Volume 4, Issue 2, February 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Hybrid Machine Learning Implementations for Classifying Disease-Treatment Relations in Short Texts

Kandimalla.Srinivasarao*
Master of Technology
Department of CSE
Dvr & Dr.Hs Mic College Of Technology
Jntuk, Andhra Pradesh, India.

RamaSatish.A
Assoc. Professor
Department of CSE
Dvr & Dr.Hs Mic Ollege Of Technology
Jntuk, Andhra Pradesh,India.

Abstract: Machine Learning (ML) field is gaining reputation in almost any domain of research and just recently has become a reliable tool in the medical domain. ML is envisioned as a tool by which computer-based systems can be integrated in the healthcare field in order to get a better, more efficient medical care. This empirical domain of automatic learning drives the creation of intelligent and automated application that assists health care personals to undertake task such as medical decision support, medical imaging, protein-protein interaction, extraction of medical knowledge, and for overall patient management care. This paper describes a Hybrid ML-based methodology that is fused with an SVM classifier in combination with Bag-of-Words Representation and NLP tasks for building an application that is capable of identifying and disseminating healthcare information. In its crude form it extracts sentences from medical information sources such as published medical papers, patient case sheets that mention diseases and treatments, and identifies semantic relations that exist between the diseases and treatments. This fundamental approach obtains reliable outcomes that could be integrated in an application to be used in the medical care domain. An implementation of the proposed approach validates the claim.

Key words: Natural Language Processing, SVM Classifiers, Electronic Health care systems, decision-based models, Machine Learning.

I. INTRODUCTION

Data mining is the process of extracting and discovering information in the form of patterns in large data sets involving methods at the intersection of artificial Intelligence, Machine learning and data base system applications. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). The traditional healthcare [4] [8] system is also becoming one that embraces the Internet and the electronic world. Electronic Health Records (hereafter, EHR) are becoming the standard in the healthcare domain.

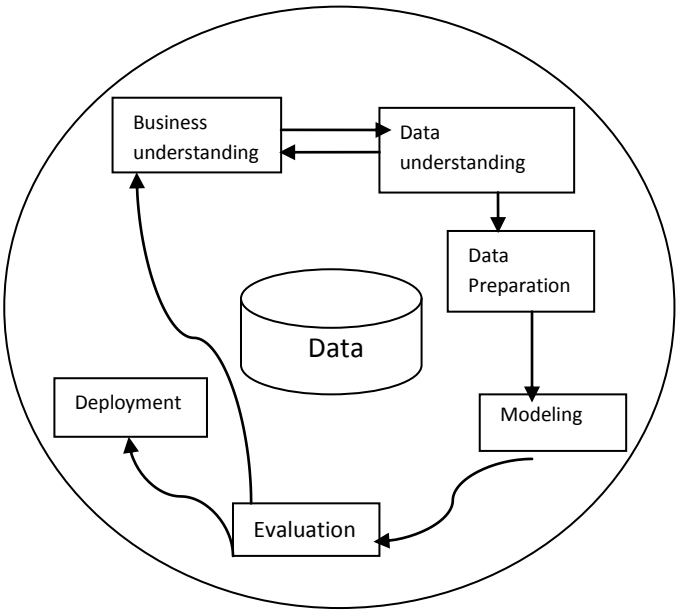

Figure 1: Data mining procedure from different primitives.

Pervasive computing is the concept that incorporates computation in working and living environment in such a way so that the interaction between human and computational devices such as mobile devices or computers becomes extremely natural and the user can get multiple types of data in a totally transparent manner.

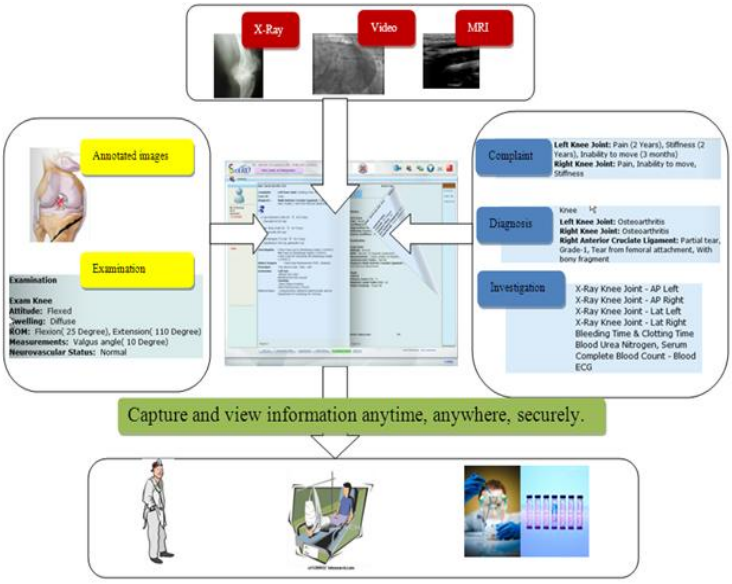

Figure 2: Case sheet classification for fast understanding of patient's profile[12].

In order to embrace the views that the EHR system has, we need better, faster, and more reliable access to information. In the medical domain, the richest and most used source of information is Medline a database of extensive life science published articles. All research discoveries come and enter the repository at high rate (Hunter and Cohen, making the process of identifying and disseminating reliable information a very difficult task. [1] [2] [3] The work that we present in this paper is focused on two tasks: automatically identifying sentences published in medical abstracts (Medline) as containing or not information about diseases and treatments, and automatically identifying semantic relations that exist between diseases and treatments, as expressed in these texts. The second task is focused on three semantic relations: *Cure, Prevent, and Side Effect.*

The tasks that are addressed here are the foundation of an information technology framework that identifies and disseminates healthcare information. People want fast access to reliable information and in a manner that is suitable to their habits and workflow. Medical care related information (e.g., published articles, clinical trials, news, etc.) is a source of power for both healthcare providers and laypeople.[5] [7] Using the analysis of the application property regions in articles we describe the machine learning approach for extracting and building an application that is capable of automated identification and dissemination of healthcare information[1] [5]. This machine learning approach consists following data representations such as NLP, classification algorithms, Bag-of-Words Representation, NLP and Biomedical Concepts Representation, Medical Concepts (UMLS) 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. NLP tasks that has sentence selection and relation identification. The tasks addressed in our research are information extraction and relation extraction.[2] [4] [7] From the wealth of research in these domains, we are going to mention some representative works. The task of relation extraction or relation identification is previously tackled in the medical literature, but with a focus on biomedical tasks: sub cellular location, gene-disorder association, and diseases and drugs. Usually, the data sets used in biomedical specific tasks use short texts, often sentences. [1] [2] This is the case of the first two related works mentioned above. The tasks often entail identification of relations between entities that co-occur in the same sentence. In this paper we describe different process generations of Pub Med articles based on Medline constraints. We apply SVM and CNB Classifiers for extracting relevant data sets comparison process with suitable process generations.

II. BACKGROUND WORK

The data set consists of sentences from Medline five abstracts annotated with disease and treatment entities and with weight semantic relations between diseases and treatments. The main focus of their work is on entity recognition for diseases and treatments. The authors use Hidden Markov Models [11] and maximum entropy models to perform both the task of entity recognition and the relation discrimination. Their representation techniques are based on words in context, part of speech information, phrases, and a medical lexical ontology—Mesh six terms. Compared to this work, our research is focused on different representation techniques, different classification models, and most importantly generates improved results with less annotated data. [7] [8] [9] Syntactic rule-based relation extraction systems are complex systems based on additional tools used to assign POS tags or to extract syntactic parse trees. It is known that in the

biomedical literature such tools are not yet at the state-of-the-art level as they are for general English texts, and therefore their performance on sentences is not always the best. A good comparison of different syntactic parsers and their contribution to extracting protein-protein interactions can be found. Statistical methods tend to be used to [5] [6] [7] solve various NLP tasks when annotated corpora are available. Rules are automatically extracted by the learning algorithm when using statistical approaches to solve various tasks. In general, statistical techniques can perform well even with little training data. For extracting relations, the rules are used to determine if a textual input contains a relation or not. Taking a statistical approach to solve the relation extraction problem from abstracts, the most used representation technique is bag-of-words. [3] [4] It uses the words in context to create a feature vector and other researchers combined the features of BOW, extracted from sentences with other sources of information like POS used two sources of information: sentences in which the relation appears and the local context of the entities, and showed that simple representation techniques bring good results.

III. EXISTING SYSTEM

Tasks and Data Sets:

The potential for pervasive computing is evident in almost every aspect of our lives including the hospital, emergency and critical situations, industry, education, or the hostile battlefield. The use of this technology in the field of health and wellness is known as pervasive health care. 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.

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. [4] [8] the product value also stands in the fact that it can provide a dynamic content to the consumers, information tailored to a certain user. The pipeline is similar to a hierarchy of tasks in which the results of one task is given as input to the other. We believe that this can be a solution for identifying and disseminating relevant information tailored to a specific semantic relation because the second task is trying a finer grained classification of the sentences that already contain information about the relations of interest.

Classification Algorithms and Data Representations:

In ML, as a field of empirical studies, the acquired expertise and knowledge from previous research guide the way of solving new tasks. The models should be reliable at identifying informative sentences and discriminating disease-treatment semantic relations. The research experiments need to be guided such that high performance is obtained.

Table 1: Data Sets Used for the two tasks.

	Training		Test	
	Positive	**Negative**	**Positive**	**Negative**
Cure	554	531	276	266
Prevent	42	531	21	266
SideEffect	20	531	10	266

There are at least two challenges that can be encountered while working with ML techniques. One is to find the most suitable model for prediction. The ML field offers a suite of predictive models (algorithms) that can be used and deployed. The task of finding the suitable one relies heavily on empirical studies and knowledge expertise. The second one is to find a good data representation and to do feature engineering because features strongly influence the performance of the models. Identifying the right and sufficient features to represent the data for the predictive models, especially when the source of information is not large, as it is the case of sentences, is a crucial aspect that needs to be taken into consideration.

IV. PROPOSED SYSTEM

As classification algorithms, we use a set of two representative models: Complement Naive Bayes (CNB), which is adapted for text with imbalanced class distribution with a linear classifier support vector machine (SVM) with polynomial kernel, and a classifier that always predicts the majority class in the training data (used as a baseline). We decided to use these classifiers because they are representative for the learning algorithms [4] [5] [6] in the literature and were shown to work well on both short and long texts. Decision trees are decision-based models similar to the rule-based models that are used in handcrafted systems, and are suitable for short texts. In this sequence of process with data representations.

SVM ALGORITHM:
Candidate SV = { closest pair from opposite classes }
while
there are violating points
do
Find a violator
Candidate SV = candidate SV U violator
if any $\alpha_p < 0$ due to addition of c to S
then
candidate SV = candidate SV / *p*
repeat till all such points are pruned
end if
end while

Figure 3: Algorithm for SVM Classifier process generation.

SVM has following advantages.

- Prediction accuracy is generally high
- Robust, works when training examples contain errors
- Fast evaluation of the learned target function

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. [4] [2] 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.

NLP and Biomedical Concepts Representation:

The second type of representation is based on syntactic information: noun-phrases, verb-phrases, and biomedical concepts identified in the sentences. Experiments are performed when using as features only the final set of identified noun-phrases, only verb-phrases, only biomedical entities, and with combinations of all these features. When combining the features, the feature vector for each instance is a concatenation of all features.

Medical Concepts (UMLS) Representation:

In order to work with a representation that provides features that are more general than the [6] [9] words in the abstracts (used in the BOW representation), we also used the Unified Medical Language system (hereafter, UMLS) concept representations. UMLS is a knowledge source developed at the US National Library of Medicine (hereafter, NLM) and it contains a metathesaurus, a semantic network, and the specialist lexicon for biomedical domain.

V. EXPERIMENTAL RESULTS

The test set on which the models are evaluated contain the true classes and the evaluation tries to identify how many of the true classes were predicted by the model classifier. In the ML settings, special attention needs to be directed to the evaluation measures that are used. For data sets that are highly imbalanced (one class is [6] [2] overrepresented in comparison with another), standard evaluation measures like accuracy are not suitable. Because our data sets are imbalanced, we chose to report in addition to accuracy, the macro averaged F-measure.

For taking the datasets as follows in the sequence of downloaded from Pub Med data articles. Download 2700 articles from National Center for Biological Institution using Med line process abstracts. Test data is 2700 retrieved from original data sets.

Specify Training record Count: 50
Specify data record count: 5

Table 2: Training and testing datasets comparison results.

Abstract Number	SVM Classifier	CNB Classifier
1	95	92
2	90	94
3	97	90
4	90	90
5	96	97

Fig.4 presents the best results obtained so far. An increase of almost 5 percentage points, for both accuracy and F-measure is obtained when using as representation features biomedical entities extracted by the Genia tagger and [6] [7] CNB as classifier. An increase in results for the other classifiers can be also observed. This increase can be caused by the fact that, when present in sentences, the biomedical entities have [5] a stronger predicting value. The entities identify better if a sentence is informative or not and this is something that we would hope to happen.

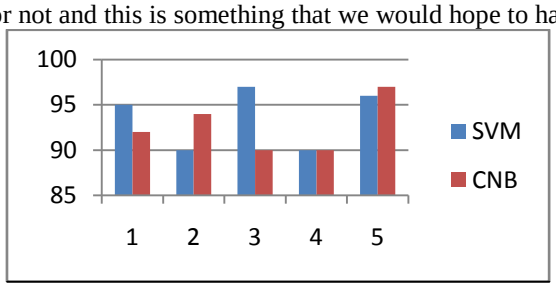

Figure 4: Accuracy and F-measure results when using training and testing medical datasets.

The improvement over the other settings can be due to the fact that the combination of classifier and features has a good predicting value for a model trained on the three relations. The results show that probabilistic models based on Naıve Bayes formula, obtain good results. [1] [4] the fact that the SVM classifier performs well shows that the current discoveries are in line with the literature. These two classifiers have always been shown to perform well on text classification tasks. Even though the independence of features is violated when using Naive Bayes classifiers, they still perform very well. The best results obtained are: 98 percent F-measure for the class Cure, 100 percent F-measure for the class Prevent, and 75 percent F-measure for the Side Effect class.

VI. CONCLUSION

Machine Learning (ML) field is gaining reputation in almost any domain of research and just recently has become a reliable tool in the medical domain. We included the sentences that did not contain any of the three relations in question and the results were lower than the one when we used models trained only on sentences containing the three relations of interest. These discoveries validate the fact that it is crucial to have the first step to weed out uninformative sentences, before looking deeper into classifying them. This paper describes a Hybrid ML-based methodology that is fused with an SVM classifier in combination with Bag-of-Words Representation and NLP tasks for building an application that is capable of identifying and disseminating healthcare information. In its crude form it extracts sentences from medical information sources such as published medical papers, patient case sheets that mention diseases and treatments, and identifies semantic relations that exist between the diseases and treatments.

REFERENCES

[1] Oana Frunze, Diana Inkpen, and Thomas Tran, "A Machine Learning Approach For Identifying Disease-Treatment Relations in Short Texts", IEEE Transactions On Knowledge and Data Engineering, Vol. 23, No. 6, June 2011.

[2] L. Hunter, Z. Lu, J. Firby, W.A. Baumgartner Jr., H.L. Johnson, P.V. Ogren, and K.B. Cohen, "OpenDMAP: An Open Source, Ontology-Driven Concept Analysis Engine, with Applications to Capturing Knowledge Regarding Protein Transport, Protein Interactions and Cell-Type-Specific Gene Expression," BMC Bioinformatics, vol. 9, article no. 78, Jan. 2008.

[3] M. Yusuke, S. Kenji, S. Rune, M. Takuya, and T. Jun'ichi, "Evaluating Contributions of Natural Language Parsers to Protein-Protein Interaction Extraction," Bioinformatics, vol. 25, pp. 394-400, 2009.

[4] S. Novichkova, S. Egorov, and N. Daraselia, "MedScan, A Natural Language Processing Engine for MEDLINE Abstracts," Bioinformatics, vol. 19, no. 13, pp. 1699-1706, 2003.

[5] T. Mitsumori, M. Murata, Y. Fukuda, K. Doi, and H. Doi, "Extracting Protein-Protein Interaction Information from Biomedical Text with SVM," IEICE Trans. Information and Systems, vol. E89D, no. 8, pp. 2464-2466, 2006.

[6] C. Giuliano, L. Alberto, and R. Lorenza, "Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature," Proc. 11th Conf. European Chapter of the Assoc. for Computational Linguistics, 2006.

[7] J. Ginsberg, H. Mohebbi Matthew, S.P. Rajan, B. Lynnette, S.S. Mark, and L. Brilliant, "Detecting Influenza Epidemics Using Search Engine Query Data," Nature, vol. 457, pp. 1012-1014, Feb. 2009.

[8] S.V.N. Vishwanathan, M. Narasimha Murty, "SSVM : A Simple SVM Algorithm", In *Advances in Neural Information Processing Systems, (NIPS*2000)*, volume 13. NIPS, Cambridge MA: MIT Press, 2001.

[9] S. V. N. Vishwanathan and M. Narasimha Murty. Geometric SVM: A fast and intuitive SVM algorithm. Technical Report IISC-CSA-2001-14, Dept. of CSA, Indian Institute of Science, Bangalore, India, November 2001. Submitted to ICPR 2002.

[10] L. Hunter and K.B. Cohen, "Biomedical Language Processing: What's beyond PubMed?" Molecular Cell, vol. 21-5, pp. 589-594, 2006.

[11] S. Ray and M. Craven, "Representing Sentence Structure in Hidden Markov Models for Information Extraction," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI '01), 2001.

[12] https://www.google.co.in/search?q=er+diagram+for+user+filter+message&source=lnms&tbm=isch&sa=X&ei=Yv_6UvPwCIbUrQeV_4HIBA&ved=0CAcQ_AUoAQ&biw=1366&bih=624#q=patient+case+sheet+fot+mri+scan&tbm=isch.

