

Data Science

Section

PREDICTION ON MARKET PRICE OF ONION IN PAKOKKU DISTRICT USING MARKOV CHAIN MODEL

Myint Myint Toe ⁽¹⁾, Kyi Kyi Myint ⁽²⁾, Khin Phyu Thant ⁽³⁾

⁽¹⁾⁽²⁾⁽³⁾University of Computer Studies (Meiktila), Myanmar

⁽¹⁾*myintmyinttoe2017@gmail.com*, ⁽²⁾ *myint.kyikyi@gmail.com*, ⁽³⁾ *khinphyuthant31@gmail.com*

ABSTRACT

Onion is one of Myanmar agricultural products and it is planted at the central part of Myanmar. This paper represents the studying and prediction of the future market price of onion produced in Pakokku region. The data related to the market prices of onion are collected and categorized the frequency of the days with same intervals of prices to predict the near future price. This paper uses every day prices of onion on the particular day's price intervals for one year. The steady state probability is calculated using **Chapman–Kolmogorov Equations and predict the price of the onion near future** by using Markov Chain Model.

KEYWORDS: *Chapman–Kolmogorov Equations, Markov Chain Model, Prediction, Intervals, Steady State Probability*

1. INTRODUCTION

Pakokku is the largest district situated at the central part of Myanmar. There are many farmers in that district who plant onion. The variation of the price of onion may effect on the farmers' lives and economies. The market price of onion is changing day by day. They are facing with the difficulties to predict the future market price of onion. This paper intends to solve that problem. Firstly, an approximate evaluation of data is created. To find the different data set, Markov Chain is applied. These differences are going to be what Markov Chains was applied too. Using **Chapman–Kolmogorov equations** with the transition probability matrix multiplied by the transition probability matrix, the steady states probability is calculated.

The Markov Chain Model can be used to apply many predictions of market prices of various products

and many scopes of predictions. This paper studies Markov Chain model:

- To predict the coming market price movement of onion in near future
- To develop the transition probability matrix for the price of onion and to find the steady state probability using **Chapman–Kolmogorov Equations**

2. RELATED WORK

Markov Chain is used to model many economic and social problems of brand loyalty, prediction of market prices for stock and vegetable etc. Bairagi and Kakaty (2017) used matrix, determined state transition vector, computed state transition probabilities that are calculated and attempted to find out the long term behavior of the market price of potatoes.

P. Jasinthan, A. Laheetharan and N. Satkunanathan (2016) modeled for Vegetable Price Movement in Jaffna using Markov Chain and calculated the transition probability matrix, mean recurrence times and the stationary probability vectors and expected recurrent times the regional market approach leads to superior results.

M. K. Bhusal (2017) applied a Markov chain model to forecast the behavior of Nepal Stock Exchange (NEPSE) index. It derived transition probability matrix, determined state transition vector, computed state transition probabilities for forecasting the NEPSE index and then makes decisions using long run behavior of the index.

Zhu and Xu (2012) used Markov Chain to analyze and predict on the fluctuation cycle of vegetable

price. First, they construct Markov Chain model and calculate the steady state probabilities and the expected recurrence times. Then predict the price of vegetables.

3. RESEARCH METHODOLOGY

The Markov Chain Model has been applied in this study. The study is based on the collected data from Agricultural trading's, which was launched by the radio broadcasting and TV channels.

3.1 Markov Chain Concepts

Markov Chain is a stochastic process that has Monrovia property to be an indexed collection of random variables $\{X_t\}$ where the index t runs through a given set T , where T is the set of nonnegative integers, and X_t represents a measurable characteristic of interest at time t . Consider a finite discrete time homogeneous stochastic process with index set $Z^+ = \{0, 1, 2, \dots\}$; that is, a sequence $\{X_n: n \in Z^+\}$ of random variables. As usual the subscript n in X_n stands for the time and X denotes the state of the process at time n . If $X_n \in S$, then S is called state space of the stochastic process considered here satisfy the Markov property. Given the conditional probability of any future "event," given any past "event" and the present state, the future of the process is independent of the past. That is,

$$\begin{aligned} \text{For } i, j: x_0, \dots, x_{n-1} \in S, \\ P(X_{n+1} = j | X_0 = x_0, \dots, X_{n-1} = x_{n-1}, X_n = i) \\ = P(X_{n+1} = j | X_n = i) = P_{ij}. \end{aligned}$$

A conventional way to note stationary transition probabilities that will be seen later in this paper is,

$$\begin{aligned} P_{ij} &= P(X_{t+1} = j | X_t = i), \\ P_{ij}^{(n)} &= P(X_{t+1} = j | X_t = i) \end{aligned}$$

i and j are said to communicate.

3.2 Chapman-Kolmogorov equations

Chapman-Kolmogorov equations are the method for computing these n -step transition probabilities:

$$\begin{aligned} P_{ij}^{(n)} &= \sum_{k=0}^M P_{ik}^{(m)} P_{kj}^{(n-m)} \text{ for all } i, j = 0, 1, \dots, M \\ \text{and any } m &= 1, 2, \dots, n-1, n = m+1, \\ m+2, \dots \end{aligned}$$

The n^{th} step transition matrix is

$$P^{(n)} = \begin{bmatrix} P_{11}^{(n)} & P_{12}^{(n)} & P_{13}^{(n)} & \dots & P_{1k}^{(n)} \\ P_{21}^{(n)} & P_{22}^{(n)} & P_{23}^{(n)} & \dots & P_{2k}^{(n)} \\ P_{31}^{(n)} & P_{32}^{(n)} & P_{33}^{(n)} & \dots & P_{3k}^{(n)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ P_{k1}^{(n)} & P_{k2}^{(n)} & P_{k3}^{(n)} & \dots & P_{kk}^{(n)} \end{bmatrix}$$

For any irreducible ergodic Markov Chain, $\lim_{n \rightarrow \infty} P_{ij}^{(n)} = \pi_j > 0$, where the π_j uniquely satisfy the following steady-state equations

$$\begin{aligned} \pi_j &= \sum_{i=0}^M \pi_i p_{ij} \text{ for } j = 0, 1, \dots, M \\ \sum_{j=0}^M \pi_j &= 1 \end{aligned}$$

π_j 's the steady-state probabilities of the Markov Chain. These π_j can estimate the near future market and the other with probabilities.

4. PROCEDURAL APPROACH

The procedural Approach of the research work is shown in figure 1.

First, we collect the data of onion price for one year. Classify the frequency of days with the same interval of the price. Next, we specify the states of Markov Chain by analyzing the days depend on the prices raise or down to create the transition probability matrix and construct state diagram. Finally we calculate the steady state probabilities using **Chapman-Kolmogorov Equations. Based on this result, we can predict the price of onion near future.**

Market Data of 365 days from August, 2018 to July, 2019 are categorized as shown in Table 1.

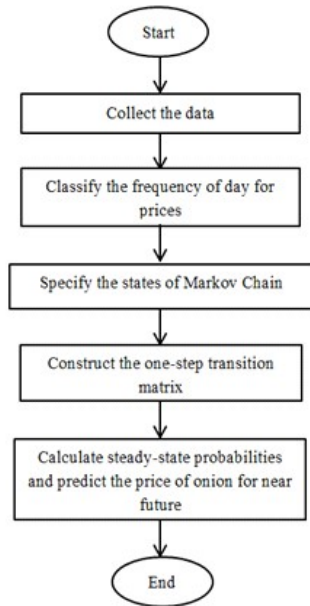


Fig 1. The procedure of research work

Table1. Frequency data

	Price below	Frequency
1	$200 < \text{price} \leq 550$	92
2	$550 < \text{price} \leq 600$	70
3	$600 < \text{price} \leq 650$	65
4	$650 < \text{price} \leq 700$	30
5	$700 < \text{price} \leq 750$	25
6	$750 < \text{price} \leq 800$	9
7	$800 < \text{price} \leq 850$	13
8	$850 < \text{price} \leq 900$	11
9	$900 < \text{price} \leq 950$	10
10	$950 < \text{price} \leq 1000$	16
11	$1000 < \text{price} \leq 1050$	9
12	$1050 < \text{price} \leq 1100$	7
13	$1100 < \text{price} \leq 1150$	8

Identify the above categorized data to four states of Markov Chain as follows:

C_1 = the state that the price of onion per peittha less than or equal to 550.

C_2 = the state that the price of onion per peittha on $550 < \text{price} \leq 750$.

C_3 = the state that the price of onion per peittha on $750 < \text{price} \leq 950$.

C_4 = the state that the price of onion per peittha above 950.

The transition frequencies for the price of onion are counted for each interval shown in Table 2.

Table2. Frequency transition data for each state

States	C_1	C_2	C_3	C_4
C_1	52	24	6	10
C_2	27	139	17	7
C_3	10	15	13	5
C_4	3	12	7	18

The transition probability matrix is obtained by dividing the frequency of each class by total frequency. Using these probabilities, the one-step transition matrix is obtained.

$$P = \begin{pmatrix} 0.5652 & 0.2609 & 0.0652 & 0.1087 \\ 0.1421 & 0.7316 & 0.0895 & 0.0369 \\ 0.2326 & 0.3488 & 0.3023 & 0.1163 \\ 0.075 & 0.3 & 0.175 & 0.45 \end{pmatrix}$$

The state transition diagram is shown in Figure 2.

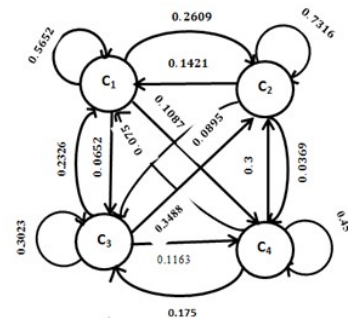


Fig 2. State transition diagram

The two step transition matrix and the next transition matrices can obtain by using Chapman-Kolmogorov equations.

$$P^2 = P \cdot P$$

$$P^2 = \begin{pmatrix} 0.3798 & 0.3937 & 0.0989 & 0.1276 \\ 0.2079 & 0.6146 & 0.1083 & 0.0695 \\ 0.2601 & 0.4562 & 0.1581 & 0.1256 \\ 0.1595 & 0.4351 & 0.1634 & 0.2421 \end{pmatrix}$$

And then we obtained the 32nd step transition probability matrix.

$$P^{32} = \begin{pmatrix} 0.2524 & 0.5214 & 0.1180 & 0.1098 \\ 0.2524 & 0.5214 & 0.1180 & 0.1098 \\ 0.2524 & 0.5214 & 0.1180 & 0.1098 \\ 0.2524 & 0.5214 & 0.1180 & 0.1098 \end{pmatrix}$$

This obtained the steady state probabilities of the followings:

$$\pi_1 = 0.2524,$$

$$\pi_2 = 0.5214,$$

$$\pi_3 = 0.118 \text{ and}$$

$$\pi_4 = 0.1098$$

are steady state probabilities at the 32-step transition matrix.

Markov chain with finite states has long run behaviors under relatively general conditions. The steady-state equations can also be expressed as:

$$\pi_1 = \pi_1 P_{11} + \pi_2 P_{21} + \pi_3 P_{31} + \pi_4 P_{41}$$

$$\pi_2 = \pi_1 P_{12} + \pi_2 P_{22} + \pi_3 P_{32} + \pi_4 P_{42}$$

$$\pi_3 = \pi_1 P_{13} + \pi_2 P_{23} + \pi_3 P_{33} + \pi_4 P_{43}$$

$$\pi_4 = \pi_1 P_{14} + \pi_2 P_{24} + \pi_3 P_{34} + \pi_4 P_{44}$$

$$1 = \pi_1 + \pi_2 + \pi_3 + \pi_4$$

Substituting values for p_{ij} into these equations leads to the following equations and solve for steady state probabilities. Solving these equations simultaneously provides the solution:

$$\pi_1 = 0.2524,$$

$$\pi_2 = 0.5214,$$

$$\pi_3 = 0.118 \text{ and}$$

$$\pi_4 = 0.1098$$

These are the same with the solutions obtained by using **Chapman-Kolmogorov** equations. It does not express the time to be get these probabilities.

As the result, the higher order of transition probability matrix shows that the probability of the market price after the next 32nd day trading. The overall analysis of the results can be summarized as follows. Based on the probability, the prediction of the price can express as 25%, 52%, 11% and 11%. The market price of onion will be after 32 days as follows. The percentage that price of onion per peittha below 550 will be 25%, the percentage that price of the onion per peittha on $550 < \text{price} \leq 750$ will be 52%, price of the onion per peittha on $750 < \text{price} \leq 950$ will be 11% and that above 950 will be 11%. Based on the previous research data, the market price of onion in Pakokku will be most likely between 550 and 750. Thus, the farmers in Pakokku will be predicted the market price of the local onion.

We can express the result as the histogram in the Figure 3.

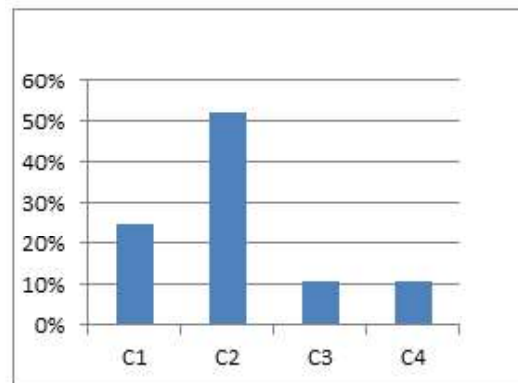


Fig 3. Histogram for percentage of price

5. CONCLUSION

This paper predicts the price of onion in Pakokku District in Myanmar for near future using Markov Chain. The prediction is for the day after 32 days. The actual market price of onion is in the prediction of the price interval by our result. The daily price of onion should exact value not interval for prediction. Our product is one of export products of Myanmar. Products market price is changing on day by day. The prediction of the price is very important for the farmers of Myanmar. The fluctuation of onion price is influenced by many factors, therefore, the prediction made for the near future price by any one method may not be adequate but the results obtained by Markov Chain model was quite encouraging. The prediction of our result has an error. Future plan also be to find that error and also to predict the general prediction that the price will be increase the price or decrease the price or remain the same.

ACKNOWLEDMENT

I would like to give thanks to Dr. Mie Mie Khin, the Rector, University of Computer Studies (Meikhtila) for her guide and support.

REFERENCES

- [1] M. K. Bhusal, "Application of Markov Chain Model in the Stock Market Trend Analysis of Nepal," vol. 8, no. 10, p. 13, 2017.
- [2] X. Zhu, X. Liu, L. Bao, and X. XU, "Markov Chain Analysis and Prediction on the Fluctuation Cycle of Vegetable Price," in *LISS 2012*, Berlin, Heidelberg, 2013, pp. 1457–1463.
- [3] A. Bairagi and S. C. Kakaty, "Markov Chain Modelling for Prediction on Future Market Price of Potatoes with Special Reference to Nagaon District," p. 7.
- [4] P. Jasinthan, A. Laheetharan, and N. Satkunanathan, "A Markov Chain Model for Vegetable Price Movement in Jaffna," *Sri Lankan J. Appl. Stat.*, vol. 16, no. 2, pp. 93–105, Oct. 2015.
- [5] "Hillier - Introduction to Operations Research 9th c2010 solutions ISM," *StuDocu*. [Online]. Available: <https://www.studocu.com/en/document/northeastern-university/operations-research/other/hillier-introduction-to-operations-research-9th-c2010-solutions-ism/3711980/view>. [Accessed: 12-Dec-2019].

STUDYING THE FP-GROWTH METHOD AND FP-TREE METHOD WITH CASE STUDY FOR ASSOCIATION RULES MINING

Aye AyeThein⁽¹⁾, Nan Yu Hlaing⁽²⁾

⁽¹⁾University of Computer Studies (Banmaw), Myanmar

⁽²⁾Myanmar Institute of Information Technology (Mandalay), Myanmar

⁽¹⁾ *ayethein14@gmail.com*

ABSTRACT

Association rules are the knowledge that to finding of recurring relationship among item sets. Finding the frequent item sets transaction in massive databases has been supported famously in data mining technology. Frequent patterns Information leads to the discovering of interesting path of tree graph that informative frequently occurrence set of items execution and visualization. FP-growth which mining the transaction database without candidate generation has been developed. This method is used to project and partition databases based on the discovered frequent item sets and grow of patterns to longer it's in the projected databases. The graph structure of FP growth is designed that it can be easily maintained and compressed into smaller data structure for mining the complete set of items by pattern fragment growth. This paper describes to reduce size of database and perform mining recursively. The sample dataset take as a input in FP-Growth algorithm to get association rules of that transaction data in this paper implementation.

KEYWORDS: *Algorithm, FP-Growth, FP-Tree, Item sets, Conditional Pattern Base*

1. INTRODUCTION

There are many association rule mining algorithms are available such as Apriori, AprioriTID, CASTS Tree, FP-growth and FIN algorithm etc. Extraction of association rules mines frequently patterns, correlation, associations between item-sets

in data repositories. An association rules have two parts namely suffix pattern and prefix pattern. The result is an item set that is found in the sequence of prefix. Rules are observed by splitting data for frequent suffix/prefix patterns and confidence value to extract the correlation. Support refers to how often the items appear in the database and confidence refers to the number of times the suffix/prefix statements have been found. The growth in frequent item-sets has been developed in data mining methodology that for extraction of knowledge from data. It performs store and management in database and affects the process of decision making is more important. Most of the developed methods adopt a candidate generation approach, which scans the transaction db multiple times. The first scan of the transaction database is to find all of the length-1 frequent items sets based on satisfying a minimum support threshold value 2. The k-th item sets for $k > 1$ scan starts with test of items sets length(k-1) frequent patterns found and generates new items sets length-k patterns called candidate patterns. Arrange these items sets into a list L in sorted according to descending support count which derives the set of frequent items. The candidates which assess the minimum support threshold are identified and to find for next pass. The execution of terminates where there is no candidate pattern that can be generated in any pass. But a challenge in mining frequent item sets from a large data set, when the min_sup threshold is set low or frequent item sets of length of pattern to be generated is long; the candidate generation algorithm may suffer from non-trivial costs. An interesting method in this strategy attempts to count information of frequency relevant data sets.

2. METHODOLOGY

2.1 A Motivating Statement

Frequent item set mining is a popular data mining technique. Apriori, and FP-Growth are among the most common algorithms for frequent item set mining.

Considerable research area has been performed to compare the relative performance between these two algorithms, by evaluating the scalability of each algorithm as the dataset size increases. While scalability as data size increases is important, the performance impacts datasets that contain different item set feature. The Apriori candidate generate-and-test method reduces the size of candidate sets, leading to good performance gain. However, it can suffer from two nontrivial cost:

- 1) It may need to generate a large number of candidate sets.
- 2) It may need to repeatedly scan the whole database and check a large set of candidates by pattern matching.

So, can it interesting design a method that mines the complete set of frequent item sets without a costly candidate generation process?

An interesting method of this paper describes which is called FP-growth. With the help of the mining technique implemented in this paper, Tree and Graph method can determine the relationship between frequent patterns how frequent item sets can be mined efficiently using algorithm.

2.2 Related Work

Data mining is also known as extracting or mining information from data set. It is called a method of knowledge discovery where visualization and knowledge mining method is used to present the mined knowledge to the user behavior [4]. There are many papers discussing and implementing about the FP-Growth algorithm. The frequent pattern mining has active applications which incorporate clustering, classification and various problems.

Jiawei Han and Jian Pei [3] representative of FP-Growth for constructing the tree node link approach and binary tree approach. They derived from the frequent pattern-growth method are more efficient and scalable than other pattern mining methods.

AmanvirKaur, Dr. GagandeepJagdev [4] elaborated the advantages and disadvantages associated with the FP-Growth algorithm.

Ms. ShilpaGuranani and Prof. M. Vijaylakshmi [7] discovered and extracted infrequent patterns in the application between relevant items and irrelevant items within each transaction.

Sanjay Patel, Dr.K.Kotecha, Viral Patel, harsh Shukla, Dhyey Shah, Sharan Raghu [8] compared Apriori, FP-Growth and extension of FP-Growth by requiring time for different minimum support value.

Frequent pattern mining in data classification has become an important data mining task. The set of items at the dataset, each item has Boolean variable. That value can be analyzed for corresponding patterns that reflect items that are frequently associated together. These patterns can be represented in the form of association rules. Support and confidence of association analysis are two measures of interesting rules. Association rules are interesting if it is satisfy both a minimum support threshold and minimum confidence threshold. The aim of this paper isto avoid the costly process of candidate generation used by Apriori.

2.3 FP-GrowthAlgorithm

Input:

1. D , transaction database
2. Threshold values for each item
3. Support and confidence threshold values for association rules

Output:

1. Frequent item s
2. Association rules

Procedure

Step: 1 Consider a database $D = \{T_1, T_2, \dots, T_n\}$

Transaction $T_1 = \{I_1, I_2, \dots, I_p\}$, where each

T_1 has set of item s $T_{1j} = \{T_{11}, T_{12}, \dots, T_{1n}\}$,

$j = 1, 2, \dots, n$ each I_n has individual threshold

Step: 2

If Tree contains a single path R then

for each combination β of the node link in R


```

generate pattern  $\beta \cup \alpha$  with
 $TIn = \{TI1, TI2, \dots, TIN\}$ , individual threshold
values of nodeslink in  $\beta$ 
else for each headerlink  $ai$  in the
header of Tree
do {
generate pattern  $\beta = ai \cup \alpha$  with
 $S_{supportcount} = ai.S_{supportcount}$ ;
Construct  $\beta$ 's conditional
pattern
base;
 $C_k =$  Construct  $\beta$ 's conditional FP-tree
 $S_{supportcount} = \text{Min}(TIn)$ , select the
minimum threshold item
from the combination of
items
K=K+1}
    
```

Fig. 1 FP-Growth for finding frequent item sets for extracting association rules

3. PERFORMANCE AND EXPERIMENT SETTING

This system considered to constructing tree using the following transaction database TDB for experiment.

Table1. A transaction DB as running data

TID	ITEMLISTS
T001	TV, SETTOPBOX, ANTENNA
T002	SETTOPBOX, SKYNET
T003	SETTOPBOX, PSI
T004	TV, SETTOPBOX, SKYNET
T005	TV, PSI
T006	SETTOPBOX, PSI
T007	TV, PSI
T008	TV, SETTOPBOX, PSI, ANTENNA
T009	TV, SETTOPBOX, PSI

Step1: Counting the numbers of occurrences of each item,

ITEMLISTS	Sup_count
{ TV }	6
{ SETTOPBOX }	7
{ PSI }	6
{ SKYNET }	2
{ ANTENNA }	2

Step2: Sorting the support count order

TID	ITEMLISTS
T001	SETTOPBOX, TV, ANTENNA
T002	SKYNET, SETTOPBOX
T003	SETTOPBOX, PSI
T004	SETTOPBOX, TV, SKYNET
T005	TV, PSI
T006	SETTOPBOX, PSI
T007	TV, PSI
T008	SETTOPBOX, TV, PSI, ANTENNA
T009	SETTOPBOX, TV, PSI

Then FP-tree is constructed as follows. First, create the root of the tree, labeled with "null". Scan database D a second time. In each transaction, the items are processed in L order, and a sub branch is created for each transaction. The branch to be added for a transaction, the count of each node through a common prefix is incremented by 1, and nodes for the items following prefix are created and linked accordingly. The FP-tree is mined from each frequent item sets length-1 pattern as suffix pattern, construct its conditional pattern base as a sub-database, which contain the set of prefix paths in the FP-tree, perform mining recursively on the tree. The pattern growth is acquired by the concatenation of the suffix pattern with the frequent patterns generated from a conditional FP-tree. FP-tree mining is the following Table1. The scan of database which derives the set of frequent items (1-itemsets) and support counts

threshold value 2. The second transaction contains the item “SETTOPBOX” and “SKYNET” in L order, which result in a path where “SETTOPBOX” is linked to the root and “I4” is linked to “SKYNET”. This path shares a common prefix “SETTOPBOX”, with the existing path for identifier T001. So, increment the count of the “SETTOPBOX” node by 1 and create a new node {SKYNET: 1} that links a sub node to {SETTOPBOX: 2}. And then finally the transaction TID009, Consider “ANTENNA” is the last item in L because for starting at the end of the list will become apparent as the FP-tree mining process “ANTENNA” occurs in two FP-tree paths. The occurrences of “ANTENNA” can be found by its chain of node-links. For “PSI” instance, Conditional pattern base = {{SETTOPBOX, TV:2}, {SETTOPBOX:2}, {TV:2}}. For this item of conditional FP-tree have 2 paths, {SETTOPBOX:4, TV:2}, {TV:2}

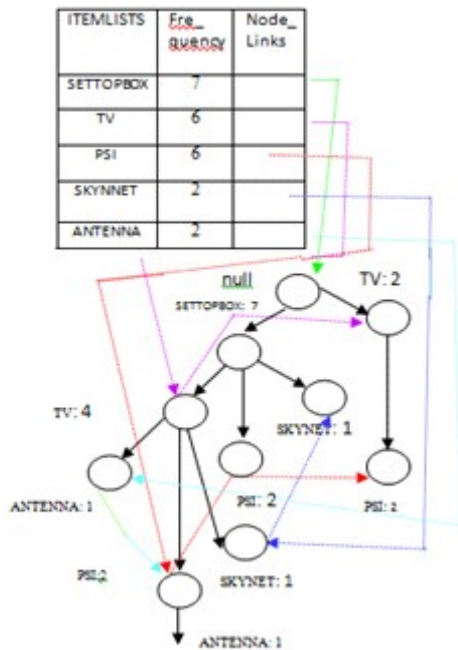


Fig.2 Frequent Pattern Information based FP-Tree

Table 2. Mining the FP-tree by creating conditional pattern bases

STEREOLISTS	Conditional Pattern Based
ANTENNA	{{(SETTOPBOX_TV=1), (SETTOPBOX_TV_PS=1)}}
SKYNET	{{(SETTOPBOX_TV=1), (SETTOPBOX=1)}}
PSI	{{(SETTOPBOX_TV=2), (SETTOPBOX=2),(TV=1)}}
TV	{{(SETTOPBOX=4)}}

ITEMISTS	Conditional F#-rec
ANTENNA	{SETTOPBOX:2, TV:2}
SKYNET	{SETTOPBOX:2}
PSI	{SETTOPBOX:4, TV:2}, {TV:2}
TV	{SETTOPBOX:4}

ITEM/ITEMS	Frequent Pattern Generated
ANTENNA	{SETTOPBOX, ANTENNA-2}, {TV, ANTENNA-2}, {SETTOPBOX, TV, ANTENNA-2}
SEWNET	{SETTOPBOX, ANTENNA-2}
PSI	{SETTOPBOX, PSI-6}, {TV, PSI-6}, {TV, PSI-6}, {SETTOPBOX, TV, PSI-6}
TV	{SETTOPBOX, TV-6}

The branch has one item on the left-hand-side. An FP-tree registers constructed, frequent pattern information is more efficient strategy because it reduces scanning time period to transitional database. Each of the items in the frequent pattern generated table has an association link which provides association rules information about categories. All data in this transaction,

two frequent item sets $X = \{\text{SETTOPBOX, TV, ANTENNA: 2}\}$ and $\{\text{SETTOPNOX, TV, PSI: 2}\}$, the resulting association rules six rules that satisfy minimum support count = 2 and confidence from these two item sets are the following:

- $$\begin{aligned} \{TV, SETTOPBOX\} \circ! ANTENNA, & \text{—————} (1) \\ \{TV, ANTENNA\} \circ! SETTOPBOX & \text{—————} (2) \\ \{SETTOPBOX, ANTENNA\} \circ! TV, & \text{—————} (3) \\ \{TV\} \circ! \{SETTOPBOX, ANTENNA\}, & \text{—————} (4) \end{aligned}$$

{SETTOPBOX,PSI} → {TV,ANTENNA}, ————— (5)
 {ANTENNA,PSI} → {TV,SETTOPBOX}, ————— (6)
 For rule (1), confidence=50%
 For rule (2), confidence=100%
 For rule (3), confidence=100%
 For rule (4), confidence=33%
 For rule (5), confidence=29%
 For rule (6), confidence=100%

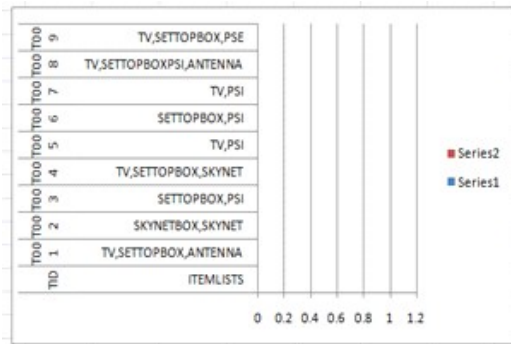


Fig.3 Number of Frequent Items

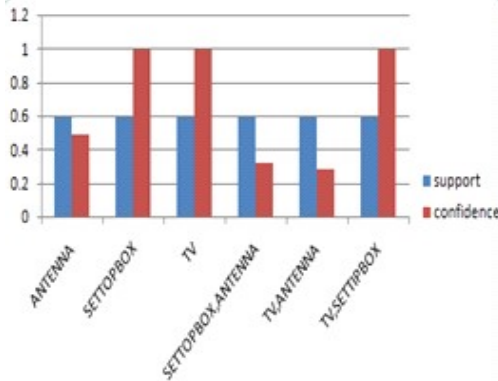


Fig. 4 Number of Rules Generation

4. CONCLUSIONS AND FUTURE WORK

This implementing system provides decision what kind of pattern are association rules. As a result, association rule extracted from the second frequent item set {TV, SETTOPBOX, ANTENNA} that satisfy the confidence threshold value, so rule 2, 3, 6 are result of output. In this paper, system implemented FP-Tree is mined and FP-growth method transforms the finding frequent patterns in conditional

databases recursively and corresponding the suffix offering pattern information. Based on this method, system demonstrated on FP-Growth by reducing the size of candidate sets. This performance derived from the FP tree method that is efficient and scalable for mining frequent pattern by using algorithm. These processes that manipulate sets of instances are classified corresponding to in any context. Moreover, this paper is intended to extract association rules. The experiment results perform using item set in database and thresholds by using FP-Growth algorithm. For future implementation will comparison results of another various algorithm namely Apriori algorithm of association rule mining.

ACKNOWLEDGEMENT

We would like to express special thanks to all the people and all member of UCSMTLA organizing committee "Journal of Research & Applications (JRA), UCSMTLA, 2019, Volume-01, Issue-01" for paper invitation. We are grateful to all teachers and reviewers who provided helpful during the development of this paper and the faith they have had in me.

REFERENCES

- [1] J. Han, M. Kamber and J. Pei. Data Mining Concept and Techniques, University of Illinois at Urbana-Champaign, 2012.
- [2] Jian Pei. "Pattern-growth Methods for Frequent Pattern Mining," Peking University, 1999.
- [3] Jiawei Han, Jian Pei, "Mining Frequent Patterns by Pattern-Growth: Methodology and Implications," Simon Fraser University, December 2000.
- [4] Amanvir Kaur¹, Dr. Gagandeep Jagdev², "Analyzing Working of FP-Growth Algorithm for Frequent Pattern Mining," Research Scholar (M.Tech.), Yadavindra College of Engineering, Talwandi Sabo (PB), "International Journal of Research Studies in Computer Science and Engineering (IJRSCSE) Volume 4, Issue 4, 2017.
- [5] J. Breckling, Ed., The Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser. Lecture Notes in Statistics, Manasa G. Mrs. Kulkarni Varsha, "IAFP: Integration of Apriori and FP-Growth Techniques to Personalize Data in Web Mining," International

Journal of Scientific and Research Publications, Volume5, Issue 7, July 2015. ISSN 2250-3153.

- [6] T. Seidl, "Knowledge Discovery in Databases, SS2016," Ludwig-Maximilians.
- [7] S. Gurnani, M. Vijaylakshmi, "Frequent Pattern Growth Method For Infrequent Weighteditemset Mining," International Journal of Engineerng Sciences & Research Technology, ISSN: 2277-9655, Impact Factor: 4.116, CODEN: IJESS7.
- [8] S. Patel, K.Kotecha, V. Patel, H. Shukla, D. Shah, S. Raghu, "Frequent Pattern mining UsingNovel FP-GrowthApproach" ,IJCSC, Vol 5, 2014 .
- [9] V.Ramya, M.Ramakrishnan, " Usage of Dimension Tree and Modified FP-Growth Algorithm for Association Rule Mining on Large Volumes of Data," Journal of Engineering and Applied Sciences 13(7): 1670-1675,2018,ISSN:1816-949X, Medwell Journals,2018.
- [10] J. Heaton, "Comparing Dataset Characteristics that Favor theApriori, Eclat or FP-Growth Frequent Itemset Mining Algorithms," arXiv:1701.09042v1[cs.DB] 30 Jan 2017.
- [11] S. Sharmila, S. Vijayarani, "Association Rule Mining Using Enhanced FP-Growth and H-Mine Algorithms," International Journal of Electrical Electronics & Computer Science Engineering , Vol 4, Issue 4, 2017.

PREDICTION TELECOM CLIENT ATTRITION USING CART ALGORITHM

Su Mon Ko ⁽¹⁾, Su Myat Sandar Win⁽²⁾, Theint Win Lai ⁽³⁾

⁽¹⁾⁽²⁾⁽³⁾University of Computer Studies (Meiktila), Myanmar

sumonko@ucsmtla.edu.mm

ABSTRACT

The quick development of the market in each division is prompting unrivaled subscriber base for specialist organizations. In this aggressive world, mobile telecommunications market will in general arrive at a saturation state and faces a furious challenge. This circumstance powers the telecom organizations to concentrate on keeping the clients intact as opposed to building a large client base. Client Relationship Management frameworks are utilized to empower organizations to obtain new clients, build up a continuous relationship with them and increment client retention for greater benefit. CRM frameworks use machine learning models to analyze client's personal and behavioral data to give organization a competitive advantage by increasing client retention rate. This paper attempts to apply machine learning technique that is utilized for churn forecast issue. Decision Trees (Classification and Regression Trees, CART) method is picked for this examination. The exploratory outcome accomplished 67% exactness.

KEYWORDS: *Customer Relationship Management (CRM), Decision Tree, Customer Churn Prediction, Machine Learning*

1. INTRODUCTION

Client churn can be characterized as the loss of clients, and it is brought about by an adjustment in taste, absence of appropriate client relationship system, change of living arrangement and a few different reasons. Overseeing client churn is a major issue confronting organizations, particularly those that offer subscription-based services. The organizations are interested on recognizing jobs of the clients in light of the fact that the cost for obtaining another client is typically higher than holding the old one. To decrease client churn, the

organization ought to have the option to foresee the conduct of client effectively and set up relations between client attrition and keep factors under their control [1]. Client churn forecast has been progressively researched in numerous business areas, including, however not restricted to, media transmission, banking, retail and cloud administrations subscriptions. Distinctive measurable and machine learning methods are utilized to address this issue [2]. Machine Learning is strategy for information examination which computerizes analytical model building. Using algorithms that iteratively learn from data, machine learning allows systems to explore hidden patterns without being explicitly programmed where to look [1]. Decision trees are used in data mining to study historical data and on the basis of the data analysis and its rules, one can predict the result. This paper proposed a basic expectation model utilizing decision tree calculation to anticipate telecommunication client churn attrition.

2. RELATED WORK

Abdelrahim Kasem Ahmad and Assef Jafar and Kadan Aljoumaa developed a telecom customer churn prediction model using machine learning techniques on big data platform. The dataset contained all customers' information over 9 months, and was used to train, test, and evaluate the system at SyriaTel. Four algorithms are experimented: Decision Tree, Random Forest, Gradient Boosted Machine Tree "GBM" and Extreme Gradient Boosting "XGBOOST". The best results were obtained by applying XGBOOST algorithm.

Federico Castanedo, Gabriel Valverde, Jaime Zaratiegui and Alfonso Vazquez reported results for predicting customer churn using four-layer feedforward architecture. In this work, they used billions of call records from an enterprise business

intelligence system and present their work towards using deep learning for predicting churn in a prepaid mobile telecommunication network. Their model achieved 77.9% AUC on validation data.

V. Umayaparthi and K. Iyakutti developed three deep neural network architectures and built the corresponding churn prediction model using two telecom dataset. In this work, they used Cell2Cell dataset and CrowdAnalytix dataset. Their results show that deep-learning based models are performing as good as traditional classification models, without even using the hand-picked features.

3. THEORETICAL BACKGROUND

Machine Learning techniques have been generally utilized for assessing the probability of client to churn. In view of a review of the writing in churn prediction, the methods utilized in the bulk of written works can be categorized as one of the accompanying classes 1) Regression examination; 2) Tree – based; 3) Support Vector Machine; 4) Bayesian calculation; 5) Ensemble learning; 6) Sample – based learning; 7) Artificial neural network; and 8) Linear Discriminant Analysis [2].

3.1 Decision Trees Learning

Decision Tree (DT) is a model that creates a tree-like structure that speaks to set of decisions. DT restores the probability scores of class participation. DT is made out of: 1) **internal Nodes**: each node refers to a single variable/feature and represents a test point at feature level; 2) **branches**, which represent the outcome of the test and are represented by lines that finally lead to 3) **leaf Nodes** which represent the class labels. That is the manner by which decision guidelines are set up and used to group new cases. DT is an adaptable model that supports both all categorical and continuous data. Because of their adaptability they picked up popularity and became one of the most commonly used models for churn prediction [2]. DT has no great performance on capturing complex and non-linear relationships between the attributes. Yet, in the customers churn problem, the accuracy of a DT can be high, depending on the form of the data [Praveen Asthana (2018)]. There are couples of algorithms there to build a decision tree, which are: CART (Classification and Regression Trees) uses Gini Index (Classification) as metric and ID3 (Iterative Dichotomiser 3) uses Entropy function and Information gain as metrics.

3.2 Classification and Regression Trees (CART) Algorithm

The algorithm is based on Classification and Regression Trees by Breiman et al (1984). A CART tree is a binary decision tree that is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole learning sample. The representation for the CART model is a binary tree. Each root node represents a single input variable (x) and a split point on that variable (assuming the variable is numeric). The leaf nodes of the tree contain an output variable (y) which is used to make a prediction. Gini index is a metric for classification tasks in CART. Gini index or Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen. The degree of Gini index varies between 0 and 1, where 0 denotes that all elements belong to a certain class or if there exists only one class, and 1 denotes that the elements are randomly distributed across various classes. A Gini Index of 0.5 denotes equally distributed elements into some classes. It stores sum of squared probabilities of each class. It can illustrate below.

$$Gini = 1 - \sum_{i=1}^n (p_i)^2 \quad (1)$$

4. METHODOLOGY

The initial step before applying the chose analytical models on the dataset, explanatory data analysis for more insights into dataset was performed. In view of the perceptions data was preprocessed to be increasingly reasonable for analysis.

4.1 Dataset

Telco dataset is used for the experiment of this study. It is a database of client information of a telecommunication organization. There are 7032 clients in the dataset and 19 features without customerID (non- informative) and Churn column (target variable). The greater part of the categorical features has 4 or less unique values. This work is experimented using Python programming language.

4.2 Data Preprocessing

Telco dataset has one client for every line with numerous columns (features). There aren't any rows with every missing values or duplicates in the dataset. There are 11 samples that have TotalCharges set to “”, which appears to be a mistake in the data. Firstly those samples are removed and set the type to numeric (float).

4.3 Exploratory Data Analysis

There are 2 kinds of features in the dataset: categorical (two or more values and without any order) and numerical. A large portion of the feature names are self-explanatory, except for: Partner: whether the client has a partner or not (Yes, No), Dependents: whether the client has dependents or not (Yes, No), OnlineBackup: Whether the client has an online backup or not (Yes, No, No internet service), tenure: number of months the client has remained with the company, MonthlyCharges: the sum charged to the client month to month, TotalCharges: the aggregate sum charged to the client.

4.4 Features Distribution

Prior to demonstrate preparing, feature distribution is one of the most significant factors that can influence the performance of models. Numeric summarizing techniques (mean, standard deviation, and so on.) don't show spikes, states of distribution and it is difficult to observe outliers with it. That is the explanation histograms are utilized in this study. At first glance, there aren't any outliers in the data. No data point is disconnected from distribution or too far from the mean value. To confirm that interquartile range (IQR) is calculated and show that values of each numerical feature are within the 1.5 IQR from first and third quartile. Numerical features are converted to ordinal intervals. Figure 2 illustrates the histograms of numerical features.

The result is observed that the greater TotalCharges and tenure are the less is the probability of churn at distributions of numerical features in relation to the target variable. Figure 3 shows the numerical features in relation to the target variable.

To analyze categorical features, bar graphs are utilized in this examination. The outcome is seen that Senior citizens and customers without phone service are less represented in the data.

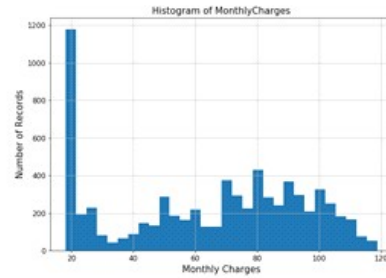


Figure 1: Histograms of monthly charges numerical feature

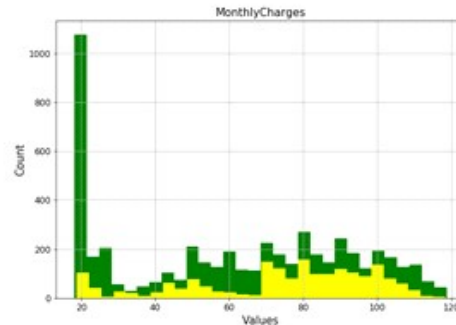


Figure 2: Numerical monthly charges feature in relation to the target variable

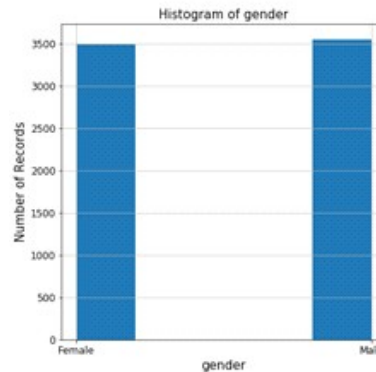


Figure 3: Distribution of gender categorical feature

The following stage is to look at categorical features in connection to the target variable. This is done distinctly for contract feature. Clients who have a month-to-month contract are more likely to churn than users with long term contracts.

4.5 Decision Tree

Decision Tree model uses Contract, MonthlyCharges, InternetService, TotalCharges, and tenure features to settle a decision if a client will churn or not. These features separate churned clients from others very much dependent on the split criteria in the decision tree. Every client test crosses the tree and last node gives the forecast. If contract_month-to-month is equivalent to 0, keeps crossing the tree with true branch, equivalent to 1, keeps traversing the tree with false branch, and not defined, it yields the class 0. This is an incredible way how the model is settling on a decision or if any features sneaked in this model shouldn't be there. Categorical and numerical features are separately processed. Categorical features are one-hot encoded and scaled numerical features by removing the mean and scaling them to unit variance. A decision tree model is chosen because of its interpretability and set max depth to 3 (arbitrarily).

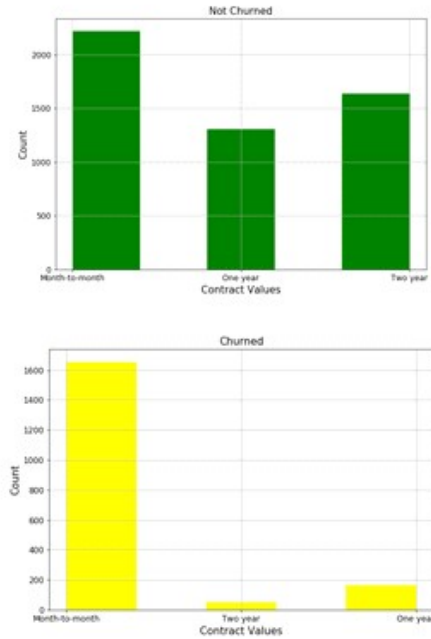


Figure 4: Contract feature in relation to the target variable

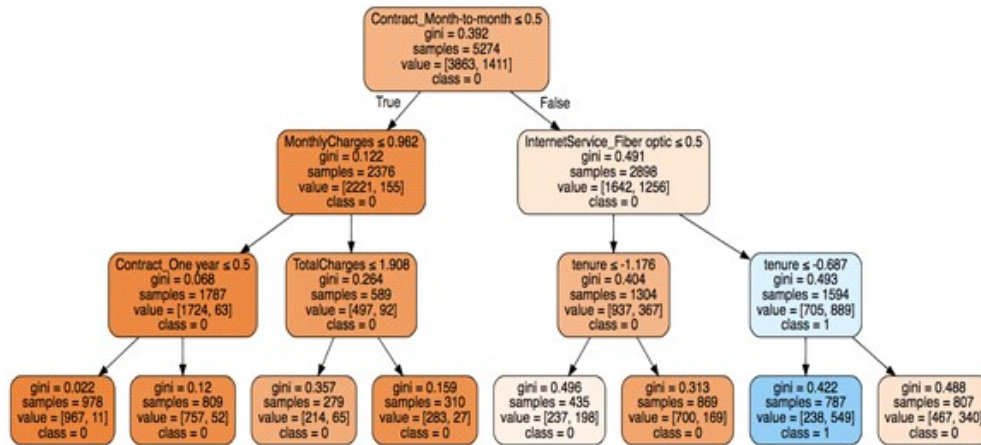


Figure 5: Interpretation of the Decision tree model

5. RESULTS

Accuracy is utilized to assess the model performance. It shows the capacity to separate the credible and non-credible cases accurately. It's the proportion of true positive (TP) and true negative (TN) in all evaluated news:

$$\text{Precision} = \frac{tp}{tp + fp} \quad (2)$$

$$\text{Recall} = \frac{tp}{tp + fn} \quad (3)$$

Where,

tp: is the absolute number of clients effectively distinguished as churn.

fp: is the absolute number of clients inaccurately distinguished as churn.

tn: is the absolute number of clients effectively distinguished as no- churn.

fn: is the absolute number of clients inaccurately distinguished as no- churn.

Table 1. The precision and recall results

	Precision	Recall
No	0.81	0.94
Yes	0.67	0.37

Table 1 shows the precision and recall results. The dataset is part into train (75% samples) and test (25% samples). With classification report precision and recall is determined with actual and predicted values. Precision tells what number of churned clients did the classifier anticipated effectively. On the opposite side, recall tells what number of churned clients it missed. The classifier isn't precise for churned clients. For class 1 (churned clients) model accomplishes 0.67 precision and 0.37 recall.

6. CONCLUSION

There are simple decision rules based models and a complex classification model for churn prediction task has been proposed in the literature. The exactness of the chose models was assessed on the clients of Telco dataset. The investigation can

be stretched out by including hybrid models and deep learning models. Other execution measurements can be utilized for execution assessment. Timing measures of the models can also be a major indicator for performance. Models can also evaluate against various datasets from various domains.

ACKNOWLEDMENT

I would like to express my special thanks to all my teachers who gave me their time and guidance, and all my friends who helped in the task of developing this paper. Finally, I would like especially to thank my parents for their continuous support and encouragement throughout my whole life.

REFERENCES

- [1] S. Kumar and C. D., "A Survey on Customer Churn Prediction using Machine Learning Techniques," *Int. J. Comput. Appl.*, vol. 154, no. 10, pp. 13–16, Nov. 2016.
- [2] S. F., "Machine-Learning Techniques for Customer Retention: A Comparative Study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 2, 2018.
- [3] "Customer churn prediction in telecom using machine learning in big data platform," *ResearchGate*. [Online]. Available: https://www.researchgate.net/publication/331908752_Customer_churn_prediction_in_telecom_using_machine_learning_in_big_data_platform.
- [4] "Customer Churn Warning with Machine Learning," *ResearchGate*. [Online]. Available: https://www.researchgate.net/publication/329893308_Customer_Churn_Warning_with_Machine_Learning.
- [5] F. Castanedo, "Using Deep Learning to Predict Customer Churn in a Mobile Telecommunication Network," 2014.
- [6] V. Umayaparthi and K. Iyakutti, "Automated Feature Selection and Churn Prediction using Deep Learning Models," vol. 04, no. 03, p. 9.
- [7] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. Ch. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simul. Model. Pract. Theory*, vol. 55, pp. 1–9, Jun. 2015.

DATA MINING CLASSIFICATION TECHNIQUES FOR CARDIOVASCULAR DISEASE DIAGNOSIS

Hnin Ei Ei Cho⁽¹⁾, Nan Yu Hlaing⁽²⁾

⁽¹⁾⁽²⁾ Myanmar Institute of Information Technology, Mandalay, Myanmar

⁽¹⁾ *hnin_ei_ei_cho@miit.edu.mm*, ⁽²⁾ *Email: nan_yu_hlaing@miit.edu.mm*

ABSTRACT

The huge amounts of data generated by healthcare transactions are complex and voluminous. We process and analyze them by using different traditional methods. The healthcare industry collects huge amounts of healthcare data, which, unfortunately, are not “mined” to discover hidden information. Mining Techniques offer a principled approach for developing sophisticated, automatic, and objective algorithms for analysis of high dimensional and multimodal biomedical data. Medical diagnosis is the process of determining which disease or condition explains a person's symptoms and signs. In this study, we briefly examine the potential use of classification-based data mining techniques to massive volume of healthcare data. Aim of the paper is to propose a model for early detection and correct diagnosis of the disease, which will help the doctor in saving the life of the patient.

KEYWORDS: *Cardiovascular Disease, Classification, Random Forest (RF), Artificial Neural Network (ANN), Support Vector Machine (SVM)*

1. INTRODUCTION

Cardiovascular disease (CVDs) are a group of disorders of the heart and blood vessels. They include coronary heart disease, cerebrovascular disease, peripheral arterial disease, rheumatic heart disease, congenital heart disease, deep vein thrombosis and pulmonary embolism. This disease attacks a person so instantly that it hardly gets any time. s[1]. CVD is the number one cause of death globally: more people die annually from this disease than from any other cause. An estimated 17.9 million people died from CVDs in 2016, representing 31% of all global deaths. Of these deaths, 85% are due to heart attack and stroke. Over three quarters of CVDs,

deaths take place in low- and middle- income countries. People with cardiovascular disease or with at high cardiovascular risk need early detection and management using counselling and medicines, as appropriate. Recent advances in health related studies are concentrating on risk prediction of diseases.

In the study of risk prediction from patients' health records, different classification techniques are used. Classification problems are prediction of class labels where number of classes is fixed and pre-defined. There is nothing like a particular classification method is accurate to classify the data in all situations. The accuracy of classification method is depends on the data we want to classify. Effective and efficient automated cardiovascular disease prediction can benefit healthcare sector and this automation will save not only cost but also time [2]. This research paper highlights the utility and application of three different classification models of data mining techniques for prediction of cardiovascular disease to facilitate experts in the healthcare domain.

We divide this paper into the following sections: section two contains the related work, section three describes the methods and materials which include a description of the datasets, section four illustrates data preprocessing techniques used and how the system works, and section five, which contains the conclusion and future scope. For this research, we referenced these works:

- Chaithra and Madhu, (2018)
- Dhiyaa ALjdiao, Hamed Monkaresi, (2017)
- Deepti Sisodia, Dilip Singh Sisodia, (2018)
- Rahul Joshi , Minyechil Alehegn, (2017)

- Dipti N. Punjani, Kishor Atkotiya, (2018)

2. RELATED WORK

Chaithra and Madhu proposed a system for diagnosis of cardiovascular patients [3]. They predicted heart disease from echocardiography dataset and analyzed by applying techniques prospectively. They investigated three different classification models: J48 Decision Tree, Naive Bayes and Neural Network on cardiovascular disease prediction. That analysis showed that Neural Network performed better in predicting the heart disease with 97.91% of accuracy.

Dhiyaa ALjdiaoi, Hamed Monkaresi in [4] designed the system based on Cleveland Heart Disease Dataset, that consists of 13 features are considered as input. That research being carried out using the data mining techniques to enhance heart disease diagnosis and prediction including decision trees, Naive Bayes classifiers, K-nearest neighbor classification (KNN) and support vector machine (SVM). Results show that NB classifier achieve 87.45% of classification accuracy.

Deepti Sisodia, Dilip Singh Sisodia in [5] investigated three machine-learning classification algorithms and evaluated on various measures. They performed experiments on Pima Indians Diabetes Database. Their experimental results determine the adequacy of the designed system with an achieved accuracy of 76.30 % using the Naive Bayes classification algorithm.

In [6], they settled on employing four most known machine-learning algorithm (Random Forest (RF), KNN, Naive Bayes, and J48) classification algorithm and ensemble/combined them in to one using base learner. They found that single algorithm provided less accuracy than ensemble one. In their study, hybrid system Weka and java are the tools to predict diabetes dataset.

Dipti N. Punjani, Kishor Atkotiya discussed how they used Naive Bayes algorithm to perform classification in [7]. They studied that Naive Bayes is not suitable when features are dependent on each other and therefore they advised to develop a method checking for the appropriateness of an algorithm before used in real life.

3. METHODS AND MATERIALS

The researchers have found that no classifier that generates the best result for each dataset. In this

paper, we have investigated three data mining techniques. We used these algorithms to predict the survivability rate of cardiovascular disease data set. We selected these three classification techniques to find the most suitable one for predicting cardiovascular survivability rate.

In this work, we use Python programming language. Python provides a variety of efficient tools for data mining and data analysis. Among them, we use scikit-learn. It is a free software machine-learning library for the Python programming language. It features various classification, regression and clustering algorithms.

3.1. Random Forest (RF)

Random Forest is the one of the Classifier for Classifications problems. Random Forest is ensemble classifier made using many decision trees where ensemble means that uses multiple machine-learning algorithm to obtain the predictive performance. It is better than other classifiers for the prediction of cardiovascular disease but it takes more learning time [8].

Random Forest can handle large set of data with high dimensionality. It is useful in the case of missing data and fit for some datasets with noisy classification/regression tasks. Classifications made by random forests are difficult to interpret.

3.2. Artificial Neural Network (ANN)

Artificial Neural networks are those systems modeled based on the human brain working. Multi-Layer Perceptron (MLP) network models are the popular network architectures. It is used in most of the research applications in medicine, engineering, mathematical modeling, etc. Artificial Neural Network Have the ability to work with inadequate knowledge and fault tolerance. It also has numerical strength that can perform more than one job at the same time. The realization of the equipment is dependent. It is difficult to know how many neurons and layers we need to process and duration is unknown.

The challenge of training neural networks involves carefully selecting the learning rate. It may be the most important parameter for the model. It controls the rate or speed at which the model learns. In this system, we use the learning rate of 0.1; weights in the network are updated 10% of the estimated weight error at each time to update the weights. By using this rate, it allows the model to learn optimal but may take significantly longer to train.

3.3. Support Vector Machine (SVM)

Support vector machine is closely relating to neural networks. The aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data [9]. This classifier is very effective in high dimensional spaces. In SVM, the risk of overfitting is less. It works well with unstructured and semi structured data. Training time is long for large datasets. It is hard to visualize their impact.

SVM is capable of generating very fast classifier functions following a training period. Three approaches to classification problems are multiclass ranking, one-against-all and pairwise. In this system, we use pairwise classification because the used dataset has large records and we found that this approach performed reasonably well for this work.

3.4 Dataset Description

We design the system to integrate multiple indicators from many data sources to provide a comprehensive picture of the public health burden of CVDs and associated risk factors in the United States. There are two different data sets. Table 1 shows the details of the datasets.

Table 1. Dataset description

Database	No. of Attributes	No. of Instances
Cleveland	14	303
Cardio Train	12	700,000

The datasets used for experiments are:

1. Cleveland dataset provided by DHDSP, the National Cardiovascular Disease Surveillance System.
2. Cardio train dataset from Kaggle.

Cleveland data set has 75 attributes, but all published experiments refer to using a subset of 14 of them. This database has concentrated on simply attempting to distinguish presence or absence of cardiovascular disease. We are available the original dataset at [10]. By analyzing this dataset, 165 instances (54.46%) are having cardiovascular disease. We describe attributes descriptions in Table 2.

Table 2. Cleveland dataset attributes and their description

Attribute	Description
1. Age	Age in years
2. Sex	1=male;0=female
3. CP	Chest pain type(4 values)
4. trestbps	Resting blood pressure(mm Hg)
5. chol	Serum cholestoral in mg/dl
6. fbs	Fasting blood sugar(1=true;0=false)
7. restecg	Resting electrocardiographic results
8. thalach	Maximum heart rate achieved
9. exang	Exercise induced angina (1=yes;0=no)
10. oldpeak	ST depression induced by exercise relative to rest
11. slope	The slope of the peak exercise ST segment
12. ca	Number of major vessels(0-3)colored by flourosopy
13. thal	3=normal;6=fixed defect;7=reversible defect
target	0 or 1

The other one, cardio train dataset from Kaggle consists of 70,000 records of patients with 11 features plus target. We can download this dataset at [11].The amount of 34,979 instances (49.97%) have the disease. The attribute “cardio” describes the predictable attribute with value “1” for patients with cardiovascular disease and value “0” for patients with no disease. The attribute description for this dataset is as follows:

Table 3. Cardio train dataset attributes and their description

Attribute	Description
1. id	ID Number
2. age	Age in days
3. gender	1=women, 2=men
4. height	Height in cm
5. weight	Weight in kg
6. ap_hi	Systolic blood pressure
7. ap_lo	Distolic blood pressure
8. cholestrol	1=normal, 2=above normal, 3=well above normal
9. gluc	1=normal, 2=above normal, 3=well above normal
10. smoke	Whether patient smokes or not
11. alco	Binary Feature
12. active	Binary Feature
cardio	Target variable (0 or 1)

4. PROPOSED SYSTEM

Proposed research work introduces a framework to develop a classifier based on data mining techniques. Another objective is to perform cross validation of different framework designed for different category of data. In this frameworks datasets are given to preprocessing stage which further classified by selected classifier. This approach involves:

1. Classify dataset
2. Data Cleaning
3. Data Transformation
4. Splitting training and testing data
5. Select classifier with the best performant
6.
 - i. RF
 - ii. ANN
 - iii. SVM
7. Interpret Results

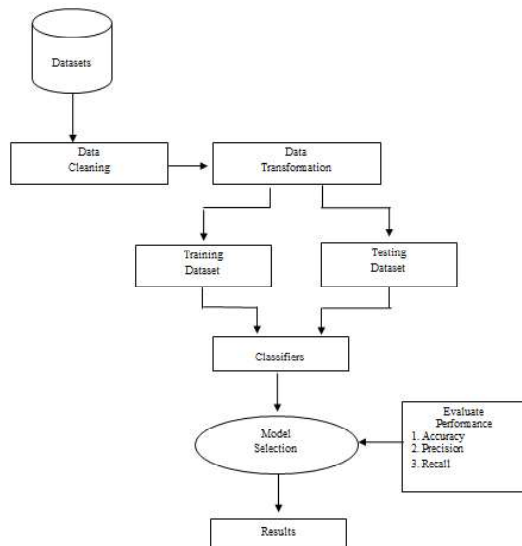


Fig 1. Cardiovascular disease diagnosis system

4.1 Data Preprocessing

Identification of the unnecessary attributes, which impedes the processing task, is crucial before the application of the classification technique. Besides acting as noise and disturbing the process, they also affect the classifier performance. To identify these, employ the statistical methods. We apply the data cleaning techniques first. Identify the missing values and replaced by the group median. Further, apply the min-max scaling technique to have the features value range between zero and one.

4.1.1 Data Cleaning

Some instances have missing data for some of the features. Machine learning algorithms cannot work very well with missing data. To find a solution to “clean” the data, the easiest option is to eliminate all those records, but in this way, we would eliminate many important data. Another option is to calculate the **median** value for a specific column and substitute that value everywhere in the same column that have missing data. The steps and techniques for data cleaning will vary from dataset to dataset.

4.1.2 Data Transformation

Most of the machine learning **algorithms do not work very well if the features have a different set of values**. The solution is to apply the **feature scaling technique**. Feature Selection Technique (FST) eliminates the less important features and reduces the time complexity of the machine learning technique. The type of scaling depends on the data fed to which model, so there is no universally best approach. In this paper, min-max normalization techniques is used. Min-max normalization preserves the relationships among the original data values. It always boosts the classification accuracy and minimizes the computational cost.

4.2 Splitting the Dataset

To check the performance of classifiers, part each dataset into two divisions – training and testing. Test a classifier using a testing dataset, is chosen based on its performance in comparison to other available classifiers. In this paper, we use the K-fold cross-validation method. It partitions the original data set into equal-sized sub-segments. The number of segments depends upon the value of k taken; in our case, we have taken k to be 3, 5, or 10. We use the first part to train the model ignoring the column with the pre assigned label. Then we use the trained model to make predictions on new data, which is the test

dataset, not part of the training set, and compare the predicted value with the preassigned label.

The advantage of using this validation is that we can use every single data is for training as well as in testing the model and each entry in the dataset is used for validation of the result at least once. This helps to increase the accuracy of the model.

4.3. Comparison of different Algorithms

We compare the accuracy of multiple algorithms with two different datasets. To understand classifier's behavior, we should calculate metric Confusion Matrix. This matrix is a visualization tool that present the accuracy of the classifiers in classification [12]. Based on data mining techniques as explained above, evaluated all the developed models in terms of following error measures. True positive (TP) denotes the number of identified positive samples in the positive set. True negative (TN) represents the number of classification negative samples in the negative set. False positive (FP) is the number of identified positive samples in the negative set. False negative (FN) means the number of identified negative samples in the positive set. The accuracy is as the ratio of the number of samples correctly classified by the classifier to the total number of samples.

Table 4. Performance measures

Measures	Definitions	Formula
Accuracy (A)	Determine the accuracy of the algorithm in predicting instances.	$A = (TP + TN) / (TN + TP + FP + FN)$
Precision (P)	Measure the classifier's correctness/accuracy.	$P = TP / (TP + FP)$
Recall (R)	Measure the classifiers' completeness or sensitivity.	$R = TP / (TP + FN)$

In this research, we apply the K-Fold cross-validation technique by considering the different value of k to be 3, 5 and 10. We present the resulted output of the three classifiers that predict the cardiovascular disease using small and large datasets in table 5.

Table 5. Performance comparison of three classifiers

Database	Kth Validation	Classification Model		
		RF	ANN	SVM
Cardio Train	10-fold	71.46	65.51	70.57
	5-fold	69.58	72.99	64.79
	3-fold	70	64.59	72.96
Cleveland	10-fold	83.87	87.10	87.10
	5-fold	88.52	85.25	86.86
	3-fold	84.62	81.32	81.32

By analyzing this result, we should apply K-fold cross-validation if the dataset is small because of getting more performant than larger dataset. According the comparison results, we consider that RF with 5-fold cross-validation gives the most performant algorithm for the Cleveland dataset. Then, using cardio train dataset, ANN with 5-fold have the highest accuracy in all of them. Therefore, we plot the confusion matrix of all classifiers based on accuracy measures using 5-fold cross-validation, which gives better accuracy.

Table 6. Confusion Matrix for the two datasets

Model	Actual Class	Cleveland		Cardio Train	
		Predict Class		Predict class	
		Correct	Incorrect	Correct	Incorrect
RF	Yes	85	15	74	26
	No	91	9	65	35
ANN	Yes	85	15	65	35
	No	88	12	64	36
SVM	Yes	81	19	70	30
	No	88	12	70	30

Classification Matrix displays the frequency of correct and incorrect predictions. It compares the actual values in the test dataset with the predicted values in the trained model. Table. 6 shows the results of the Classification Matrix for all the three algorithms. In Cleveland, we diagnosed that 85%, 85% and 91% patients have disease and can correctly classify 74%, 65%, 70% patients in cardio train. Then, we found that 15%, 15% and 19% patients for Cleveland and 26%, 35%, 30% for cardio train do not have cardiovascular disease but the model incorrectly classified that they had disease, it is very dangerous.

In this study, we evaluate the performances of the models using the standard metrics of accuracy, precision, recall. We conducted three different experiments on the different size datasets using three algorithms: Random Forest, Neural Network and support vector machine as given in Table 7 and 8.

Table 7. Performance measures for Cleveland dataset

Model	Cleveland				
	Accuracy (%)	Precision		Recall	
		TP	TN	TP	TN
RF	89	0.86	0.88	0.91	0.85
ANN	85	0.86	0.85	0.88	0.81
SVM	87	0.88	0.85	0.88	0.85

According to Table.7, the True positive rate for Random Forest algorithm (0.89), Artificial Neural Network (0.86) and Support Vector Machine (0.88). Whereas Random Forest is best in True Positive Rate and Artificial Neural Network performed lowest in True Positive Rate. The True Negative Rate for Random Forest (0.91), Artificial Neural Network (0.88) and Support Vector Machine (0.88), we observed that all the three algorithms performed best in True Positive Rate.

For True positive, Support Vector Machine can predict with the lowest rate in Cardio Train dataset and Random Forest and Artificial Neural Network are better performance. By analyzing the result of performance measures for two datasets, we discover the models are best in identifying Negative cases for Cardio Train and best in Positive cases for Cleveland.

Table 8. Performance measures for Cardio Train dataset

Model	Cardio Train				
	Accuracy (%)	Precision		Recall	
		TP	TN	TP	TN
RF	70	0.72	0.68	0.65	0.74
ANN	73	0.72	0.70	0.70	0.72
SVM	65	0.65	0.65	0.64	0.65

In this paper, we used two different datasets, one consists of more than three hundred records and the other contains seventy thousand records. We found by the result with the highest accuracy of 88.52% was achieved by using the Random Forest with 5-fold cross-validation. From the results, we also observed that the Neural Network with 72.99% accuracy was performing better compared to all the other classification algorithms for larger dataset. The results show that Random Forest is best in Cleveland and Artificial Neural Network is performing better amongst all the other algorithms for Cardio Train.

To build the model, it took 2.21, 0.60 and 0.05 seconds time for Cleveland, and 2.16, 73.08 and 303.37 seconds for Cardio Train respectively. We described

the comparisons of performance based on accuracy percentage with bar chart as in Figure 2.

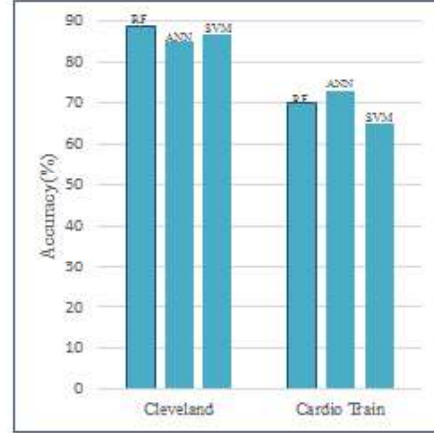


Fig 2. Performance evaluation of classifiers

5. CONCLUSIONS

This research was using two different datasets. First, replace the group median values in all missing values. Further, apply the data transformation technique with a proper feature scaling method. We use three learning algorithms along with k -fold cross-validation with $k=3, 5$ and 10 . This enabled to perform data analysis to obtain the optimal result. Every model can have best performance for specific dataset. The accuracy depends on the nature of dataset. From above study, we observed that the accuracy for the prognosis analysis of various applied data mining classification techniques is highly acceptable and can help the medical professionals in decision making for early diagnosis and avoid the disease. We also observed that Artificial Neural Network has much impressive power. It works best in large dataset and more robust when encountering with missing values. Future work will include trying a study with different data transformations or trying algorithms that we have not tested yet for further analysis of the dataset.

REFERENCES

- [1] M. Dey and S. S. Rautaray, "Study and Analysis of Data mining Algorithms for Healthcare Decision Support System," vol. 5, p. 8, 2014.
- [2] N. Bhatla and K. Jyoti, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques," *Int. J. Eng. Res.*, vol. 1, no. 8, p. 4, 2012.

- [3] C. Gowda and M. B, "Classification Models on Cardiovascular Disease Prediction using Data Mining Techniques," *J. Biodivers. Endanger. Species*, vol. 09, Jan. 2018.
- [4] D. Hammad and H. Monkaresi, "Using Data Mining Techniques to Enhance Heart Disease Diagnosis," 2017.
- [5] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, pp. 1578–1585, Jan. 2018.
- [6] I. Journal, "Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach."
- [7] D. Punjani and K. Atkotiya, "A Comprehensive Study of Various Classification Techniques in Medical Application using Data Mining," *Int. J. Comput. Sci. Eng.*, vol. 6, pp. 1039–1042, Jun. 2018.
- [8] W. Almayyan, "Lymph Diseases Prediction Using Random Forest and Particle Swarm Optimization," *J. Intell. Learn. Syst. Appl.*, vol. 08, pp. 51–62, Jan. 2016.
- [9] J. Kim, J. Lee, and Y. Lee, "Data-Mining-Based Coronary Heart Disease Risk Prediction Model Using Fuzzy Logic and Decision Tree," *Healthc. Inform. Res.*, vol. 21, no. 3, pp. 167–174, Jul. 2015.
- [10] "UCI Machine Learning Repository: Heart Disease Data Set." [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>. [Accessed: 13-Dec-2019].
- [11] "Kaggle: Your Home for Data Science." [Online]. Available: <https://www.kaggle.com/sulianova/cardiovascular-diseasedataset/download/>. [Accessed: 13-Dec-2019].
- [12] J. Padmavathi, "A Comparative study on Breast Cancer Prediction Using RBF and MLP," vol. 2, no. 1, p. 5, 2010.

OPINION MINING FROM TOURIST REVIEWS IN BAGAN

Ohnmar Aung ⁽¹⁾, Nang Win Phyu Phyu Naing ⁽²⁾

⁽¹⁾⁽²⁾University of Computer Studies (Taunggyi), Myanmar

⁽¹⁾aung.ohnmar777@gmail.com, ⁽²⁾winphyuphyunaimg@ucstgi.edu.mm

ABSTRACT

The main tourist destination in Myanmar is Bagan, one of Myanmar's best attraction and one of the richest archaeological sites in South-East Asia. According to the hotel user reviews, hotel managers collected insight information about the hotel conditions that was perceived the hotel user and tourists based on online reviews in TravellingToBagan.com. This study used text mining and aspect-based sentiment analysis approaches to obtain the hotel user opinion. Aspect-based sentiment analysis is the task of identifying fine-grained opinion polarity towards a specific aspect associated with a given target. Recursive Neural Tensor Network (RNTN) algorithm is applied in this study to perform these tasks. RNTN was commonly used for classifying sentiment in sentence level. In this study, the average accuracy 92% is achieved and the classification result on the sentiment of words. Moreover, the result of this study can be used for evaluation of improving the hotelier industry as well as supporting the tourism industry in Bagan, Myanmar.

KEYWORDS: *Recursive Neural Tensor Network; Opinion Mining; Aspect-based Analysis; Text Mining; Text Summarization; Sentiment Analysis*

1. INTRODUCTION

Bagan, temple town is the capital of a powerful ancient kingdom and UNESCO World Heritage Site located in the Mandalay Region of Myanmar. The Bagan dynasty, the first dynasty in the history of Myanmar, flourished from the 11th to 13th centuries and built the foundation of Myanmar culture. Buddhism was introduced throughout the coastal region; people endeavored to build pagodas in order to perform religious practices. Many pagodas were maintained by successive kings. Even today, Bagan is loved by many people both at home and abroad as

the hometown of Myanmar culture and efforts for conservation are continuing.

The number of visiting tourists to Bagan has been increasing for years. This amount contributes nearly 60% of the total visiting tourists to Myanmar. The tourists' arrival was also supported by the adequate infrastructure and hotel facilities there in. There are 155 hotels in Bagan as seen from TravellingToBagan.com. Although Bagan has been known as the icon of world travel, it needs to make continuous improvements quality for tourism industry by devoting individual attention to the hotels as an important part of tourism. Hotels can also be represented as an attraction for tourist because of its services and facilities that had a role for determining whole traveler experiences. To improve the quality of service, hotels must provide the best services for the user that met users' expectation, thus the hotels need user opinion on hotel. Hotel user opinions are really needed to record an overview of hotel for managers about the current condition of hotels in Bagan.

The improvement of communication techniques and website about travelling tourism domain represented by the development of online tourism forum. These forums serve as the primary tools for the traveler to search for travelling information [1]. According to the number of online reviews, travel forums are improving every day because of the increasing number of travelers who are willing to share their travelling experiences. These forums achieved popularity among international travelers a leading source of information in the field of tourism [2]. As a data source of this research, TravellingToBagan.com is one of the travel website that provides freely online reviews of Bagan's hotels. Additionally, other data sources are collected from the various travelling website of Bagan. In this

system, 2000 reviews are collected for all hotels in Bagan.

Considering the large number of text reviews, it needs an effective way to get knowledge from these reviews. Text mining approach became important and useful since it offers solutions for handling unstructured text data from large volume [3]. Li et al. used the tourism domain data source to identify user current preference towards aspects related to services and facilities of hotel with Emerging Pattern Mining (EPM) approach. The other studies were discovering for the classification of words that are considered for sensitive and important factors of hotel user satisfaction level. In these studies, sentiment analysis and text summarization approach are used to obtain summarization from 900 hotel user reviews. This study used text mining approach to process on the large amount of text data automatically.

Moreover, sentiment analysis is utilized to obtain the hotel user opinion deal with the services and facilities of the hotel. Sentiment analysis (also known as opinion mining) is a rule that combine the information retrieval process, text mining and computation to detect the opinion described in the text reviews. There are main tasks in sentiment analysis which include the identification and classification of opinions to positive, negative and neutral [4].

Three research directions are included in sentiment analysis: document level, sentence level and aspect level. Based on Hu and Liu's research, both document and sentences level of sentiment analysis showed the orientation of sentiment for each document or sentence but could not provide the result about the feature that writers like or dislike in detail [5]. The main objective of this study is to obtain the opinion of hotel user in the form of sentiment towards the facility and service of hotel in Bagan using Recursive Neural Tensor Network (RNTN) algorithm.

2. RESEARCH METHODOLOGY

2.1 Data Collection

Hotel user reviews in Bagan were collected from TravellingToBagan.com and several travel website which display large amount of online reviews from all tourists travelling to Bagan. An automated program was used to collect the entire reviews data. All data are extracted from the web page or web scraping by using this program. Web scraping is the retrieving process of semi-structured document from the Internet, usually in the form of web pages with

markup language like HTML or XHTML and then data from a specific page is obtained by analyzing these documents. Python programming language was used to create this study.

2.2. Text Pre-Processing

A series of text pre-processing is required for the downloaded text reviews before they can be analyzed using text mining. There are several stages in the text pre-processing which applied in accordance with the needs of the research. Spelling, Normalization, Filtering, Case Folding, Lemmatization and Sentence Boundary are specific stages of text pre-processing. The stages of text pre-processing are described in Fig 1.

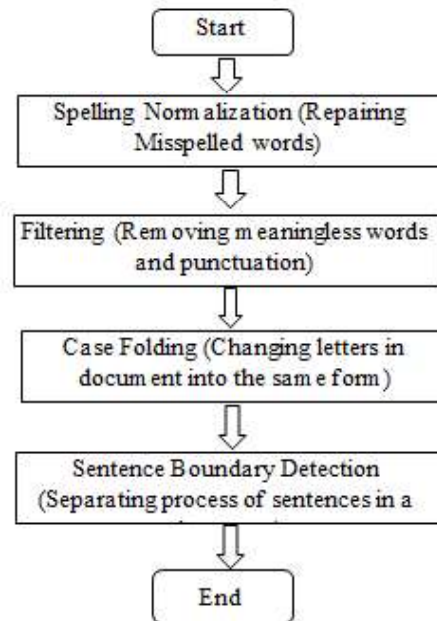


Fig 1. Stages of Text Pre-Processing

2.3. Hotel Aspects Determining

The main objective of this stage is to obtain the hotel's services and facilities aspects which provided in hotel user reviews, still with text mining approach. The four steps are processed in this stage: POS Tagging, Word Indexing, Lemmatization and Taxonomy Formulation. POS stage aims to discover the nouns from the text reviews. Then nouns contained in reviews text are indexed in the word indexing step. Lemmatization phase synchronizes the

words that same meaning into one diction (eg. Meal and Food). The final step of section, Taxonomy Formulation group the nouns into the various categories concerned with services and facilities of hotel. The processes of this stages are shown in Fig 2.

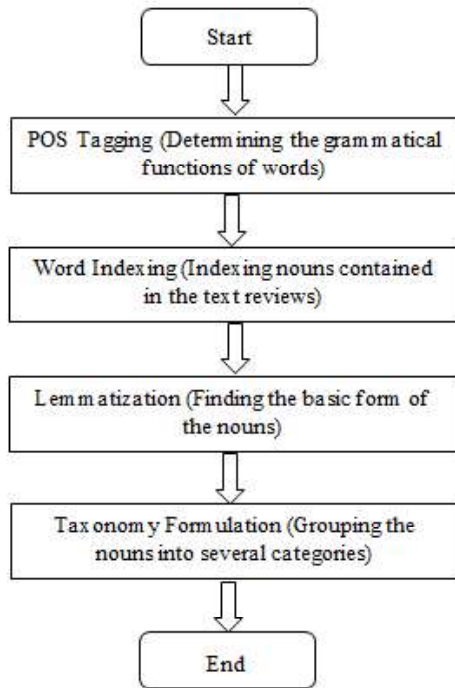


Fig 2. Stages of Determining Aspects

2.4. Aspect Based Sentiment Analysis

This step is to determine the sentiment of every defined aspect in the previous stages using opinion mining approach. The classification of aspect-based sentiment was developed using the help of text mining program. The name of this program is Prameswari v1.4.0 which is created by Stanford University and works based on RNTN model [6]. When text reviews processing, the program will break down the words from a review sentence and then compile these words into a tree, as shown in Fig 3.

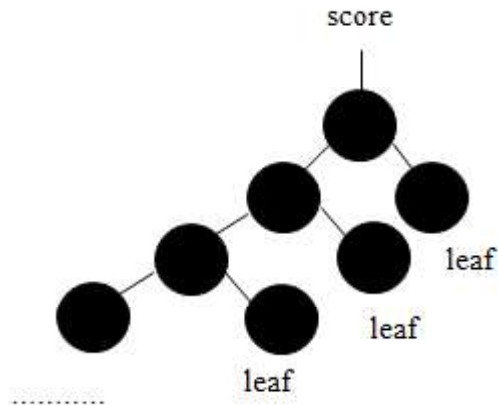


Fig 3. Parse Tree of Components

The database on sentiment Treebank was created sentiment analysis. Positive and negative of the words are defined in this step which is called opinion words inside the database. The parse tree structure detects the opinion words which are positive or negative sentiment. Then the opinion word is moved towards the root, which is the hotel aspect word defined in the previous stage. The sentiment calculation performances are shown in Fig 4.

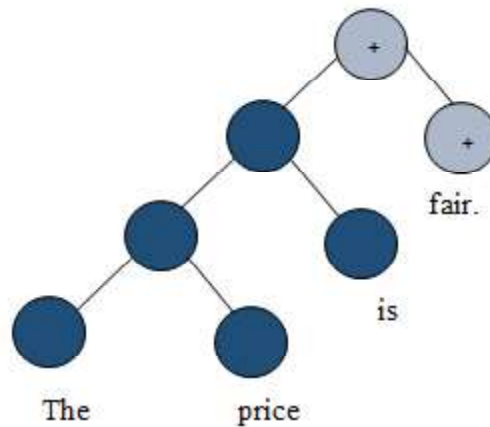


Fig 4. Parse Tree from Sentence "The price is fair."

3. ANALYSIS AND RESULT

In Table 1 and 2, the aspects of hotel services and facilities are shown that were selected and grouped into the eight categories.

Table 1. Facilities and Services Aspect of Hotel in Bagan

Categories			
Food	Accessibil ity	Human resource	Activitie s
Meat	Location	Staff	Sunset
Fruit	Distance	Owner	Swim
Dinner	Main Road	Reception	Rest
Lunch	Airport	Manager	Club
Breakfast		Service	Spa
Restaurant		Driver	
Drink		Chef	
Menu			
Bar			
Café			

Table 2. Facilities and Services Aspect of Hotel in Bagan (Cont'd)

Categories			
Guest's Perspective	Transpo rtation	Room	Environ ment
Experience	Car	Bedroom	Lobby
Value	Boat	Bed	River view
Privacy	Taxi	Air conditioner	Place
Style		Water	Garden
Price		Kitchen	Swimmi ng pool
Expectation		Bathroom	toilet
Security		Wifi	
accommodat ion		towel	
		Fridge	
		balcony	

The sentiment graphs were obtained by calculating all the results of the sentiment analysis. Fig 5 shows the results of aspect-based sentiment analysis. In this figure the blue bar represents for the positive sentiment and the red bar represents for the otherwise. Generally, the hotel users in Bagan

were not satisfied with more than four categories of the hotel services and facilities.

Each category consists of aspects, in which these aspects have an influence on the resulting sentiment. The more occurrence of the aspect in the document, the greater the contribution to the analysis results. The high frequency of the word as an aspect also indicated that these aspects were important by the hotel users.

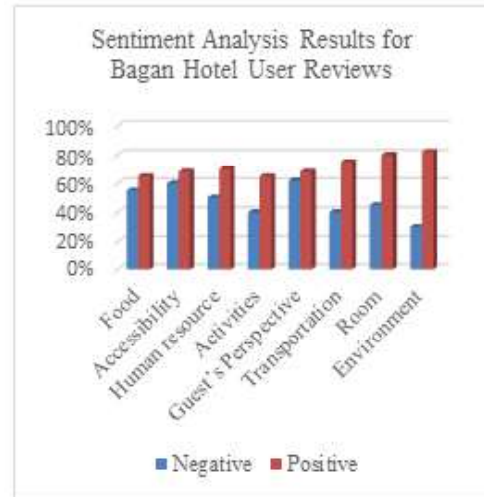


Fig 5. Sentiment Analysis Results of Bagan Hotel

4. DISCUSSION

Based on the computation performed in this study, five of the eight categories related to hotel services and facilities in Bali reaped negative sentiment. Negative sentiment showed dissatisfaction, disappointment, as well as incompatibility of hotel services and facilities with user expectations. Consequently, hotel managers need to prioritize improvements to the categories with the predominance of negative sentiment.

5. CONCLUSION

In Table 3, the confusion matrix of the classification model is shown in which the classifier model used in this study had an average accuracy of 88%, while the average value of F1 Measure is 79%. The positive category has the highest value of F1 Measure that is 92%. This suggests that the model works best in categorizing aspects to positive

sentiment. Meanwhile, the average value obtained for categorizing negative sentiment was 58%.

Table 3. Confusion Matrix of the Classifier Model

Observed	Predicted		Total
	Positive	Negative	
Positive	76	14	90
Negative	0	16	16
Total	76	30	106
Positive	Precision	100%	
	Recall	85%	
	F1	92%	
Negative	Precision	48%	
	Recall	100%	
	F1	58%	
F1 Measure		79%	
Accuracy		88%	

The success of Bagan as a world tourism icon was inseparable from the role of the hotel as the important tourism superstructures. Therefore, the hotels need to be maintained by improving the quality, taking into account the voice of customers from hotel users towards the services and facilities that they perceived.

This study utilized RNTN algorithm, which usually used for sentence-level sentiment analysis in previous studies, yet the result showed that RNTN functioned properly in classifying the sentiment of words or aspects.

The collection of data process through web scraping has not considered the upload time of the hotel user reviews from online, so there are no visible changes from time to time. Future researcher may limit the upload time of the reviews of the hotel users to gain deeper analysis.

REFERENCES

- [1] K. Khan, et al., "Mining opinion components from unstructured reviews: a review," Journal of King Saud University-Computer and Information Sciences, vol. 26, pp. 258-275, 2014.
- [2] N. Ur-Rahman and J. Harding, "Textual data mining for industrial knowledge management and text classification: a business oriented approach," Expert Systems with Applications, vol. 39 (5), pp. 1-11, 2013.
- [3] Z. Xiang, et al., "What can big data and text analytics tell us about hotel guest experience and satisfaction?" International Journal of Hospitality Management, vol. 44, pp. 120-130, 2015.
- [4] G. Li, et al., "Identifying emerging hotel preferences using emerging pattern mining technique," Tourism Management, vol.46, pp. 311-321, 2015..
- [5] R. Socher, et al., "Recursive deep models for semantic compositionality over a sentiment treebank," in Empirical Methods in Natural Language Processing, Palo Alto, 2013.
- [6] B. Pang and L. Lee, "Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales," in The 43rd Annual Meeting on Association for Computational Linguistics, Stroudsburg, Pennsylvania, 2005, pp. 115-124.

COMMUNITY OUTLIER DETECTION IN SOCIAL NETWORKS USING EXTENDED LOCAL OUTLIER FACTORS ALGORITHMS

Hlaing Phyu Phyu Mon⁽¹⁾, War War Myint⁽²⁾, Thin Thin San⁽³⁾

⁽¹⁾⁽²⁾⁽³⁾Faculty of Information Science, University of Computer Studies (Meiktila), Myanmar

⁽¹⁾*drhppmon.is@ucsmtda.edu.mm*

ABSTRACT

Social networks have gained very popularity among internet users since recent year ago. Consequently, analysis on social network community has received many attentions of research scholars. Among different approaches of community detection, this paper focuses on detection of outlier factor of community in social networks, especially Twitter. In graph-based community network, we perform local outlier factors algorithm to detect abnormal or anomalies behaviors of the users within the community group. The proposed system is considered over static network graphs with a dynamic evolution of node attributes (different users or different behaviors of the users) within predefined periods. We then perform experiments in order to prove the efficiency of proposed system. According to the results, the proposed system achieves the promising results compared with traditional KNN or LOF algorithms.

KEYWORDS: *community detection, local outlier factors, social network, anomalies behaviors*

1. INTRODUCTION

The popularity of the Web has been growing over the past years and it significantly affects the way of our daily routines because we become more like to connect with people and share everything we go; we eat and so on. As a result, the community of social networks becomes bigger, and there are various kinds of community in social networks such as online selling community, beauty discussion groups, fashion talks and so forth.

This transition of the Web makes the business people to attract their customers by selling or

advertising their products on social networks. Subsequently, the business applications need to find out what kinds of users should be attracted depending on their products by exploring their connections with others, behaviors and interest.

Regarding users' community detection in social network, there are some challenges to handle. The social network is intuitively very dynamic, network patterns and user's behaviors are changing in unprecedented rates. Therefore, detection algorithms must be able to adapt with this changing environment to produce the accuracy and instant results depending on changing users' tastes to follow the latest trends of social networks.

In order to challenges of outlier community detection in social network, we proposed outlier detection in social network community with these following significant points.

- Extend traditional LOF algorithm from detection only on static data to adaptive detection on dynamic data with K-nearest neighbor algorithms
- Execute group-based attributes calculations for both inter- and intra-group attributes within the time frame
- Observe how the system can dynamically adapt the changes of network structures and internal attributes of the users

The rest of the paper is organized as follows. The related research works are studies in Section 2 to understand underlying theories of community outlier detection of social network. The background theory is glimpsed in Section 3 regarding community

networks and outliers. The proposed system is explained in detail and presented corresponding algorithm to shed light on how system works to detect the outlier's features of social network users.

2. RELATED WORKS

Within the social sciences literature, we found a number of papers focusing on the concept of network change [1]-[2], that try to characterize the evolution of social networks. There are many approaches of outlier's detection algorithm over years to be proven in either statistical data or graph-based structures. Each work has its own efficient and effective effects in this area. After the time of Hawkins [1] definition, many different points of view in outliers were appeared.

A recent cutting-edge technology regarding spatio-temporal data mining highlights the roles of anomaly detection techniques for the dynamic social network's domain [3]-[2]. Due to the complex nature of the data, most existing approaches treat spatial and temporal components of the data independently.

Grouping anomaly detection techniques, previous reviews describe a number of challenges, mainly associated with the problem of defining normal behavior, particularly in the face of evolving systems, or systems where anomalies result from malicious activities [4]-[5]. In particular, the study [4] presented that general solution to anomaly detection remains some significant gaps and it reveals constant demanding to solve specific problems, accommodating the specific requirements of these problems and the specific representation of the underlying systems.

The paper [2] proposed the approach of using similarity measure based on neighborhood overlapping of nodes to organize communities and to identify outliers which cannot be grouped into any of the communities based on Edge Structure. There are three main processes in this paper, firstly finding the intersect value of related node to detect outliers, then use the method of the degree centrality to determine seed nodes and finally use the method of neighborhood overlap based on vertex similarity for detecting the communities.

Referring related works mentioned in the above, this paper searches the outlier user accounts that are dissimilar with the members in some community by means of their preferences, behaviors and connected community, and so forth.

3. PROPOSED OUTLIER DETECTION IN SOCIAL COMMUNITY DETECTION

3.1 Background Theory

Anomalies in social networks are often can be seen as illegal and unwanted behavior between social network users. The increasing explosion of social media and online social systems means that many social networks have become key targets for malicious individuals attempting to illegally profit from, or otherwise cause harm to, the users of these systems.

Many users of online social systems such as Facebook, Weibo, Twitter, etc. are regularly subjected to a barrage of spam and otherwise offensive material. Moreover, the relative anonymity and the unsupervised nature of interaction in many online systems provide a means for sexual predators to engage with young, vulnerable individuals.

Since the perpetrators of these behaviors often display patterns of interaction that are quite different from regular users, they can be identified through the application of anomaly detection techniques. These anomalies or abusive types of structures can be identified by examining a range of network features or through the use of trained classifiers.

Communities may be groups of friendship in social networks, sets of web pages concerning with the same topic and groups of cells with similar functions. While identifying the communities in graphs, nodes which cannot groups with any communities and need not be necessary group, will be identified as outliers.

The social network can be seen as an undirected graph G that contains nodes (users) and edges (communications) between users. It is traditionally symbolized as $G = (V, E)$, where V is the total number of nodes and E represents total number of edges. There might have zero or more edges between any two nodes.

3.3 Proposed System

Community network is as same as a heterogeneous network with K types of objects. A community is a probabilistic collection of similar objects, such that similarity between objects within the community is higher than the similarity between objects in different communities. For example, a research area is a community in a social network. For heterogeneous networks, one is often interested in

identifying heterogeneous communities which contain objects of different types. We will consider connected network to denote the number of communities.

Among anomaly or outlier detection processes, local outlier factors (LOF) algorithm is one of the most popular techniques to find, however, it usually works in static graphs or statistical data. In this paper, to the best of our knowledge, we bring LOF-algorithm in this social outlier detection system by presenting K-LOF algorithm. This algorithm adds K-nearest neighbor (KNN) in the front of LOF algorithm.

There are two main reasons why we need to mix KNN with LOF. The first factor is we would like to introduce the dynamic features of the community group to LOF algorithm. To do so, we firstly group the community with their crisp values without loading to LOF algorithm, that takes a long time in first step of distance calculation (with Euclidean or Manhattan distance). After getting first round group, we use only LOF algorithms which filter the social account members which act abnormally against with the others in the same community. However, whenever we detect the network changes within periodic time interval, we use KNN for every first group with fresh members. Another reason of taking into account KNN with LOF is due to highly dynamic structure of social network. This is to say, instead of individuals matching each other in the whole network, we can just need to explore only inter- or intra-group-based calculation that can consequently reduce the computational time and increase the performance as well.

The proposed KNN based LOF an algorithm is presented as follows. The internal process of proposed algorithm is referred to traditional LOF algorithm. The first time grouping the members depending on the crisp values of the nodes are performed as also traditional KNN algorithm. The fact beyond mixture of KNN and LOF is that, we observe how nodes and attributes are changed within the time frame.

The detailed process of our proposed algorithm can be seen as follows.

Algorithm: Extended LOF Outlier Detection in Dynamic SN

Input : Social Network Dataset [6]

Output : Name of Outlier Nodes

Algorithm:

1. DataSet inquisition with nodes (N) structures (N_s) and attribute (N_A)
2. KNN Algorithm for first group partition within minimum inter-group distance
3. if network changes or nodes' behavior changes
4. Go to Step 2
5. Else
6. LOF algorithms
7. Check network and nodes' status and attributes
8. if changes occur
9. Go to Step 2
10. Else
11. Froze the detection and Go to Step 8
12. End
13. End

The abovementioned algorithm is finding the outlier nodes of the input network graph. By observing the network's changing patterns, the algorithm decides which algorithms should be called, that is to say, KNN or LOF algorithms to reflect the network changes. To be best of our experience, the longer time frame makes more calculation process because the social network's changes are unpredictable in unprecedented rates.

<p>Algorithm: LOF algorithm</p> <p>Let D be a data set, p, q, and o are some nodes in D, and k be any positive integer.</p> <p>1. Computing intra-group distance:</p> <p>Calculate distance between elements of intra-groups distance by splitting the dataset into k^{th} nearest neighbor groups randomly but this must satisfy the following conditions.</p> <p>(a) For at least k objects $o \in D \setminus \{p\}$ it holds that $d(p, o) \leq d(p, o)$ and</p> <p>(b) For at most $k-1$ objects $o \in D \setminus \{p\}$ it holds that $d(p, o) < d(p, o) \cdot k\text{-distance}(p)$</p> <p>2. Finding $(k\text{-inter-group distance of } p)$: The k-distance neighborhood of p contains every object whose distance for p is not greater than the k-distance. $N_k\text{-distance}(p) = \{q \in D \setminus \{p\} \mid d(p, q) \leq k\text{-distance}(p)\}$</p> <p>3. Computing the (reachability distance of p wrt object o): For each data point q in the k-distance neighborhood of p, define the reachability distance of p with respect to q as $\max\{k\text{-distance}(q), d(p, q)\}$, where $d(p, q)$ is the distance between p and q.</p> <p>4. Computing (the local reachability density of p): The local reachability density of an object p is the inverse of the average reachability distance from the k-nearest neighbors of p as shown in eq (1):</p> $lrd_k(p) = \left[\frac{\sum_{o \in N_k(p)} \text{reach-dist}_k(p, o)}{ N_k(p) } \right]^{-1}$ $lrd_k(p) = \frac{\sum_{o \in N_k(p)} \text{reach-dist}_k(p, o)}{ N_k(p) }$ <p>5. Computing the local outlier factor of p: The local outlier factor is a ratio that determines whether or not an object is an outlier with respect to its neighborhood. $LOF_k(p)$ is the average of the ratios of the local reachability density of p and that of p's k-nearest neighbors.</p> $LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{ N_k(p) }$
--

Inspired to traditional LOF algorithm [7], we perform LOF algorithm to find outlier nodes from social networks. The readers are referred to study [8] for detailed calculation of LOF algorithms.

4. EXPERIMENT DATASETS AND RESULTS

In this paper, a synthetic dynamic network used in social network analysis, Zachary Karate Club Dataset [6] is used. In this Dataset statistics, nodes represent the number of friends while edges are a sign of friendship between two friends. There are 128 member nodes and 258 edges. A brief illustration of a network is shown in Figure 1.

Regarding experimental results, we evaluate the results in Python programming by inputting the nodes attributes and their links with imagination of they are formed as graph.

To evaluate the effectiveness of the proposed system, we evaluate how well the system save the computation time and can find the accurate outlier factors and in different observing time frames. According to our result, not more than 3 seconds for predefined time frames gets acceptable computational time 4.5 seconds in minimum.

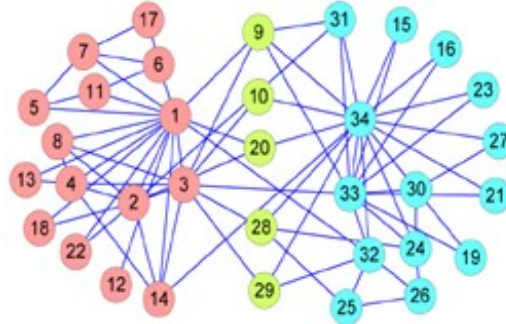


Fig 1. Example social network graph of dataset [6]

Table 1. Computational Time Evaluation with different time frame

Number of Nodes	Time Frame	Computation Time
10	0.1s-3.0s	4.5s-5.6s
40	3.1s-5.0s	6.s-8.8s
70	5.1s-7.0s	10.2s-24.0s
90	7.1s-9.0s	56.7s-1.0min
120	9.1s-12.0s	1.2 min – 1.5 min

Another interesting factor of proposed system's performance, it could find accurate outlier nodes compared to using LOF algorithm alone as shown in following table Table 2. The accuracy value is calculated depending on the number of accurate

Table 2. Accuracy Result Evaluation

Number of Nodes	KNN+LOF algorithm	LOF algorithm
40	89.5%	76.4%
70	82.3%	67.7%
120	76.4%	49.3%

We alternatively measure how our proposed algorithm saves more time in comparison with LOF algorithm with the same setting in previous evaluations. We get the following satisfactory results in following Table 3. What we see in Table 3, LOF algorithm is nowhere near as less computation time as our proposed KNN+LOF algorithm in different number of nodes.

Table 3. Computation Time Result Evaluation

Number of Nodes	KNN+LOF algorithm	LOF algorithm
40	4.5s-5.6s	23.0s-46.s
70	5.1s-7.0s	45.0s-2.4min
120	9.1s-12.0s	4.5min-7.8min

5. CONCLUSIONS

The evolution of internet and popularity of social networks takes some parts of our daily lives. Instead of connecting to real friends, shopping at the market, reading the newspapers, the people become more like to do everything in online social networks as one stop service such as reading news from social network posts, checking the friend's updated information, prices of online items, etc. Therefore, this paper, as part of community detection in social network, presented an anomaly detection system with local outlier factors so as to explore some of user's accounts are far interesting in some communities while they are more interested in other communities. According to the numerical calculation results, our system could detect outlier accounts, which are assumed to be out of some communities having a mere communion lines between the communities.

As future work, we plan to extend outlier detection system to cope with dynamic synthetic networks and real social networks. The aim is to observe rapidly changing patterns of social networks and users' behaviors by using dynamic outlier features [9] extractions[10] and [11]detections[12]-[13].

REFERENCES

- [1] D. Hawkins, *Identification of Outliers*. Springer Netherlands, 1980.
- [2] H. N. Win and K. T. Lynn, "Community Detection in Social Network with Outlier Recognition," *Adv. Sci. Technol. Eng. Syst. J.*, vol. 3, no. 2, pp. 21–27, Mar. 2018.
- [3] V. Miz, B. Ricaud, K. Benzi, and P. Vanderghenst, "Anomaly detection in the dynamics of web and social networks," *ArXiv190109688 Cs*, Jan. 2019.
- [4] V. Chandola, "Anomaly Detection for Symbolic Sequences and Time Series Data," p. 154.
- [5] V. Hodge, "A Survey of Outlier Detection Methodologies," *Artif. Intell. Rev.*, vol. 22, pp. 85–126, Oct. 2004.
- [6] W. W. Zachary, "An Information Flow Model for Conflict and Fission in Small Groups," *J. Anthropol. Res.*, vol. 33, no. 4, pp. 452–473, 1977.
- [7] M. Alshawabkeh, B. Jang, and D. Kaeli, "Accelerating the local outlier factor algorithm on a GPU for intrusion detection systems," presented at the International Conference on Architectural Support for Programming Languages and Operating Systems - ASPLOS, 2010, pp. 104–110.
- [8] "LOF(Local Outlier Factor) Example," p. 12.
- [9] D. Savage, X. Zhang, X. Yu, P. Chou, and Q. Wang, "Anomaly detection in online social networks," *ArXiv160800301 Phys.*, Jul. 2016.
- [10] "Enron Email Time-Series Network | Zenodo." [Online]. Available: <https://zenodo.org/record/1342353#Xfm4HBsVSM8>. [Accessed: 13-Dec-2019].
- [11] M. Mongiovì, P. Bogdanov, R. Ranca, E. E. Papalexakis, C. Faloutsos, and A. K. Singh, "NetSpot: Spotting Significant Anomalous

Regions on Dynamic Networks - Supplemental material,” p. 4.

- [12] C.-T. Lu, Y. Kou, J. Zhao, and L. Chen, “Detecting and tracking regional outliers in meteorological data,” *Inf. Sci.*, vol. 177, no. 7, pp. 1609–1632, Apr. 2007.
- [13] R. McEliece, E. Posner, E. Rodemich, and S. Venkatesh, “The capacity of the Hopfield associative memory,” *IEEE Trans. Inf. Theory*, vol. 33, no. 4, pp. 461–482, Jul. 1987.

APPLIED FP-GROWTH FOR TRADITIONAL HERBAL REMEDIES FOR HEALTH CARE

Hnin Yu Hlaing

University of Computer Studies (Meiktila), Myanmar

hninyuhlaing489@gmail.com

ABSTRACT

Over 1000 years traditional medicine has been used continuously in Myanmar. It is very useful, valuable and powerful. Our antecedents lived healthily, wealthily and happily by only using it. Today, an evolutionary era, Myanmar Traditional Medicine should not be late. Because traditional medicine is formed by combining parts that are gained from trees, herbs and animals and not chemical products, it's free from danger. Treatments with medicinal plants are considered very safe as there is no side effect. These remedies are fit with nature, which is the biggest advantage. This study focuses on useful traditional medicinal plants or herbs used in producing traditional medicine. In this study, Frequent Pattern (FP) Growth algorithm is applied to be fast and be efficient. Frequent Pattern (FP) Growth is one of frequent pattern mining comprises a set of techniques able to uncover hidden patterns from the data. The preprocessed database is mined to extract frequent patterns related to symptoms by using Frequent Pattern (FP) Growth algorithm.

KEYWORDS: Keywords: Frequent Pattern (FP) Growth, Traditional Medicinal Plants.

1. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. It is the process of sorting through large amounts of data and picking out relevant information through the use of certain sophisticated algorithms. Data mining is a multidisciplinary field drawing works from statistics, database technology, artificial intelligence, pattern recognition, machine learning, information theory, knowledge acquisition, information retrieval, high-performance computing and data visualization [10].

The rapid growth and integration of databases provides scientists, engineers, and business people with a vast new resource that can be analyzed to make scientific discoveries, optimize industrial systems, and uncover financially valuable patterns. This takes these large data analysis projects, researchers and practitioners have adopted established algorithms from statistics, machine learning, neural networks, and databases and have also developed new methods targeted at large data mining problems [4].

Data mining is one component of the exciting area of machine learning and adaptable computation. The goal of building computer systems that can adapt to their environments and learn from their experience has attracted researchers from many fields, including computer science, engineering, mathematics, physics, neuroscience, and cognitive science. Out of this research has come a wide variety of learning techniques that have the potential to transform many scientific and industrial fields. Several research communities have converged on a common set of issues surrounding supervised, unsupervised, and reinforcement learning problems [4]. Data Mining is the process of discovering new correlations, patterns, and trends by digging into large amounts of data stored in warehouses. It is related to the subareas of artificial intelligence called knowledge discovery and machine learning. Data mining can also be defined as the process of extracting knowledge hidden from large volumes of raw data i.e. the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [10].

In this study, symptoms occurred in people are taken into account from the Traditional Medicine Hospital, Mandalay, and plants' data are collected from "Collection of Commonly Used Herbal Plants". The accurate information of transaction data of

symptoms and herbal plants is to be found in about 73% with 200 symptom transactions.

2. PROBLEM DEFINITION

Let $I = \{i_1, i_2, \dots, i_m\}$ be the set of items and D be the transactional data source which contains the set of transactions. Each transaction T is a set of items such that $T \subseteq I$ and is associated with an identifier called TID. An association rule is an implication of the form $X \Rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$ and $X \cap Y = \emptyset$ [3]. In general, every association rule must satisfy two user specified constraints, one is support and the other is confidence. The support of a rule $X \Rightarrow Y$ is defined as the fraction of transactions that contain $X \cup Y$, while the confidence is defined as the ratio of $\text{support}(X \cup Y) / \text{support}(X)$. An itemset is frequent if its support satisfies at least the minimum support, otherwise it is said to be infrequent. A frequent itemset is a Maximal Frequent itemset if it is a frequent set and no superset of this is a frequent set. The paper aims to find the Maximal Frequent itemset from a huge data source.

3. RELATED WORKS

The solution is the frequent-pattern growth, or simply FP-growth, which mines the complete set of frequent itemsets without candidate generation. This method adopts a divide-and-conquer strategy as follows: first it compresses the database representing frequent items into frequent-pattern tree, or FP-tree, which retains the itemset association information. It then divides the compressed database into set of conditional databases; each associated with one frequent item or pattern fragment and mines each such database separately. FP-tree is created from the root and labels it null.

The FP-growth algorithm: (mine frequent itemsets using an FP-tree by pattern fragment growth):

Input:

1. D , a transaction database.
2. min_sup , the minimum support count threshold.

Output: the complete set of frequent patterns.

Method:

- (1) The FP-tree is constructed.

- (2) The FP-tree is mined by calling FP-growth (FP_tree, null):

Procedure FP_growth (Tree, α)

```

if Tree contains a single path P then
    for each combination (denoted as  $\beta$ ) of the
    nodes in the path P
        generate pattern  $\beta \cup \alpha$  with
        support_count = minimum support
        count of nodes in  $\beta$ ;
    else for each  $a_i$  in the header of Tree {
        generate pattern  $\beta = a_i \cup \alpha$  with
        support_count =  $a_i.\text{support\_count}$ 
        construct  $\beta$ 's conditional pattern base and then  $\beta$ 's
        conditional FP_tree Tree $\beta$ ;
        if Tree $\beta \neq \emptyset$  then
            callFP_growth(Tree $\beta$ ,  $\beta$ ); }
    
```

Based on the above algorithm, association rules can be generated as follows:

1. For each frequent itemset l , generate all nonempty subsets of l .
2. For every nonempty subset s of l , output the rule " $s \Rightarrow (l-s)$ " if $\text{support_count}(l) / \text{support_count}(s) \geq \text{min_conf}$, where min_conf is the minimum confidence threshold.

Support and confidence are defined as:

$\text{Support}(A \rightarrow B) = P(A * B)$

$\text{Confidence}(A \rightarrow B) = P(A/B)$ [6].

Symptom-set Implementation

Let D be a database of transaction. Each transaction consists of a transaction identifier and a set of symptoms $\{S_1, S_2, S_3, S_4, S_5, \dots, S_{10}, S_{11}, \dots, S_{100}, \dots\}$ selected from the universe symptom of all possible descriptive symptoms of diseases. Table 1 shows the transaction data of symptoms.

Table 3.1 Transaction Symptoms and Their Frequency

SymptomId	Symptoms	Support count
S018	ခါးနာကျင်ခြင်း	11
S048	ကိုယ်လက်အင်္ဂါကိုက်ခဲခြင်း	10
S047	ကိုယ်လက်လေးလံခြင်း	10
S216	ကိုယ်ပူခြင်း	9
S199	ခေါင်းဖူးခြင်း	8
S147	ဈာန်ခြင်း	8
S014	ခေါင်းကိုက်ခြင်း	8
.	.	.
S204	မျက်သားဖြူခြင်း	1
S252	လည်ချောင်းယားခြင်း	2

There are transactions of symptoms of diseases in this database. In the process of mining frequent itemsets. The support count of an itemset is the length of the TID set of the itemset. Suppose that the minimum transaction support count is 2.

Table 3.2 Symptoms-Frequency Table with Minimum Support Count 2

SymptomId	Symptoms	Support count
S018	ခါးနာကျင်ခြင်း	11
S185	ကိုယ်လက်အင်္ဂါကိုက်ခဲခြင်း	10
S047	ကိုယ်လက်လေးလံခြင်း	10
S216	ကိုယ်ပူခြင်း	9
S199	မျက်သားညှပ်ခြင်း	8
S147	ဈာန်ခြင်း	8
S252	ခေါင်းကိုက်ခြင်း	8
.	.	.
S253	ခံတွင်းအနံ့ခြင်း	2

In table 3.2, frequencies for each symptom are included after pruning with minimum support count 2.

FP-Growth extracts frequent symptom-sets from the FP-tree by using Bottom-up algorithm - from the leaves towards the root. It uses divide and conquer approach.

Divide and conquer:

- Compress the database (build FP-tree) to retain item-sets association information.
- Divides the compressed database into a set of conditional database.

Once the frequent itemsets from transaction in the database have been found, it is straightforward to generate association rules from them.

Table 3.3 Conditional Pattern-base for Frequent-Symptom Sets

SymptomID	Conditional Pattern Base
S185	{S047, S186, S184: 2}
S131	{{S018, S013, S094, S199: 2}}
S261	{{S109, S111, S114: 1}, {S020, S129, S111, S114: 1}}
S160	{{S047, S186, S184, S185: 1}, {S047, S186, S184: 1}}
S147	{{S018, S146, S148, S255, S221: 1}, {S018, S146, S148, S255: 1}, {S018, S146, S148: 1}}
S252	{{S253, S251: 1}, {S253: 1}}
S253	{{S003, S090, S212, S042: 1}, {S047: 1}, {S030, S032, S049, S042: 1}}
.	.
.	.

This can be done using the following equation for the confidence, can be shown for completeness.

$$\text{confidence}(A \cup B) = \frac{\text{support_count}(A \cup B)}{\text{support}(A)}$$

Table 3.4 Support and Confidence of Symptom-set

Subser (A)	Subser (B)	Sup(AUB) /Sup(A)	Confidence(%)
S185	S184	2/2	100
S131	S199	2/2	100
S194	S048, S216	2/2	100
S215	S048, 216, S003	2/2	100
S041	S048, 216, S003	2/3	66.67
S261	S184	2/2	100
S090	S048, 216, S003	2/4	50
S147	S148	3/4	75

4. DESIGN AND IMPLEMENTATION

This system implemented for retrieving information of Myanmar traditional medicinal plants by FP-Growth algorithm works the following procedure. In implementing this system, a database of herbal plants on the traditional medicine is used. The database is used to send out learning. The database describes attributes of the herbalplants, such as their usage, botanical name, growing place, feature, effect and family of eachplant.

In this system, fp-growth algorithm is used together with divide and conquers approach.

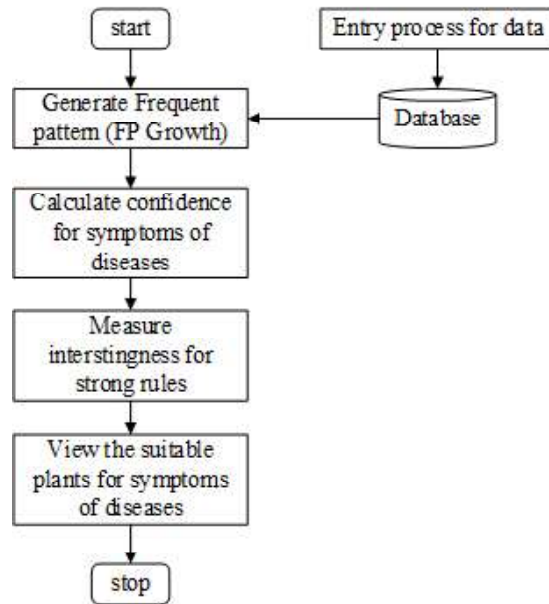


Figure 4.1 System Flow Diagram

This system focuses on the association rule mining of data mining according to the related data of Myanmar traditional medicinal plants. Firstly, data about the information of Myanmar herbal plants is stored into the herbal database. In one transaction, symptoms of disease occurred in patients are contained. By applying Frequent-Pattern growth algorithm of association rule mining, frequency of diseases of same symptoms that can be cured by the herbal plants from the transactions of symptoms are numbered with the minimum support count defined by the user-specified minimum support count and sorted by descending frequency order. There are many oriented-programming languages. Among them, this system is implemented by using Microsoft Visual Studio 2008 and Microsoft Office Access Database. This system works as follows.

- Entry process for data of traditional medicinal plants, diseases, symptoms, symptoms occurred in the disease, symptoms that can be cured by the plants.
- FP-tree is constructed with symptoms by the FP-growth algorithm.
- Generate frequent pattern of symptoms.

- Calculate confidence for symptoms of diseases in patients.
- Search the suitable plants for the symptoms of diseases.
- Display the related plant's information.

When the user starts the program, the form named "Start Application" will appear. In this form, there are "System View" and "User View". When the user chooses the "User View", the "User View" form is going to appear. In this form, one symptom can be chosen by the user who wants symptoms can be occurred with this symptom together. In the "System View", "File", "FP-Growth", and "View" menu are involved.

4.1 ENTRY PROCESS

Symptoms that are commonly occurred in disease of patients are shown. These particular symptoms are 255 and patient's records are totally 414 transactions. User can fill the symptoms of disease found in patients with their PatientId, Disease and Symptoms. When the user wants to view lists of data of patient's symptoms, patient data list..." stored in database can be retrieved.

When the user wants to add or view what the symptoms can be cured by which plant, the new data about herbal plant can be added in the herbal table in the database. Plants and symptoms that can be cured by these plants are related to each other.

In all tables, the data of plant, symptoms, diseases and patient's symptom-records can be added or deleted respectively if the user wants to add or delete these data. When the user does not want to run to continue the system, it can be closed.

4.2. DISPLAY THE RELATED PLANT'S INFORMATION

Then, plant table in the database can be retrieved to view their information as knowledge containing their description, usage, location and taste. There are 120 kinds of plants taken from "Collection of Commonly Used Herbal Plants, Ministry Of Health, Department of Traditional Medicine, January 2003".

4.3 RULE INTERESTINGNESS MEASURE BY CORRELATION ANALYSIS

The accurate information of transaction data of symptoms above the 75% of confidence is to be

found in about 73% with 414 symptom transactions by support count of 2 and 68% by support count of 3. A correlation measure can be used to augment the support-confidence framework for association rules. There are various correlations that measure to determine which would be good for mining large data sets. Lift is a simple correlation measure that is given as follows. The occurrence of itemset A is independent of the occurrence of itemset B if $P(A \cup B) = P(A) P(B)$; otherwise, itemsets A and B are dependent and correlated as events. This definition can easily be extended to more than two itemsets. If the resulting value shown in table 4.3.1 is less than 1, then the occurrence of A is negatively correlated with the occurrence of B. If the resulting value of a rule is greater than 1, then A and B are positively correlated, that is meaning that the occurrence of one implies the occurrence of the other. The lift between the occurrence of A and B can be measured by computing $Lift(A, B) = P(A \cup B) / P(A)P(B)$.

There are rule interestingness measures for above strong rules by lift as correlation analysis.

Table 4.3.1 Rule Interestingness Measurement

No of Transactions	With Support counts and Confidence above 75%	No of Rules (lift value >1)	No of Rules (lift value <1)
414	2	32	5
414	3	15	4
300	2	29	5
300	3	13	4
250	2	32	8
250	3	8	5

The values of above calculated lift that are greater than 1, these rules are positively correlated, meaning that the occurrence of one implies the occurrence of the other. And the others are negatively correlated.

5. CONCLUSION

The association rules play a major role in many data mining applications, trying to find interesting patterns in data bases. Apriori is the simplest algorithm which is used for mining of frequent patterns from the transaction database. The main drawback of Apriori algorithm is that the candidate

set generation is costly, especially if a large number of patterns and/or long patterns exist. Apriori algorithm uses large item set property, easy to implement, but it repeatedly scan the database. The frequent pattern tree (FP-tree) is used for storing compressed, crucial information about frequent patterns, and developed a pattern growth method, FP-growth, for efficient mining of frequent patterns in large databases. Fp algorithm uses divide and conquer approach and it is more efficient than apriori algorithm and also takes lesser time and gives better performance.

Symptoms of diseases are built by scanning the patient's symptom records in the Herbal database with FP-Growth Algorithm. Furthermore, other data related to this system such as other herbal information, patient's symptom transaction, collection of diseases can be stored. The association rules play a major role in many data mining application, trying to find interesting patterns in data bases. However, it is sometimes unrealistic to construct a main memory-based FP-tree.

Fp-Growth algorithm decomposes patient's symptom records according to the frequent patterns obtained so far. It leads to focused search of smaller databases and compresses database called FP-tree structure.

6. LIMITATION AND FURTHER EXTENSION

This system is implemented only for the related information of Myanmar traditional medicinal plants. The numbers of herbal plants are 120. The user can add other herbal plants which are related to diseases and symptoms that can be cured by these plants. This software is implemented by using C#.NET programming language.

In future, this system can be extended by adding the number of herbal plants and can be improved by other association rule mining algorithms or frequent itemset mining algorithms.

ACKNOWLEDMENT

I would like to take this opportunity to express my sincere thanks to all who gave me a lot of valuable advice and information. I am also grateful to all respectable people who directly or indirectly helped towards the success.

Thanks are also extended especially to my colleagues and all my friends for their encouragement and cooperative help for the completion of this paper.

REFERENCES

- [1] Bhavesh V. Berani, Dr.ChiragThaker, Assistant Professors, "FP Growth Algorithm for finding patterns in Semantic Web", Shantilal shah engineering college, Bhavnagar.
- [2] Charu C. Aggarwal, Jiawei Han Editors, "Frequent Pattern Mining".
- [3] C.I. Ezeife and Dan Zhang, "TidFP: Mining Frequent Patterns in Different Databases with Transaction ID", School of Computer Science, University of Windsor, Windsor, Ontario, Canada N9B 3P4 zhang3d@uwindsor.ca, <http://www.cs.uwindsor.ca/~cezeife>.
- [4] David Hand, HeikkiMannila and Padhraic Smyth, "Principles of Data Mining" ISBN: 026208290xThe MIT Press © 2001 (546 pages)A comprehensive, highly technical look at the math and science behindextracting useful information from large databases.
- [5] Jiawei Han and MichelineKamber, "Data Mining Concepts and Technique, Second Edition".
- [6] Jiawei Han and MichelineKamber, "Frequent Item set Mining Methods, Data Mining– Concepts and Techniques", Chapter 5.2, Julianna KatalinSipos.
- [7] Jiawei Han hanj@cs.uiuc.edu, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach", University of Illinois at Urbana-Champaign.
- [8] Lai Lai Win, Khin Myat Myat Moe, Computer University (Magway), "Mining Association Rules by using Vertical Data Format", lailaiwin.myn@gmail.com.
- [9] Springer, "Principle of Data Mining, Undergraduate Topics in Computer Science".
- [10] "Application of FP Tree Growth Algorithm in Text Mining", Project Report Submitted In Partial Fulfillment Of The Requirements for the Degree Of Master of Computer Application, Department of Computer Science and Engineering, Faculty of Engineering and Technology Jadavpur University, Kolkata-700032, India.

- [11] “A Frequent Pattern Mining Algorithm Based on FP-growth without Generating Tree”,
Universiti Putra Malaysia, Serdang,
MALAYSIA, Itohidi.h@gmail.com,
2hamidah@fsktm.upm.edu.my.
- [12] “FP-Tree Based Algorithms Analysis,
FPGrowth, COFI-Tree”, Thapar University,
Patiala, India, bharatgupta35@gmail.com.
- [13] လက်တွေ့အသုံးချဆေးဖက်အပင်များပေါင်းချုပ်” ,
Collection of Commonly Used Herbal Plants,
MinistryOf Health, Department of Traditional
Medicine, January 2003.
- [14] "မြန်မာ့တိုင်းရင်းဆေးပညာ၊ စံညွှန်းကုထုံး၊ ကျန်းမာ .
ရေး ဝန်ကြီးဌာန၊ တိုင်းရင်းဆေးပညာဦးစီးဌာန"
- [15] “ကလေးရောဂါပညာ, စတုတ္ထနှစ်(B.M.T.M)”.

A STUDY ON FEATURE ANALYSIS USING SPARSE REPRESENTATION FOR MUSIC CLASSIFICATION

May Thu Myint ⁽¹⁾, Phyu Phyu Khaing ⁽²⁾

⁽¹⁾University of Computer Studies (Hpa-an), Myanmar

⁽²⁾ Myanmar Institute of Information Technology, Mandalay, Myanmar

⁽¹⁾*mthumyint@gmail.com*

ABSTRACT

This paper presents the first attempt to classify for Myanmar ethnic music by using sparse representation classification method to define the class label according to their ethnic traditional style. In this system, the only five myanmar ethnic groups are considered such as Kachin, Kayin, Mon, Shan, Rakhine. The classification system describe the better accuracy by analysing the temporal features and spectral features, the best outcome by calculating all the results based on Sparse Representation classifier in compared with K Nearest Neighbors classifier. Therefore, this audio classification achieved the results by evaluating feature combination and the best feature combination by using SRC and KNN classifier. With all the features of all ethnic classes, the overall outcome of the SRC 64%, which is better than 54% of the overall KNN accuracy. All features (114) combination give the best results of 82% for Kayin ethnic songs than other ethnic songs. The feature combination of MFCC(std, deltamean) are tested on all of five ethnic classes which is the best classification results of 75.00% accuracy from SRC classifier than the classification results of 51.33% from KNN classifier.

KEYWORDS: *myanmar ethnic music, sparse representation classifier, k nearest neighbors, temporal feature, spectral feature*

1. INTRODUCTION

Culture has a great impact on music in term of creation, performance and interpretation. People with a certain cultural background often like a particular cultural music style. Therefore cultural style information is very helpful for listening, searching

and recommending music. It can be see that a piece of cultural style music has similar features, similar attributes such as the tuning system, musical scale and instrumentation. On the basis of this observation, machine learning techniques can be used to classify music signals according to the cultural style of music. Automatic music classification is a high-level task that refers to the process of automatically assigning class labels or genre for various tasks, including, but not limited to categorization, organization and browsing. Most audio classification systems combine two processing stages: feature extraction and classification expressed in Liu and others [1]. In this paper, the audio files contained music of different singer, different ethnic song classes (Kachin, Kayin, Mon, Shan, Yakhine etc.) and several other kinds of data such as sounds produced by others cultural people. With these music collections, it would be possible to classify the correct song of the system. In order to provide high quality results and to discriminate more audio music from the huge amount of music collection in the music recognition tasks, the selected features must reflect basic information about the music. In this proposed system, various signal features are used for this purpose including 114 features (zcr, centroid, skew, kurtosis, bandwidth, MFCC mean, MFCC std, MFCC deltamean, MFCC delta- std) have been used for feature extraction in intro songs experiment. Jothilakshmi et.al [2] stated these studies use several different classification strategies, including multivariate Gaussian models, Gaussian mixture models, self-organizing maps, neural networks, k-nearest neighbor schemes and hidden Markov models. The system can be described as a performance evaluation of the proposed system. In addition, evaluation methods are used in a comparative way to measure whether certain changes

lead to an improvement in system performance. The sparse representation classifier evaluate songs by testing several important parameters of them. Many audio classification problems involve high dimensional, noisy data. Sparse representation by l-norm minimization is robust to noise and even incomplete measurements. Now, there are some methods to be compared to the proposed ethnic music classification system, K- Nearest Neighbors classifier (well-known classifier) has to be selected in compared with the sparse representation classifier.

2. EXPERIMENT

2.1 Experiment apparatus

Classification of music based on cultural style can help for music analysis and used in search and recommendation systems. The human perception of the sounds is a way of creating music label during the identification process and as long as they are heard. Technically, people also rely on the features derived from the sound they hear to identify the sound. Deshmukh.et al [3] initiated there are various methods for music classification to classify the ethnic songs and folk songs. Many different types of audio feature extraction have been proposed of the tasks of folk song classification. The system are firstly preprocessed the audio data, from these audio samples are extracted the nine features (zcr, centroid, shew, centroid kurtosis bandwidth, MFCC mean, MFCC std, MFCC delta-mean and MFCC delta std) which may be calculated based on the basic samples after preprocessing. All of the features are 114 features in which 1-57 features are mean value and 58-114 features are standard deviation value of nine features. Finally, after feature extraction these extracted features are classified by using SRC in comparing with KNN that is to define each ethnic class label.

2.2 Audio Dataset

The audio dataset contained music of different singer, different ethnic song classes (Kachin, Kayin, Mon, Shan, Yakhine) and several other kinds of data such as sounds produced by others cultural people. The dataset contains 250 songs from the popular ethnic music songs categorized as 50 audio recordings of each ethnic classes respectively. The input audio signal (wav file) is resampled at 44100 hz with 16 bits per sample. These songs are from MRTV radio station. Each music piece lasts about 3 to 5 minutes in length but the input music pieces nearly last 60 seconds segment. In all experiments, the intro pieces of music is used for evaluation of the system.

2.3 Experiment Setup

In all experiment, there are three main components, preprocessing, feature extraction and classification. After converting the input audio from stereo to mono channel and is divided into frames with 100ms, these audio samples frame was taken 50% overlapping frame between the successive frames. The major features of audio sample are extracted from the overlapped frame, and then the extracted features are classified by SRC. The system was implemented by matlab programming language.

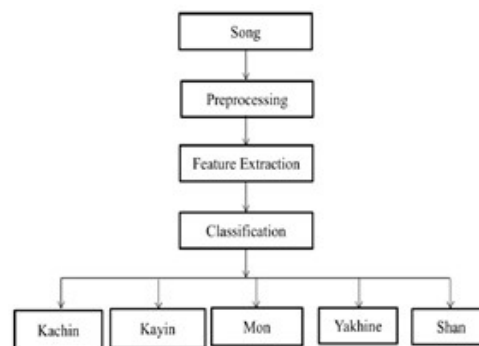


Fig 1. Architecture of myanmar ethnic music classification

From all of the five ethnic classes, in experiment I, the intro pieces of music used for evaluating two feature combination, experiment II, this ethnic songs dataset is used for analysing three feature combination, also in experiment III for testing four feature combination and the results of experiment IV was achieved by combining five features in comparison SRC with KNN. In experiment V, this experiment is to achieve the best feature combination All songs were trained on the different sets of feature vectors for each consists of nine major features as mentioned in section (2.3).

2.4 Feature Used

The classification system are mainly used spectral and temporal features. At the highest level, music is considered to have four key properties : the melody, or sequence of pitches; the harmony, or the combinations of pitches; the rhythm or organization of sounds in time;; and the timbre or tone color, which is the property that gives each instrument or combination of instruments its distinctive sound. Classification of music would ideally proceed based on these four properties.

2.4.1 Zero Crossing Rate (ZCR)

Acoustic feature ,time-domain feature, and number of times, the signal value crosses zero axis in time domain within a frame. The ZCR also makes it possible to differentiate between voiced and unvoiced speech components: voiced components have much smaller ZCR values than unvoiced ones. The average short-time zero-crossing rate can also be useful in combination with other features in general audio signal classification systems. The ZCR curves are calculated as follows: It is computed as-

$$Z_n = \sum_m |\text{sgn}[x(m)] - \text{sgn}[x(m-1)]|w(n-m),$$

(1)

where sign is the signum function, which returns 1 when the argument is positive, -1 when it is negative, and 0 otherwise.

2.4.2 Mel Frequency Cepstral Coefficients (MFCCs)

The MFCCs stands for the shape of the spectrum with few coefficients. The cepstrum is the Fourier Transform (or Discrete Cosine Transform DCT) of the spectral logarithm. The audio signal is first divided into number of periodic frames. Frames may contain samples that overlap with the previous frame. A window function is also used to minimize deviations at the beginning and end of the frame a windowing function (Hamming window is the most widely used one) is also applied on the frame. The amplitude spectrum for each frame (windowed) is obtained by applying Discrete Fourier Transform (DFT). It has led to the development of Mel frequency. The relation can be expressed as followed in Jothilakshmi et al. [2]:

$$\text{Mel}(f_m) = 2595 \times \log 1 + \frac{f}{700} \quad (2)$$

2.4.3 Spectral Centroid

It represents the balancing point or the midpoint of the spectral power distribution of a signal. Music involves high frequency sounds which means it will have higher spectral centroid values and the brighter. It is computed as

$$C = \frac{\sum_{f=1}^N f * M[f]}{\sum_{f=1}^N M[f]} \quad (3)$$

where the ratio of the sum of spectral magnitude weighted by frequency to the sum of spectral magnitude. A spectral centroid predicts how the dominant frequency of a signal changes over time expressed in Madjarov et al [4].

2.4.4 Skewness

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or even undefined described in Bantialebi-Dehkord et.al [5].

- Negative skew: The left tail is longer; the mass of the distribution is concentrated on the right of the figure. The distribution is said to be left-skewed, left-tailed, or skewed to the left.
- Positive skew: The right tail is longer; the mass of the distribution is concentrated on the left of the figure. The distribution is said to be right-skewed, right-tailed, or skewed to the right.

$$\text{skewness} = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{ns^3} \quad (4)$$

2.4.5 Kurtosis

Basically, because kurtosis is a mass movement that does not affect the variance. Consider the case of positive kurtosis, where the heavier tails has a higher peak. In a case of a negative kurtosis, if the mass moves from the tails and center of the distribution to its shoulders, the variance remains the same and the tail is lighter and flatness.

$$\text{kurtosis} = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{ns^4} - 3 \quad (5)$$

where \bar{x} is the sample mean and s is the sample standard deviation and n is the number of samples. The kurtosis of a normal distribution is 0.

2.5 Classifiers

Many audio classification problems involve high dimensional, noisy data. In this paper, two classifiers are chosen for audio classification as follows:

2.5.1 Sparse Representation Classifier (SRC)

Sparse representation by l_1 -norm minimization is robust to noise and even incomplete measurements. For the very large dataset, the SRC classifier is appropriate and thus optimization of the system is crucial. SRC first encodes a query sample into a linear combination of several atoms in a pre-defined dictionary. Then, it identifies the label by assessing which class results in the minimum reconstruction error. The SRC classification methods first calculates the sparse decomposition of test sample on training data set and then calculates the reconstruction residual errors which reconstruct test sample by sparse decomposition coefficients through each class of training samples respectively as in Bantialebi-Dehkord et al. [5]. J. Wright et al. [6] proposed with the SRC, the theoretical step to finding sparse representation is fast if the sparsest solutions are found.. The system finds the optimal description of the music parts from the feature set in respect to more similar function defined in Sparse Representation Classifier (SRC) method.

2.5.2 k-Nearest Neighbors Classifier (kNN)

It is a non-linear classifier and the idea is that a small number of neighbors will influence the identification on a point. More precisely, for a specific feature vector in the target set, the k closest vectors in the training set are selected and the target feature vector is the label of most representation of k neighbours (there is actually no other training than storing the features of the training set). KNN is the most popular for classification which means the training data is stored so that the classification for unclassified new data will be compared with the training data by taking the data of the most common training. To determine the similarity, the distance function is needed to test the size stated in Jothilakshmi et al. [2].

3. ANALYSIS AND RESULTS

In all experiment, the proposed system is evaluated with the feature combination. There are major nine features in this system. So, the zcr, centroid, bandwidth, MFCC-mean, MFCC-std, MFCC delta-mean, MFCC delta-std are represented by mean value of these features, and also zcr (2), centroid (2), bandwidth (2), MFCC-mean (2), MFCC-std (2), MFCC delta-mean (2), MFCC delta-std (2) are represented by standard deviation value of these features are tested in all experiment. The following classification results are described the feature

combination results by combining each feature for each ethnic class .

Experiment I: Evaluation of Two Feature Combination

According to the table 1, fig 2. and fig 3., feature combination of MFCC(std, deltamean) are tested on all of five ethnic classes which is the best classification results of 75.00% accuracy from SRC classifier than the result of 51.33% from KNN classifier. Jothilakshmi et al. [2] observed that performance of KNN is better than GMM for feature combination of MFCC and entropy in the indian music dataset but the low performance is obtained from feature combination of MFCC and spectral centroid for both of KNN and GMM. In our experiment I, the performance of SRC is achieved the best classification accuracy for all two combination of MFCC features. But, the feature combination of MFCC(std, deltastd(2)) give the low classification accuracy of 52.33% from SRC classifier and also 40.33% achieved from KNN classifier for all ethnic songs. In this two feature combination table, all feature combination are achieved the best by using SRC in compared with KNN classifier.

Experiment II: Evaluation of Three Feature Combination

According to table (2), only all feature combination of MFCC are used for myanmar ethnic music dataset which are achieved the better accuracy of 68.33% and 68.30% from SRC than KNN classifier. The performance of SRC is achieved the lowest accuracy of 54% for MFCC(std, mean(2), deltamean(2)) but also KNN classifier gives the low accuracy of 46.67%. In these three feature combination, SRC gives the better accuracy than KNN classifier as shown in fig 4 and fig 5. Jothilakshmi et al. [2] expressed the three feature combination (MFCC, Spectral centroid, Skewness) are used for the indian music dataset in which the performance of KNN and GMM are equally achieved the classification accuracy.

Experiment III: Evaluation of Four Feature Combination

According to table 3 and fig. 6, in the four feature combination of MFCC (mean, std, std(2), deltastd(2)), it obtained the better accuracy of 72.47% and also MFCC (deltamean, deltastd, std(2), deltamean(2)) achieved 70.69% of higher accuracy than the performance of KNN in which 62% and 54.33% for all ethnic songs respectively. In Jothilakshmi et al. [2],

the another combination MFCC, Spectral centroid, skewness, kurtosis, the performance of KNN is achieved 51.25% than 48.13% of GMM. When MFCC(deltamean, deltastd, mean(2), deltastd(2)) are tested with SRC classifier, it achieves the low accuracy of 62.27% but these accuracy is higher than the results of KNN classifier.

Experiment IV: Evaluation of Five Feature Combination

In table (4) and fig 7, the performance of SRC is obtained 70.68% for the feature combination of MFCC (mean, std, deltamean, deltastd, std(2)) that is better than the accuracy of 57% from KNN. When MFCC(mean, std, deltamean, deltastd, deltamean(2)) are combined, the higher accuracy of 66.83% achieved for all ethnic songs than the accuracy of 55.67% from KNN. In Jothilakshmi et al. [2], by adding the flatness feature, the performance of KNN is better than GMM in indian music classification. According to experiment IV, SRC classifier is achieved the better accuracy for all of MFCC feature combinations than KNN classifier for all ethnic song classes.

Experiment V: Evaluation of Best Feature Combination

According to the table (5), SRC is achieved the best classification result for two feature combination of MFCC (std, delta-mean) and four feature combination of MFCC (mean, std, std(2), deltastd(2)) gives than KNN classifier. Jothilakshmi et al [2],

pointed the performance of KNN is drastically decreased while combining MFCC, Spectral centroid, Skewness, Kurtosis, Flatness, Entropy, Irregularity, Rolloff, Spread. Similarly, the results of intro piece of music is decreased by combining all features (114). The performance of SRC is achieved 64% than KNN as shown in the following table 6. In fig 8, SRC gives the best accuracy for Kayin ethnic songs by using all feature combination(114).

4. CONCLUSIONS

In my observation, the system was evaluated by combining two features, three features, four features and five features in many ways. The influence of SRC can behave well these high dimensional audio data compared to other statistical or machine learning methods. Although the classification accuracy is increased to 75% by using the combination of MFCC(std, deltamean), the classification accuracy of SRC is decreased to 64% when all features (114) are used to classify all ethnic classes. From the analysis, the classification system are achieved the best outcomes by combining these timbre features than other features. Therefore, SRC gives the best accuracy of 82% for kayin ethnic songs than other ethnic songs. In conclusion, the obtained results have clearly shown that the system was achieved more representation flexibility and efficiency to classify the ethnic songs. Finally, all of the songs have cultural styles that are played with their respective traditional instruments, in otherwise the classification can't get the better accuracy.

Table 1. Two Feature Combination Classification Accuracy (%) for SRC Vs KNN

Intro Piece of Music (All ethnic classes)					
Feature Combination	SRC	KNN	Feature Combination	SRC	KNN
MFCC(mean, std)	67.00	54.00	MFCC(std, deltamean(2))	55.67	46.00
MFCC(mean, deltamean)	61.00	55.33	MFCC(std, deltastd(2))	52.33	40.33
MFCC(mean, deltastd)	65.60	56.00	MFCC(deltamean, mean(2))	59.00	53.00
MFCC(std, deltamean)	75.00	51.30	MFCC(deltamean, std(2))	69.67	59.00
MFCC(std, deltastd)	59.00	43.33	MFCC(deltamean, deltamean(2))	67.00	59.30
MFCC(mean, mean(2))	60.00	57.00	MFCC(deltamean, deltastd(2))	63.67	51.30
MFCC(mean, std(2))	63.30	52.33	MFCC(deltastd, mean(2))	60.00	53.66
MFCC(mean, deltamean(2))	64.30	57.33	MFCC(deltastd, std(2))	58.66	57.33
MFCC(mean, deltastd(2))	64.66	58.33	MFCC(deltastd, deltamean(2))	61.00	51.33
MFCC(std, mean(2))	57.00	44.66	MFCC(deltastd, deltastd(2))	56.67	48.33
MFCC(std, std(2))	62.34	57.00			

Table 2. Three Feature Combination Classification Accuracy (%) for SRC Vs KNN

Intro Piece of Music (All ethnic classes)					
Feature Combination	SRC	KNN	Feature Combination	SRC	KNN
MFCC(mean,std,deltamean)	66.60	55.30	MFCC(std,deltamean(2),deltastd(2))	56.60	43.66
MFCC(mean,Std,deltastd)	63.33	57.33	MFCC(deltamean,mean(2),std(2))	67.30	53.00
MFCC(mean,deltamean,deltastd)	67.30	56.67	MFCC(deltamean,mean(2),deltamean(2))	61.00	44.60
MFCC(std,deltamean,deltastd)	68.30	57.00	MFCC(deltamean,mean(2),deltastd(2))	61.60	45.00
MFCC(mean,mean(2),std(2))	66.00	56.33	MFCC(deltamean,std(2),deltamean(2))	68.33	60.33
MFCC(mean,mean(2),deltamean(2))	62.33	53.66	MFCC(deltamean,std(2),deltastd(2))	59.67	52.00
MFCC(mean,mean(2),deltastd(2))	60.00	51.67	MFCC(deltamean,deltamean(2),deltastd(2))	64.00	57.33
MFCC(mean,std(2),deltamean(2))	62.30	56.33	MFCC(deltastd,mean(2),std(2))	63.00	52.33
MFCC(mean,std(2),deltastd(2))	62.33	53.66	MFCC(deltastd,mean(2),deltamean(2))	63.67	47.66
MFCC(mean,deltamean(2),deltastd(2))	64.66	57.66	MFCC(deltastd,mean(2),deltastd(2))	64.67	53.00
MFCC(std,mean(2),Std(2))	60.30	48.00	MFCC(deltastd,std(2),deltamean(2))	66.34	50.66
MFCC(std,mean(2),deltamean(2))	54.00	46.67	MFCC(deltastd,std(2),deltastd(2))	61.33	51.33
MFCC(std,mean(2),deltastd(2))	58.30	48.66	MFCC(deltastd,deltamean(2),deltastd(2))	60.00	45.66
MFCC(std,std(2),deltamean(2))	65.90	52.33	MFCC(mean(2),std(2),deltamean(2))	59.70	54.33
MFCC(std,std(2),deltastd(2))	63.00	51.33	MFCC(mean(2),std(2),deltastd(2))	58.19	51.00

Table 3. Four Feature Combination Classification Accuracy (%) for SRC Vs KNN

Intro Piece of Music (All ethnic classes)		
Feature Combination	SRC	KNN
MFCC(mean,std,deltamean,deltastd)	68.38	59.00
MFCC(mean,std,mean(2),std(2))	65.00	54.66
MFCC(mean,std,mean(2),deltamean(2))	68.34	57.67
MFCC(mean,std,mean(2),deltastd(2))	64.41	53.67
MFCC(mean,std,std(2),deltamean(2))	63.54	58.67
MFCC(mean,std,std(2),deltastd(2))	72.47	62.00
MFCC(mean,std,deltamean(2),deltastd(2))	66.42	60.67
MFCC(deltamean,deltastd,mean(2),std(2))	65.93	49.00
MFCC(deltamean,deltastd,mean(2),deltamean(2))	65.45	47.33
MFCC(deltamean,deltastd,mean(2),deltastd(2))	62.27	50.00
MFCC(deltamean,deltastd,std(2),deltamean(2))	70.69	54.33
MFCC(deltamean,deltastd,std(2),deltastd(2))	65.13	55.00
MFCC(deltamean,deltastd,deltamean(2),deltastd(2))	70.65	47.67

Table 4. Five Feature Combination Classification Accuracy (%) for SRC Vs KNN

Intro Piece of Music (All ethnic classes)		
Feature Combination	SRC	KNN
MFCC(mean,std,deltamean, deltastd,mean(2))	66.17	55.33
MFCC(mean,std,deltamean, deltastd,std(2))	70.68	57.00
MFCC(mean,std,deltamean, deltastd,deltamean(2))	66.83	55.67
MFCC(mean,std,deltamean, deltastd,deltastd(2))	63.13	58.33

Table 5. Best Feature Combination Classification Accuracy (%) for intro piece of music

Feature Combination	SRC	KNN
MFCC(std,deltamean)	75.00	51.33
MFCC(deltamean,std(2))	69.67	59.00
MFCC(deltamean,std(2),deltamean(2))	68.33	60.33
MFCC(mean,deltamean,deltastd)	67.30	56.67
MFCC(mean,std,std(2),deltastd(2))	72.47	62.00
MFCC(deltamean,deltastd,std(2),deltamean(2))	70.69	54.33
MFCC(mean,std,deltamean,deltastd,std(2))	70.68	57.00
MFCC(mean,std,deltamean,deltastd,deltamean(2))	66.83	55.67

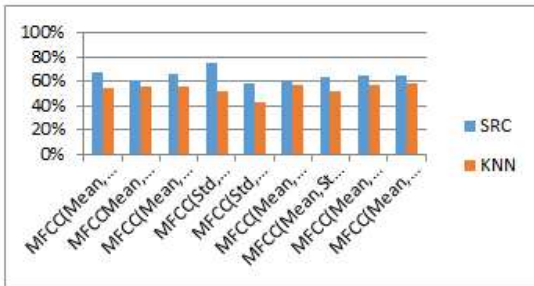


Fig 2. Charts of two features combination for intro piece of music classification

Table 6. Classification accuracy (%) of all features(114) for all ethnic classes

Intro Piece of Music Classification Accuracy (%)		
Ethnic Classes	SRC	KNN
Kachin	78.00	63.33
Shan	58.00	60.00
Mon	52.00	46.67
Kayin	82.00	65.00
Yakhine	50.00	35.00
All	64.00	54.00

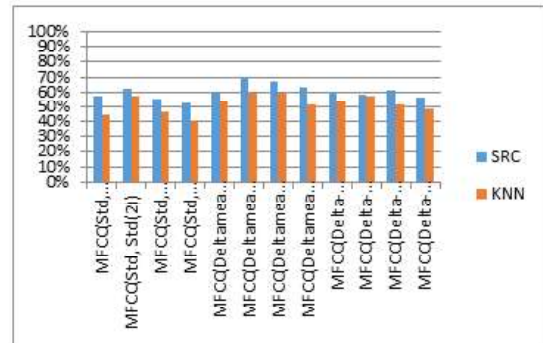


Fig 3. Charts of two features combination for intro piece of music classification

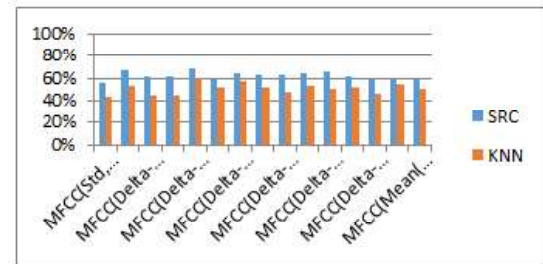


Fig 4. Charts of three features combination for intro piece of music classification

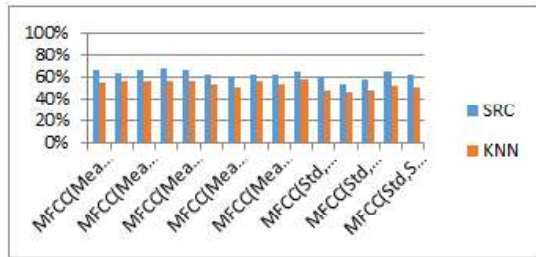


Fig 5. Charts of three features combination for intro piece of music classification

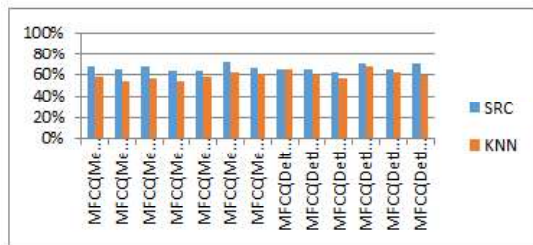


Fig 6. Charts of four features combination for intro piece of music classification

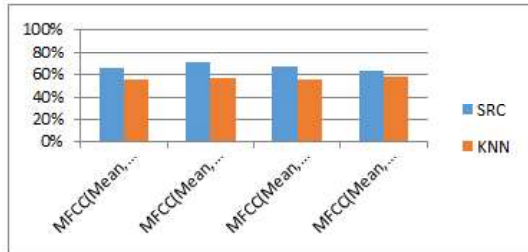


Fig 7. Charts of five features combination for intro piece of music classification

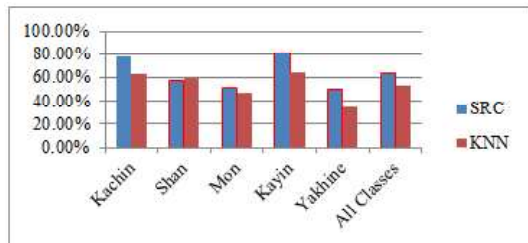


Fig 8. Charts of music classification for all features

ACKNOWLEDMENT

There has been an inspiring, often exciting, sometimes challenging, but definitely always interesting experience from the commencement until the completion of this research. The authors wish to thank my sincere gratitude to my supervisor Dr. Nu War, University of Computer Studies, Mandalay for kind and constant encouragement, close guidance, patient supervision and detail technical assistance throughout my research work. Additional thanks are extended to all my friends, for their help and support during my research.

REFERENCES

- [1] Y. Liu, Q. Xiang, Y. Wang, and L. Cai, "Cultural style based music classification of audio signals," 2009, pp. 57–60.
- [2] S. Jothilakshmi and N. Kathiresan, "Automatic Music Genre Classification for Indian Music," p. 5.
- [3] S. H. Deshmukh and D. S. G. Bhirud, "Analysis and application of audio features extraction and classification method to be used for North Indian Classical Music's singer identification problem," vol. 3, no. 2, p. 6, 2014.
- [4] G. Madjarov, G. Pesanski, and D. Spasovski, "Automatic Music Classification into Genres," *ICT Innov.*, p. 10, 2012.
- [5] M. Banitalebi-Dehkordi and A. Banitalebi-Dehkordi, "Music Genre Classification Using Spectral Analysis and Sparse Representation of the Signals," *ArXiv180304652 Cs Eess*, Mar. 2018.
- [6] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Yi Ma, "Robust Face Recognition via Sparse Representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

COMPARISON OF SUPPORT VECTOR MACHINE AND LOGISTIC REGRESSION CLASSIFIER FOR TOMATO FRUITS DETECTION

Hsan Lynn⁽¹⁾, Thae Nu Nge⁽²⁾, Moh Moh Khaing⁽³⁾

⁽¹⁾⁽²⁾⁽³⁾Technological University (Taunggyi), Myanmar

hsanlynn82@gmail.com

ABSTRACT

Vision is one of the most important senses used by humans. The human brain can recognize many objects effortlessly, even though the objects may appear differently depending upon the viewpoint, illumination, etc. Objects can even be detected if there are occlusions caused by other objects. In computer vision and robotics field object detection is vital. In this system, Histograms of Oriented Gradients (HOG) descriptor is used to training a Support Vector Machine (SVM) classifier or Logistic regression (LR) classifier. This system involves images acquisition and preprocessing. HOG is used for extracting the shape features. Finally fruit detect or not by using SVM or LR. In this paper analyses the performance of SVM and LR classifier by comparing the detection of tomato fruits on plant. According to the F1 score, the result of tomato detection in the test images, SVM classifier is more accuracy than LR classifier. The recall, precision and F1 score of the SVM classifier were 89 %, 85 % and 87 % respectively.

.KEYWORDS: *object detection, HOG, SVM, LR.*

1. INTRODUCTION

Nowadays, precision agriculture is widely applied in developed countries, while the information technologies are applied in precision agriculture. To get high quality yield production, many factors influence. Knowing the exact number of fruits, flower and tree helps farmers to make benefit decision on cultivation practices, plant disease prevention, the size of harvest labour force and etc. In agriculture sector the problem of identification and detection fruits on trees plays an important role in crop estimation work. There are various challenges faced by computer vision algorithms for detection fruits

for yield estimation, namely, illumination variance, and occlusion by foliage, varied degree of overlap amongst fruits, fruits under shadow and the scale variation. In this study, a novel and simple technique is discussed to detect tomato fruits in the tree canopy under natural outdoor conditions by utilizing shape and texture information.

This paper is based on computer vision and machine learning in precision agriculture. Computer vision is commonly used for feature extraction from images. Machine learning is one of the intelligent methodologies that have shown promising results in the domains of classification. A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .

There are various classifiers in machine learning. SVM and LR are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis. Classification of images can also be performed by using SVM or LR methods. In this paper, the performance and accuracy of SVM and LR classifier are compared. The F1 score was used to measure the performance of the classifiers. This helped provide meaningful information given that the F1 score is the harmonic mean between precision and recall.

2. LITERATURE REVIEW

Subhajit Sengupta and Won Suk Lee [2014] proposed an improved model Identification and Determination of the Number of Green Citrus Fruit under Different Ambient Light Conditions. The model proposes a novel approach was presented to detect green citrus fruit from images taken by a typical

digital color camera under natural outdoor lighting conditions in the tree canopy. The approach utilized the circular Hough transform, texture classification with a support vector machine, and keypoints by scale invariant feature transform algorithm. Its performance was evaluated with a set of test images taken from a citrus grove. The Hough circle detection and texture classification based on SVM were implemented. Texture feature classification by SVM performed well in removing false positives. The algorithm was able to accurately detect and count over 81% of citrus fruit.

InkyuSa and ZongyuanGe [2016] presented a methodology for fruit detection using deep convolutional neural networks. The aim is to build an accurate, fast and reliable fruit detection system, which is a vital element of an autonomous agricultural robotic platform; it is a key element for fruit yield estimation and automated harvesting. Recent work in deep neural networks has led to the development of a state-of-the-art object detector termed Faster Region-based CNN (Faster R-CNN). This model, was adapted through transfer learning, for the task of fruit detection using imagery obtained from two modalities: color (RGB) and Near-Infrared (NIR). Early and late fusion methods are explored for combining the multi-modal (RGB and NIR) information. This leads to a novel multi-modal Faster R-CNN model, which achieves state-of-the-art results compared to prior work with the F1 score, which takes into account both precision and recall performances improving from 0.807 to 0.838 for the detection of sweet pepper.

P.J.Tamos, F.A.Prieto, E.C. Montya and C.E. Oliverors [2017] focused on to count the number of fruits on a coffee branch by using information from digital images of a single side of the branch and its growing fruits. A Machine Vision System (MVS) was constructed, which was capable of counting and identifying harvestable and not harvestable fruits in a set of images corresponding to a specific coffee branch was constructed. This MVS consists of an image acquisition system, based on mobile devices, and an image processing algorithm to classify and detect each one of the fruits in the acquired images. After obtaining information regarding the number of fruits identified by the MVS, linear estimation models were constructed between the detected fruits automatically and the ones observed on the coffee branch. These models were calculated for fruits in three categories: harvestable, not harvestable, and fruits whose maturation stage were disregarded.

3. BACKGROUND THEORY

3.1 HOG Feature Extraction

Histogram of Oriented Gradient descriptor is actually a feature descriptor widely used in machine vision field for the purpose of detection of object. It evaluates the number of occurrences of gradient orientation in localized parts of an image. The main point of HOG is the distribution of intensity gradient and edge directions. It could be done by dividing the image into small connected regions, each compiled with histogram of gradient direction and edge orientations for the pixels involved. Hence, the histograms merged to become a descriptor. The HOG can capture the shape information of an object and is invariant to geometric and photometric transformations. It can also deal with slight occlusion. However, there is little research on fruit detection using HOG. Thus, HOG features are used in tomato detection in this work. There five steps in HOG Features Extraction.

Step 1: Gradient calculation: Calculate the x and the y gradient images, g_x and g_y , from the original image. This can be done by filtering the original image with the following kernels, Fig. 1.

$$\begin{bmatrix} -1 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

Fig. 1Kernels

Using the gradient images g_x and g_y , we can calculate the magnitude and orientation of the gradient using the following equations (1) –(2).

$$g = \sqrt{g_x^2 + g_y^2} \quad (1)$$

$$\theta = \arctan \frac{g_x}{g_y} \quad (2)$$

The calculation of gradients is unsigned and θ is in the range 0 to 180 degrees.

Step 2: Cells: Divide the image into 8×8 cells.

Step 3: Calculate histogram of gradients in these 8×8 cells: At each pixel in an 8×8 cell gradient

(magnitude and direction), and therefore it has 64 magnitudes and 64 directions 128 numbers. Histogram of these gradients will provide a more useful and compact representation. Next, these 128 numbers are converted into a 9-bin histogram. The bins of the histogram correspond to gradients directions 0, 20, 40 ... 160 degrees. Every pixel votes for either one or two bins in the histogram.

Step 4: Block normalization: The histogram calculated in the previous step is not very robust to lighting changes. Multiplying image intensities by a constant factor scales the histogram bin values as well. To counter these effects the histogram is normalized as a vector of 9 elements and divide each element by the magnitude of this vector. In the original HOG paper, this normalization is not done over the 8×8 cell that produced the histogram, but over 16×16 blocks. The idea is the same, but now instead of a 9 element vector you have a 36 element vector.

Step 5: Feature Vector: In the previous steps calculate histogram over an 8×8 cell and then normalize it over a 16×16 block. To calculate the final feature vector for the entire image, the 16×16 block is moved in steps of 8 and the 36 numbers calculated at each step are concatenated to produce the final feature vector. The input image is 64×64 pixels in size, and moving 8 pixels at a time. Therefore, 7 steps move in the horizontal direction and 7 steps in the vertical direction which adds up to 7 x 7 = 49 steps. At each step 36 numbers are calculated, which makes the length of the final vector 49 x 36 = 1764. Fig. 1 shows a visualizing example of HOG features of a tomato.

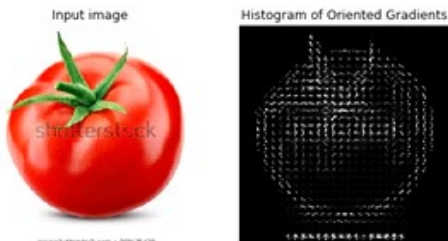


Fig. 1 Testing for visualizing Histogram of Oriented Gradients

3.2 CLASSIFICATION

3.2.1 Support Vector Machine

Support Vector Machine (SVM) is one of the most popular supervised binary classification algorithms. Although the ideas used in SVM have

been around since 1963, the current version was proposed in 1995 by Cortes and Vapnik.

In classification, support vector machines separate the different classes of data by a hyperplane

$$\langle w, \phi(x) \rangle + b = 0 \quad (3)$$

corresponding to the decision function

$$f(x) = \text{sign}(\langle w, \phi(x) \rangle + b) \quad (4)$$

$$w = \sum_i \alpha_i \phi(x_i) \quad (5)$$

The hyperplane is constructed by solving a constrained quadratic optimization problem whose solution w has an expansion in terms of a subset of training patterns that lie on the margin. These training patterns, called support vectors, carry all relevant information about the classification problem in Fig 2

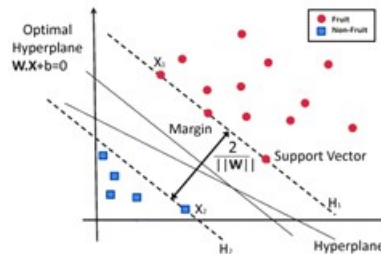


Fig. 2 Classification of data by SVM

3.2.2 Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no). Mathematically, a logistic regression model predicts $P(Y=1)$ as a function of X . It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

Logistic regression is used to model the probability of classification into an 'fruit' or 'Not-fruit'. Let Y_i indicate the classification of the i th fruit

same such that $Y_i=1$ if the fruit sample is classified as 'a' and $Y_i=0$ otherwise. Then let $\pi_i=P(Y_i=1|X_i)$ where X_i is a $1 \times (p+1)$ vector with first element equal to 1 and the remaining elements corresponding to the p fruit characteristics for fruit sample i . The LR model relates π_i to the fruit sample characteristics by the logit function:

$$(6) \quad \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = X_i\beta$$

Where $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ is the vector of regression coefficients. A simple and common predictive classification procedure is to allocate an unobserved response 'Fruit' if $\pi_i > 0.5$ and 'Not-Fruit' otherwise. Fig. 3 shows Logistic regression function.

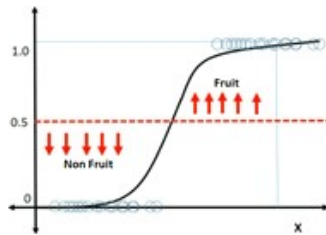


Fig. 3 Classification of data by Logistic Regression

4. Proposed System

There are two phase in the proposed systems: Training and Testing. In the training phase has four states and the testing phase has eleven states.

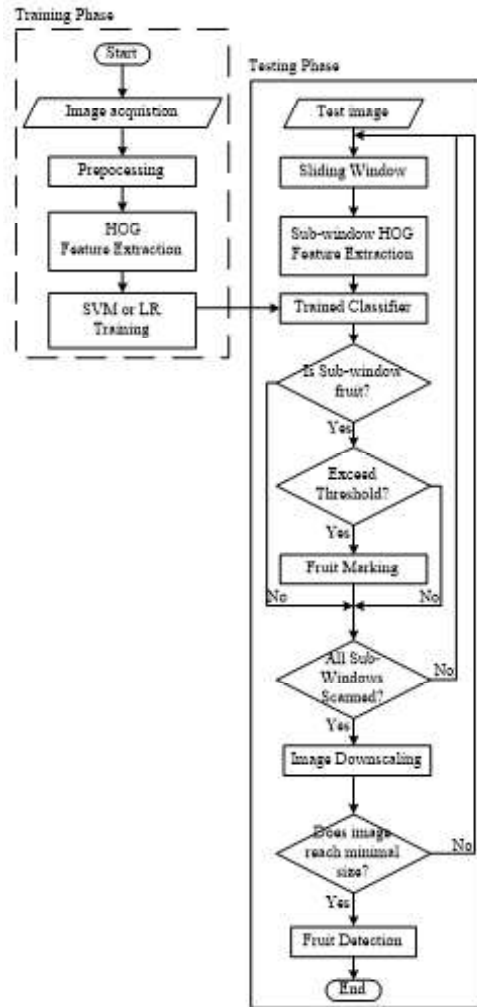


Fig. 4 System flow chart of Tomato detection framework

In training phase, image acquisition can be done in two ways: on-line and off-line. The system is used by off-line image with jpg format taken at day time from sunrise to sunset under different lighting conditions using a digital camera. To develop and evaluate the proposed algorithm, images of tomatoes in a greenhouse were acquired Mya Thein Tan Village, Nyaung Shwe Township, Southern Shan State. A total of 98 images were captured using a color digital camera (Sony DSC-S930) with a resolution of 3648 x 2736 pixels. The photographs were taken at distances of 2-3 feet, which is in accordance with the best operation distance for the

harvesting robot. A total of 50 images were used for the experiment. To speed up the image processing, all of the images were resized to 720 x 540 pixels using Photoshop software. From the training images, 320 tomato samples and 621 background samples were manually cropped to construct a training set. All of the cropped samples were resized to 64 x 64 pixels to unify the size. The background samples were randomly cropped to contain leaves, stems, strings, and other objects in Fig. 5.



Fig. 5 TomatoFruits, leaves and stems

Feature extraction state is very important state of our proposed system. The First step of feature extraction is to define the parameter of HOG and then refer the path of positive images (fruits) and negative images (leaves and stems). After that each image is converted RGB to grayscale. The final step is calculating the HOG feature and labeling 1 for positive images and 0 for negative images.

The SVM and LR classifier state is used to implement the classification model. The tomato dataset was divided into training size = 0.8 and testing size = 0.2. The process of training uses the open source Scikit-Learn packet contain the SVM and LR libraries in python 2.7. After training process the Linear SVM classifier and Logistic regression the result of training model is shown in Fig. 6 and 7.

IPython console

```

Console I/A
Constructing training/testing split...
Training Linear SVM classifier...
Evaluating classifier on test data ...

```

	precision	recall	f1-score	support
0	0.95	1.00	0.97	466
1	0.80	0.27	0.40	30
micro avg	0.95	0.95	0.95	496
macro avg	0.88	0.63	0.69	496
weighted avg	0.95	0.95	0.94	496

Fig. 6 Training with SVM

IPython console

```

Console I/A
Constructing training/testing split...
Training Logistic Regression classifier...
Evaluating classifier on test data ...

```

	precision	recall	f1-score	support
0	0.95	1.00	0.97	466
1	0.86	0.20	0.32	30
micro avg	0.95	0.95	0.95	496
macro avg	0.90	0.60	0.65	496
weighted avg	0.95	0.95	0.93	496

Fig. 7 Training with Logistic Regression

The first step of testing phase is input an image and then sliding sub-windows on the test image. After that HOG Features of each sub-window are extracted using same as the training model parameter. The extract feature result of each sub-window is classification with SVM train model to make a prediction on the HOG feature extracted from the image. If the result is 1, it is a fruit. If the result is 0, it is non fruit. If it is a fruit and the probability of fruit is also exceeding threshold value, the system will detect the fruit. Although is a sub-windows fruit, the probability of fruit is under the threshold value, the system will not detect the fruit. The previous stages are repeated again until all sub-windows scanned are finished. If all sub-windows scanned is finished, reduced the size of image and the previous states is also repeat until reach the minimal size of the image. Finally the system shows the fruit detection result.

5. TESTING AND RESULTS

In this study, all experiments of the developed algorithm were performed on Python version 2.7 with Intel® Core™ i7-4770 3.40 GHz and RAM 16 G.

HOG features with different cell sizes, block sizes, and number of orientation bins were tested on the training sample set. In this study, the HOG feature with 8 x 8 pixel cells, 2 x 2 cell blocks, and 9 orientation bins was used for experiments.

In this paper, five images are randomly selected from the testing image dataset. Table 1 and 2 show 5 representative images along with their actual and predicted count. In Table 1 and 2 column Images contains the test images, column AC contain the actual count of ground truth, column DFC contain the detected fruit count and column DNFC contain the detected non fruit count.

Three indexes will use to analysis the performance of the proposed system: recall, precision

and F1 score, which are defined by the Equations (7) to (9).



$$\text{Recall} = \frac{\text{Correctly Identified Tomato Count}}{\text{Total Number of Tomatoes}} \times 100\% \quad (7)$$

$$\text{Precision} = \frac{\text{Correctly Identified Tomato Count}}{\text{Total Number of Detections}} \times 100\% \quad (8)$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \quad (9)$$

There are 64 tomatoes fruit in selected image. To evaluate the performance of results, the predicted count of proposed system is compared with actual count. The accuracy of SVM and LR was calculated as shown in Table 1 and 2:

Table. 1 Result of SVM Classifier






IMAGE	AC	DFC	DNFC
	16	16	1
	15	13	3
	13	11	2
	11	10	2
	9	7	2

$$\text{Recall for SVM} = \frac{57}{64} \times 100\% = 89\%$$

$$\text{Precision for SVM} = \frac{57}{57+10} \times 100\% = 85\%$$

$$F_1 \text{ for SVM} = \frac{2 \times 89 \times 85}{89 + 85} \times 100\% = 87\%$$

Table. 2 Result of Logistic regression Classifier

IMAGE	AC	DFC	DNFC
	16	15	0
	15	12	2
	13	10	3
	11	10	2
	9	6	3

$$\text{Recall for LR} = \frac{53}{64} \times 100\% = 82\%$$

$$\text{Precision for SVM} = \frac{53}{53+10} \times 100\% = 84\%$$

$$F_1 \text{ for LR} = \frac{2 \times 82 \times 84}{82 + 84} \times 100\% = 83\%$$

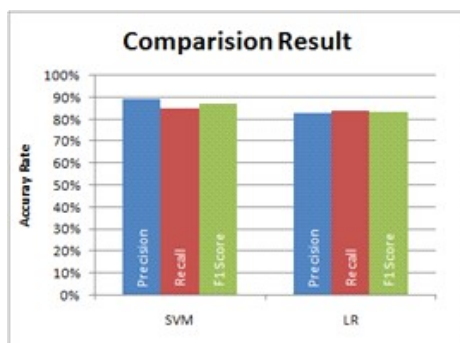


Fig. 8 Comparison Result of SVM and LR Classifier

According to the experimental results shown in Fig.8, the result of tomato detection in the test images showed that the recall, precision and F1 score of the SVM were 89 %, 85 % and 87%. Logistic regression result is 82%, 84% and 83% respectively. The detection performance of SVM is more accuracy than Logistic regression.

6. CONCLUSION

This research is proposed to overcome the difficulties that harvesting robot face in fruit detection. This system can handle the variation in illumination, size, and shadow, even also images with overlapped and partially-occluded fruits. Object detection is a key ability for most computer vision and robot system. This paper presents the comparison of SVM and LR classifier using HOG feature extractor for tomatoes fruit detection on plant. According to F₁ score, the detection performance of SVM is more accuracy than LR.

Although our proposed system is trained to detect tomatoes, it can be applied to other fruits. After detection using the classifier, there are many sub-windows classified as fruits and some to them correspond to the same one. Our proposed system future results on: (1) accelerate the fruit detection speed (2) increase the detection accuracy (3) detection fruit on real time and (4) Merging the detection results using the Non-Maximum Suppression.

ACKNOWLEDGEMENT

The author would like to thank Dr. Aye Zarchi Minn, Professor and Head, Department of Information Technology, Technological University (Taunggyi) for her helpful and valuable guidance. The author also has to say thank Dr. Thae Nu Nge, Daw Moh Moh

Khaing and all the teachers from Department of Information Technology, Technological University (Taunggyi) for their support and guidance.

REFERENCES

- [1] S. Sengupta and W. S. Lee, "Identification and determination of the number of immature green citrus fruit in a canopy under different ambient light conditions," *Biosyst. Eng.*, vol. 117, pp. 51–61, Jan. 2014.
- [2] I. Sa, Z. Ge, F. Dayoub, B. Upcroft, T. Perez, and C. Mccool, "DeepFruits: A Fruit Detection System Using Deep Neural Networks," *Sensors*, vol. 16, p. 1222, Aug. 2016.
- [3] P. J. Ramos, F. A. Prieto, E. C. Montoya, and C. E. Oliveros, "Automatic fruit count on coffee branches using computer vision," *Comput. Electron. Agric.*, vol. 137, pp. 9–22, May 2017.
- [4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 1, pp. 886–893 vol. 1.
- [5] "Histogram of Oriented Gradients | Learn OpenCV."
- [6] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
- [7] S. Patel, "Chapter 2/ : SVM (Support Vector Machine) — Theory," *Medium*, 04-May-2017. [Online]. Available: <https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72>. [Accessed: 12-Dec-2019].
- [8] "GitHub - nrsyed/svm-vehicle-detector: HOG-based linear SVM for detecting vehicles (or any other object) in videos." [Online]. Available: <https://github.com/nrsyed/svm-vehicle-detector>. [Accessed: 12-Dec-2019].

COMPARATIVE TESTING OF FACE DETECTION METHODS

Thi Thi Khaing ⁽¹⁾, Khin Thein Su ⁽²⁾

⁽¹⁾⁽²⁾ Technological University (Taunggyi), Taunggyi, Myanmar

⁽¹⁾thithikhaing181015@gmail.com

ABSTRACT

Face and eye detection is the most challenging and popular problem in the machine vision area. After Viola and Jones proposed a work, it has been solved reasonably well by traditional feature-based techniques, such as the cascade classifier. More recently deep learning algorithms have attained state of the art results on standard face datasets. The purpose of this work is to study the face detection algorithms and make comparative testing in respect of their performance and accuracy. In this paper, the two algorithms were tested and analyzed: Viola-Jones algorithm and Deep Belief Network which is the modern method to solve wild faces problem. Finally, the results are presented which proves the performance and precision of the compared face detection methods.

KEYWORDS: *Face detection, Viola-Jones, Deep Learning, Deep Belief Network, Restricted Boltzmann Machines*

1. INTRODUCTION

The advance of computing technology has facilitated the development of real-time vision modules that interact with humans in recent years. Face detection is a biometrics system that determines the sizes and location of faces in digital images, which is a key technology in face information processing. It has been widely used in various computer vision areas such as automatic video surveillance, pattern recognition, human computer interface and identity authentication, etc.

There are many methods has been researched for a long time and much progress has been proposed in literature. These methods can be divided into four main types; template matching, feature invariant,

knowledge based and appearance based. Most of the face and face parts detection methods concentrated on detecting frontal faces with enough lightning condition.

- Knowledge based methods create facial feature model using human coding including two symmetric parts such as mouth, nose, eye, etc.
- Feature invariant method finds the features which are invariant to pose and lighting.
- Template matching methods use the correlation between features of trained image and test image.
- Appearance based method uses machine learning algorithms to extract particular features from a pre-labeled image set.

The face detection problem is challenging as it needs to account for all possible appearance variation caused by partial occlusions, facial features, change in illumination, etc. In addition, it has to detect faces that appear at different scale, occlusion and multi-pose. Therefore, it is worthwhile to a further investigating for face detection. Currently, Deep Learning is a new area of computer vision and machine learning research, which has been successfully introduced to image dimensional reduction and recognition. Deep learning originates from artificial neural networks and consists of multi-layer perceptron (MLP) of multi-hidden layers which is a deep learning structure.

2. ALGORITHMS

For this experiment, Viola-Jones and Deep Belief Network (DBN) methods were tested using OpenCV image processing library and Python programming

language. The Viola-Jones method [1]-[2] is a widely used algorithm for face and object detection. The main character of this method is that training is very slow, but detection is fast. This method uses Haar basis feature extractors, so it does not use multiplications. By using the integral image technique, efficiency of the Viola-Jones object detection algorithm can be significantly increased. It makes integrals for the Haar feature filters [3] to be calculated by summing only four numbers. This avoids repeatedly summing of the pixel intensities within a rectangle into only three operations with four numbers. The sum of pixels in the rectangle ABCD can be derived from the values of points A, B, C, and D, using the formula $D - B - C + A$. It is easier to understand this formula visually:

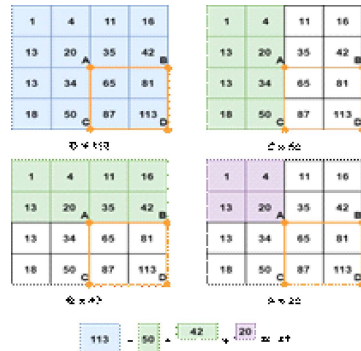


Fig.1 Image area integration using integral image

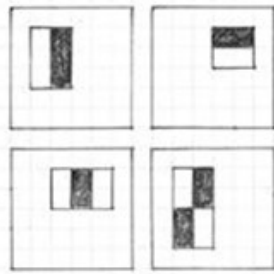


Fig.2 Example rectangle features shown relative to the enclosing detection window

Each face detection classifier is composed of weak classifiers which are Haar feature extractors. Each Haar feature is the weighted sum of 2-D integrals of small rectangular areas attached to each other. The weights may take values +1 or -1. In figure 2, white areas represent a negative weight and gray areas hold a positive weight. Rectangular subset of Haar feature extractors are scaled to fit with face

detection window size. The cascade classifier is an efficient classifier which applied the fewest features along the process of the cascade. The result is minimizing the total requirement of computation. The most popular algorithm for features training is AdaBoost [4].

In machine learning, Deep Belief Network (DBN) [5] is a type of deep learning architecture [6] and described as a graphical model as shown in figure 3. The DBN reconstructs its inputs probabilistically and build a neural network layers starting with the input layer and end with output layer. Between input and output layers, different numbers of hidden layers are present. Every layer should be trained with Restricted Boltzmann Machines (RBM) [7]. An RBM is an extract feature to reconstruct inputs. By combining all RBM's there should be produce a new power full collaboration method for our problem which is called DBN.

DBN is considered as a stack of RBM's with a back propagation to obtain a powerful trained neural network. The training process is the key task for developing powerful neural networks. A DBN is a stack of RBN's. The first RBM is constructed as input layer. When the second RBM is trained with the supporting of first RBM also called visible layer, this process is repeated until all layers are completed the training process. In detection process, the job of hidden layers is to find the edges of the face and visible layer finds the facial features. The output layer will forward as the input of next layer after its going to end. If the training process is not complete, it comes back to start and train by considering previous output training to input for present training.

$$P(x, h^1, \dots, h^\ell) = (\prod_{k=0}^{\ell-2} P(h^k | h^{k+1})) P(h^{\ell-1}, h^\ell)$$

Where, x is the observed vector, -1 is the no of hidden layers, h^k is represent each hidden layer, $P(h^{\ell-1}, h^\ell)$ is visible layer and hidden layer joint for the RBN at level k , $P(h^{k-1} | h^k)$ at $x = h^0$ is the conditional distribution for visible units.

Face detection happens inside a predefined detection window of an image. A minimum and maximum size of detection window and sliding step size for each window size is defined. Then the detection window with detection filters is moved across the image. The detection steps of Viola-Jones algorithm are as follows.

- Define the minimum window size, and sliding step corresponding to that window size.
- Slide the detection window horizontally and vertically with the size of sliding step. A set of detection filters is applied at each step. If a filter output a positive result, a face is detected in the current detection window.
- Stop the procedure when the size of the window is reached at the maximum. Otherwise increase the window size to the next chosen size and sliding step size and then go second step.

The face detection filter which is applied at each detection step contains a set of cascade of booted classifiers. Each classifier looks at a rectangular subset of the detection window and determines if it looks like a face. If it does, the next classifier is applied. If all classifiers give a positive answer, the filter gives a positive answer and the face is recognized. Otherwise the next filter in the set of N filters is run.

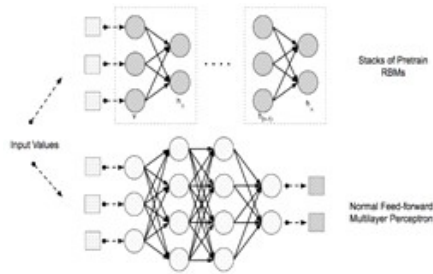


Fig.3 Architecture of BDN

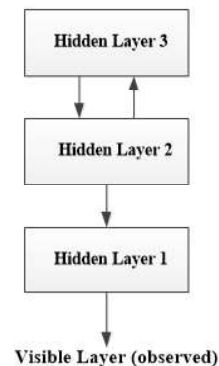


Fig.4 DBN with 3 hidden layers

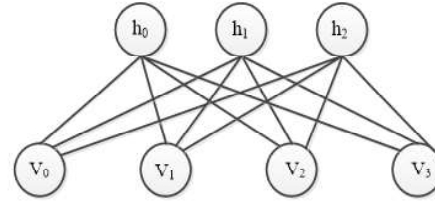


Fig.5 RBM connections between visible and hidden layers

3. EXPERIMENT

In this experiment, the performances of two algorithms are evaluated. The following optimized tools and image processing libraries are used by author for implementation of presented algorithms.

3.1 Tool

OpenCV (Open-source Computer Vision) is highly optimized open source computer library, which is mainly focus on real-time application. It has a wide range of modules that can help us with a lot of computer vision problems. But perhaps the most useful part of OpenCV is its architecture and memory management. It provides you with a framework in which you can work with images and video in any way you want, using OpenCV's algorithms or your own, without worrying about allocating and reallocating memory for your images. Open CV libraries and functions are highly optimized and can be used for real time image and video processing. OpenCV's highly optimized image processing function are used by author for real time image processing of live video feed from camera.

3.2 Data

In this work, the training data set consists of 2825 images which are obtained from the Labeled Faces in the Wild (LFW) dataset and manually annotated. This can be applied to research the problem of face detection under non-limitation condition. The LFW is a rich-content face database. It consists of the face images of various angles, postures and occlusions under the different situations.

3.3 Results

Based on the results of the experiment, the traditional classic Viola-Jones algorithm is perhaps the best of the algorithms for implementation in frontal face detection and video processing. The

accuracy rates leave much to be desired. The major disadvantage of this method is that it gives a lot of false predictions. It can work on frontal images only and doesn't work under occlusion. The major advantage is works almost real-time on CPU and simple architecture. It can detect faces at different scales. The DBN is able to recognize faces correctly but when tried for videos, it takes more time for processing. The advantages of DBN are able to recognize blurred images and multi-pose face images also which other models (like Viola-Jones) are incapable of recognizing in such case.

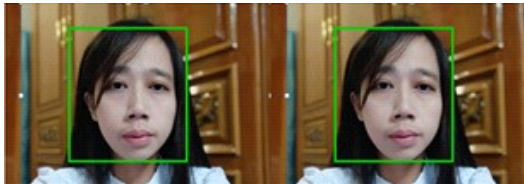


Fig.6 Frontal Face(left-VJ and right-DBN)

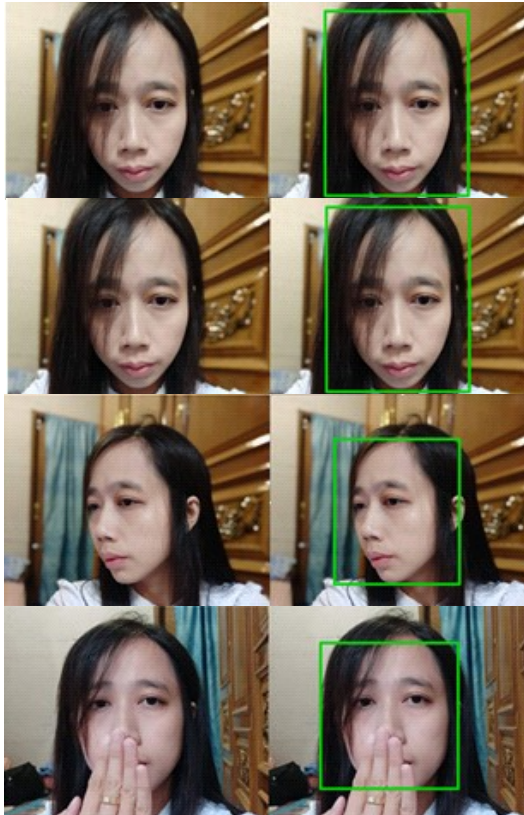


Fig.7 Multi-pose (left-VJ and right-DBN)

The following tables are performance comparison of Viola-Jones and DBN face detection algorithms on multi-criterion faces. Both algorithms work well at frontal faces. But DBN do better than Viola-Jone for multi-pose and partial occlusion.

Table.1 Performance comparison of two face detection algorithms

Algorithm	TP	FP	Detection Rate (%)	False Detection Rate (%)	Precision (%)
Viola-Jones	916	98	92.6	9.8	90.3
DBN	987	25	98.7	2.5	97.5

Table 2 shows the face detection accuracy of the two algorithms using 1000 images of LFW dataset.

Table.2 Precision comparison of two face detection algorithms

Criterion	Viola-Jones	DBN
Frontal Face	97.1 %	96.9 %
Group	82.5 %	93.4 %
Multi-Pose	21.4 %	61.6 %
Partial Occlusion	12.7 %	63.3 %

4. CONCLUSION

Based on the results, the DBN is more efficient for performing in face detection than Viola-Jones method. The Viola-Jones is simple and it can correctly detect the frontal and scale faces. The advantages of DBN are able to recognize multi-pose face images also which other methods like Viola-Jones are incapable of detection in such case. The drawbacks of DBN are that it fails to recognize eyes with glasses and CPU expensive in video processing. The key problems can be identified for any face detection systems; one is illumination problem and second is pose problem. Finally, there is still no method to believe for all unconstraint real-world applications.

ACKNOWLEDGEMENT

The author would like to express heartfelt gratitude to Dr. Aye Aye Nwe, Professor, Department of Electronics and Communication, Taunggyi Technology University, for her helps, directly or indirectly supports for my paper.

The author is greatly and especially thankful to Daw Khin Thein Su, Lecturers, Department of Electronics and Communication, for her guidance and support through the research.

REFERENCES

- [1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, Kauai, HI, USA, 2001, vol. 1, pp. I-511-I-518.
- [2] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," p. 19.
- [3] T. Mita, T. Kaneko, and O. Hori, "Joint Haar-like features for face detection," presented at the IEEE Int Conf Comp Vis, 2005, vol. 2, pp. 1619-1626 Vol. 2.
- [4] S. Yoosaf, "Face Detection & Smiling Face Identification Using Adaboost & Neural Network Classifier," vol. 4, no. 8, p. 3, 2013.
- [5] Y. Bengio, "Learning Deep Architectures for AI," *Found Trends Mach Learn*, vol. 2, no. 1, pp. 1–127, Jan. 2009.
- [6] M. Li, C. Yu, F. Nian, and X. Li, "A Face Detection Algorithm Based on Deep Learning," *Int. J. Hybrid Inf. Technol.*, vol. 8, no. 11, pp. 285–296, Nov. 2015.
- [7] A. Fischer and C. Igel, "Training restricted Boltzmann machines: An introduction," *Pattern Recognit.*, vol. 47, no. 1, pp. 25–39, Jan. 2014.
- [8] "Face Recognition-Based Mobile Automatic Classroom Attendance Management System - IEEE Conference Publication." [Online]. Available: <https://ieeexplore.ieee.org/document/8120338>. [Accessed: 13-Dec-2019].
- [9] M. M. K. Kavita, "A Survey paper for Face Recognition Technologies," vol. 6, no. 7, p. 5, 2016.
- [10] S. Bajpai, A. Singh, and K. V. Karthik, "An Experimental Comparison of Face Detection Algorithms," *Int. J. Comput. Theory Eng.*, pp. 47–51, 2013.