DATA ANALYSIS – BETWEEN THEORY AND PRACTICE

Nicoleta ENACHE-DAVID, Livia SANGEORZAN,
George-Alex STELEA

"Transilvania" University, Braşov, Romania (nicoleta.enache@unitbv.ro, sangeorzan@unitbv.ro, george@stelea.ro)

DOI: 10.19062/1842-9238.2017.15.2.11

Abstract: *In our paper we highlight some aspects of the text classification problem using the Naïve Bayes Multinomial classifier. We use Weka software for modeling this problem and we study the conditions that allow the classifier to obtain the highest prediction.*

Keywords: *classifier, prediction, text classification.*

1. INTRODUCTION

Document classification is a very actual issue and is a continuous challenge; it is based on different techniques of machine learning including Bayesian classification [6], SVM classifiers (Support Vector Machine) [9],[11] k-NN (k-Nearest-Neighbor) classifier [10], classification based on association rules [20], decision tree [16] etc.

Such machine learning techniques can be applied on complex information systems like in [1],[2],[3],[5], and for mathematical models like [6],[7],[8]. In the economic field, quantitative analysis of risk represents one of the phases that have to be followed in order to evaluate risks that an organization may face while developing its business [4], [12],[13]. This kind of analysis aims to numerical assessment of the probability and impact of each risk upon the organization's objectives. For this purpose there are used quantitative techniques such as the decision tree method [14],[15].

In literature there exist also statistical machine learning methods that can be applied to document clustering, document classification and predictive modeling. For testing the model inference, one can use the Monte Carlo method [18].

Our application for text classification takes into account the training set with synonyms and without synonyms. Synonyms are words having similar meaning. In our study we use the Naïve Bayes Multinomial classifier and we study the conditions that allow the classifier to obtain the highest prediction.

There are many studies on extracting synonyms automatically including the use of machine learning. Some studies analyze synonyms using similarity without machine learning. In [21] the authors present an automatic selection of synonyms using machine learning.

In our paper we are looking after synonyms for the four training sets for different categories (spam, sport, social-media, and travel). It is important to obtain high performance for automatic selection of synonyms using machine learning for the language of interest.

2. NAÏVE BAYES CLASSIFIER

2.1 Bayes' theorem

Let us consider an experiment denoted by E. We denote by S the sample space as the set of all possible outcomes of E. [17]

Definition 1

Let be A and B two events associated with an experiment E.

We denote by $P(B \mid A)$ the conditional probability of the event B, given that A has occurred.

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}, \text{ provided that } P(A) > 0$$

Definition 2

Events $B_1, B_2, ..., B_k$ represent a partition of the sample space S if

(a) $B_i \cap B_j = \varnothing$ for all $i \neq j$

(b) $\bigcup_{i=1}^{k} B_i = S$

(c) $P(B_i) > 0$ for all $i = \overline{1,k}$.

Definition 2 means that when the experiment E is performed one and only one of the events B_i occurs.

Let A be an arbitar event generated from S and let $B_1, B_2, ..., B_k$ the partition of S. Let $B_1 + B_2 + ... + B_k = S$, where $B_1, B_2, ..., B_k$ are mutually exclusive and exhaustive events. Each term $P(A \cap B_j)$ may be expressed as $P(A \mid B_j)P(B_j)$ and hence we obtain what is called the theorem on total probability:

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + ... + P(A \mid B_k)P(B_k) \quad (1)$$

Bayes' theorem

Let $B_1, B_2, ..., B_k$ be a partition of the sample space S and let A be an event associated with S. From the definition of conditional probability we obtain

$$P(B \mid A) = \frac{P(A \mid B_i)P(B_i)}{\sum_{j=1}^{k} P(A \mid B_j)P(B_j)} \quad (2)$$

The previous formula is called the formula for the probability of "causes". Bayes's theorem is the basis of statistical decision theory in some situations. [19]

Let us consider the event $B_i, i = \overline{1,k}$. The probability $P(B_i)$ is called prior probability, $P(B_i \mid A)$ is called posteriori probability and $P(A \mid B_i)$ is called the likelihood.

The Bayes' theorem provides a way of calculating the posterior probability, $P(B_i \mid A)$ if $P(B_i)$, $P(A)$ and $P(A \mid B_i)$ are known.

2.2 The classification problem

We consider a set of n classes $c_1, c_2, ..., c_n$. The problem is to determine which class(es) a given object belongs to. If this collection will increase, we must repeat the task but we want that repetitive task be automated. This process is called standing query, it is like any other query except that it is periodically executed on a collection to which new documents are incrementally added over time. [5]

If the standing query serves to divide the collection into the two classes, we refer to this as two-class classification.

Many systems support standing queries. When we use a classification with standing queries it is called routing or filtering.

2.3 Bayes classifiers

The Bayes classifiers are also called: Idiot Bayes, Naïve Bayes, Simple Bayes. The Bayes classifiers use Bayes' theorem.

The Naive Bayes classifier is a simple probabilistic classifier which is based on Bayes theorem with strong and naïve independence assumptions. It is one of the most basic text classification techniques with various applications.

Naive Bayes performs well in many complex real-world problems, even if it has a naïve design and oversimplified assumptions. This classifier is superior in terms of memory consumption and in several cases its performance is very close to more complicated and slower classification techniques. Overall, the Naive Bayes classifier is used as a baseline in many researches.

There are several types of Naïve Bayes classifiers: Multinomial Naive Bayes, Binarized Multinomial Naive Bayes and the Bernoulli Naive Bayes. Naïve Bayes and multinomial Naïve Bayes model are both supervised learning methods. They are also probabilistic learning methods.

Each type of Naïve Bayes classifiers can have as output different results since they use completely different models.

Multinomial Naive Bayes is used when the multiple occurrences of the words is very important in the classification problem. The Binarized Multinomial Naive Bayes is used when the frequencies of the words don't have a very important role in the classification. Bernoulli Naive Bayes can be used when is important the absence of a particular word matters. Bernoulli is usually used in spam detection and good results are obtained.

2.4 Text classification problem

Text classification is intended to assigning subjects to categories and can be used for spam detection, age or gender identification, language identification, etc.

In practice, it is possible to have more than two classes and the naïve Bayesian classifiers estimate the probability of class c_j generating instance d. Generally, the Naïve Bayes attributes have independent distributions. The assumption to have all attributes independent because of the meaning of the word naïve does not fit in real world situations. Though, the classifier works well in many practical situations.

A text classification definition can be: we have as input a document d, a fixed set of classes $C = \{c_1, c_2, ..., c_n\}$ and as output a predicted class $c \in C$ [16].

A method that we can use to predict the class c is using a Supervised Machine Learning Method. This means we have as input a document d, a fixed set of classes $C = \{c_1, c_2, ..., c_n\}$, a training set of m hand-labeled documents $(d_1, c_1), (d_2, c_2), ..., (d_m, c_m)$. The output will be a learned classifier $g : d \to c$.

We denote by X the document space. In text classification, we are given a description $d \in X$ of a document and a fixed set of classes $C = \{c_1, c_2, ..., c_n\}$. Classes are called categories or labels.

The Naive Bayes classifiers can be represented as this type of graph:

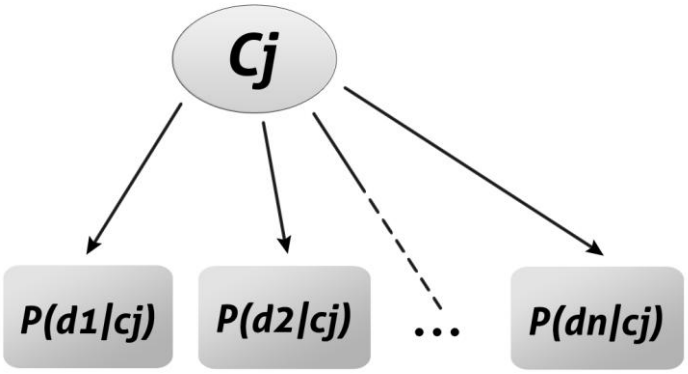

FIG. 1. Naïve Bayes classifier

The directions of the arrows indicate which state that each class causes certain features, with a certain probability.

In text classification, the goal is to find the best class for the document. The best class in Naïve Bayes classification is the most likely or MAximum a Posteriori (MAP) class noted with c_{MAP}. We refer at c_{MAP} as "MAximum a Posteriori"; the most likely class $c \in C$.

$$c_{MAP} = \arg\max P(c \mid d) = \frac{\arg\max(P(d \mid c)P(c))}{P(d)} = \frac{\arg\max (\mathrm{P}(\mathrm{t}_1, \mathrm{t}_2, \ldots, \mathrm{t}_n \mid \mathrm{c})\mathrm{P}(\mathrm{c}))}{P(d)} =$$

$$= \frac{\arg\max (\mathrm{P}(\mathrm{c})\mathrm{P}(\mathrm{t}_1 \mid \mathrm{c})\,\mathrm{P}(\mathrm{t}_2 \mid \mathrm{c}) \ldots \mathrm{P}(\mathrm{t}_n \mid \mathrm{c}))}{\mathrm{P}(\mathrm{d})} \tag{3}$$

There will be used only argmax(P(d|c)P(c) because it analyzes the same document d, which is test set. The document d is consisting of up of $t_k, k = \overline{1, n_d}$ terms, where n_d is the number of terms in document d. These t_k terms are tokens in document d [5].

The notation $P(t_n \mid c)$ represents the relative frequency of term t_n in document d belonging to class c. The situation when there is a term with zero frequencies is not possible and one use Laplace smoothing (add-1) for Naïve Bayes which adds one. This way one eliminates zeros.

We compare the calculated probabilities that the document belongs to a certain class and we choose the class with the higher probability.

3. APPLICATION - DOCUMENTS CLASSIFICATION WITH NAÏVE BAYES CLASSIFIER

In our application we use WEKA (Waikato Environment for Knowledge Analysis) that is a collection of machine learning algorithms for solving real-world data mining problems. Features of Weka are: machine learning, data mining, preprocessing, classification, regression, clustering, association rules, attribute selection, visualization [22].

WEKA software allows us to calculate the probability that a document belongs to a particular class, in which case we have the four categories of spam, sport, social media and travel but can not analyze the synonyms in a test set or training set for a foreign language.

The paper proposes to study the belonging of a document to a so-called class from two points of view, namely:

1. Pairs of words that are synonymous;
2. Words totally, without taking into account pairs of synonyms.

The scheme below suggests the proposed layout.

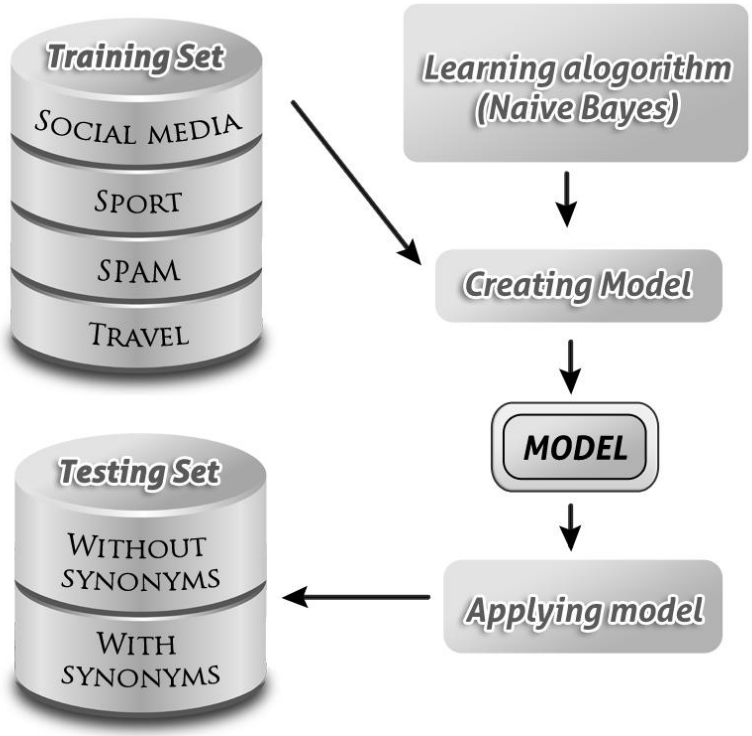

FIG. 2. Creating and testing the model

In our study we use the Naïve Bayes classifier from WEKA that has as output different files with results.

Our training dataset consists of 40 documents having 2131 words, respectively 10 documents for each category: sport, social media, spam and travel. We want to study which class the following document that includes synonyms (*holiday* and *vacation* belongs to: 'I want a great holiday and vacation'.

In WEKA we made the following settings: we have chosen Meta - Filtered Classifier and classifier: Naïve Bayes Multinomial, filter: StringToWordVector.

In *testHV* dataset *holiday* and *vacation* were considered distinct tokens and in *testH* and *testV* datasets they were considered as synonyms. The training set contains both tokens several times. We have obtained the results from the Table 1.

Table 1. Output for three datasets-Naïve Bayes

Test set	**Document content**	**Category**	**Prediction**	**Time (sec)**
testHV	I want a great *holiday* and *vacation*	Travel	0.429	0.02
testH	I want a great *holiday*	Travel	0.352	0.02
testV	I want a great *vacation*	Sport	0.515	0.01

In Table 2 we highlight the number of appearance of tokens holiday and vacation in training dataset.

Table 2. Number of appearance of words *holiday* and *vacation* in training dataset

Category	**Number of appearance in training dataset**	
	holiday	*vacation*
Sport	0	1
Social media	2	0
Spam	0	0
Travel	8	3

Weka output of probability of word given the class from the training dataset given by class is highlighted in Table 3. Table 3 contains only the number of occurrence of words in training set from the test datasets.

Table 3. Weka output: probability of word given by class

Word	**Sport**	**Social media**	**Spam**	**Travel**
I	0.00205	0.00071	0.00069	0.00067
want	0.00068	0.00213	0.00069	0.00135
a	0.00410	0.00498	0.00484	0.00610
great	0.00136	0.00071	0.00069	0.00067
holiday	0.0068	0.00142	0.00069	0.00203
and	0.00546	0.00570	0.00553	0.00542
vacation	0.00136	0.00071	0.00069	0.00013

In the *testH* document we have two synonyms *holiday* and *vacation*, and we have replaced all the occurrences of *vacation* with *holiday*. Applying the Naïve Bayes Multinomial classifier we have obtained that this document belongs to *Travel* category with prediction 0.352. This prediction is for *Travel* because the word *holiday* has the highest probability of word given the class in *Travel* (0.00203).

In the *testV* document we have have two synonyms *holiday* and *vacation*, and we have replaced all the occurrences of *holiday* with *vacation*. Following the application of Naïve Bayes Multinomial classifier we have obtained that this document belongs to the *Sport* category with prediction 0.515. This prediction is for *Sport* because the word *vacation* has the best probability of word given the class in *Sport* (0.00136).

The synonym selection in the dataset test is very important and should be made according to the significance of the information in the document, because in our training dataset we have both the words *holiday* and *vacation*. If we have replaced *vacation* with *holiday*, the *Travel* category has the highest probability of word given by the class, which has classified this document in *Travel*. Instead, when we replaced *holiday* with *vacation*, the probability of word given by class was for the *Sport* category.

Our proposal to do these two types of analysis is important if it is desired to classify documents in *Travel* or *Sport*, to take into account the meaning of synonyms for that language.

We want to emphasize that *holiday* and *vacation* in English language are synonymous, but have different meanings depending on the context: *vacation* means "planned time spent not working"; *holiday* means "celebration day or time off". In this analysis of the ambiguous intent document, the human factor must work to specify which synonyms will be used. Such an analysis is useful for books translations, newspaper articles and another related domains.

CONCLUSIONS

In this paper we highlighted some theoretical aspects regarding the text classification problem using the Naïve Bayes Multinomial classifier. In our application we uses WEKA software for modeling this problem.

We applied the Naïve Bayes Multinomial classifier on a training dataset containing a pair of words that are synonymous. We have studied the conditions for obtaining the highest prediction, taking into account the meaning of synonyms. From this point of view, our conclusion is that the human factor is decisive in choosing the proper synonyms.

REFERENCES

[1] C. Carstea, *Control and management in complex information systems*, Bulletin of the Transilvania University of Braşov, Series III Mathematics • Informatics • Physics, pp 73-87, 2013;
[2] C. Carstea, *Optimization Techniques in Project Controlling*, Ovidius University Annals Economic Sciences Series, 1: 428-432, 2013;
[3] C.G.Carstea, *Modeling System's Process for Control Of Complex Information Systems*, In Proceedings of The 25th International Business Information Management Association Conference (IBIMA), Soliman KS (ed). Amsterdam, Netherlands, May 2015, pp 566-574, 2015;
[4] C. Carstea, *Optimization Techniques in Project Controlling*, OVIDIUS UNIVERSITY ANNALS, ECONOMIC SCIENCES SERIES, Volume XIII Issue 1, pp. 428-432, Ovidius University Press, 2013;
[5] C.G. Carstea, *IT Project Management – Cost, Time and Quality*, ISSN 2067-5046, Economy Transdisciplinary Cognition International , Volume17, Issue 1/2014, pg.28-34, 2014;
[6] O. Florea, I. Rosca, *The Mechanical Behavior and the Mathematical Modeling of an Intervertebral Disc*, Acta Technica Napocensis Series-Applied Mathematics Mechanics and Engineering, 58: 213-218, 2015;
[7] O. Florea, I.C Rosca,, *Analytic study of a rolling sphere on a rough surface*, AIP ADVANCES, Volume: 6, Issue: 11, 2016;
[8] O. Florea, I.C. Rosca, *Stokes' Second Problem for a Micropolar Fluid with Slip*, PLOS ONE, Volume: 10, Issue: 7, Article Number: e0131860, 2015;
[9] A. Khashman, N.I. Nwulu, *Support vector machines versus back propagation algorithm for oil price prediction*, Advances in Neural Networks–ISNN 2011, pp 530-538, 2011;
[10] A. Khashman, *Blood Cell Identification using Emotional Neural Networks*, J. Inf. Sci. Eng. 25 (6), pp 1737-1751, 2009;
[11] A. Khashman, *IBCIS: Intelligent blood cell identification system*, Progress in Natural Science 18 (10), PP1309-1314, 2008;
[12] L. Mandru, *How to Control Risks? Towards A Structure of Enterprise Risk Management Process*, Journal of Public Administration, Finance and Law, pp 80-92, 2016;
[13] M. Popescu, L. Mandru**,** *Relationship between Quality Planning and Innovation*, Bulletin of the Transilvania University of Braşov, Series V, Economic Sciences, Vol. 9 (58) No. 2, pp.203-212, 2016;
[14] L. Mandru, D. Pauna, *Application of 'Small Steps Strategy' in the Management of European Companies*, Acta Universitatis Danubius Journal, Vol 8, no.3, October, 2012;
[15] L. Mandru, *Managementul integrat calitate-risc pentru societăţile comerciale cu profil industrial*, Editura Universităţii Transilvania, Brasov, 2011;

[16] C.D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008;
[17] P.L. Meyer, *Introductory probability and statistical applications. Second edition*, Addison Wesley Publishing Company, 1970;
[18] B.G. Munteanu, A. Leahu, S. Cataranciuc, *On the limit theorem for life time distribution connected with some reliability systems and their validation by means of the Monte Carlo method*, AIP Conference Proceedings 1557, pp. 582-588, 2013;
[19] H. Robbins, J. Ryzin, *Introduction to statistics*, Science research associates Inc., 1975;
[20] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Amsterdam, 1990;
[21] T. Wang, G. Hirst, *Exploring patterns in dictionary definitions for synonym extraction*, Natural Language Engineering, 18: 313–342, 2012;
[22] http:// http://www.cs.waikato.ac.nz/ml/weka/.