IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

Analysing Success Possibility of a Mobile Application Using Data Mining Technique

Sharvil Kadam[1], Shantanu Kawlekar[2], Anish Nair[3]

B.E. Student, Information Technology, Dwarkadas J. Sanghvi College of Engineering, Mumbai, India [1, 2, 3]

Abstract: In this world of smartphones, everything is now being ported to the mobile in the form of mobile applications. A huge number of developers and entrepreneurs are starting their venture directly via a mobile application. Also many existing businesses and desktop application are now shifting to the mobile application world. This rapid rise of the use of smartphone applications has put a doubt in the minds of the developers that whether their app will be a success on the play store or not. In this paper we are going to propose the use of data mining techniques to analyse whether the newly hosted application will be a success or not. The results from the techniques cited in this paper can be used by various venture capitalist and potential investors for the application to make an informed decision about the funding. This paper uses analogy for this purpose. The analysis of attributes of many successful application hosted on the play store is done and then the most important attributes are found out using data mining. The resultant attributes can be used to compare with the new apps to find out the success possibility.

Keywords: Data Mining, Ideal Attributes, Raw Data, Application, Classification, Algorithms.

I. INTRODUCTION

Thousands of new Applications are launched per day on the app stores for various platforms such as Google play for android and iTunes for iOS. New ideas are cooked-up every day in some part of the world. A successful application can benefit a large group of masses. Those who benefit from the success of the application idea executed are:

- Developers and entrepreneurs
- Investors funding the app
- The audience using the application
- Budding developers who want to follow the footsteps of the successful application
- Advertisers putting their ads on the application

For an application to be successful, a variety of factors come to play in determining the potential success rate of the application. These factors are the characteristic of the application. These characteristics or attributes contribute majorly in the success of the application. Similarly attributes of the development team members and the developing environment also play a major factor in the success of the app. The marketing team will also be a factor. Thus identifying these different attributes in a successful application will give us a comprehensive view to analyse the factors responsible for the success of an application.

In this thesis, we are proposing a method to anticipate the success prospects of the application. This method has various steps that we have proposed.

1. Data Collection: We will find out all the attributes and the characteristics of the known popular applications in various domains.
2. Data Integration: After collecting the data, the data is cleaned by applying various algorithms and it is integrated, removing incomplete, inconsistent data from the data sets.
3. Knowledge: Ultimately, a set of attributes that most affect the success of the application are listed. This serves as an ideal attribute set for the success of the application.
4. Data of new application: Now the attributes of the new application are gathered and these are the raw data sets.
5. Comparison: The amount of similarity between the raw set and the ideal attribute set can give us the possibility of the success of the new application.

Such an analysis of the raw attribute data and the ideal attribute set can help us predict the success factor of the new application. It can help the developers and entrepreneurs to decide on the authenticity of the idea. It will also help the venture capitalist to decide on whether to fund the application or not.

II. APPROACH

The approach to be followed to find out the success possibility of a new application can be divided into three phases as shown the figure 1:

A. Classification of the application based on their category and use.
B. Defining and Formation of Ideal parameter set.
C. Comparison between Ideal and raw parameter set.

A. Classifying the applications based on their category and use:

Firstly we will be classifying a limited number of applications from different categories and further from sub categories. This limited number of applications will selected on the basis of their popularity, negative/positive reviews and considering the number of downloads wherein we can set a limit for a particular app to be eligible for analysis.

DOI 10.17148/IJARCCE.2015.41136

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

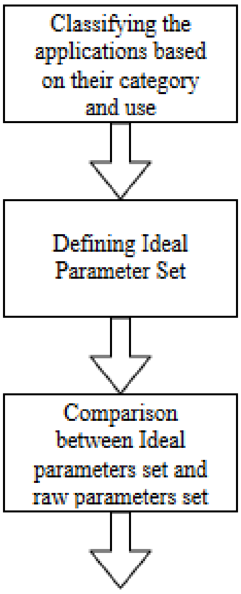
Classifying the applications based on their category and use
Defining Ideal Parameter Set
Comparison between Ideal parameters set and raw parameters set

Figure 1 – Process Flow

B. Defining the ideal Parameter set:

This phase is the crucial as the ideal parameter set is found out after identifying and applying the algorithms on the data sets of successful apps. Determining ideal parameters to be used in analysis is significant as this is the phase which provides most of the input information. Defining parameters is crucial part so that overall characteristics of an application is known and which will help in producing more accurate results during predictions and suggestions. The step involves in the formation and defining of the ideal parameters is shown in figure 2.

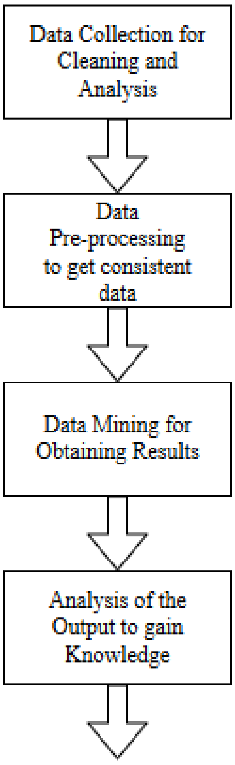
Data Collection for Cleaning and Analysis
Data Pre-processing to get consistent data
Data Mining for Obtaining Results
Analysis of the Output to gain Knowledge

Figure 2 – Formation of ideal attribute/parameter set for analysis

1. Data Collection for Cleaning and analysis:

This step consists of collecting the attributes of a successful application hosted on the store. These parameters may be associated with the team as a whole or the individual members of the team or may be the environmental factors or the technical factors.

The parameters of the successful applications are not clean and maybe inconsistent or incomplete. This data is send for pre-processing later where it will be cleaned. Many parameters may be irrelevant to the domain as well. [3]

1.1. Numerical and direct attributes:

These attributes are those which are visible to us and can be obtained without much effort.

Attributes under this category will help in determining what is expected and what is being delivered by the application developers as far as the end users are concerned.

Attributes	Description
Graphics quality	How much graphical content is present and what is its quality. Quality can be measured on a scale 1 to 10 with 0.5 increments.
Number of downloads	The number of application downloads will enable us to determine the reach of the application.
Rating	The star rating will be required to know how the application is.
Number of screens	Determining the number of user screens or options within the application.
Size of the Application	Size of the application is important as it will help in determining the space complexity of the application.
Application's Device Requirements	Many phones have different OS installed onto them and with different hardware specs, so it is essential in knowing what requirements are necessary to be satisfied by the application to run smoothly.
Application Developer Brand value	Popular Application developers have followers which helps them in spreading their application's reach. But does this truly affect the application success? This why this should be considered as an attribute. (Possible values : novice ,professional , student developer)

Table 1 – Direct Attributes

1.2. Rational Attributes:

The logical attributes that are not directly visible are required to be collected from the application. These attribute which can be converted into numeric values by mapping them on a scale with relevant points.

DOI 10.17148/IJARCCE.2015.41136

Attributes	Description
Engaging	The amount of time a user spends on the application. This can be obtained by user surveys.
Usability	Whether the application is operating as specified and performing all its functions as expected.
User Interface Complexity	Information about how easy, intuitive is the application to use.
Difficulty level (this attribute is required for games)	Level of difficulty a particular game based application can have, based on user reviews.
Marketing & Advertisements	This attribute is a Boolean attribute where we can understand whether the application is being advertised on T.V, newspapers or the web.

Table 2 – Rational Attributes

1.3. Member Attributes:

These parameters are the defined as how the qualities and experience of the members of the development team are. There are many abstract attributes associated with members such as member motivation, hard work, persistence and interest. However such attributes cannot be seen and hence cannot be calculated individually. Such attributes can be grouped into the term ‘Abstract attributes’ and can be rated. Apart from these, there are other important attributes that come under Member attributes.

- **Experience:** Experienced developers can build the application a lot fasters and with lesser bugs initially. Also no training is required for them as amateurs require. The success rate also depends on the experience of the members. Statistics are shown below in table 3.

Experience	Success Percentage
First-timer	18%
Repeat-Player	20%
Veteran	30%

Table 3 – Role of Experience in success

- **Familiarity and cognition level**: Unlike established applications, a new under developed application comes across new problems every day. The problem becomes severe if familiarity with these problems is low. Thus, familiarity with unknown situation candiminish the rate of application success. A fuzzy function S can be used to find the familiarity F (t) [2].

1.4. Environmental Attributes:

These attributes have an indirect effect on the application’s success. These are attributes that are not directly associated with the application but can affect the success rate. Following table give us a hint about environmental parameters.

Attributes	Description
Location	Location where application is developed as well as audience location is considered.
Competition	Other similar application already existing or in the making can affect the success rate.
Uniqueness of the idea and need of it	The idea behind the application is important, also whether application is needed or not for the masses.

Table 4 – Environmental Attributes

2. Data Pre-processing to get consistent data:

The information or the data that is gathered as the attributes in the data compilation period are not completely clear. The data might be lacking attribute values or containing discrepancies and errors. This data that is not clean may be completely insignificant. Data mining is not possible on inconsistent or noisy data. To deal with this raw data, cleaning is performed. Cleaning the data provides valuable result on mining. Other than just cleaning the data, variable transformation and dimensionality reduction are the other actions are performed to convert this unstructured data to a structured format.

Data pre-processing is a very important phase in data mining. The quality of data does affect the result. The different mining techniques like classification and clustering give most optimum and efficient results if the data that is used is thoroughly cleaned by pre-processing.

Initially, the process of collecting data was either by adding all the details in a particular file or folder or updating the excel sheets regularly. After which this data that was collected was pre-processed. Preprocessing deals with the removal of incomplete, noisy and inconsistent data from the file or the sheet.

While evaluating the different applications on the play store, new attributes can be seen, which may leave the data collected in an imbalanced format. A dataset is imbalanced if the classifications categorized are not equally represented. [10] Imbalance problem occur when one of the datasets contain more sample than the other dataset. Balancing of this imbalanced data has to be done before applying the data mining concepts to enhance the performance of the data. The most appropriate way to deal with imbalanced data is by using the hybrid method. The hybrid method contains one of the boosting algorithms like the RUSBoost or the SMOTEBoost along with the Feature Selection for the Classification of the Imbalanced data. [11]
After pre-processing of the data is done, the different mining algorithms can be applied on the processed data. These selective algorithms will help in selecting and estimating the attributes initially identified in the application. A table of these attributes is then made as per the frequency of occurrence of the attribute as shown below table 4.

DOI 10.17148/IJARCCE.2015.41136

Attributes	Frequency
Number of downloads	27
Uniqueness of idea	23
Location	17
Competition	11

Table 5 – Example Attributes and their frequency of occurrence

3. Data Mining for Obtaining Results:

Data mining (sometimes called data or knowledge discovery) is the process of analysing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analysing data. It allows users to analyse data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. [6][9]

The techniques of gaining knowledge involve various types of algorithms. These various algorithms used for data mining are clustering, classification, association, etc. [12]. The classification algorithm aptly works in this our case. Various data mining tools such as R-Programming or WEKA tool can be used for applying these algorithms to the data sets. To apply the algorithms various data sets as gathered previously can be used.

This can be done in the following ways:
1. The attributes that were identified in the data gathering tool are applied with the algorithms.
2. The attributes applied with selection algorithms first and then selected set with the classification algorithm.
3. The attributes obtained after balancing the set and then applied with the classification algorithms.

After applying these algorithms, accuracy and the cost effectiveness of the cases above can be calculated to get the best results. The classification algorithms are as follows:

3.1. Decision Tree Algorithm:

These are algorithm to represent information of classification algorithm classically. Tools of data mining provide with built-in algorithms [12]. Decision Tree algorithms such as ID3, C4.5, and CART adopt a greedy (i.e., no backtracking) approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner. [1]

3.2. Rule Based Algorithm:

These classifier algorithms utilize a set of IF_THEN rules for classification. [12] The rule can be written as:
IF <condition>
THEN <conclusion>
The IF part is the precondition or the antecedent while the THEN part is rule consequent. The IF part consists of condition that has the one or more testing attributes that are then ANDed while class prediction is done in consequent part. [1]
Such classification provides with induction rules and decision tree to determine the prediction/classification model.
Thus after getting the decision trees and induction rules, we interpret them and then the attributes that influence the classification most are found out, thereby removing the parameters that don't affect the prediction so much.
Hence, we obtain the attributes after data mining in a cost-efficient way. These attributes/parameters form the ideal attribute set.

4. Analysis of the output to gain knowledge:

Thus, the final step in KDD (Knowledge Discovery process) iscomplete and we get the desired output.Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. [7]

As we get the ideal attribute set from the previously successful application, now we can proceed with the comparison of the attributes of the new application (i.e. the raw data set) with the ideal attributes data set.

C.Comparison between Ideal and raw parameter set:

In this step, attributes from the new application are identified. This is similar to the data collection in the step one. Data from the new application are classified into four clusters as Numerical and directattributes, Rational Attributes, Member attributes and Environment attributes. Then the attributes are identified in all the three groups. These identified attributes are not clean. They have similar problem like inconsistency and incompleteness. Thus these attributes are pre-processed before comparing them with ideal attribute sets.

The attributes collected from all new applications may be unbalanced. This is there may be a few attributes more or less than the ideal attribute set. Balancing of the attributes is done and then these attributes are ready for the comparison study. This set forms the raw data set.

Thus, two data sets are available one is the ideal attribute set and other is the raw data sets. Any of the data set comparison technique used in statistics can be applied to compare the given data. This would thus determine the success possibility of the application. During comparison, the ideal data may be changed as per requirement of statistical comparison methods.
Standard error Method: Statistics consists of many algorithms for comparison. One such algorithm is standard error method. Here we assume the ideal attribute set in 0 to 1 range attributes and the raw data is given a 0 to 1 rating for the attributes. Assuming the mean to be ideal attribute set, standard deviation or variance of the raw set with ideal set is calculated and then prediction is made. The standard deviation is inversely proportional to success possibility. [5]
T-test: T-test is another such test that assumes the mean of two data set E and O and helps in data comparison. [6] The t-test is used to compare the values of the means from two samples and test whether it is likely that the samples are from populations having different mean values. [8]

DOI 10.17148/IJARCCE.2015.41136

After obtaining the results from analysis, the application undergoing development or is ready to publish can be compared to similar category based popular and successful application which will enable the system to suggest possible changes which can be made in the application. Similar applications are used for comparison because they will reflect characteristics which are required or should be present in that kind of an application (i.e. an application from game category can be compared with another application from a similar category).

Thus, by using this process, success possibility of an application can be found out.

III. CONCLUSION

The prediction of possibility of a new application to gain success is very difficult to anticipate since these statistics are very dynamic in nature. The procedure of determining the possibility of success is very extensive and time consuming. The information is gathered by examining the various prosperous applications across all platforms. This data thus obtained is undergone cleaning and balancing the imbalance attributes is done. There are various different Selection algorithms that are used in the pre-processing stage to determine accurate comparative data set. After analysing the ideal optimal data attribute set, data from the brand new endeavour is collected. This data also goes through various stages in pre-processing to form the raw attribute data set. Finally, when we have these two data sets for comparison of these two data sets so as to predict the probability of prosperity of the new application.

IV. FUTURE SCOPE

This approach of determining success can be further enhanced with the use of more parameters and extending this process across different fields. For example, same methodology can be used to predict success of a new car produced or rather any product which is new the market.
Also, Machine learning techniques can be used to learn the previous data and then comparison techniques can be applied. This would further reduce work and increase efficiency.

REFERENCES

[1] Data Mining Concepts and Techniques: By Jiawei Han, MichelineKamber and Jian Pei
[2] Gancho Vachkov, Hidenori Ishihara, “Classification of Process Data and Images by Human Assisted Fuzzy Similarity Analysis”, ICROS-SICE International Joint Conference 2009 August 18-21, 2009, Fukuoka International Congress Center, Japan
[3] W. Yathongchai, C. Yathongchai, K. Kerdprasop, N. Kerdprasop, “Factor Analysis with Data Mining Technique in Higher Educational Student Drop Out”, Latest Advances in Educational Technologies, 2003.
[4] http://www.surveystar.com/startips/jan2013.pdf
[5] http://www.statisticallysignificantconsulting.com/Ttest.html
[6] Data Mining: What is Data Mining?: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
[7] Overview of the KDD Process: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html
[8] T-test for data analysis: http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1438
[9] From Data Mining to Knowledge Discovery in Databases: - http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230/1131
[10] Data mining for imbalance data sets: An Overview By Nitish Chawla
[11] Class Imbalance Problem in Data Mining - http://arxiv.org/pdf/1305.1707.pdf
[12] Hina Gulati, “Predictive analysis using data mining concepts”, Institute of Electrical and Electronics Engineers.

BIOGRAPHIES

Mr. Sharvil Kadam is a Bachelor’s Student pursuing his B.E. in Information Technology from Dwarkadas J. Sanghvi College of Engineering, Vile Parle, Mumbai, Maharashtra, India. His fields of interests are Big Data, Databases, Artificial Intelligence, Data Mining and Security.

Mr. Shantanu Kawlekar is a Bachelor’s Student pursuing his B.E. in Information Technology from Dwarkadas J. Sanghvi College of Engineering, Vile Parle, Mumbai, Maharashtra, India. His fields of interests are Security, Data Mining, Networking and Software Development.

Mr. Anish Nair is a Bachelor’s Student pursuing his B.E. in Information Technology from Dwarkadas J. Sanghvi College of Engineering, Vile Parle, Mumbai, Maharashtra, India. His fields of interests are Information Management, Data Mining, Networking and Software Development.

DOI 10.17148/IJARCCE.2015.41136