Journal of Biomedical Informatics 47 (2014) 153–159

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

An SVM-based high-quality article classifier for systematic reviews

Seunghee Kim, Jinwook Choi [*]

Department of Biomedical Engineering, College of Medicine, Seoul National University, 103 Daehak-ro, Jongno-gu, Seoul 110-799, Republic of Korea

ARTICLE INFO

Article history:
Received 4 January 2013
Accepted 17 October 2013
Available online 29 October 2013

Keywords:
Classification
Artificial intelligence
Evidence-based medicine
Review literature as topic

ABSTRACT

Objective: To determine whether SVM-based classifiers, which are trained on a combination of inclusion and common exclusion articles, are useful to experts reviewing journal articles for inclusion during new systematic reviews.
Methods: Test collections were built using the annotated reference files from 19 procedure and 4 drug systematic reviews. The classifiers were trained by balanced data sets, which were sampled using random sampling. This approach compared two balanced data sets, one with a combination of included and commonly excluded articles and one with a combination of included and excluded articles. AUCs were used as evaluation metrics.
Results: The AUCs of the classifiers, which were trained on the balanced data set with included and commonly excluded articles, were significantly higher than those of the classifiers, which were trained on the balanced data set with included and excluded articles.
Conclusion: Automatic, high-quality article classifiers using machine learning could reduce the workload of experts performing systematic reviews when topic-specific data are scarce. In particular, when used as training data, a combination of included and commonly excluded articles is more helpful than a combination of included and excluded articles.

1. Introduction

Evidence-based medicine (EBM), the philosophical origins of which extend back to mid-19th century Paris and earlier, is the conscientious, explicit, and judicious use of the current best evidence in making decisions regarding individual patient care [1]. EBM is an important development in clinical practice and scholarly research because this approach aims to provide better care with better outcomes through referring clinical decisions mainly on solid scientific evidence [2]. Because EBM directly applies the knowledge gained from large clinical trials to patient care, it promotes consistency in individual patient treatments, optimal clinical outcomes and quality of life [3]. The practice of EBM integrates individual clinical expertise with the best available external clinical evidence from systematic research.

Systematic review (SR) plays a key role in EBM [4] and attempts to identify, appraise, and synthesize all empirical evidence that meets the pre-specified eligibility criteria to answer a given question [5]. SR is processed in four distinct steps [6]. In the first step, the review topic and key questions are defined; then, all relevant studies are retrieved from a number of different databases, such as MEDLINE and EMBASE. In the second, experts select the retrieved abstracts that are most likely to meet the inclusion criteria (abstract triage step). In the next step, the experts closely read the selected articles, classify the articles as inclusion or exclusion in the SR using pre-specified eligibility criteria (full text triage step), and assess quality of inclusion articles. Finally, if included articles are sufficiently similar, their results are synthesized.

The new Health Technology Assessment (nHTA) center in the National Evidence-based Healthcare Collaborating Agency assesses new medical technologies introduced into Korean healthcare markets. It evaluates the safety and effectiveness of new medical technologies in real clinical settings. It systematically reviews all evidence relevant to the evaluation of those technologies. To date, 126 evidence reports have been completed and published [7].

The scientific literature is growing extremely fast (500,000 new abstracts are added to Medline every year), but only a minority of trials have been suggested in SRs [4,8]. The Cochrane Collaboration, which coordinates the creation and update of SRs, estimates that at least 10,000 reviews are needed to cover a substantial proportion of the studies relevant to health care [9]. However, creating a new SR or updating an existing one takes considerable time and effort. Using current methods, we have not been able to cover new issues and keep even half of existing reviews up-to-date [10]. We need to reduce avoidable processes in the production of research evidence [11]. Advanced information technologies can be developed and implemented to support SRs by reducing the labor required while still capturing high-quality evidence [4].

[*] Corresponding author. Fax: +82 2 745 7870.
E-mail address: jinchoi@snu.ac.kr (J. Choi).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.10.005

Researchers conducting SRs try to provide a more precise estimate and reduce uncertainty aimed at minimizing bias [5]. To produce more reliable findings, they exclude studies having high risk of biases (i.e., opinion pieces). In this study, we divided exclusion articles of the SRs into two parts: common and topic-specific. Common exclusion articles cannot be included in any SRs, because their results are definitely biased. Topic-specific articles can be included or excluded in SRs according to the topics. We hypothesize that by using commonly excluded articles across all SRs, we can automatically classify rigorous articles with better results than previous works. We propose a method that creates classifiers through training on articles that are included and commonly excluded.

2. Background

To find high-quality articles concerning internal medicine, Aphinyanaphongs and colleagues [2] applied machine learning (ML) techniques using data derived from the ACP Journal Club. They used a variety of ML techniques and found that the support vector machine (SVM) achieved the best performance with those data. They showed an efficient and improved means for identifying high-quality articles in internal medicine.

Cohen [12] also applied ML technique to systematic drug class reviews. They evaluated various feature combinations to classify rigorous articles based on SVM^{light} [13]. They showed that the overall top scoring combination among three feature combinations was the combination of unigram and n-gram with a length of 2 extracted from the title/abstract of MEDLINE and MeSH(Medical Subject Headings). On the other hand, the lowest scoring combination was the combination of unigram, MeSH, and UMLS CUI (Unified Medical Language System Concept Unique Identifier). Among the two feature combinations, the combination of unigram and MeSH showed the best score, and the combination of unigram and UMLS CUI showed the worst score. They also compared the classification performance of each feature. MeSH performed the best, and UMLS CUI performed the worst. They concluded that the MeSH feature was essential for good performance.

Kilicoglu and colleagues [14] evaluated various ML techniques and feature sets to recognize scientifically rigorous research evidence. They showed that combining commonly used classifiers (Naïve Bayes, SVM, boosting) and disparate features in various ways using stacking improved recognition of rigorous studies. They also found that manually assigned metadata like MeSH and publication types improved classification effectiveness.

If an SR already is created in a given topic, a set of associated included/excluded article judgments accumulates. These judgments can serve as input to an ML algorithm for updating the SR. However, when an SR of another topic is first created, no data for training the ML algorithm are available. To solve this problem, Cohen and colleagues [6] proposed a method that creates a model by training on the data from a combination of other SR topics that already has a base collection of included/excluded article judgments. They compared this method with three systems, a baseline system using only topic-specific training data obtaining from included/excluded article judgments in associated SRs, a non-topic system using only the non-topic data sampled from other SR topics, and a hybrid system combining topic-specific training data with data from other SR topics. On average, the hybrid system improved the mean AUC over the baseline system by 20% when topic-specific training data were scarce. In addition, the system performed better than the non-topic system on all but the two smallest fractions of topic-specific training data. However, with very sparse topic-specific training data, the performance of the non-topic system on individual topics is often better than the baseline system and is, at times, better than that of the hybrid system.

3. Methods

We present our methods in three parts. In the first, we describe the data set used to evaluate our system. In the second, we show the classifier system. Finally, we describe our evaluation process.

3.1. Data collection

3.1.1. Procedure data sets

In this study, the procedure data corpus was collected, which was based on SR inclusion/exclusion judgments of the expert reviewers of the nHTA center. We analyzed criteria to judge inclusion or exclusion articles of 126 SRs and classified common and topic-specific ones. Common criteria are the same judgment criteria across all SRs. The criteria are shown in Table 1. Common exclusion criteria were as follows: gray literature (i.e., conference papers), non-original articles (i.e., review articles, editorials, letters, and opinion pieces), non-human (animal) articles, and pre-clinical studies.

Among 126 SRs performed by the nHTA, we selected 19 procedure SRs with more than 10 inclusion articles for this experiment. Table 2 shows the 19 review topics with the number of articles included and excluded in each study. The number of articles excluded by the common exclusion criteria (code 1-4) is shown in parenthesis. The number before parenthesis means number of articles excluded by all exclusion criteria (code 1-5).

3.1.2. Drug data sets

We also used publicly available drug data to confirm our method [15]. Table 3 gives information on the inclusion/exclusion criteria of the drug topics [16]. We selected codes 8 and 9 as common exclusion criteria across all drug SRs because we thought that background articles (code 8) might be non-original articles (i.e., review articles, editorials, letters, and opinion pieces) and that articles with only the abstracts available (code 9) might be gray literature (i.e., conference papers). In the drug sets, the number of common exclusion articles excluded by codes 8 and 9 were small because most articles were excluded by code E. In this study, we selected four topics (*Atypical antipsychotics, Beta blockers, Calcium channel blockers, Urinary incontinence*) with more than 10 common exclusion articles. We also separated common exclusion articles (code 8-9) from exclusion articles (code E-9) (Table 4).

3.1.3. Training and test sets

As shown in Fig. 1, our data sets consisted of a small number of inclusion articles (I) and a large number of exclusion articles (E = TE + CD). This class imbalance issue corresponds to domains for which one class is represented by a large number of examples, whereas the other is represented by only a few [17]. This problem causes a significant bottleneck in the performance attainable by standard learning methods, which assume a balanced distribution of the classes [18]. The classifiers, which are trained on the imbalanced data set, tend to have a bias toward the majority class data because ML techniques cannot work well with such data for building an accurate classifier [19].

Table 1
Article triage decisions in procedure topics.

Code	Criteria	Article type
1	Excluded due to gray literature	Common exclusion article
2	Excluded due to non-original articles	Common exclusion article
3	Excluded due to non-human articles	Common exclusion article
4	Excluded due to pre-clinical studies	Common exclusion article
5	Excluded due to topic-specific reasons	Topic-specific exclusion article

Table 2
Number of articles included and excluded across 19 procedure systematic review topics.

Topics	Included	Excluded (com[a])	Total
Auditory brainstem implant	14	156(46)	170
Autologous noncultured epidermal cellular transplantation	18	126(23)	144
Continuous intraarticular pain control	22	742(38)	764
Endoscopic cryotherapy of lung tumors	14	334(172)	348
Glaucoma aqueous tube insertion	10	500(102)	510
Hand transplantation	10	227(113)	237
Holmium laser treatment of benign prostatic hyperplasia	34	155(93)	189
Impedance controlled endometrial ablation	11	55(22)	66
Intrastromal corneal ring surgery for keratoconus	31	140(26)	171
Magnetic navigation assisted catheter technique	14	365(86)	379
Radiofrequency ablation of primary and secondary lung malignancy	18	506(192)	524
Small bowel transplantation	27	911(184)	938
Somatic nerves stimulation	12	378(42)	390
Surgical ablation of atrial fibrillation	13	185(66)	198
Therapeutic temperature management with endovascular catheters	16	293(62)	309
Therapeutic use of autologous bone marrow cells in peripheral arterial disease	28	249(143)	277
Transanal endoscopic microsurgery	10	246(43)	256
Transarterial radioembolization	32	473(156)	505
Trigeminal nerve stimulation	11	730(50)	741
Totals	345	6,771(1,659)	7116

[a] The number of common exclusion articles excluded by the common exclusion criteria among excluded articles excluded by all exclusion criteria.

Table 3
Article triage decisions in drug topics.

Code	Criteria	Article type
E	Nonspecifically excluded	Exclusion article
1	Excluded due to foreign language	Exclusion article
2	Excluded due to wrong outcome	Exclusion article
3	Excluded due to wrong drug	Exclusion article
4	Excluded due to wrong population	Exclusion article
5	Excluded due to wrong publication type	Exclusion article
6	Excluded due to wrong study design	Exclusion article
7	Excluded due to wrong study duration	Exclusion article
8	Excluded due to background article	Common exclusion article
9	Excluded due to only abstract being available	Common exclusion article

Table 4
Number of articles included and excluded across 4 drug systematic review topics.

Topics	Included	Excluded (com[a])	Total
Atypical antipsychotics	146	974(11)	1120
Beta blockers	42	2030(104)	2072
Calcium channel blockers	100	1118(25)	1218
Urinary incontinence	40	287(21)	327
Totals	328	4409(161)	4737

[a] The number of common exclusion articles excluded by the common exclusion criteria among excluded articles excluded by all the exclusion criteria.

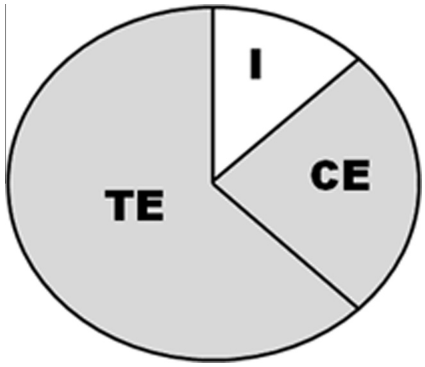

Fig. 1. Article collection diagram. Our collection consists of inclusion articles (I), topic-specific exclusion articles (TE), and common exclusion articles (CE). Each area depicts the proportion of article types.

In order to build the balanced training sets, we proceeded in two phases. In the first phase we randomly selected exclusion articles with the same number of inclusion articles regarding each topic. Exclusion articles were sampled five times with replacement for the future five times of the evaluation test. They were combined with inclusion articles to form the balanced data set. In the second phase, in order to make a training set we combined all of the non-topic balanced data (Fig. 2). For example, regard to *Auditory brainstem implant*, we trained our model using the combined balanced data set which consist of articles not related to the *Auditory brainstem implant*.

We made two kinds of the balanced training sets for each topic; the first set (IE training set) has inclusion articles (I) and exclusion articles randomly selected from total exclusion articles (E), and the second set (ICE training set) consists of inclusion articles (I) and common exclusion articles randomly selected from CE.

We used articles of each topic as test sets. Topics of each test set are shown in Table 2 and 4. We also made two kinds of test sets for each topic; one set (IE test set) has inclusion (I) and total exclusion articles (E), and the other set (ICE test set) consists of inclusion (I) and common exclusion articles (CE). For example, to classify *Auditory brainstem implant*, the IE test set has 170 articles (14 inclusion and 156 exclusion articles) and ICE test set has 60 articles (14 inclusion and 46 common exclusion articles).

3.2. Classifier System

We used four basic feature types, listed as follows:

- Words in the titles of a MEDLINE citation.
- Words in the abstracts of a MEDLINE citation.
- Medical Subject Headings (MeSH) indexing terms from a MEDLINE citation.
- Publication types assigned manually by NLM indexers.

The titles and abstracts were parsed into tokens. MeSH indexing terms and publication types were encoded as phrases. Individual words of the titles and abstracts were further processed by the

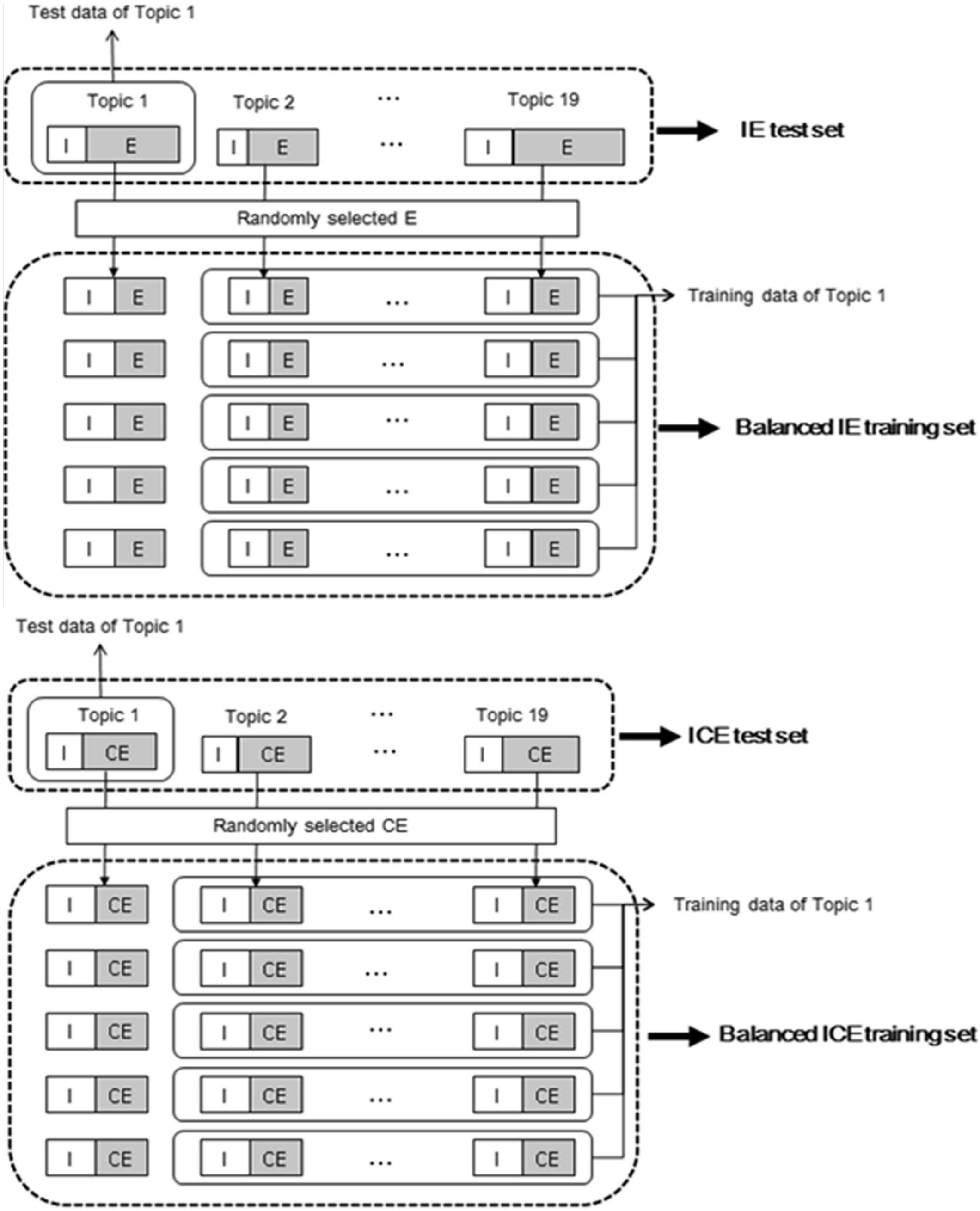

Fig. 2. Building the two kinds of training set and test set.

removal of stop words such as ‘the’, ‘an’, and ‘other’, which are not likely to add semantic value to the classification [20]. The words were also stemmed by the Porter stemming algorithm [21].

As titles and abstracts are narrative texts, the frequency-based representation is more appropriate. Conversely, because MeSH indexing terms and publication types do not occur in an article more than once, the binary representation method might be more suitable for the feature types [4]. Therefore, we represented the titles and abstracts by word frequencies and the MeSH indexing terms and publication types as binary.

In some tasks, it made sense to give some features weights greater than those of other features [4]. However, we did not weigh the features by term frequencies because, in an earlier study, the authors found that weighting features by intra-document frequency and/or inverse document frequency (TF, IDF, TFIDF) decreased the performance of the classifier system [16].

The ML system presented here was motivated by interesting results observed in earlier studies on the best evidence for SRs [2,12–14]. The authors found that using the SVM rather than other ML techniques led to improved classification performance. In the present work, our basic ML system was the SVM^{light} [22] implementation of the SVM algorithm, with linear kernel and default settings [23].

3.3. Evaluation

We evaluated how well our classifiers, which were trained on a combination of inclusion and common exclusion articles, classified rigorous articles for new procedure or drug SRs. We performed three experiments.

- Experiment 1: IE training set + IE test set.

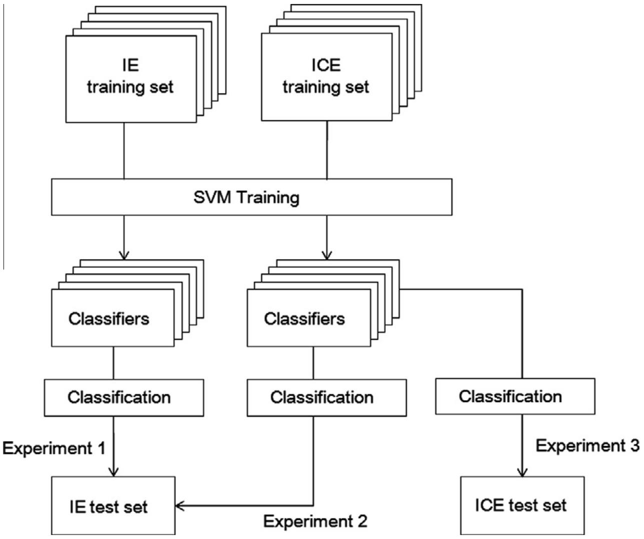

Fig. 3. Process diagram of the experimental procedure.

- Experiment 2: ICE training set + IE test set.
- Experiment 3: ICE training set + ICE test set.

All topics were tested using the same processes (Fig. 3). We used AUC as an evaluation metric. The average of five training data sets gave the final performance estimate. We applied one-way ANOVA to compare classifier performances.

As an additional experiment, we performed 10-fold cross validation tests. The goal of cross validation is measuring the generalizability of an algorithm, comparing the performance of two or more different algorithms, and finding out the best algorithm for the available data [24]. In 10-fold cross validation, we used IE and ICE data sets. A single subsample is used as the validation data for testing the model, and the remaining nine subsamples are used as training data. We applied independent *t*-test to compare the results of IE and ICE data sets.

4. Results

In order to evaluate the performance of the classifiers, we made receiver operating curves. The area under the curve (AUC) was calculated for the Procedure with inclusion and exclusion/common exclusion articles. The same evaluation was done using AUCs for the Drug with inclusion and exclusion/common exclusion articles. Table 5 shows the performances using 19 Procedure data by topics. The mean AUC of the Experiment 1 was 0.81 (range 0.66–0.91), Experiment 2 was 0.83 (range 0.70–0.90), and Experiment 3 was 0.95 (range 0.89–1.00). The performances of the classifiers trained on the Procedure with inclusion and common exclusion articles (Experiments 2, 3) were significantly higher than those of the classifiers trained on the Procedure with inclusion and exclusion articles (Experiment 1) ($p < 0.05$). The classifier showed the

Table 5
Mean AUCs of the three experiments across 19 procedure systematic reviews.

Topics	Experiment 1	Experiment 2	Experiment 3
Auditory brainstem implant	0.80 (0.78–0.81)	0.82 (0.81–0.82)	0.89 (0.88–0.90)
Autologous noncultured epidermal cellular transplantation	0.82 (0.74–0.86)	0.80 (0.78–0.82)	0.91 (0.90–0.92)
Continuous intraarticular pain control	0.75 (0.68–0.80)	0.76 (0.70–0.78)	0.92 (0.91–0.93)
Endoscopic cryotherapy of lung tumors	0.84 (0.83–0.86)	0.84 (0.84–0.85)	0.95 (0.94–0.96)
Glaucoma aqueous tube insertion	0.67 (0.62–0.70)	0.70 (0.68–0.72)	0.97 (0.96–0.97)
Hand transplantation	0.85 (0.81–0.88)	0.87 (0.86–0.88)	0.96 (0.95–0.97)
Holmium laser treatment of benign prostatic hyperplasia	0.82 (0.79–0.85)	0.84 (0.83–0.85)	0.96 (0.96)
Impedance controlled endometrial ablation	0.66 (0.62–0.72)	0.74 (0.73–0.75)	0.95 (0.93–0.95)
Intrastromal corneal ring surgery for keratoconus	0.91 (0.88–0.93)	0.90 (0.88–0.93)	0.98 (0.97–1.00)
Magnetic navigation assisted catheter technique	0.89 (0.86–0.92)	0.90 (0.88–0.90)	1.00 (0.99–1.00)
Radiofrequency ablation of primary and secondary lung malignancy	0.78 (0.76–0.81)	0.79 (0.77–0.80)	0.97 (0.97–0.98)
Small bowel transplantation	0.88 (0.85–0.90)	0.87 (0.85–0.88)	0.94 (0.93–0.95)
Somatic nerves stimulation	0.81 (0.79–0.82)	0.80 (0.78–0.81)	0.93 (0.91–0.96)
Surgical ablation of atrial fibrillation	0.83 (0.76–0.90)	0.88 (0.88–0.89)	0.98 (0.98)
Therapeutic temperature management with endovascular catheters	0.88 (0.85–0.93)	0.87 (0.86–0.88)	0.96 (0.95–0.96)
Therapeutic use of autologous bone marrow cells in peripheral arterial disease	0.86 (0.86–0.88)	0.86 (0.85–0.87)	0.91 (0.90–0.92)
Transanal endoscopic microsurgery	0.68 (0.66–0.71)	0.75 (0.72–0.77)	0.98 (0.97–0.98)
Transarterial radioembolization	0.84 (0.81–0.86)	0.87 (0.86–0.87)	0.98 (0.98)
Trigeminal nerve stimulation	0.86 (0.85–0.88)	0.84 (0.83–0.85)	0.95 (0.94–0.95)
Mean	0.81 (0.66–0.91)	0.83 (0.70–0.90)	0.95 (0.89–1.00)

The range of AUC is shown in parenthesis.

Table 6
Mean AUCs of the three experiments across 4 drug systematic reviews.

Topics	Experiment 1	Experiment 2	Experiment 3
Atypical antipsychotics	0.75(0.75–0.77)	0.73(0.72–0.74)	0.90(0.88–0.91)
Beta blockers	0.82(0.79–0.85)	0.78(0.76–0.83)	0.83(0.81–0.85)
Calcium channel blockers	0.71(0.69–0.74)	0.56(0.54–0.57)	0.77(0.76–0.87)
Urinary incontinence	0.83(0.80–0.85)	0.87(0.86–0.87)	0.87(0.86–0.87)
Mean	0.78(0.71–0.83)	0.73(0.56–0.87)	0.84(0.77–0.90)

The range of AUC is shown in parenthesis.

best performance in the Experiment 3, in which common exclusion articles were used in training set and test set both.

Table 6 indicates that the performances using 4 Drug data by topics. The mean AUC of the Experiment 1 was 0.78 (range 0.71–0.83), Experiment 2 was 0.73 (range 0.56–0.87), and Experiment 3 was 0.84 (range 0.77–0.90). The performances of the classifiers trained on the Drug with inclusion and common exclusion articles (Experiment 3) were significantly higher than those of the classifiers trained on the Drug with inclusion and exclusion articles (Experiments 1, 2) only if test data consisted of inclusion and common exclusion articles ($p < 0.05$).

The 10-fold cross validation results showed that the mean AUC of IC data was 0.93 with a standard deviation of 0.02 and ICE data was 0.97 with a standard deviation of 0.01 in Procedure data sets. In Drug data sets, the mean AUC of IC data was 0.84 with a standard deviation of 0.03 and that of ICE data was 0.92 with a standard deviation of 0.05. In both of Procedure and Drug SRs, the results of ICE data which used common exclusion articles were significantly better than those of IC data ($p < 0.05$).

5. Discussion

Our results show that AUCs of the classifiers trained on the Procedure with common exclusion articles were significantly higher than those of the classifiers trained on the Procedure with exclusion articles. In Drug, the AUCs of the classifiers trained with common exclusion articles were significantly higher than those of the classifiers trained with exclusion articles when test data consisted of inclusion and common exclusion articles.

When we performed the cross validation experiments, the AUCs of the classifiers showed slightly better results than the previous experiments. For instance the mean AUC of the Experiment 1 was improved from 0.81 to 0.93 in the case of Procedure data set. The reason for the good performance in cross validation test might be understood that in cross validation experiment, some topics in test data were included into the training data sets. As the SVM classifier was already trained with a specific topic, so it would classify better. However, in the real world, above situation happens rarely. There will be many topics that are not systematically reviewed, in which our method can be helpful.

We focused on reducing the labor required for capturing high-quality articles to the new SRs that are not systematically reviewed. If experts performing SRs use our classifiers after manually filtering out topic-specific exclusion articles, their workload of literature review will be reduced. Further research is required an extended learning system that classifies topic-specific exclusion articles of the new SRs.

Our evaluation has several limitations. The sample sizes were small. Although the data corpus included 19 topics and expert judgments, there were approximately 7200 articles overall. We used the data generated by an SR-producing organization, which is a limitation even if the nHTA uses the most rigorous processes to maximize quality and consistency. We tried to confirm our method using drug SRs generated by DERP (Drug Evidence Review Project) [6], but we could not properly evaluate our method with those articles because, in drug SRs, most articles were classified as E (nonspecifically excluded).

6. Conclusion

Our research has provided that SVM-based high-quality article classifiers can support new SRs by reducing the labor required to experts reviewing journal articles for inclusion. We focused on data in training sets to improve the performance of classifiers and showed the method of constructing training sets. The performance of classifiers, which were trained on a combination of inclusion and common exclusion articles, were significantly higher than those of the classifiers, which were trained on other data.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2010-0028631).

References

[1] Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996;312:71–2.
[2] Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models for high-quality article retrieval in internal medicine. J Am Med Inform Assoc 2005;12:207–16.
[3] Lewis SJ, Orland BI. The importance and impact of evidence-based medicine. J Manag Care Pharm 2004;10:S3–5.
[4] Matwin S, Kouznetsov A, Inkpen D, Frunza O, O'Blenis P. A new algorithm for reducing the workload of experts in performing systematic reviews. J Am Med Inform Assoc 2010;17:446–53.
[5] The Cochrane Library [cited 07.08.12]. <http://www.thecochranelibrary.com/view/0/AboutCochraneSystematicReviews.html>.
[6] Cohen AM, Ambert K, McDonagh M. Cross-topic learning for work prioritization in systematic review creation and update. J Am Med Inform Assoc 2009;16:690–704.
[7] nHTA [cited 07.08.12]. <http://nhta.or.kr/nHTA/english/>.
[8] Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med 2010;7:e1000326.
[9] Mallett S, Clarke M. How many Cochrane reviews are needed to cover existing evidence on the effects of health care interventions? ACP J Club 2003;139:A11.
[10] Gg K, editor. No improvement – still less than half of the Cochrane reviews are up to date. Dublin, Ireland: XIV Cochrane Colloquium; 2006.
[11] Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet 2009;374:86–9.
[12] Cohen AM. Optimizing feature representation for automated systematic review work prioritization. AMIA Annu Symp Proc 2008:121–5.
[13] Joachims T. Text categorization with support vector machines: learning with many relevant features. In: Proceedings of the 10th European conference on machine learning; 1998. p. 137–42.
[14] Kilicoglu H, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes RB. Towards automatic recognition of scientifically rigorous clinical research evidence. J Am Med Inform Assoc 2009;16:25–31.
[15] Systematic drug class review gold standard data [cited 07.08.12]. <http://davinci.ohsu.edu/~cohenaa/systematic-drug-class-review-data.html>.
[16] Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. J Am Med Inform Assoc 2006;13:206–19.
[17] Tom F, Foster P. Adaptive fraud detection. Data Mining Knowledge Discov 1997;3:291–316.

[18] Japkowicz N. The class imbalance problem: significance and strategies. In: Proceedings of the 2000 international conference on artificial intelligence; 2000. p. 111–7.

[19] Yin HL, Leong TY. A model driven approach to imbalanced data sampling in medical decision making. Stud Health Technol Inform 2010;160:856–60.

[20] Onix Text Retrieval Toolkit [cited 07.08.12]. <http://www.lextek.com/manuals/onix/stopwords1.html>.

[21] Porter MF. An algorithm for suffix stripping. Program 1980;14:130–7.

[22] SVMlight [cited 07.08.12]. <http://svmlight.joachims.org/>.

[23] Joachims T. Making large-scale SVM learning practical, advances in kernel methods-support vector learning, Cambridge, MA, USA; 1999.

[24] Refaeilzadeh P, Tang L, Liu H. Cross-validation, encyclopedia of database systems; 2009. p. 532–8.