Signaling Adverse Drug Reactions with Novel Feature-based Similarity Model

Fan Yang*
School of Computer Science & Technology
Shandong University, China
Email: yang4084@umn.edu

Xiaohui Yu
School of Computer Science & Technology Shandong University
School of Information Technology
York University, Toronto, Canada
Email: xhyu@yorku.ca

George Karypis
University of Minnesota, Twin City
Minneapolis, Minnesota, USA
Email: karypis@umn.edu

***Abstract*—Adverse drug reactions (ADRs) are a main cause of hospitalization and deaths worldwide. These unanticipated episodes are generally infrequent, but almost all existing ADR signaling techniques are designed to use dataset extracted from spontaneous reporting systems or employed a predefined type of information (e.g., drugs), which suffer from failures to detect unexpected and latent ADRs. In this paper, we propose a novel Feature-based Similarity model (FS) to detect the potential ADRs for medical cases using the electronic patient dataset. FS is tested on the real patient data retrieved from the US Food Drug Administration that includes 54,070 patients detail information and 9,567 ADRs records. Our model ranked all ADRs for the given medical case that combined the information of drugs, medical conditions, and patient profiles and can be applied in therapy decision support systems and unexpected ADR warning systems. The experimental results show that FS outperforms comparing methods. This paper clearly illustrates the great potential along the new direction of ADR signal generate from health care administrative database.**

I. Introduction

An adverse drug reaction (ADR) is a harmful reaction to a currently administered drug [1]. In the U.S. alone, over 21K ADRs were reported in 2004 [2], and it is estimated that each year $6\%\sim7\%$ of the hospitalized patients develop ADRs [3]. ADRs can lead to a potential of 100,000 deaths, making it the fourth largest cause of death in the U.S. [4]. Therefore, it is important to signal and predict a drug's ADRs from preclinical screening phase to post-market surveillance.

The ADRs signal algorithms can mainly be categorized in two broad areas: (1) signal associations between ***S***ingle-***D***rug and ***A***DR (SDA); (2) signal associations between ***M***ultiple-***D***rugs and ***A***DRs (MDA). In the SDA category, disproportionality analysis (DPA) is the mainly used technique to detect ADRs signals. DPA methodologies use frequency analysis of 2×2 contingency tables to estimate surrogate measures of statistical association between specific drug-event combinations mentioned in spontaneous reports (SRS), such as Gamma Poisson Shrinker, Relative reporting ratio, and Reporting odds ratio [5]. However, the statistical detection methods are not effective discover associations of multiple drugs and ADRs. The ADRs prediction methods in the SDA area can be categorized into drug-drug interactions (DDIs), protein-target based, and drug-ADR based. Thomas et al. [6] used dependency parsing and combine linguistic information in an machine learning-based approach to identify DDIs. However, this methods relies on information about known DDI which is a significant bottleneck, due to DDIs is usually lacking. Jacob et al. [7] take pairs of drug-target as inputs for SVMs with pairwise kernels to predict the association of drug and target. However the prior knowledge of association between drugs and proteins are requested. Lin et al. [8] developed an "external link prediction" approach for predicting new association of drug-ADR with using two snapshots of data based on the intersection of SIDER and FAERS. However, it also needs prior knowledge of specific association between drugs and ADRs. There have some studies on detecting new MDA, Caster et al. [9] used bayesian logistic regression on the WHO SRS to address confounding by co-medication in ADR surveillance. Harpaz et al. [10] applied association rule mining with multiple-drugs to signal associations of drugs and ADRs. However, these methods do not consider patient-profile (age, gender, and indications), which might cause false positive problems [11].

The focus of this paper is to develop methods that can predict the ADRs that a patient may develop given three types of information: patient profile, medical conditions, and drugs. For predicting the association between specific case and ADRs with high probability, we use our model to rank all ADRs based on the above mentioned information and select the top $\boldsymbol{n}$ ADRs could be developed by this patient.

This technique can be applied in two scenarios. $(\mathbf{1})$ Therapy Decision Support: almost drugs could cause ADRs for patient with a certain probability, so the risk-benefit analysis is necessary whenever a doctor makes a prescription. $(\mathbf{2})$ Early Warning System: ADR monitor agents like FDA and WHO can use such a method to detect unexpected ADRs.

To solve the ADRs problem, we developed a method that borrows the idea from recommender systems and referring to ADR and medical case as "user" and "item" respectively in our story. Our aim is to detect credible ADRs for a specific medical case that consists of patient profile, medical conditions and drugs information using the administrative database, FDA Adverse Event Reporting System (FAERS). For achieving this, we propose the $\underline{F}$eature-based $\underline{S}$imilarity (FS) model that extends the Feature-based method to the ADR detection problem to effectively mine the unexpected/potential ADRs.

* Corresponding author Email: fanyang.sdu@gmail.com

We evaluate the performance of FS by using a dataset obtained from FAEAR that contains 9,567 ADRs and 54,070 medical cases. It is compared to the commonly used Naive Feature-based similarity function, Cosine similarity function and Popularity-based function. The experimental results show that FS outperforms other competing methods in term of prediction quality of ADRs.

The remainder of the paper is organized as follows. Section 2 defines the notation that will be used throughout the paper. Section 3 describes FS model that is used to learn the model. Section 4 describes the evaluation methodology and the characteristic of dataset. In section 5, the results and discussion are provided. Finally, Section 6 provides the concluding remarks.

II. Definitions and Notations

Throughout the paper, all vectors are row vectors and are represented by bold lower case letters (e.g. $\boldsymbol{f}_i$). We will let $\boldsymbol{M}$(n, m) denote the space of real $n \times m$ matrices with n rows and m columns. The ADR and medical case relation matrix is represented by $\boldsymbol{R}$. Each row in $\boldsymbol{R}$ corresponds to an ADR and each column corresponds to an identified medical case. The r^{th} row of a matrix $\boldsymbol{R}$ is represented as $\boldsymbol{r}_i$. The entries of $\boldsymbol{R}$ are binary, indicating the ADR is developed by the medical case. The symbol $r_{i,u}$ represents the score computed by the FS model.

All ADRs are denoted as the set $\mathcal{U}$ with size of n_U, and all medical cases are denoted as the set $\mathcal{C}$ with the size of n_C. $\mathcal{R}_i^+$ represents the set of ADRs that is developed by $Case_i$. $\mathcal{R}_i^-$ represents the set of ADRs that is not developed by $Case_i$. $\mathcal{C}_u$ represents the set of medical cases that developed ADR_u. $\mathcal{C}_v$ represents the set of medical cases that developed ADR_v. Each case has a feature vector that represents some intrinsic characteristics which compose patient profile, medical conditions, and drugs. The feature vectors of all cases are represented as the matrix $\boldsymbol{F}$ whose row $\boldsymbol{f}_i$ corresponds to the $Case_i$. The number of feature vectors is referred to as n_F.

III. Methods

A. FS Model for Adverse Drug Reaction Prediction

Given the information associated with a medical case, a way of determining if it could develop a particular ADR is to compare the features describing that medical case against the features of the known medical cases that have already developed the ADR. That is, if $\boldsymbol{f}_i$ is a vector containing various features of the medical case under consideration, and u is a certain ADR, then we can compute a score as $\sum_{k \in C_u} sim(\boldsymbol{f}_k, \boldsymbol{f}_i)$, where C_u is the set of historical medical cases that developed ADR u, and $sim(\boldsymbol{f}_k, \boldsymbol{f}_i)$ is a function that measures the similarity between the two medical cases' feature vectors. The FS method that we developed uses an approach, initially developed in the context of solving the cold-start recommendation problem [12], to learn $sim(\boldsymbol{f}_k, \boldsymbol{f}_i)$.

Specifically, for each ADR it estimates its own similarity function that is defined as $Equation$ 1.

$$r_{i,u} = sim(i, u), \tag{1}$$

where $sim(i, u)$ is the ADR-specific similarity function that is given by:

$$sim(i, u) = \sum_{d=1}^{l} \sum_{k \in \mathcal{C}_u} m_{u,d} \times gsim_d(i, k), \tag{2}$$

where $gsim_d(\cdot)$ is the d^{th} global similarity function, l is the number of global similarity functions, and $m_{u,d}$ is a scalar that determines how much the d^{th} global similarity function contributes to the ADR-specific similarity function.

The similarity between two medical cases i and k under the d^{th} global similarity function $gsim(\cdot)$ is estimated using the medical cases' feature vectors as $Equation$ 3.

$$gsim_d(i, k) = \boldsymbol{w}_d \cdot (\boldsymbol{f}_i \odot \boldsymbol{f}_k)^T, \tag{3}$$

where $\odot$ is the element-wise Hadamard product operator [13], $\boldsymbol{f}_i$ and $\boldsymbol{f}_k$ are the L_2 normalized feature vectors of medical case i and k respectively, $\boldsymbol{w}_d$ is a weight vector and each entry $w_{d,c}$ holds the weight of feature c under $gsim_d(\cdot)$. Putting it all together, the score $r_{i,u}$ of the given medical case i for the ADR u is given by:

$$\begin{aligned} r_{i,u} = sim(i, u) &= \sum_{d=1}^{l} \sum_{k \in \mathcal{C}_u} m_{u,d} \times gsim_d(i, k) \\ &= \sum_{d=1}^{l} \sum_{k \in \mathcal{C}_u} m_{u,d} \times \boldsymbol{w}_d \cdot (\boldsymbol{f}_i \odot \boldsymbol{f}_k)^T, \end{aligned} \tag{4}$$

Notice that the global similarity functions $gsim_d(\cdot)$ and their weight vectors are shared among all ADRs, while each similarity functions has its own membership vector $\boldsymbol{m}_d$ that determines its contribution to the medical case similarity computations for each ADR. The weight vector $\boldsymbol{w}_d$ is orthonormal in order to ensure the independence of the different similarity functions. Since FS incorporates each ADR's own set of medical cases while applying the feature weights of the similarity functions, our model can signal the unexpected ADRs with high credibility.

The feature weights that are embodied by the similarity functions can account for the relative importance of different cases' features with respect to different ADRs. On the other hand, in the case of a low-dimensional matrix $\boldsymbol{F}$ where the number of features is limited. That is, $\boldsymbol{F}$ is a long thin matrix, which means the number of required similarity functions can be potentially high as some features can be duplicated appearing in many cases and accordingly appear in a large number of ADRs' records. This in turn results in the need for multiple similarity functions.

B. Learning for FS Model

FS learns a model $\Theta = [\boldsymbol{M}, \boldsymbol{W}]$, where $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_i, \cdots, \boldsymbol{w}_l]$, $\boldsymbol{w}_i$ represents parameters of the similarity functions and $\boldsymbol{M}$ is a $|\mathcal{U}| \times l$ matrix of "ADRs' memberships". Prior to learning the model Θ, we need to specify the number of global similarity functions that should be learned. The inputs to the learning process are: matrix $\boldsymbol{R}$, matrix $\boldsymbol{F}$, and number of ADR memberships l that we want to learn.

The FS model uses the Bayesian Personalized Ranking (BPR) loss function proposed by Rendle et al. [14]:

$$\mathcal{L}_{bpr} = -\sum_{i\in\mathcal{I}} \sum_{\substack{u\in\mathcal{R}_i^+ \\ v\in\mathcal{R}_i^-}} ln\sigma(r_{i,u}(\Theta) - r_{i,v}(\Theta)), \tag{5}$$

where $\sigma(x)$ is the sigmoid function that increases as x increases. In this paper, x is the relative rank between ADR_u (that develops for the given medical case i) and ADR_v (that does not develop for the given medical case i) estimated as:

$$r_{i,uv}(\Theta) = r_{i,u}(\Theta) - r_{i,v}(\Theta). \tag{6}$$

The relative rank $r_{i,uv}$ must be greater than 0, because $\sigma(r_{i,uv})$ increases with $r_{i,uv}$ and the loss function minimizes the negative of $\sigma(r_{i,uv})$ aggregated over all triplets (i, u, v) in the entire training set. In the learning process of the BPR function, the ranking score $r_{i,u}$ should be higher than $r_{i,v}$, as the ranked scores are used to select the highest n ADRs.

IV. Experiments and Evaluation

We generated similarity matrices by incorporating the patient profiles, medical conditions, and drugs. We then check how well the FS model ranks the ADRs and how much this performance can be improved by adding similarity matrices to the interaction matrix.

A. ADRs Data

We evaluated the performance of the FS model on the large-scaled real-world Quarterly Data extracted from the FDA Adverse Event Reporting System. We used the latest database with ASCII format that covers the period from January 1st, 1980 through 2013Q1. The dataset contains 54,070 medical cases, 738,407 drugs, 454,435 medical conditions, and 9,567 ADRs. Each ADR is associated with a feature vector, whose elements correspond to the historical medical cases. Each medical case is associated with three types of feature: patient profile, indications, and drugs.

B. Competing Methods

Personalized User Modelling Technique: the totally personalized baseline method that we compare against is inspired by the ideas from Billsus and Pazzani [15].

(1) *CoSim*: Cosine-Similarity with user profile is a personalized neighborhood-based user modeling technique. The score $r_{i,u}$ for ADR_u over the given medical case i is estimated as:

$$r_{cos} = \sum_{k=1}^{n_{C_u}} \frac{\boldsymbol{f}_i \cdot \boldsymbol{f}_{u,k}^T}{||\boldsymbol{f}_i||_2 \times ||\boldsymbol{f}_{u,k}||_2}\ , \tag{7}$$

the $\boldsymbol{n}$ ADRs with highest scores are selected as the ones that develop for the given medical cases i.

(2) *Feature-based Similarity*: it is equivalent to FS with having a single global similarity function and the weight assigned to each feature is 1.

$$r_{dot} = \sum_{k=1}^{n_{C_u}} \frac{\boldsymbol{f}_i \cdot \boldsymbol{f}_{u,k}^T}{|\mathcal{C}|}\ , \tag{8}$$

where n_{C_u} is the number of medical cases that have developed ADR_u, $\boldsymbol{f}_i$ is the feature vector of the given medical case, and $\boldsymbol{f}_{u,k}^T$ is the feature vector of $Case_k$ that has developed ADR_u.

Popularity-based Similarity: in this case, we calculate the ADR's popularity across the medical cases. Then we compute the intersection of each, given the medical case's top-$\boldsymbol{n}$ ADR list:

$$r_{pop} = \sum_{i\in\mathcal{C}} \frac{\frac{|M_i \cap N_i|}{|N_i|}}{|\mathcal{C}|}\ , \tag{9}$$

where N_i represents the set of actual ADRs that developed for the given medical case, M_i is $\boldsymbol{n}$ times of N_i be selected from the top-$\boldsymbol{n}$ ADR list predicted for the given medical case.

C. Evaluation Metric

We propose a new metric to evaluate the signaling quality, which is defined as:

$$\texttt{HitRate@}\boldsymbol{n} = \sum_{i=1}^{n_C} \frac{|ADR_i \cap ADR_p|}{|\mathcal{C}|} \tag{10}$$

where ADR_i is the set of actual ADRs that developed for the given medical case, and ADR_p is $\boldsymbol{n}$ times of ADR_i that be selected from this medical case's ranked ADR list. $\texttt{HitRate@}\boldsymbol{n}$ measures how many selected ADRs from the medical case's top $\boldsymbol{n}$ list actually developed. This metric is estimated for each medical case and averaged over all cases.

D. Experimental Settings

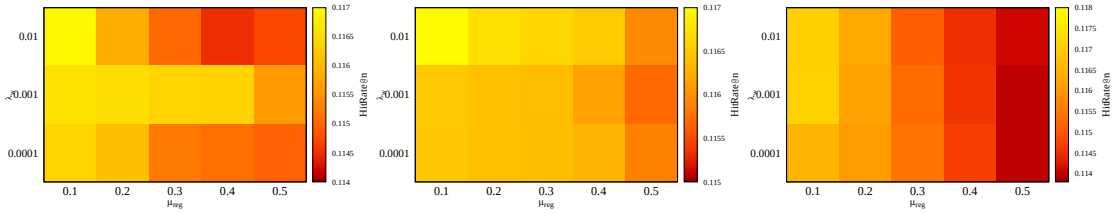
We employed the 5-fold Leave-One-Out-Cross-Validation (LOOCV) to evaluate the performance of the proposed model. The FS model is learned on the training set and estimated over the validation set via the $\texttt{HitRate@}\boldsymbol{n}$ metric. The FS model contains seven parameters, λ_w, λ_u, λ_v, μ_{reg}, μ_{orth}, γ_u, and γ_v. At the step of model selection, we tested all combinations of the following values: $\{0.0001, 0.001, 0.01\}$ for λ_w, $\{0.001, 0.01\}$ for λ_u and λ_v, $\{0.1 \sim 0.5\}$ for μ_{reg} and μ_{orth}, $\{0.01, 0.05, 0.1\}$ for γ_u and γ_v, $\{1 \sim 3\}$ for l. The relation of model's parameter selection and the evaluation performance will be discussed in detail in the next section.

V. Performance Results and Discussion

In this section, we first present the effect of employing different model parameters on signaling quality of FS. Next, we compare the results of our model and competing methods.

A. Effects of Model Parameters

In the FS model, penalization parameter μ_{orth} is responsible for orthogonality among the feature weight vector $\boldsymbol{w}$ with d^{th} membership, and $\mu_{orth} \sum_{\substack{d=1 \\ d'\neq d}}^{l} (\boldsymbol{w}_d \cdot \boldsymbol{w}_{d'}^T)$ does not play a role in the learning process. In this case the penalty parameter μ_{reg} controls the $\boldsymbol{w}_d$'s complexity which plays the key role in the model learning process. Figure 1 shows how the $\texttt{HitRate@}\boldsymbol{n}$ ($\boldsymbol{n} = 2$) varies with λ_w and μ_{reg} associated with the different number of memberships over the FAER dataset. In general, μ_{orth} should be less than μ_{reg}. That is because the number



(a) $l = 1$ (b) $l = 2$ (c) $l = 3$

Fig. 1: Effect of the Learning Rate λ_w, Penalty Parameter μ_{reg} and Number of Membership l for the FAER Dataset

of constraints accompanied with μ_{reg} is equals to l while the number of constraints accompanied with μ_{orth} is on the order of $O(l^2)$.

B. Comparison Results

TABLE I: Comparison on HitRate@$\boldsymbol{n}$ with Evaluating l

HitRate@$\boldsymbol{n}_{pop}$	HitRate@$\boldsymbol{n}_{cos}$	HitRate@$\boldsymbol{n}_{dot}$	l	HitRate@$\boldsymbol{n}_{\text{FS}}$
0.0421128	0.0425058	0.058686	1	0.116384
			2	0.116478
			3	**0.116539**
			4	0.116489
			5	0.116522

We investigate the effect of employing $\boldsymbol{n} = 2$, in estimating HitRate@$\boldsymbol{n}$ on the performance over FS and other competing methods as shown in Table I. For the FS model, we varied the membership $l = [1 \sim 5]$ with an increment of 1 and highlighted the best performance. For the high dimensional FAER's dataset, we set parameters as $\gamma_u = 0.1$, $\gamma_v = 0.1$, $\lambda_u = 0.001$, $\lambda_u = 0.001$, $\mu_{reg} = 0.1$, $\mu_{orth} = 0.1$, and $\lambda_w = 0.0001$.

VI. Conclusion

We presented a new feature-based similarity model FS for signaling adverse drug reactions, and propose a new evaluation metric HitRate@$\boldsymbol{n}$ to evaluate the model. This can help doctors reduce the risk when prescribing and also can assist ADR monitor agents or pharmaceutical industries by providing early warnings for unexpected adverse drug reactions. The key feature of our approach is to use a neighborhood-based method in learning Feature-based similarity models. The FS model learns feature weights with the purpose of stressing the effect of the features that will lead to better accuracy. The model was compared against the Personalized Popularity method, the Simple Feature-based method, and the Cosine Similarity method, the experimental results show that FS outperforms all the other methods in terms of signaling quality.

Acknowledgment

This work was supported in part by NSF (IOS-0820730, IIS-0905220, OCI-1048018, CNS-1162405, and IIS-1247632) and the Digital Technology Center at the University of Minnesota. Access to research and computing facilities was provided by the Digital Technology Center and the Minnesota Supercomputing Institute. Fan Yang was supported in part by China Scholarship Council.

References

[1] I. R. Edwards and J. K. Aronson, "Adverse drug reactions: definitions, diagnosis, and management," *The Lancet*, vol. 356, no. 9237, pp. 1255–1259, 2000.

[2] D. S. Budnitz, D. A. Pollock, K. N. Weidenbach, A. B. Mendelsohn, T. J. Schroeder, and J. L. Annest, "National surveillance of emergency department visits for outpatient adverse drug events," *Jama*, vol. 296, no. 15, pp. 1858–1866, 2006.

[3] J. Lazarou, B. H. Pomeranz, and P. N. Corey, "Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies," *Jama*, vol. 279, no. 15, pp. 1200–1205, 1998.

[4] M. Liu, M. E. Matheny, Y. Hu, and H. Xu, "Data mining methodologies for pharmacovigilance," *ACM SIGKDD Explorations Newsletter*, vol. 14, no. 1, pp. 35–42, 2012.

[5] M. Suling and I. Pigeot, "Signal detection and monitoring based on longitudinal healthcare data," *Pharmaceutics*, vol. 4, no. 4, pp. 607–640, 2012.

[6] P. Thomas, M. Neves, I. Solt, D. Tikk, and U. Leser, "Relation extraction for drug-drug interactions using ensemble learning," *1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011)*, pp. 11–18, 2011.

[7] L. Jacob and J.-P. Vert, "Protein-ligand interaction prediction: an improved chemogenomics approach," *Bioinformatics*, vol. 24, no. 19, pp. 2149–2156, 2008.

[8] J. Lin, Q. Kuang, Y. Li, Y. Zhang, J. Sun, Z. Ding, and M. Li, "Prediction of adverse drug reactions by a network based external link prediction method," *Analytical Methods*, vol. 5, no. 21, pp. 6120–6127, 2013.

[9] O. Caster, G. N. Norén, D. Madigan, and A. Bate, "Large-scale regression-based pattern discovery: The example of screening the who global drug safety database," *Statistical analysis and data mining*, vol. 3, no. 4, pp. 197–208, 2010.

[10] R. Harpaz, H. S. Chase, and C. Friedman, "Mining multi-item drug adverse effect associations in spontaneous reporting systems," *BMC bioinformatics*, vol. 11, no. Suppl 9, p. S7, 2010.

[11] A. Lee, *Adverse drug reactions*. Pharmaceutical press, 2006.

[12] A. Elbadrawy and G. Karypis, "Feature-based similarity models for top-n recommendation of new items," Department of Computer Science, University of Minnesota, Minneapolis, Minnesota, Tech. Rep. 14-016, June 2013. [Online]. Available: http://www.cs.umn.edu/research/technical_reports/view/14-016

[13] R. A. Horn, "The hadamard product," in *Matrix theory and applications, Series: Proc. of Symposia in applied mathematics*, vol. 40, 1990, pp. 87–169.

[14] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "Bpr: Bayesian personalized ranking from implicit feedback," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2009, pp. 452–461.

[15] D. Billsus and M. J. Pazzani, *A hybrid user model for news story classification*. Springer, 1999.