Co-Training and Ensemble Based Duplicate Detection in Adverse Drug Event Reporting Systems

Wen-Yang Lin and Chiao-Feng Lo
Dept. of Computer Science and Information Engineering
National University of Kaohsiung, Kaohsiung, Taiwan
wylin@nuk.edu.tw; m1005507@mail.nuk.edu.tw

***Abstract*—Nowadays, many countries have established spontaneous reporting systems (SRSs) to facilitate postmarketing surveillance of listed drugs and collect enough data for detecting unknown adverse drug reactions. Due to data in SRSs coming from different sources of reporters, there heralds the problem of duplicate reporting; even a small amount of duplicate records would bias the detection results. Although lots of works have been conducted on duplicate record detection, very few of them have been devoted to dataset about adverse drug reactions, and none of them have considered the existence of follow-up reports. In this study, we investigated the problem of identifying duplicate ADR reports in SRSs with the presence of follow-ups. We proposed an ensemble and co-training based detection method that is capable of detecting for a given report not only its duplicates but also its initial or earlier linkage cases.**

Keywords- Adverse drug events; duplicate records detection; ensemble learning; co-training learning; follow-up

I. INTRODUCTION

Adverse drug reactions (ADRs) refer to harmful reactions associated with the normal usage of a given medicine. Due to the limited number of patients that can participate in clinical trials, it is impossible to discover all potential ADRs. Therefore, many countries or world organizations have established various spontaneous reporting systems (SRSs) for adverse drug events (ADEs), acting as a repository researches or drug-related agencies to analyze and to identify potential adverse reactions, e.g., UK Yellow Card, Food and Drug Administration (FDA), and Health Canada (HC).

Although different SRSs follow different reporting regulations for cases reporting, most of them asset voluntary activity from different source reporters (e.g., health care practitioners, patients), which arises the problem of duplicate reporting. For example, in an ADR evaluation of quinine-induced thrombocytopenia conducted by FDA researchers in 2002 [5], 20% of 141 reports in the FAERS database were identified as duplicates.

The presence of duplicate records may bias the results of ADR signal detection, either causing false signals or overlooking important signals. Contemporary measurement used for ADR signals considers the number of occurrences; a specific ADR-drug pair supported even by three cases might be reported as a suspicious ADR signal [6]. Therefore, a small amount of duplicate records would have a significant impact on the detection results.

Although the problem of duplicate detection has been well studied in the database and statistics communities [3], little work has been devoted to the detection of duplicate ADR reports. Additionally, previous work neglected the presence of follow-up reports [6], which complement the information of initial report and so have to be merged with the initial report to form a more accurate and complete version. Wrongly identifying a follow-up as a duplicate or failure to link it to earlier cases obviously will deteriorate the results of detected ADR signals.

In this study, we investigated the problem of identifying duplicate reports with the presence of follow-ups, using the FAERS database [4] as a case study. We proposed an ensemble and co-training based detection method that is capable of detecting for a given report not only its duplicates but also its initial or earlier linkage cases. Experiments conducted using example FAERS datasets show that our algorithm outperforms contemporary classification methods, having better accuracy and F-measure.

II. DUPLICATE REPORTING PROBLEM IN FAERS DATASET

FAERS [4] is a database designed to assist the FDA to monitor the safety of drugs after the listing. The problem of duplicate detection in the FAERS database is complicated by the existence of follow-up reports, which contains update information, such as information modification, medication changes, and adverse reactions changes. Action taken for dealing with duplicate reports is different from that for initial / follow-up or follow-up / follow-up cases. If two records are identified as duplicate, only one record should be retained, while a follow-up should be merged with its initial report or preceding follow-up to form a more accurate report. Confusing these two situations will bias the case occurrences and result in incorrect ADR signals. As such, we view the problem of duplicate detection in the FAERS database as a tri-class classification problem. That is, given any two ADE records, we would like to classify this record pair into the following three scenarios (classes):

1) Real duplicate (label is II): This denotes that two records are true duplicate of each other. In the FAERS data format, this case corresponds to records with identical CASE no., different ISR no., and both code “I” in the I_F_COD field.

2) Follow-up linkage (label is IF&FF): This denotes the relationship between the record pair is initial / follow-up or follow-up / follow-up. According to the FAERS coding format, the former corresponds to record of the same CASE no., but different ISR no., and having diffierent I_F_CODs, one with "I" and another with "F".

3) Others (label is OTHER): This denotes all cases other than real duplicate and follow-up linkage. Intuitively, real duplicate and follow-up linkage are rare situations. Most of the record pairs belong to "OTHER" category.

Following the above class specification, we transform the selected FAERS dataset into labeled record pairs, each of which is represented by a similarity vector:

$$l \quad \mathrm{sim}(f_{i1}, f'_{j1}) \quad \mathrm{sim}(f_{i2}, f'_{j2}) \quad \ldots \quad \mathrm{sim}(f_{ih}, f'_{jh})$$

where l denotes the label, f_{ik} and f'_{jk} the values of attribute A_k of two records r_i and r_j, respectively, and sim(·) is the similarity function.

III. Co-training and Ensemble based Method

Our algorithm is a hybrid of ensemble [7] and co-training learning [1] strategies. The idea is motivated by the fact that we have to build learning models from the data set with very small amounts of labeled instances and large amounts of unlabeled data. We adopt multiple, diverse base classifiers, as a whole to determine the class of a given unknown record pair. Meanwhile, our algorithm utilize unlabeled record pairs, choosing those receiving high commitment among the initial base classifiers for being classified as II or IF&FF into the training set, and retain the initial models using the updated training set.

Our algorithm proceeds in three stages. First, the training set with assigned labels is processed to build the initial ensemble classifier. Next, some of the unlabeled data, which refer to all record pairs not identified as II or IF&FF and having high similarity over the threshold are added into the original training set. The choice follows the principle of majority teaching minority. That is, a record pair from the unlabeled set is chosen if classified as IF or IF&FF by over half of the base classifiers. We consider only II or IF&FF class because these two classes are relatively small compared with the OTHER class. Each chosen unlabeled record pair is given a pseudo label of II or IF&FF, depending on the predicting result. Finally, the updated training set is used to build the final ensemble classifier.

IV. Expermental Results

A recent work in [2] has provided a widespread comparison of contemporary classification methods. Accordingly, we selected four representatives from different categories of classifiers, including Bayesian classifier (Bayesian Network), instance-based classifier (IBk), rule-based classifier (JRip), and decision tree (J48), all of which are available on the Weka package. We also implemented an ensemble of these four classifiers, serving as an additional comparator to our algorithm.

We chose seven quarters of datasets from FAERS, including 05Q4, 06Q3, 07Q2, 08Q1, 09Q4, 10Q3, and 11Q2. Two commonly used measures, accuracy and F-measure, were adopted in this experiment. In average, our method exhibits 93.17% of accuracy and 0.94 of F-measure, which are better than all other competitive methods (Bayesian Network, IBK, JRip, J48, Ensemble) by 1% to 16% (accuracy) and 1% to 14% (F-measure).

We conducted an additional experiment to inspect how our algorithm can identify duplicates that were known in the literature. Specifically, we referenced the work conducted by Hauben et al. [5], which showed the existence of extreme duplication in the FAERS database; 20 reports about aortic dissection associated with a drug of interest referred to the same cases. Our algorithm predicted that all pairs belong to IF&FF, implying that all these cases were follow-up reports that refer to the same patient. This result does not totally conform to the result in [5]. The best reason for explaining this phenomenon is that Hauben et al. did not differentiate follow-up linkage from real duplicate while our algorithm did.

V. Conclusions

We have presented the problem of duplicate detection in the ADR reporting systems, highlighting the effect of follow-up, and conduct a study on the FAERS database. We formalized the problem as a tri-class classification problem, and proposed an ensemble with co-training based method to accomplish the task, identifying a record pair as real duplicate, follow-up linkage or others. We conducted experiments on several quarterly reported datasets from the FAERS database to compare our method with some representative classifiers. The experimental results showed that our method exhibits better accuracy and F-measure than other methods. Although further investigation is needed, this preliminary result has shown the feasibility of our proposed method.

Acknowledgment

This work is supported by the National Science Council of Taiwan under grant No. NSC100-2221-E-390-024.

References

[1] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," Proc. Conference on Computational Learning Theory, pp. 92–100, 1998.

[2] L. Cagliero and P. Garza, "Improving classification models with taxonomy information," Data and Knowledge Engineering, vol. 86, no. 1, pp. 85–101, 2013.

[3] A.K. Elmagarmid, P.G. Ipeirotis, and V.S. Verykios, "Duplicate record detection: A survey," IEEE Trans. Knowledge and Data Engineering, vol. 19, no. 1, pp. 1–16, 2007.

[4] FDA Adverse Event Reporting System, Available: http://www.fda.gov/cder/aers/default.htm,

[5] M. Hauben, L. Reich, J.D. Micco, and K. Kim, "'Extreme duplication' in the US FDA adverse events reporting system database," Drug Safety, vol. 30, no. 6, pp. 551–554, 2007.

[6] G.N. Norén, R. Orre, A. Bate, and I.R. Edwards, "Duplicate detection in adverse drug reaction surveillance," Data Mining and Knowledge Discovery, vol. 14, no. 3, pp. 305–328, 2007.

[7] L. Rokach, "Ensemble-based classifiers," Artificial Intelligence Review, vol. 33, no. 1-2, pp. 1–39, 2010.