Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2020) 28: 1179 – 1196

doi:10.3906/elk-1906-46

Chemical disease relation extraction task using genetic algorithm with two novel voting methods for classifier subset selection

Stanley Chika ONYE[1,2], **Nazife DİMİLİLER**[2,*], **Arif AKKELEŞ**[3]

[1]Department of Applied Mathematics and Computer Science, Faculty of Arts and Sciences, Eastern Mediterranean University, Famagusta, North Cyprus

[2]Department of Information Technology, School of Computing and Technology, Eastern Mediterranean University, Famagusta, North Cyprus

[3]Department of Mathematics, Faculty of Arts and Sciences, Eastern Mediterranean University, Famagusta, North Cyprus

Received: 08.06.2019 • **Accepted/Published Online:** 12.11.2019 • **Final Version:** 28.03.2020

Abstract: Biomedical relation extraction is an important preliminary step for knowledge discovery in the biomedical domain. This paper proposes a multiple classifier system (MCS) for the extraction of chemical-induced disease relations. A genetic algorithm (GA) is employed to select classifier ensembles from a pool of base classifiers. Moreover, the voting method used for combining the members of each of the ensembles is also selected during evolution in the GA framework. The performances of the MCSs are determined by the algorithms used for selecting the classifiers, the diversity among the selected classifiers, and the voting method used in the classifier combination. The base classifiers are represented in the form of chromosomes, where each chromosome contains all information on the ensemble it represents: the subset of classifiers voting and the voting method. The chromosomes are evolved using a variety of genetic selection, mating, and mutation techniques in order to find an optimal solution. The aim of the proposed system is to select the subset of classifiers with diverse abilities while maximizing the strengths of the best classifiers in the classifier ensemble for a given voting method. Two main contributions of this work are the evolution of the voting bit as part of the GA and the novel approach of using two different decision-making under uncertainty techniques as voting methods. Furthermore, two different selection algorithms and crossover operators are employed as ways of increasing variations during evolution. We validated our proposed method on nine different experimental settings and they produced good results comparable to the state-of-the-art systems, thereby justifying our approach.

Key words: Multiple classifier systems, genetic algorithm, chemical disease relation, relation extraction, text mining, classifier ensemble

1. Introduction

There has been an increase in the scientific efforts dedicated to improving knowledge discovery in biomedical texts in the last decade. Biomedical relation extraction, which aims at extracting and identifying the relationships or associations among the predefined biomedical entities, is one of the most important prerequisites of knowledge discovery in this domain. Some of the specific biomedical relations that are currently targeted in text mining are chemical-induced disease (CID) relation extraction and protein-protein and gene-gene interactions. Most of the research in this domain has been treated as a classification task where single classifier-based systems [1–4] as well as multiple classifier systems (MCSs) have been proposed [5–7]. In the classification tasks of

*Correspondence: nazife.dimililer@emu.edu.tr

multiple nontrivial pattern recognition problems, the MCS reportedly provides better performances [8–10]. One of the core modules of an MCS is classifier subset selection (CSS). In the CSS module, a subset of classifiers is selected from the base classifiers to form an ensemble such that the performance of this new ensemble is better than that of the complete set of base classifiers and the best individual classifier [7, 11–13]. Previous research on improving the performance of an MCS indicated, among other factors, diversity or complementarity of the base classifiers as well as their individual performances [7, 10–12, 14–17]. Complementarity or diversity among the base classifiers in an MCS can be improved through the variations of the parameters of the classifiers, the use of different subsets of the training dataset, and the use of different feature subsets [7, 11, 18]. Moreover, the selection and combination of features for training influences the performance of the individual base classifiers [12]. Diversity within the set of base classifiers may be improved through the use of different classifiers and the tuning of classifiers to different parameter settings in the base classifier, and the use of different feature subsets for training [19].

The advancement of relation extraction (RE) tasks in the biomedical domain has been hindered due to the lack of a comprehensive dataset to serve as a benchmark for the comparison of different RE systems and methods [20]. However, the Critical Assessment of Information Extraction in Biology presented a challenge, the BioCreative V challenge, which is a major formal evaluation event for biomedical natural language processing research in order to improve the present state-of-the-art systems by providing a benchmark dataset [21]. The aim of this challenge is for participants to develop an automated system for the extraction of possible CID relations between entity pairs in the corpus (BioCreative V corpus) provided [20]. The overall aim of the CID relation extraction task is to assist future relation extraction tasks by ensuring that the mined entities and their relations are given unique concept identifiers saved in the database for efficient curation and also limiting costs from the need of infrastructure and text-mining tools [20]. The CID relation extraction task discussed in our paper is a subtask of this BioCreative V challenge.

Relation extraction tasks in the biomedical domain are generally performed as a binary classification task to predict the existence of a relation between a pair of chemical and diseases [4, 22–26] and were previously tackled on a sentence or co-occurrence level [26–29]. On the co-occurrence level, the possibly related entities exist in the same sentences. However, based on the loss of information on the co-occurrence level, RE tasks have now evolved to be widely handled on a document or non-co-occurrence level [4, 22, 24, 25]. According to [30], the non-co-occurrence level accounts for one-third of the total CID relations present in the BioCreative V corpus. Interestingly, some systems have been developed with a great degree of success to handle the CID relation extraction task on both of the levels before then merging the results from the two levels to get the final CID relation extraction result [22, 24, 31–34]. The merging process can be performed simply on the entity level [32] or with a more advanced combination approach by using a voting method [24].

The ML method is the most frequently implemented method used in RE tasks [22]. In order to aid the performance of ML-based relation extraction systems, feature extraction has become an integral part of them [2, 4, 23–25, 35–38]. With the use of some natural language processing (NLP) toolkits, different types of features (such as shortest paths, path-of-speech tags, and path-of-speech paths) can be extracted to help provide vital information needed by the ML-based systems for efficient classification tasks [39]. The ML-based systems can also be referred to as feature-based systems. The popularly employed ML method for relation extraction tasks is the support vector machine (SVM) method [36, 40] and it has been widely employed in the CID relation extraction task [24, 25, 36, 38]. Currently, the knowledge-based [4, 23, 25] and rule-based [27]

systems are being employed as an improvement for biomedical relation extraction tasks. Although successful, the drawbacks of the rule-based system are its need for a defined experimental set of rules for a target task and the huge computational time it requires for execution. Other methods such as the convolutional neural network (CNN) [20, 22, 41] and recurrent neural network (RNN) [28] models that have the ability of learning feature representations are also being implemented.

In this paper, we propose a novel approach of using the MCS for CID relation extraction. Our MSC framework employs GA as the optimization technique. GA, which is an evolutionary algorithm, provides a variety of options to deal with the complexity between the search algorithm used and the solution found [12]. In our work we improved the diversity among the base classifiers by using different feature subsets for training. We originally increased the number of the base classifiers by tuning the initial base classifiers to multiple parameter settings. Additionally, we added some variations during evolution by using two different randomly selected selection techniques and two types of crossovers.

To different degrees of success, the ML-based, knowledge-based, rule-based, CNN, and RNN systems have been employed for CID relation extraction tasks. However, to our best knowledge, this is the first time that MCS is applied in the CID relation extraction task. Our proposed method in this work employs a multidimensional classifier selection approach through GA. The major contributions of our work involve the implementation of a genetic algorithm framework, which consists of three novel features: (i) each chromosome has an extra bit attached of 1 or 0, called the voting bit, to determine the voting method used for the combination of the classifiers in an ensemble; (ii) as the voting method for classifier combinations, one of two decision-making under uncertainty techniques is used randomly; and (iii) two different selection algorithms and two types of crossovers are used randomly during the evolution process.

2. Materials and methods

2.1. Dataset

The dataset used in this paper is the gold annotated corpus from BioCreative V CDR Task Corpus [42], which is made available by the organizers.[1] The corpus was manually annotated with chemical and disease mentions as well as the chemical-disease relations by domain experts at document level. It contains a total of 1500 PubMed articles that have been divided into three sets, namely training, development, and test sets, each containing 500 abstracts. The number of chemical and disease mentions as well as the number of relations in all three sets is very close, providing a consistent dataset for training and testing a new system. This dataset was used to allow comparison of the proposed system with other systems designed for the biomedical domain.

2.2. Features

Feature selection and extraction demand a special process since the use of the best feature set can improve classification results as well as reduce computational cost compared to when a classifier is fed with redundant features [12, 43]. Previous work in biomedical relation extraction have employed lexical features, information derived from parse trees, statistical information, and bags of words derived from the dataset and/or other resources for training their systems [2, 23, 27, 31, 35–37, 44–46]. In this work, we use contextual, dependency, and statistical feature sets all derived from the dataset. We use a contextual feature set that includes entity mentions and relations or clue words to provide information on the context of each sample. A dependency parse

[1] http://www.biocreative.org/media/store/files/2016/CDR_Data.zip.

tree was generated using the Spacy parser [2] to extract features such as part of speech tags and the height of the tree. The statistical feature set includes quantitative information such as the total number of entity mentions and a number of action/relationship words in each sample. The complete list of features used in this study can be found in our previous work [24]. Table 1 presents a brief description of the feature set combinations employed for training classifiers. The combinations of features are performed in order to compare the effects of different features on the performance of the relation extraction system [24, 31, 37].

Table 1. Feature sets used in the experiments.

Sets	Features
A	All (contextual, dependency, and statistical)
B	Contextual and dependency
C	Contextual and statistical
D	Dependency and statistical

2.3. Classifiers

The base classifiers used in this paper are SVM, two implementations of the Bayes algorithm (naive Bayes and Bayes network), and three implementations of decision trees (J48, random forest (R4), and random tree (R3)). Table 2 gives brief information about the base classifiers and their settings used in our work.

Table 2. List of classifiers.

S. no.	Classifier	Settings
1	Naive Bayes (NB)	1. NB 2. NB Kernel
2	J48	3. J48
3	Bayes net (BN)	4. BN hill climber 5. BN K2 6. BN TAN
4	Random tree (R3)	7. R3
5	Random forest (R4)	8. R4
6	SVM	9. SVM1 10. SVM2

The six different ML classifiers reported in Table 2 are used in either different parameter settings or implementations to produce an initial number of 10 base classifiers. An introduction of the four different feature sets for training the initial base classifiers increased the number of the base classifiers to 40, as each of the classifiers was trained separately on the four feature sets. Therefore, a total of 40 classifiers were used as our base classifiers. The base classifiers are trained using the BioCreative V training dataset and the outputs of the base classifiers from the BioCreative V development dataset are used during the evolution process of our system. The optimum ensemble generated after the evolutionary process is used to evaluate the performance of our system on the BioCreative V test dataset. The performance of each classifier is evaluated by their F-scores,

[2] https://spacy.io/docs/usage/.

which are calculated using the following common metrics: recall (R), precision (P), and F-score (F1). The precision and recall are measured by four metrics: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The final F-score is calculated by the following procedure:

$$Recall(R) = TP/((TP + FN)) \tag{1}$$

$$Precision(P) = TP/((TP + FP)) \tag{2}$$

$$F - score(F1) = 2RP/((P + R)) \tag{3}$$

2.4. Proposed genetic algorithm framework

In the MCS settings for the GA framework, a subset of classifiers is represented by a string of binary values called chromosomes, with '1' or '0' at a location i denoting the presence or absence of classifier C_i. A group of chromosomes is termed a population and the population evolves in every generation through the application of selection, crossover, and mutation processes. These processes are employed to generate possibly better chromosomes in every generation while aiming for an eventual convergence towards an optimal solution. During selection, some of the chromosomes are randomly selected for reproduction. This selection is performed mainly using the fitness level of the chromosomes, such that the fittest ones have a greater probability for reproduction. The chromosomes selected for reproduction are called parents and the products of their reproduction are called offspring. In our approach, the two selection methods employed are the roulette wheel and the tournament selection methods.

- Roulette wheel selection: The parents are selected based on their relative fitness within the population. Therefore, chromosomes with better fitness have more chances to be selected. The probability $prob_i$ of selecting an individual i is given by:

$$\mathbf{prob_i} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{f_i} \tag{4}$$

  where f_i is the fitness of the individual and N is the population size.

- Tournament selection: This method selects the chromosome with the highest fitness from a randomly selected subset of the population. The size of the subset controls the selection pressure as a bigger subset size causes an increase in the selection pressure.

The processes of crossover and mutation are performed after the selection of the parents. These two operations are performed to increase the variety of individuals in the population, thus increasing the chances of avoiding a convergence towards the local optimum [47]. When the termination condition is met or after a predefined number of generations, the fittest chromosome in the population is considered as the optimal MCS solution. Figure 1 describes the flowchart of our GA system and Algorithm 1 provides further description of our GA framework.

Classifier ensembles can be represented as chromosomes where each bit represents the participation of a classifier in the ensemble as reported in [48]. For a population of size N, C_i (where $1 \leq i \leq N$) are the chromosomes representing classifier ensembles where each chromosome contains M bits such that the first M - 1 bits represented by 0 or 1 in location i denotes the absence or presence of a classifier respectively and the last bit

Algorithm 1 Pseudocode of our GA framework.

Given:
    N: Total number of chromosomes in a generation
    M: Number of bits in each chromosome
    Termination_criteria: Desired performance of the classifier ensemble
    $\mathbf{Max_{iteration}}$: Maximum number of iterations
    $\mathbf{r_{crossover}}$: Mutation rate for crossover
    $\mathbf{r_{classifier}}$: Mutation rate for classifier section of the chromosome
    $\mathbf{r_{voting}}$: Mutation rate for voting bit section of the chromosome
Initialization:
    Initial_population <- Randomly generate N chromosomes with M bits
    Use Mth bit to select the voting method for classifier combination: 0: Minimax Regret, 1: Hurwicz Criterion
    Current_generation <- Initial_population
    New_generation <- {} /*empty set*/
Evaluate fitness:
    Calculate the fitness of the offspring chromosomes using voting method indicated by the voting bit
while Iteration $<$ $\text{Max}_{\text{iteration}}$ or Best Fitness of Current_generation $\geq$ Termination_criteria **do**
  while size(New_generation) $<$ N **do**
    Selection:
    Select two parents from the Current_generation using either Roulette Wheel or Tournament Selection
    Crossover:
    Mate the selected parents using one of the randomly chosen 1- or 2-point crossover techniques to create two new offspring with $\text{r}_{\text{crossover}}$
    Add offspring chromosomes to New_generation
    Mutation:
    Mutate first M - 1 bits (i.e. classifier section) of the offspring chromosomes with mutation rate $\text{r}_{\text{classifier}}$
    Mutate Mth bit (i.e. voting bit) of the offspring chromosomes with mutation r_{voting}
    Evaluate fitness:
    Calculate the fitness of the offspring chromosomes using voting method indicated by the voting bit
  end while
  Elitism:
    Current_generation <- Top 5% of the Current_generation + Top 95% of the New_generation
end while
Output: Display the final generation
The fittest chromosome in the output is the classifier ensemble generated by our GA framework

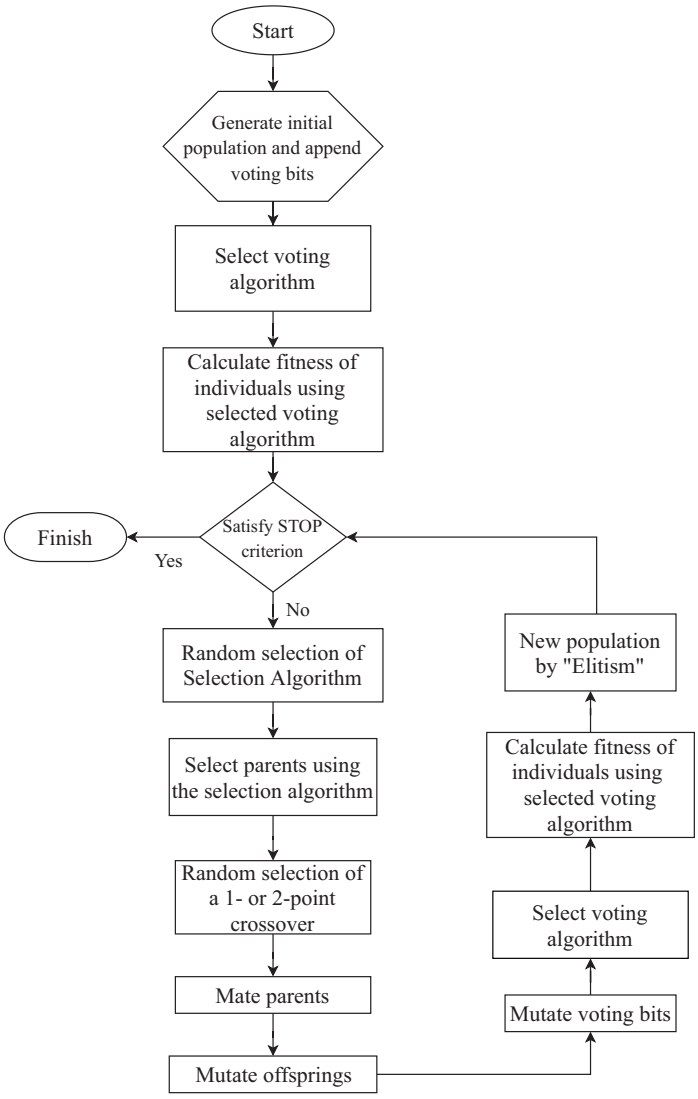

Figure 1. Flowchart of our GA framework.

shows the voting method used in the ensemble as shown in Figure 2. The choice of voting method depends on the voting bit where the bits 0 and 1 represent "Minimax Regret" and "Hurwicz Criterion" methods, respectively. For chromosome C_1 in Figure 2, classifiers 1, 2, … , 38, and 40 are selected in the classifier ensemble and the voting method used is the Minimax Regret algorithm.

The population size in this study is N = 100 chromosomes, each represented by binary strings of length M = 41 including the voting bit. The number of generations for the GA evolution is set at 100. For every generation, the selection of the pair of chromosomes to partake in reproduction takes place through a tournament or roulette wheel selection as described in Figure 1. The pair of chromosomes selected is then passed through a process of crossover and mutation at a rate of 0.9 and 0.01, respectively. For a crossover, the system randomly chooses between a 1- or 2-point crossover based on a split decision. After the crossover, the offspring are considered for mutation. The voting bit in an individual chromosome is subjected to mutation at a rate of 0.2. The fitness, which is the F-score, of each chromosome in the population is calculated by combining the classifier

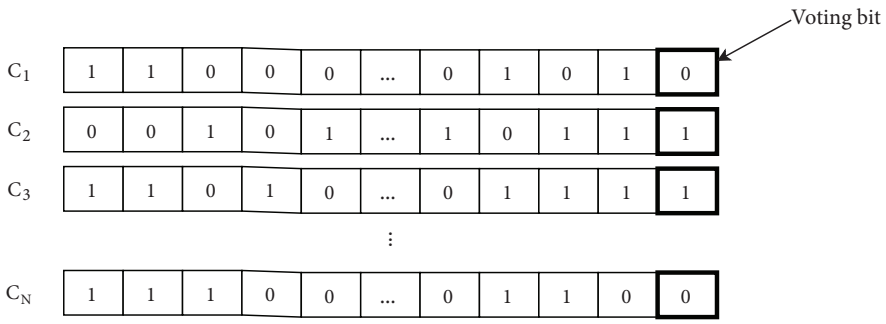

Figure 2. Description of population and voting bit.

ensemble and the chromosomes are ranked according to their fitness. We employed elitism where 5% of the best individuals from the previous generation are propagated to the new generation as long as they are not already present in the new generation. This is to avoid the fittest chromosomes from quickly taking over the entire population.

There exist different voting methods used to compute the performance of a classifier ensemble, including the simple majority, the weighted majority, and percentage majority. However, in this paper, we introduce the novel use of two decision-making under conditions of uncertainty techniques to calculate the fitness of chromosomes. These two voting methods are employed as they overcome the limitations of the conventional methods where decisions are made from a single opinion of either the strength, weight, or percentage of the alternatives considered. They consider multiple opinions from all the alternatives considered before making a decision.

- **Hurwicz criterion (HC):** This is a pessimistic approach suggested by Leonid Hurwicz in 1951. It selects the maximum and minimum payoff from each alternative and tries to find a middle ground between the extremes of the optimistic and pessimistic criteria [49]. It also employs a measure of assigning a given percentage weight to optimism and the balance to pessimism in a bid to avoid an assumption of total optimism or pessimism. This percentage weight is called the coefficient of realism (α) and the balance is called the coefficient of pessimism (1 - α), where $0 \leq \alpha \leq 1$ [49]. In our implementation of this method, due to the pessimism about the actual outcome, we set α to 0.6, which is slightly in favor of the optimistic alternative. The subsets of classifiers are grouped into the two alternatives. The best and worst F-scores from both alternatives are selected and used to calculate an HC weighted average for both alternatives, Yes (A_Y) and No (A_N), as follows:

$$HC(A_Y) = \alpha(A_Y \max) + (1-\alpha)(A_Y \min) \tag{5}$$

$$HC(A_N) = \alpha(A_N \max) + (1-\alpha)(A_N \min) \tag{6}$$

The best HC (A_j) such that HC= max (HC (A_Y), HC (A_N), where $_j$ signifies one of the two alternatives is chosen as the decision of the ensemble. However, in cases of a tie in the decision-making, we apply the reverse of the process to the same value of α, such that:

$$HC(A_j) = (1-\alpha)(A_j \max) + \alpha(A_j \min) \tag{7}$$

- **Minimax regret (MR):** This approach seeks to minimize the maximum regret and it is useful in executing risk-neutral decision-making [49]. The subsets of classifiers are grouped into two alternatives of "Yes" and "No" and the best and worst F-scores from both alternatives are selected as $A_{j\ max}$ and $A_{j\ min}$, where $_j$ signifies one of the two alternatives. This method selects the alternative with the least opportunity loss using the following formula:

$$MR = min[max((A_{Y,N}\ \text{max}) - A_j\ \text{max}, (A_{Y,N}\ \text{min}) - A_j\ \text{min})] \tag{8}$$

where $A_{(Y,N)}$ max or $A_{(Y,N)}$ min respectively represent the maximum or minimum from both alternatives. In this method, in cases of a tie, we break the tie by the use of a coin toss.

In the proposed framework, the dataset is prepared for training by extracting the entity mentions and features. The label of each sample provided in the dataset is also used for training. The features used for training the classifiers are categorized into four different sets as shown in Table 1. The individual classifiers are trained separately using different combinations of these feature categories with the aim of creating diversity among the pool of classifiers. At the CSS stage of the framework, the GA is employed, and after a series of evolution and a number of generations, a solution is found as the best classifier ensemble. The classifiers are represented by the first M - 1 bits of the chromosome, while the voting bit that is to determine the voting method to be used for combining the members of the ensemble is represented by the last bit of the chromosome. At the end of the evolution, the best chromosome is selected from the final population generated based on its fitness. In the fittest chromosome, the bits represented by "1" signify the base classifiers that participated in the decision-making process.

We performed the experiments using the 9 different settings in order to determine which of the settings generates the best classifier ensemble adaptable to the test sample in the last stage of our experiment. The initial population consists of 100 unique chromosomes that were generated randomly. However, the same set of randomly generated chromosomes is used as the initial population throughout the experiment for proper comparison of the evolution and results of the different settings employed. Table 3 shows the settings used in this experiment.

Table 3. Experimental settings.

S. no.	Settings	Selection algorithm		Voting method	
		Roulette wheel	Tournament	Hurwicz criterion	Minimax regret
1	RTHM	X	X	X	X
2	RTH	X	X	X	
3	RTM	X	X		X
4	RHM	X		X	X
5	RH	X		X	
6	RM	X			X
7	THM		X	X	X
8	TH		X	X	
9	TM		X		X

The acronyms (RTHM, RTH, RTM etc.) shown in Table 3) used to describe the settings are derived from the first characters of the names of the selection algorithms and voting methods employed in that setting.

For example, RTHM shows that both roulette wheel and tournament selection algorithms along with Hurwicz criterion and minimax regret voting methods are used in the setting. Likewise, THM shows that only the tournament selection algorithm was used while both voting methods were employed in the setting. In settings RTHM, RTM, and RTH, where the two selection algorithms are employed, a coin toss is used to select the algorithm used for the selection process of paring all the chromosomes to be considered for reproduction in each generation. The main objective of using different selection methods is to create more variations in the type of chromosomes selected for reproduction. Additionally, in settings RTHM, RHM, and THM, where the two voting methods are employed for classifier combination, the choice of voting method used depends on the voting bit on the individual ensemble. Since the fitness of a given chromosome can be affected by the quality of the voting method used, this seeks to achieve the best possible combination solution over time during the evolution.

The classifier results obtained from the development dataset are used throughout the GA evolution process. The test dataset was only employed after the evaluation process for the validation of our GA framework.

3. Results and discussion

3.1. Results evaluation

Using the development dataset, the base classifiers are tested with the four different feature subsets described in Table 1 and the results are reported in Table 4.

The evolution process was performed on the development dataset using the 9 experimental settings described in Table 3. After the evolution process, we selected from all 9 settings the two fittest chromosomes produced after 100 generations and the results are reported in Table 5. Additionally, the performance of the base classifiers was evaluated by our GA framework using the HC and MR voting methods. A classifier ensemble containing the base classifiers was combined using the MR and HC voting methods and it produced F-scores of 45.37% and 48.78%, respectively, which are also reported in Table 5 as FullM and FullH.

Table 5 shows that the fittest classifier ensembles (chromosomes) are produced in setting RH, where the roulette wheel selection and HC voting method are employed, and in setting TM, where the tournament selection and MR voting method are employed. These fittest ensembles are then applied to the test dataset for validation and the results are reported in Table 6. Validation on the test dataset is performed to determine the performance of our evolution system in order to compare our system with the state-of-the-art systems and to determine the ability of the fittest classifier ensembles generated after the evolution to handle generalization tasks on the test dataset.

After validating the fittest classifier ensembles, the best performing ensemble produced an F-score of 64.45% and was generated from setting RTHM when the MR voting method was employed for classifier combination. When we validated the settings FullM and FullH, which signify the ensembles of all classifiers, they produced F-scores of 46.13% and 46.58% for MR and HC, respectively.

As presented in Table 4 for the development dataset, the individual performances of the base classifiers when tested on the four different feature subsets using the test dataset are reported in Table 7.

3.2. Discussion

The classifier ensembles in Table 5 reveal that they are composed of different classifier algorithms trained on different feature sets. For example, the second classifier ensemble "000001000100000100000000000011010000001000" from RTHM with MR employed as the voting method shows that the classifiers are selected from three different feature sets, which are Sets A, B, and D, and comprise BN K2, BN TAN, R4, and SVM2 classifier algorithms.

Table 4. Results obtained from the individual classifiers using the development set.

S. no.	Classifier	Feature subsets	P (%)	R (%)	F1 (%)
1	BN Hill	A	62.25	52.47	56.94
2		B	48.84	6.23	11.05
3		C	65.12	48.52	55.61
4		D	61.39	52.47	56.58
5	BN K2	A	62.31	52.27	56.85
6		B	48.84	6.23	11.05
7		C	65.12	48.52	55.61
8		D	61.69	52.67	56.82
9	BN TAN	A	78.95	40.02	53.12
10		B	47.45	6.42	11.31
11		C	87.68	35.18	50.21
12		D	79.11	40.42	53.50
13	J48	A	72.73	50.59	59.67
14		B	54.82	26.98	36.16
15		C	83.55	37.65	51.91
16		D	71.19	49.80	58.60
17	NB	A	74.49	36.07	48.60
18		B	41.26	14.23	21.16
19		C	73.76	32.21	44.84
20		D	80.75	31.92	45.75
21	NBK	A	74.25	43.87	55.15
22		B	49.59	17.98	26.39
23		C	76.16	37.25	50.03
24		D	69.79	47.04	56.20
25	R3	A	40.92	34.98	37.72
26		B	35.81	26.68	30.58
27		C	42.37	42.79	42.58
28		D	53.95	51.98	52.95
29	R4	A	94.67	35.08	51.19
30		B	79.17	9.39	16.79
31		C	93.55	28.66	43.88
32		D	89.54	46.54	61.25
33	SVM1	A	57.67	49.80	53.45
34		B	43.18	35.38	38.89
35		C	47.66	34.19	39.82
36		D	68.84	37.55	48.59
37	SVM2	A	72.28	42.00	53.13
38		B	66.05	28.06	39.39
39		C	71.35	37.15	48.86
40		D	68.50	44.27	53.78

The individual classifiers in the ensemble produced average performances on the test dataset in terms of recall and F-score as reported in Table 7. However, by employing the classifier combination method, this classifier

Table 5. The fittest chromosomes from the 9 settings on the development dataset.

S. no.	Settings	Chromosomes	P (%)	R (%)	F1 (%)
1	RTHM	00000100010000000000000000001101000001000	78.94	54.45	64.45
		00000100010000010000000000001101000001000	68.57	59.29	63.59
2	RTH	00000100110001000100000011001101000010001	76.38	55.93	64.57
		00000100110001000100010011001101000010001	76.38	55.93	64.57
3	RTM	01000000010001000000110010001101000010000	47.99	65.02	62.58
		01000000010001000000110000001101000011000	61.99	62.06	62.58
4	RHM	00000100001000000100010001001101000011000	71.53	58.10	64.12
		00000100001000000100000001001101000011000	71.36	58.10	64.05
5	RH	01000100110001000100000010000001000010001	76.42	56.03	64.66
		01000100110001000100010011000001000011001	76.42	56.03	64.66
6	RM	01000000010010001010010000000101010011000	50.97	67.29	60.68
		01000000010010001010010000000101010001100	50.71	67.19	60.20
7	THM	01000100010101000100000001001101000010000	71.09	58.79	64.52
		01000100010101000100000001001101000010001	76.08	55.93	64.47
8	TH	01000100110001000100010010000111000011001	75.68	55.34	63.93
		01000100110001000100010010000111010011001	75.68	55.34	63.93
9	TM	01000000110001000100010001000001000010000	71.70	59.58	64.66
		01000000110001000100010001000001000011000	69.44	60.18	64.66
	FullM	11111111111111111111111111111111111111110	31.73	79.55	45.37
	FullH	11111111111111111111111111111111111111111	36.14	75.00	48.78

ensemble produced the best F-score of 64.45% on the test dataset as shown in Table 6. The improved performance of the classifier ensembles from the average performances of the individual classifiers is due to the voting methods employed. These voting methods help to improve the complementarity in the classifiers and maximize the strengths of the best performing classifiers in the ensemble. Furthermore, unlike conventional methods, these voting methods handle the diversity of classifiers in an ensemble better in order to make a more accurate decision.

Although these voting methods show good decision-making ability and efficiency in classifier combination, they also have some drawbacks. Consider the first classifier ensemble from setting RTH presented in Table 5. The chromosome "00000100110001000100000011001101000010001" shows that the classifiers are selected from feature sets A, B, and D and comprise 7 different classifiers (BN K2, BN TAN, J48, NB, R3, R4, and SVM2). The ensemble when applied to the test dataset reported in Table 6 is used to discuss the limitations of combining the two voting methods employed in our approach. This classifier ensemble is a collection of classifiers 6, 9, 10, 14, 18, 25, 26, 29, 30, 32, and 37.

From the test dataset, consider the abstract excerpts from the documents with PubMed IDs 23433219 and 24100257, respectively.

Excerpt 1: "…for **methamphetamine**-induced psychosis and other Axis I **psychiatric disorders**."
Excerpt 2: "…Extensive literature search revealed multiple cases of **coronary artery vasospasm** secondary to **zolmitriptan**."

Table 6. Results obtained by applying the fittest classifier ensembles on the test dataset.

S. no.	Settings	Chromosomes	P (%)	R (%)	F1 (%)
1	RTHM	00000100010000000000000000001101000001000	88.93	48.97	63.16
		00000100010000010000000000001101000001000	73.91	57.13	64.45
2	RTH	00000100110001000100000011001101000010001	84.79	50.19	63.06
		00000100110001000100010011001101000010001	84.79	50.19	63.06
3	RTM	01000000010001000000110010001101000010000	49.03	63.70	55.41
		01000000010001000000110000001101000011000	63.05	62.10	62.57
4	RHM	00000100001000000100010001001101000011000	75.79	53.75	62.90
		00000100001000000100000001001101000011000	85.12	49.91	62.92
5	RH	01000100110001000100000010000001000010001	84.54	50.28	63.06
		01000100110001000100010011000001000011001	53.09	59.66	56.18
6	RM	01000000010010001010010000000101010011000	52.38	68.11	59.22
		01000000010010001010010000000101010001100	49.97	71.29	58.76
7	THM	01000100010101000100000001001101000010000	75.65	54.22	63.17
		01000100010101000100000001001101000010001	84.54	50.28	63.06
8	TH	01000100110001000100010010000111000011001	51.11	62.57	56.26
		01000100110001000100010010000111010011001	44.49	67.45	53.62
9	TM	01000000110001000100010001000001000010000	75.93	53.85	63.01
		01000000110001000100010001000001000011000	65.58	60.23	62.79
	FullM	11111111111111111111111111111111111111110	31.67	84.90	46.13
	FullH	11111111111111111111111111111111111111111	32.61	81.52	46.58

The aim of the classifier ensemble is to decide whether or not there exists a CID relation between the entity mentions shown as boldfaced in the excerpts. In Excerpt 1, there exists no true CID relation between the chemical "methamphetamine" and the disease "psychiatric disorders" mentions with concept identifier D008694 and D001523, respectively. However, based on the HC and MR methods, the decision for this relation instance differs between the two methods. This scenario is explained by applying Equations (5), (6), and (8). From the classifier ensemble, only classifier 25 with an F-score of 37.72% predicted "Yes", while the others predicted "No". The best and worst F-scores of the classifiers that predicted "No" are 61.25% and 11.05%, respectively. Applying Equations (5) and (6) for the HC method where $\alpha = 0.6$ gives:

$$H_{Yes} = 0.6\ (37.72) + 0.4\ (37.72) = 37.72$$
$$H_{No} = 0.6\ (61.25) + 0.4\ (11.05) = 41.17$$

Since H_{No} is better than H_{Yes}, the decision from HC is "No". However, when applying Equation (8), the MR method gives:

$$MR_Y = \max\ [\max\ (37.72, 61.25) - 37.72), (\max\ (37.72, 11.05) - 37.72)] = 23.53$$
$$MR_N = \max\ [\max\ (37.72, 61.25) - 61.25), (\max\ (37.72, 11.05) - 11.05)] = 26.67$$
$$MR = \min\ (MR_Y, MR_N) = 23.53 = MR_Y$$

The decision from MR is "Yes" since MR_Y gives a lower opportunity cost compared to MR_N.

Table 7. Results obtained from the individual classifiers using the test set.

S. no.	Classifier	Feature subsets	P (%)	R (%)	F1 (%)
1	BN Hill	A	52.53	62.64	57.14
2		B	7.13	61.29	12.77
3		C	49.34	67.61	57.05
4		D	52.53	62.29	57.00
5	BN K2	A	52.25	63.22	57.21
6		B	7.13	61.29	12.77
7		C	49.44	67.48	57.07
8		D	52.44	62.67	57.10
9	BN TAN	A	40.90	79.85	54.09
10		B	7.60	57.86	13.44
11		C	36.21	90.40	51.71
12		D	41.18	80.11	54.40
13	J48	A	52.25	76.83	62.20
14		B	28.71	61.32	39.11
15		C	39.02	86.85	53.85
16		D	53.38	75.17	62.43
17	NB	A	38.74	79.73	52.14
18		B	14.92	47.46	22.70
19		C	34.24	77.49	47.49
20		D	35.37	82.14	49.45
21	NBK	A	46.06	76.72	57.56
22		B	23.83	61.80	34.40
23		C	39.12	81.45	52.85
24		D	47.37	70.73	56.74
25	R3	A	39.02	42.15	40.52
26		B	25.89	33.99	29.39
27		C	44.28	42.29	43.26
28		D	52.44	51.71	52.07
29	R4	A	35.74	97.69	52.33
30		B	10.60	81.88	18.77
31		C	30.68	95.61	46.45
32		D	48.41	88.81	62.66
33	SVM1	A	50.47	56.87	53.48
34		B	39.40	46.56	42.68
35		C	39.49	51.28	44.62
36		D	38.27	71.20	49.78
37	SVM2	A	9.38	38.31	15.07
38		B	0.38	100	0.76
39		C	39.59	71.77	51.03
40		D	10.23	23.80	14.31

Furthermore, in Excerpt 2, a true CID relation exists between the chemical "zolmitriptan" and the disease "coronary artery vasospasm" mentions with concept identifier C089750 and D003329, respectively. Only

classifier 14 predicted "Yes" to these entities having a relationship. When calculating the decisions of the HC method using Equations (5) and (6) and the MR method using Equation (8), HC predicted "No", while MR predicted "Yes". These examples show the limitations of HC handling the scenarios where an alternative has only a single option (classifier) and MR handling the scenarios where an alternative produces both the best and the worst performances. We handled these limitations during the evolution by allowing the classifiers chosen in a chromosome to be combined using one of the two voting methods over the course of the evolution. The voting bit on the chromosome is mutated with a probability of 0.2 and this helps to improve the performance of the ensembles generated during the evolution.

3.2.1. Comparison of results

In Table 8, we compare the best performing ensemble produced by our GA framework with other state-of-the-art systems that used the BioCreative V corpus test dataset. Zheng et al. [22] achieved an F-score of 54.30% by integrating long-short term memory units (LSTM) into their CNN for the extraction of high-level semantic relations between the chemical and disease mentions. Zhou et al. [35] achieved an F-score of 55.05% by extracting direct semantic and syntactic relations between the entity mentions by using the shortest dependency path tree. Alam et al. [23] used features extracted from CTD [45] alongside other linguistic features to achieve an F-score of 56.60%. Xu et al. [36] achieved an F-score of 57.03% by using KB features in training their SVM classifier. They also introduced relation labels of chemicals and diseases available in CTD. Gu et al.

Table 8. Comparisons with the related works.

Systems	Descriptions	P (%)	R (%)	F1 (%)
Zheng et al. [22]	CNN, LSTM	45.20	68.10	54.30
Zhou et al. [35]	Tree kernel, three parsers	58.63	51.87	55.05
Alam et al. [23]	Knowledge approach	43.68	80.39	56.61
Xu et al. [36]	SVM, KB features	55.67	58.44	57.03
Gu et al. [31]	CNN, ME	60.90	59.50	60.20
Lowe et al. [27]	Heuristic rules	59.29	62.29	60.80
Peng et al. [37]	SVM, KB	62.07	64.17	63.10
Onye et al. [24]	SVM, KB	76.90	56.50	63.10
Our system	MCS using GA	73.91	57.13	64.45

[31] achieved an F-score of 60.20% by introducing CNN and linguistic features alongside maximum entropy (ME) models in their ML-based system. Lowe et al. [27] achieved good results (an F-score of 60.80%), but the computational cost, as well as the huge amount of time their system requires for this task, makes their system limited. Peng et al. [37] used a set of linguistic knowledge and statistic features to attain an F-score of 63.1%. They also trained their system with an additional 500 BioCreative V development dataset and 18,410 CTD-Pfizer documents from [46] to improve the performance of their system to 71.83%.

4. Conclusion

Our paper has discussed a novel approach for improving an MCS framework. This includes: (1) the random selection of two different selection algorithms and two types of crossovers during the evolution process introduc-

tion, (2) the use of a voting bit to determine the voting method used for the combination of the classifiers in the ensemble, and (3) the introduction of two decision-making under uncertainty techniques (minimax regret and Hurwicz criterion) used as the voting methods.

Our approach produced results comparable to the current state-of-the-art CID relation extraction systems. The two voting methods consider multiple opinions about every alternative before carefully making a decision, unlike the conventional ones that normally make decisions from a single opinion of the strength, weight, or percentage of the alternatives. Due to the critical nature of the decisions to be made, their decision-making is pessimistic as they try to avoid making costly decisions at every point. These methods also showed that by increasing the diversity and complementarity among the classifiers in the MCS without the literature requirement of making up the MCSs with well-performing individual classifiers, efficiency and results are most importantly not sacrificed.

Despite the success of this approach, there is a need for further improving the system. For instance, this can be achieved by increasing the pool of classifiers to determine the effect a larger pool can have, and also to apply a control function to the voting methods to help them overcome their limitations and to further improve their performances.

4.1. Conflict of interest

The authors declare that there is no conflict of interest.

References

[1] Airola A, Pyysalo S, Björne J, Pahikkala T, Ginter F et al. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. BMC Bioinformatics 2008; 9 (Suppl. 11): S2. doi: 10.1186/1471-2105-9-S11-S2

[2] Bui QC, Katrenko S, Sloot PM. A hybrid approach to extract protein-protein interactions. Bioinformatics 2010; 27 (2): 259-265. doi: 10.1093/bioinformatics/btq620

[3] Miwa M, Sætre R, Miyao Y, Tsujii J. A rich feature vector for protein-protein interaction extraction from multiple corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing; Singapore; 2009. pp. 121-130.

[4] Pons E, Becker BF, Akhondi SA, Afzal Z, Van Mulligen EM et al. Extraction of chemical-induced diseases using prior knowledge and textual information. Database 2016; 2016: baw046. doi: 10.1093/database/baw046

[5] Kuncheva LI. Combining Pattern Classifiers: Methods and Algorithms. Hoboken, NJ, USA: John Wiley & Sons, 2004.

[6] Giot R, Rosenberger C. Genetic programming for multibiometrics. Expert Systems with Applications 2004; 39 (2): 1837-1847.

[7] Ruta D, Gabrys B. An overview of classifier fusion methods. Computing and Information Systems 2000; 7 (1): 1-10.

[8] Arasu A, Götz M, Kaushik R. On active learning of record matching packages. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data; 2010. pp. 783-794.

[9] Gabrys B. Combining neuro-fuzzy classifiers for improved generalisation and reliability. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN2002), A Part of the WCCI2002 Congress; Honolulu, HI, USA; 2002. pp. 2410-2415.

[10] Kuncheva L. Fuzzy Classifier Design. Berlin, Germany: Springer Science & Business Media, 2000.

[11] Hao H, Liu CL, Sako H. Comparison of genetic algorithm and sequential search methods for classifier subset selection. In: IEEE 2003 Seventh International Conference on Document Analysis and Recognition; 2003. pp. 765-769. doi: 10.1109/ICDAR.2003.1227765

[12] Gabrys B, Ruta D. Genetic algorithms in classifier fusion. Applied Soft Computing 2006; 6 (4): 337-347.

[13] Beitia IM. Contributions on distance-based algorithms, multi-classifier construction and pairwise classification. PhD, Universidad del País Vasco-Euskal Herriko Unibertsitatea, San Sebastián, Spain, 2015.

[14] Bellare K, Iyengar S, Parameswaran A, Rastogi V. Active sampling for entity matching with guarantees. ACM Transactions on Knowledge Discovery from Data (TKDD) 2013; 7 (3): 12.

[15] Bennett PN, Carvalho VR. Online stratified sampling: evaluating classifiers at web-scale. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management; 2010. pp. 1581-1584.

[16] McDowell LK, Gupta KM, Aha DW. Cautious collective classification. Journal of Machine Learning Research 2009; 10: 2777-2836.

[17] Sen P, Namata G, Bilgic M, Getoor L, Galligher B et al. Collective classification in network data. AI Magazine 2008; 29 (3): 93.

[18] Zenobi G, Cunningham P. Using diversity in preparing ensembles of classifiers based on different feature subsets to minimize generalization error. In: European Conference on Machine Learning; Berlin, Germany; 2001. pp. 576-587.

[19] Onye SC. Novel approaches for relation extraction in biomedical domain. PhD, Eastern Mediterranean University, Famagusta, Northern Cyprus, 2018.

[20] Wei CH, Peng Y, Leaman R, Davis AP, Mattingly CJ et al. Overview of the BioCreative V chemical disease relation (CDR) task. In: Proceedings of the Fifth BioCreative Challenge Evaluation Workshop; Seville, Spain; 2015. pp. 154–166.

[21] Huang CC, Lu Z. Community challenges in biomedical text mining over 10 years: success, failure and the future. Briefings in Bioinformatics 2016; 17(1): 132-144. doi: 10.1093/bib/bbv024

[22] Zheng W, Lin H, Li Z, Liu X, Li Z et al. An effective neural model extracting document level chemical-induced disease relations from biomedical literature. Journal of Biomedical Informatics 2018; 83: 1-9. doi: 10.1016/j.jbi.2018.05.001

[23] Alam F, Corazza A, Lavelli A, Zanoli R. A knowledge-poor approach to chemical-disease relation extraction. Database (Oxford) 2016; 2016: baw071. doi: 10.1093/database/baw071

[24] Onye SC, Akkeleş A, Dimililer N. relSCAN - a system for extracting chemical-induced disease relation from biomedical literature. Journal of Biomedical Informatics 2018; 87: 79-87. doi: 10.1016/j.jbi.2018.09.018

[25] Peng Y, Wei CH, Lu Z. Improving chemical disease relation extraction with rich features and weakly labeled data. Journal of Cheminformatics 2016; 8: 53. doi: 10.1186/s13321-016-0165-z

[26] Jiang Z, Jin LK, Li LS, Qin M, Qu C et al. A CRD-WEL system for chemical-disease relations extraction. In: Proceedings of the Fifth BioCreative Challenge Evaluation Workshop; Seville, Spain; 2015. pp. 317-326.

[27] Lowe DM, O'Boyle NM, Sayle RA. Efficient chemical-disease identification and relationship extraction using Wikipedia to improve recall. Database 2016; 2016: baw039. doi: 10.1093/database/baw039

[28] Zhou H, Deng H, Chen L, Yang Y, Jia C et al. Exploiting syntactic and semantics information for chemical-disease relation extraction. Database (Oxford) 2016; 2016: baw048. doi: 10.1093/database/baw048

[29] Li Z, Yang Z, Lin H, Wang J, Gui Y et al. CIDExtractor: A chemical-induced disease relation extraction system for biomedical literature. Bioinformatics and Biomedicine 2016; 2016: 994-1001. doi: 10.1109/BIBM.2016.7822658

[30] Wei CH, Peng Y, Leaman R, Davis AP, Mattingly CJ et al. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database 2016; 2016: baw032. doi: 10.1093/database/baw032

[31] Gu J, Sun F, Qian L, Zhou G. Chemical-induced disease relation extraction via convolutional neural network. Database 2017; 2017 (1): /bax024. doi: 10.1093/database/bax024

[32] Gu J, Qian L, Zhou G. Chemical-induced disease relation extraction with lexical features. In: Proceedings of the Fifth BioCreative Challenge Evaluation Workshop; Seville, Spain; 2015. pp. 220-225.

[33] Gu J, Qian L, Zhou G. Chemical-induced disease relation extraction with various linguistic features. Database 2016; 2016: baw042. doi: 10.1093/database/baw042

[34] Xu J, Wu Y, Zhang Y, Wang J, Lee HJ et al. CD-REST: a system for extracting chemical-induced disease relation in literature. Database (Oxford) 2016; 2016: baw036. doi: 10.1093/database/baw036

[35] Zhou HW, Deng HJ, He J. Chemical-disease relations extraction based on the shortest dependency path tree. In: Proceedings of the Fifth BioCreative Challenge Evaluation Workshop; Seville, Spain; 2015. pp. 214-219.

[36] Xu J, Wu Y, Zhang Y, Wang J, Liu R et al. UTH-CCB@ BioCreative V CDR task: identifying chemical-induced disease relations in biomedical text. In: Proceedings of the Fifth BioCreative Challenge Evaluation Workshop; Seville, Spain; 2015. pp. 254–259.

[37] Peng Y, Wei CH, Lu Z. Improving chemical disease relation extraction with rich features and weakly labeled data. Journal of Cheminformatics 2016; 8 (1): 53. doi: 10.1186/s13321-016-0165-z

[38] Miwa M, Sætre R, Miyao Y, Tsujii J. Protein–protein interaction extraction by leveraging multiple kernels and parsers. International Journal of Medical Informatics 2009; 78 (12): e39-e46. doi: 10.1016/j.ijmedinf.2009.04.010

[39] Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M et al. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of Biomedical Informatics 2012; 45 (5): 885-92. doi: 10.1016/j.jbi.2012.04.008

[40] Le HQ, Tran MV, Dang TH, Collier N. The UET-CAM system in the BioCretive V CDR task. In: Proceedings of the Fifth BioCreative Challenge Evaluation Workshop; Seville, Spain; 2015. pp. 1-20.

[41] Zhou H, Yang Y, Liu Z, Liu Z, Men Y. Integrating word sequences and dependency structures for chemical-disease relation extraction. In: Sun M, Wang X, Chang B, Xiong D (editors). Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Cham: Springer, 2017. pp. 97-109. doi: 10.1007/978-3-319-69005-6_9

[42] Li J, Sun Y, Johnson R, Sciaky D, Wei CH et al. Annotating chemicals, diseases and their interactions in biomedical literature. In: Proceedings of the Fifth BioCreative Challenge Evaluation Workshop; Seville, Spain; 2015. pp. 173-182.

[43] Kim S, Yoon J, Yang J, Park S. Walk-weighted subsequence kernels for protein-protein interaction extraction. BMC Bioinformatics 2010; 11: 107. doi: 10.1186/1471-2105-11-107

[44] Murugesan G, Abdulkadhar S, Natarajan J. Distributed smoothed tree kernel for protein-protein interaction extraction from the biomedical literature. PLoS One 2017; 12 (11): e0187379. doi: 10.1371/journal.pone.0187379

[45] Davis AP, Murphy CG, Saraceni-Richards CA, Rosenstein MC, Wiegers TC et al. Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical–gene–disease networks. Nucleic Acids Research 2008; 37: D786-D792. doi: 10.1093/nar/gkn580

[46] Davis AP, Wiegers TC, Roberts PM, King BL, Lay JM et al. A CTD–Pfizer collaboration: manual curation of 88 000 scientific articles text mined for drug–disease and drug–phenotype interactions. Database (Oxford) 2016; 2016: bat080. doi: 10.1093/database/bat080

[47] Rocha M, Neves J. Preventing premature convergence to local optima in genetic algorithms via random offspring generation. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems; Springer, Berlin; 1999. pp. 127-36.

[48] Dimililer N, Varoglu E, Altançay H. Vote-based classifier selection for biomedical NER using genetic algorithms. In: Martí J, Benedí JM, Mendonça AM, Serrat J (editors). Pattern Recognition and Image Analysis. IbPRIA 2007. Lecture Notes in Computer Science. Berlin: Springer, 2007, pp. 202-209.

[49] Pažek K, Rozman Č. Decision making under conditions of uncertainty in agriculture: a case study of oil crops. Poljoprivreda 2009; 15 (1): 45-50.