Volume 6, Issue 2, February 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Survey on Text Classification Methods

Nisha Mariam Daniel, Karthik K
Computer Science, Christ University, Bengaluru,
Karnataka, India

Abstract*— Classification is an important area in the field of Machine Learning, Data Mining and Data Categorization and Analysis. Massive amount of data being collected and stored in databases everywhere every day. The amount of information is heading for rapid increase year after year. Every enterprise and research facilities have databases with terabytes of data. That is over 1,099,511,627,776 bytes of data. There is vital information and hidden knowledge in such databases and it's basically impossible to mine for them without programmed methods for extracting this information. Many algorithms are created to extract nuggets of knowledge from large data sets. The different methodologies to approach this problem include classification, association rule, clustering, etc. This survey paper will center on few methods of classification. The classification problem is as follows, for a given set of categories, the mission is to predict the category of objects by using the examples of objects or documents whose category is already known. The classification process is supervised as it uses earlier developed knowledge from training sets. The knowledge developed from training sets is known as the classification model. The classification model should have high accuracy in prediction of the category of unknown objects.

Keywords*— Data Mining, Data Classification Algorithms, Classifiers, Text Classifications, Text documents

I. INTRODUCTION

Data analysis has two forms, firstly those that can be used for extracting models to describe important classes and others which predicts future data. Classification models predict definite class labels whereas the prediction models predict continuous valued functions. With the vast amount of information these days and the limitations in the storage of data in an organized way, various classification models are created to address the problem of storing data in various meaningful categories by predicting the various possible categories of new unknown data documents

II. DATA CLASSIFICATION

Some domains that uses Text Classification

A. News filtering and Organization:

Information filtering system is a system that helps to remove redundant information from an information source using automated or computerized methods preceding user presentation.The main goal is the management of the information overload. To do this user's profiles are compared to some reference features. These characteristics may originate from the information item or the user's social environment.

News services today are mostly electronic in nature. Large volume of news articles are created every single day by the organizations. It is difficult to organize such huge lumps of these news articles manually. And so, automated methods can be helpful for news categorization in a variety of web portals.

B. Document Organization and Retrieval:

A number of supervised methods may be used for document organization. Various domains for document retrieval that include digital libraries, web collections, and scientific literature or even social feeds. Organizing the vast amount of information requires a lot of time and automated methods.

C. Opinion Mining:

Customer opinions and reviews are often text documents which can be mined to determine valuable information. Opinion mining is a type of natural language processing for tracking the mood of the public about a particular product. Opinion mining, which is also called sentiment analysis, involves building a system to collect and categorize opinions about a product. Automated opinion mining often uses machine learning, a type of artificial intelligence (AI), to mine text for sentiment [1]

D. Email Classification and Spam Filtering:

Email filtering is the processing of email to organize it according to specified criteria. Most often this refers to the automatic processing of incoming messages, but the term also applies to the intervention of human intelligence in

addition to anti-spam techniques, and to outgoing emails as well as those being received. Email filtering software inputs email. For its output, it might pass the message through unchanged for delivery to the user's mailbox, redirect the message for delivery elsewhere, or even throw the message away. Some mail filters are able to edit messages during processing. [2]

III. COMMONLY USED CLASSIFIERS

A. Decision Tree Classifiers

Decision tree builds classification model in the form of a tree. It basically breaks down a dataset into smaller subsets at the same time an associated decision tree is developed incrementally. Finally it results in a tree with decision nodes and leaf nodes. The decision node has two or more and the leaf node represents a classification or a decision. The topmost decision node is called root node. Decision trees handles both categorical and numerical data.

This classification method is simple and widely used. It gives a straightforward idea to solve the classification problem by posing a series of questions about the various attributes. When an answer is received, a follow-up question is asked until a conclusion is obtained. Few of the issues of the decision tree include binning, avoiding overfitting, super attributes, working with missing values etc.

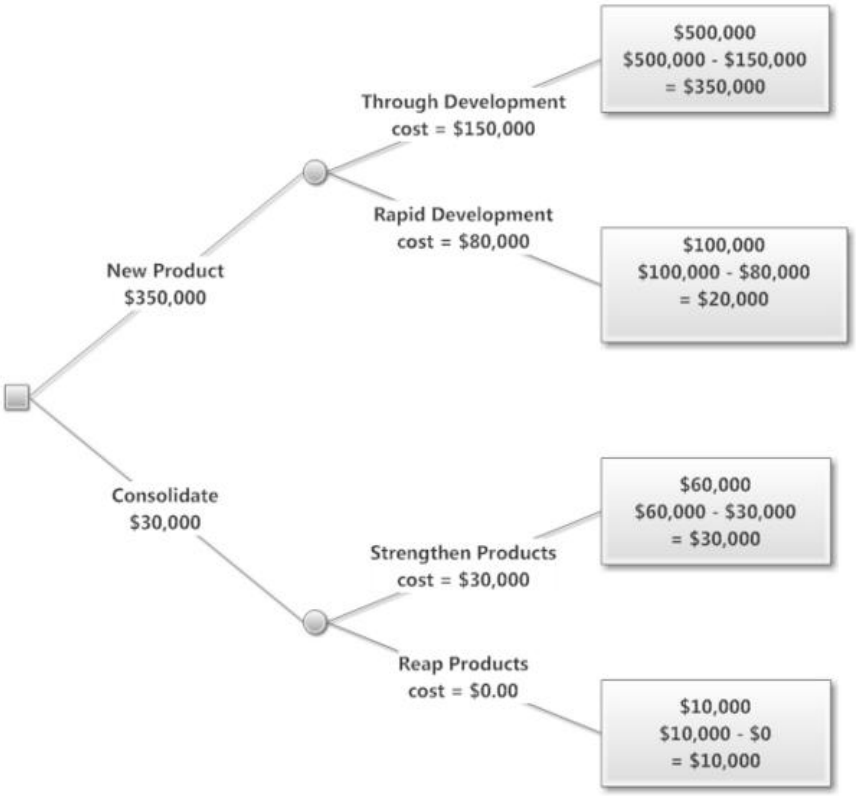

Fig -1: Decision Tree to develop a new product or consolidate

Fig 1 shows a decision tree for a development process in which the major task is to identify is a new component is to be developed or a consolidated product would be enough to fulfil a given requirement or functionality.

B. Rule-based Classifiers

Pattern recognition is a branch of machine learning that focuses on the recognition of patterns and regularities in data, although it is in some cases considered to be nearly synonymous with machine learning.[3]Pattern recognition systems follow supervised learning, but if no labeled data are available then other algorithms can be used to discover unknown patterns. Pattern recognition is generally categorized based on the type of learning procedure. Rule-based classifier uses the IF-THEN rules for classification. The rule is expressed as

IF condition THEN conclusion

Pattern mining are helpful to obtain models for domains such as graphs and sequences. It has been proposed as a means to obtain and more interpretable models. The key idea of pattern-based classification is to help create classification model by using patterns to define new features.

C. Probabilistic and Naive Bayes Classifiers

Naïve Bayes is a very simple probabilistic model that tends to work well on text classifications and usually takes orders of magnitude less time to train when compared to models like support vector machines. A Naive Bayes classifier is a simple probabilistic model based on the Bayes rule along with a strong independence assumption. The Naïve Bayes

model involves a simplifying conditional independence assumption. That is given a class (positive or negative), the words are conditionally independent of each other. This assumption does not affect the accuracy in text classification by much but makes really fast classification algorithms applicable for the problem. Rennie et al discuss the performance of Naïve Bayes on text classification tasks in their 2003 paper. In our case, the maximum likelihood probability of a word belonging to a particular class is given by the expression:

$$P(x_i|\, c) = \frac{Count\ of\ x_i\ in\ documents\ of\ class\ c}{Total\ no\ of\ words\ in\ documents\ of\ class\ c}$$

The frequency counts of the words are stored in hash tables during the training phase. According to the Bayes Rule, the probability of a particular document belonging to a class ci is given by,

$$P(c_i|d) = \frac{P(d\ |c_i) * P(c_i)}{P(d)}$$

If we use the simplifying conditional independence assumption, that given a class (positive or negative), the words are conditionally independent of each other. Due to this simplifying assumption the model is termed as "naïve".

$$P(c_i|d) = \frac{(\prod P(x_i|c_j)) * P(c_j)}{P(d)}$$

Here the xi s are the individual words of the document. The classifier outputs the class with the maximum posterior probability.[4] We also remove duplicate words from the document, they don't add any additional information; this type of Naïve Bayes algorithm is called Bernoulli Naïve Bayes. Including just the presence of a word instead of the count has been found to improve performance marginally, when there is a large number of training examples.

D. Regression-Based Classifiers

Regression is a kind of machine learning data mining technique used to fit an equation to a dataset. The, linear regression which is the simplest form of regression, uses straight line (y = mx + b) formula and determines the values for m and b to calculate the value of y for a given value of x. Multiple regression, uses more than one input variable and uses more complex equations such as a quadratic equation. Regressions is used to learn relationship values between real valued attributes and not binary attributes.

E. Proximity-based Classifiers

Proximity-based classifiers use distance-based procedures to perform the classification. The main idea is that documents which belong to the same class are more close to one another based on similarity measures like the dot product. Normally two methods are used to perform the classification for a given distance:

- Determine the k-nearest neighbors in the training data.
- Perform training data aggregation during pre-processing, in which groups of documents belonging to the same class are formed.

Proximity is the basic quality which identifies and characterizes groups of objects in various domains and contexts. When objects are compared to a set of chosen prototype examples, proximity can be used as a natural ingredient to build a numerical representation. Pattern classes may be learned from such proximity representations by the traditional nearest neighbor rule, as well as by other alternative strategies. These encode the proximity information in suitable representation vector spaces in which statistical classifiers can be trained. Such recognition techniques can be successful, provided that the measure is informative, independently whether it is metric or Euclidean, or not. [5]

F. Neural Network Classifiers

Neural networks have begun as advanced data mining tools where other techniques do not produce acceptable predictive models. It resembles the biological modelling capability with neurons. A neural network is made up of units (neurons), arranged in different layers, which converts the input vector into an output. Each unit takes an input, applies a nonlinear function to it and then passes the output to the next layer. The networks are defined to be feed-forward ie. A unit feeds its output to the units on the next layer from lower to upper and no feedback is given the other way. Weights are applied to the signals passing from one unit to another, and these weights are tuned in the training phase to adapt a neural network to the particular problem.

IV. CONCLUSIONS

Data mining offers favorable ways to uncover buried patterns within tons of data. All these hidden patterns can possibly be used to predict future behaviors. New data mining algorithms must have certain aspects in consideration like, these techniques are good as the data that has been collected. The prime requirement for good data exploration is worthy good data. Once good data is available, the next step is to choosing the most appropriate technique to mine the data. However, this comes with compromises while considering appropriate data mining technique to be used in a certain application. Some techniques even include many other techniques in combination to solve a problem. Different classification algorithms are needed with high accuracy to obtain categorize unknown objects and documents.

REFERENCES

[1] Article on Opinion Mining in the Search Business Analytics Forum
[2] From Wikipedia, the free encyclopedia (Email Filtering).
[3] From Wikipedia, the free encyclopedia (Pattern Recognition)
[4] Paper on "Fast and accurate sentiment classification using an enhanced Naive Bayes model." Vivek Narayanan , Ishan Arora , Arjun Bhatia
[5] Paper on "Learning with general proximity measures." by El˙zbieta P Ekalska , Robert P. W. Duin

