Research Journal of Applied Sciences, Engineering and Technology 5(2): 421-426, 2013
ISSN: 2040-7459; E-ISSN: 2040-7467

Submitted: May 04, 2012 Accepted: June 08, 2012 Published: January 11, 2013

Implementation of ML Using Naïve Bayes Algorithm for Identifying Disease-Treatment Relation in Bio-Science Text

T.F. Michael Raj and S. Prasanna
Department of Computer Science Engineering, SASTRA University, India

Abstract: In recent years many successful machine learning applications have been developed, ranging from data-mining programs to information-filtering systems that learn users' reading preferences. At the same time, there have been important advances in the theory and algorithms that can be used identify the diseases and treatment relations in a Bio-Science text. Imagine a computer learns from medical records which treatments are most effective for new diseases. Having the machine learning concept behind we have proposed a Machine Learning (ML) approach based on Naïve Bayes (NB) algorithm to improve the automatic disease identification in the medical field. And also we have improved text classification by using an integrated model.

Keywords: Health care information, machine learning, natural language processing

INTRODUCTION

"The goal of machine learning is to build computer systems that can adapt and learn from their experience" said Thomas G. Dietterich. (http://web.engr. oregonstate.edu/~tgd/) Machine Learning is a natural outgrowth of the intersection of Computer Science and Statistics. We might say the defining question of Computer Science is "How can we build machines that solve problems, and which problems are inherently tractable/intractable?" The question that largely defines Statistics is "What can be inferred from data plus a set of modeling assumptions, with what reliability?" The defining question for Machine Learning builds on both, but it is a distinct question (Tom, 2006).

The attention currently being given to the computer based disease-treatment relations adds impetus to the development of representational structures for 'clinical data. It is hoped that standardized structures can be developed to serve as a framework for combining disease-treatment, prevent-cure data from multiple sources.

In the evolution of computerized record systems, the controversy between free text and preset categories for recording patient data has not been resolved (Sager *et al*., 1994). The need for standards pushes toward reset categories and controlled vocabularies, while the need for expressive power, so as not to distort the disease data, speaks for allowing some amount of free-text reporting. It is the aim of this study to show that the techniques of linguistic analysis and Natural-Language Processing (NLP) can contribute to this effort.

A necessary supplement to syntactic analysis, in order to arrive at a representation of the specific information in a text or query is a method of determining the semantic categories of the discourse. The linguist Z. Harris provided a basis for developing the relevant categories and relations in technical and scientific subject matters.

Medical areas of study and/or computer applications using the LSP system have included radiology, nuclear medicine, pharmacology, sickle-cell disease, pneumonia, bacterial meningitis, anatomic pathology, rheumatoid arthritis, digestive surgery and asthma. Experience with this variety of medical language material has shown us that the principles of sublanguage analysis apply.

In our approach we have used NLP and ML to process the Medline text and web based application is developed for future use. The system is application-independent in the sense that the processor and the major components (tokenizer, adoptive procedure) are geared to the processing of Medline abstract. As a result, the medical sublanguage grammar and the medical word classification scheme have remained stable over a range of clinical areas. Nevertheless, a period of adaptation for any particular application will be necessary. It means that necessary update should be done for the application. The amount of effort required to develop an application that uses the system depends on the complexity of the information that is desired from the Medline documents and how tolerant (or intolerant) of error these demands are.

Related concepts: The main objective of study is to propose a model with suitable procedures. There are so many model have been proposed to classify the text. The challenges in ML techniques are:

- Identifying the suitable model for prediction
- Finding good data representation

Corresponding Author: T.F. Michael Raj, Department of Computer Science Engineering, SASTRA University, India

Here we have discussed the various models for prediction and data representation:

- Decision based models (decision trees)
- Adoptive learning algorithm-Ada boost
- SVM
- Probabilistic model-Naïve Bayes

Analysis of representative models:

Decision based models (decision trees): Decision trees are powerful and popular tools for classification and prediction (Azmy *et al*., 2005). Decision trees represent *rules*, which can be understood by humans and used in knowledge system such as database. Key requirement of this model is:

- **Attribute-value description:** Object or case must be expressible in terms of a fixed collection of properties or attributes (e.g., hot, mild, cold)
- **Predefined classes (target values):** The target function has discrete output values (bollean or multiclass)
- **Sufficient data:** Enough training cases should be provided to learn the model

Decision trees classify instances or examples by starting at the root of the tree and moving through it until a leaf node. It is used in hand crafted models and I s suitable for short texts. Decision tree is a classifier in the form of a tree structure:

- **Decision node:** Specifies a test on a single attribute
- **Leaf node:** Indicates the value of the target attribute
- **Arc/edge:** Split of one attribute
- **Path:** A disjunction of test to make the final decision

Adoptive learning algorithm-ada boost: Ada Boost is an algorithm for constructing a "strong" classifier as linear Combination (Freund, 1999). It focuses on hard to learn concepts and characteristics that appears in our short texts and imbalanced data sets:

- Advantages
  - Very simple to implement
  - Does feature selection resulting in relatively simple classifier
  - Fairly good generalization
- Disadvantages
  - Suboptimal solution
  - Sensitive to noisy data and outliers

SVM: SVMs are a new learning method introduced by Vapnik (1995). A Support Vector Machine (SVM) is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the input, making the SVM a non-probabilistic binary linear classifier. They are well-founded in terms of computational learning theory and very open to theoretical understanding and analysis.

Support vector machines are based on the Structural Risk Minimization principle from computational learning theory. The idea of structural risk minimization is to find a hypothesis h for which we can guarantee the lowest true error. The true error of h is the probability that h will make an error on an unseen and randomly selected test example.

An upper bound can be used to connect the true error of a hypothesis h with the error of h on the training set and the complexity of H (measured by VC-Dimension), the hypothesis space containing h (Vapnik, 1995). Support vector machines find the hypothesis h which (approximately) minimizes this bound on the true error by effectively and efficiently controlling the VC-Dimension of H.

Probabilistic model-Naïve Bayes: Naive Bayes is one of the most efficient and effective inductive learning algorithms for machine learning and data mining. Its competitive performance in classification is surprising, because the conditional independence assumption, on which it is based, is rarely true in real world applications (Qin, 2006).

Naive Bayes is the simplest form of Bayesian network, in which all attributes are independent given the value of the class variable. This is called conditional independence. It is obvious that the conditional independence assumption is rarely true in most real-world applications (Harry, 2004). Advantage of NB's algorithm is that the state of the art in the text classification.

LITERATURE REVIEW

"The goal of machine learning is to build computer systems that can adapt and learn from their experience." Every machine learning algorithm has both a computational aspect (how to compute the answer) and a statistical aspect (how to ensure that future predictions are accurate).

The ultimate aim of our study is to provide base of information technology model called as Health care Information Model (HIM) or frame work that helps the

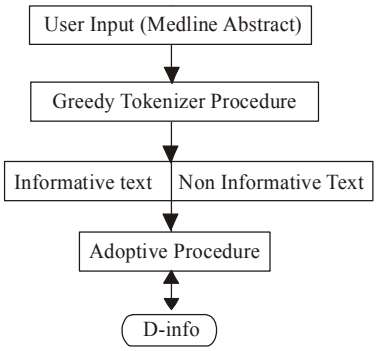

Fig. 1: Integerated model for medline text extraction

society to identify and disseminates health care information. Currently medical document is reality-based medicine, so that is the most important richest and most realistic source of medical and health information (Khozoie, 2012). In this study we have proposed a model to train the machine and to identify the diseases based on the NB's classifiers algorithm.

There are various approaches (Rosario *et al*., 2004) of classifying the Medline abstracts (http: // structuredabstracts.nlm.nih.gov/).

In our model the text classification and selection is done by the syntactic rule based approach. There are two steps in the text classification:

- Identifying the text related to disease and treatment and semantic relations between them
- Storing the information in the disease-treatment repository called D-Info

NLP is an approach that can be computerized to analyze text. NLP is based on set theories and set of technologies. NLP techniques have been long used to extract medical knowledge from narrative reports. These NLP approaches roughly fall into four categories symbolic, statistical, connectionist and hybrid.

Figure 1 shows our model called integrated model for Medline text extraction where we have used Greedy tokenizer algorithm followed by an adoptive intelligent string matching algorithm that split the Medline text into words informative and non informative text. Informative text can be used in our proposed model to identify the disease-treatment relation. Non-informative text will not consider further.

The informative text can be compared with medical dataset if it is matched with D-info it can be added to the client information to identify the cure, prevent relations.

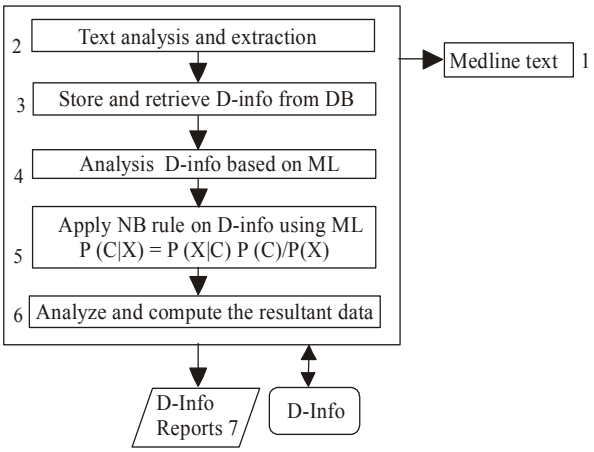

Fig. 2: Health care Information Model (HIM)

Algorithm Word_Classifier (MedLine-Abstarct)

- CALL Greedy_Tokenizer (Medline-Abstract)
- CALL Intelligent_Word_Classifier (Tokens[])

Algorithm Greedy_Tokenizer (Medline-Abstract)

- Tokenizes text into an array of words, using whitespace and all punctuation as delimiters
- <param name="text"> the text to tokenize</param>
- <returns> an array of resulted tokens</returns>

Algorithm Intelligent_ Word_Classifier(Tokens[])

For each word in Tokens

- Compare with D-info
- if the word(disease/symptom) exists then
    - Apply the NB's classifier algorithm
    - Analyse and compute the result

Else

    - Discard it

End if

Figure 2 shows the HIM model, contain a component which process the Medline abstracts using several steps. These steps are:

- Which provides the medical abstract to the model
- Text classification and extraction
- Store the classified text in the DB and retrieve the same for the process of identification of diseases
- Analyse D-Info based of ML

Table 1: Symptoms-disease relations

Test no	Symptom 1	Symptom 2	Symptom 3	Disease?
1	Fever	Weakness	Headache	Typhoid
2	Fever	Weakness	Headache	Malaria
3	Fever	Weakness	Headache	Typhoid
4	Fever	Weakness	Headache	Malaria
5	Stomach pain	Weight loss	Loss of appetite	Typhoid
6	Stomach pain	Weight loss	Loss of appetite	Malaria
7	Stomach pain	Weight loss	Loss of appetite	Typhoid
8	Stomach pain	Weight loss	Headache	Malaria
9	Fever	Weight loss	Loss of appetite	Malaria

- Apply the NB's classifier algorithm
- Analyse and compute the result
- And produce the result/report for the future use

Medical abstracts are processed in such a way that in can be manipulated by the machine and it is ready for Machine Learning (ML). The NB's Classifier algorithm processes the D-info and identifies the disease-treatment relations and suggests the users to take decision regarding their inputs.

The main purpose of modern HIM is to assist clinicians at the point of care. This means that a clinician would interact with a HIM to help determine diagnosis, analysis, etc., of patient data. Previous theories of HIM were to use the HIM to literally make decisions for the clinician. The clinician would input the information and wait for the HIM to output the "right" choice and the clinician would simply act on that output.

The main task in decision analysis is to construct decision models, which are mathematical frameworks with graphical representations.

Naïve bayes text classifier-related example: The Naive Bayes classifier selects the most likely classification V_{nb} given the attribute values $a_1, a_2, \ldots a_n$. These attributes are depends upon the user inputs. This results in:

$$V_{nb} = \text{argmax}_{vj \in V} P(v_j) \Pi P(a_i|v_j) \quad (1)$$

We generally estimate $P(a_i|v_j)$ using m-estimates:

$$P(a_i|v_j) = n_c + mp \quad (2)$$

$$n + m \quad (3)$$

where,
n : The number of training examples for which $v = v_j$
n_c : Number of examples for which $v = v_j$ and $a = a_i$
p : A priori estimate for $P(a_i|v_j)$
m : The equivalent sample size

Disease identification: Attributes are Symptom 1, Symptom 2, Symptom 3, and the subject, Disease can be either Typhoid or Malaria.

Data set: The following Table 1 shows that the dataset used in the disease identification. This is an example of symptoms-disease relations.
NB's classifier uses the above dataset finds the probabilities for disease:

- P (Fever|Typhoid)
- P (Weight loss|Typhoid)
- P (Headache|Typhoid)
- P (Fever|Malaria)
- P (Weight loss|Malaria)
- P (Headache|Malaria)

We can estimate these values of P (Typhoid) and P (Malaria), respectively using above Eq. (1).

Typhoid Malaria:

n = 5 n = 5
n-c = 3 n-c = 2
p = 0.5 p = 0.5
m = 3 m = 3

Weight loss:

n = 5 n = 5
n-c = 1 n-c = 3
p = 0.5 p = 0.5
m = 3 m = 3

Headache:

n = 5 n = 5
n-c = 2 n-c = 3
p = 0.5 p = 0.5
m = 3 m = 3

Looking at P (Fever|Typhoid), we have 5 cases where, v_j = Typhoid and in 3 of those cases a_i = Fever. So for P (Fever|Typhoid), n = 5 and n_c = 3. Note that all

Table 2: Probability-disease values

Probability	Disease
0.2000	Malaria
0.1660	Typhoid
0.0833	Dengue
0.1846	Tuberculosis
0.1111	Hepatitis B

attribute are binary (two possible values). We are assuming no other information so, p = 1/ (number-of-attribute-values) = 0.5 for all of our attributes. Our m value is arbitrary, (We will use m = 3) but consistent for all attributes. Now we simply apply the Eq. (3). Using the pre computed values of n, n_c, p and m.

P (Fever | Typhoid) = 3 + 3 ∗ 0.5
5 + 3 = 0.56
P (Fever | Malaria) = 2 + 3 ∗ 0.5
5 + 3 = 0.43
P (Weightloss | Typhoid) = 1 + 3 ∗ 0.5
5 + 3 = 0.31
P (Weightloss | Malaria) = (3 + 3 ∗ 0.5) / (5 + 3) = 0.56
P (Headache | Malaria) = (2 + 3 ∗ 0.5) / (5 + 3) = 0.43
P (Headache | Malaria) = (3 + 3 ∗ 0.5) / (5 + 3) = 0.56

We have P (Typhoid) = 0.5 and P (Malaria) = 0.5, so we can apply Eq. (2). For v = Typhoid, we have

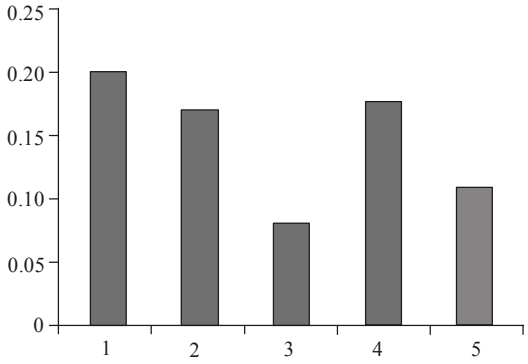

Fig. 3: Disease and probability relationships

(Typhoid) * P (Fever | Typhoid) * P (Weightloss | Typhoid) * P (headache | Typhoid) = 0.5 * 0.56 * 0.31 * 0.43 = 0.037 and for v = Malaria, we have P (Malaria) * P (Fever | Malaria) * P (Weightloss | Malaria) * P (headache | Malaria) = 0.5 * 0.43 * 0.56 * 0.56 = 0.069

Since 0.069>0.037, our example gets classified as 'Malaria' Table 2 Shows that probability values of the disease and probability of the above example.

Form the above data we can say that as the number of symptoms increases the accuracy of the result also

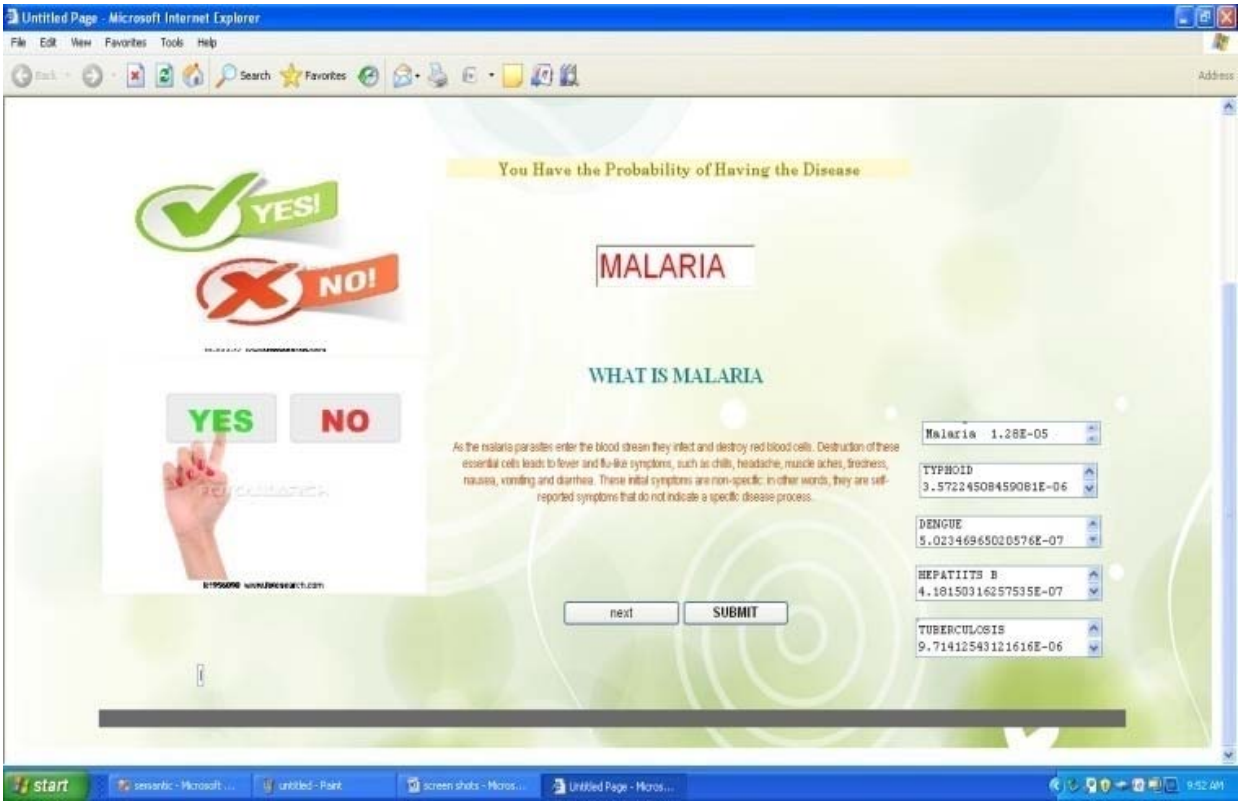

Fig. 4: Screen shots of the web version

increase. The following Fig. 3 shows that disease and probability relationships for the above example. increase. The following Fig. 3 shows that disease and probability relationships for the above example.

The above model is implemented as a web based application using Dot net framework. It will be useful for the human society and provides user friendly interface to access it. The following Fig. 4 shows that one of the screen shots of the web version.

CONCLUSION

Here we have proposed the model that trains the machine and it proves the probabilistic models are stable and reliable to identify the disease-treatment relation. We developed a web based application that automatically find and extract the medical related information. The future product may be e-commerce based product where we can apply various marketing strategies to prove the information that is presented is trustful.

This system may be integrated with web resources like web services and Ontologies (OWL) so that it can available as semantic based web diagnosis system. The domain based ontologies can understand the medical text entered by the users and retrieves the various symptoms-prevent-treatment relations from the various zones.

ACKNOWLEDGMENT

Authors want to thank friends and colleagues at Srinivasa Ramanujan Centre (SRC) SASTRA University for their suggestions and encouragement.

REFERENCES

Azmy, A.M., M.R. Mohamed and I. Erlich, 2005. Decision tree-based approach for online management of fuel cells supplying residential loads. IEEE Power Tech Conference Proceedings, June 2005 St. Petersburg, Russia and Univ. of Duisburg-Essen, Essen, pp: 27-30.

Freund, Y., 1999. An adaptive version of the boost by majority algorithm. Proceeding COLT '99 Proceedings of the 12th Annual Conference on Computational Learning Theory, ACM New York, USA, pp: 102-113.

Harry, Z., 2004. The Optimality of Naive Bayes. Proceedings of the 17th International FLAIRS Conference (FLAIRS2004), pp: 1-7.

Khozoie, N., 2012. Health Information Management on Semantic web: Semantic (HIM). IJWesT, 3(1): 1-8.

Qin, Z., 2006, Naive bayes classification given probability estimation trees. ICMLA '06 Proceedings of the 5th International Conference on Machine Learning and Applications, IEEE Computer Society Washington, DC, USA, pp: 34-42.

Rosario, B. and M.A. Hearst, 2004. Classifying semantic relations in bioscience texts. Proceeding ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Article No. 430, Association for Computational Linguistics Stroudsburg, PA, USA.

Sager, N., M. Lyman, C. Bucknall, N. Nhan and L.J. Tick, 1994. Natural language processing and the representation of clinical data. J. Am. Med. Inform. Assoc., 1(2): 143-144.

Tom, M.M., 2006. The Discipline of Machine Learning. School of Computer Science, Carnegie Mellon University, Pittsburgh, pp: 12.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Retrieved from: http: // cscs. umich. Edu /~ crshalizi/reviews/ vapnik-nature / 2nd Edn., Springer-Verlog, Berlin, pp: 187.