Feature Extraction for Disease Prediction Using Machine Learning Techniques

D. Vishnu Vardhan Raju[1], Dr. M. Humera Khanam[2], A. Khudhus[3]
[1]*M.Tech, PG Scholar, Dept. Of CSE, Sri Venkateswara University College of Engineering, Sri Venkateswara University, Tirupathi.*
[2]*Associate Professor, Dept. Of CSE, Sri Venkateswara University College of Engineering, Sri Venkateswara University, Tirupathi.*
[3]*M.Tech, Tirupathi*

***Abstract*- Feature selection is an important technique for data mining. Despite its importance, most studies of feature selection are restricted to batch learning. Unlike traditional batch learning methods, online learning represents a promising family of efficient and scalable machine learning algorithms for large-scale applications. The machine learning field has gained its momentum in almost any domain of research and just recently has become a reliable tool in the medical domain .Machine Learning (ML) is envisioned as a tool by which computer-based systems can be integrated in the health care field in order to get more efficient medical care. An ML-based methodology is described to build an application that is capable of identifying disseminating health care information .It extracts sentences from published medical papers that mention diseases and treatments, and identifies semantic relations that exist between diseases and treatments. In this paper we have focused on Machine Learning (ML) techniques and the Classification algorithms that are suitable to use for identifying and classifying relevant medical information in short texts. The encouraging results of our experiments validate the efficiency of the proposed techniques.**

***Index Terms*- Big data Analytics, Machine Learning , Health Care.**

I. INTRODUCTION

People care deeply about their health and want to be, now more than ever, in charge of their health and healthcare. Life is more hectic than has ever been, the medicine that is practiced today is an Evidence-Based Medicine (hereafter, EBM) in which medical expertise is not only based on years of practice but on the latest discoveries as well [1]. Tools that can help us manage and better keep track of our health such as Google Health and Microsoft HealthVault 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. Electronic Health Records (hereafter, EHR) are becoming the standard in the healthcare domain [4]. Researches and studies show that the potential benefits of having an EHR system are:

Health information recording and clinical data repositories [10]**:** immediate access to patient diagnoses, allergies, and lab test results that enable better and time-efficient medical decisions.

Medication management**:** rapid access to information regarding potential adverse drug reactions, immunizations, supplies, etc.

Decision support**:** the ability to capture and use quality medical data for decisions in the workflow of healthcare.

Obtain treatments that are tailored to specific health needs**:** rapid access to information that is focused on certain topics.

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 [2], a database of extensive life science published articles. All research discoveries come and enter the repository at high rate, making the process of identifying and disseminating reliable information a very difficult task. The work presented here 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 [5]. 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 [8]. Studies reveal that people are searching the web and read medical related information in order to be informed about their health. The proposed objective for this work is to show the different Natural Language Processing (NLP) and Machine Learning (ML) techniques [11] are used for the representation of information and classification algorithms that are suitable to use for identifying and classifying relevant medical information in short texts.

II. RELATED WORK

The data set consists of sentences from Medline abstracts annotated with disease and treatment entities and with eight semantic relations between diseases and treatments. All the way the information is provided as abstracts in form of the documents. It is possible that the abstracts are made largely available to the viewers,but it is inconvenient to measure the ratio of facts , our research is focused on different representation techniques, different classification models, and most importantly generates improved results with less annotated data.

Most existing studies of online learning require accessing all the attributes/features of training instances. Such a classical setting is not always appropriate for real-world applications when data instances are of high dimensionality or it is expensive to acquire the full set of attributes/features. To address this limitation, we investigate the problem of Online Feature Selection (OFS) in which an online learner is only allowed to maintain a classifier involved only a small and fixed number of features. The key challenge of Online Feature Selection is how to make accurate prediction for an instance using a small number of active features. This is in contrast to the classical setup of online learning where all the features can be used for prediction.

1) Prediction of heart disease at early stage using data mining and big data analytics: A survey

In this paper, the various technologies of data mining (DM) models for forecast of heart disease are discussed. Data mining plays an important role in building an intelligent model for medical systems to detect heart disease (HD) using data sets of the patients, which involves risk factor associated with heart disease. Medical practitioners can help the patients by predicting the heart disease before occurring. The large data available from medical diagnosis is analyzed by using data mining tools and useful information known as knowledge is extracted. Mining is a method of exploring massive sets of data to take out patterns which are hidden and previously unknown relationships and knowledge detection to help the better understanding of medical data to prevent heart disease. There are many DM techniques available namely Classification techniques involving Naive Byes (NB), Decision tree (DT), Neural network (NN), Genetic algorithm (GA), Artificial intelligence (AI) and Clustering algorithms like K-NN, and Support vector machine (SVM). Several studies have been carried out for developing prediction model using individual technique and also by combining two or more techniques. This paper provides a quick and easy review and understanding of available prediction models using data mining from 2004 to 2016. The comparison shows the accuracy level of each model given by different researchers.

2) The eICU Research Institute - Collaboration between Industry, Health-Care Providers, and Academia

In this paper as the volume of data that is electronically available proliferates, the health-care industry is identifying better ways to use this data for patient care. Ideally, these data are collected in real time, can support point-of-care clinical decisions, and, by providing instantaneous quality metrics, can create the opportunities to improve clinical practice as the patient is being cared for. The business-world technology supporting these activities is referred to as business intelligence, which offers competitive advantage, increased quality, and operational efficiencies. The health-care industry is plagued by many challenges that have made it a latecomer to

business intelligence and data-mining technology, including delayed adoption of electronic medical records, poor integration between information systems, a lack of uniform technical standards, poor interoperability between complex devices, and the mandate to rigorously protect patient privacy. Efforts at developing a health care equivalent of business intelligence (which we will refer to as clinical intelligence) remains in its infancy. Until basic technology infrastructure and mature clinical applications are developed and implemented throughout the health-care system, data aggregation and interpretation cannot effectively progress. The need for this approach in health care is undisputed. As regional and national health information networks emerge, we need to develop cost-effective systems that reduce time and effort spent documenting health-care data while increasing the application of knowledge derived from that data.

III. METHODOLOGY

In this paper focus is on different representation techniques, different classification models, and most importantly generates improved results with less annotated data. The tasks addressed in our research are information extraction and relation extraction. From the wealth of research in these domains, There are three major approaches used in extracting relations between entities Co-occurrences analysis, Rule based approaches, Statistical methods. Statistical methods tend to be used to 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

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.

IV. SYSTEM ARCHITECTURE

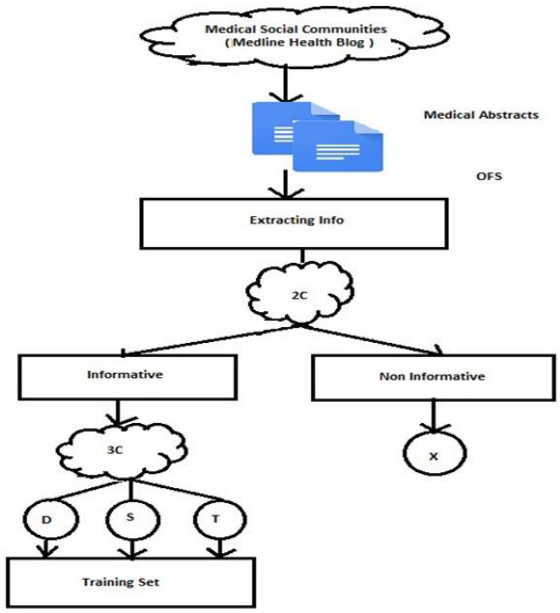

Fig 3.1 Architecture

Figure 3.1 shows the details of the online performance of the compared online feature selection algorithms with varied fractions of selected features. The proposed OFS algorithms outperform the other two baselines for most cases.

This encouraging result further verifies the efficiency of the proposed technique Application Modules.

V. ALGORITHM

BEGIN

1.Collect abstracts

A={a1,a2,ai,…}

2.Pre-processing

aiHtml=toHtml(ai_xml)

3.Text extraction

aiText=extract(siHtml)

4.2c - ct<-r_stop_words(aiText)

5.If 2c(Informative) then

6.3c-feature_extraction >--->D(d1,d2,di,..),T(t1,t2,ti…),S(s1,s2,si,…)

7.Else

8.Non-informative

9.goto step 3

10.End if

11.Update Training Set Ts<-D,T,S

12.Calculate F-measure
13.Display R
END

VI. MODULE DESCRPTION

MODULE 1: INFORMATION EXTRACTION
The first task (task 1 or sentence selection) 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 article in order to present to the user-only sentences that are identified as containing relevant information (disease treatment information)

MODULE 2: RELATION IDENTIFICATION
The second task (task 2 or relation identification) 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 1 is applied first). We focus on three relations: Cure, Prevent, and Side Effect, a subset of the eight relations that the corpus is annotated with.

VII. EXPERIMENTAL RESULTS

In this paper we investigate the problem of Online Feature Selection (OFS) in which an online learner is only allowed to maintain a classifier involved only a small and fixed number of features. The key challenge of Online Feature Selection is how to make accurate prediction for an instance using a small number of active features.

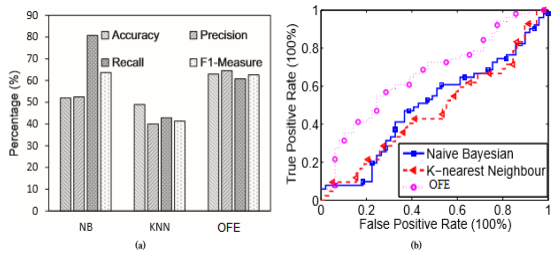

FIGURE 7.1. (a) Comparison of accuracy, precision, recall and F1-Measure for NB, KNN and OFE, in which NB naive Bayesian, KNN k-nearest neighbour, and OFE online feature extraction (b) ROC curves for NB, KNN and OFE.

Moreover,compared to the several prediction algorithms the prediction accuracy of our proposed algorithm reaches 94% i.e the better is the feature description of the disease,the higher the accuracy will be.

VIII. CONCLUSION

This project suggest that domain-specific knowledge improves the results. Probabilistic models are stable and reliable for tasks performed on short texts in the medical domain. The representation techniques influence the results of the ML algorithms, but more informative representations are the ones that consistently obtain the best results.

FUTURE ENHANCEMENT

As future work, we would like to extend the experimental methodology when the first setting is applied for the second task, to use additional sources of information as representation techniques, and to focus more on ways to integrate the research discoveries in a framework to be deployed to consumers. In addition to more methodological settings in which we try to find the potential value of other types of representations, we would like to focus on source data that comes from the web. Identifying and classifying medical-related information on the web is a challenge that can bring valuable information to the research community and also to the end user.

REFERENCES

[1] M. Goadrich, L. Oliphant, and J. Shavlik, "Learning Ensembles of First-Order Clauses for Recall-Precision Curves: A Case Study in Biomedical Information Extraction," Proc. 14th Int'l Conf. Inductive Logic Programming, 2004.

[2] M. Craven, "Learning to Extract Relations from Medline," Proc. Assoc. for the Advancement of Artificial Intelligence, 1999.

[3] 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.

[4] P. B. Jensen, L. J. Jensen, and S. Brunak``Mining electronic health records:Towards better research applications and clinical care," NatureRev. Genet., vol. 13, no. 6, pp. 395405, 2012.

[5] A.M. Cohen and W.R. Hersh, and R.T. Bhupatiraju, "Feature Generation, Feature Selection, Classifiers, and Conceptual Drift for Biomedical Document Triage," Proc. 13th Text Retrieval Conf. (TREC), 2004.

[6] I. Donaldson et al., "PreBIND and Textomy: Mining the Biomedical Literature for Protein-Protein Interactions Using a Support Vector Machine," BMC Bioinformatics, vol. 4, 2003.

[7] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky, "GENIES: A Natural Language Processing System for the Extraction of Molecular Pathways from Journal Articles," Bioinformatics, vol. 17, pp. S74-S82, 2001.

[8] O. Frunza and D. Inkpen, "Textual Information in Predicting Functional Properties of the Genes," Proc. Workshop Current Trends in Biomedical Natural Language Processing (BioNLP) in conjunction with Assoc. for Computational Linguistics (2008).

[9] R. Gaizauskas, G. Demetriou, P.J. Artymiuk, and P. Willett, "Protein Structures and Information Extraction from Biological Texts: The PASTA System," Bioinformatics, vol. 19, no. 1, pp. 135-143, 2003.

[10] 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.

[11] R. Kohavi and F. Provost, "Glossary of Terms," Machine Learning, Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, vol. 30,pp. 271-274, 1998.

[12] R.Bunescu and R.Mooney,"A Shortest Path Dependency Kernel for Relation Extraction," Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing (HLT/ EMNLP), pp. 724-731, 2005.