A Survey of Machine Learning Approaches for Relation Classification from Biomedical Texts

Remya K R¹, Ramya J S²
¹PG Student, ²Asst.Professor, MCET, Anad, Trivandrum

Abstract— The Machine learning field has many applications in almost every area. In this digitized world, everything has to be automated to improve quality, time complexity, accuracy etc. In the biomedical area, information is mainly in natural language text format. The biomedical researchers need fast information accessing tools for extracting useful information from huge amount of biomedical repositories. This paper provides a survey of machine learning methods and various natural language text representations that are used currently. Two main challenges in the area of machine learning are the selection of a good learning algorithm and feature representation technique. The study is mainly concentrated in relationship extraction from biomedical text. This will help to understand appropriate feature representation and machine learning technique suitable for the biomedical domain.

Keywords— classifiers, feature selection, machine learning, multiple classifiers, relation extraction.

I. INTRODUCTION

Machine learning means the construction and study of systems that can learn from data. It is a method of teaching computers to make and improve predictions or behaviors based on some data. Machine learning is a huge field with hundreds of different algorithms for solving different problems. Machine learning poses challenging problems in terms of algorithmic approach, data representation, computational efficiency, and quality of the resulting program.

Relation extraction from natural language text is a long standing research area. Biomedical information and its updates are stored in natural language text format. Information is growing considerably every day. With the increased amount of biomedical publications, it is becoming more and more challenging to access useful and relevant information about a specific topic. All research discoveries come and enter the repository at high rate (Hunter and Cohen [12]), making the process of identifying and disseminating reliable information a very difficult task. Manual inspection of such large amount of data will be very difficult and time consuming.

Also, since the information is in natural language text which is not structured, efficient natural language processing methods need to be applied in order to map the information into structured useful format. In this paper, a survey of different machine learning techniques and feature representation techniques for biomedical literature are conducted.

II. RELATED WORKS FOR RELATION EXTRACTION

The biomedical relation extraction is an active interesting research area. There are many related researches undergone in this field. Some of the prominent works are discussed here. Usually, the data sets used in biomedical specific tasks use short texts, often sentences.

In the work done by Rosario and Hearst [3], Hidden Markov models are used for entity recognition. This includes mapping biomedical information into structural representation. It converts natural language text into structural format. Their work uses machine learning for information extraction. The extraction of medical abstract is obtained through text classification. Semantic lexicons of words labelled with semantic classes so associations can be drawn between words which helps in extracting the necessary sentences related to the query. In this research paper the author used sentence co-occurrence and naive bayes algorithm to extract semantic relation like Gene-Protein from medline abstract.

In [2] the authors C. Giuliano, L. Alberto, and R. Lorenza, used a combination of kernel functions to represent two distinct information sources: the global context where entities appear and their local contexts. The whole sentence where the entities appear is used to discover the presence of a relation between two entities. The problem considered here is that of identifying interactions between genes and proteins from biomedical literature. They performed experiments on extracting gene and protein interactions from two different data sets. The first data set used in the experiments is the Almed corpus. There was no distinction between genes and proteins and the relations are symmetric. They used SVM for using kernel methods.
The gene/protein interactions are annotated without distinguishing the type and roles of the two interacting entities. The second one was, Learning Language in Logic (LLL) challenge on Genic Interaction extraction. A shallow linguistic kernel is constructed by combining the two types of kernel functions and found the combined kernel outperformed the basic kernels.

Another [5] paper proposes to use kernel-based learning methods to automatically extract biomedical relations from literature text. The authors modified the standard tree kernel function by incorporating a trace kernel to capture richer contextual information. In their experiments on a biomedical corpus, they compared different kernel functions for biomedical relation detection and classification. The experimental results showed that a tree kernel outperformed word and sequence kernels for relation detection, the new trace-tree kernel outperformed the standard tree kernel, and a composite kernel outperformed individual kernels for relation extraction.

In [6] a multiple classifier system is employed with majority voting. When a multiple classifier system is employed, one of the most popular methods to accomplish the classifier fusion is the simple majority voting. However, when the performance of the ensemble members is not uniform, the efficiency of this type of voting generally results affected negatively. In this paper, new functions for dynamic weighting in classifier fusion are introduced. Experimental results with several real-problem data sets from the UCI Machine Learning Database Repository demonstrate the advantages of these novel weighting strategies over the simple voting scheme. It improved the accuracy of the overall system.

Oana Frunza.et.al [1] in her work performed two task in pipelined manner for identifying and extracting the relationship between the given MEDLINE abstract. First task involves finding most suitable model for prediction, the second task is to find good data representation. To achieve this two task various predictive algorithm and textual representation techniques are considered. A set of six classification algorithm namely decision based models, probabilistic models(Naïve Bayes, Complement Naïve Bayes), Adaptive learning, linear classifier namely support vector machine and a classifier that always predicts the majority class in training data are used. Three representation technique namely Bag-Of-Word representation, NLP and Biomedical Concept representation and Medical concept representation are used for feature representation. When using all the features that is, the three representation techniques are combined, probabilistic classifiers performed more accurate.

There are many other works related to relation classification. The next section discusses about various feature representation techniques that can be used for biomedical texts.

III. BIOMEDICAL RESOURCES AND ANNOTATED CORPORA

For biomedical information, the primary resource is MEDLINE which is a bibliographic database of life sciences and biomedical information. The database contains more than 21.6 million records from 5,639 selected publications covering biomedicine and health from 1950 to the present. It includes bibliographic information for articles from academic journals covering medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care. Compiled by the United States National Library of Medicine (NLM), MEDLINE is freely available on the Internet and searchable via PubMed and NLM's National Center for Biotechnology Information's Entrez system[15]. There are many sources for annotated MEDLINE abstracts.

Annotation of biomedical text adds some information labeled with the biomedical text that will help the data mining process. Annotation can improve the learning process. The GENIA corpus[13] is a collection of Medline abstracts which is intended to represent the literature of molecular biology. It contains 1,999 MEDLINE abstracts retrieved using the MeSH terms “human,” “blood cells,” and “transcription factors.” The GENIA corpus is currently the most thoroughly annotated collection of MEDLINE abstracts. It is annotated for part-of-speech, syntax, coreference, biomedical concepts and events, cellular localization, disease-gene associations, and pathways. Other sources are BioCreAtIve collections [18,19] and PennBioIE corpus[21,24]. The BioCreAtIve Databases contains sequence and annotation data for genes and proteins. The PennBioIE corpus contains 1100 abstracts for cytochrome P-450 enzymes and 1157 oncology abstracts with annotations for paragraphs, sentences, tokens, parts-of-speech, syntax, and biomedical entities.

IV. FEATURE REPRESENTATION

Feature selection is very important in machine learning. Important features are used to distinguish between the relation types or entity pairs. From the related study, it is to be stated that, the more the features are included in the representation, the more accurate will be the classification.
A. Bag of Words Representation

In BOW representation, the features are represented by their presence or absence. The binary values 1 and 0 are used to indicate whether a feature is present or not present respectively. The features are selected from the words in the training data using different selection methods. It is the most commonly effective representation techniques used in many classification tasks. That is it represents the entire document as a sequence of words that are present in the document. Sometimes frequency information is also taken into consideration. In that case, if a feature is present, its value is indicated by its frequency of occurrence, otherwise the value will be zero. This contains more information that the feature having largest value is occurred more in the document and hence is important. Ould et al. [20] experimented with bag-of-word features on sentences.

B. NLP and Biomedical Concepts Representation

This is based on the noun phrases, verb phrases and biomedical concepts representation. That is it takes into account of the syntactic information. There are specialized tools available for extracting such features. The Genia tagger is tool which analyzes English sentences and outputs the base forms, part-of-speech tags, chunk tags, and named entity tags. General-purpose part-of-speech taggers do not usually perform well on biomedical text because lexical characteristics of biomedical documents are considerably different from those of newspaper articles, which are often used as the training data for a general-purpose tagger. The GENIA tagger is trained not only on the Wall Street Journal corpus but also on the GENIA corpus and the PennBioIE corpus, so the tagger works well on various types of biomedical documents. It requires some preprocessing steps before using the feature set; for example, removing the stop words, removing the features that contain only punctuation etc. In this method, the features represented will be the noun phrases or verb phrases or biomedical entities.

C. Medical Concepts (UMLS) Representation

UMLS is a knowledge source developed at the US National Library of Medicine. It contains over 1 million medical concepts and 5 million concept names. It contains a metathesaurus, a semantic network and a specialist lexicon for the biomedical domain. The Metathesaurus is a very large, multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships among them.

It is built from the electronic versions of many different thesauri, classifications, code sets, and lists of controlled terms used in patient care, health services billing, public health statistics, indexing and cataloguing biomedical literature, and /or basic, clinical, and health services research. The Metathesaurus is organized by concept or meaning. Its purpose is to link alternative names and views of the same concept together and to identify useful relationships between different concepts. All concepts in the Metathesaurus are assigned to at least one semantic type from the Semantic Network. This provides consistent categorization of all concepts in the Metathesaurus at the relatively general level represented in the Semantic Network. Many of the words and multi-word terms that appear in concept names or strings in the Metathesaurus also appear in the SPECIALIST Lexicon. The lexical tools are used to generate the word, normalized word, and normalized string indexes to the Metathesaurus. For example, UMLS contains six meanings for “cold.” One is an abbreviation expansion (chronic obstructive lung disease). However, each of these meanings carries a unique concept identifier (CUI) within the UMLS that can be invoked to represent a context-independent statement. MetaMap is a highly configurable program developed by Dr. Alan (Lan) Aronson at the National Library of Medicine (NLM) to map biomedical text to the UMLS Metathesaurus or, equivalently, to discover Metathesaurus concepts referred to in text. MetaMap uses a knowledge-intensive approach based on symbolic, natural-language processing (NLP) and computational-linguistic techniques.

V. RELATION CLASSIFICATION

Relation Classification task aims to identify occurrences of particular types of relationships between pairs of given entities and classify them according to the type of relationship. This task is already done for the medical literature with a focus on biomedical tasks such as: subcellular location (Craven, [4]), gene-disorder association (Ray and Craven, [23]), and diseases and drugs (Srinivasan and Rindflesch, [26]). For relation classification mainly there are three types of approaches. The Co-occurrence based analysis, rule based approaches, and statistical learning methods.

A. Co-Occurrence Analysis

The co-occurrence analysis is based on the fact that if two entities both occur in the same sentence or in the same document, then there exists some relation between the two entities.
It is based only on lexical knowledge and words in context. Based on the word co-occurrences, a relation can be identified. It is the simplest method of identifying relations between biomedical entities in which we have to collect instances where the entities co-occur. If the entities are repeatedly mentioned together, then there is a greater chance that they may be related in some way, although the type and direction of this relation typically cannot be determined by co-occurrence statistics alone. Since it is very simple and flexible it is widely used for relation classification and can achieve high recall. However, since it can capture little syntactic or semantic information, co-occurrence analysis cannot distinguish relation types, and often achieves low precision. The work done by Jenssen et al. [14] and Stapley and Benoit [27] are good examples.

B. Rule Based Approaches

Rule based approaches are flexible in terms of the applicable domain. Researchers have to manually develop rules based on syntactic or semantic information to parse relations from text. There are Syntactic rule based systems, semantic rule based systems and a Hybrid system which consider both syntactic and semantic information. Syntactic information such as part-of-speech and syntax structures can be represented via data structures such as parse trees. Syntax parsing approaches extensively utilize syntactic information and rely on syntactic rules for relation extraction and generally achieve poor precision for relation extraction. The works done by Thomas et al. [28], Yakushiji et al. [29], and Leroy et al. [16] are based on syntactic rule based systems. In [11] Miyao et al. , different semantic parsers are compared for extracting protein-protein interaction.

In semantic rule based systems, certain semantic indicators are used which are selected by domain experts. A semantic parser represents these indicator words as various relation types. A pair of biomedical entities whose contextual information satisfies a certain predefined semantic type is identified as a relation. Since these are largely based on domain-specific lexicons, they have lower portability across different domains. Works done by Friedman et al. [16], Pustejovsky et al. [17], Rindflesch et al. [22] are examples of semantic rule based systems.

In hybrid approaches, syntactic analysis takes place first to create possible parses from the original sentence. Next, they eliminate incorrect parses and identify domain words (such as genes) based on semantic information. McDonald, Chen, Su, and Byron (2004)[25] combined access to syntactic and semantic information via a single grammar and reported higher precision and recall.

Such hybrid parsers maintain both the flexibility of syntax parsing and high precision of semantic analysis. Another example is the works of Gaizauskas et al. [28] and Novichkova et al. [29]. However, rule-based approaches require manual encoding of syntactic and semantic rules based on domain knowledge, which is very labor intensive and time-consuming.

C. Statistical Methods

Using statistical methods 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. Relation extraction can be formulated as a text-classification problem. Each entity pair from a sentence is regarded as a candidate relation instance. Sentences that contain multiple entities can derive multiple relation instances. Statistical methods require less training data and can be used when annotated corpora are available. Statistical methods are the currently most using supervised machine learning algorithms. And the most used are SVM with different kernel functions. The works [30][31][32] are examples of different kernel functions applied for relation extraction.

VI. DISCUSSIONS AND FUTURE WORK

Different algorithms have its own advantages and disadvantages depending on the application domain. There is no algorithm that is best suited to every situation. A given algorithm may perform better than all others for a specific subset of situations. Here we discuss about some of the algorithm's advantages and disadvantages.

The decision trees are easy to implement, easy to understand, and offer many advantages compared to other decision-making tools. One of the decision tree advantages are its listing comprehensive information and all possible solutions to an issue. The reliability of the information in the decision tree depends on feeding the precise internal and external information at the onset. Even a small change in input data can at times, cause large changes in the tree. Changing variables, excluding duplication information, or altering the sequence midway can lead to major changes and might possibly require redrawing the tree. Decision trees are also prone to errors in classification, owing to differences in perceptions and the limitations of applying statistical tools.
Neural networks offer a number of advantages, including requiring less formal statistical training, ability to implicitly detect complex nonlinear relationships between dependent and independent variables, ability to detect all possible interactions between predictor variables, and the availability of multiple training algorithms. Disadvantages include its "black box" nature, greater computational burden, proneness to over-fitting, and the empirical nature of model development. Neural networks, modify their parameters automatically in response to prepared stimuli and expected response. This allows them to produce much behaviour. It only keeps the current "state" of the approximation, which is updated as learning occurs. Rather than remembering what happened and how to react to it, neural networks build a sort of "model" of their "world." The model tells them how to react to certain inputs, even if the inputs are something that it has never seen before.

Adaptive boosting uses a committee of weak base classifiers to vote on the class assignment of a sample point. The base classifiers can be decision stumps, decision trees, SVMs, etc. It takes an iterative approach. On each iteration - if the committee is in agreement and correct about the class assignment for a particular sample, then it becomes down weighted (less important to get right on the next iteration), and if the committee is not in agreement, then it becomes up weighted (more important to classify right on the next iteration). Adaboost is known for having good generalization. It improves accuracy of classification. Goadrich et al. [12] used inductive logic techniques for information extraction from Medline abstracts.

Another system currently used for improving accuracy is multiple classifier system or classifier ensemble systems. Our future work will implement MCS for improving classifier accuracy. Also we will use a representation technique which is including all of the important features. That is as used by OanaFrunza and Diana Inkpen[1] in their work, a combination of different representation techniques are used. The richer, more informative representation technique can bring better performance results. Different algorithms have different decision points. Using MCS can reduce the local different behaviours of individual classifiers by averaging the results of each member classifier. So we hope both this techniques can improve the performance to a better level. The conditions that are to be noted when using multiple classifier system are, the classifiers used should be accurate and diverse. Each classifier should perform better than a minimum performance. Combining identical classifiers is useless. The classifiers should be different, and should make different errors for the same data.

For comparing the performances of different classifiers a tool called WEKA [32] can be used. It is open source data mining software that contains many algorithms for classification, regression, association rule mining etc.

VII. CONCLUSION

Different machine learning techniques and their advantages and disadvantages when applied to biomedical literature are discussed here. Different feature representation techniques are also listed. For two class classification tasks SVM is found to be the best and for multiple classification problems, Naïve Bayesian techniques are performed well. And we suggest multiple classifier system for improving the accuracy of the system. In the multiple classifier system different member classifiers used should have different errors independent of each other and should have performance better than a minimum level. That is, for the method to be successful each member classifier should keep a minimum level of disagreement. By averaging the results of a large number of such classifiers, decision boundary can be approximated with some accuracy. That is uncorrelated errors of individual classifiers can be eliminated by averaging.

REFERENCES


