives: refers to the positive tuples that were correctly labeled by the classifier,

TN= True Negatives: refers to the negative tuples that were correctly labeled by the classifier,

P= no. of positive tuples,

N=no. of negative tuples.

Training time is referred to the time that an algorithm takes to build a model on datasets. Minimum training time is desirable.

This experiment mainly focuses on the test of the number of instances that are decreased on binary class or multiclass data. Binary classification refers to predicting one of two classes and multiclass classification involves predicting one of more than two classes .In this testing, two types of datasets with various properties can be seen.

Table 1. Composition of the datasets

No	Datasets	instances	Attributes	class
1	Diabetes	768	9	2
2	Soybean	683	35	19

In table 1, it can be seen, firstly, that the number of classes are different in forms. Secondly, the numbers of instances decrease and finally, that run through the various algorithms. The numbers of instances are decreased 10 percent at each time. The data sets are obtained for the accuracy following results. It can be seen as follows:

Table 2. Accuracy for diabetes

Decrease the no. of instances Diabetes	No. of instance	J48 (%)	Naive Bayes (%)	Bagging (%)	OneR (%)
100%	768	85.1	80.2	89.8	76.8
90%	691	80.9	77.8	89.6	76.4
80%	614	82.1	77.6	89.4	75.9
70%	538	82.1	77.2	89.2	75.4
60%	461	78.3	74.6	89	74.6
50%	384	76.4	74.3	88.6	74.3
40%	307	79.2	74.9	87	74.6
30%	230	78.7	73	86.5	77.4
20%	154	77.4	74.2	86.2	75.5
10%	77	75.6	71.8	85	79.5

Table 3. Accuracy for soybean

Decrease the no. of instances Soybean	No. of instance	J48 (%)	Naive Bayes (%)	Bagging (%)	OneR (%)
100%	683	97.3	94.2	92.8	40.8
90%	614	96.3	93.9	92.8	40.2
80%	546	96.2	93.1	92.3	43.6
70%	478	95.6	92.5	90.6	44.7
60%	409	95.8	92.7	91.4	38.1
50%	341	95	92.5	87.4	38.1
40%	273	95.6	89.2	89.3	40.4
30%	204	97.5	88.2	96.6	49
20%	136	97.8	83.8	97.6	59.6
10%	68	76.5	83	97.8	89.9

The result shows that, the performance of decrease in number of instances on binary class and multiclass of the decision tree classifier, Bagging, OneR and Naïve Bayes classifier. After testing the training datasets, four algorithms come out in different forms.

The performance of OneR classifier was very poor when the number of classes was large (in soybean). Its performance was even poor than the baseline that provide that OneR was not good at all for the multiclass datasets. The performance improvement is better when the number of classes was binary class (in Diabetes). Moreover, its accuracy and the time of execution increase as the number of instances decreases. Also, the performance of OneR was good and more consistent in the

binary class Diabetes as compared to the multiclass Dataset (in Soybean).

Table 4. Training time for diabetes (second)

Decrease the no of instances Diabetes	No. of instance	J48	Naive Bayes	Bagging	OneR
100%	768	0.07	0.02	0.13	0.05
90%	691	0.07	0.05	0.09	0.03
80%	614	0.06	0.02	0.07	0.02
70%	538	0.06	0.03	0.11	0.02
60%	461	0.05	0.03	0.11	0.02
50%	384	0.05	0.02	0.09	0.01
40%	307	0.03	0.02	0.09	0.01
30%	230	0.03	0.02	0.07	0.01
20%	154	0.02	0.01	0.05	0.0
10%	77	0.01	0.01	0.03	0.0

Table 5. Training time for soybean (second)

Decrease the no of instances soybean	No. of instance	J48	Naive Bayes	Bagging	OneR
100%	683	0.13	0.02	0.18	0.05
90%	614	0.11	0.02	0.18	0.05
80%	546	0.13	0.02	0.11	0.05
70%	478	0.09	0.01	0.11	0.03
60%	409	0.09	0.01	0.09	0.03
50%	341	0.06	0.01	0.09	0.02
40%	273	0.03	0.01	0.06	0.01
30%	204	0.02	0.01	0.03	0.0
20%	136	0.01	0.01	0.02	0.0
10%	68	0.01	0.01	0.01	0.0

Therefore, it is beneficial to use OneR only if it is a binary class or has a small number of classes and the small number of instances. In Diabetes dataset, the performance increased a little when 90% of dataset was used as compared to 100 %. In this WEKA tool, the training of the build model shows in second. So, in this experiment, their training time of OneR classifier has at least 0.0 second. If it is mini second, it may present 0.001 second.

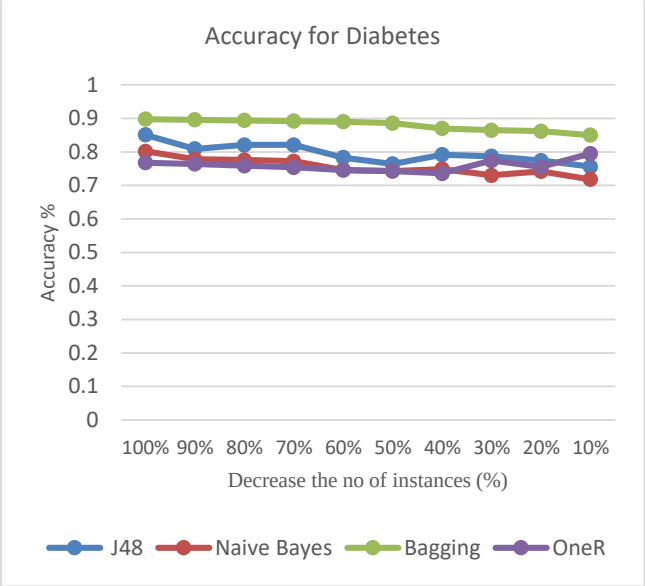

Figure 1. Accuracy of the diabetes dataset

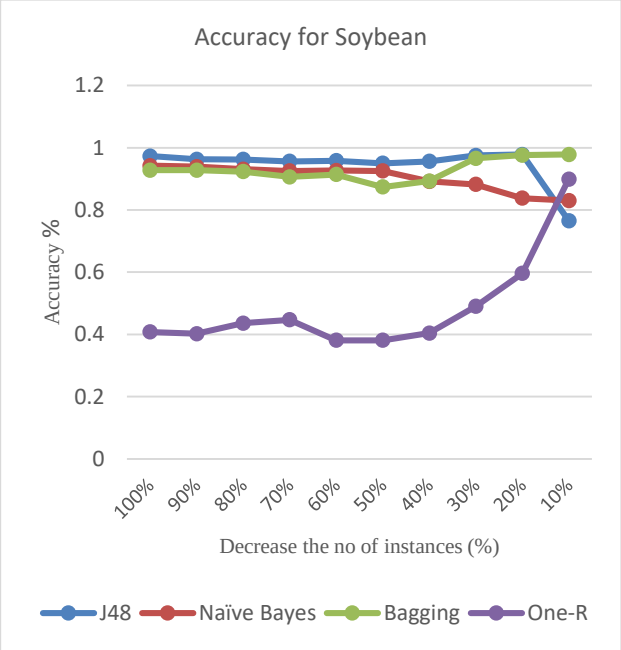

Figure 2. Accuracy of the soybean dataset

Bagging classifier performed as a standard trend where it increases for a few readings and then started decreasing. It performed the same trend in both two cases. As a result, it gives the result in accuracy for the reading with more amount of dataset and finally starts decreasing as the size of dataset decreases. However, the training time of the build model is less than the other classifier.

Decision Tree classifier (J48) depends on the size of dataset and their performance comparatively decreases first and then increases as the number of instances were decreased as in Soybean and Diabetes. The accuracy of decision tree is 75-85% for diabetes and 76-97% for soybean respectively. The time of the build model is better than the bagging classifier.

Naive Bayes classifier also depends on the size of datasets. However, the accuracy of the multiclass class datasets (Soybean) is better than binary class (Diabetes). And also, its time taken for build model is good for binary class datasets. However, when the number of instances was decreased the training time of the build model was small. Sometimes the accuracy of Naive Bayes classifier is unstable.

The performance of all the algorithms decreases dramatically at the end because only 10 percent dataset is remaining that does not provide sufficient information to classify the dataset.

6. CONCLUSION

In this paper, the result of the various classifiers performed differently with the decrease in the number of instances. It also presents how the outcome is different when working with binary class or multiclass dataset. OneR works poorly with multiclass data but increase its performance with decrease in number of instances in a binary class. When the numbers of class are decrease, the accuracy and time are good. Its performance is good for binary class. So, OneR method should be used in binary class datasets.

Bagging follows the trend where the accuracy increases at the start but then gradually decreases as the size of the dataset decreases. However, its time performance does not really bad, so it should be used only when the timing is not a major concern.

In decision tree, the performance first decreases and later on increased point out that decision tree is dependent on the size of datasets. So, it is advisable not to use it when the size of dataset is variable.

Naive Bayes depends on the size of dataset. The accuracy of the multiclass datasets is more proper than binary class. Sometimes, Naive Bayes accuracy is not stable. So, it should not be used when the size of dataset is variable.

The comparison of four different classification algorithms, OneR classifier can be used in binary class datasets. In addition, Naive Bayes and Decision tree (J48) can be used in binary class datasets. Bagging classifier tends to be the same for the binary class or multiclass datasets.

7. FUTURE WORK

There are also many classification algorithms that can be tested for the effect of decreasing the number of instances. The effect of the same class can also be checked within the various algorithms e.g., J48 algorithm, C4.5 algorithm, and ID3 algorithm of the decision tree.

REFERENCES

[1] https://www.mayoclinc.org/diseases-conditions/diabetes/symptoms-cause/syc-203711444, “Diabetes”

[2] Jiawei Han, Micheline Kamber, Jian Pei “Data Mining Concepts and Concepts Techniques” Book, Third Edition

[3] Medeswara Rao, Kondamudi, Sudhir Tirumalasetty, "Improved Clustering And Naïve Bayesian Based Binary Decision Tree With Bagging Approach, International Journal of Computer Trends and Technology (IJCTT) " volume 5 number 2 -Nov 2013

[4] Amit Gupta, Ali Syed, Azeem Mohammad, Malka N. Halgamuge, “A Comparative Study of Classification Algorithms using Data Mining: Crime and Accidents in Denver City the USA” (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 7, 2016

[5] Dr Hemlata Chahal “COMPREHENSIVE ANALYSIS OF DATA MINING CLASSIFIERS USING WEKA” International Journal of Advanced Research in Computer Science-Volume 9, No. 2, March-April 2018

[6] Dr. Vaishali S. Parsania*1, Dr. N. N. Jani2, Navneet H Bhalodiya3 “Applying Naïve bayes, BayesNet, PART, JRip and OneR Algorithms on Hypothyroid Database for Comparative Analysis**”** INTERNATIONAL JOURNAL OF DARSHAN INSTITUTE ON ENGINEERING RESEARCH & EMERGING TECHNOLOGIES -Vol. 3, No. 1, 2014

[7] D.Sheela Jeyarani, G.Anushya, R.Raja rajeswari, A.Pethalakshmi “A Comparative Study of Decision Tree and Naïve Bayesian Classifiers on Medical Datasets” International Conference on Computing and information Technology (IC2IT-2013)

[8] Ahmad Ashari, Iman Paryudi, A Min Tjoa “Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool” (IJACSA) Vol. 4, No. 11, 2013

[9] https://www.coursehero.com/file/43983002

[10] Md. Nurul Amin1, Md. Ahsan Habib2 “Comparison of Different Classification Techniques Using WEKA for Hematological Data” American Journal of Engineering Research (AJER) Volume-4, Issue-3, pp-55-61

[11] https://medium.com/fintechexplained/machine-learning-bagging

[12] https://www.knowledgehut.com/blog/data-science/bagging-and-random-forest-in-machine-learning

[13] Finch-Savage, W. E. & Bassel, G. W. Seed vigor and crop establishment: Extending performance beyond adaptation.

[14] S. Asha Kiranmail 1 * and A. Jaya Laxmi2 “Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy” Original Research, Open Access, 2018.

Introduction to Popular Hadoop Analytics Tools for Big Data

Aye Myat Nyo[1], Day Si Win[2], Khin Nyein Myint[3]
[1]*FCS, UCS (Monywa),* [2]*FCS, UCS (Banmaw),* [3]*FCS, UCS (Monywa)*
[1]*ayemyat.n@gmail.com,* [2]*dayday.day4@gmail.com,* [3]*khinnyeinmyint@ucsmonywa.edu.mm*

ABSTRACT: *Big data is so huge that traditional tools for data processing do not allow the data to be stored, handled or processed. In Web data, e-commerce and point of sale at stores, bank transactions, social networks, real-time plant data, etc., and a large number of data are collected and processed. In addition, big data processing is now the gateway to academic field, research and IT industry. We can solve the analytical problem with massive datasets using Hadoop analytics tools. The aim of this paper is to offer a comprehensive review and tool comparison. We provide the detailed information on various tools and valves in order to decide the most suitable and successful tools. We can choose which tools are suitable for our needs to enhance the output of big data.*

Keywords: Big Data, Hadoop, Map Reduce, Apache Spark, Apache Storm

1. INTRODUCTION

Nowadays, the increasing digital information now moves with quick lightning speed through the internet and billions of people capture, upload and exchange information through smart phones, laptops, tablets, etc. on social media and other platforms. Its information comes from structure, semi structure or unstructured forms. This huge amount of data are complicated and cannot handle using conventional database management techniques [9]. During these years, e-commerce and digital marketing have become so popular that online transactions processing and services are widely used. Big data analysis is evidence of an advantage for such a business. Big data Analytics concern with storage and process of various kind of datasets. Big data can be characterized by ten Vs [8].

Volume: Volume, in the term of "big data" which extremely relates to large data that decides whether a set of data is big data or not. So, among various parameter, "volume" is the important one that can be assumed as dealing with 'big data'. Volume of the Hadoop tool is about 250Pb [2]. By 2020, it is predicated that 44ZB (Zeta bytes) of data will be generated [9].

Variety: The variety is the sort of data in which the data is being collected and progressed inside the big data. Structured, semi and unstructured data are kinds of data that are applied in Hadoop tool [2]. In the past, primary sources of data collection were from databases and spread sheet like excel, now the data becomes as the form like images, audios, videos, sensors etc.

Velocity: It is the processing speed needed for accessing the data from the database and sending to the designated goal [2].

Veracity: The Veracity of the data is uncertainty in data due to inherent inconsistencies and usually result from huge volume. It is impossible to develop a central veracity-checking mechanism.

Variability: In big data's context, variability concerns with a few different things. In big data, there is only one that is inconsistencies. It can be seen in anomaly and outlier detection methods.

Validity: Validity means how the data is accurate and correct for its intended use.

Value: One of the foremost property of big data is value. It is important for infrastructure system of IT, to store huge amount of values in their databases.

Volatility: A parameter of big data is volatility that relates to the measurement of statistics of separation for a given set of returns.

Visualization: A recent feature of big data is visualization which concerns with illustration of data.

Vulnerability: Vulnerability deals with the safety characteristics of data. And then, a data breaks into big data that violates.

There are many tools and techniques to analyzed big data. In order to study these analytics tools, the remainder sections are given below. The related words is described in section 2 and section 3 discusses Apache Hadoop. In section 4 describes tools for big data processing and study of Map Reduce, Apache Spark and Apache Storm. In section 5 presents comparison of big data analytic tools with basic parameters and section 6 discuss conclusion and of the system.

2. RELATED WORKS

There are many papers discussing of big data, analytics tools. In the first paper, the authors presented a study of big data analytics tools. This paper review four different tools and the parameter of each tool is being described with table. It can be used to determine the greatest and efficient tools in big data [2]. In the second paper, the authors discussed seven Vs of big data characteristics. Besides, the authors described big data analytics prospects, challenges and barriers. There is strong evidence that business performance can be improved via data-driven decision making, big data technologies analytical tools and techniques on big data [3]. In the third paper, the authors presented big data visualization: tools and challenges. They identified why big data visualization is important for and what are the

challenges and issues related to this. And then they reviewed some of the popular visualization tools and observed their merits and demerits. This paper showed that businesses want to review what all are their requirement and which tool(s) suite the best for them [7].

3. APACHE HADOOP

Apache Hadoop is an open source framework for writing and distributing applications to process large amount of diverse data. It can compute where data is located, typically data is moved to compute servers. It can replicate data for redundancy and fault tolerant architecture such as restart failed compute jobs and no shared resources [10]. The aim of Apache Hadoop is to commoditize data processing. It is typically used a cluster to process data and used by significant number of large enterprises like Yahoo, Amazon, Twitter etc. [10].

3.1. Working of Hadoop

Hadoop works a master slave design for data storage and distributed data processing using HDFS [4] and Map Reduce respectively as shown in Figure 1. In Hadoop HDFS, the master node for data storage is the name node and the master node for parallel processing of data is the Job Tracker using Hadoop Map Reduce. In the Hadoop architecture, the slave node is the other machine in the Hadoop cluster which accumulates data and performs complex computations. Job Tracker breaks down computation logic into tasks and distributes tasks across slaves. Name node manages files and relevant metadata across slave.

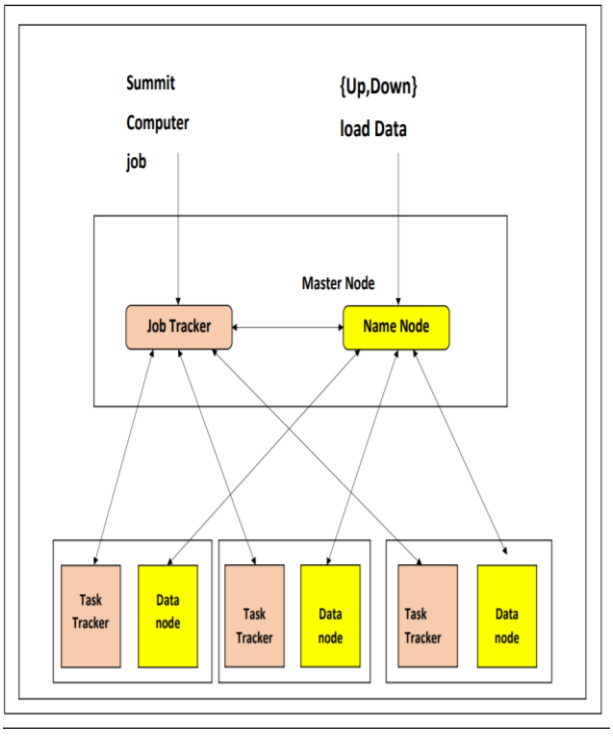

Figure 1. Working of Hadoop

4. TOOLS FOR BIG DATA PROCESSING

Large numbers of tools are available to process big data. Some present ways are discussed for analyzing big data with the focus on three crucial tools: Map Reduce, Apache Spark and Apache Storm.

4.1. Map Reduce

Map Reduce is a programming paradigm, using in division and conquer approach to perform distributed data processing. It partition tasks into smaller sub tasks and execute multiple sub tasks in parallel. Map Reduce for Big data is use to solve analytical problems on large data sets efficiently. The key idea is paralyzing tasks on smaller datasets. Map Reduce uses the Hadoop framework for distributed data processing. Map tasks consist of splitting and mapping. Reduce tasks consists of consolidation and reducing. Map Reduce application needs to "express" distributed computations over data. Furthermore, Hadoop and Map Reduce plays as a dominant software framework for solving big data problems.

4.1.1. Map Reduce Applications

Map Reduce Applications consists of mapper and reducer functions. In Figure 2, it specified multiple instances of mappers. Function output depends only on input and does not store state in functions.

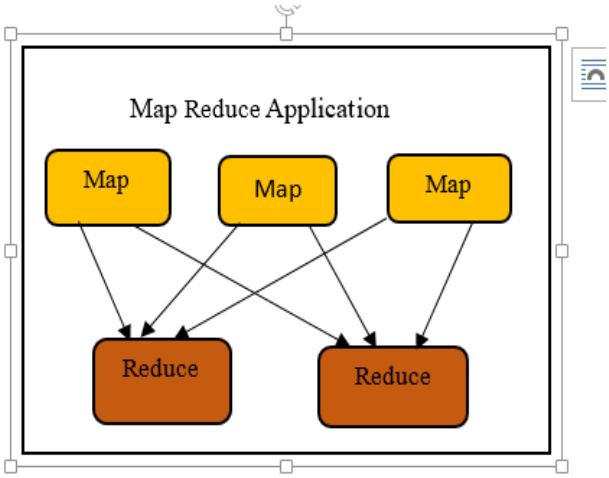

Figure 2. Map reduce applications

4.1.2. Hadoop and Map Reduce

Apache Hadoop and Map Reduce is the most set up software stage for big data analysis. It contains Hadoop kernel, Map Reduce, and Hadoop distributed file system (HDFS). Map Reduce program puts forward to job tracker [8]and configuration parameter is a number mappers. Input data is a HDFS files. Mapper reads input data from HDFS and generate intermediate results as key-value pair. Hadoop Map Reduce framework groups all values with same key. Reducer combine all values of the same key into one key value pair and write data to HDFS [7]. Hadoop Map Reduce is not suitable for inter active processing and random data access. It is designed for batch processing and optimize for sequential access.

4.1.3. Traditional Apache Hadoop

Apache Hadoop consists of two parts.

1. Hadoop Map Reduce: Map Reduce is a computational model and software framework for writing applications running on Hadoop. These Map Reduce programs are capable of processing vast data in parallel on large clusters of computing nodes.
2. HDFS (Hadoop Distributed File System): The storage aspect of Hadoop applications is taken care of by HDFS. Map Reduce applications consume data from HDFS.It generates and distributes several replica of data blocks on clustered computing nodes. This distribution enables reliable and extremely rapid computations.

Hadoop is best known for Map Reduce and it is distributed file system. The components of HDFS are Map Reduce, Hive, Apache pig, and Apache HBase as shown in Figure3.

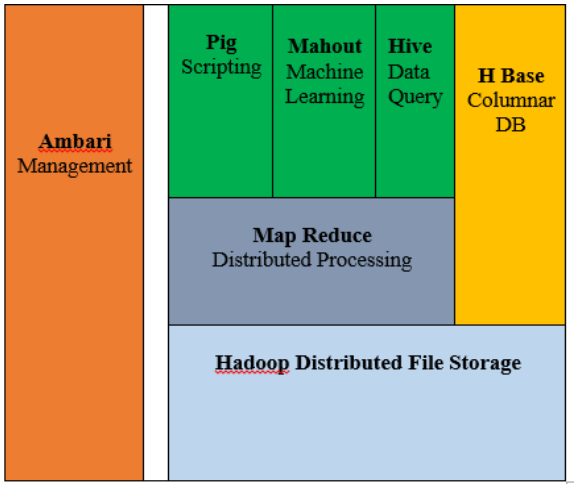

Figure 3. Traditional apache Hadoop

4.2. Apache Spark

Apache Spark is the next generation for big data. It not only offers batch processing capabilities but also streaming capability. It is a general function for the open source that offers a great deal of data for the computing engine used for processing and analysis. Apache Spark is used in memory in order to store all the data. Just like Hadoop Map Reduce, it also works with the system to distributed data across the cluster and process the data in parallel. The master /slave architecture is applied for Spark. It supports an interface for programming language entire clusters with implicit data parallelism and fault-tolerance.

4.2.1. Important features of Apache Spark

Some of the important characteristics of Apache Spark are as follows and shown in Figure 4.

Spark Streaming is a trivial API that allows developers to perform batch processing and real- time streaming of data with simplicity.

Spark MLib is a low-level machine learning library that is simple to use, is scalable, and compatible with various programming languages.

Spark SQL framework component is used for structured and semi-structure data processing.

Graph X is Spark's own Graph Computation Engine and data storage.

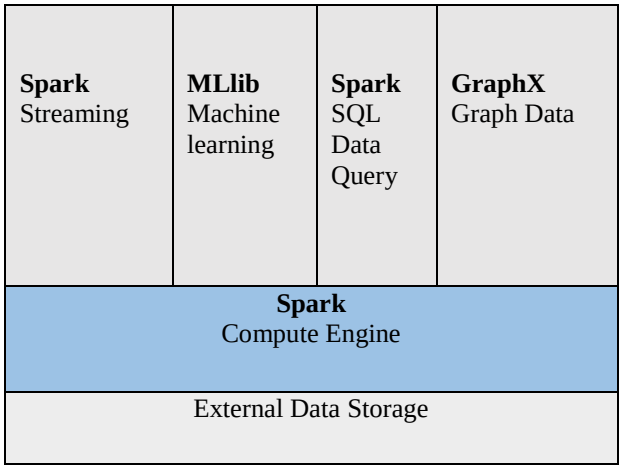

Figure 4. Traditional apache spark

4.2.2. Spark Applications

Driver process runs the main () function It maintains housekeeping information about application and responds to user code or inputs from interactive shell. It analyzes, distributes and schedules work across executors. Cluster Manager manages set of machines which run Spark tasks. Multiple executer processes executing code assigned by driver and reporting status of code computation back to driver. Driver "driven" by applications written Spark supported languages. Executors run only Spark code shown in Figure 5.

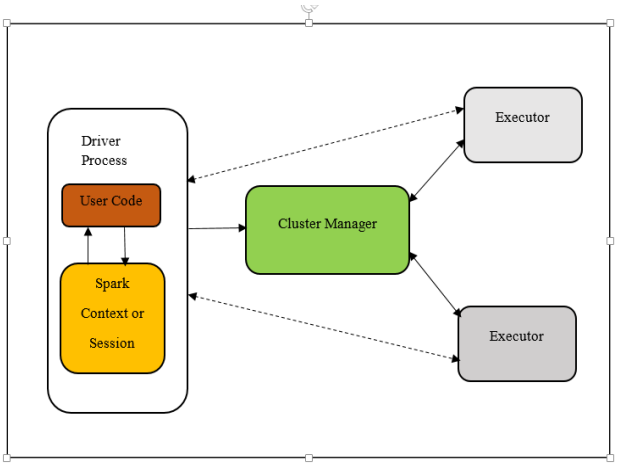

Figure 5. Spark applications

4.3. Apache Storm

Apache Storm is a stream processing framework that is open-source, defect-tolerant and scalable. It provides the basis for the real-time processing of data. It concentrates on event processing or stream processing. Storm updates the fault-tolerant mechanism for computing

or scheduling multiple event computations. The Apache storm is based on streams and tuples. It continues to be principal in real-time data analysis. Storm is east to set up, operate, and ensures that every message is processed through the topology at least once.

In Figure 6, the storm cluster consists of two kinds of nodes, such as the master node and the slave node. They perform two kinds of roles, such as nimbus and supervisor. The two roles have similar functions in accordance with the framework of the work tracker and the map task tracker. Nimbus is responsible for distributing code across the storm cluster, scheduling and assigning tasks to the slave nodes, and monitoring the entire structure. The supervisor shall perform the tasks assigned to them by the nimbus [2] .In addition, it will start and stop the process, as necessary, on the basis of the demand for nimbus. All computational technology is partitioned and distributed to a number of workers' processes.

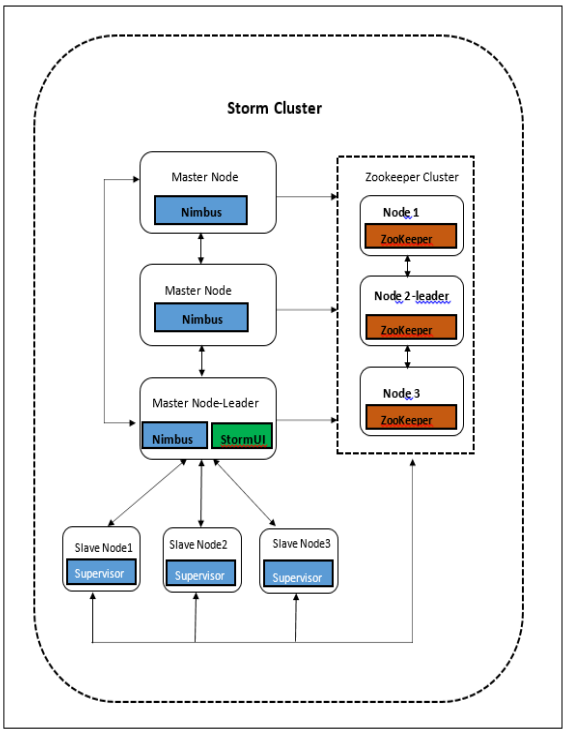

Figure 6. Working of apache storm

4.3.1. Apache Storm Applications

The Storm big data environment consists of streams, spout, and bolts shown in Figure 7. In real-time computing, Storm creates topologies. Topology is a graph of computation, and the Spout is the entry point in the Topology of Storm. It is the data source in Storm. In general, the Spout will read tuples from an external source and emit them to the topology. Each note in topology bolt processing and link between nodes indicate how data should be passed around between nodes (streams).

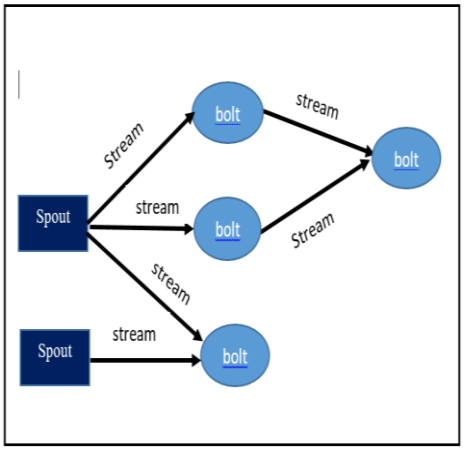

Figure 7. Apache storm application

5. Comparison Of Big Data Analytic Tools With Basic Parameters

The comparative study on big data analytics tools will be discussed with basic features parameters [2] of the big data given in the table 1.

Processing model: Big data analysis includes the valuable useful knowledge in decision-making and following a range of techniques or programming models for accessing large.

Latency: latency, showing how fast it is, is one data of the main factors for system and network success.

Cost: To specify the amount of money needed to develop the project with the tools. In the above comparison, Spark is more expensive than Map Reduce and Apache Storm due to a large amount of memory.

Scalability: Data won't be the same all the time, but it will grow as your organization grows. With big data tools, this is always easy to scale as soon as new data is collected for the company and can be analyzed as expected.

Security: In any case, it is essential to save their data. Big data analysis should provide the data with security and security. In addition, data encryption is also a key feature that Big Data Analytics tools should provide. Data analytics tools are also offered for encryption.

Performance: This is used to measure the amount of result obtained byte the use of the tool by applying it in different fields.

Table 1. Comparison of big data analytics tools with basic parameter

Tools	Characteristics of tools						
	Processing model	**Language support**	**latency**	**Performance**	**Costs**	**Scalability**	**Security**
Apache Spark	Mini/ micro batches, streaming	Scalar, Java, Python, R	seconds	100xfaster than Map Reduce in-memory processing	Expensive due to large amount of memory	Pretty scalable 8000 node production cluster	Only shared secret authentication Relies on data storage for enhanced security
Map Reduce	Parallel Processing	Java, Ruby, Python, C++	More (seconds)	Hard-disk based processing	Cheaper	Highly scalable 42000 node production cluster	Kerberos authentication supported
Apache Storm	Micro-batch processing	Java,Clojure ,Scalar(Multiple Language Support)	milli-second	stream processing via core storm layer	less expensive	scalable million tuples processed per second per node	Kerberos authentication supported

The information used in Table 1 are also described in [9] [10].

Among these approaches, Apache Spark is a fast cluster computer computing technology developed for fast computing and often commonly used by industries. It is very popular with data scientists because of its speed. Spark is 100 times faster in memory than Hadoop Map Reduce for large scale data processing [9]. In addition, Spark can also satisfy all kinds of requirements. Spark is also a platform for data mining and one of the favourite solutions of the data scientist.

6. CONCLUSIONS

This paper introduces some of the most common Hadoop analytics tools for big data and then compares these tools. Based on a comparison, users or data analysts may select any of the tools they need to enhance, or they can create new functionality to the invented tools that can be used for potential empowerment. When choosing big data analytics tool, people who want to learn what all their needs are and which technology suite is ideally suited for them. This paper will direct the selection of Apache Spark as the best tools for data science and large data processing.

REFERENCES

[1] D.P. Acharjya, Kauser Ahmed P, "A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools", IJACSA,Vol 7, No.2, 2016, pp-511–518.

[2] J. Vijayarij, R.Saravanan, P.Victer Paul, R. Raju "A Comprehensive Survey on Big Data Analytics Tools", Conference Paper, November 2016.

[3] Konstantin Vassakis, Emmanuel Petrakis and Ioannis Kopanakis,"Big Data Analytics:Applications, Prospects and Challenges", Springer International Publishing AG 2018 ,January 2018,pp-3-20.

[4] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system", in Proceedings of the 2010 IEEE

[5] Ms. Komasl, "A Review Paper on Big Data Analytics Tools", International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES) Volume 4, Issue 5, May-2018,pp-1012-1017.

[6] Tulasi. B,"Learning Analytics and Big Data in Higher Education",International Journal of Engineering Research & Technology (IJERT) Vol. 3 Issue 1, January - 2014,pp-3377-3385.

[7] Syed Mohd Ali, Noopur Gupta , Gopal Krishna Nayak, Rakesh Kumar Lenka , "Big Data Visualization : Tools and Challenges", 2016 2nd International Conference on Contemporary Computing and Informatics (ic3i) 2016,IIIE,pp.709-713.

[8] https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx.

[9] https:// swayam.gov.in.

[10] https://www.itecogoi.in

Documents Retrieval using Solr

Ei Marlar Win[1], Shwe Thinzar Aung[2]
[1,2]*Faculty of Computer Science (University of Computer Studies, Taungoo)*
[1]*eimarlarwin@ucstaungoo.edu.mm*, [2]*shwethinzaraung@ucstaungoo.edu.mm*

ABSTRACT: *Nowadays, people kept the data digitally. The more the data, the harder it is to search. Searching contents of document in many folders and extraction for the needed information spent much time. The system can search the needed documents by using Solr to reduce the difficulties concerned with time quickly and speedily. Solr (Searching On Lucene Replication) provides the indexing and searching of the data such as structured data, semi-structured data and unstructured data. Solr can index multiple documents as doc, pdf, xml, json, csv, xlsx and so on. Furthermore, the system can search the documents that used in Myanmar language. The aim of the system is to search a document that contains the words the user wants to desire from the indexed documents in Solr core. The system can search the required documents by using search criteria (the content and metadata: title and date). The system is implemented by using Java and Apache Solr.*

Keywords: solr; indexing; searching; metadata; java;

1. INTRODUCTION

This is the age of Information Technology. Information Technology is applied in data collection, storage, search, retrieval, and availability. The main technology of search engines is Information Retrieval (IR), a field about obtaining information from a collection of files. The main focus of IR is to retrieve information from text and text documents. The following steps are applied as general information retrieval functions.

- Crawling: The document collection is browsing and fetching documents.
- Indexing: An index of the documents is built.
- Searching: The documents that are relevant to the query are retrieved from the index.

Indexing data is one of the most important things. The searched results will be poor when data is not indexed properly. It's almost sure that users will not be satisfied with the application when the searched results are poor. Therefore, the application that uses Solr is applied. Solr is that uses data to be prepared and indexed as timely and correctly as much as possible. Multiple formats of data from multiple sources are indexed.

The age of development technology, Solr is one of the most popular IRSs which are an open source. Solr is a search server built on top of Apache Lucene, an open source, Java-based and information retrieval library. It is designed to drive powerful document retrieval applications that need to serve data to users based on their queries.

There are many related works concerned with Solr. Dong Yi and Wu Youyu, stated that the efficient search system is built by using the open source search engine Solr. This paper proves the system that the Solr search system is better than the query function of database in terms of search time and accuracy [8].

To achieve Chinese word segmentation, employed Vector Space Model (VSM) based on keywords to implement topic relevance, extended the user search module and the tourism domain word library to collect information, filter information retrieval, and related word by using various stages. Experiments were also conducted in order to evaluate the algorithm and the result show that the vertical search engine based on Nutch and Solr which is used for tourism information retrieval. It can improve the user retrieval precision and meet the professional demand of user retrieval [3].

Searching documents in many folders are very complex and time consuming. This system can provide to solve this problem by using Solr. This system searches the document quickly that the user desired from the indexed documents.

This paper is organized as follows: Section 2 expresses search engine. Section 3 discusses the architecture of Solr. Section 4 explains the proposed system and the implementation of this system. Section 5 gives the evaluation of this system that used Solr. Section 6 gives the conclusions of the paper.

2. SEARCH ENGINE

The main technology of search engines is Information Retrieval (IR), a field about obtaining information from a collection of files. The main focus of IR is to retrieve information from text and text documents [5].

2.1. Elastic Search

Elastic search is a search engine and data analysis tool that has been developed from Apache Lucene substructure and that is light, easily installed, open source coded and scalable. This search engine that offers service over Restful API is very fast and practical tools and safety options. The elastic search engine has multi-tenancy as the key feature. Elastic search has major features: distributed search, multi-tenancy, an analyzer chain, analytical search, grouping and aggregation. Elastic search is less complex and less detailed contrast to Solr search engine developed from Apache Lucene. It is strong and flexible. One of the most important advantages is that it is real-time and

distributed. Elastic search has fast index building, it only supports JSON format.

Today Elastic search is used in content searching, data analysis and queries in the projects such like Mozilla, Foursquare, GitHub Elastic search has many full text search capacities such like multi-language option, a strong query language and autocomplete. Elastic search is full-text search which is generally used in single page application. Data is explored at a speed and scale. It is used for full-text search, structured search, analytics, and all in combination. Stack Overflow combines full-text search with uses more-like-this to find related questions and answers and geolocation queries. Wikipedia uses Elastic search to provide full-text search with highlighted search snippets [1].

2.2. Apache Solr

Solr is scalable, ready to deploy, search/storage engine optimized to search large volumes of text-centric data. SOLR technology mainly follows indexing methodology i.e. it maintains indexes for each and every document which can be used for fast search and retrieval [4]. Solr is enterprise-ready, fast and highly scalable, built on a Java library called Lucene [7]. Solr is a standalone/cloud enterprise search server with a REST-like API. Documents are applied in it (called "indexing") via XML, JSON, CSV or binary over HTTP. The users query it via HTTP GET and receive XML, JSON, CSV or binary results. It's an open-source search engine. Anyone can contribute and thus there is availability of more features: Highlighting, Full-text search, real-time indexing, faceted search, dynamic clustering, database integration, NoSQL features and handles documents in a better way [6]. Solr integrates several open source tools:

- Jetty: Solr runs on a Jetty server and supports REST-like HTTP request and JSON API.
- Apache Lucene: Apache Lucene is a full-text indexing and search development kit. Solr uses Lucene in its core.
- Apache Tika: Solr accepts various file formats as input, including plain text, HTML file, PDF, Microsoft Word and so on. These files will be handled by Tika.

3. THE ARCHITECTURE OF SOLR

The block diagram of the architecture of Apache Solr is shown in "Figure 1".

Major blocks of Apache Solr are

- Request Handler – Solr processed the request by using request handlers. The request is either query request or index update request. For passing a request to Solr, the handler to a certain URI end-point and the specified request will be mapped.
- Search Component – Search component feature is used in spell checking, query, faceting, hit highlighting etc.
- Query Parser – The queries are parsed by the Apache Solr query parser and are verified for syntactical errors. After parsing the queries, queries are translated to a format that Lucene understands.
- Response Writer – A response writer generates the formatted output for the user queries. Respnse formats such as XML, JSON, CSV, etc are supported.
- Analyzer/tokenizer – Apache Solr analyzes the content, divides it into tokens, and passes these tokens to Lucene and examines the text of fields and generates a token stream. A tokenizer breaks the token stream prepared by the analyzer into tokens.
- Update Request Processor - The update request run through a set of plugins and collectively. This processor is applied for modifications such as deleting a field, adding a field, etc.

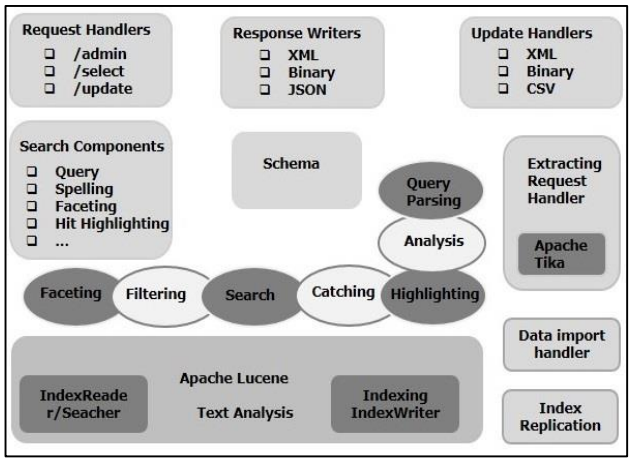

Figure 1. Architecture of Solr

3.1. Core in Solr

Core in the Solr is a separate index and configuration. Cores are used for data partitioning and supporting multiple applications. A single server can support multiple cores. A core is created in Solr by using command prompt.

3.2. Indexing Documents

A Solr index can accept data from many different sources, including CSV files, XML files, common file formats such as PDF or Microsoft Word and data extracted from a database. The following "Figure 2" shows the Solr architecture indexing. Indexing is an arrangement of documents or (other entities) systematically and to locate information in a document. Indexing collects, parses, and stores documents. A document is composed various information such as fields and records. For the type of information, the analyzer analyzed the document fields and records as numerical info, binary info, alphanumeric text and other such information.

Loading data in a document into a Solr index has many ways: Solr Cell framework integrated Apache Tika can be used for ingesting structured files or binary files such as Office, Word, PDF, and other proprietary formats.

Data can be ingested by using a custom Java application that uses Solr's Java Client API.

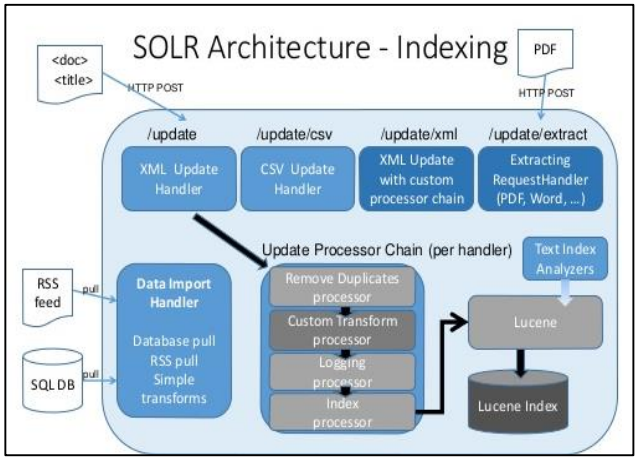

Figure 2. Solr indexing architecture

There is a common basic data structure to ingest data into a Solr index: a document containing multiple fields that has a name and content which may be empty. As a unique ID field, one of the fields is identified. Although the use of a unique ID field, it is not strictly required by Solr. In Apache Solr, Various document formats (xml, csv, pdf, etc) can be indexed [8].

3.3. Searching

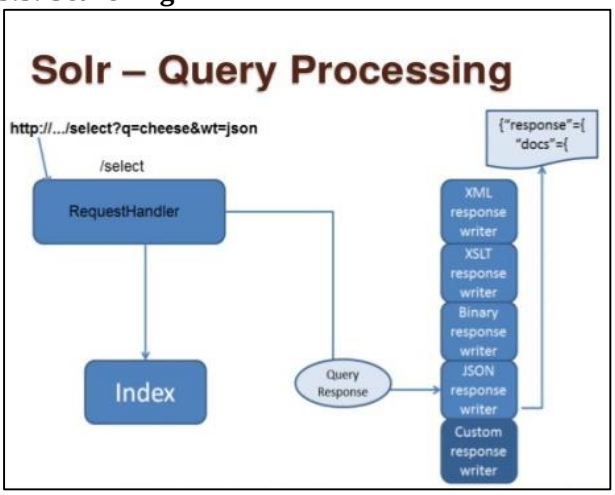

Figure 3. Solr search

The processing of Solr query is shown in "Figure 3". Request handlers are handling incoming requests and performing particular processing of requests. A RequestHandler will parse the search parameter, tokenize it and perform searches.

For fast search, inverted index are built for the whole document set. An inverted index (also referred to as a postings file or inverted file) is an index data structure storing a mapping from content, such as numbers or words, to its location in a database file, or in a document or set of documents (named in contrast to a Forward Index, which maps from documents to content). With inverted index, the documents that a word occurs can be directly retrieved [2].

Search engines search documents by checking the inverted index. Documents contain all keywords are returned. There is a search box or a query box where the user can type any expression or string or text string to be searched. There are different options next to the search box also allows the user to filter the search criteria. These features are the search components of the search application.

4. Implementation Of The System

This system can search the related documents by using keywords. There are two types of process for implementing this system. They are the indexing process for saving document on Solr core and the searching process for retrieving related documents that the user wants.

In the indexing process of this system, the document is uploaded for indexing. The documents can be parsed by using Apache Tika. Indexed document can be stored in Solr core. This system is not used database for the storage.

In the searching process, this system can search the desired documents even the user knows only a keyword in the documents. User can enter the keyword by using metadata (content, title and date). According to the entered keyword, the related documents can be parsed by using XMLResponseParser.

4.1. Indexing Document of the System

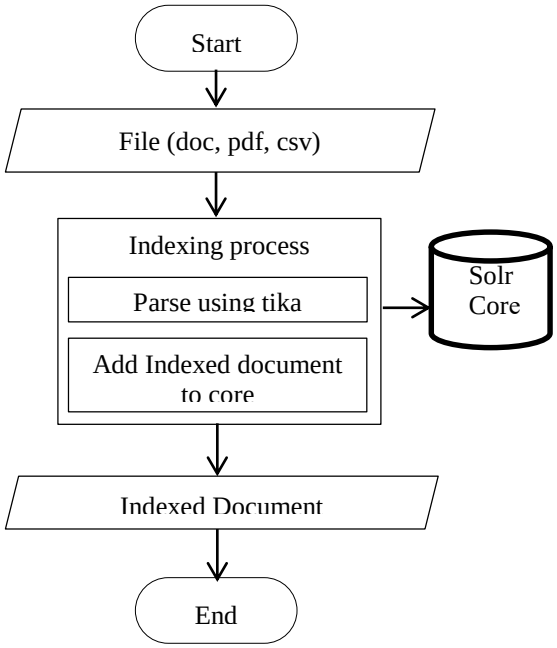

Figure 4. Indexing process of the system

For this system, the input documents are Microsoft Word and PDF. The user can use any docx file and pdf file. The indexing process is shown in "Figure 4".

To be indexed documents, it can be uploaded to Solr by HTTP request and handled by UpdateHandler. Solr uses different UpdateHandlers to handle different document formats, including JSON, XML and complex

proprietary formats. Apache Tika handles these formats, e.g. PDF and Microsoft Word [2].

Solr is integrated with java and apache tika is applied for extracting content and metadata of the documents.

4.1.1. Indexing Document Implementation

Solr uses an inverted index that indexes words. By indexing a word, it maps this word with a similar content in the document.

For indexing, the following scripts are updated in the solrconfig.xml to implement the system.

```
<lib dir="${solr.install.dir:../../..}/contrib/extraction/lib"
regex=".*\.jar"/>
<lib dir="${solr.install.dir:../../..}/dist/" regex="solr-cell-\d.*\.jar"/>
<requestHandler name="/update/extract"
class="solr.extraction.ExtractingRequestHandler" >
<lst name="defaults">
<str name="fmap.content">text</str>
<str name="lowernames">true</str>
 <str name="uprefix">attr_</str>
<str name="captureAttr">true</str>
</lst>
```

4.2. Searching Document of the System

To search document, search query is used by three criteria (content, date, title). In querying process, when the user entered search key, related documents are searched in Solr Core. Then, the resulted documents are parsed by XMLResponseParser and response as document list. The system displays the resulted related document list. The searching process of the system is shown in "Figure 5".

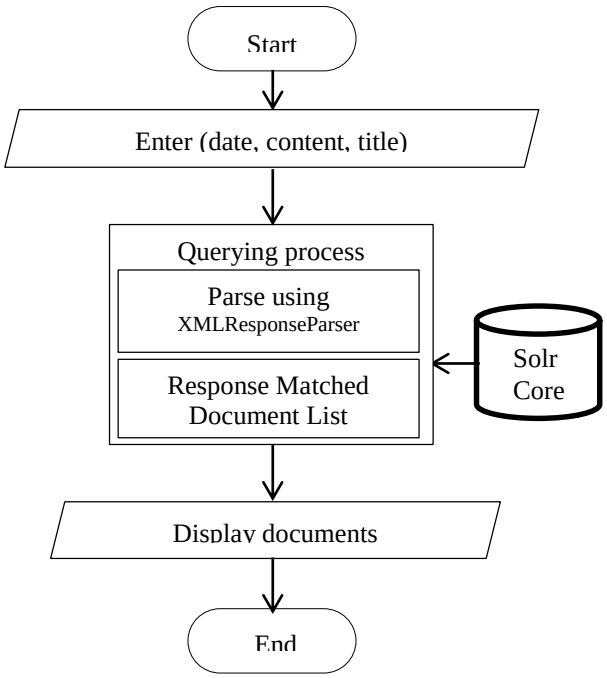

Figure 5. Searching process of the system

5. EVALUATION

The evaluation processes of indexing and searching are described by using Solr. The system is evaluated with the document that has characters less than 100000. Word documents and pdf documents for evaluation are applied.

Four different numbers of characters in the documents are applied for indexing and searching evaluation process.

Table 1. Document size for indexing and searching

Document No:	Indexing	Searching
1	300kb (characters 10,000)	300kb (characters 10,000)
2	250kb (characters 20,000)	250kb (characters 20,000)
3	180kb (characters 30,000)	180kb (characters 30,000)
4	150kb (characters 40,000)	150kb (characters 40,000)

Indexing process in Solr has more time than searching process. Searching process in Solr can save time. Although the document has different numbers of character, Solr searching process is slightly more time gap.

The following figure "Figure 6" shows the indexing and searching time (second) of documents. Evaluation result is the performance time of documents indexing and documents searching.

The processing time (second) of each document was recorded. This process was done 10 times. Then, performance time was measured by average.

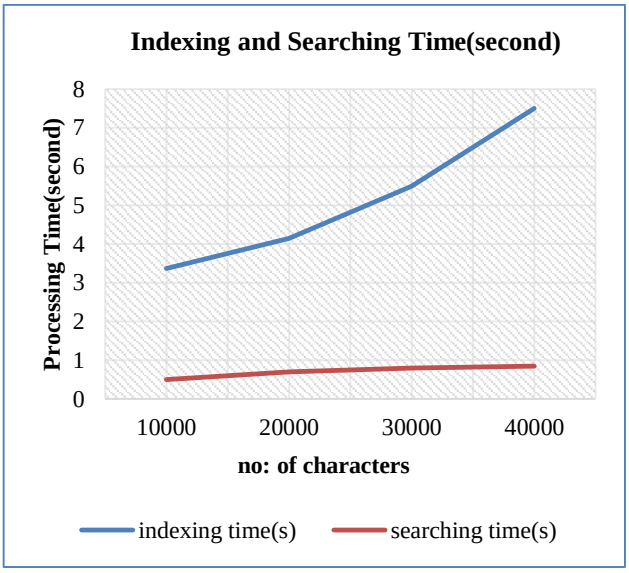

Figure 6. Indexing and searching evaluation

5.1. Comparison of Searching Performance on MySQL and Solr

When searching the related documents, MySQL searches the related documents by basing on each record in database. In contrast, Solr uses the inverted index words. In this situation, Solr can search the documents faster than MySQL. Although the documents become more, Solr searching time is slightly increased. A number of applied documents are 40 for this comparison. The result is evaluated by using MySQL and Solr integrated with java.

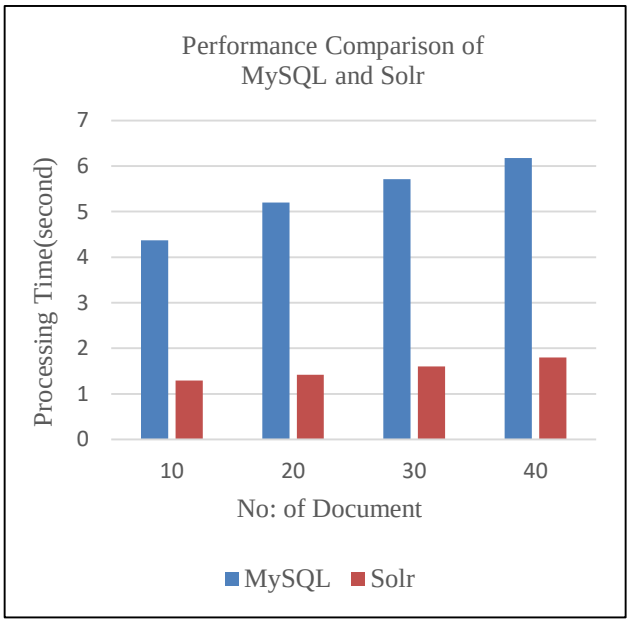

Figure 7. Comparison of performance on MySQL and Solr

6. CONCLUSIONS

This system is a process of searching and retrieving the needed documents from collection of documents. This paper expresses the searching time less even large documents by evaluating. This system can search quickly the documents that contain many characters (large document size). This system can also search the documents that use the Myanmar language. Although Myanmar content in pdf documents can be indexed by using Solr, it cannot be searched correctly according to the limitation of Myanmar encoding in pdf. The accuracy for MySQL and Solr will be compared more detail as further extension.

REFERENCES

[1] K. U˘gur and K. I¸sıl, "Comparison of Solr and Elasticsearch Among Popular Full Text Search Engines and Their Security Analysis", Oct 2016

[2] L. Zhongmiao, "A Domain Specific Search Engine with Explicit Document Relations"

[3] M. Huawei and D. Wencai, "Searching Tourism Information by Using Vertical Search Engine Based on Nutch and Solr"

[4] P. Sanket and R. Rajeshkannan, “Search Engine Optimization of E-Commerce Website using Apache SOLR”, International Journal of Advanced Research in Computer Science, Volume 4, No. 8, May-June 2013

[5] R. Akram, “Review: Information Retrieval Techniques and Applications”, International Journal of Computer Networks and Communications Security, VOL. 3, NO. 9, SEPTEMBER 2015, pp-373–377

[6] R. Padmavathy, "Enterprise Search Technology Using Solr and Cloud", Spring 2015

[7] Y. Divakar, S. Sonia, “An approach for spatial search using SOLR”, January 2013

[8] Y. Dong and Y. Wu, "Shopping Website based on solr", 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), 2019

Certificate Index	***Author***
VOL02IS01001	*Aung Kyi, Zaw Moe Aung* "A Study of Reading Comprehension Questions From Matriculation English Question Papers By Using Bloom's Taxonomy"
VOL02IS01002	*Aye Aye Lwin, Win Kyaw, Thaik Thaik, San San Wai* "Characterization of Advanced Superionic Conducting Materials of Lithium Cobalt-Nickel Oxides for Solid Oxide Fuel Cell (SOFC) Application"
VOL02IS01003	*Aye Lae Maw* "The Concept of Speech Function in the Articles in the Reader`s Digest"
VOL02IS01004	*Aye Mon Win, Yu Yu Khaing, Lei Yi Htwe* "Performance Analysis of Classification Algorithms in Data Mining Technique"
VOL02IS01005	*Aye Myat Nyo, Day Si Win, Khin Nyein Myint* "Introduction to Popular Hadoop Analytics Tools for Big Data"
VOL02IS01006	*Ei Marlar Win, Shwe Thinzar Aung* "Documents Retrieval using Solr"
VOL02IS01007	*Hla Thein Maung, Aung San Min, Aye Aye Khine* "Construction of Motor Control System Using Fingerprint Sensor"
VOL02IS01008	*Khaing Khin Aye, Thi Thi Swe, Nyein Ei Lwin* မိုးမိုး(အင်းလျား)၏ "မေတ္တာကမ်းနားအချစ်သစ်ပင်"ဝတ္ထုမှ ဇာတ်ဆောင်စရိုက်ဖန်တီးမှုအတတ်ပညာ
VOL02IS01009	*Khaing Su Wai, Hsu Mon Maung* "Analysis of Routing Protocols over TCP in Mobile Ad-hoc Networks using Random Way Point Model"
VOL02IS01010	*Khin Khin Maw* "ဦးပုည၏ တေးထပ်များမှ နှစ်သက်မှုသက်သက်ကိုသာပေးသော ရသမြောက်အဖွဲ့များ လေ့လာချက်"
VOL02IS01011	*Khin Moh Moh Thin, Hla Yin Moe, Lin Lin Aye* "Applying the Queue Theory in Bank Service Centers"
VOL02IS01012	*Khin Pyone, Phyu Phyu Khaing* "ရှင်ဥတ္တမကျော်တောလားလာ ရာသီဘွဲ့၏ သဘောသဘာဝများ လေ့လာချက်"
VOL02IS01013	*Khine Khine* "မြန်မာသဒ္ဒါသမိုင်းကြောင်းလေ့လာချက်"
VOL02IS01014	*May Oo Mon, Su Myat Aye, Naing Naing Maw* "An Investigation into Adolescents' Emotional Creativity and the Influence of Personality Traits on it"
VOL02IS01015	*May Wah Linn, Naing Naing Maw* "An Investigation into Personality and Career Interest of High School Students"
VOL02IS01016	*Moe Ei Swe, Mar Mar Lwin, Than Than Sint*

	Illocutionary Acts Of The Main Character's Utterances In The Movie "Beyond Rangoon"
VOL02IS01017	*Mya Mya Win* "ကယန်း (ပဒေါင်) တိုင်းရင်းသားတို့၏ ကြေးပတ်ခြင်းဓလေ့နှင့် ဘာသာစကား သဘောလက္ခဏာ"
VOL02IS01018	*Myat Theingi Kyaw, Wai Wai Phyo* "Grammatical Parallel Structure In Inaugural Addresses By George W. Bush"
VOL02IS01019	*Myat Yu Yu Mon , Su Myat Aye* "Development of Numeracy Test For Adolescents"
VOL02IS01020	*Naing Naing Maw* "Reducing Students' Anxiety in Learning Reading Passages through Reciprocal Teaching"
VOL02IS01021	*Nang Khin Pyone Myint, Thin Nu Nu Win, Nang Sabae Phyu* "Multilevel Association Rules Mining using Apriori Algorithm"
VOL02IS01022	*Nilar Htun, Nang Seint Seint Soe* "Sentiment Analysis of Students' feedback from Coursera Online Learning Using Bernoulli Naïve Bayes Classifier"
VOL02IS01023	*Nyein Ei Lwin, Thi Thi Swe, Khaing Khin Aye* "မောင်ချောနွယ်၏ရထားကဗျာပေါင်းချုပ်မှုနိမိတ်ပုံအသုံးများ လေ့လာချက်"
VOL02IS01024	*Phue Wai Ko Ko, Su Myat Aye, Naing Naing Maw* "Analysis of Gender and Grade Differences on Abstract Reasoning Test for High School Students"
VOL02IS01025	*Phyo Wai Hlaing* "ပန်းမွေ့ရာ ရွှေကော်ဇော၀တ္ထုတိုကန့်သတ်သိ ရှုထောင့်များ"
VOL02IS01026	*Phyu Thwe, Cho Cho Lwin, Hnin Pwint Myu Wai* "Detection of Diabetes Using Classification Methods "
VOL02IS01027	*San Myint Yi* "Evaluating the Effect of Environmental Education and Awareness on Solid Waste Management within Elementary Students"
VOL02IS01028	*San San Nwe, Hla Yin Moe, Lin Lin Aye* "Student-Centered Approach is more Effective than Teacher-Centered Approach on Mathematics"
VOL02IS01029	*Sandar Shein* "ဦးကြီး၏လွမ်းချင်းကဗျာများမှ မြန်မာ့ကျေးလက်ဓလေ့များ"
VOL02IS01030	*Shine Maung Maung* "မြိတ်ဒေသိယစကားကို အတ္ထဗေဒရှုထောင့်မှလေ့လာချက်"
VOL02IS01031	*Shwe Sin Win, Khin Htwe Myint, Zar Zar Thin* "ကာတွန်းများမှ ထုတ်ဖော်မပြောသောအနက်"
VOL02IS01032	*Shwe Zin Aung* "Scanning System for Light Transmission of Glasses"

VOL02IS01033 *Su Myat Aye, Naing Naing Maw*
"An Investigation into Non-verbal Intelligence of Primary Students"

VOL02IS01034 *Than Than Naing*
"တ-ဝဂ်အတွင်းရှိအသံစွဲစကားလုံးများ၏အနက်ကိုစိစစ်ခြင်း"

VOL02IS01035 *Thandar Kyi, Htun Htun Oo*
"Study on Neutron-proton Scattering using CD-Bonn Potential"

VOL02IS01036 *Thet Lwin Oo , Win San Win, Hnin Ei Maung, Daw Lwin Lwin Soe*
"Design And Construction of Android Phone Controlled Electrical Household Appliances"

VOL02IS01037 *Tin Tin Maw*
"Generation of Orthogonal Polynomials in Least-Squares Approximations"

VOL02IS01038 *Tint Tint Ei, Mar Mar Lwin, Hnin Lae Win*
"A Survey On Preferences Towards Learning Styles of Undergraduate Arts And Science Students"

VOL02IS01039 *Wai Wai Phyo, Myat Theingi Kyaw*
"Grammatical Collocations Found in The Selected Academic Texts"

VOL02IS01040 *Wai Wai Tin*
"တရုတ်သံရောက်မော်ကွန်းမှ နန်းဓလေ့ယဉ်ကျေးမှုများ"

VOL02IS01041 *Ye Ye Cho*
"မြန်မာဝါကျရိုးဖွဲ့စည်းပုံရှိ စကားမြုပ်စနစ်နှင့်ဘာသာဗေဒအမြင်"

VOL02IS01042 *Yi Mon Aung, Nwet Nwet Than, Linn Linn Htun*
"Genetic Algorithm-Based Feature Selection and Classification of Breast Cancer Using Bayesian Network Classifier"

VOL02IS01043 *Yu Yu Tun, Aye Thandar Win, Htar Hlaing Soe*
ငြိမ်းချမ်းချိုးဖြူဆက်၍ကူ"ကဗျာမှဘာသာစကားတာဝန်များစိစစ်ချက် (လူမှုဘာသာဗေဒ)

VOL02IS01044 *Zaw Naing*
"နည်းပညာဆိုင်ရာဝေါဟာရများနှင့် မော်ဒန်ကဗျာ"

VOL02IS01045 *Zin Nwe Khaing, Hla Yin Moe, Aye Mya Mya Moe*
"Optimizing Integer Programming Problem Using Branch and Bound Method"

University of Computer Studies (Taungoo)
Website: https://ucstaungoo.edu.mm
Email:jites.admin@ucstaungoo.edu.mm