International Journal Of

Recent Scientific Research

ISSN: 0976-3031

Volume: 7(6) June -2016

TEXT CLASSIFICATION USING ITERATIVE DICHOTOMISER (ID3) ALGORITHM

Shruthi E and Deepika N

THE OFFICIAL PUBLICATION OF
INTERNATIONAL JOURNAL OF RECENT SCIENTIFIC RESEARCH (IJRSR)
http://www.recentscientific.com/ recentscientific@gmail.com

ISSN: 0976-3031

Available Online at http://www.recentscientific.com

International Journal of Recent Scientific Research
Vol. 7, Issue, 6, pp. 12153-12157, June, 2016

International Journal of Recent Scientific Research

Research Article

TEXT CLASSIFICATION USING ITERATIVE DICHOTOMISER (ID3) ALGORITHM

Shruthi E[1] and Deepika N[2]

[1,2]Department of Computer Science and Engineering New Horizon College of Engineering Bangalore, India

ARTICLE INFO

Article History:

Received 05th March, 2016
Received in revised form 21st April, 2016
Accepted 06th May, 2016
Published online 28th June, 2016

Key Words:

WhatsApp Messenger, ID3 Algorithm, Machine learning, Decision tree.

ABSTRACT

WhatsApp Messenger is cross-platform messaging, where text, video, images etc can be sent to anyone by just having the contact of the person. Today WhatsApp Messenger usage is increasing day by day. Hence an approach to classify the text into bizarre (abuse) or customary using machine learning algorithm. The decision tree algorithm Iterative Dichotomiser ID3 is used to learn WhatsApp text and evaluated to measure the accuracy. This paper explains an approach to classify the WhatsApp text messages into various categories that give an insight of real world application service. The paper explains that WhatsApp is a multi tool used for personal, communication, business etc.

INTRODUCTION

The ID3 algorithm is used in various fields to classify large data into various category. Fields like in medical research, facebook, twitter, email etc. Hence a new approach to use machine learning algorithm to classify WhatsApp text data. Since WhatsApp application is used by everyone for their personal, professional, communication and even for business.
This paper explains how well the algorithm is able to classify the dataset into various categories. In decision tree the nodes are the attributes of the dataset and the leaf node is the targets attribute which give the output of the instance. ID3 uses top-down, greedy search through the space of all decision trees. It used statistical property called Entropy and information gain to select the attribute to be node of the tree. [12]

LITERATURE SURVEY

WhatsApp is an application that is use through mobile software and desktop with has restricted features. Since it is use in large population large data is collected, hence an effort to classify the text data of WhatsApp in to various category.[14][15]

S. Appavu alias Balamurugan and Ramasamy Rajaram, [1] Suspicious E-mail Detection via Decision Tree: A Data Mining Approach, that classify the email as deceptive or not.

Suhas Pandhe, Sahil Pawar, Volume[2] Algorithm to monitor Suspicious activity of social networking sites using data mining approach: Proactive posts related to renowned people, sexuality, countries and any related topic.

Marco Vanetti, Elisabetta Binaghi, Elena Ferrari, Barbara Carminati, and Moreno Carullo, [3]A System to Filter Unwanted Messages from OSN User Walls. Online social network (OSN) users to have a direct control on the messages posted on their walls.

Twitter content classification by Stephen Dann.First [4] new classification framework that offers a deeper insight into Twitter content for analyzing individual timelines.

K. Saruladha, L. Sasireka [7] explains the problem of Spam and gave an overview of Text Classification algorithms based on Spam filtering techniques. Rule based classifiers, linear classifiers and example based classifiers were briefly discussed. Machine learning classifiers such as J48, Alternate Decision tree, Decision Stump, Boosting Algorithms, Naïve Trees, CART can also be used in Spam detection, to enhance the accuracy and good performance evaluation.

METHODOLOGY

Preprocessing of Whatsapp Data

A. Extracting data from WhatsApp

The large data from the WhatsApp Messenger is extracted is many ways. First the text messages, calls, videos etc. Are Encrypted and stored in WhatsApp folder.

Using Tools: There are many tools that is really available to extract the data from the WhatsApp. To read the message you need to root your device in order to decrypt it. There are many tools available to root and decrypt the encrypted file.

Corresponding author: **Shruthi E*
Department of Computer Science and Engineering New Horizon College of Engineering Bangalore, India

Email the conversation of the chat: Another very easy method that is provide by WhatsApp is that you can mail your chat message to your mail id.

Once you have the large messages from the WhatsApp, next we need to remove the unwanted inconsistent data before converting it to the attribute values of the dataset. Hence the unwanted words are moved from the sentences before converting to values.

B. Features selection from sentences

The following features are selected from each sentence if it's present.

Customary words: customary words are the goods or good words that are present in the sentences. Each sentence is scanned for the customary words by comparing each word in the sentence with that of customary words provide. Like, good, happy are the some of the customary words considered.

Converting to value as below

Number for good words/total number of words in the sentence.

Bizarre words: Words that bizarre (abuse) that are present in the sentences. Each sentence is scan for the bad words by comparing each word in the sentence with that of bizarre words provide. Attack, kill, hate are some of the bizarre words considered.

Converting to value as below

Number for bad words/total number of words in the sentence.

Capital words: The uppercase letters that is present in the sentences. Each word in the sentence is checked for uppercase letter if the capital letter is present and is greater or equal then half of the word, is consider. This is repeated for all the words in the sentence.

Converting to value as below.

The total number of words consider as explained above / total number of words in the sentence.

Numbers: The number if present in the sentence. Each word in the sentence is checked for numbers if the number is present and is greater or equal then half of the word, is consider. This is repeated for all the words in the sentence.

Converting to value as below.

The total number of words consider as explained above / total number of words in the sentence.

Punctuation: Punctuation is symbols other than alphabets and numbers like comma, pullstop, brackets that are used to clarify the meaning. Each word of the sentence is evaluated for presence of punctuations and repeat for the complete sentence.

Converting to value as below.

The total number of punctuation / total length of the sentence.

Links: Any universal links in the sentences. In WhatsApp even location, links of the website can be sent. Hence the sentence is checked for such links by evaluating the presence of http or www which means it's a link. If the sentence contains links then it will return one otherwise zero.

Exclamation: A punctuation mark (!) is considered as exclamation, the exclamation is usually used to express strong feeling. Hence each sentence is checked for exclamation mark if present.

Converting to value as below.

The total number of exclamation marks/total length of the sentence.

Questions Marks: A punctuation mark (?) is consider as question mark, the question mark is usually used in place of query, missing or unknown data .hence each sentence is checked for question mark if present,

Converting to value as below.

The total number of question marks/total length of the sentence.

Hence all this features are looked for each sentences or text of WhatsApp. If any feature is present it is converting to values or instances as explained above.

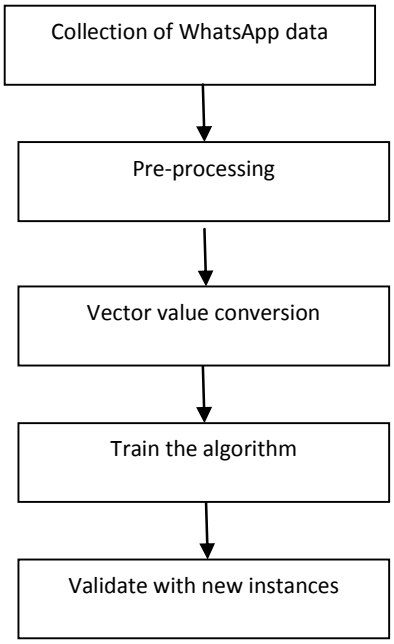

Fig. 1 System Architecture

Vector Space Model

A. Example: consider the sentence.

- Good morning, how are you?
- Customary Word: 0.166.
- Bizarre Word: 0.0.
- Capitalletter: 0.0.
- Punctuation Marks: 0.1466.
- Exclamation Marks: 0.0.
- Question Marks: 0.0.
- Numbers: 0.0.
- Links: 0.0.

This forms the vector space model. With form the dataset to train the algorithm.

Classifying the Sentences into Various Classes

Conversation class

Since WhatsApp is use for communication between the users and in conversation message @symbol is use to refer another user while communicating in WhatsApp group. Conversation message can be query, action, referral, response.

Example

- What time does the match start?
- www.indiagottalent.in check this link video it's very nice.
- @Raj please send me the mail.

Status class

WhatsApp also contain short texts which give the present doing or situation of the user. Status can be personal positive or negative thoughts, or location status that indicate user traveling, or mechanical status that is related to technology, work status that indicate user of his/her business work or activity status the present work done by user.

Example

- Think positive
- I liked modest house after they became famous.
- Waiting for my 3pm performance review.
- Spend 3.5mb out of 6 GB.
- In movie
- Happy birthday

Phatic class

Phatic means it refers to basic interaction rather than convey or asking question, expressing feelings, goodwill rather than information.

Examples

- You are welcome
- Thank you

Spam class

Since WhatsApp have feature of sending links video which can spread malware on downloading or opening. Some links can contain cookies that will help to track the user interest.

Examples

- www.cookies.com
- www.devial.com

Algorithm

Decision tree algorithms perform top down greedy search through set of decision tree present in the hypothesis, a tree is generated for the learnt function, where the node represent the attribute and values of the dataset and leaf node represent the class, decision tree is robust in nature to noisy data which means can be use for missing values. Decision tree can have more than one class or cluster. [5][11]

ID3 algorithm

ID3 generate the tree for the given dataset by doing statistical tests for each attribute to find which attribute will best classify the instances, that attribute is made as the root of the tree, the subtree is then created for the values of the root attribute. The process is then repeated with each decedent node.[9]

How the best attribute is selected:[6]

The best attribute is the one that classify maximum instances, hence we use statistics method to select best attribute. To understand information gain we need to know another method called entropy.

Entropy

Entropy measures the impurities, the higher the entropy the higher the uncertainty, entropy give the average amount of information that is present in the sample.

$\text{Entropy}(S) = \sum_{i=1}^{k} p_i \log_2 p_i$

S denotes the dataset. K denotes the number of output variable classes, and Pi the possibility of the class i.

Information gain

Information gain measures the expected reduction in entropy, the information gain give the accuracy of the attribute in classify the instances hence the attribute that as the highest gain is selected as the node of the tree.

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Sv represents the subdivision of dataset S which contains value v in S.

Steps of the ID3 algorithm

id3(dataset, target_attributes, attribute)

- dataset: the examples that is use to train the algorithm.
- traget_attribute: the classes that are algorithm should output from the tree.
- attribute: other attributes that are used for classification.
- If there is only one class then the output is a tree with a single root node label with that class.
- If the attributes is empty, then output a tree with root with the class which classify maximum instances.
- If the above two condition do not hold, then repeat the below steps.
- Find the information gain of all the attributes, by using the above formula.
- Select the attribute which as the highest gain as the root of the tree.
- For each value of the root, create a branch node below root.
- Now consider the subset of dataset that have the value of the root.
- If no such subset dataset is present or the subset of dataset is empty then,
- Create leaf node below the branch node with the class name that classify maximum instances.
- Otherwise below the branch node add a subtree, with ID3 (dataset having the value of branch node, target_attribute, attribute-attribute selected as the root).
- Return the tree.

- This process is repeated until all the attributes are covered or all the dataset is covered.

After the ID3 is trained with the dataset, next instance is given with unknown class, and the algorithm should classify the instance based on the generated tree, the class of the instance is predicted but travelling through the tree from top down. In our paper we are using WhatsApp dataset, hence after training with the vector space model, some sentences are given to the algorithm to classify into four classes that is considered. [11],[13].

Experimental Results

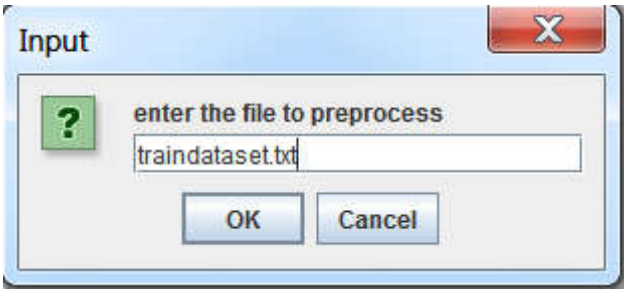

Fig. 2 Entering the Input Dataset File

The file to preprocess is the text file, which is preprocessed as explained above, after the preprocessing the text data is converted to vector space value as shown below.

After training new set of dataset is given called as validating dataset with is use to measure the accuracy of the algorithm, it is found that the accuracy is increased as the algorithm learns all new dataset.

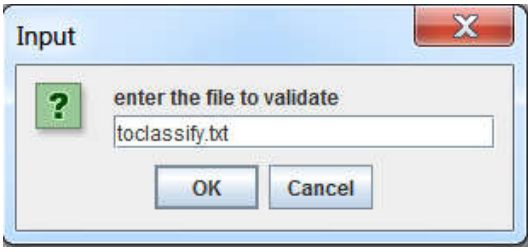

Table 2 Time taken to execute different size dataset.

Number of sentences	Time(milliseconds)
10 sentences	5975ms
20 sentences	6287ms
80 sentences	11897ms

CONCLUSION

This paper explains how ID3 decision tree algorithm can be used to classify the WhatsApp text data, and the paper also explains how the text sentences are converted to vector values, the classes that are classified can be used to add more features to the application and in future it can also be used to analysis the risk of suspicious messages sent by unauthorized users.

Table I Example of the dataset of whatapp text data.

Target_Attribute	Good	Bad	Capital	Punctuation	Exclamation	Question	Numbers	Links
Conversation	.35	0	0.0	0.0	0.5	0.0	0.5	0.1
Phatic	0	0.1	0.2	0.5	0.6	0.1	0.9	0.0
Spam	0.04	0.5	0.6	0.9	0.8	0.3	0.12	0.1
Status	0.15	0.13	0.0	0.13	0.4	0.5	0.16	0.1
Conversation	0.16	0.18	0.0	0.5	0.3	0.2	0.0	0.0
Spam	0.18	0	0.0	0.6	0.1	0.0	0.0	0.0
Spam	0.0	00.1	0.0	0.0	0.0	0.0	0.0	0.0

This vector values is given input to the ID3 algorithm and the output of the algorithm is the hypothsis of decision trees, as showen below.

```
      =spam
0.08
  =ham
0.09
  =ham
0.18
  pun--->
   0.11
     =ham
   0.02
     =spam
   0.03
     =spam
   0.04
     =spam
0.15
  =ham
0.16
  qmark--->
   0
     =spam
   0.01
     =ham
0.29
```

Fig. 3 Example For Decision Tree Generated By The Algorithm

The algorithm classify the text into spam, if it is not spam [10] then it classify the text into conversation class, status class and phatic class.[8]After the algorithm is trained with the set of vector space values then the algorithm is validated with the different set of sentences called as validating set to test the accuracy of the algorithm to classify the unknown class.

References

1. S. Appavu alias Balamurugan and Ramasamy Rajaram, *Suspicious E-mail Detection via Decision Tree: A Data Mining Approach, Journal of Computing and Information Technology* - CIT 15, 2007.
2. Suhas Pandhe, Sahil Pawar, *Algorithm to monitor suspicious activity of social networking sites using data mining approach*, *International Journal of Computer Applications* (0975 –8887)Volume 116 –No. 12, April 2015
3. Marco Vanetti, Elisabetta Binaghi, Elena Ferrari, Barbara Carminati, and Moreno Carullo, *A System to Filter Unwanted Messages from OSN User Walls*,IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 2, February 2013.
4. Stephen Dann. First, *Twitter content classification*.
5. Tom M. Mitchell, *Machine Learning*, McGraw-Hill Science/Engineering/Math;(March 1, 1997) ISBN:0070428077

6. N. Deepika,Dr. N GuruPrasad, *Empirical Study of Combinatorial Learning Method for Data Clustering*, IJIFC, Vol.3, Issue.4, Dec.2015.
7. K. Saruladha And L. Sasireka, *Survey Of Text Classification Algorithms For Spam Filtering,* I J I T E, ISSN: 2229-73673(1-2), 2012.
8. Shahar Yifrah, Guy Lev, *Machine Learning Final Project Spam Email Filtering*, March 2013.
9. Kasra Madadipouya, *A New Decision Tree Method For Data Mininging Medicine*, Advanced Computational Intelligence: An International Journal (ACII), Vol.2, No.3, July 2015.
10. Aman Kumar Sharma, Suruchi Sahni, *A Comparative Study of Classification Algorithms for Spam Email Data Analysis*, IJCSE.
11. Wikipedia website [Online]Available: https://en.m.wikipedia.org/wiki/Decision_tree
12. (2002) The IEEE website. [Online]. Available: http://www.ieee.org/
13. www.cise.ufl.edu website.
14. WhatsApp website[Online]Available: https://www.whatsapp.com
15. Wikipedia website [Online] Available: https://en.m.wikipedia.org/wiki/WhatsApp.

****

How to cite this article:

Shruthi E and Deepika N.2016, Text Classification Using Iterative Dichotomiser (Id3) Algorithm. *Int J Recent Sci Res.* 7(6), pp. 12153-12157.

TSSN 0976-3031