Improvement of the clustering technique to classify medicines based on indications or efficacies

Kouta Katano\textsuperscript{a,}\*, Masaomi Kimura\textsuperscript{b} Michiko Ohkura\textsuperscript{b}, Fumito Tsuchiya\textsuperscript{c}

\textsuperscript{a}Graduated School, Shibaura Institute of Technology, 3-7-5 Toyosu Koto-ku, TOKYO 135-8548, JAPAN
\textsuperscript{b}Shibaura Institute of Technology, 3-7-5 Toyosu Koto-ku, TOKYO 135-8548, JAPAN
\textsuperscript{c}International University of Health and Welfare, 2600-1 Kitakanemaru Ohtawara, TOCHIGI 324-8501, JAPAN

Abstract

A current Japanese standard for drug efficacy classification is comprised of words from both pharmaceutical and external fields. A previously reported method was based on clustering of descriptions in medical package inserts for circulatory drugs. In this report, a new criteria is proposed to extract clusters at each junction of a dendrogram relative to the number of drugs and the nouns in the drug description. Results of a comparative analysis indicate that the proposed method provides more detailed classification than the previous method.

© 2014 The Authors. Published by Elsevier Ltd.

Keywords: clustering, dendrogram, medical package insert, indication or efficacy

1. Introduction

In Japan, a variety of drug efficacies are classified by therapeutic category and designated with a three-digit number. This number is based on the \textit{Japan Standard Commodity Classification} \cite{1}. In recent years, doctors have utilized a computerized drug ordering system. The development of a crosschecking method for drug ordering is important to prevent incorrect prescription input. We expect that drug classification based on efficacy can be crosschecked with the drug order because the efficacy information tightly connects with disease conditions.

\* Corresponding author. Tel.: +81-3-5859-8507.
\textit{E-mail address:} ma13027@shibaura-it.ac.jp
However, the classification includes non-efficacy viewpoints; therefore, the crosscheck might function inappropriately. The World Health Organization [2] recommends the Anatomical Therapeutic Chemical Classification System, which is a classification method for drugs based on the consumption statistics. However, it is classified from both efficacy and non-efficacy viewpoints [3]. Therefore, this classification system is irrelevant for crosschecking purposes.

Ishida et al. [4] proposed a new classification method based on nouns that appear in the description of medical package inserts of circulatory drugs (over 1,300 documents). They reported that the method is useful for circulatory drugs; however, they did not clarify whether the method is useful for all medical drugs. Additionally, it is not clear how to determine a threshold value in their method (see below for details). Our aim is to improve their classification method based exclusively on drug efficacy and to develop a method that is applicable to all ethical drugs (over 10,000 documents). The remainder of this paper is organized as follows: First, Section 2 describes how we treated target data in this study. Section 3 reviews the method of classification and results presented by Ishida et al. [4]. Section 4 describes our method and results obtained by applying the proposed method application to all ethical drugs. In Section 5, we evaluate our method, and in Section 6, we present our conclusions and suggestions for future work.

2. Target data

We analyzed 9,358 efficacy documents from 10,054 medical package inserts registered in Standard Generalized Markup Language (SGML) format by the Pharmaceuticals and Medical Devices Agency, Japan [5]. Nabeta et al. [6,7,8,9,10] proposed a method to extract data from medical package inserts in SGML format, and we have extracted data from efficacy documents using the same method. Each efficacy document is described in Japanese. In this paper, we have translated Japanese text to English to facilitate easy comprehension.

3. Classification by the previous study

3.1. Noun vector

The efficacy documents frequently include specific nouns, such as names of diseases and symptoms; therefore, Ishida et al. [4] extracted nouns by means of morphological analysis [11]. They utilized a noun vector to obtain a set of concepts contained in the efficacy document for each drug. The noun vector has a value of 1 for an element if the corresponding word appears in the document and has a value of 0 if it does not appear. Table 1 shows an example of the noun vectors for the specific efficacy phrase, “頭部外傷に伴う意識障害” (consciousness disturbance induced by external head injury). As the nouns in the third and fifth columns do not appear in that phrase, values of 0 are assigned to them. Ishida et al. [4] obtained a dendrogram based on the distance between the noun vectors given by Euclidean distance and the cluster distance measured by the Ward method [12,13].

Ishida et al. [4] proposed a division index to determine which cluster is extracted from the dendroagam. The division index of a cluster, \( I_c = N_c \left( E_{before} - E_{after}(c) \right) \), is calculated at each junction of the dendrogram and exhibits the increase of noun cohesion in efficacy among drugs caused by the cluster division. The number \( N_c \) counts how many nouns are in the cluster \( c \) after division. The vale \( E_{before} \) is the entropy of the cluster before division (i.e., the degree of word assembly in the cluster), and \( E_{after}(c) \) is the entropy of the cluster \( c \) after division. The entropy \( E \) is given as follows:
\[ E = -\sum_{w=1}^{N} q_w \{ P_w^c \log P_w^c + (1 - P_w^c) \log(1 - P_w^c) \}, \]

\[ P_w^c = \frac{N_w^c}{N_c}, \quad q_w = \frac{n_w}{n_{all}}, \]

where \( N_w^c \) is the number of drugs in the cluster \( c \) including noun \( w \) in the efficacy documents, \( n_{all} \) is the total number of nouns appearing in the efficacy documents in the cluster before division, and \( n_w \) is the total number of noun \( w \) in the efficacy documents. If the value of \( [E_{before} - E_{after}(c)] \) is a large value, the division decreases the variation of words in the cluster.

They gave a condition to extract the clusters with \( I_c \) greater than threshold \( \alpha \):

\[ I_c = N_c \left( E_{before} - E_{after}(c) \right) > \alpha. \]

They applied the abovementioned equations to the clustering of circulatory drugs. Figure 1(a) plots \( I_c \) for the drugs. In the figure, the blue dots represent the cluster of the larger number of drugs after division, and the red dots are smaller clusters of drugs after division. The division index \( I_c \) of each cluster is on the vertical axis, and the horizontal axis shows the ratio \( M \) of the number of drugs in the cluster before division to the total number of drugs.

We can see that large \( I_c \) values in Fig. 2(a) appear in both the higher and lower ratio regions of the drugs. The latter case includes clusters that are not extracted even though \( I_c \) is greater than \( \alpha \). There were no obvious criteria to exclude these clusters from the extraction and determine the value of \( \alpha \).

4. Improvement of the classification method

4.1. The relationship between the number of drugs and words

Figure 2 shows the dendrogram of the clustering method applied to efficacy documents of all ethical drugs in Japan.

Naive application of Eq. (2) to the dendrogram in Fig. 2 gives Fig. 1(b), which shows a relationship between \( I_c \) and the ratio \( M \) of the number of drugs similar to that shown in Fig. 1(a). To investigate the results in detail, we also plotted the distribution of \( (E_{before} - E_{after}(c)) \) and the distribution of \( N_c \) on the drug number ratio \( M \) (Fig. 3, Fig. 4).

Figure 3 shows that \( (E_{before} - E_{after}(c)) \) tends to be large for a small cluster \( c \). This result is easy to understand because a small cluster has a small number of words with small variation.

In Fig. 4, we can observe large \( N_c \) at the lower \( M \) region. It should be noted that \( N_c \) is introduced in \( I_c \) to reduce the small size effect of clusters on \( (E_{before} - E_{after}(c)) \). Ishida et al. [4] expected that \( N_c \) would be small for a small cluster \( c \), and multiplied it by the entropy difference to overcome this difficulty. Their assumption is right for the relatively higher \( M \) region. However, Fig. 3 suggests that their assumption is wrong in the lower \( M \) region. This suggests that there is a new mechanism that affects the lower region.

To investigate change in the number of words against the ratio of the number of drugs in detail, we transformed Fig. 4 into a bi-logarithmic form. Figure 5 shows the relationship of the number of drugs and words in bi-logarithmic form. The horizontal axis represents \( \log M \), where \( M \) is the ratio of the number of drugs in the cluster before division to the total number of drugs. The vertical axis represents \( \log N_c \), where \( N_c \) is the same as in Eq. (2).

In Fig. 5, we can see that there are two parts in this graph, the downward slope from the point of \( \log M = 0 \) and the region with crowded dots. The right part of the slope and the crowded dots form straight lines. This indicates that \( N_c \) and \( M \) satisfy the relationship:

\[ \log N_c = \rho \log M + \log D, \]

where \( \rho \) and \( D \) are constants. This leads to the power law equation, \( N_c = DM^\rho \). The right part of the slope and the crowded dots have approximately the same \( \rho \) value.
Fig. 1. (a) Division index of each cluster for circulatory drugs; (b) Division index of each cluster for all of medical drugs.

Fig. 2. Dendrogram applied to efficacy documents of all ethical drugs in Japan.

Fig. 3. The relationship of the ratio of the number of drugs and the entropy difference.
Fig. 4. The relationship of the ratio of the number of drugs and the number of words.

Fig. 5. The bi-logarithmic graph transformed from Fig.4.

Fig. 6. Schematic dendrogram representing the corresponding relationship between clusters in the dendrogram and lines in Fig.5.
Figure 6 illustrates how the clusters in the dendrogram are divided. At each junction, a cluster is divided into two sub-clusters denoted by a blue circle and a red square. The blue circle indicates a larger sub-cluster, and the red square indicates a smaller sub-cluster. We confirmed that the sequence of sub-clusters marked with blue circles (surrounded by a green dotted line) corresponds to the downward slope in Fig. 5. Each branch of sub-clusters marked with blue circles starting from a sub-cluster with red square circles (surrounded by red dotted lines) corresponds to a line of crowded dots.

These facts suggest that $N_\alpha$ has a fractal feature, i.e., cluster division corresponding to the slope and the lines of crowded dots share the same mechanism to decompose clusters independently from the size of clusters.

### 4.2. Extraction of the clusters

The above consideration suggests that smaller sub-clusters derived from larger clusters on the slope give us the sets of drugs whose efficacies are similar because such smaller sub-clusters have small variation of words and should be coherent in terms of efficacy descriptions. Moreover, further divisions of such sub-clusters can result in fragments with the same efficacies. The junctions corresponding to the right part of the red line in Fig. 5 give the required sub-clusters. With reference to Eq. 2, this condition is equivalent to $\alpha = 7$.

The threshold $\alpha = 7$ provides 26 clusters for all ethical drugs in Japan, as shown in Table 2. The primary efficacies in Clusters 1 to 26 do not overlap significantly. Cluster 26, which represents 7% of the total number of drugs, is comprised of miscellaneous efficacies. We should also note that 60% of all drugs were classified into the 26 clusters. The remaining 40% are insufficiently classified to be denoted as “a mixed cluster” (not indicated in Table 2).

<table>
<thead>
<tr>
<th>Cluster No.</th>
<th>Main efficacy</th>
<th>Cluster No.</th>
<th>Main efficacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>nutrient supplement</td>
<td>14</td>
<td>cold medicine</td>
</tr>
<tr>
<td>2</td>
<td>analgesic, antiphlogistic</td>
<td>15</td>
<td>allergic, vasomotor nerve</td>
</tr>
<tr>
<td>3</td>
<td>eczema, dermatitis</td>
<td>16</td>
<td>depression</td>
</tr>
<tr>
<td>4</td>
<td>bronchitis, asthma</td>
<td>17</td>
<td>anesthetic</td>
</tr>
<tr>
<td>5</td>
<td>antiseptic, dermatomycosis</td>
<td>18</td>
<td>abnormal digestion</td>
</tr>
<tr>
<td>6</td>
<td>anorexia, nausea</td>
<td>19</td>
<td>cataract</td>
</tr>
<tr>
<td>7</td>
<td>hypertension</td>
<td>20</td>
<td>insomnia, mental instability</td>
</tr>
<tr>
<td>8</td>
<td>herpes simplex, metabolic disorder of Vitamin D</td>
<td>21</td>
<td>pain</td>
</tr>
<tr>
<td>9</td>
<td>epilepsy, diabetes</td>
<td>22</td>
<td>solvents</td>
</tr>
<tr>
<td>10</td>
<td>prescription</td>
<td>23</td>
<td>ulcerative</td>
</tr>
<tr>
<td>11</td>
<td>constipation</td>
<td>24</td>
<td>schizophrenia</td>
</tr>
<tr>
<td>12</td>
<td>respiratory, dermatomycosis</td>
<td>25</td>
<td>glaucoma, ocular hypertension</td>
</tr>
<tr>
<td>13</td>
<td>supplicative, ophthalmic</td>
<td>26</td>
<td>miscellaneous</td>
</tr>
</tbody>
</table>

### 5. Evaluation

To evaluate the improvement of the proposed method, we applied it to the data from the efficacy documents for circulatory drugs used in the previous study [4].

The results presented in Table 3 were obtained under the condition that the threshold $\alpha = 42$. Using the proposed method, we obtained 22 clusters; contrastingly, in the previous study, only 13 clusters were obtained. In Table 3, the clusters in the red column (Clusters 1 to 11) were extracted using both methods.

One cluster extracted using the previous method was further divided into four clusters by our proposed method: Clusters 17 and 20 to 22 (blue). These four clusters included the word “hypertension.” Consequently, the previous
result put them in one cluster because it treated “hypertension” in a broad sense. The proposed method shed light on pharmaceutically different efficacies in a narrower sense: “hypertension and pulmonary arterial hypertension” in Cluster 21, “hypertension and angina” in Cluster 22, and so on.

A mixed cluster in the previous results was divided into seven different clusters: Clusters 12−16 and Clusters 18−19 (green). Each cluster has a specific description in addition to the common word “hypertension”:

- “bronchial asthma” in Cluster 14, “congestive heart failure” in Cluster 16, and so on. This suggests that the proposed method generates more detailed clusters than the previous method.

Table 3. Efficacies of drugs included in each cluster for circulatory drugs.

<table>
<thead>
<tr>
<th>Cluster No.</th>
<th>Main efficacy</th>
<th>Cluster No.</th>
<th>Main efficacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>disturbance of consciousness through head injury</td>
<td>12</td>
<td>thrombus, embolus, nephrosis</td>
</tr>
<tr>
<td>2</td>
<td>hypertension, chronic cardiac failure</td>
<td>13</td>
<td>tachyarrhythmia, arteriosclerosis obliterans</td>
</tr>
<tr>
<td>3</td>
<td>chronic arterial obstruction, ductus arteriosus</td>
<td>14</td>
<td>bronchial asthma, bronchitis</td>
</tr>
<tr>
<td>4</td>
<td>acute circulatory failure, uremic</td>
<td>15</td>
<td>hypertensive emergency</td>
</tr>
<tr>
<td>5</td>
<td>hypertension, renal edema</td>
<td>16</td>
<td>congestive cardiac failure</td>
</tr>
<tr>
<td>6</td>
<td>intracranial hypertension, brain infarction and hemorrhage</td>
<td>17</td>
<td>essential hypertension, renal hypertension</td>
</tr>
<tr>
<td>7</td>
<td>tachyarrhythmia</td>
<td>18</td>
<td>cerebral vasospasm, bronchial asthma</td>
</tr>
<tr>
<td>8</td>
<td>hyperlipemia</td>
<td>19</td>
<td>cerebral infarction sequelae</td>
</tr>
<tr>
<td>9</td>
<td>essential hypertension</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>congestive and congenital cardiac failure</td>
<td>20</td>
<td>essential hypertension</td>
</tr>
<tr>
<td>11</td>
<td>hypertension, angina, renal parenchymal hypertension</td>
<td>21</td>
<td>hypertension, pulmonary arterial hypertension</td>
</tr>
<tr>
<td></td>
<td></td>
<td>22</td>
<td>hypertension, angina</td>
</tr>
</tbody>
</table>

6. Conclusion

First, we improved a previous method based on the nouns that appear in the indications and efficacy statements in medical package inserts. The target data was comprised of 9,358 efficacy documents from 10,054 medical package inserts. The previous work by Ishida et al. [4] focused only on circulatory drugs; comparatively, our target was all ethical drugs in Japan. We extracted the nouns by morphological analysis and applied noun vectors that were used to generate a dendrogram.

Next, we examined the division index used to extract groups of drugs with the similar efficacies. As in the previous study, naive application of the division index to the dendrogram was problematic. To investigate the results in detail, we observed the distribution of \(E_{\text{before}} - E_{\text{after}}(c)\) and the distribution of \(N_c\) on the drug number ratio \(M\) and found that large \(N_c\) values are obtained at the lower \(M\) region. We also found that there are two components in the relationship between \(\log N_c\) and \(\log M\), the downward slope from the point of \(\log M = 0\) and the region with crowded dots. The right part of the slope and the crowded dots form straight lines, which indicates that \(N_c\) and \(M\) satisfy the power law, \(N_c = D M^\beta\). This, in turn, suggests that \(N_c\) has a fractal feature, and cluster division corresponding to the slope and the lines of crowded dots share the same mechanism to decompose clusters independent of cluster size. Based on this consideration, we proposed a standard to determine the threshold value in the division index. As a result, we extracted 26 clusters for all ethical drugs.

Finally, we evaluated the extent to which the proposed method has been improved by using data for circulatory drugs and comparing the results to those of the previous study. The previous method extracted 13 clusters; our improved method was able to extract 22 clusters. By comparing the content of each cluster, we determined that the proposed method provided more finely divided clusters than the previous method. Accordingly, the effectiveness of the dividing criteria used in this study was shown. In the proposed drug classification method, the introduction of classification criteria provides more precise clustering.

However, the proposed method can be improved further. In our results, particular clusters contained drugs with a variety of efficacies. In future, we will propose a method to provide classification with greater precision.
References