International Journal of Future Generation Communication and Networking
Vol. 13, No. 4, (2020), pp. 275–283

Predicting demographics features of the SNS by using Amalgamative Classification Algorithm (ACA)

[1] Kotaiah Swamy Kakulla [2] Gouse Baig Mohammad [3] Ramu Kuchipudi

[1, 2] Assistant Professor, Department of CSE, Vardhaman College of Engineering, Shamshabad, Hyderabad, T.S.
[3] Associate Professor, Department of CSE,Vardhaman College of Engineering,Shamshabad,Hyderabad, T.S.

Abstract

The internet has become unmanageably big and day by day it is increasing exponentially mainly through social media, blogs and reviews. Most of this information is written by various authors in different contexts. The availability of information put a challenge to researchers and information analysts to develop automated tools for analyzing such information. In this regard, Author Profiling is a popular technique attracted by several researchers to extract as much information as possible from the texts by analyzing author's writing styles. Writing styles are many types such as messages in social networking sites such as facebook timeline messages, twitter posts etc. Based on the given social networking data of every user it is easy to predict the demographics features of the various users. In this paper, a new amalgamative classification algorithm (ACA) which classify the various types of users with different features and their ideas on latest trends. To improve the performance of the proposed system pruning is the technique adopted. The dataset used in this paper is facebook dataset for classification. Results show the performance of proposed system.

Keywords: pruning, demographics, classification, OSN.

Introduction

The data analysis is the way toward analyzing the crude information and applying factual or consistent procedures to outline and assess the information to know decisions about the data. It encourages us to settle on a superior choice and to confirm the current speculations and models. The primary motivation behind breaking down the information is to secure fundamental data, paying little respect to whether the data is abstract or quantitative. Utilizing information examination we will almost certainly relate and condense the information, perceive associations among elements and dissect factors with one and the other.

ISSN: 2233-7857 IJFGCN

The information gained from Facebook is the subjective sort of information that depicts something in words, for example, sentiments, feelings or abstract view of something.

Social media platforms offer (close) continuous spontaneous data on clients' musings, sentiments, and encounters, enabling specialists to follow demeanors and practices as they rise and after some time. The individual and populace size of this information empowers the investigation of the progression of data through complex systems that are hard to evaluate utilizing customary methods for information gathering [1]. This information are likewise generally ease contrasted with conventional methods of information accumulation, for example, reviews.

The demographics is the social networking sites of every stage is a basic advance to take before figuring out which stages bode well for your image to use. Every social media life system comes fit as a fiddle and size, with its very own substance methodology and client base, so there's ordinarily nobody size-fits-all procedure. You would prefer not to burn through your time making content for a stage where your group of spectators doesn't really have an enormous nearness. Rather, you need to contribute your important time advancing the correct sorts of substance for each applicable stage [2]. On the off chance that that is on facebook, you'll need to invest energy making and curating illustrations and pictures to share, etc. Adjusting your substance procedure for every stage your group of spectators uses is an extraordinary method to connect with them. Also, stage one is realizing who involves a group of spectators inside those stages [3].

In this paper, the random forest (RF) is adopted with pruning called as ACA

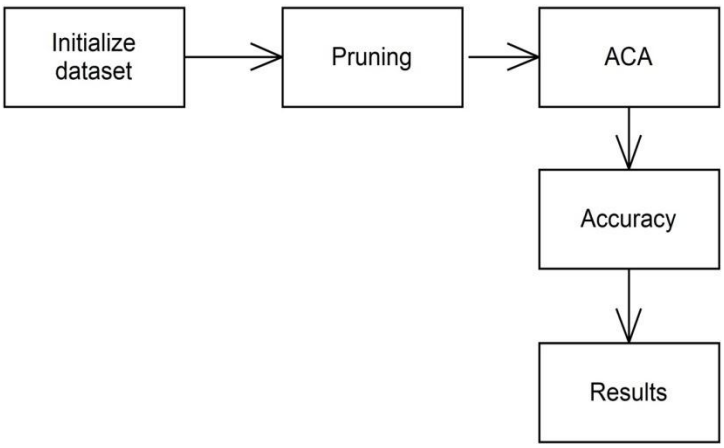

Figure: 1 System Architecture

Literature Survey

ISSN: 2233-7857 IJFGCN

The literature survey and progress on the data investigation on Facebook profiles. We spread these themes into two classifications forecast of phony profiles utilizing AI systems and supposition extraction and arrangement of the Facebook status. The canvassed subjects are accounted for in the middle of 2012 and 2016 Facebook is an online social media network (OSMN) organize which is broadly utilized over web, where the client needs to enroll first utilizing his/her email-id or portable number and make a record, next they can make companions and talk with them on the web, share recordings, messages, posts statuses, and so forth they can likewise mess around and keep up fan pages and make a gathering among their companions and have a dialog. It is a noteworthy system where we can discover tremendous arrangements of information for examination [4].

All the fake profiles on Facebook are predicted, it alludes to the theoretical model which distinguishes phony profiles of Facebook utilizing distinctive AI calculations to discover the clients profile whether it is a phony or genuine one. The proposed model applies different standardization methods on the datasets, and a system has been utilized to perceive the invalid traits in datasets and deduct them in like manner by applying certifiable procedures. This model uses distinctive machine calculations for both the informational indexes independently, for example, genuine and counterfeit and furthermore, by utilizing Ensemble Classifier the model can discover all the more precisely. Also, cross-overlap approval is utilized to locate the various methods for information where the information is isolated into preparing and testing, and every gap part is called crease [5]. The execution of a considerable number of folds is half of the calculation.

CLASSIFICATION USING DECISION TREE

The classifier which very easy to implement is called Decision Tree (DT). The main feature of the DT is this will extract the decision in depth. DT constructs the classification in the form of structure this will makes easy to debug and handle. DT can handle any type of data such as categorical and numerical data. This will find the information gain from the attributes and extracting the attributes for dividing the branches in trees. From the bellow steps, the process of DT explained [6].

Eq.(1). E(S) = -P (P) log2P (P)-P (N) log2P (N)

(1) The algorithm steps as follows:

ISSN: 2233-7857 IJFGCN

Step 1: From the dataset the data gained for the attributes.

Step 2: data gain is done by sorting for the selected datasets in descending order.

Step 3: After the recognition of the data gain assign the best attribute of the dataset at the root of the tree.

Step 4: With the same formula the information gain is done.

Step 5: The nodes are divided based on highest data gain.

Step 6: This process will be repeated until every attribute are set as child nodes.

Amalgamative Classification Algorithm (ACA)

In this paper, Amalgamative Classification Algorithm (ACA) is the combination of random forests (RF) and pruning technique which is used to predict the DT such that every tree depends on the random vector simplified individually and with the same distribution for all trees in the forest. This algorithm consists of various trees in different forests and the strength of every individual tree in the forest and the correlation between them. Due to large number of trees, the prediction of values is complicated to get the accurate results. To improve the performance pruning will help the RF for accurate results for large datasets and also dataset used in this paper [7].

The algorithm steps as follows:

Step 1: From entire y features select the x features randomly from the datasets, where x << y.

Step 2: From the nearest x features, calculate the node “d” using the best split point.

Step 3: divide the node into child nodes using the best split.

Step 4: the steps 1 to 3 are repeated.

Step 5: the forest is constructed by repeating steps 1 to 4 for n number times to create n number of trees.

Role of Pruning in ACA

If the dataset contains the complicated data or huge datathen the pruning will help us to get better results. And this will used to limit the no of trees in RF is called as Pruning of

ISSN: 2233-7857 IJFGCN

Random Forest. This will generate the most efficient random forest for both learning and classification.

Advantages of Random Forest algorithm:

- High predictive accuracy.
- Efficient on large data sets.
- Ability to handle multiple input features without need for feature deletion.
- Feature selection is possible.

Dataset Description

The dataset is https://bigml.com/dashboard/dataset/5d71347eeba31d5272000492. This the facebook data used for the demographics. The overall data is 961 and according to the algorithm, the missing values are to be reduced and improve the accuracy of the algorithm with random tree adopted with pruning technique [8].

Name	Count	Missing	Instances
Sex	961	0	494 (Females) 467 (Males)
Age	961	0	18-35 (671) 36-70 (290)
Do you use internet?	961	0	No-159 Yes-802
Access internet on mobile handheld device?	961	0	No-425 Yes-536
Do you ever use Facebook?	821	140	No-318 Yes-503
Ever voluntarily taken a Facebook break for several weeks or more?	503	458	No-199 Yes-304
Over the last year Facebook has become	503	458	More-57 Less-139

ISSN: 2233-7857 IJFGCN

MORE / LESS important to you?			Unknown-307
The amount of time you spend using Facebook has increased/decreased/stayed over the last year?	503	458	Decreased-171 Increase-66 Unknown-266
Do you expect to spend MORE/LESS time on Facebook the upcoming year?	503	458	More-132 Less-18 Unknown-353

Table: 1 Dataset attributes

Performance Evolution

Various parameters are implemented in this system. The basic count values such as True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN), Sensitivity, Specificity and Accuracy are used by these measures.

False Positive Rate (FPR)

The percentage of cases where an image was classified to normal images, but in fact it did not.

$$FPR = \frac{FP}{FP + TN}$$

False Negative Rate (FNR)

The percentage of cases where an image was classified to abnormal images, but in fact it did.

$$FNR = \frac{FN}{FN + TN}$$

Accuracy

We can compute the measure of accuracy from the measures of sensitivity and specificity as specified below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + TN} * 100$$

ISSN: 2233-7857 IJFGCN

Experimental Results

In this section, the results are evaluated by using the java programming language with jdk 1.8. This will provide the better library functions to implement the proposed system efficiently compare with the proposed system. In this dataset, the missing values are very high for some of the attributes. With this missing values the result is not accurate. To overcome this, pruning is adopted with the ACA to improve the accuracy of the results.

Name	**Count**	**Missing**	**Instances**	**DT-Accuracy**	**ACA-Accuracy**
Sex	961	0	494 (Females) 467 (Males)	100%	100%
Age	961	0	18-35 (671) 36-70 (290)	100%	100%
Do you use internet?	961	0	No-159 Yes-802	100%	100%
Access internet on mobile handheld device?	961	0	No-425 Yes-536	100%	100%
Do you ever use Facebook?	821	140	No-318 Yes-503	87%	98%
Ever voluntarily taken a Facebook break for several weeks or more?	503	458	No-199 Yes-304	88.97%	97.78%
Over the last year Facebook has become MORE / LESS important to you?	503	458	More-57 Less-139 Unknown-307	89.98%	98.76%
The amount of time you spend using Facebook has increased/decreased/stayed over the last year?	503	458	Decreased-171 Increase-66 Unknown-	90.87%	97.87%

ISSN: 2233-7857 IJFGCN

			266		
Do you expect to spend MORE/LESS time on Facebook the upcoming year?	503	458	More-132 Less-18 Unknown-353	86.87%	96.87%

Table: 2 Shows the Performance of the ACA

	Accuracy
DT	**89.87 %**
ACA	**98.76 %**

Table: 3 Overall accuracy performances

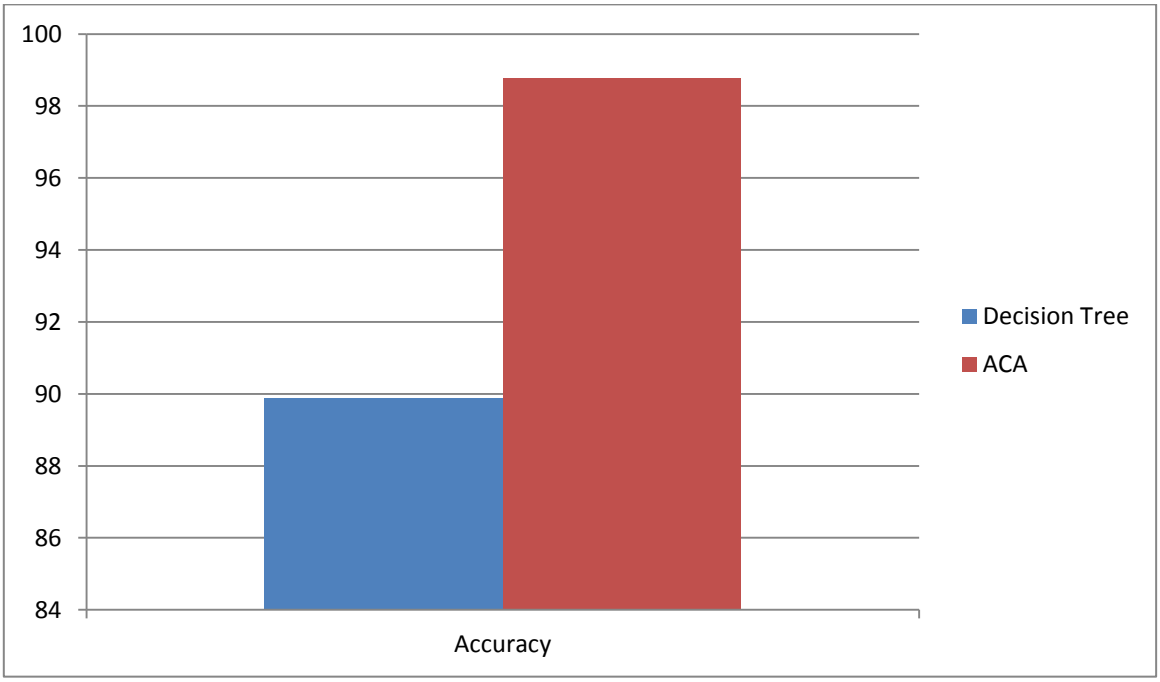

Figure: 2 Representation of Graph Performance

Conclusion

In this paper, the classification of data plays a major role. Now a days, the usage of the facebookmost widely used in many applications and for various purposes. It is very important to use these types of demographics datasets for the classification of data. In this

ISSN: 2233-7857 IJFGCN

paper, the ACA is the classification algorithm used to check the accurate values of the analysis of facebook datasets based on user's habits. The ACA adopted with pruning technique improves the performance of the proposed system.

References:

[1] Mitchell, (1997). "Machine Learning", The McGrawHill Companies, Inc.

[2] J. R. Quinlan, (1993). "C4.5: Programming for Machine Learning". San Francisco, CA: Morgan Kaufman.

[3] S. K. Murthy (1998). "Automatic construction of decision trees from data: a multi-disciplinary survey". DMKD, Vol. 2, No. 4, pp. 345-389.

[4] Moshe Ben-Bassat (1987). "Use of distance measure, Information measures and error bounds on feature evaluation". In Sreerama Murthy (1), pp. 9-11.

[5] Mark Last and OdedMaimon (2004). "A compact and accurate model for classification". IEEE Transactions on KDE, Vol. 16, No. 2, pp. 203-215.

[6] Byung Hwan Jun, Chang Soo Kim, Hong-Yeop Song and Jaihie Kim (1997). "A new criterion in selection and discretization of attributes for the generation of decision trees". IEEE Transactions on PAMI, Vol. 19, No. 12, pp. 1371- 1375.

[7] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone (1984). "Classification and Regression Trees". Wadsworth International Group, Belmont, California.

[8] S. K. Murthy, Simon Kasif and Steven Salzberg (1994). "A system for induction of oblique decision trees". JAIR 2, pp.1-33.ss

ISSN: 2233-7857 IJFGCN