Int. J. Adv. Res. Sci. Technol. Volume 3, Issue3, 2014, pp.140-144 # ICV 5.14

IJARST

International Journal of Advanced Research in Science and Technology

journal homepage: www.ijarst.com

ISSN 2319 – 1783 (Print)

ISSN 2320 – 1126 (Online)

An Adaptive Machine Learning Approach for Semantic Analysis to Extract Medical Knowledge

*Anushaa Putta*and RamaJanaki Devi Ramireddy*

Department of Computer Science and Engineering, Vignan's Institute of Engineering for Women, Visakhapatnam, India.

****Corresponding Author's Email: hasini.view@gmail.com,***

ARTICLEINFO

Article history:

Received 10 Oct. 2014

Accepted 04 Nov. 2014

Available online 06 Nov. 2014

Keywords:

Machine Learning,
Data Mining & Extraction,
Natural language Processing,
Healthcare,
Treatment Deciding System.

ABSTRACT

Machine learning approach over medical datasets is still an important research issue in recent days of technology in medical field. In our approach we are proposing an efficient classification approach for analysis of testing samples with training samples. Initially we train the medical abstracts by identifying disease and treatment and then forwards these informative and non informative sentences towards word of bag (Cure, prevent, etc.) to extract positive and negative sentences and finally update them in database for future classification of testing samples.

© 2014 International Journal of Advanced Research in Science and Technology (IJARST). All rights reserved.

Introduction:

The two tasks that are undertaken in this paper provide the basis for the design of an information technology framework that is capable to identify and disseminate healthcare information. The first task identifies and extracts informative sentences on diseases and treatments topics, while the second one performs a finer grained classification of these sentences according to the semantic relations that exists between diseases and treatments. The problems addressed in this paper form the building blocks of a framework that can be used by healthcare providers (e.g., private clinics, hospitals, medical doctors, etc.), companies that build systematic reviews (hereafter, SR), or laypeople who want to be in charge of their health by reading the latest life science published articles related to their interests. The final product can be envisioned as a browser plug-in or a desktop application that will automatically find and extract the latest medical discoveries related to disease-treatment relations and present them to the user. The product can be developed and sold by companies that do research in Healthcare Informatics, Natural Language Processing, and Machine Learning, and companies that develop tools like Microsoft Health Vault.

The value of the product from an e-commerce point of view stands in the fact that it can be used in marketing strategies to show that the information that is presented is trustful (Medline articles) and that the results are the latest discoveries. For any type of business, the trust and interest of customers are the key success factors. Consumers are looking to buy or use products that satisfy their needs and gain their trust and confidence. Healthcare products are probably the most sensitive to the trust and confidence of consumers. Companies that want to sell information technology healthcare frameworks need to build tools that allow them to extract and mine automatically the wealth of published research. For example, in frameworks that make recommendations for drugs or treatments, these recommendations need to be based on acknowledged discoveries and published results, in order to gain the consumers' trust. The product value also stands in the fact that it can provide a dynamic content to the consumers, information tailored to a certain user (e.g., a set of diseases that the consumer is interested in).

The first process i.e. task A or sentence selection which identifies sentences from Medline published abstracts that talk about diseases and treatments. The task is similar to a scan of sentences contained in the abstract of an journal in order to present to the user-only sentences that are identified as containing relevant information disease-treatment information.

The second process i.e. task B or relation identification, which has a deeper semantic dimension and it is focused on identifying disease-treatment relations in the sentences already selected as being informative (e.g., task A is applied first). We focus on three relations: Cure, Prevent, and Side Effect, Remedies, diet, Suggestion from Experts (doctors) a subset of the eight relations that the corpus is annotated with. We decided to focus on these three relations

because these are most represented in the corpus while for the other five, very few examples are available. Table 1 presents the original data set, the one used by Rosario and Hearst that we also use in our research. The numbers in parentheses represent the training and test set size. For example, for Cure relation, out of 810 sentences present in the data set, 636 are used for training and 174 for testing. The approach used to solve the two proposed tasks is based on NLP and ML techniques. In a standard supervised ML setting, a training set and a test set are required. The training set is used to train the ML algorithm and the test set to test its performance. The objectives are to build models that can later be deployed on other test sets with high performance. Extracting informative sentences is a task by itself in the NLP and ML community. Research fields like summarization and information extraction are disciplines where the identification of informative text is a crucial task. The contributions and research value that are brought with this task stand in the usefulness of the results and the insights about the experimental settings for the task in the medical domain.

For the first process, the data sets are annotated with the following information: a label indicating that the sentence is informative, i.e., containing disease-treatment information, or a label indicating that the sentence is not informative. Table 2 gives an example of labeled sentences. For the second process, the sentences have annotation information that states if the relation that exists in a sentence between the disease and treatment is Cure, Prevent, or Side Effect. These are the relations that are more represented in the original data set and also needed for our future research. We would like to focus on a few relations of interest and try to identify what predictive model and representation technique bring the best results.

The task of identifying the three semantic relations is expressed in two ways: 1. Three models are built. Each model is focused on one relation and can distinguish sentences that contain the relation from sentences that do not. This setting is similar to a two-class classification task in which instances are labeled either with the relation in question (Positive label) or with non-relevant information (Negative label); Setting 2. One model is built, to distinguish the three relations in a three-class classification task where each sentence is labeled with one of the semantic relations.

Table1: Data Set description:

Relationship	Definition and Example
Cure 810(636,174)	Treat cures Disease Intravenous immune globulin for recurrent spontaneous abortion
Only Treat 606(490,116)	Disease not mentioned
Only Disease 616(492, 124)	Treat not mentioned
Prevent 63(50, 13)	Treat prevents the Distains for prevents of stroke

Literature Review:

Machine learning is a subfield of computer science (CS) and artificial intelligence (AI) that deals with the construction and study of systems that can learn from data, rather than follow only explicitly programmed instructions. Besides CS and AI, it has strong ties to statistics and optimization, which deliver both methods and theory to the field. Machine learning is employed in a range of computing tasks where designing and programming explicit, rule-based algorithms is infeasible. Example applications include spam filtering, optical character recognition (OCR), search engines and computer vision. Machine learning, data mining, and pattern recognition [citation needed] are sometimes conflated.

Machine learning tasks can be of several forms. In supervised learning, the computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs. Spam filtering is an example of supervised learning, in particular classification, where the learning algorithm is presented with email (or other) messages labeled beforehand as "spam" or "not spam", to produce a computer program that labels unseen messages as either spam or not.

In unsupervised learning, no labels are given to the learning algorithm, leaving it on its own to groups of similar inputs (clustering), density estimates or projections of high-dimensional data that can be visualized effectively.

Table2: Example of annotated sentence for the sentence selection task.

Label	Sentence
Informative sentence	Urgent colonoscopy for the diagnosis and treatment of severe hemorrhage.
Non-informative sentence	In all cases a testing diagnosis study was performed.

Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end. Topic modeling is an example of unsupervised learning, where a program is given a list of human language documents and is tasked to find out which documents cover similar topics. In reinforcement learning, a computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle), without a teacher explicitly telling it whether it has come close to its goal or not.

Machine learning algorithms can be organized into a taxonomy based on the desired outcome of the algorithm or the type of input available during training of the machine.[citation needed]. Supervised learning algorithms are trained on labeled examples, i.e., input where the desired output is known. The supervised

learning algorithm attempts to generalize a function or mapping from inputs to outputs which can then be used speculatively to generate an output for previously unseen inputs.

Unsupervised learning algorithms operate on unlabeled examples, i.e., input where the desired output is unknown. Here the objective is to discover structure in the data (e.g. through a cluster analysis), not to generalize a mapping from inputs to outputs.

Semi-supervised learning combines both labeled and unlabeled examples to generate an appropriate function or classifier .Transduction, or transductive inference, tries to predict new outputs on specific and fixed (test) cases from observed, specific (training) cases .Reinforcement learning is concerned with how intelligent agents ought to act in an environment to maximize some notion of reward. The agent executes actions which cause the observable state of the environment to change. Through a sequence of actions, the agent attempts to gather knowledge about how the environment responds to its actions, and attempts to synthesize a sequence of actions that maximizes a cumulative reward.

Learning to learn learns its own inductive bias based on previous experience. Developmental learning, elaborated for robot learning, generates its own sequences (also called curriculum) of learning situations to cumulatively acquire repertoires of novel skills through autonomous self-exploration and social interaction with human teachers, and using guidance mechanisms such as active learning, maturation, motor synergies, and imitation. Machine learning algorithms can also be grouped into generative models and discriminative models.

Research Approach & Objectives:

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. *Association rule learning* is a method for discovering interesting relations between variables in large databases.

An *artificial neural network* (ANN) learning algorithm, usually called "neural network" (NN), is a learning algorithm that is inspired by the structure and functional aspects of biological neural networks. Computations are structured in terms of an interconnected group of artificial neurons, processing information using a connectionist approach to computation. Modern neural networks are non-linear statistical data modeling tools. They are usually used to model complex relationships between inputs and outputs, to find patterns in data, or to capture the statistical structure in an unknown joint probability distribution between observed variables.

Inductive logic programming (ILP) is an approach to rule learning using logic programming as a uniform representation for input examples, background knowledge, and hypotheses. Given an encoding of the known background knowledge and a set of examples represented as a logical database of facts, an ILP system will derive a hypothesized logic program which entails all the positive and none of the negative examples. Inductive programming is a related field that considers any kind of programming languages for representing hypotheses (and not only logic programming), such as functional programs.

Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other.

A *Bayesian network*, belief network or directed acyclic graphical model is a probabilistic graphical model that represents a set of random variables and their conditional independencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. Efficient algorithms exist that perform inference and learning.

Reinforcement learning is concerned with how an agent ought to take actions in an environment so as to maximize some notion of long-term reward. Reinforcement learning algorithms attempt to find a policy that maps states of the world to the actions the agent ought to take in those states. Reinforcement learning differs from the supervised learning problem in that correct input/output pairs are never presented, nor sub-optimal actions explicitly corrected.

Sparse Dictionary Learning, a datum is represented as a linear combination of basic functions, and the coefficients are assumed to be sparse. Let x be a d-dimensional datum, D be a d by n matrix, where each column of D represents a basis function. r is the coefficient to represent x using D. Mathematically, sparse dictionary learning means the following $x \sim D * r$ where r is sparse. Generally speaking, n is assumed to be larger than d to allow the freedom for a sparse representation. Methods for sparse dictionary learning include K-SVD.

Sparse dictionary learning has been applied in several contexts. In classification, the problem is to determine which classes a previously unseen datum belongs to. Suppose a dictionary for each class has already been built. Then a new datum is associated with the class such that it's best sparsely represented by the corresponding dictionary. Sparse dictionary learning has also been applied in image de-noising. The key idea is that a clean image path can be sparsely represented by an image dictionary, but the noise cannot.

Natural language processing (NLP) is a field of computer science, artificial intelligence, and linguistics

concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation.

Bag of Words Representation:

The bag of words (BOW) representation is commonly used for text classification tasks. It is a representation in which features are chosen among the words that are present in the training data. Selection techniques are used in order to identify the most suitable words as features. After the feature space is identified, each training and test instance is mapped to this feature representation by giving values to each feature for a certain instance. Two most common feature value representations for BOW representation are: binary feature values are 1. The value of a feature can be either zero or one. 2. Where one represents the fact that the feature is present in the instance and 0 otherwise, or frequency feature values, the value of the feature is the number of times it appears in an instance, or 0 if it doesn't appear

We deal with short texts with an average of 20 words per sentence the difference between a binary value representation and a frequency value representation is not large. In our case, we choose a frequency value representation. This has an advantage that if a feature appears more than once in a sentence, this means that it is important and frequency value representation will capture this, the features value will be feature than that of other features. The selected features are words delimited by spaces and simple punctuation marks such as (,) , [,] . We keep only the words that appeared at least 3 times in the training collection, contain at least 1 alpha numeric character, are not part of an English list of stop words[1] and are longer than 3 characters. The frequency threshold of 3 is commonly used for text collection because it removes non-informative feature and also strings of characters that might be the result of a wrong tokenization when splitting the text into words. Words that have length two or one character or not consider as feature because of two other reasons; possible incorrect tokenization and problems with very short acronyms in the medical domain that could me highly ambiguous could be an acronym or an abbreviation of a common word).

Classification:

Naive Bayesian classification

For classification purpose we are using naïve Bayesian classification, in this approach we will forward the testing samples over training samples and compute the posterior probability for both positive class label and negative class labels.

The following algorithm shows Naive Bayesian classification. Each data sample is represented by an n-dimensional feature vector,

X=(x1, x2, x3.....x n), depicting n measurements made on the sample from n attributes, respectively A, A2, A3....An.

- Suppose that there are m classes, C1, C2, C3.....Cm, given an unknown data sample, X which has no label class, the classifier will predict that X belongs to the class having the highest posterior probability conditioned on X. Then Naive Bayesian classifier assigns an unknown sample X to the class Ci, if and only if

P(Ci|X) > P(Cj|X) for 1 < j < m,
j not equal to **i**
Then **P(Ci|X) = P(X|Ci) P(Ci) / P(X)**
As **P(X)** is constant for all classes, only
P(X|Ci) P(Ci) need to be maximized.
Prior probability can be estimated by **P(Ci) = si/s.**
where si = no of training samples,
s = total no of samples.

Flexibility:

Preliminary investigation examine project flexibility, the likelihood the system will be useful to the organization. The main objective of the flexibility study is to test the Technical, Operational and Economical flexibility for adding new modules and debugging old running system. All system is flexible if they are unlimited resources and infinite time. There are aspects in the flexibility study portion of the preliminary investigation:

- Economical Flexibility
- Technical Flexibility
- Operational Flexibility

Economic Flexibility:

As system can be developed technically and that will be used if installed must still be a good investment for the organization. In the economical feasibility, the development cost in creating the system is evaluated against the ultimate benefit derived from the new systems. Financial benefits must equal or exceed the costs.

The system is economically feasible. It does not require any addition hardware or software. Since the interface for this system is developed using the existing resources and technologies java SDK 1.6 open source, there is nominal expenditure and economical feasibility for certain.

Operational Feasibility:

Proposed projects are beneficial only if they can be turned out into information system. That will meet the organization's operating requirements. Operational feasibility aspects of the project are to be taken as an

important part of the project implementation. Some of the important issues raised are to test the operational feasibility of a project includes the following: -

- Is there sufficient support for the management from the users?
- Will the system be used and work properly if it is being developed and implemented?
- Will there be any resistance from the user that will undermine the possible application benefits?

This system is targeted to be in accordance with the above-mentioned issues. Beforehand, the management issues and user requirements have been taken into consideration. So there is no question of resistance from the users that can undermine the possible application benefits.

The well-planned design would ensure the optimal utilization of the computer resources and would help in the improvement of performance status.

Evaluation:

The fast responses from the system are expected by the user. To increase the quality of response, the interface must provide effective feedback, perfect training to the system and predictable system responses.

Results:

It is an efficient machine learning approach by classifying training samples from the short texts or medical abstracts, initially abstracts can be segmented into sentences and finds informative and non informative sentences based on the disease and treatment relationship, then it can be forwarded to relation identification based on word of bags to extract the useful sentences from medical abstracts and these extracted sentences can be updated to training dataset.

To analyze the testing sample initially sentences can be preprocessed by eliminating the unnecessary information from raw sentences and extracts disease, treatment and relation identification from preprocessed sentences and then these sentences can be forwarded for classification to classify the testing samples by using training samples.

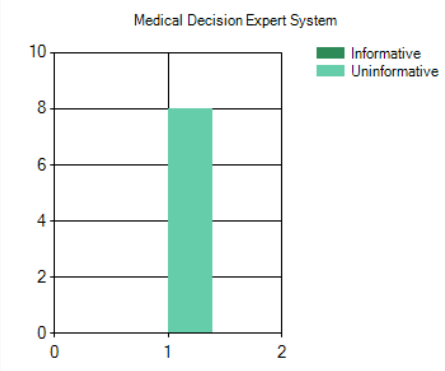

Fig: 1. Results for testing set

Conclusion:

This paper proposed an research work with relationally extracted sentences (including semantic relations) from the medical abstracts and separates' the sentences into informative and non-informative sentences by using bag of words and that data is been saved in the database and will be used as an sample data in testing phase, an machine learning approach for testing samples using training datasets with Naïve Bayesian classification algorithm. Our experimental result shows efficient and relevant information extraction on sample medical abstracts than the traditional approaches. And it also helps the user and medical representatives, and it is also good for large and mission-critical projects, it also takes less amount of time to extract informative sentences from journals and documents.

References:

1. J. Ginsberg, H.Matthew, S.P. Rajah, B. Lynnette, S.S Mark, and L. Brilliant, "Detecting Influenza Epidemics Using Search Engine Queer Data, " *Nature, vol.457, pp.1012-1014, feb-2009.*
2. M. Craven, "Learning to Extract Relations from Medline," Prof. Assoc. for the Advancement of Artificial Intelligence, 1999.
3. M. Goad rich, L. Oliphant, and J. Slavic, "Learning Ensembles First-Order Clauses for Recall-Precision, " Proc. 14th Int'l Conf. Indy Logic Programming, 2004.
4. G. Leroy, and J.D Martinez, "A Shallow Parser Based on Closed-Class Words to Capture Relations in Biomedical Text," J. *Biomedical Informatics*, vol.36, no.3, pp. 145-158,2003.
5. R. Buns, R. Mooney, Y. Weiss, and J. Platt, "Subsequence Kernels for Relation Extraction, " Advances in Neural Information Processing Systems, vol.18, pp. 171-178, 2006.
6. O. Frunze and D. Ink pen, "Textual Information in Predicting Functional Properties of the Genes, " Proc. Workshop current Trends in Biomedical Natural Language Processing(Bio NLP) in conjunction with assoc. For computational Linguistics (ACL'08), 2008.
7. R. Buns and R. Mooney, "A Shortest Path Dependency Kernel for Relation Extraction, "Proc. Conf. Human Language Technology and Empirical Methods in Natural Processing (HLT/EMNLP), pp. 724-731, 2005.

About Authors:

Anushaa Putta is a M.TECH student of Department of Computer Science and Engineering, in VIGNAN'S Institute of Engineering for Women, Visakhapatnam, Andhra Pradesh, India

Rama Janaki Devi Ramireddy, Masters in Computer Science & Engineering specialized in AI&R, Andhra University. Currently working as Assistant Professor in Department of Computer Science & Engineering, VIGNAN'S Institute of Engineering for Women, Visakhapatnam, Andhra Pradesh, India.