Proceedings of the 2007 IEEE Symposium on Computational
Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Analysis and Prevention of Dispension Errors by Using Data Mining Techniques

Vincent S. Tseng[1, 3], Chun-Hao Chen[1], Hsiao-Ming Chen[1], Hui-Jen Chang[2], and Chin-Tai Yu [3]
[1]*Department of Computer Science and Information Engineering, National Cheng-Kung University*
{vincent, chchen, nick}@idb.csie.ncku.edu.tw
2 Department of Pharmacy, National Cheng Kung University Hospital
hjchang@mail.ncku.edu.tw
3 Department of Medical Information, National Cheng-Kung University Hospital
yujt@mail.ncku.edu.tw

***Abstract*—Medical treatment techniques have been improved continuously in the past years. However, the better approaches are still needed to solve medical treatment problems. One important topic in this field is the analysis and prevention of medication errors. In this paper, we focus on the problem of dispensing error that is one important problem of medication errors and we proposed a prevention model by using three approaches. The proposed dispensing error mining framework consists of two phases, namely the modeling and prediction phases. Firstly, Statistical approach (logistic regression) and data mining approaches (C4.5 and SVM) are used to analyze dispensing error problem and to build classification models. Three kinds of factors, namely drug-names factor, drug-properties factor and environmental factor, with totally thirteen attributes are used in the modeling phase. In prediction phase, new drugs thus can be analyzed for the probability of dispensing error by the model so as to prevent dispensing error. At last, experimental results on real dataset showed that the proposed approach is effective and the considered factors can actually increase the accuracy of the model.**

I. INTRODUCTION

Medical treatment techniques have been improved continuously in the past years. However, the better approaches are still needed to solve medical treatment problems. One topic in this field is medication errors. For example, in general, when people sick, they go to see the doctors and take medicine according to the prescription. However, before a patient get the drugs, it is needed some processes. Those processes can not be done automatically such that some mistakes might be arisen. The problem is called medication error problem. In 1998, Lazarou *et al.* point out that there are about 1,300,000 peoples suffered medication error in America of each year [14]. In 2000, Kohn *et al.* also point out that there are about 44,000 to 98,000 peoples died for medication error [5]. The medication error is thus became an important topic for researches.

In 2001, Cavell *et al.* said that medication error can be found in four possible processes, including prescribing, transcription, dispensing, and administration [4]. In this paper, we focus on the third part, dispensing, and try to decrease the dispensing error by using different techniques. In previous works, the main criterion which is used to prevent dispensing error is the similarity (distance) values of drug names [14, 11, 23]. For example, two drugs, *Ephedrine* and *Epinephrine*, are recognized as high similarity because of its high orthographic similarity. In this case, it means that the probability of dispensing error is high. However, we can know that the similarity (distance) of the drug names is one of the reasons of dispensing error [12, 15]. In this paper, we thus use not only similarity and distance of drug names (drug-names factor) as criterions like previous works, but also take drug-properties factor and environmental factor of drugs into consideration (e.g. the color or shape of drugs and location of drugs, etc.) to analyze the problem.

Statistical approaches have always been used for analyzing problems in different fields, including dispensing error [13]. Besides, data mining techniques are also applied to many topics in recent years. Decision makers can retrieve useful information by using data mining techniques to make appropriate strategies. The well known data mining techniques, including association rules [1], clustering [18], and classification [20, 21]. In medication error, Rudman *et al.* applied data mining tools to recognize and analyze near miss and adverse drug reaction [23].

In this paper, we focus on the problem of dispensing error that is one important problem of medication errors and we proposed a prevention model by using three approaches. The proposed dispensing error mining framework consists of two phases, namely the modeling and prediction phases. In modeling phase, statistical approach (logistic regression) and data mining approaches (C4.5 and SVM) are used to analyze dispensing error problem and to build classification models with real data. In prediction phase, new drugs can be analyzed to avoid dispensing error according to the model and feedback mechanism is also set up for continuous improvement of the models. Three contributions of this paper are stated as follows.

1. We take three factors, namely drug-names factor, drug-properties factor and environmental factor, of drugs into consideration in building prevention model.
2. We proposed a dispensing error mining framework to

prevent dispensing error and three approaches were used to analyze them.

3. The derived rules can be used for dispensing error prevention and be referenced by medical experts.

The remaining parts of this paper are organized as follows. The problem definition is described in Section 2. The related works is stated in Section 3. The proposed framework is described in Section 4. Experiments to demonstrate the performance of the proposed algorithm are described in Section 5. Conclusions and future works are given in Section 6.

II. Problem Definition

The problem we target in this paper is as shown in Fig. 1. Each pair in the left of Fig. 1. means a dispensing error case. For example, the pair (*A*, *B*) represents drug *A* which is prescribed by doctor, but drug *B* is given to patient. The selected attributes are shown in the top of Fig. 1. They are divided into three factors, drug-names factor, drug-properties factor and environmental factor. Drug-names factor contains attributes that can be derived form drug names. For example, orthographic similarity of drug names is one of drug-names factor. Drug-properties factor means attributes that can be presented the drugs. For example, color of drugs is one of drug-properties factor. Environmental factor means the location of drugs in administration.

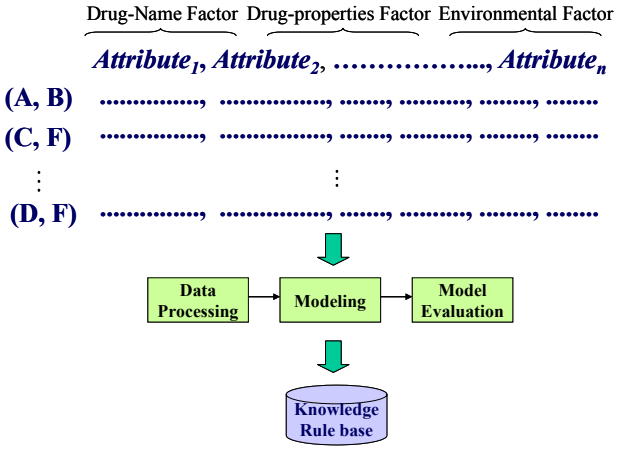

Fig. 1. Problem definition

Form Fig. 1, the problem that we want to solve is using dispensing error cases and selected attributes to build a prevention model and decrease dispensing error rate. The processes include appropriate pre-processing, modeling and evaluation.

III. Related Works

In this section, we first define what is dispensing error. The reasons of why dispensing error occurs are then discussed. Thirdly, different approaches whose are used to evaluate similarity of drug names are introduced. At last, three classification methods, logistic regression, C4.5 and support vector machine, are described.

A. Dispensing Error

In general, dispensing error means there have some mistakes during dispensing drugs process in administration. There thus have many definitions of the word "dispensing error" in different point of views [2, 6, 19]. In 1999, Peterson *et al.* defined that if the wrong drugs can not be found by dispensers in administration, then it is called dispensing error. Otherwise, it is called near miss [19]. In 2003, Chua *et al.* defined that dispensing error should be jugged by patient gets the medicines or not [6]. However, Ashcroft *et al.* defined that dispensing error should be jugged by patient already takes medicines or not [2]. In this paper, we adapted the Peterson's definition as dispensing error.

B. Reasons of Why Dispensing Error Happen

In fact, there exist many factors that can cause dispensing error. Many researches think that the main reason of dispensing error is the drug names and package similarity [11, 14, 23]. Long *et al.* think dispensing errors are made by employees who violate the strand workflow. Kistner *et al.* think the error is due to the heavy workload. According to the reasons discuss above, the selected attributes in this paper thus contain three factors, drug-names factor, drug-properties factor and environmental factor.

C. Similarity Evaluation Approaches

The dispensing error is caused because of confusing of phonetic or orthographic similarity of drug names. Phonetic similarity of drugs is measured by comparing phonetic codes of drug names. Orthographic similarity of drugs is evaluated by the required effort when transferring two drugs name into identical one. Two approaches, namely similarity-based approach and distance-based approach, are used to evaluate phonetic or orthographic similarity of drug names and different algorithms whose have been proposed are listed in Table I.

TABLE I
PHONETIC AND ORTHOGRAPHIC SIMILARITY APPROACHES

	Similarity-based approaches	Distance-based approach
Orthographic similarity	edit [26] gram-dist [24]	lcs-similarity [17] DICE [16]
Phonetic similarity	Soundex [8] Phonix [7]	ALINE [10]

Each of the approaches listed in Table I have different advantages. However, Lambert *et al.* point out that Trigram-2b and normalized edit distance (NED) have better accuracy in similarity-based approaches and distance-based approach, respectively [13]. Phonetic similarity is hard to measure. Two criterions of orthographic similarity, Trigram-2b and NED, are thus used in this paper and described as follow.

1. Trigram-2b: String (drug name) is divided into continuous substrings. Each of substrings has three letters (trigram). DICE is then used to calculate the similarity as in (1).

$$DICE(A, B) = 2 N_C / (N_B + N_A) \quad (1)$$

where N_A is number of trigrams of drug A, N_B is number of trigrams of drug B, N_c is number of trigrams appear in both drugs A and B. Take two drugs, *acthar* and *acular*, as an example. The number of trigrams of each of them is six (*acthar*: □□a, □ac, act, cth, tha, har; *acular*: □□a, □ac, acu, cul, ula, lar). Two of them are identical (□□a, □ac). Thus, the similarity value is 0.33 (=2 * 2 / (6+6)).

2. Normalized edit distance: NED is calculated by using (2).

$$NED(A, B) = D(A, B) / MaxLen(A, B) \quad (2)$$

where D(A, B) is edit distance which means that the required number of steps to transfer drug A into drug B. Edit means insert, delete or replace a letter. MaxLen(A, B) represents maximum length of drugs A and B. For example, transfer *ambient* into *amen*, we need to delete letters *b*, *i* and *t*. Edit distance of *ambient* and *amen* is 3. Maximum length of them is 6. Thus, the NED is 0.5 (= 3/6).

D. Classification Models

In this subsection, three models, logistic regression, C4.5 and SVM, are introduced as follows.

Logistic Regression Model

The main concept of regression model which is a mathematic model is using independent variable to estimate dependent variable. When there is only an independent variable and a dependent variable, the regression model is called linear regression model. When there are many independent variables and a dependent variable, the regression model is called multiple regression model. However, the disadvantage of the two regression models is that they allow only one dependent variable. Logistic regression model is then proposed to deal with this problem. Comparing with traditional regression models, it allows two dependent variables in the model by using a function such that independent variables can be mapped into "1" or "0". Hence, logistic regression model is often applied to different problems, including dispensing error [13].

C4.5 Model

Classification is commonly used in data mining techniques. The goal of classification is trying to find rules that can classify new data correctly. Classification is a supervised learning approach which is learning useful rules from labeled data. The derived rules are stored in specified data structure is called classifier. The well known and most used approach is decision tree (ID3) which has been proposed by Quinlan in 1992 [21]. The main concept is using the difference of data distribution (e.g. entropy) as criterion to build the model. The derived classification rules thus can be used to classify new data. In the same time, the derived classification rules are represented as tree structure. Such a tree structure is called classifier. The improved approach, C4.5, is used in this paper [20].

Support Vector Machine Model

Another well known classification approach is SVM (support vector machine) which has been proposed by Vapnik in 1990 [25]. The main concept of SVM is based on mathematic approach to find an optimal marginhyperplane that can classify positive and negative instances clearly. There are two advantages of SVM. The first is that it is a powerful approach for classification when there are only two classes. The second is that it is useful when continuous variables are used to build the model. The most famous approaches are *LibSVM* which have been developed by Chang *et al.* [3] and *LightSVM* which have been implemented by Joachims *et al.* [9]. In this paper, the *LibSVM* is used in to analyze the dispensing error problem.

IV. FRAMEWORK FOR DISPENSING ERROR MINING

In this section, the proposed framework for dispensing error mining as shown in Fig. 2 will be described in details.

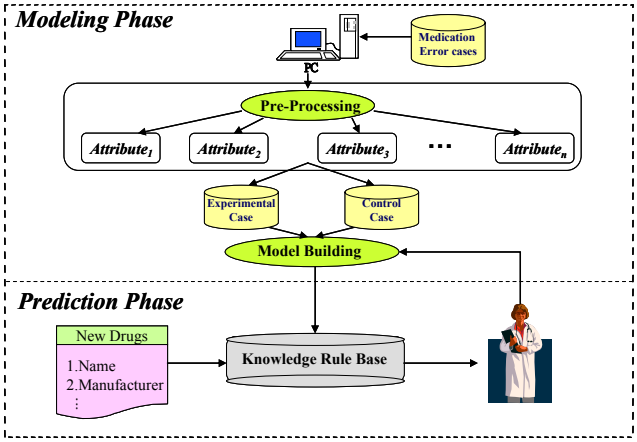

Fig. 2. Framework for dispensing error mining.

From Fig. 2, the proposed framework consists of two phases, modeling phase and prediction phase. In modeling phase, the pre-processing of dispensing error cases are made to generate experimental cases. The control cases were then generated from experimental cases. Experimental cases are dispensing error cases. On the contrary, control cases are not. From the framework, we can thus use different approaches to build the classification models. In this paper, three approaches, logistic regression, C4.5 and SVM, are adapted in the first phases. In perdition phase, when new drugs arrived, it can be alert to related people according to the model if dispensing error may be arisen. Besides, related people at administration can also have feedback to the model to improve the model.

V. EXPERIMENTAL RESULTS

In this section, the experiments were made with real data which was collected form a hospital at Taiwan. The following subsections including experimental dataset, comparison of three models and analysis of three factors, are described.

A. Experimental Dataset

The experimental dataset were collected form dispensing error system of a hospital at Taiwan. This dataset was used to analyze the reflection of different considered factors to discuss dispensing error problem. The dataset contains 219 records of experimental cases and 219 records of control cases. The 219

records in experimental cases are dispensing error cases whose were collected form the system. The control cases were generated form experimental cases as follow. Firstly, all drug names in experimental cases were gathered together into drug-names pool and duplicated drug names were removed from it. Two drug names were then selected randomly as a pair from the drug-names pool to form the control cases. Three factors whose have thirteen attributes and one class label are shown in Table 2.

TABLE II
THE SELECTED ATTRIBUTES

Attributes	Description	Type
subject	Class label	Discrete
totaloca	Location score of drugs	Continuous
dose	Dose score of drugs	Continuous
pharma	Pharmacology score of drugs	Continuous
totaform	The form score of drugs	Continuous
size	Size score of drugs	Continuous
shape	Shape score of drugs	Continuous
color	Color score of drugs	Continuous
T1	Similarity of scientific name of drugs	Continuous
T2	Similarity of product name of drugs	Continuous
T4	Similarity of scientific and product names of drugs	Continuous
ned1	Distance of scientific name of drugs	Continuous
ned2	Distance of product name of drugs	Continuous
ned4	Distance of scientific and product names of drugs	Continuous

In Table II, *totaloca* is an attribute of environmental factor. The attributes from *dose* to *color* are drug-properties factor. The last six variables are drug-names factor. Since the variables from *totaloca* to *color* whose belong to environmental factor and drug-properties factor are not easily to score, they were scored by senior dispensers. In drug-properties factor, Trigram-2b which is an orthographic similarity measure and NED (normalized edit distance) which is an orthographic distance measure are used in this paper due to its accuracy [13]. Besides, Lambert *et al.* also point that if the value of trigram-2b is larger than 0.116 or NED is less than 0.659, then they are similar drugs [13]. In this paper, discretizations of the last six variables were made for model building according to Lambert's suggestion.

B. Comparison of Models

In modeling phase, three techniques, C4.5, logistic regression and SVM, are used to build models. The *J48* (C4.5) which is a component of famous data mining tool *weka* [27] was used to generate classification tree. Logistic regression model was derived by using *SPSS* which is a statistical tool. The *libSVM* which has been developed by Chang *et al*. is used to build SVM model. In order to get reliable results, the experiments were made to show average values of accuracy of ten runs with four different training and testing proportion. The results are shown in Table III.

TABLE III
AVERAGE VALUES OF ACCURACY (%)

	Training:Testing (%)			
Model	6:4	7:3	8:2	9:1
C4.5(Training)	**85.6**	85.26	85.57	85.38
LR(Training)	**86.05**	85.71	86.02	85.57
SVM(Training)	**85.19**	84.77	85.14	84.56
C4.5(Testing)	79.77	82.04	81.47	**83.86**
LR(Testing)	82.81	**84.32**	82.28	82.74
SVM(Testing)	82.10	83.71	83.068	**85.68**

From Table III, it is easily to observe that most of three models have good results among four different proportions of training and testing. When the ratio of training and testing is 6:4, it has the best accuracy in training phase. However, the accuracy of testing data is the lowest in testing phase. When the ratio of training and testing is 9:1, *J48* and *SVM* have the highest accuracy in testing phase. Due to avoid over fitting problem, we do not suggest building the model with this proportion. When the ratio of training and testing is 7:3, the logistic regression has the best results. Hence, the ratio of training and testing is 7:3 or 8:2 is appropriate choice for building the model.

When they are applied to real application, the best model is always first priory. The best model of each of three models with different proportion of training and testing were picked among ten runs. The results are shown in Fig. 3.

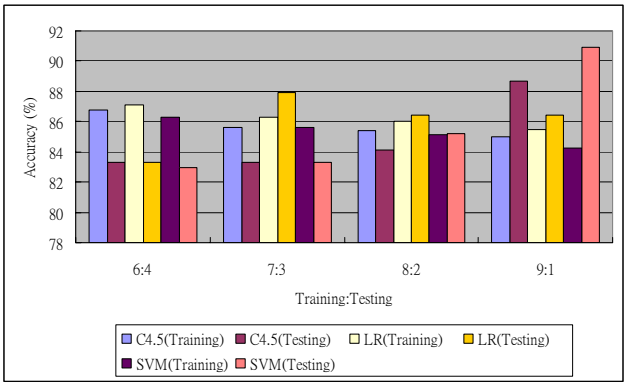

Fig. 3. The accuracy of the models.

From Fig. 3, it can easily be observed two things. Firstly, although the three models have similar accuracy, logistic regression is a good choice of three models. However, C4.5 is suggested when user want have clear prevention rules. The accuracy of SVM is between logistic regression and C4.5. It is a surprising observation because SVM is a well know approach and suitable for continuous attributes. Secondly, when the proportion of the training and testing is 7:3 or 8:2, the accuracy of three models have good results. We thus suggest that proportion of the training and testing is 7:3 or 8:2 is suitable

used to build models. The decision tree which was derived form C4.5 is shown in Fig. 4.

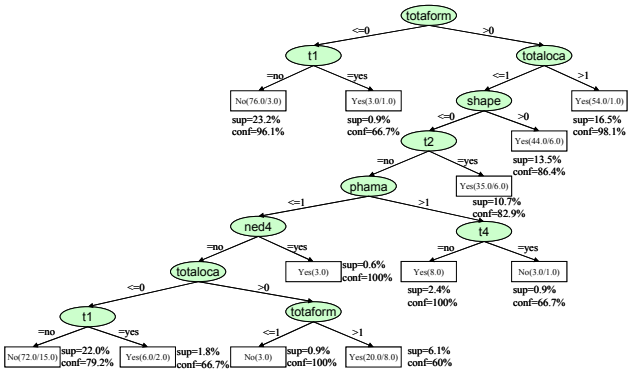

Fig. 4. The generated decision tree.

From Fig. 4, it is easily to know that the drugs-properties factor and environmental factor are important. Take the rule "IF *totaform* >0 and *totaloca* >1, Then dispensing error, sup. = 16.5%, conf. = 98.1%" as an example, the attributes *totaform* and *totaloca* were used to determide the dispensing error and support and confidence value are 16.5% and 98.1%, respectively, which mean the rule is reliable. We thus conclude that the drugs-properties factor and environmental factor are important. The rules whose support values are large 10% are listed in Table IV and can be referenced by medical experts.

TABLE IV
DISPENSING ERROR RULES

ID	Rules
1	IF *totaform* > 0 and *totaloca* > 1, Then dispensing error, sup. = 16.5%, conf. = 98.1%
2	IF *totaform* > 0 and *totaloca* ≤ 1 and *shape* > 0, Then dispensing error, sup. = 13.5%, conf. = 86.4%
3	IF *totaform* > 0 and *totaloca* ≤ 1 and *shape* ≤ 0 and *T2* = yes, Then dispensing error, sup. = 10.7%, conf. = 82.9%

In logistic regression model, the backward approach is used to build the model. The result is shown in (3).

$$Z = 0.101 + 1.204(totaloca) + 1.355(totaform) + 11.545(T1) + 7.869(T2) - 4.078(ned4) + 1.961(shape) \quad \cdots\cdots (3)$$

From (3), three variables belong to the drugs-properties factor and environmental factor of the six variables. The effect of the considered factors is demonstrated again.

C. Analysis of Drug-Names factor, drug-properties factor and environmental factor

In order to discuss the reflections of drugs-properties factor and environmental factor, the experiments were made to show the comparison of the accuracy of the three models with and without drugs-properties factor and environmental factor. The reulsts are shown in Fig. 5.

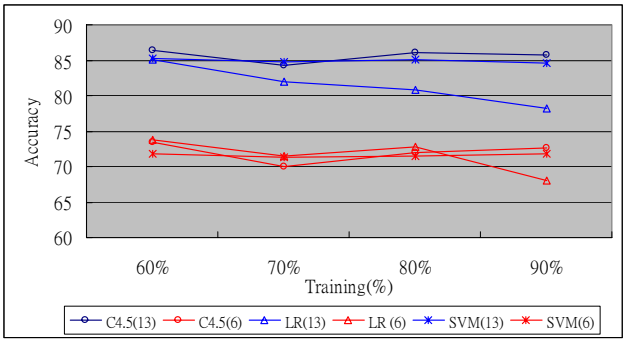

Fig. 5. Comparison of the accuracy of the three models with and without drugs-properties factor and environmental factor.

From Fig. 5, it is easily to observe that the three modes have high accuracy when all attributes are used to build the models. On the contrary, when only the drug-names factor is considered in the models, the accuracies are not good enough. From the results, we can conclude that the drugs-properties factor and environmental factor are useful for classifying dispensing error.

VI. CONCLUSIONS AND FUTURE WORKS

In this paper, three approaches, namely logistic regression, C4.5 and SVM, have been used to analyze dispensing error problem. There exist three main contributions in this paper. Firstly, the drugs-properties factor and environmental factor are used together with drug-names factor to analyze dispensing error. Second, a dispensing error mining framework has been proposed and three approaches are used to build the models. At last, the derived rules can be used for dispensing error prevention and be referenced by medical experts.

For the experimental results, comparison of three models with different proportion of training and testing dataset were made to provide appropriate setting. The ratio of training and testing as 7:3 or 8:2 is an appropriate choice for building an effective model. Besides, the experimental results also showed that the accuracies of three models are improved form 70% to 80% when drugs-properties factor and environmental factor are considered. In the future, we will explore further improvement on the proposed framework. For example, a fusion model of C4.5, logistic regression and SVM can be considered in modeling phase. Meanwhile, we will continue to enhance the proposed framework to retrieve better knowledge for utilization by medical experts so as to reduce the medication error problems.

ACKNOWLEDGMENT

This research was supported by National Science Council, Taiwan, R.O.C., under grant number NSC94-2218-E-006-043

REFERENCES

[1] R. Agrawal, T. Imielinksi and A. Swami, "Mining association rules between sets of items in large database," The 1993 ACM SIGMOD Conference, Washington DC, USA, 1993, pp.207-216.

[2] D. M. Ashcroft, P. Quinlan and A . Blenkinsopp, "Prospective study of the incidence, nature and causes of dispensing errors in community

pharmacies," Pharmacoepidemiology and drug safety, Vol. 14, No. 5, 2005, pp. 327-332.

[3] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[4] G. F. Cavell and C. A. Oborne, "Anonymously reported medication errors: the tip of the iceberg," The Internal Journal of Pharmacy Practice, R52, 2001.

[5] C. D. Kohn, To error is human: building a safer health system. National Academy Press, 1999.

[6] S. S. Chua, I. C. Wong et al., "A feasibility study for recording of dispensing errors and near misses' in four UK primary care pharmacies," Drug safety, Vol. 26, No. 11, pp. 803-813, 2003.

[7] T. N. Gadd, "PHONIX: The Algorithm," *Program: Automated Library and Information Systems*, Vol. 24, No. 4, 1990, pp. 222-237.

[8] P. A. V. Hall and G. R. Dowling, "Approximate String Matching," *Computing Surveys*, Vol. 12, No. 4, 1980, pp. 381-402.

[9] T. Joachims, *Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

[10] G. Kondrak, "A new algorithm for the alignment of phonetic sequences," *In Proceedings of NAACL-2000: First Meeting of the North American Chapter of the Association for Computational Linguistics*, 2000, pp. 288-295.

[11] J. W. Kenagy and G. C. Stein, "Naming, labeling, and packaging of pharmaceuticals," *American Journal of Health-System Pharmacy*, Vol. 58, No. 21, 2001, pp. 2033-2041.

[12] U. A. Kistner and M. R. Keith, K. A. Sergeant and J. A. Hokanson, "Accuracy of dispensing in a high-volume, hospital-based outpatient pharmacy," *American Journal of Hospital Pharmacy*, Vol. 51, No. 22, 1994, pp. 2793-2797.

[13] B. L. Lambert, S. J. Lin, K. Y. Chang and S. K. Gandhi, "Similarity as a risk factor in drug-name confusion errors: the look-alike (orthographic) and sound-alike (phonetic) model," Med Care, Vol. 37, No. 12, 1999, pp. 1214-1225.

[14] J. Lazarou, B. H. Pomeranz, and P. N. Corey, "Incidence of adverse drug reactions in hospitalized patients," *Journal of the American Medical Association*, 1998, pp. 1200 - 1205.

[15] G. Long and C. Johnson, "A pilot study for reducing medication errors," *QRB Quality review bulletin*, Vol. 7, No. 4, 1981, pp. 6-9.

[16] A. McEnery and M. P. Oakes, Sentence and Word Alignment in the CRATER Project: Methods and Assessment in J. Thomas & M. Short (eds) *Using Corpora for Language Research*, London: Longman, 1996, pp 211-231.

[17] Dan I. Melamed, "Bitext Maps and Alignment via Pattern Recognition," *Computational Linguistics*, Vol. 25, No. 1, 1999, pp.107-130.

[18] J. B. McQueen, "Some Methods of Classification and Analysis of Mutivariate Observations," Proceedings of the 5th Berkeley Symposium on Mathematical Satistics and Probability, pages 281-297,1967.

[19] G. M. Peterson, M. S. Wu and J. K. Bergin, "Pharmacists attitudes towards dispensing errors: their causes and prevention," *Journal of clinical pharmacy and therapeutics*, Vol. 24, No. 1, pp. 57-71.

[20] J. R. Quinlan, *C4.5 : programs for machine learning*, The Morgan Kaufmann series in machine learning. Morgan Kaufmann Publishers, San Mateo, Calif., 1993.

[21] J. R. Quinlan, *Induction of Decision Trees*, Machine Learning, Kluwer Academic Publishers, Vol. 1, Issue 1, 2003, pp. 81–106.

[22] W. J. Rudman et al, "The use of data mining tools in identifying medication error near misses and adverse drug events," *Topics in Health Information Management,* Vol. 23, 2002, pp. 94-101.

[23] N. Tuohy and S. Paparella, "Look-alike and sound-alike drugs: errors just waiting to happen," *Journal of emergency nursing*, Vol. 31, No. 6, 2005, pp. 569-571.

[24] E. Ukkonen, "Finding approximate patterns in strings," *Journal of Algorithms*, Vol. 6, 1985, pp.132-137.

[25] V. Vapnik, *The Nature of Statical Learning Theory*, Springer Verlag, New York, 1995.

[26] Robert A. Wagner and Michael J. Fischer, "The string-to-string correction problem," *Journal of the ACM*, Vol. 21, No. 1, 1974, pp. 168–173.

[27] Ian H. Witten and Eibe Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.