Reducing Free-Text Communication Orders Placed by Providers Using Association Rule Mining

Zahra Hajihashemi, Master[1], Paul Pancoast, MD, MBA[2]
[1]University of Missouri, Computer Science Department, Columbia, MO; [2]Health Quest System, Poughkeepsie, NY

Abstract

Electronic health record (EHR) systems are used to collect, store and retrieve the details of patient care. Computer Provider Order Entry (CPOE) is a process by which providers directly enter patient care orders into the EHR. Providers may enter free-text orders when they are unable to find standard orders. These free-text orders require translation into a structured order which reducing efficiency, may bypass duplicate checking and can be associated with medical errors. To overcome these problems we developed a system to automatically detect free-text orders and assign them to the appropriate order categories. This system applies association rule mining on structured orders to extract the patterns of orders in the related categories. The extracted patterns were tested on a set of free-text orders for evaluation and to determine the closest matching category of structured orders. This process may be used to improve future iterations of CPOE applications.

Introduction

CPOE is widely considered as an essential informatics solution to reduce medical and prescribing errors and which may also result in increased system efficiencies[6,2]. Clinicians, hospital administrators, pharmacists, as well as business councils, researchers, the institute of medicine, health care agencies, and state legislatures are proponents of CPOE[7]. The importance of CPOE in health care is due to its significant reductions in medication errors, reductions in translation or transcription errors (from handwritten orders), and reduction in through-put time to complete orders[,17]. However, the efficiency of CPOE from the provider viewpoint is highly dependent on the provider's knowledge and familiarity with the application for selection of structured orders. Projects including implementation of CPOE applications result in major changes to provider workflow and impact multiple stakeholders in the healthcare system, including providers, nurses, and ancillary departments. These are often known as 'clinical transformation projects' because they involve redesign of complex clinical process and integrate technology at key points to improve and optimize the ordering process.

Physicians and other health care providers at Health Quest Systems are currently using CPOE to enter orders and requests for patient care, including orders for laboratory testing, diagnostic imaging, medication administration, consult to other providers and health professionals, and direct patient care orders. The providers using CPOE select specific (structured) orders from a defined order catalog which is categorized by order type and specialty. When providers are unable to find the exact order they want to place, they may place a free-text order. In addition to requiring another clinician (usually a nurse or pharmacist) to review the free-text orders to discover exactly what the provider wants, these free-text orders bypass the clinical decision support safety checks and may be a source of inefficiency, duplicate testing and therapies. They may even contribute to inappropriate therapy resulting in avoidable medical errors.

Natural Language Processing (NLP) is an important tool for health care providers and decision makers. The wide variety of useful techniques and methods to overcome different and complex problems in medical documents coming from day by day activities of health care organizations makes NLP an essential and competitive solution[9,10]. Text mining techniques, as part of NLP solutions, can process free and unstructured text to extract meaningful knowledge automatically. Among different methods, Association Rule Mining is widely used by researchers to drive relations between words and semantics of the medical text[8,11,12]. We determined that NLP techniques could be used to help solve the issues with free-text medical orders.

Applying association rule mining on structured electronic health record, associations between medications, laboratory results and problems has been determined. These associations have been used to identify possible gaps in patient problem list, an important component of clinical medicine[18]. Temporal association mining is suitable for finding "interesting patterns" in the temporal data. This technique can be efficiently applied to different health care problems such as biomedical and clinical problems, patient monitoring, health care administrative data management

and molecular biology. Consequently, those interesting patterns will be mined for the sake of temporal association rules[1,13].

The ability of association rule mining to extract meaningful combination of terms related to a similar subject makes it a useful solution for mining free-text documents in medical domain. The results of applying association rule mining in extracting patterns of Adverse Drug Reactions (ADR) from user comments and/or medical reports makes it very interesting as well as useful. Extracted rules are applied to different numbers of conditions related to user comments, laboratory results, disease conditions, and drug administration for ADR detection. Using these derived rules where some conditions trigger a conclusion, the knowledge about ADRs can be explained[4,5]. The results from applying these rules to clinical data can demonstrate the underlying patterns associated with ADRs. After review of the results from these researches, we determined that association rule mining could be helpful to extract the patterns of CPOE orders.

The patterns of the similar structured orders can be found by using association rule mining in CPOE orders. In the data set provided by Health Quest, the majority of orders placed are selected from specific order catalogs; the details of the orders are included in the order sentences. When a provider cannot find the appropriate order, they may select a "communication order" or "miscellaneous nursing task" order, which allows free-text entry of the order that they wish to place without the constraints of the codified order catalog entries. We propose to use association rule classification to review communications orders and miscellaneous nursing task orders and place them into logical grouping. The results from these automated logical groupings will help us identify patterns of provider use of communication orders for specific tasks, creation of new orders or modification of existing orders in the future.

Method

There are two main research approaches to obtain association patterns: knowledge-based and corpus-based approaches. In knowledge based approach expert knowledge is used to design patterns. A main drawback of this approach is significant time and effort requirement to design handcrafted patterns. Corpus-based approach, however, uses supervised learning techniques to extract association patterns among a collection of data. To label this data set, the domain specific knowledge is needed. A wide variety of statistical methods can then be applied to discover association patterns from possible combination of words in the text corpora. Association rule mining can be used for this purpose[19].

Basically, the idea of association rule mining derived from the "shopping cart" problem that try to find which set of items-called frequent item sets- are more likely to be bought together. Consequently, to control costumers traverse in the supermarket, managers will use this information to place frequent item sets in the shelves. Likewise, the challenge of finding patterns in text can be modeled as association rule mining problem in which the sentences in the text are transactions and the words in the sentences are considered as items in the transactions. In this case, frequent item sets would be as set of terms (or words) which are more likely to appear in a sentence together, called co-occurred words.

Among different algorithms that can be used to derive frequent item sets, FP-growth (frequent pattern growth) uses an extended prefix-tree (FP-tree) structure to store the database in a compressed form[14]. FP-growth adopts a divide-and-conquer approach to discover frequent item sets without generating all possible candidates. It uses a pattern fragment growth method to avoid the costly process of candidate generation and testing used by Apriori[22]. For this purpose, first a compact data structured called FP-tree is build using two passes over the data set. Then, frequent item sets will be directly extracted from the FP-tree by traversing through that. These generated frequent items sets will be used to derive rules.

After extracting frequent item sets for each specific category, an association rule of the form X => c can be constructed, where X is a frequent item set containing combination of different terms and c is a pre-assigned category. For each rule, the number of transaction that include all the terms in X among the text labeled as c called *support* of the rule. Another measure which is a probability representing how frequently the rule is occurred among all the categories containing the rule body is called the *confidence* of the rule. The higher the value, the more often this set of items is associated together. The goal is to find out association rules with high values of support, confidence.

Basically, a threshold for support and confidence is set to cut down thousands of generated rules to the more significant ones. However, in most cases not all of them would be significant enough to accurately classify data. More attentions need to be considered in this regard. In the literature, different types of rule selection techniques have been done. In a research, authors defined a new characteristic for each rule which determines how significant a

rule can determine a class. Accordingly, a threshold based on this definition need to be set[21]. This threshold is used to cut down insignificant rules, called noisy rule. Moreover, another measurement for the significant rules is the number of times they are hit. Even though the method is well designed in terms of the time complexity and accuracy, the performance highly depends on the threshold. In most real case problem, setting such a threshold is not easy and needs lots of experiments. On the other hand, the second measure, number of rule hits, might have some drawback in terms of failing to find rules which can be useful to discover rare events.

For ranking rules, during one pass of algorithm rules are ranked if they pass the minimum support and minimum confidence thresholds according to a series of parameters, including confidence, support, antecedent cardinality, class distribution frequency, item row order and rule antecedent length[21]. The global sorting method on the rules is effective for rule selection and building the classifier. The proposed algorithm is shown to be highly competitive when compared with the C4.5,CBA,CMAR and CPAR algorithms in terms of classification accuracy. However, parameters such as rule antecedent length cannot guarantee the selection of all possible significant rules.

In a study [20], based on the set theory, significant rules have been selected. Let X1 represent a set of items in rule R1 and X2 represent a set of items in rule R2. Following, three different types of relationship between R1 and R2 are explained:

1. R1 is equal to R2 if all items in X1 are occurred in X2 and vice versa.
2. R1 is super set of R2 if X1 include all the items in X2 plus at least one item which is not appeared in X2.
3. R1 is subset of R2 if all items in X1 are occurred in X2 and at least one item in X2 does not exist in X1.

Rules that are a super-set of the others are selected while the rules which are subset are discarded. The significant rules are then selected based on the confidence and lift. In some complex data set, such as medial documents, this technique is not accurate enough.

In this study, we applied the rule selection techniques which is based on the methods described by Marukatat[20] and a new scoring method to compensate its drawback. However, we used a weighted scoring function which is a combination of support and confidence and entropy of terms occurred in the rule. In information theory, entropy is a measure of the uncertainty associated with a random variable. The lower the entropy of a term represents the more certainty about that term that leads to better classification task. Even though the entropy has been used before for classification, we applied this measure here in terms of rule scoring and selection[15]. This weighted scoring use partial match to get more insight of the nature of the free-text data, since a significant part of the orders in CPOE are free text entries.

After matching rules and scoring, the next step is to make a decision about the category which can be assigned to the test data. Otsu's method has been used before for finding threshold in image processing[16]. Otsu's thresholding method involves iterating through all the possible threshold values and calculating a measure of spread for the pixel levels each side of the threshold, i.e. the pixels that either falls in foreground or background. The aim is to find the threshold value where the sum of foreground and background spreads is at its minimum. We used this technique for finding the best threshold in this step. Moreover, this method is applied for each category separately to find the threshold score for that category. Figure 1 illustrates the general architecture of our system.

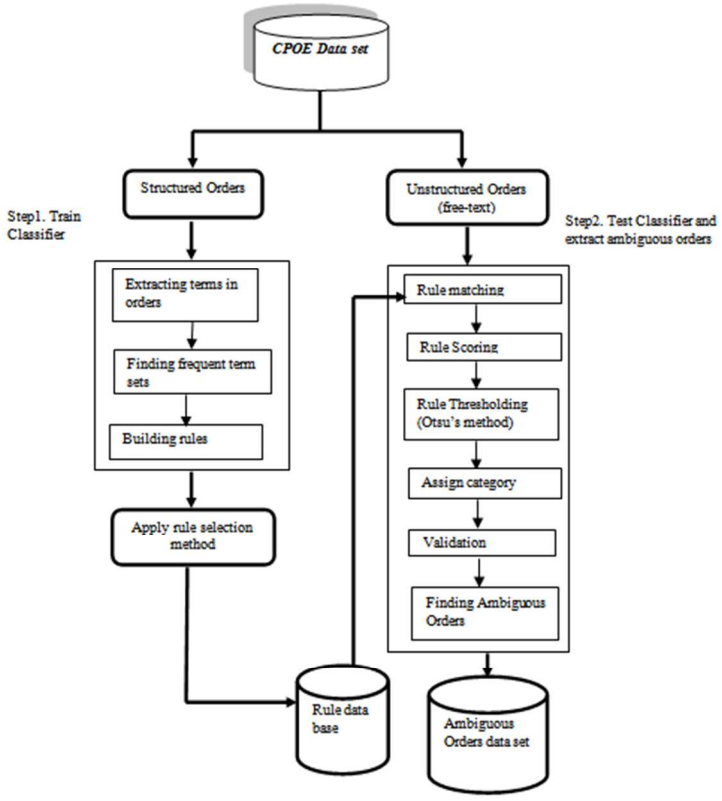

Figure 1. Overall general architecture of our system

Results

Providers at Health Quest hospitals enter approximately 12000-15000 orders every day, and about 5% or 750 orders per day are classified as communication order or miscellaneous nursing task orders. Of these, a significant number are pre-built within Cerner as communication orders, such as "D/C Foley in AM Post-op day 1", " Discontinue IV when patient stable" and "if POC blood glucose level is >400 or <40 mg/dl, collect a STAT blood glucose level". These are standard orders that give directions to perform actions at specific time intervals or if when predefined parameters are met. Other communication orders are free-text created by providers. Some communication orders are necessary because there is no appropriate standard order to select. In other cases, these free-text orders are created because the provider doesn't want to navigate the system to select the appropriate order because of time constraints or unfamiliarity with the system, or because the system is too cumbersome to use.

Orders regardless of structured or unstructured can be divided into several different categories. Our goal is to develop a system that can parse a free-text order and automatically assign it to the most likely category. Table 1 shows the numbers of free-text orders that were manually parsed and assigned to their appropriate categories by one of the investigators.

Table 1. Statistics of structured data.

Order Category	Number of Orders
Activity	158
Laboratory	334
Diagnostic Imaging	264
Medications	698
Diet	93
Urinary Catheter	45
NG Tube	30

Using association rule mining, frequent term sets for each category are extracted and the related rules are constructed and inserted into the rule data base. In the training step, minimum support and confidence were set as 0.05 and 60% respectively. The numbers of generated rules depend significantly on the size of the training data and the defined threshold of support and confidence of the rules. For testing association rule classifier, unlabeled free-text orders were used. Table 2 demonstrates the result of the classifier for free-text categorization on the test set of the size 270 orders. Correct and incorrect classification of our rules-based classifier were determined by manual review of the results by one of the investigators.

Table 2. Association rule classifier results.

Order Category	Percentage
Correct Classification	40
Incomplete/multiple orders	10
Ambiguous orders	32
Incorrect Classification	18

Discussion

We collected a large dataset of free-text orders for our test dataset. The performance of the association rule classifier presented here is dependent on the training data and the portion of the test data that includes structured orders which matched the generated rules. Having a large corpus of data for train and test increases the performance significantly. Moreover, some issues of test data set related to the behavior of health care providers who place orders have effects on the classifier accuracy. Difficulties with the dataset include large numbers of non-standard abbreviations used inconsistently by providers as well as frequent miss-spellings. Putting multiple orders in a single sentence without using punctuation to separate the orders also contributes to inaccuracies. Classification errors are caused when providers use 'cath' to refer to a urinary catheter as well as a directive to obtain a urine specimen by mechanical means and other similar ambiguities. Other errors are made by inconsistencies in the way measurements and units are recorded in a free-text order. Finally, orders such as 'ambulate in hall remove cath when amb w/o diff' are parsable by a knowledgeable clinician but require more complex rules than the ones we built in our initial system.
A pre-processing step could be helpful in overcoming some of these problems, but we would need a comprehensive dictionary of the related medical words in the specific domains, and some method of classifying and identifying the non-standard abbreviations which are used. Our association rule classifier is built based on structured order sentences from the order catalog with only one order sentence per order; free-text orders are often written such that a single sentence contains multiple order requests. These kinds of orders lead to ambiguity in the data and reduce the accuracy of the system. Many of these problems may be solved using additional NLP techniques, and will be incorporated in future research.

A side benefit we will realize is the identification of free-text orders that do not match structured orders in our catalog. This will direct our efforts in developing new structured orders for inclusion in the catalog, and should reduce the need for physicians to free-text orders which recoups the losses in efficiency and regains the inherent safety of clinical decision support during the ordering process. Ultimately our goal is to increase patient safety by reducing the potential for medical errors and improve the efficiency of providers delivering care.

This study can show how comprehensive the current structured orders are compare to what the health care providers really need. As it is shown in the table 2, the considerable percentage of ambiguous orders demonstrate the necessity of developing and expanding structured orders based on the user needs. Having a developed structured orders system will reduce the number of free-text orders and increase the efficiency of health care organizations.

Conclusion

CPOE is recognized to improve efficiency in processing orders placed by providers and allows clinical decision support rules to help guide the ordering process. When health care providers use a CPOE system to enter free-text orders they bypass the inherent safety mechanisms of clinical decision support and increase the workload on other caregivers by requiring manual interventions to process the orders. Using a system to automatically classify free-text orders into logical groupings may reduce the loss of efficiencies. Reviewing free-text orders that do not logically match existing structured order sentences can be used to improve and expand the current order catalog and pre-defined order sentences. Our experiments demonstrate the potential for accomplishing both of these goals. Using

the proposed association rule classifier on free-text orders, the ambiguous orders can be selected and will be used in the future to extract the pattern of the new structured data.

References

1. Zhou XH, Li SL, Tian F, Cai BJ, Xie YM, Pei Y, Kang S, Fan M, Li JP. Building a disease risk model of osteoporosis based on traditional Chinese medicine symptoms and western medicine risk factors. Statistics in Medicine 2012; 31; 643–652.(sim4382)
2. Stone WM, Smith BE, Shaft JD, Nelson RD, Money SR. Impact of a Computerized Physician Order-Entry System. J Am Coll Surg. 2009;208;5;960-967. (CPOEstone0509)
3. Koppel R, Metlay JP, Cohen A, Abaluck B, Localio AR, Kimmel SE, Strom BL. Role of Computerized Physician Order Entry Systems in Facilitating Medication Errors. JAMA. 2005;293;10;1197-1203. (JAMIA CPOE 1197)
4. Chazard E, Ficheur G, Bernonville S, Luyckx M, Beuscart R. Data Mining to Generate Adverse Drug Events Detection Rules.IEEE Transaction on Information Technology in Biomedicine. 2011;15;6;823-830. (05995169)
5. Nikfarjam A, Gonzalez GH. Pattern Mining for Extraction of mentions of Adverse Drug Reactions from User Comments. AMIA Annu Symp Proc. 2011; 1019-1026. (1019_amia_2011_proc)
6. Ash JS, Gorman PN, Seshadri V, HErsh WR. Perspectives on CPOE and patient care. J Am Med Inform Assoc. 2004; 11; 207-216. (17)
7. Ash JS, Berg M, Coiera E. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. J Am Med Inform Assoc. 2004;11:104-112. (22)
8. Lesar TS, Lomaestro BM, Pohl H. Medication prescribing errors in a teaching hospital: a 9-year experience. Arch Intern Med. 1997;157:1569-1576. (1)
9. Kohn LT, Corrigan J, Donaldson MS, eds. To Err Is Human: Building a Safer Health System. Washington, DC: National Academy Press; 2000.(2)
10. Eslami S, Abu-Hanna A, de Keizer NF. Evaluation of outpatient computerized physician medication order entry systems; a systemic review. J Am Med Inform Assoc 2007;14:400–406. (3)
11. Kanjanarat P, Winterstein AG, Johns TE. Nature of preventable adverse drug events in hospitals: a literature review. Am J Health Syst Pharm. 2003;60:1750-1759. (4)
12. Kaushal R, Jha AK, Franz C. Return on investment for a computerized order entry system. J Am Med Inform 2006;13:261–266. (5)
13. Bellazzi R, Sacchi L, Concaro S. Methods and Tools for Mining Multivariate Temporal Data in Clinical and Biomedical Applications. 31st Annual International Conference of the IEEE EMBS Minneapolis; 2009;5629-5632.
14. Jiawei H, Jian P, Yiwen Y, Runying M. Mining frequent patterns without candidate generation. Data Mining and Knowledge Discovery.2004; 8;53-87.[18 fp-tree]
15. Witten IH, Bray Z, Mahoui, M, Teahan B. Text mining: a new frontier for lossless compression. Data Compression Conference, 1999. Proceedings. DCC '99;1999;198-207.
16. Liao PS, Tse-Sheng Chen TS, Chung PC.A Fast Algorithm for Multilevel Thresholding. J. Inf. Sci. Eng; 2001; 17;5;713-727.
17. Kuperman GJ, Gibson RF. Computer Physician Order Entry: Benefits, Costs, and Issues. Ann Intern Med. 2003;139;1;31-39.
18. Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. Journal of biomedical infromatics;2010;43;6;891-901.
19. Zhang S, Zhou Q. A Novel Efficient Classification Algorithm Based on Class Association Rules. Applied Mechanics and Materials. 2012;135-136;106-110.
20. Marukatat R. Structure-Based Rule Selection Framework for Association Rule Mining of Traffic Accident Data. IEEE Computational Intelligence and Security;2006;781-784.
21. Yu LC, Chan CL, Lin CC, Lin IC. Mining association language patterns using a distributional semantic model for negative life event classification. J Biomed Inform. 2011;44;4;509-518.
22. Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, Santiago, Chile. 1994; 487-499.