Hybrid imbalanced data classifier models for computational discovery of antibiotic drug targets

Yucel Kocyigit and Huseyin Seker*

Abstract—**Identification of drug candidates is an important but also difficult process. Given drug resistance bacteria that we face, this process has become more important to identify protein candidates that demonstrate antibacterial activity. The aim of this study is therefore to develop a bioinformatics approach that is more capable of identifying a small but effective set of proteins that are expected to show antibacterial activity, subsequently to be used as antibiotic drug targets. As this is regarded as an imbalanced data classification problem due to smaller number of antibiotic drugs available, a hybrid classification model was developed and applied to the identification of antibiotic drugs. The model was developed by taking into account of various statistical models leading to the development of six different hybrid models. The best model has reached the accuracy of as high as 50% compared to earlier study with the accuracy of less than 1% as far as the proportion of the candidates identified and actual antibiotics in the candidate list is concerned.**

I. Introduction

Despite advances in the bio-medical technology that further aid understand biological systems, drug discovery is still one of the most difficult and long-lasting processes, which are not only costly but also with low success rate of new therapeutic outcome [1]. In order to address this problematic process, computational methods have been proposed to identify potential drug targets and candidates [2]. For example, interaction between drugs and target proteins is predicted in order to identify potential new drugs or novel targets for the existing drugs [3]. In addition, there have been attempts to discover drug candidates from protein sequence information by using computational intelligence and statistical predictive methods [4]. Due to smaller number of drug target available, this problem can be regarded as imbalanced data classification problem. A major problem is that a classifier developed with the imbalanced data set tends to classify an object to a class with the highest number of samples, in which case it is the non-drug target set. The outcome of such study generally results in higher overall classifier accuracy but lower sensitivity. Therefore, the classifier is more biased towards the non-drug set and cannot be reliable [5].

Yucel Kocyigit is with Electrical and Electronics Engineering Department, Celal Bayar University, 45140, Manisa, Turkey (e-mail: yucel.kocyigit@cbu.edu.tr)

Huseyin Seker is with the Bio-Health Informatics Research Group, Faculty of Technology, De Montfort University, Leicester, LE1 9BH, UK..(email: hseker@dmu.ac.uk).

*Corresponding author: H. Seker (email: hseker@dmu.ac.uk)

The data sets used for antibiotic drug discovery are of similar nature and therefore should be analysed by using computational predictive model that is more capable of dealing with such imbalanced data sets.

II. Materials and Methods

In this work, the drug data bank was used to form the basis of this study [6]. In order to compare our study with previous studies, the data set presented in [4] was first utilized. It consists of 22 E.Coli (strain K-12) drug target proteins and 4243 E. Coli proteins used as non-drug targets.

The bacterial target dataset was downloaded from the Drug-Bank [6] in March 2011. As one of the most common species is the E. coli (strain K-12) used as antibiotic drug targets, this current study has therefore particularly focused on the E. coli (strain K-12).

The non-targets dataset was downloaded from the High-quality Automated and Manual Annotation of microbial Proteomes (HAMAP) at Expasy [7]. One of the challenging parts of the study was to determine non-drug targets. In order to maintain consistency with the drug target set, this current study has taken into account of the E. coli (strain K-12) only as this is found to be one of the most common species for all the targets. The E. coli proteome currently contains highly accurate and complete sequences for the K-12 strain with 4282 entries [6]. This set was culled to maintain 40% sequence identity resulting in 3753 proteins for the non- redundant dataset. Furthermore, there were 22 approved antibiotic drug targets found in the data set. They were then removed from the set to form a non-redundant non-drug-target dataset of 3731 entries. Final data set now contains 22 drug-targets and 3731 non-drug targets proteins along with their sequence information. The proteins were then represented by using their amino acid composition.

In this study, a basic sample ordering procedure used for pre-processing the imbalanced data set for the classification is proposed. The pre-processing procedure consists of calculating the Euclidean distance between targets and non-targets and ordering the distances by means of two methods, *ttest* and *the sum of distances*, for each non-target samples.

Due to smaller number of drug targets available, non-drug targets need to be separated into a number of sub-sets with smaller number of proteins in order to overcome the problem of imbalanced data classification. This is achieved by calculating the Euclidean distances between drug targets and non-drug targets based on amino acid composition of the proteins. These sub-sets are constructed by applying the

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

"ttest statistical" and "sum of distance" methods. For the sum of distance method, sum of all the distances between non-drug targets and drug targets for each drug target is calculated, averaged and ranked in descending order. These sub-sets of proteins that contain non-drug targets can then be easily separated from each other. Alternativley, the ttest statistical method is also utilised instead of the average distance to determine the sub-groups where the ranking is based on the statistical *p* values of each sub-group.

Having identified the sub-groups for non-drug target proteins, a base classifier is applied to perform the classification task. In order to show independence of the method of a specific classifier, the analysis is carried out by using three different base classifiers, namely, Support Vector Machine, Linear Discriminative and Naïve Bayesian classifiers, which have been selected due to differences in their algorithmic structures.

The proposed models to be utilized are listed in Table I and explained in the subsequent sections. For example, the model M1 as noted in the subsequent tables is constructed by using both ***ttest*** and ***LDA***.

In order to assess the performance of all the models, a 5-fold cross validation is used.

TABLE I. PROPOSED MODELS FOR THE IMBALANCED DATA CLASSIFICATION

	Ordering Methods		**Classification Methods**		
	ttest	***sum of distances***	***LDA***	***SVM***	***Bayesian***
M1	√		√		
M2		√	√		
M3	√			√	
M4		√		√	
M5	√				√
M6		√			√

A. Support Vector Machine

Support vector machine (SVM) is one of the widely used classifiers and has been shown to yield better generalization ability for high-dimensional data sets [8]. For the SVM classifier, it is assumed that a two-group data set can always be separated by a hyper-plane provided that a suitable non-linear mapping to a sufficiently high dimension is found. In addition, one of the main tasks during the construction of SVMs is to find separating hyper-plane(s) with the largest possible margin in order to result in a classifier with better generalization ability. The data points that highly represent the hyper-planes are the support vectors derived from the samples during training are. These points can then be regarded as the most representative data samples that could help build a robust classifier [9].

B. Linear Discriminant Analysis

For a two-class classification problem, linear discriminant analysis (LDA) tries to find one hyper-plane to separate one group from another one. There are various parameters that affect performance of LDA including distance metric such as Euclidean and Mahalanobis distances [9]. As LDA is one of the well-known and simple classifier models, it is also used as one of the base classifiers in order to show applicability and robustness of the method developed in this study. For the sake of simplicity, the Euclidean distance metric was selected to implement LDA-based classifier.

C. Naïve Bayes

Bayesian classifier is created as to minimize the overall misclassification using a cost function. Naive Bayes classifier is one of the well-known and widely used Bayesian classifiers and easy to implement [9]. As its architecture is different from both SVM and LDA, it is also chosen as one of the base classifiers in order to show independence of the developed method.

D. Performance Evaluation of the Classifiers

As in this study, the binary classification yields two discrete results, namely positive and negative for which there are four possible outcomes; if a positive instance is classified correctly, it is counted as a true positive (TP), otherwise a false negative (FN) whereas if the negative instance is classified correctly, it is counted as a true negative (TN), otherwise a false positive (FP).

Overall accuracy is generally used to assess overall performance of classifiers. However, it is alone not a reliable metric for imbalanced data classification problem as the influence of negative samples on overall accuracy is much higher than that of positive samples. Therefore, along with the accuracy, the following metrics should also be presented in for more objective evaluation [10]

$$\text{accuracy}=(TP+TN)/(TP+TN+FP+FN)$$
$$\text{sensitivity}=TP/(TP+FN)$$
$$\text{specificity}=TN/(TN+FP)$$
$$\text{precision}=TP/(TP+FP)$$
$$\text{gmeans}=(\text{sensitivity}*\text{specificity})^{1/2}$$
$$\text{f-measure}=2*\text{sensitivity}*\text{precision}/(\text{sensitivity}+\text{precision})$$

In this study, the minority and majority classes represent positive and negative groups, respectively.

III. RESULTS AND DISCUSSION

Experiments were carried out using three base classifiers (SVM, LDA and Naïve Bayes) with 22 drug-targets and 3731 non-drug targets. In order to compare our study with previous studies, the drug targets presented in [4] were utilized, which consists of 22 E.coli (strain K-12) proteins and the non-drug targets with 3731 E.coli proteins.

The drug and non-drug targets samples were trained and divided into sub-groups by using the proposed methods, which subsequently produced 37 sub-groups. Each sub-group has 22 drug-targets and 100 non-drug targets except for the last group that consists of 130 non-drug targets. The results obtained to assess the training performance of all the six predictive models (M1-M6) are presented in Table II.

TABLE II. THE RESULTS OF TRAINING PERFORMANCE FOR EACH 37 GROUP

M1	Statistical results of training performance for M1 and M2					M2				
	Accuracy	*Sensitivity*	*Specificity*	*g-means*	*f-measure*	*Accuracy*	*Sensitivity*	*Specificity*	*g-means*	*f-measure*
Max	0.810	0.864	0.870	0.815	0.594	0.787	0.909	0.820	0.815	0.594
Min	0.607	0.273	0.570	0.431	0.200	0.566	0.136	0.620	0.318	0.102
Mean	0.685	0.592	0.705	0.639	0.399	0.684	0.585	0.705	0.632	0.396
Median	0.680	0.545	0.700	0.629	0.394	0.672	0.591	0.700	0.632	0.400
SD	0.044	0.165	0.050	0.086	0.085	0.056	0.191	0.046	0.121	0.120

M3	Statistical results of training performance for M3 and M4					M4				
	Accuracy	*Sensitivity*	*Specificity*	*g-means*	*f-measure*	*Accuracy*	*Sensitivity*	*Specificity*	*g-means*	*f-measure*
Max	0.869	0.636	0.970	0.740	0.560	0.852	0.727	0.954	0.800	0.640
Min	0.689	0.091	0.800	0.273	0.095	0.689	0.000	0.800	0.000	0.065
Mean	0.767	0.281	0.873	0.479	0.292	0.771	0.285	0.876	0.463	NaN
Median	0.779	0.227	0.870	0.450	0.263	0.770	0.227	0.880	0.450	NaN
SD	0.038	0.156	0.036	0.126	0.126	0.037	0.200	0.038	0.181	NaN

M5	Statistical results of training performance for M5 and M6					M6				
	Accuracy	*Sensitivity*	*Specificity*	*g-means*	*f-measure*	*Accuracy*	*Sensitivity*	*Specificity*	*g-means*	*f-measure*
Max	0.967	0.864	0.990	0.925	0.905	0.984	0.909	1.000	0.953	0.952
Min	0.779	0.273	0.850	0.493	0.308	0.770	0.182	0.880	0.413	0.250
Mean	0.856	0.534	0.926	0.691	0.560	0.880	0.515	0.960	0.686	0.592
Median	0.844	0.545	0.930	0.693	0.522	0.869	0.455	0.970	0.650	0.563
SD	0.056	0.209	0.037	0.145	0.186	0.068	0.239	0.034	0.174	0.237

TABLE III. TEST RESULTS FOR M1

# of Groups Supporting Drug Candidates	*# of Approved Drug Candidates*	*# of Total Drug Candidates*
21	15	571
22	14	476
23	12	387
24	10	273
25	7	172
26	6	58

TABLE IV. TEST RESULTS FOR M2

# of Groups Supporting Drug Candidates	*# of Approved Drug Candidates*	*# of Total Drug Candidates*
17	16	803
18	15	658
19	15	534
20	15	394
21	12	254
22	10	115

TABLE V. TEST RESULTS FOR M3

# of Groups Supporting Drug Candidates	*# of Approved Drug Candidates*	*# of Total Drug Candidates*
18	5	132
19	5	94
20	4	68
21	2	44
22	1	24
23	0	6

TABLE VI. TEST RESULTS FOR M4

# of Groups Supporting Drug Candidates	*# of Approved Drug Candidates*	*# of Total Drug Candidates*
18	6	119
19	5	91
20	5	63
21	5	37
22	3	13
23	2	4

TABLE VII. TEST RESULTS FOR M5

# of Groups Supporting Drug Candidates	# of Approved Drug Candidates	# of Total Drug Candidates
18	16	571
19	15	499
20	15	425
21	9	221
22	7	165
23	6	69

TABLE VIII. TEST RESULTS FOR M6

# of Groups Supporting Drug Candidates	# of Approved Drug Candidates	# of Total Drug Candidates
16	20	715
17	18	627
18	17	522
19	16	446
20	6	165
21	6	107

As seen in Table II, in order to select the sub-groups that yield the best performance, the median values of the accuracies, gmeans, and f-measures are considered. Having chosen the sub-groups for testing all data and removing the non-drug target samples belonging to that training group, several thresholds that refer to the number of sub-groups supporting drug candidates were determined to obtain potential drug candidates. As a result of the analyses presented in Tables III - VIII, number of potential drug candidates can be seen for several thresholds.

As mentioned earlier, the drug candidates are compared with those found in the Drug Databank [6]. The 2nd column in the tables gives the number of approved drug candidates that we found in Databank while the 3rd column lists the number of total drug candidates. For example, if the number of groups supporting drug candidates is selected as 26 for M1 in Table III, the number of approved drug candidates is found to be 6 out of 58. Comparing this current study with the previous study [4], it can be seen that proportions of approved drug candidates to total drug candidates are between 0.0199 (~2%) and 0.5 (50%), whereas the earlier study presented just 0.016 corresponding to only 1 out of 64 as the number of approved drug candidates.

For the application of all the methods, the best model is found to be M4 with an accuracy of as high as 50%. The outcome suggests that this study not only increased number of potential drug candidates identified but also narrowed down the search space for experimental study for validation and verification for such potential drug candidates.

As far as the algorithmic side of the models is considered, the results appear to suggest that the ordering method "ttest" used for pre-processing is less successful to obtain the drug candidates compared to "the sum of distance" approach. It is also observed that SVM tends to identify much more drug targets compared to Bayes and LDA making the in-silico drug discovery unreliable. Finally, it may be proposed that the hybrid method with "the sum of distances" and SVM is potentially useful in identifying such drug targets successfully.

Applying the proposed method to the E. coli proteome identifies various sets of proteins that have similar properties to known antibiotic drug targets. These proteins may therefore be considered as potential new targets that may demonstrate antibacterial activity and therefore should be verified by means of lab-based experiments.

IV. CONCLUSIONS

In-silico discovery of antibiotic drug targets was studied by using sequence information of proteins and imbalanced data classifier. Comparison of antibiotic protein targets with non-target proteins from E. Coli has yielded highly comparable results when compared to the previous study [4] although only the amino acid composition of the drug target proteins was utilized as a feature set. This simple hybrid but effective method allows accurate identification of potential drug targets with an accuracy of as high as 50% compared to earlier study with an accuracy of 1.6%. The method is also shown to be independent of any base classifier and even works well with simple classifiers such as LDA.

Given the promising results, further works are now being carried out in order to (i) develop similar but better hybrid methods that should be more capable of dealing with highly imbalanced data sets and (ii) incorporate feature extraction and selection methods (e.g., for sequence-driven feature set). This will be further applied to other drug candidate applications.

REFERENCES

[1] B.D. Anson, J. Ma, and J. He, "Identifying CardiotoxicCompounds", *Genetic Engineering & Biotechnology News*, TechNote (Mary AnnLiebert), vol. 29(9), pp. 34-35, 2009.

[2] M. Bantscheff and G. Drewes, "Chemoproteomic approaches to drug target identification and drug profiling", *Bioorganic & Medicinal Chemistry*, vol.20, pp. 1973–1978, 2012.

[3] S.J. Haggarty, K.M. Koeller, J.C. Wong, R.A. Butcher, and S.L. Schreiber, "Multidimensional chemical genetic analysis of diversity oriented synthesis-derived deacetylase inhibitors using cell-based assays", *Chem. Biol.*, vol.10, pp. 383–396, 2003.

[4] T. M. Bakheet and A.J. Doig, "Properties and identification of antibiotic drug targets", *BMC Bioinformatics*,vol. 11, pp. 195-204 , 2010.

[5] Y. Kocyigit and H. Seker, "Imbalanced data classifier by using ensemble fuzzy c-means clustering ", *Proc. of BHI2012* China, Jan 2012, pp. 952-955.

[6] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A.C. Guo, D.S. Wishart, "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs", *Nucleic Acids Res.*, vol. 39, pp.1035-41, 2011

[7] http://www.expasy.ch/sprot/hamap/ [Last accessed 10 July 2013]

[8] V. Vapnik, *The Nature of Statistical Learning*. Springer-Verlag, 1995.

[9] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J.A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, and V. Robles, "Machine learning in bioinformatics", *Briefings in Bioinformatics*, vol.7, no.1, pp. 86-112, 2006.

[10] H. He and E. A. Garcia, "Learning from imbalanced data", *IEEE Transactions on Knowledge and Data Engineering*, vol.21 (9), pp.1263-1284, 2009.