IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Classifying gene disease entities using Relevance Vector Machine Classifier

S. Vijaya[1], R. Radha[2]

[1]Dept. of Computer Science, S.D.N.B. Vaishnav College for Women, Chromepet, Chennai, India
[2]Dept. of Computer Science, S.D.N.B. Vaishnav College for Women, Chromepet, Chennai, India

Available online at: www.ijcseonline.org

Accepted: 16/Jul/2018, Published: 31/July/2018

***Abstract*:** An increased interest has been noticed in Named Entity Recognition, particularly in the field of biomedical domain as identifying and extracting biomedical entities such as genes, diseases, proteins and drugs are tedious and demanding due to its ambiguity in biomedical terms. Named entity Recognition task consists of two phases. The first phase recognizes and extracts the entities whereas the second phase classifies the extracted entities under its associated classes. This research work is focused on the second phase of NER that is classifying the extracted entities with its associated class. In order to classify the entities, Relevance Vector Machine is trained and tested on two different datasets. For comparison purpose, HMM and SVM methods have been applied on the same datasets. The evaluation results shows that the RVM classifier performs better than HMM and SVM with high accuracy and less period of execution time.

***Keywords* :** Named Entity Recognition, Biomedical domain , Biomedical entities, Entities Classification, RVM

I. INTRODUCTION

In this recent era, interest on automatic extraction of information using Named Entity Recognition task is increased exponentially. General NER methods classify the entities into broad categories such as Person, location, time and organization[1]. These NER are useful in many areas of research that uses natural language text. More generalized domain dependent categorization such as genes, disease names , proteins and drugs would play a vital role in automated extraction of useful information from large amount of biomedical literature, automated construction of Ontologies and Question Answering applications. Finding the gene that is associated with a particular disease is of great support in improving the treatment of disease as the same drug works differently on the type of genes [2]. For instance, if the gene that is associated with 'Breast Cancer' is known, it would assist Oncologists to propose the suitable treatment for better result.

As the Support Vector Machine(SVM) takes highest time complexity[3], The RVM method has been employed as a supervised machine learning technique for various high performance. Owing to its relaxation over feature independence assumptions NER system has the benefit of dealing with high dimensional arbitrary feature sets in more efficient way over existing machine learning techniques such as Hidden Markov Models(HMM)[4] and Support Vector Machine(SVM)[5].

Rest of the paper is organized as follows, Section II contains the related work , Section III discusses the modelling of the system, Section IV presents evaluation measures, section V describes the experiment results and Section VI concludes research work.

II. RELATED WORK

To extract entities from Biomedical texts without relying on handcrafted rules, heuristics methods or annotated data, S.Zhang and N.Elhadad[6] has been proposed an unsupervised approach and experimented the approach on Clinical notes and biological literature.

Christopher Bowd et al.,[7] used SVM and RVM methods to classify healthy, normal and Glaucomatous eyes and concluded that both the classifiers classified with high accuracy. Though SVM directly reduces the classification error and resulted higher accuracy the time complexity is more than RVM.

Stefan Andelic et al.[8] analyzed the impact of named entities on text classification using multiple classifiers. Rebholz-Schuhman et al. [9], proposed Silver Standard Corpus(SSC) as an alternative to Gold Standard Corpus(GSC) as the production of GSC is expensive and time-consuming.

Michael E.Tipping[10] used RVM to obtain sparse solutions for both regression and classification by utilizing lineal model in the parameters. Michael Fleischman and Eduard Hovy[1] presented a supervised learning model to classify named entities with global semantic information.

II.1 Hidden Markov Model(HMM):

Hidden Markov models are usually used in applications such as speech recognition and the alignment of biological sequence. Rong xu et al[11], used HMM for sentence classification. The sentences were categorized based on the Background, objective, the methods used result and conclusion part. As a result, it was concluded that the HMM achieved better performance than Naïve Bayes and Maximum Entropy [11] whereas , the accuracy of gene disease classification is less than RVM.

II.2 Support Vector Machine (SVM)

Support Vector Machine is a well-known classification method because of its implementation method with high accuracy. To minimize Classification Error the SVM method does not require a statistical data model. This method can be applied on both classification and regression models. Moreover, it has been used by many researchers for various classification. For classifying gene disease entities SVM takes more time. Hence to increase accuracy and reduce time complexity RVM classifier is trained and used for classifying gene and diseases from DisGeNET dataset.

III. MODELLING OF THE SYSTEM

III.1 DisGeNET Dataset

DisGeNET is a discovery of platform for human diseases and their associated genes, available at http://disgenet.org. It consists of the columns such as GeneID, GeneName,Description,DiseaseID, DiseaseName,Score, NofPmids,NofSnps and Sources. This dataset consists of 4,29,036 different gene names and names of their associated disease with additional information such as Description, Score,etc[12].

In this research work, the subset of DisGeNET that is, the genes associated with two diseases namely Adenocarcinoma and Breast Cancer alone taken and the additional information like Description, Sources, NofPmids, NofSnps were removed using dimensionality reduction.

Totally 8410 different genes associated with Adenocarcinoma and Breast Cancer were taken as subset. With this subset of DisGeNET v4.0 data RVM classifier is modeled and the diseases ‘Adenocarcinoma’ and ‘Breast Cancer’ were considered as class labels.

However, the inputs to the RVM were different gene names that are associated with these two diseases. Irrelevant rows and columns for this analysis were filtered before modeling. For example, Disease ID, Description and Source were removed.

Training RVM:

The algorithm used for training RVM is given in Fig.1.

```
Algorithm:
    1. Start
    2. Give training set as input ${\{x_i , t_i\}}_{i=1}^{n}$, where
       $x_i \in R^n, t_i \in \{0,1\}$ and n is the number of entity samples.
    3. Calculate $\sigma^2$ using the formula,
            i. $\sigma(a) = \frac{1}{1+\exp(-a)}$
    4. Calculate weight based posterior probability using the following formula
            a. $p(t = 1|\hat{x}) = y(\hat{x}, w)$ for class 1.
            b. $p(t = 0|\hat{x}) = y(\hat{x}, w)$ for class 2.
    5. If $w_i$ is not zero
            a. Take the corresponding point$(x_i , t_i)$ as a relevance vector
    6. Repeat step 5 for all i values till it gets the value n.
    7. Calculate $y(\hat{x}, w)$ using the formula
            $y(\hat{x}, w) = \sigma\left( \sum_{i=1}^{n} \omega_i k(x_i , \hat{x}) \right)$
            $= \sigma(w^T K)$
    8. End
```

Fig.1 Algorithm to train Relevance Vector Machine

Several parameters were set to build this RVM model. Sigmoid function was selected as Kernel function. The list of parameters and the setting values are given in Table 1.Various performance measures were tested and the results were noted.

Table 1. Parameters for RVM model

Parameters	Values
Rvm type	Classification
Kernel type	Sigmoid
Kernel ‘a’	1.0
Kernel ‘b’	0.03
Max iteration	1000
Min delta log alpha	0.001
Alpha max	1.0E12

III.1 Dataset

DisGeNET v5.0[13] dataset is used in order to evaluate the entity classification. Entities considered in this work are Gene names and Disease names. As the research work focused only two kinds of diseases namely Breast Cancer and Adenocarcinoma, only a subset of records(8410 records) related to the above mentioned diseases were taken from the DisGeNET dataset among 4,29,036 records. This phase focus on categorizing the gene with the disease associated with it that is ‘Breast Cancer’ disease or ‘Adenocarcinoma’.

IV. EVALUATION MEASURES

Various evaluation measures are used to evaluate the performance of the classifier. For instance, Accuracy, Kappa, Classification error, Weighted mean precision and Weighted mean recall are considered to evaluate the performance of RVM classifier.

IV.1 Accuracy

The accuracy value is calculated by taking the percentage of correct predictions over the total number of examples. The term correct prediction means the examples where the value of the prediction is equal to the value of label attributes.

$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$ ------ (4.1)

IV.2 Kappa

Cohen's Kappa is a very useful statistics metric. In some cases, the measures such as Precision/ Recall and accuracy do not provide the complete performance measure of the classifier in multi class classification problem. This Kappa statistic is a very good measure to handle both multi-class and imbalanced class problems efficiently.

Cohen's Kappa is defined as,

$K = \frac{Po-Pe}{1-Pe}$ -------- (4.2)

Where Po is Observed value and Pe is the expected value.

Kappa value is always less than or equal to 1. Table 2 summarizes the categorized values provided by Landis and Koch(1977)[14].

Table 2. Categorized Kappa values

Values	Category
<0	No agreement
0 – 0.20	Slight
.21 – 0.40	Fair
.41 – 0.60	Moderate
.61 – 0.80	Substantial
.81 – 1	Perfect agreement

IV.3 Classification Error

Classifying an entity belonging to one class when it belongs to another class is considered as Classification Error. The error rate can be calculated as the percentage of misclassified examples out of total number of examples in the dataset.

$Error\ rate = \frac{FP+FN}{TP+FP+FN+TN}$ --------(4.3)

IV.4 Weighted mean precision & Weighted mean recall

The Weighted mean recall is calculated by taking the average recall values of the classes. Similarly, the Weighted mean precision is calculated by taking the average of precision values of every class.

V. EXPERIMENT RESULTS AND DISCUSSION

The proposed method was implemented by training Relevance Vector Machine. The Precision, Recall values were obtained and summarized in Table 3.

Table 3. Precision and Recall values obtained by RVM

Dataset	Class (Disease)	Class Precision	Class Recall
DisGeNET	Adenocarcinoma	94.46	92.94
	Breast Cancer	94.79	95.93
PubMed abstracts	Adenocarcinoma	99.01	93.48
	Breast Cancer	92.91	98.92

The classifier performance is evaluated with various measures. The values obtained for various measures(Accuracy,Classification error, Kappa, ROC) were observed and analyzed. The RVM classifier is , tested on two different datasets, where one dataset consists of entities extracted from the corpus constructed using the abstracts retrieved from PubMed abstracts and another one is DisGeNET dataset. The results obtained by RVM classifier are shown in Table 4. & Table 5.

Table 4. Confusion matrix obtained by RVM for PubMed abstracts collection.

	True Adenocarcinoma	True Breast Cancer
Predicted Adenicarcinoma	602	6
Predicted Breastcancer	42	550
Class Recall	93.48	99.92

The class precision for the above confusion matrix is 99.01 to predict Adenocarcinoma and 92.91 Breast Cancer disease based on their associated gene names. Accuracy obtained is 96.00

Table 5 Confusion matrix obtained by RVM for DisGeNET dataset.

	True Adenocarcinoma	True Breast Cancer
Predicted Adenicarcinoma	3343	196
Predicted Breastcancer	254	4617
Class Recall	92.94	95.93

The class precision for the above confusion matrix is 94.65 to predict Adenocarcinoma and 94.79 to predict Breast Cancer based on their associated gene names. Accuracy obtained is 94.65.

The results of various measures obtained by testing the RVM classifier on both datasets are listed in Table6.

Table 6. Results of RVM performance

Dataset	**Accuracy**	**Kappa**	**Weighted mean recall**	**Weighted mean precision**
PubMed abstracts	96	0.92	95.99	96.19
DisGeNET	94.65	0.890	94.43	94.62

To train and test the RVM system Ten-fold cross validation was used. The subset of DisGeNET v5.0 dataset is taken as input and randomly divided into 10 approximately equal subsets. In the next step, classifier was trained on the remaining 9 subsets and tested on the 10th subset. Likewise, the sequence was iterated 10 times, with each subset serving as test set one time. The test results were then used to plot the ROC curve. The ROC curve is presented in Fig. 3. The red line represents the graph for the true positive value to classify the disease and the blue line shows the various thresholds used to test the true positive value.

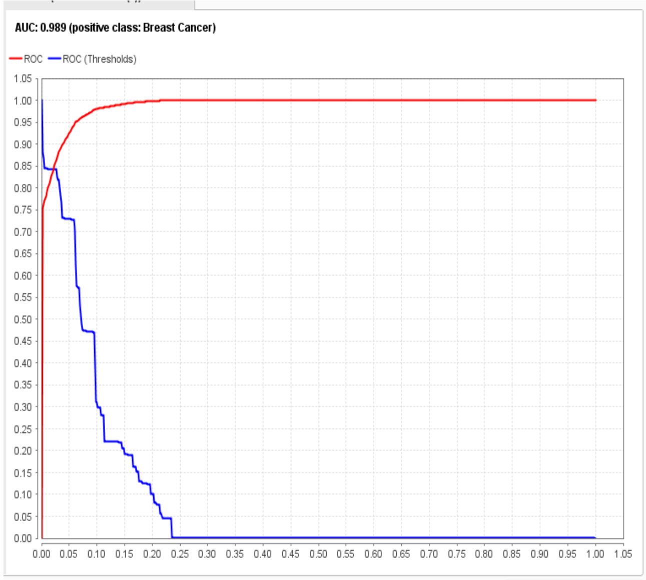

Fig. 1. ROC curve for Positive class

To evaluate RVM the same datasets were tested with HMM and SVM classifiers. The values obtained are presented in Table 7 and Table 8.

Table 7. Accuracy and time obtained by HMM , SVM and RVM Classifiers on PubMed abstracts

Method	Accuracy	Time in seconds
HMM	88.17	5
SVM	93.56	6
RVM	96.00	3

Table 8. Accuracy and time obtained by HMM , SVM and RVM Classifiers on DisGeNET dataset.

Method	Accuracy	Time in seconds
HMM	87.17	7
SVM	92.56	8
RVM	94.65	5

The measures obtained by HMM ,SVM and RVM methods are shown as graphical representation in Fig 2. & Fig 3.

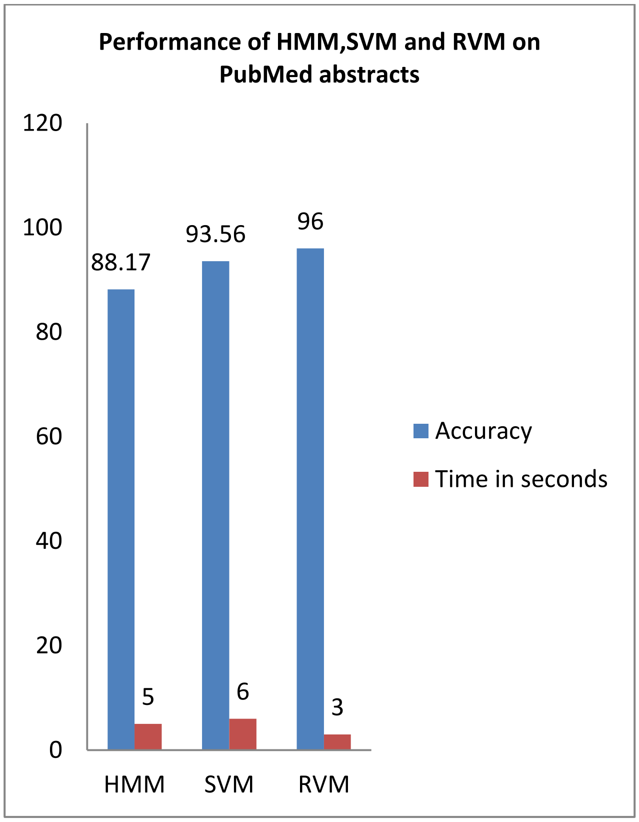

Fig.2. Performance of HMM,SVM and RVM on PubMed abstracts

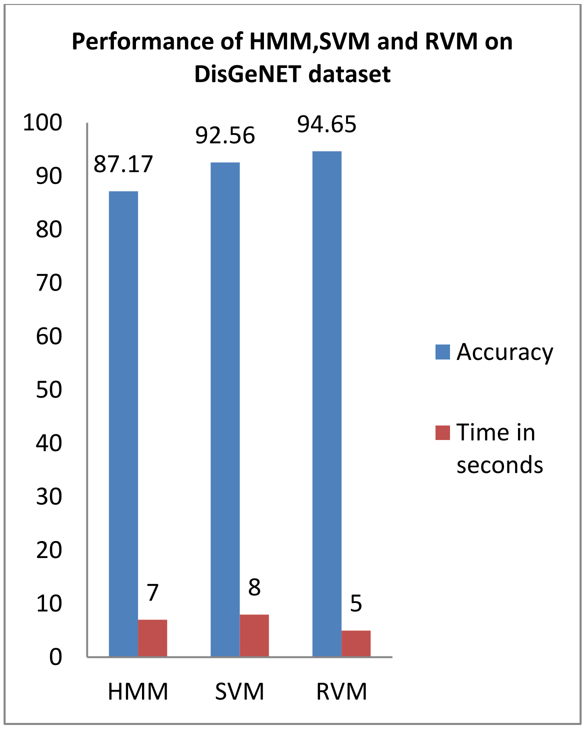

Fig.3. Performance of HMM,SVM and RVM on DisGeNET dataset

From the results summarized in Table 7 and Table 8, it is clearly noted that RVM classifier is performed better than HMM and SVM classifiers in terms of high accuracy and less execution time. Therefore, it is concluded that the RVM classifier constitute an alternative to the classification of entities with high level of accuracy and less time complexity. From the results obtained , it is clearly noted that the RVM classifier was able to classify the entities with higher accuracy of 94.65% on DisGeNET dataset and 96% on PubMed abstracts and with less time duration 5 and 3 seconds than SVM and HMM respectively.

VI. CONCLUSION

Named Entity Recognition on Biomedical domain is a tedious and most challenging task. Two phases in NER are i) Extracting Named entities from collection of text and ii) Classifying the entities in the class associated with them. This work focused on the second phase of Named Entity recognition that classifies entities. Experimental results on two different dataset demonstrated the effectiveness of RVM classifier. The performance of the classifier is tested with various performance measures and from the results obtained it is concluded that this RVM method performs well with high accuracy and less time complexity.

REFERENCES

[1] Michael Fleischman and Eduard Hovy,”Fine Grained Classification of Named Entities”, USC Information Science Institute, U.S.A.

[2] S.Vijaya, Dr.R.Radha,”Named Entity Recognition and Gene Disease Relationship Extraction Using Relevance Vector Machine(RVM) Classifier”, International Journal for Research in Applied Science & Engineering Technology (IJRASET), ISSN: 2321-9653; IC Value:45.98, SJ Impact Factor:6.887, Volume 5, Issue XII December 2017].

[3] Nor Liyana Mohd Shuib et al. (2014), “Data Mining Approach: Relevance Vector Machine for the Classification of Learning Style based on Learning Objects”, UKSim-AMSS 16th International Conference on Computing Modelling and Simulation.

[4] G.D.Zhou,”Recognizing names in biomedical texts using mutual information independence model and SVM plus sigmoid”, Int. J. Med. Inform, Vol 75,no. 6, pp.456-67, Jun 2006.

[5] S.Jonnalagadda, T.Cohen et al., “Using empirically constructed lexical resources for named entity recognition “, Biomed. Inform. Insights,vol 6, no.Suppl.1, pp.17-27,Jan 2013.

[6] S.Zhang and N.Elhadad,”Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts”, Journal of Biomedical Informatics 46 (2013) 1088-1098.

[7] Christopher Bowd et al.,”Relevance Vector Machine and Support Vector Machine Classifier Analysis of Scanning Laser Polarimetry Retinal Nerve Fiber Layer Measurements”, Machine Classifier Analysis of SLP Measurements, Investigative Ophthalmology &Visual Science, April 2005, Vol. 46, No.4.

[8] Stefan Andelic et al., “Text classification Based on Named Entities”, 7th International Conference on Information Society and Technology ICIST (2017)

[9] Rebholz-Schuhman et al.,”The CALBC Silver Standard Corpus for Biomedical Named Entities: A study in Harmonizing the

contributions from Four Independent Named Entity Taggers", Proc. LREC 2010.

[10] Michael E.Tipping,"Sparse Bayesian Learning and the Relevance Vector Machine", Journal of Machine Learning Research 1 (2001) 211-244.

[11] Rong Xu et al, "Combining Text Classification and Hidden Markov Modeling Techniques for Structuring Randomized Clinical Trial Abstracts", AMIA 2006, Symposium Proceedings Page-824.

[12] Janet Piñero et al., "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants." Nucl. Acids Res. (2016) doi:10.1093/nar/gkw943

[13] Janet Piñero et al. , " DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes", Database (2015) doi:10.1093/database/bav028

[14] Landis,J.R, Koch.G.G(1977), "The measurement of observer agreement for categorical data", Biometrics 33(1):159-174.

Authors Profile

Mrs.S.Vijaya completed M.Sc (Computer Science) from Annamalai University in 2005 , M.Phil (Computer Sciecne) from Bharathidasan University in 2007. She is currently pursuing Ph.D in S.D.N.B.Vaishnav College for Women, Chromepet,Chennai. She has published 6 national/ international journals.

Dr.R.Radha is working as Associate Professor, Department of Computer Science, SDNB vaishnav college for women, chennai. She has 26 years of teaching experience and has published 35+ national and international journals.