IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 5, May 2020

Inculcating the Application of Web Mining in E- Mail Spam and Traveling Salesman Problem

Dr. S. Prema[1], Dr. K.M. Sharavana Raju[2], C. Supriya[3], P. Gayathri[4]

Assistant Professor, Department of Computer Science (PG), K.S.Rangasamy College of Arts and Science[1]

Department of Computer Science, College of Computer Science & Information Technology,

JAZAN University, Kingdom of Saudi Arabia[2]

M. Sc (CS), Department of Computer Science (PG), K.S.Rangasamy College of Arts and Science[3,4]

Abstract: Web Mining is playing a major role in Networks. This paper reflects the project work of two important problems identified during master degree. The first case study is about spamming. E-mail spam. Spammed content may contain many copies of the same message. In existing work, various filtering techniques are used to detect these e-mails such as Random Forest, Naive Bayesian, Support Vector Machine (SVM) and Neutral Network. In this project work Naive Bayes algorithm is chosen for e-mail spam filtering. Two datasets Spam Data and SPAMBASE datasets was selected for this project work. The execution of the datasets is evaluated based on their accuracy, recall, precision and F-measure. This work use VB.Net and C# for the implementation of Naive Bayes algorithm for e-mail spam filtering on both datasets. The outcome shows that the type of email and the number of instances of the dataset has an influence towards the performance of Naive Bayes.
The second case study is about finding optimization solution for shortest path. MLT (Machine Learning Techniques) finds potentially useful patterns in the data. The famous example is probably the Traveling Salesman Problem (TSP) in which a salesperson intends to visit a number of cities exactly once, and returning to its starting point, while minimizing the total distance traveled or the overall cost of the trip. TSP is used in combinatorial optimization. This paper proposed genetic algorithm based cuckoo search technique for TSP. Result significantly improves the performance of the TSP. The algorithm is implemented using MATLAB.

Keywords: Naive bayes algorithm, Genetic algorithm, Cuckoo search.

I. INTRODUCTION

Case study1:
Spam brings financial damage to companies and frustrating individual users. Spam filter is used to find the spam message. Spam filters can take various roles including content filtering, blacklist filtering, malware filtering, virus detection, etc. Several machine learning algorithms have been used in spam e-mail filtering, but Naive Bayes algorithm is mainly popular in business and open-source spam filters. This is because of its plainness, which make them easy to implement and just need short instruction time or fast evaluation to filter email spam.
The Project objectives are: (i) to implement the Naive Bayes algorithm for e-mail spam filtering (ii) to evaluate the performance of Naive Bayes algorithm for e-mail spam filtering on the chosen two datasets.
Solution of the Problem:
To solve the problem of previous study Naive Bayesian Classifier is used in this project to classify the spam and non-spam mails. Text pre-processing based on regular expression and spam message feature extraction based on word segmentation and the TF-IDF algorithm. The accuracy of system increases using Naïve Bayesian Classifier.

Case study2:
Data mining is used to explore the contents from large data. Machine learning (ML) models that power modern artificial intelligence (AI) applications such as search engine algorithms and recommendation systems. The Project objectives are: (i) to implement the Naive Bayes algorithm for e-mail spam filtering (ii) to evaluate the performance of Naive Bayes algorithm for e-mail spam filtering on the chosen two datasets.
Solution of the Problem:
To solve the problem of previous study Naive Bayesian Classifier is used in this project to classify the spam and non-spam mails. Text pre-processing based on regular expression and spam message feature extraction based on word segmentation and the TF-IDF algorithm. The accuracy of system increases using Naïve Bayesian Classifier.
Data mining is used to explore the contents from large data. Machine Learning (ML) models that power modern artificial intelligence (AI) applications such as search engine algorithms and recommendation systems.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 5, May 2020

II. LITERATURE REVIEW

Case study1:

Learning-based methods of spam filtering: The most popular method for anti-spam technique is spam filtering according to the study of Mikko Siponen and Carl Stucke (2006). Spam filtering classifies the messages into spam and legitimate email. Existing filtering algorithms have quite affective results even close to 90% accuracy and it was found that integrating different learning algorithms actually seems to be a promising way (the evaluation performed by Lai & Tsai, 2004). Spam filtering is an application which implements a function with binary output, spam or legitimate. Machine learning classification techniques are the main type for the spam filters.

Naive Bayes: Naive Bayes Classifier is the mainly used classifier in spam filtering (Pantel & Lin, 1998; Sahami, Dumais, Heckerman & Horvitz, 1998). After Paul Graham's 'A plan for spam' (Graham) article it becomes widely known method. This can be mainly classified as a learning based keyword filter when used for the text content. Bayesian method uses d dimensional x vectors to classify the email as spam or legitimate. Here d is the independent features of x, used for estimating the probabilities the email classification. Several variants of Naïve Bayes were applied to spam filtering, an overview and comparison of them can be found in the article by Metsis et al. (Metsis, Androutsopoulos & Paliouras, 2006).

Web spam: Web spam which is a major issue throughout today's web search tool; consequently it is important for web crawlers to have the capacity to detect web spam amid creeping. The Classification Models are designed by machine learning order algorithm. [2] The one machine learning algorithm is Naïve Bayesian Classifier which is also used in [1] to separate the spam and non-spam mails. Big Data analyzing framework which is also outline for spam detection. Extricating the feeling from a message is a way for get the precious data. Paul Graham's Naive Bayesian Machine learning approach is used to develop the competence of Bayesian approach.

Case study2:

Rohit Chaturvedietal (2014) proposed a competent method which improves ACO in terms of iteration count and ability to find better solutions for TSP so that it can be used in different locales like industrial and educational, for solving NP problems more efficiently. This paper proposes a modified ant colony optimization (MACO) algorithm which uses the peculiarity of Elitist Ant System (EAS) and Ant Colony System (ACS). Using EAS property, the union speed is optimized by added pheromone evidence on the arcs of the best tour and pseudorandom relative rule and local pheromone update of ACS tunes the level of examination and prevents the algorithm from stagnation. The experiments done on benchmark datasets from TSPLIB manifest clearly that MACO has an upper hand in terms of performance on conventional ACO, ACO-GA and ACO-PSO. This paper proffer how an improved ant colony algorithm can solve travelling salesman problem competently and thus help in taking up other NP complex problems without any hitch. The improvements centralize on pseudorandom proportional rule and local pheromone update used in Ant Colony System and additional pheromone deposition to the arcs belonging to the best tour used in Elitist Ant System. Even though traveling salesman problem (TSP) is used here, this algorithm can be applied to other optimization problems which occur in industrial and educational environments where computational resources and time are limited.

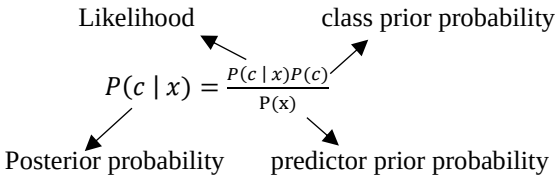
Ms. RinkyDwivedietal (2014) focused the efficiency of machine learning algorithm, ACO has been improved for solving Travelling Salesman Problem (TSP) in finding optimal path by varying its parameters (α and β explained later) using Fuzzy inference system. TSP is a NP hard combinatorial optimization problem which means no algorithm is known to solve it in polynomial time. Ant colony optimization algorithm is first applied and then studied for the parameters that highly affect its performance. When altered slightly, it appears as a sub-problem in many problems. One such application is that of DNA sequencing. Furthermore in the cases when the graph may change dynamically, the ant colony algorithm can be run incessantly and acclimatize to the alterations in real time. This lures the researchers in the field of network routing and urban transportation systems to optimize the performance of ACO.

III. STUDY WORKS

Case study1:

Naive bayes classifiers are a collection of classification algorithms based on bayes' theorem. It is not a single algorithm but a family of algorithms where all of them all of them share a common principle, i.e. each match up of features being classified is independent of each other. It is a classification method based on bayes' theorem with a hypothesis of liberty along with predictors. In plain conditions, a naive bayes classifier presumes that the existence of a particular attribute in a class is unconnected to the occurrence of any other attribute. Along with simplicity, Naive Bayes is known to outperform highly sophisticated classification methods. Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 5, May 2020



$P(c \mid x) = P(x_1 \mid c) \times P(x_2 \mid c) \times \ldots\ldots \times P(x_n \mid c) \times P(c)$

- $P(c|y)$ is the posterior probability of *class* given *predictor* (y, *attributes*).
- $P(c)$ is the prior probability of *class.*
- $P(y|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(y)$ is the prior probability of *predictor.*

How Naive Bayes algorithm works?

Every day, we collect tons of unrelated emails, some of which are plainly business emails while others are fake ones. The peak tech companies like Google, etc. spend a lot of manpower in developing tough spam classifiers simply because customers are annoyed with such emails. Here is an example spam email:

Dear Mr. Aman,

You have won a lottery offer for $2,000,000!!! Click here to claim it now.

Clearly, this is a spam email for most cases. The dispatcher of such email needs you to hit it off on the link so that the sender could fool the user by:

Capturing your personal details.

Making you download a malware/virus.

Spam detection problem is therefore quite important to solve. More formally, are given an email or an SMS and are required to classify it as a spam or a no-spam (often called ham).

Case study 2:

Enhanced genetic algorithm inspired cuckoo search (gacs) algorithm travelling salesman problem: MLT finds potentially useful patterns in the data. Optimization problems, either single-objective or multi-objective, are generally difficult to solve. Many of them are said to be NP-hard and cannot be resolved proficient by any famous algorithms in a suitable quantity of time. The most famous example is probably the traveling salesman problem (TSP) in which a salesperson intends to visit a number of cities exactly once, and returning to its starting point, while minimizing the total distance travelled or the overall cost of the trip. TSP is one of the most widely studied problems in combinatorial optimization. The TSP and its variants have several important applications, such as, drilling of printed track panels, x-ray crystallography, processor electrics, car steering and preparation, and manage of robots, among others. Therefore, solving this class of troubles is both of academic interest and practical importance, and consequently it has been an important topic of active research. After the proposed strategy is carried out, results are recorded and compared.

IV. PROPOSED METHOD

Case study 1:

In this project naive bayes is applied to spam detection problem**.** In the spam discovery difficulty, there are 2 courses: C1 which is the no-spam class and C2 which is the spam class. X is fundamentally each electronic mail current in the guidance data. To convert X into a machine-readable form, it is basically needed to convert X into a vector:

Create an ordered list of all the words in the vocabulary. Convert an email into a vector, map out the figure of times each word happens in that email. Each word is produced self-determining of each other.

Consider the supervised Machine Learning:

Given:

A document d

A set of classes C = {c1, c2, …. , cn}

A teaching set of m papers that we have prearranged to fit in to an exact group.

Now train classifier using the training set, and result in a learned classifier. Then make use of this educated classifier to classify new papers. Use Y(d) = C to represent classifier, where Y() is the classifier, d is the document, and c is the class we assigned to the document. This is a easy classification scheme based on bayes regulation. It relies on a very simple representation of the document.

Precision, Recall & F-measure: We need more accurate measure than contingency table (True, false positive and negative) as talked in my blog "Basics of NLP". **Precision:** % of selected items that are correct. Tp (True Positive) / Tp + Fp (False Positive). **Recall:** % of correct items that are selected. Tp / Tp + Fp. There is a tradeoff between precision &

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 5, May 2020

recall. A standard way proposed to combine this measure is **F-measure**. F-measure is weighted harmonic mean. Mostly balanced F measure (F1 measure) is used.
F1 = 2PR / P+R
F = 1 / (k/P + (1-k) / R) = (B*B + 1)*R / (B*B*P + R)
For F1, put B(Beta) = 1
Depending on the domain we are working with, we can do things like:

Collapse part number or chemical names into a single token.

If have a sentence that contains a title word, we can up weight the sentence (multiply all the words in it by 2 or 3 for example), or we can up weight the title word itself (multiply it by a constant).

Choosing Classifier

If there is No data → Handwritten rules!

Less training data → Naive Bayes is best

Reasonable amount → SVMs & Logical Regression can also use decision trees.

V. ALGORITHM

Case study 1:
The naïve Bayes algorithm has good performance regarding the problems of classification and high latitude. Focusing on the problem of the multiple classifications of spam messages, a multi-classification naïve Bayes spam message model is presented.

```
test_data = testset// Test data set
train_data = trainset // Training data set
fori = 1:N
// Number of spam message classification
fobj = file.open(file(i))
// Read text preprocessed data of each classification
while True:
raw = fobj.readline() // Read each message
if raw:
word_cut = jieba.cut(raw)
// word segmentation
If test_data.length>0.3*fobj.length:
//Judgment, test set 30%, Training set 70%
train_data.append(word_cut,i)
//Add data for Training set, word segmentation
+classification
else:
test_data.append(word_cut,i)
//Add data for test set, word segmentation +classification
else:
break
word_features = get_features()
// Read TF-IDF feature value
test_data = document_features(test_data)
// Test set document vector process
train_data = document_features(train_data)
// Training set document vector process
classify = NaïveBayesClassifier.train(train_data)
// Carry out classification training for Training data set
classify.test(test_data)
// Carry out test inspection for classification model
```

Case study 2:
Basic Rules of Cuckoo Search:

1) Each cuckoo lays one egg at a time, and dumps it in a randomly chosen nest.

2) The best nests with high quality of eggs will carry over to the next generations.

3) The number of available host nests is fixed, and a host can discover an alien egg with a probability pa [0,1].In case, the bird can either throw the egg away or abandon the nest so as to build a completely new nest in a new location.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 5, May 2020

Genetic Algorithm:

Step 1: Start with a randomly generated population with Mutation, =0.01 and Crossover, =0.99 Generate initial population of n host nests xi (i = 1, . . . , n)
While (t <MaxGeneration) or (stop criterion) do
Step 2: Assess the fitness value of each individual in the population. The fitness may range from 0 to 1.
Step 3: Cuckoo Search, calculate Objective function f (x), x = (x1, . . . , xd) T Get a cuckoo randomly by Levy flights and Evaluate its quality/fitness Fi
Choose a nest among n (say, j) randomly
if (Fi > Fj) then
replace j by the new solution;
end if
Step 3: Select persons to replicate based on their robustness given. Compute the standard robustness of all value
Step 4: Apply crossover with probability
=0.99
Step 5: Apply mutation with probability
=0.01
A fraction of poorer nests are discarded and new ones are built; Keep the best answers; Rank the solutions and find the current best
end while
Post process results and visualization
Step 6: Restore the inhabitants by the new production of individuals after the assessment
Step 7: Go to step 2

VI. CONCLUSION

Case study 1:

This work puts forward a spam message classification model based on the naïve Bayes algorithm and estimates the performance of the naïve Bayes algorithm model based on binary classification in spam message using spam message text preprocessing based on regular expression and spam message feature extraction based on word segmentation and the TF-IDF algorithm. The experiment results show that the binary classification naïve Bayes algorithm has the best efficiency. With the development of big data technology, how to perform feature extraction and text classification for bulk data on the basis of big data computation will be the future work direction.

Case study 2:

The success of these methods to solve various problems such as the combinatorial optimization problems depends on many factors, ease of implementation, ability to consider specific constraints that arise in sensible submissions and the high quality of the results they produce. Therefore, there is no specific algorithm to solve all optimization problems. This research makes a significant contribution with novel GACS algorithm and from the results it is clearly evident that the GACS performs well.

REFERENCES

Case study 1:

[1] C. Huang, "Research on SMS filtering technology on intelligent mobile phone," M.S. thesis, Huazhong University of Science and Technology, Wuhan, China, 2012.
[2] J. Ma, Y. Zhang and Z. Wang, "A message topic model formultae grain SMS spam filtering," International Journal of Technology and Human Interaction (IJTHI), vol. 12, no. 2, pp. 83-95, 2016.
[3] S. J. Delany, M. Buckley and D. Greene, "Review: SMS spam filtering: Methods and data," Expert Systems with Applications, vol.39, no. 10, pp. 9899-9908, 2012.

Case study 2:

[1] Z. W. Geem, J. H. Kim, et al., A new heuristic optimization algorithm: harmony search, Simulation 76 (2001) 60–68.
[2] X. S. Yang, Firefly algorithms for multimodal optimization, Stochastic algorithms: foundations and applications (2009) 169–178.
[3] H. Shah-Hosseini, The intelligent water drops algorithm: a nature-inspired swarm based Optimization algorithm, International Journal of Bio-Inspired Computation 1 (2009) 71–79.
[4] X. S. Yang, S. Deb, Cuckoo search via l´evy flights, in: Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on, IEEE, pp. 210–214.
[5] E. D. Taillard, L. M. Gambardella, M. Gendreau, J. Y. Potvin, Adaptive memory programming: A unified view of metaheuristics, European Journal of Operational Research 135 (2001) 1–16.