J Inf Process Syst, Vol.15, No.6, pp.1326~1334, December 2019
https://doi.org/10.3745/JIPS.04.0146

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Knowledge Base Associated with Autism Construction Using CRFs Learning

Ronggen Yang* and Lejun Gong**

Abstract

Knowledge base means a library stored in computer system providing useful information or appropriate solutions to specific area. Knowledge base associated with autism is the complex multidimensional information set related to the disease autism for its pathogenic factor and therapy. This paper focuses on the knowledge of biological molecular information extracted from massive biomedical texts with the aid of widespread used machine learning methods. Six classes of biological molecular information (such as protein, DNA, RNA, cell line, cell component, and cell type) are concerned and the probability statistics method, conditional random fields (CRFs), is utilized to discover these knowledges in this work. The knowledge base can help biologists to etiological analysis and pharmacists to drug development, which can at least answer four questions in question-answering (QA) system, i.e., which proteins are most related to the disease autism, which DNAs play important role to the development of autism, which cell types have the correlation to autism and which cell components participate the process to autism. The work can be visited by the address http://134.175.110.97/bioinfo/index.jsp.

Keywords

Autism, Biological Molecular, Conditional Random Fields, Knowledge Base

1. Introduction

Prevalent explanation for knowledge base is a self-serve customer service library stored in computer system, including information or appropriate solutions about specific area. Knowledge of specific domain is structured representation explored or mined from unstructured data. Knowledge also can be represented by rules, and rule-based knowledge base can provide new knowledge to the users of a decision support system [1]. A deep learning approach for knowledge discovery to power system security area is reported in [2]. These knowledges can be structured and stored in knowledge base. Knowledge discovery with its base construction in many other areas [3] showed that the data exploration was more and more refined and the effectiveness of utilizing information was closer to the reality needs of specific domain.

Autism also known as autism spectrum disorder (ASD), refers to a broad range of conditions characterized by challenges with social skills, repetitive behaviors, speech and nonverbal communication. Although many research experiments [4] from different perspectives, we still know little information about it, for example, we may want to know which proteins metabolism or genes play decisive role in the

Manuscript received June 26, 2019; accepted September 3, 2019.
Corresponding Author: Ronggen Yang (rg4592@jit.edu.cn)
* School of Intelligent Science and Control Engineering, Jinling Institute of Technology, Nanjing, China (rg4592@jit.edu.cn)
**School of Computer, Nanjing University of Posts and Telecommunications, Nanjing, China (glj98226@njupt.edu.cn)

pathogenesis process and the relations between them. Fortunately, big data provided in many researches gives us a chance to deep investigate and mining from these massive literatures. We are coming closer to completely understand the disease autism with the help of prevalent machine learning algorithms. This paper focuses on the knowledge of biological molecular information extracted from biomedical texts with the aid of widespread conditional random fields (CRFs) methods. Protein, DNA, RNA, cell line, cell component, and cell type, etc., 6 classes of biological molecular information are concerned. The knowledge base can help biologists to etiological analysis and pharmacists to drug development.

2. Corpus and Preprocessing

The GENIA term annotation was provided by GENIA Project, which was founded by Prof. Jun'ichi Tsujii and ran at the Tsujii Laboratory of University of Tokyo from 1998 to 2012. The corpus is a collection of 1999 biomedical abstracts [5] in Molecular Biology Domain and 38 classes terms were annotated to help machine learn the biological knowledge. We concentrated on the 6 classes of terms such as protein molecule, DNA molecule, RNA molecule, cell line, cell type, cell component.

Original GENIA term annotation corpus is formatted in xml file. This can help us extract the useful parts by the corresponding mark, e.g., the mark <cons> indicates some knowledge including in the flowing annotation, <title> point out the following is the title of the abstract.

The MEDLINE number, title, content, protein molecule, DNA molecule, RNA molecule, cell line, cell type, and cell component, etc., six annotated terms were extracted from the sentences with the regular expressions and structured in the csv file, which can be downloaded from the website http://134.175.110.97/bioinfo/index.jsp. Each row in the file denoted an article with 6 columns corresponding to different information. In order to discover the knowledge from the corpus, the csv file was further organized in samples to make preparation for learning process.

Each row is a sample with the token and the tag, while the tag was the class label of the token. Maybe some term contains multi-tokens, we utilize the traditional representation as B-I-O methods. Naturally, the tags include 13 type of labels for the 6 classes. These samples were trained in the CRF algorithm and the model as the rules was yielded for knowledge discovery from literatures associated with the disease autism.

3. Methods

3.1 Fundamental of Mathematics

Conditional random fields, a kind of structured prediction methods, are essentially a combination of classification and graphical model [6]. Much work in learning with graphical models that explicitly model a joint probability distribution $p(y, x)$ in Eq. (1) over outputs and inputs. However, it is difficult to model joint probability for the dimensionality of x is very large and the features may have complex dependencies under many circumstances. Fortunately, CRFs as a discriminative approach, model the conditional distribution $p(y|x)$ in Eq. (2) directly. Where $Z(x) = \sum_y exp\ (\theta_y + \sum_{j=1}^{K} \theta_{y,j} x_j)$ is an normalized constant, and θ_y is a bias weight that acts like $log\ \ p(y)$ in naïve Bayes.

$$\mathrm{p(y,x)} = \mathrm{p(y)} \prod_{k=1}^{K} p(x_k|y) \tag{1}$$

$$p(y|x) = \frac{1}{Z(x)} exp\,\{\theta_y + \sum_{j=1}^{K} \theta_{y,j} x_j\} \tag{2}$$

Rather than using one weight vector per class, as in Eq. (2), we can use a different notation in which a single set of weights is shared across all the classes. A set of feature functions is defined as $f_{y',j}(y,x) = 1_{\{y'=y\}} x_j$ that is nonzero only for a single class, in practice. Naturally, we can use f_k to index each feature function $f_{y',j}$, and θ_k to index its corresponding weight $\theta_{y',j}$. The Eq. (2) can be rewritten as (3).

$$p(y|x) = \frac{1}{Z(x)} exp\,\{\sum_{j=1}^{K} \theta_k f_k(y,x)\} \tag{3}$$

CRFs combine the ability of graphical models to compactly model multivariate data with the ability of classification methods to perform prediction using large sets of input features.

3.2 Roadmap of the Method

Knowledge base construction associated autism method can be described in Fig. 1. The GENIA corpus, as the inputs, has been transformed into samples in the previous preprocessing. There are 18,749 sentences including 508,645 tokens in the corpus.

As mentioned above, feature functions are important for CRF algorithm to learn model and rules from these samples. In this work, we set two types feature functions i.e. state transition functions $f(y_{i-1}, y_i)$ and state observation functions $f(y_i, x_i)$. Although the long range dependency may give a little improved to performance, we only consider the dependency between two neighbor classes, which called Markov property. These samples are classified into 13 classes, so the number of transition functions is 169. State observation functions embrace current observations and its context. We set the following features including $f(y_i, x_{i-2}, x_i)$, $f(y_i, x_{i-1}, x_i)$, $f(y_i, x_i)$, $f(y_i, x_i, x_{i+1})$, $f(y_i, x_i, x_{i+2})$, $f(y_i, x_{i-2}, x_{i-1}, x_i)$, $f(y_i, x_i, x_{i+1}, x_{i+2})$. Dependency between the tokens in context with distance five is modeled in the work. The total number of the feature functions is 5519618 after the statistic computing.

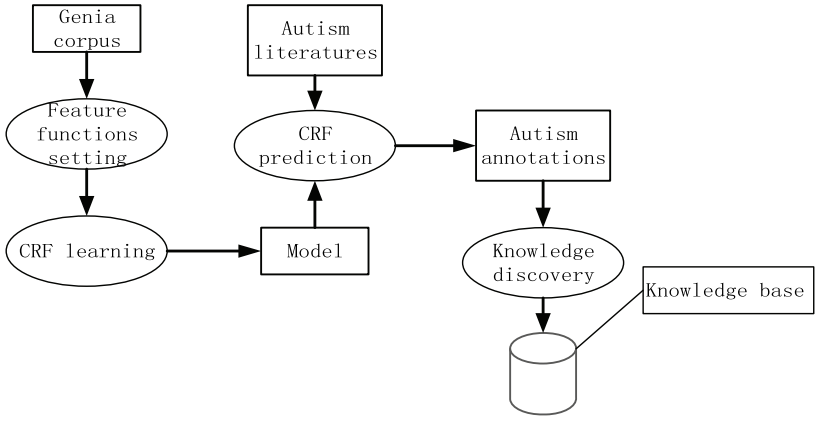

Fig. 1. Knowledge base construction associated with autism flow diagram.

CRF learning is process of utilizing toolkit to compute the weights for these feature functions and find the rules for knowledge discovery. CRF++0.58 toolkit is a simple and open source implementation of CRFs, developed for a variety of natural language processing (NLP) tasks. To evaluate the performance the method, we split the whole corpus into training set and test set with four different proportion of 0.6, 0.7, 0.8, and 0.9, and take CRF++ over each training set to learn model. At the end, we take the CRF++ tool to learn the model over all the samples and perform 453 iterations to final convergence. Fig. 2 shows the tag error and sentence error rate decrease with first 100 iterations by CRF++ tool. Tag error rate descends rapidly and converges to a little value with the aid of the set of feature functions. Although sentence error rate decrease is slowly, it also gets convergence in the final iterations. The final tag error rate is 0.0022 and sentence error rate is 0.03291.

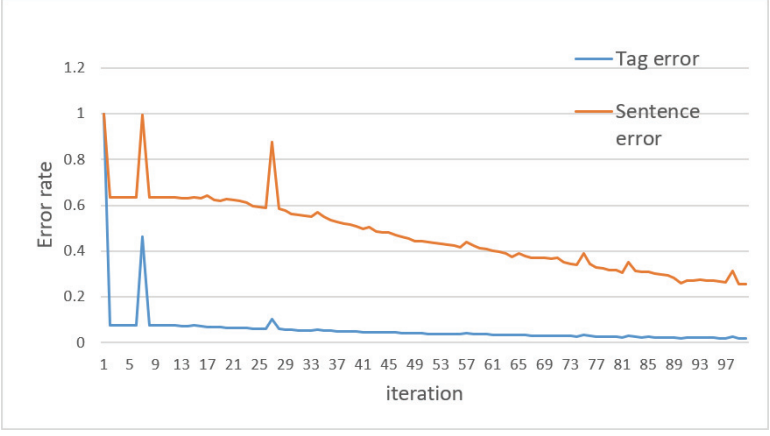

Fig. 2. Tag error and sentence error rate decrease with first 100 iterations by CRF++ tool.

Table 1 is the performance of 12 classes (not including the *O* class) identified by CRF++ tool over four different scale test sets with 0.4, 0.3, 0.2, and 0.1. Results in the table shows that there is no obvious difference between test sets, although their models are trained from the different training sets. These illustrate the stability of the performance. The performance does not depend on the number of the samples. Precision of class B-protein can achieve 0.85 and recall 0.62 in test set of 0.1 for the amount of B-protein is the most among the classes. This demonstrate that the protein identification is more faithful than others.

Results in Table 2 are the performance of completed entity merged from the test sets. Just as expected, the performance is lower than that in Table 1. Maybe, there is a lot space for improvement, but this does not affect we extract rough knowledge from the literatures. At the end, we learn the model from all the GENIA corpus to enlarge the training set extremely. Weights of each feature function are computed and stored in the model to extract knowledge from literature associated with autism.

Literatures associated with autism are downloaded from PubMed website with the key word autism, and the total number of abstracts is 42,997. There are 6,684,747 tokens in the literatures. These literatures with the model learned from the GENIA corpus as the inputs to CRF++ tool and the annotations of the literatures were predicted. Finally, the knowledge discovery programs give the knowledge related to autism from the prediction result. The model learned from GENIA corpus represent the rules of 6 classes. So, we can easily identify the protein, DNA, RNA, cell line, cell type and cell component from original

Table 1. Performance of original results through CRF++ over the GENIA corpus

Test set		B-cell_component	B-cell_line	B-cell_type	B-DNA	B-protein	B-RNA	I-cell_component	I-cell_line	I-cell_type	I-DNA	I-protein	I-RNA	Avg.
0.4	Precision	0.67	0.69	0.57	0.52	0.82	0.80	0.60	0.81	0.53	0.61	0.57	0.80	0.67
	Recall	0.48	0.32	0.34	0.28	0.52	0.58	0.26	0.64	0.31	0.40	0.43	0.60	0.43
	F1-score	0.56	0.44	0.43	0.36	0.64	0.67	0.36	0.72	0.39	0.48	0.49	0.69	0.52
	Support	134	565	788	103	8283	149	58	577	443	131	2459	181	13871
0.3	Precision	0.66	0.74	0.54	0.48	0.84	0.70	0.57	0.82	0.48	0.59	0.62	0.70	0.65
	Recall	0.48	0.34	0.39	0.29	0.54	0.60	0.24	0.65	0.30	0.38	0.43	0.67	0.44
	F1-score	0.56	0.46	0.45	0.36	0.66	0.65	0.34	0.73	0.37	0.46	0.51	0.68	0.52
	Support	108	417	621	69	6349	93	49	420	352	98	1787	103	10466
0.2	Precision	0.70	0.69	0.60	0.52	0.84	0.73	0.75	0.79	0.50	0.55	0.60	0.69	0.66
	Recall	0.51	0.34	0.39	0.40	0.59	0.71	0.26	0.61	0.32	0.48	0.46	0.83	0.49
	F1-score	0.59	0.46	0.47	0.45	0.70	0.72	0.39	0.69	0.39	0.51	0.52	0.75	0.55
	Support	72	254	479	40	4078	52	23	265	235	56	1081	52	6687
0.1	Precision	0.81	0.68	0.62	0.89	0.85	0.72	0.67	0.82	0.59	0.79	0.65	0.57	0.72
	Recall	0.59	0.33	0.43	0.42	0.62	0.72	0.20	0.61	0.36	0.41	0.50	0.74	0.49
	F1-score	0.69	0.45	0.51	0.57	0.71	0.72	0.31	0.70	0.45	0.54	0.56	0.64	0.57
	Support	37	117	237	19	2127	25	10	120	132	27	559	27	3437

text and count its frequency naturally. Table 3 is top ten of 6 classes term and its frequency, and all the terms also can be found in the project website. Four classes terms such as protein, DNA, cell type and cell component showed in Table 3 are very representativeness from the frequency perspective. In the following discussions, we will elaborate on the validation of these four classes of terms.

Table 2. Performance of completed entity merged from GENIA corpus

Test set		Cell_component	Cell_line	Cell_type	DNA	Protein	RNA	Avg.
0.4	Precision	0.67	0.69	0.57	0.52	0.82	0.80	0.67
	Recall	0.48	0.32	0.34	0.28	0.52	0.58	0.43
	F1-score	0.56	0.44	0.43	0.36	0.64	0.67	0.52
	Support	134	565	788	103	8283	149	13871
0.3	Precision	0.66	0.74	0.54	0.48	0.84	0.70	0.65
	Recall	0.48	0.34	0.39	0.29	0.54	0.60	0.44
	F1-score	0.56	0.46	0.45	0.36	0.66	0.65	0.52
	Support	108	417	621	69	6349	93	10466
0.2	Precision	0.70	0.69	0.60	0.52	0.84	0.73	0.66
	Recall	0.51	0.34	0.39	0.40	0.59	0.71	0.49
	F1-score	0.59	0.46	0.47	0.45	0.70	0.72	0.55
	Support	72	254	479	40	4078	52	6687
0.1	Precision	0.81	0.68	0.62	0.89	0.85	0.72	0.72
	Recall	0.59	0.33	0.43	0.42	0.62	0.72	0.49
	F1-score	0.69	0.45	0.51	0.57	0.71	0.72	0.57
	Support	37	117	237	19	2127	25	3437

4. Discussions

We concentrate on the 4 classes terms for its frequency is larger 30, such as protein, DNA, cell type and cell component. Among those proteins, interleukin 6 (IL-6) has the highest frequency 232. This strongly suggest that IL-6 related to the disease autism is high probability. We can find the evidence from the literature [7] directly. In this article, IL-6 is reported that it is increased in the cerebellum of autistic brain and alters neural cell adhesion, migration and synaptic formation. Full name of IL-6 is interleukin-6, which plays the crucial role in the development of autism. Recent evidence shows that localized inflammation of the central nervous system (CNS) may lead to autism and IL-6 just contribute to the process. Other proteins in Table 3 also can be proved height correlation to autism.

Among the DNAs in Table 3, we focus on the top frequency DNA, X chromosome. In fact, chromosome X is same as chromosome X. So, its frequency should be 103. [7] showed that X-linked (XL) inheritance or maternal skewed X-chromosome inactivation (XCI) is presenting with autism, using a home-made X-chromosome-specific microarray covering the whole human X-chromosome at high resolution. Evidences [8] also indicate the presence of X-linked susceptibility genes in human with autism and conclude *TBL1X* gene in X-chromosome may play a role in autism risk. These prove high correlation between autism and X-chromosome.

Microglia, a kind of cell type, is also proved important to autism in [9,10]. Microglia is critical to the development of normal neural networks, and abnormal microglia often present in autism. Maternal

Table 3. Top 10 of six classes term and its frequency

Rank	Protein	DNA	RNA	Cell line	Cell type	Cell component
1	IL-6 (232)	X chromosome (85)	FMR1 mRNA (20)	Wild-type (11)	Microglia (197)	Nucleus (133)
2	CD38 (120)	Chromosome 15 (34)	BDNF mRNA (8)	Mouse line (11)	EC (80)	Plasma membrane (32)
3	ERP (112)	Chromosome X (18)	FMRP mRNA (3)	LCL (7)	DC (35)	Endoplasmic reticulum (23)
4	ERK (61)	Chromosome 7 (15)	TH mRNA (3)	Lymphoblastoid cell line (3)	Pluripotent stem cell (32)	ER (22)
5	PI3K (57)	Chromosome 10 (15)	Local mRNA (3)	Neuronal cell line (2)	T cell (22)	Cytoplasm (17)
6	SM (54)	Chromosome 16 (15)	Auts2 mRNA (3)	Straight line (2)	Platelet (20)	Cell surface (15)
7	IL-10 (51)	Chromosome 14 (13)	GPR155 mRNA (2)	Human neuronal cell line (2)	Mast cell (18)	Nuclei (14)
8	MMP-9 (42)	Chromosome 5 (10)	Met transcript (2)	HeLa (2)	Neural stem cell (16)	Cell membrane (12)
9	IL-4 (38)	Chromosome 6 (9)	CD38 mRNA (2)	Conditional knockout mouse line (2)	Stem cell (15)	Cytosol (10)
10	IL-2 (38)	Chromosome 1 (9)	1α-hydroxylase mRNA (2)	Mutant line (2)	PBMC (13)	Membrane (9)

The values in parentheses represent frequency.

immune activation and microglial dysfunction in the developing brain have been gaining mounting evidence and leading to potential treatment options.

The component nucleus including caudate nucleus, reticular thalamic nucleus, bed nucleus supraoptic nucleus and paraventricular nucleus, etc., is a cluster of cell bodies of neurons in the central nervous system. Autism is a complex disorder of the central nervous system and the condition has a wide range of severity along its spectrum. In addition, nucleus is annotated as cell component in the GENIA corpus. Naturally, nucleus presented in the literatures can be identified precisely.

As mentioned above, we verify the validation of the knowledge extracted from the original literatures associated with autism and construct the knowledge base, which can at least answer the four questions in QA system, i.e., which proteins are most related to the disease autism, which DNAs play important role to the development of autism, which cell types have the correlation to autism and which cell components participate the process to autism.

5. Conclusions

This work attempt to construct knowledge base associated with disease autism using CRF learning, the widespread probabilistic statistic method. Firstly, we extract protein, DNA, RNA, cell line, cell type, cell component, 6 classes of molecular information from GENIA corpus and format into samples to feed to CRF++ tool. Secondly, we utilize the model learned from the GENIA corpus to find the 6 classes of molecular information from literatures related to autism. Thirdly, knowledge discovery program can seek what is the most high correlated to development of autism and its therapy. If we construct a QA system, we at least can answer the four questions, which proteins are related to the disease autism, which DNAs play important role to the development of autism, which cell types have the correlation to autism and which cell components participate the process to autism.

Acknowledgement

This paper is supported by the project (No.16KJD52003) of Jiangsu Province education department.

References

[1] A. Nowak-Brzezinska and A. Wakulicz-Deja, "Exploration of rule-based knowledge bases: a knowledge engineer's support," *Information Sciences*, vol. 485, pp. 301-318, 2019.

[2] T. E. Huang, Q. Guo, H. Sun, C. W. Tan, and T. Hu, "A deep learning approach for power system knowledge discovery based on multitask learning," *IET Generation, Transmission & Distribution*, vol. 13, no. 5, pp. 733-740, 2018.

[3] Y. Sato, K. Izui, T. Yamada, and S. Nishiwaki, "Data mining based on clustering and association rule analysis for knowledge discovery in multiobjective topology optimization," *Expert Systems with Applications*, vol. 119, pp. 247-261, 2019.

[4] M. Aarabi, E. Kessler, S. Madan-Khetarpal, U. Surti, D. Bellissimo, A. Rajkovic, and S. A. Yatsenko, "Autism spectrum disorder in females with ARHGEF9 alterations and a random pattern of X chromosome inactivation," *European Journal Of Medical Genetics*, vol. 62, no. 4, pp. 239-242, 2019.

[5] D. Q. Nguyen and K. Verspoor, "From POS tagging to dependency parsing for biomedical event extraction," *BMC Bioinformatics*, vol. 20, article no. 72, 2019.

[6] F. I. Alam, J. Zhou, A. W. C. Liew, X. Jia, J. Chanussot, and Y. Gao, "Conditional random field and deep feature learning for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 3, pp. 1612-1628, 2018.

[7] S. F. Ahmad, M. A. Ansari, A. Nadeem, S. A. Bakheet, L. Y. AL-Ayadhi, and S. M. Attia, "Elevated IL-16 expression is associated with development of immune dysfunction in children with autism," *Psychopharmacology*, vol. 236, no. 2, pp. 831-838, 2019.

[8] M. Aarabi, E. Kessler, S. Madan-Khetarpal, U. Surti, D. Bellissimo, A. Rajkovic, and S. A. Yatsenko, "Autism spectrum disorder in females with ARHGEF9 alterations and a random pattern of X chromosome inactivation," *European Journal of Medical Genetics*, vol. 62, no. 4, pp. 239-242, 2019.

[9] C. A. Edmonson, M. N. Ziats, and O. M. Rennert, "A non-inflammatory role for microglia in autism spectrum disorders," *Frontiers in Neurology*, vol. 7, article no. 9, 2016.

[10] J. W. Kim, J. Y. Hong, and S. M. Bae, "Microglia and autism Spectrum disorder: overview of current evidence and novel immunomodulatory treatment options," *Clinical Psychopharmacology and Neuroscience*, vol. 16, no. 3, pp. 246, 2018.

Ronggen Yang https://orcid.org/0000-0003-0969-4216

He received Ph.D. degree in School of Computer Science from Nanjing University of Science and Technology in 2011. He is currently an associated professor with School of Intelligence Science and Control Engineering, Jinling Institute of Technology. His current research interests include text mining and biomedical information processing.

Lejun Gong https://orcid.org/0000-0001-7062-9777

She received M.S. degree in School of Software Engineering from Yunnan University in 2005. She is currently a lecturer in School of Computer, Nanjing University of Posts and Telecommunications. Her current research interests include bioinformatics and biomedical text mining.