ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

Hybrid Machine Learning Implementations for Classifying Cure-Ailments Results

[1] B.Srujana, [2]Dr. M.R Narsinga Rao
[1] M Tech, [2]Professor
[1,2]K.L University, Vaddeswaram, Guntur (dt).

Abstract: Machine Learning (ML) implementations can be found in many domains for automation and just recently has become a reliable tool in the medical domain too. ML is envisioned as a tool by which computer-based systems can be integrated in the healthcare sector in order to get a better, faster and more efficient medical care. This empirical domain of automatic learning drives the creation of intelligent and automated applications that assists health care personals to undertake tasks such as medical decision support, medical imaging, protein-protein interaction, extraction of medical knowledge, and an overall patient management systems. 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 elementary 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.

Index Terms: *machine learning, natural language processing, classification algorithms, bag-of-words (BOW) representation.*

I. INTRODUCTION

A growing body of recent work in information extraction has addressed the problem of relation extraction (RE), identifying relationships between entities stated in text, such as LivesIn(Person, Location) or Employed By(Person, Company). Supervised learning has been shown to be effective for RE (Zelenko et al., 2003; Culotta and Sorensen, 2004; Bunescu and Mooney, 2006); however, annotating large corpora with examples of the relations to be extracted is expensive and tedious. Pervasive computing is the concept that incorporates computation in our 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. 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. Tools that can help us manage and better keep track of our health such as Google Health1 and Microsoft HealthVault2 are reasons and facts that make people more powerful when it comes to healthcare knowledge and management. The traditional healthcare system is also becoming one that embraces the Internet and the electronic world.

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. 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

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

relations that exist between diseases and treatments, as expressed in these texts. This paper describes a Machine Learning(ML)-based methodology for building an application that is capable of automated identification and dissemination of healthcare information. It extracts sentences from published medical papers that mention diseases and treatments, and identifies semantic relations that exist between diseases and treatments. Our evaluation results for these tasks show that the proposed methodology obtains reliable outcomes that could be integrated in an application to be used in the medical care domain.

We envision the potential and value of the findings of our work as guidelines for the performance of a framework that is capable to find relevant information about diseases and treatments in a medical domain repository. The results that we obtained show that it is a realistic scenario to use NLP and ML techniques to build a tool, similar to an RSS feed, capable to identify and disseminate textual information related to diseases and treatments. Therefore, this study is aimed at designing and examining various representation techniques in combination with various learning methods to identify and extract biomedical relations from literature.

In this paper we also 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 elementary 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.

II. RELATED WORK

The tasks addressed in our research are information extraction and relation extraction. 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 (Craven, [4]), gene-disorder association (Ray and Craven, [23]), and diseases and drugs (Srinivasan and Rindflesch, [26]). Usually, the data sets used in biomedical specific tasks use short texts, often sentences. 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.

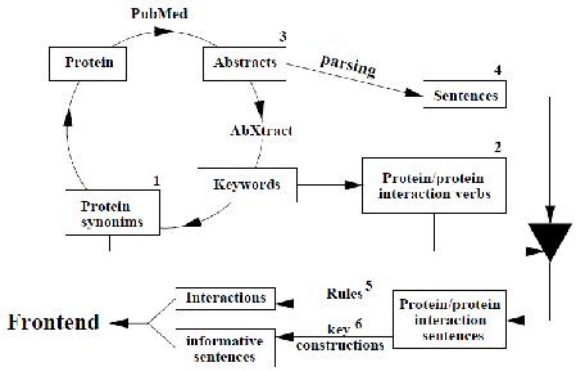

Fig 1: Extraction protein interaction from scientific text.

The abstracts are analyzed using the Abstract system. A new search may be done in Medline with the new set of synonymous protein routines.

III. EXISTING SYSTEM

a) Tasks and Data Sets

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 first task (task 1 or sentence selection) identifies sentences from Medline published abstracts that talk about diseases and treatments. The second task (task 2 or relation identification) has a deeper semantic dimension

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

and it is focused on identifying disease-treatment relations in the sentences already selected as being informative (e.g., task 1 is applied first).

b) Classification Algorithms and Data Representations

The models should be reliable at identifying informative sentences and discriminating disease- treatment emantic relations. The research experiments need to be guided such that high performance is obtained. The experimental settings are directed such that they are adapted to the domain of study (medical knowledge) nd to the type of data we deal with (short texts or sentences), allowing for the methods to bring improved performance.

As classification algorithms, we use a set of six representative models: decision-based models (Decision trees), probabilistic models (Naı¨ve Bayes (NB) and Complement Naı¨ve Bayes (CNB), which is adapted for text with imbalanced class distribution), adaptive learning (Ada- Boost), 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 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. Probabilistic models, especially the ones based on the Naı¨ve Bayes theory, are the state of the art in text classification and in almost any automatic text classification task. Adaptive learning algorithms are the ones that focus on hard-to-learn concepts, usually underrepresented in the data, a characteristic that appears in our short texts and imbalanced data sets. SVM-based models are acknowledged state-of-the-art classification techniques on text.

c) 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

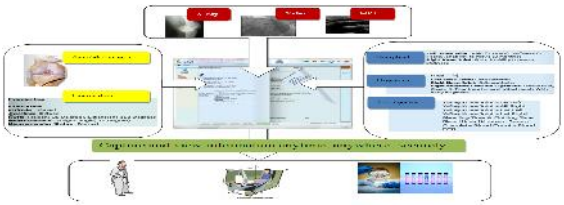

Fig 2: Automatic medical case sheet classification

We propose and discuss pros and cons of certain ML settings used these days such as NLP, classification algorithms, Bag-of-Words Representation, NLP and Biomedical Concepts Representation, Medical Concepts (UMLS) Representation. The potential value of this paper stands in using any one of the ML settings and developing an automated application that can understand and classify disease treatment classifications in shorter texts.

IV. PROPOSED WORK

Earlier approaches use the following classification algorithms. Decision trees are decision-based models similar to the rule-based models that are used in handcrafted systems, and are suitable for short texts. Probabilistic models, especially the ones based on the Naıve Bayes theory, are the state of the art in text classification and in almost any automatic text classification task. Adaptive learning algorithms are the ones that focus on hard-to-learn concepts, usually underrepresented in the data, a characteristic that appears in our short texts and imbalanced data sets. SVM-based models are acknowledged state-of-the-art classification techniques on text. It is evident that SVM-based models are most recent and advanced classification techniques that are optimized to yield better performance irrespective to the voluminous of data required for training sets. SVM has following advantages.

1. Prediction accuracy is generally high

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

2. Robust, works when training examples contain errors
3. Fast evaluation of the learned target function.

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 We propose to implement a hybrid model that uses an SVM classification in combination with Bag-of-Words Representation and NLP tasks. This combinatorial hybrid approach implements the best of all the three methods for effective disease treatment classifications in shorter texts at better time frames and performances of individual approaches.

V. PERFORMANCE ANALYSIS

The most common used evaluation measures in the ML settings are: accuracy, precision, recall, and F-measure. All these measures are computed form a confusion matrix (Kohavi and Provost that contains information about the actual classes, the true classes and the classes predicted by the classifier. 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 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.

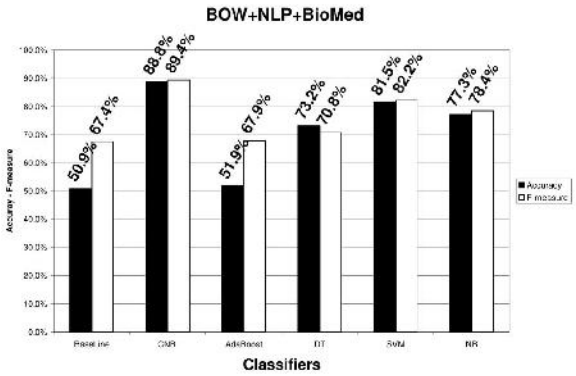

Fig 3: Accuracy and F-measure results when using BOW, NLP, and biomedical features, task 1.

For all the other classifiers, an increase in results can be observed as well. The results obtained by only a BOW representation can be further improved when these features are combined with the noun-phrases, verb-phrases, and biomedical concepts. This combinatorial hybrid approach implements the best of all the three methods for effective disease treatment classifications in shorter texts at better time frames and performances of individual approaches. By using our proposed technique we have to describe the efficient results with NLP and BOW classifiers.

VI. CONCLUSION

This paper describes a Machine Learning(ML)-based methodology for building an application that is capable of automated identification and dissemination of healthcare information. We propose and discuss pros and cons of certain ML settings used these days such as NLP, classification algorithms, Bag-of-Words Representation, NLP and Biomedical Concepts Representation, Medical Concepts (UMLS) Representation. The potential value of this paper stands in using any one of the ML settings and developing an automated application that can understand and classify disease treatment classifications in shorter texts. NLP tasks that has sentence selection and relation identification We propose to implement a hybrid model that uses an SVM classification in combination with Bag-of-Words Representation and NLP tasks. This combinatorial hybrid approach implements the best of all the three methods for effective disease

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

treatment classifications in shorter texts at better time frames and performances of individual approaches.

VII. REFERENCES

[1] 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.
[2] 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.
[3] M. Ould Abdel Vetah, C. Ne´dellec, P. Bessie`res, F. Caropreso, A.-P. Manine, and S. Matwin, "Sentence Categorization in Genomics Bibliography: A Naive Bayes Approach," Actes de la Journe´e Informatique et Transcriptome, J.-F. Boulicaut and M. Gandrillon, eds., Mai 2003.
[4] J. Pustejovsky, J. Castan˜ o, J. Zhang, M. Kotecki, and B. Cochran, "Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations," Proc. Pacific Symp. Biocomputing, vol. 7, pp. 362-373, 2002.
[5] 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.
[6] T.C. Rindflesch, L. Tanabe, J.N. Weinstein, and L. Hunter, "EDGAR: Extraction of Drugs, Genes, and Relations from the Biomedical Literature," Proc. Pacific Symp. Biocomputing, vol. 5, pp. 514-525, 2000.