Original Article

Healthc Inform Res. 2013 March;19(1):16-24.
http://dx.doi.org/10.4258/hir.2013.19.1.16
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Comparison of Machine Learning Algorithms for Classification of the Sentences in Three Clinical Practice Guidelines

Mi Hwa Song, PhD[1], Young Ho Lee, PhD[2], Un Gu Kang, PhD[2]

[1]Information and Communication Science, Semyung University, Jecheon; [2]IT Department, Gachon University, Incheon, Korea

Objectives: Clinical Practice Guidelines (CPGs) are an effective tool for minimizing the gap between a physician's clinical decision and medical evidence and for modeling the systematic and standardized pathway used to provide better medical treatment to patients. **Methods:** In this study, sentences within the clinical guidelines are categorized according to a classification system. We used three clinical guidelines that incorporated knowledge from medical experts in the field of family medicine. These were the seventh report of the Joint National Committee (JNC7) on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure from the National Heart, Lung, and Blood Institute; the third report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults from the same institution; and the Standards of Medical Care in Diabetes 2010 report from the American Diabetes Association. Three annotators each tagged 346 sentences hand-chosen from these three clinical guidelines. The three annotators then carried out cross-validations of the tagged corpus. We also used various machine learning-based classifiers for sentence classification. **Results:** We conducted experiments using real-valued features and token units, as well as a Boolean feature. The results showed that the combination of maximum entropy-based learning and information gain-based feature extraction gave the best classification performance (over 98% f-measure) in four sentence categories. **Conclusions:** This result confirmed the contribution of the feature reduction algorithm and optimal technique for very sparse feature spaces, such as the sentence classification problem in the clinical guideline document.

Keywords: Knowledge Bases, Data Mining, Information Storage and Retrieval

Submitted: October 15, 2012
Revised: 1st, December 28, 2012; 2nd, March 15, 2013
Accepted: March 21, 2013

Corresponding Author
Un Gu Kang, PhD
IT Department, Gachon University, 191 Hambangmoe-ro, Yeonsu-gu, Incheon 406-799, Korea. Tel: +82-32-820-4392, Fax: +82-32-820-4109, E-mail: ugkang@gachon.ac.kr

I. Introduction

Clinical Practice Guidelines (CPGs) are an effective tool for determining appropriate disease control methods in the medical field. To facilitate decision-making on the part of the medical staff, they provide a systematic process and minimize the gap between diagnostic judgment and scientific evidence [1]. The clinical guideline modeling service stores clinical practice processes (algorithms) in an executable format that can be run by an authoring and inference engine through a visual tool [2]. This means that the service can be optimized dynamically according to the current health sta-

tus of a patient or a behavior change, or it can be updated by new medical study results. In the process of modeling and editing this information, reference operations are conducted on a portion of the data. Therefore, it is essential that evidence-based knowledge extraction and management features are available. In this paper, we demonstrate the implementation of a knowledge extraction feature that can be searched by a medical expert with experience of the clinical guideline modeling service. To implement this feature, we realized that we could reduce the search time required by a medical expert by categorizing certain sentential elements. At this time, we consider the documents included in the search results to be a set of the sentential elements. An additional consideration during the system design is to use various machine learning models.

1. Sentence Classification

Kim et al. [3] studied automatic sentence classification for evidence-based medicine (EBM). In their study, tags such as Background, Population, Intervention, Outcome, Study Design, and Other were used to automatically classify sentences from paper abstracts. Abstracts have the feature that their sentences are listed in order of background, main contents, and results. Kim et al. [3] classified sentences using conditional random fields, which are useful for learning such sequential data. However, as the flow of sentences in the guidelines used for training data is not sequential, the guidelines were not considered in the machine learning model selection. In another study, Nawaz et al. [4] undertook a process of multi-tagging to classify bio-events. In this study, tags were applied to bio-events such as Knowledge Type, Manner, Certainty Level, Logical Type, Source, and Lexical Polarity. Among these, the knowledge type was further categorized into Investigation, Observation, Analysis, and General. Categorizing the knowledge type involves tagging representative words for knowledge types, or so-called lexical clues. Pan [5] conducted a sentence classification study using multi-label tagging of a single sentence. In this study, sentences or clauses used in medical/life science papers were considered as text instances, created by complex combinations of semantic classes, such as Focus, Polarity, Certainty, Evidence, Direction, or Trend. Various classifiers (or discrimination models), such as naïve Bayes, maximum entropy, and support vector machine (SVM), were used for the training algorithm [5]. As in the current study, Pan [5] also compared the performance between different training algorithms. The novelty of this study is the use of a transformation to reduce the dimension of the sentence instance feature space. Transformation is a feature extraction function that reduces the dimension to the curse of dimensionality [6]. This function captures and quantifies features of text, such as lexical, syntactic, and co-occurrence events. It also allows the results to be expressed as feature vectors.

2. Machine Learning Model

The maximum entropy model selects the probability distribution with the largest entropy from those that represent the current state of knowledge. To initialize the maximum entropy model, partial evidence is combined to estimate the probability of the instance class, which is generated from the specific context of the data. We obtain the conditional probability p by collecting evidence from the data via the feature function. In the maximum entropy model, the feature extraction function generally outputs a Boolean value ([0, 1]) as an indicator function. In this study, we use a real-valued feature vector to represent quantitative features alongside the Boolean feature vector.

Heckerman [7] and Tan et al. [8] studied disease classification using a Bayesian network by modeling patients with risk factors associated with heart disease, and Cho and Won [9] used a multilayer perceptron (MLP) to classify cancer. An MLP is a neural network model that is robust to noise due to its filtering of outliers, hidden variables, and errors that exist in the input vectors. This model can be used in domain problems with many uncertain factors, such as sentence category classification, so we adopt this approach in the present study.

Recently, SVMs have been the focus of a great deal of research among machine learning algorithms. An SVM uses a hyperplane to separate sets of n-dimensional data points belonging to different classes. SVM methods [10,11] then aim to optimize this hyperplane. In this study, as the feature extraction function generates five real values, we select SVM as a training algorithm for these feature vector inputs.

3. Feature Selection

Feature selection is the process of selecting a subset from an original feature set [12]. This can reduce the number of features and remove noisy data. It can also speed-up mining algorithms, and improve their performance in terms of estimation accuracy and readability. Broadly speaking, there are two types of feature selection [13]. The first uses a filter to select a subset of features with which to conduct the classification algorithm. One example of a filter is information gain (IG). This method is widely used in machine learning to evaluate the criteria of the relevance of terms. It calculates the amount of information in a term in each category by considering not only the frequency of occurrence of a term

in the document, but also the frequency of a term that does not occur in the document [14]. Lee and Lee [15] found that IG was an effective feature selection algorithm for classifying texts. In the second type of feature selection, a wrapper is used to apply a classification algorithm to a dataset, allowing the optimal features to be determined. For a large number of features, the wrapper method can take a long time. Genetic algorithms (GAs), in which a population evolves to find a better solution to an optimization problem, are a typical example of a wrapper method. Silla et al. [16] used GAs to undertake feature selection for an automatic text summary.

II. Methods

The purpose of the proposed system is to optimize the clinical care of a patient with a chronic disease based on medical papers and guidelines published by trusted institutions. Whereas existing practice models provide information only, the optimized practice model in this study provides information as well as its source and related information. Thus, we have developed a sentential classification system to categorize the characteristics of the information contained in certain sentences. As shown in Figure 1, the sentential classification process uses sentences extracted from the document as training data and creates a model to perform a classification test. The training data is formed by classifying sentences into an appropriate category using the knowledge manager to perform part-of-speech tagging and parsing.

1. Training Data Preparation

After segmenting the sentences contained in the document, we perform the process of semantic category tagging. The purpose of semantic category tagging is to enable the selective extraction of a sentence that has some semantic association with the rule in the specific algorithm node.

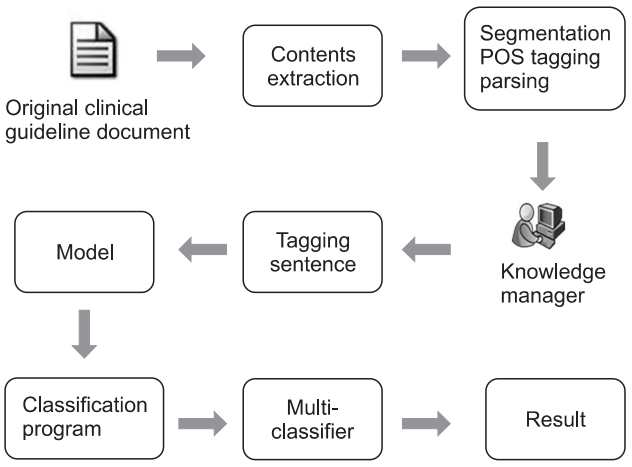

Figure 1. Overview of the sentential classification process. POS: part-of-speech.

For the training data, we used three clinical guidelines that incorporated knowledge from medical experts in the field of family medicine. These were the Seventh Report of the Joint National Committee (JNC7) on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure [17] from the National Heart, Lung, and Blood Institute; the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults [18] from the same institution; and the Standards of Medical Care in Diabetes 2010 [19] report from the American Diabetes Association.

The training data was generated by attaching a single tag to each extracted sentence. This method is different from that used by Shatkay et al. [20], who attached multi-tags. This is because the primary purpose of our system is to search for knowledge that is highly associated with the current CPGs. In addition, it was assumed that the use of a single semantic tag would be sufficient to achieve this purpose. In the semantics category classification module, we adopted the following definitions and categories of sentence: <RULE>, <RECOMMEND>, <ANALYSIS>, <GENERAL>. Further details of the sentence category tag definitions can be found in Table 1.

With regard to the three guideline documents, sentences corresponding to each semantic category were extracted by three researchers. After discussing the classification criteria with one another, 346 sentences were finally used as training data. Some of these are displayed in Table 2.

2. Training Data Representation and Feature Extraction

To classify their semantic category, each sentence should be represented by a feature vector. A feature vector extracts the feature values of a sentence, thereby enabling its use by a training algorithm. In this study, there are two feature types: five real-valued vectors and a Boolean feature vector. When training the classification model for text instances consisting of tokens, the individual token occurrence is itself considered the biggest feature element in a bag-of-words (a set of words), which is generally used as one of the feature vector expression methods. Regardless of the order of the words in the document, the token weight is calculated by the frequency of occurrence of an individual word. In a bag-of-words, the dimension of the feature space is equivalent to the size of the unique token occurring in the document. Thus, using only a general classifier training algorithm, it is difficult to estimate practical parameters for the discrimination model. In addition, if the amount of training data is small, a linearly

http://dx.doi.org/10.4258/hir.2013.19.1.16

Table 1. Sentence categories for semantic functions

Sentence category	Class description
RULE	String used in each rule in the guideline algorithm
	Rule is expressed by free text or formal representation depending on the guideline publisher
	Including inequality sign (>, <) and quantity unit, or implying medical rules semantically
	In case that the certainty level is high [21]
RECOMMEND	Sentence category which includes an expression of recommendation for practice by the author of the guideline
	Implying recommendation which is not strongly evidenced compared to RULE
	In case that the certainty level is medium [21]
	Example sentence which includes a specific scope for helping to understand the contents
ANALYSIS	Sentence category including statistical facts which were found by clinical experiment on patient cohort in the guideline document
	Study results such as randomized controlled trial, cohort study, and meta-analysis
GENERAL	As basic classification, generally accepted knowledge such as scientific facts, process, and methodology

Table 2. Sentence tagging examples by category

Category	Example of sentences
RULE	If LDL remains ≥130 mg/dL after 3 months of TLC, consideration can be given to starting an LDL-lowering drug to achieve the LDL goal of <130 mg/dL.
	Their LDL cholesterol goal is <160 mg/dL.
RECOMMEND	SMBG should be carried out three or more times daily for patients using multiple insulin injections or insulin pump therapy.
	For overall cardiovascular risk reduction, patients should be strongly counseled to quit smoking.
ANALYSIS	The Diabetic Retinopathy Study showed that panretinal photocoagulation surgery reduced the risk of severe vision loss from PDR from 15.9% in untreated eyes to 6.4% in treated eyes.
	Framingham Heart Study investigators recently reported the lifetime risk of hypertension to be approximately 90% for men and women who were nonhypertensive at 55 or 65 years and survived to age 80–85.
GENERAL	The level of evidence that supports each recommendation is listed after each recommendation using the letters A, B, C, or E.
	Diabetes care is complex and requires that many issues, beyond glycemic control, be addressed.

LDL: low-density lipoprotein, TLC: therapeutic lifestyle changes, SMBG: self-monitoring of blood glucose, PDR: proliferative diabetic retinopathy.

non-separable problem can occur if instance data points belong to different semantic categories in the vector space.

To solve this problem, we must reduce the dimension of the feature space. This involves eliminating tokens that are harmful to the classification model training. This is known as feature selection. Generally, function words (or stop words), such as articles or prepositions, are considered to be redundant features, lacking in discrimination. Therefore, these features are processed so as to be eliminated from the training data. Besides the elimination of function words, some filtering is required to eliminate unnecessary features. To this end, various algorithms exist to check whether each individual feature is significant or not. In this study, an optimal feature subset selection algorithm was implemented using a GA and IG. Through our algorithm, between 90% and 99% of features (tokens) were eliminated. It has been reported that the performance of a classifier is not degraded by this degree of feature elimination [22]. Although this reduces the dimension of the feature space, a problem occurs when the features of a token unit are extracted. This is because feature elements that exceed the token unit, such as proper nouns made of more than two tokens, the existence or occurrence

Table 3. Real-valued feature vector definition

No.	Description
1	Size of character included in the instance other than alphabet and number
2	The frequency of occurrence of phrase in the current instance which exclusively occurred in the RULE class tagged instance in the training data
3	The frequency of occurrence of phrase in the current instance which exclusively occurred in the RECOMMEND class tagged instance in the training data
4	The frequency of occurrence of phrase in the current instance which exclusively occurred in the ANALYSIS class tagged instance in the training data
5	The frequency of occurrence of token/phrase in the current instance which exclusively co-occurred in the RULE class tagged instance in the training data

Table 4. WEKA API

Feature selection method	API
Information gain	weak.attribureSelection.InfoGain-AttributeEval
Genetic search	weka.attributeSelection.Genetic-Search

WEKA: Waikato Environment for KnowledgeAnalysis, API: application programming interface.

of a phrase unit expression, or the co-occurrence of a specific token, are not considered. Therefore, it is necessary to use a function to extract frequently occurring proper nouns that belong to the specific semantic class, phrase unit token row, formal language symbols, and word unit co-occurrence.

To this end, this study uses a feature extraction function to reduce the dimension of the feature space [22]. Feature extraction utilizes components of the sentence instance as well as syntax information and pattern templates hidden in a combination of components as feature elements. A pattern template includes structural characteristics, such as the hierarchy inside sentences, repeatability, and concurrent events. As a result, it can derive a more generalized model from the instance set consisting of tokens. In addition, it has the advantage of reducing the probability of the generated model being over-fitted to the training data. To extract features such as co-occurrence between tokens, pattern templates, and the syntactic structure of sentence instances, the syntactic structures of sentences are analyzed by the Stanford Parser [23]. The feature values extracted from the parse tree are shown in Table 3.

For example, the real feature vector extracted from the sentence structure analysis tree will have the following form:

<0, 4, 0, 1, 4> → <RULE>.

The first value indicates that the number of non-alphanumeric characters in the corresponding sentence (instance) is 0. The second value indicates the number of phrase unit expressions that occur within instances tagged as the <RULE> class (e.g., "not achieved"). In the current instance, there are four such expressions. The third value indicates that the number of phrase unit expressions or phrase unit templates within instances of the <RECOMMEND> class is 0. The fourth value in the vector denotes the number of token rows, such as a phrase unit expression, e.g., "correlated with," that occur within sentences categorized as <ANALYSIS>, which in the current instance is 1. Finally, the fifth value denotes the number of co-occurring tokens/phrase expressions in instances tagged with the <RULE> class. The feature types generated in this study give a real-valued feature vector and a number of Boolean values. Whereas the feature selection algorithm is not applied to the real-valued vector, it is applied to the Boolean feature. The highest ranked sub-set and five real-valued features are then combined. Recently, our group proposed a feature transformation function for automatic sentence classification and evaluated the performance using medical guideline texts [24].

III. Results

1. Experimental Environment

In this study, the Waikato Environment for Knowledge Analysis (WEKA) was used to implement the two feature selection algorithms [25] (Table 4). When the GA was used, the following parameters were set [26]: population size, 20; number of generations, 20; probability of crossover, 0.6; probability of mutation, 0.033.

2. Results

In the experiment, we performed a 10-fold cross validation on the 346 selected sentences. In general, a model is trained and evaluated by separating training data, test data,

 http://dx.doi.org/10.4258/hir.2013.19.1.16

and validation data. However, the cross validation method is well suited to experiments with a small amount of data, as in our experiment. In this cross validation method, the entire data set was first classified into N sub-sets. The model was then trained using N-1 sets of training data before it was applied to the one remaining test set. The same process was repeated N-1 times, so that each sub-set had been used as the test set to calculate the precision, recall, and f-measure of the algorithms. The values of precision and recall determine the accuracy of the classification. The precision, recall, and f-measure are calculated by Formula 1 (performance evaluation), where TP denotes true positive, FP denotes false positive, and FN denotes false negative.

$$Precision = \frac{TP}{TP + FP}, \qquad Recall = \frac{TP}{FN + TP}$$
$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall}.$$

More than 75 instances were assigned to each individual semantic class. In addition, a discrimination model was constructed using the feature vector set extracted from each instance (sentence).

In this study, we examined the classification performance for each sentence category from the features acquired using IG and GA. Furthermore, the feature types were configured as token units, Booleans, and real values. The experiment aimed to find out which features most affect the classification performance. Table 5 shows the results using all feature types. We can see that the maximum entropy model gives the best performance with f-measures of 99.1% and 98.6%, followed by the radial basis function network (RBFN) using the information gain method with an f-measure of 98.8%.

Table 6 shows the classification performance without using the real-valued features among the token types. The results show that the performance depends on which machine learning model is used. Maximum entropy and RBFN exhibit comparable performance (over 90% f-measure) for the token units, Booleans, and real-valued features. The other machine learning models including Bayes network, MLP, naïve Bayes, and SVM showed worse performance in this experiment. Table 7 shows the classification performance using only the real-valued features without a feature selection algorithm. Compared to Tables 5 and 6, the f-measure was lower on average. However, when using BayesNet and the GA, the performance of MLP and SVM (f-measures of 77.8% and 78.1%, respectively) improved with the removal

Table 5. Classifier performance for each feature selection method

Feature selection	Token unit	Boolean	Real	Evaluation	MaxEnt	BayesNet	MLP	NB	RBFN	SVM
IG	O	O	O	Precision	0.991	0.859	0.060	0.937	0.989	0.541
				Recall	0.991	0.841	0.246	0.936	0.988	0.468
				f-measure	0.991	0.842	0.097	0.937	0.988	0.357
GA	O	O	O	Precision	0.986	0.863	0.807	0.894	0.889	0.544
				Recall	0.986	0.847	0.743	0.890	0.867	0.618
				f-measure	0.986	0.848	0.737	0.890	0.869	0.553

MaxEnt: maximum entropy, BayesNet: Bayesian network, MLP: multilayer perceptron, NB: naïve Bayes, RBFN: radial basis function network, SVM: support vector machine, IG: information gain, GA: genetic algorithm.

Table 6. Classifier performance for each feature selection method without the real-valued feature vector

Feature selection	Token unit	Boolean	Real	Evaluation	MaxEnt	BayesNet	MLP	NB	RBFN	SVM
IG	O	O	×	Precision	0.991	0.796	0.060	0.899	0.989	0.084
				Recall	0.991	0.783	0.246	0.899	0.988	0.289
				f-measure	0.991	0.784	0.097	0.899	0.988	0.130
GA	O	O	×	Precision	0.989	0.751	0.772	0.830	0.938	0.084
				Recall	0.988	0.728	0.734	0.829	0.931	0.289
				f-measure	0.988	0.731	0.734	0.830	0.931	0.130

MaxEnt: maximum entropy, BayesNet: Bayesian network, MLP: multilayer perceptron, NB: naïve Bayes, RBFN: radial basis function network, SVM: support vector machine, IG: information gain, GA: genetic algorithm.

of the feature selection algorithm.

The classification performance in relation to sentence category is shown in Table 8. The best results were obtained using the maximum entropy-based feature and IG. In terms of sentence categories, the best performance was found in the RULE and RECOMMEND classifications, whereas ANALYSIS and GENERAL showed a lower performance level. Finally, Table 9 compares the best performance values from Tables 5–7. This confirms that the best sentence classification performance with an f-measure of 99% was obtained using IG and maximum entropy.

IV. Discussion

In this study, we designed and implemented a clinical guideline sentence classifier using various models of machine learning. We conducted experiments using real-valued features and token units, as well as a Boolean feature. The results showed that the combination of maximum entropy-based learning and IG-based feature extraction gave the best

Table 7. Classifier performance without feature selection for real-value feature extraction

Performance	MaxEnt	BayesNet	MLP	NB	RBFN	SVM
Precision	0.805	0.815	0.800	0.808	0.801	0.801
Recall	0.783	0.760	0.775	0.757	0.780	0.777
f-measure	0.787	0.771	0.778	0.768	0.784	0.781

MaxEnt: maximum entropy, BayesNet: Bayesian network, MLP: multilayer perceptron, NB: naïve Bayes, RBFN: radial basis function network, SVM: support vector machine.

Table 8. Classification performance by sentence category for each feature selection method

Category	Classifier	Token unit	Boolean	Real	Feature selection	Precision	Recall	f-measure
RULE	MaxEnt	O	O	O	IG	0.988	1.000	0.994
	MaxEnt	O	O	O	GA	1.000	0.988	0.994
	MaxEnt	O	O	×	IG	1.000	0.988	0.994
RECOMMEND	MaxEnt	O	O	O	IG	1.000	0.990	0.995
	MaxEnt	O	O	×	IG	0.990	1.000	0.995
ANALYSIS	MaxEnt	O	O	O	IG	0.987	0.987	0.987
	RBFN	O	O	O	IG	1.000	0.974	0.987
GENERAL	MaxEnt	O	O	O	IG	0.988	0.988	0.988

MaxEnt: maximum entropy, RBFN: radial basis function network, IG: information gain, GA: genetic algorithm.

Table 9. Classifier performance for each feature selection method: best case

Feature selection	Token unit	Boolean	Real	Performance	MaxEnt	BayesNet	MLP	NB	RBFN	SVM
IG	O	O	O	Precision	0.991	0.859	0.060	0.937	0.989	0.541
				Recall	0.991	0.841	0.246	0.936	0.988	0.468
				f-measure	0.991	0.842	0.097	0.937	0.988	0.357
IG	O	O	×	Precision	0.991	0.796	0.060	0.899	0.989	0.084
				Recall	0.991	0.783	0.246	0.899	0.988	0.289
				f-measure	0.991	0.784	0.097	0.899	0.988	0.130
N/A	×	×	O	Precision	0.805	0.815	0.800	0.808	0.801	0.801
				Recall	0.783	0.760	0.775	0.757	0.780	0.777
				f-measure	0.787	0.771	0.778	0.768	0.784	0.781

MaxEnt: maximum entropy, BayesNet: Bayesian network, MLP: multilayer perceptron, NB: naïve Bayes, RBFN: radial basis function network, SVM: support vector machine, IG: information gain.

 http://dx.doi.org/10.4258/hir.2013.19.1.16

classification performance in four sentence categories.

Moreover, we found that transformation has the advantage of exploiting structural and underlying features which go unseen by the BOW model. From this result, we confirmed the contribution of the feature reduction algorithm and optimal technique for very sparse feature spaces, such as the sentence classification problem in the clinical guideline document. In future research, an automatic annotator for large data sets and a user-defined flexible annotation system will be implemented and evaluated. We also plan to further analyze the corpus, and in particular the guideline sentences annotated as GENERAL, to develop a more robust system.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This research was supported by Grant No. 10037283 from the Industrial Strategic Technology Development Program funded by the Ministry of Knowledge Economy.

References

1. Woolf SH, Grol R, Hutchinson A, Eccles M, Grimshaw J. Clinical guidelines: potential benefits, limitations, and harms of clinical guidelines. BMJ 1999;318(7182):527-30.
2. Buchanan BG, Shortliffe EH. Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. Reading (MA): Addison-Wesley; 1984.
3. Kim SN, Martinez D, Cavedon L, Yencken L. Automatic classification of sentences to support Evidence Based Medicine. BMC Bioinformatics 2011;12 Suppl 2:S5.
4. Nawaz R, Thompson P, McNaught J, Ananiadou S. Meta-knowledge annotation of bio-events. In: Proceedings of the International Conference on Language Resources and Evaluation; 2010 May 17-23; Valletta, Malta. p. 2498-505.
5. Pan F. Multi-dimensional fragment classification in biomedical text [dissertation]. Ottawa, Canada: Queen's University; 2006.
6. Duda RO, Hart PE, Stork DG. Pattern classification. 2nd ed. New York (NY): Wiley; 2000.
7. Heckerman D. Bayesian networks for data mining. Data Min Knowl Discov 1997;1(1):79-119.
8. Tan PN, Steinbach M, Kumar V. Introduction to data mining. Boston (MA): Pearson Addison-Wesley; 2006.
9. Cho SB, Won HH. Machine learning in DNA microarray analysis for cancer classification. In: Proceedings of the 1st Asia-Pacific Bioinformatics Conference on Bioinformatics; 2003 Feb 4-7; Adelaide, Australia. p. 189-98.
10. Vapnik VN. Statistical learning theory. New York (NY): Wiley; 1998.
11. Cristianini N, Shawe-Taylor J. An introduction to support vector machines: and other kernel-based learning methods. Cambridge (MA): Cambridge University Press; 2000.
12. Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Eng 2005;17(4):491-502.
13. Khan A, Baharudin B, Lee LH, Khan K. A review of machine learning algorithms for text-documents classification. J Adv Inf Technol 2010;1(1):4-20.
14. Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. In: Proceedings of the 14th International Conference on Machine Learning; 1997 Jul 8-12; Nashville, TN. p. 412-20.
15. Lee C, Lee GG. Information gain and divergence-based feature selection for machine learning-based text categorization. Inf Process Manag 2006;42(1):155-65.
16. Silla CN Jr, Pappa GL, Freitas AA, Kaestner CA. Automatic text summarization with genetic algorithm-based attribute selection. In: Lemaitre C, Reyes CA, Gonzalez JA. Advances in Artificial Intelligence, Lecture Notes in Computer Science, vol. 3315. Heidelberg, Germany: Springer; 2004. p. 305-14.
17. The National Heart, Lung, and Blood Institute. The seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. Bethesda (MD): National Institutes of Health; 2004.
18. The National Heart, Lung, and Blood Institute. Third report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (adult treatment panel III). Bethesda (MD): National Institutes of Health; 2002.
19. American Diabetes Association. Standards of medical care in diabetes: 2010. Diabetes Care 2010;33 Suppl 1:S11-61.
20. Shatkay H, Pan F, Rzhetsky A, Wilbur WJ. Multi-dimen-

sional classification of biomedical text: toward automated, practical provision of high-utility text to diverse users. Bioinformatics 2008;24(18):2086-93.
21. Thompson P, Venturi G, McNaught J, Montemagni S, Ananiadou S. Categorising modality in biomedical texts. In: Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining; 2008 May 28-30; Marrakech, Morocco. p. 27-34.
22. Feldman R, Sanger J. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge (MA): Cambridge University Press; 2007.
23. The Stanford Natural Language Processing Group. The Stanford parser: a statistical parser [Internet]. Stanford (CA): Stanford NLP Group; c2012 [cited at 2013 Mar 18]. Available from: http://nlp.stanford.edu/software/lex-parser.shtml.
24. Song MH, Kim SH, Park DK, Lee YH. A multi-classifier based guideline sentence classification system. Healthc Inform Res 2011;17(4):224-31.
25. Machine Learning Group at University of Waikato. Weka3: data mining software in Java [Internet]. Hamilton, New Zealand: The University of Waikato; c2012 [cited at 2013 Mar 18]. Available from: http://www.cs.waikato.ac.nz/ml/weka/.
26. Machine Learning Group at University of Waikato. Weka sources [Internet]. Hamilton, New Zealand: The University of Waikato; c2012 [cited at 2013 Mar 18]. Available from: http://weka.sourceforge.net/doc.stable/.

 http://dx.doi.org/10.4258/hir.2013.19.1.16